The Causes and Consequences of Variation in Evolutionary Processes Acting on DNA Sequences

This dissertation is submitted for the degree of Doctor of Philosophy at the University of Cambridge

Lee Nathan Marc Bofkin

Darwin College

March 2006

Acknowledgements and Declarations

Many thanks to Nick Goldman and all the members of the Goldman group, both current and former (Simon Whelan, Ari Loytynoja, Carolin Kosiol, Fabio Pardi, Nicolas Rodriguez, Irmtraud Meyer, Pietro Liò and Tim Massingham), for their advice and insight over the past few years.

This dissertation is my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university; and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified length limit of 300 single-sided pages of double-spaced text as defined by The Biology Degree Committee.

Abstract

This thesis uses mathematical models of evolutionary processes to make inferences from DNA sequences about variation in the processes of molecular evolution. Questions addressed include the differences in evolutionary processes observed at each codon position, including in overlapping reading frames; how this variation may be used to locate protein-coding open reading frames in genomic alignments; differences in evolutionary signal either side of an origin of replication; differences in evolutionary processes between regions of a chromosome and how to combine large datasets of separate genes to produce a single phylogeny and the biological factors that are important to account for when combining such data.

Chapter 1 discusses simple models of genetic change along evolutionary lineages and introduces many of the methodological principles upon which the research in this thesis is carried out. Chapter 2 uses evolutionary models to analyse the differences in evolutionary processes at each of the three codon positions of protein-coding DNA and then proceeds to study the same processes in overlapping reading frames, using the hepatitis B virus as a model organism. Chapter 3 uses the variation in the evolutionary processes at the different sites in a codon to formulate a powerful search tool for protein-coding sequences in windows across large-scale genomic alignments, which are largely non-protein-coding. Chapter 4 introduces a model of evolution that tests for the signal we might expect to observe at a eukaryotic origin of DNA replication. This is then developed into a sliding window approach, where we simultaneously move two spatially distinct windows of DNA across an alignment and test for a signal between them. Chapter 5 generalises such two-window models and observes specific differences between two sliding windows across large DNA alignments, making additional inferences about the average differences in evolutionary processes at any given distance across a genomic region alignment. Chapter 6 then addresses how spatially distinct data, such as separate aligned genes, should be combined to infer a phylogeny.

Contents

Chapter 1: Introduction

Contents

1.1 Introduction

Evolutionary studies provide an insight into how cells and genes function and how organisms adapt to their environment. The comparison of related sequences in an evolutionary context has been an effective study tool, for example identifying homologous genes within and between genomes (International Human Genome Sequencing Consortium, 2001; Mouse Genome Sequencing Consortium, 2002; Qian *et al.*, 2003; Rat Genome Sequencing Consortium, 2004; International Chicken Genome Sequencing Consortium, 2004), finding novel functional structures in large alignments (International Human Genome Sequencing Consortium, 2001; Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Consortium, 2004; International Chicken Genome Sequencing Consortium, 2004; ENCODE Project Consortium, 2004) and predicting protein structure and other aspects of molecular function (Goldman *et al.*, 1996; Suzuki and Gojobori, 1999; Yang and Bielawski, 2000; Zhu *et al.*, 2005). It is clear that several different forces of evolution act upon sequences (e.g., point mutation, selection, recombination) and the variation in these forces contains a wealth of useful and often medically relevant information in itself; as an example, we are especially keen to identify protein-encoding open reading frames since mutant forms of proteins commonly cause diseases such as cystic fibrosis (the CFTR gene; Hefferon *et al*., 2004) and thalasemia (haemoglobin genes). Our potential medical interventions depend on knowing the functions and regulation of the genes that are involved in these diseases.

The aim of this thesis is to show how mathematical models of sequence evolution can be used to understand the variation in evolutionary processes and,

further, how this information can be exploited to investigate both specific features and general trends within large genomic regions (specifically, aligned homologous regions from multiple species). Of course, this is very much a process of reciprocal illumination: we develop models to understand features of our data and we use data to test and improve our models. This first chapter begins by introducing some of the basic modelling techniques that are developed throughout this thesis with respect to substitution models at the level of nucleotides, codons or amino acids. The concept of likelihood is introduced as a framework for investigating variation of evolutionary processes. Hypothesis testing is discussed and three techniques for performing tests are explained: the Likelihood Ratio Test (LRT) using χ^2 approximations, parametric bootstrapping by simulation and non-parametric bootstrapping.

1.2 Evolutionary Models

The field of bioinformatics has undergone an unprecedented expansion in recent years due to the number of sequenced genomes that have been made publicly available, including for mammals (Birney *et al.*, 2004) and yeast (Cherry *et al.*, 1997). Whilst there is still a paucity of fully aligned chromosomes that are publicly available, there has been an increase in the numbers of large 'genomic-scale' alignments available (e.g., ENCODE Project Consortium, 2004) and small alignments available with provided phylogenies (Whelan *et al.*, 2006). The genomic-scale alignments serve as test datasets for the types of analyses that may be performed when entire aligned chromosomes become more widely available. Evolutionary models that consider sequence evolution across a phylogeny can provide more information than simple visual sequence comparison alone and phylogeny-based studies can do better than

even complex methods applied only to a single sequence. For example, Boffelli *et al.* (2003) were able to identify regions of high conservation among primates with fewer species than a simple visual sequence comparison could have allowed, after calculating which species were the most phylogenetically informative. Some evolutionary models also have appealing statistical properties that are well characterised, allowing us to test between competing evolutionary hypotheses for example and simultaneously assessing how confident we are in the reliability of our results; this is covered in more detail later in this chapter. Evolutionary analyses are equally relevant to both small and extremely large datasets and this thesis hopes to contribute to the array of methods that can be used on large datasets in particular.

Evolutionary models are just that: models of evolution. Models are simplifications and are not the same as reality. When we devise an evolutionary model we are trying to capture the aspects of biology that are relevant to a particular investigation, including evolutionary relationships and the processes of evolution. The next two sections of this chapter discuss the processes that we often model when undertaking an evolutionary analysis (section 1.2.1 and 1.2.2). Substitution models, mostly at the nucleotide level but also at the codon and amino acid levels, are used throughout this thesis to infer evolutionary relationships and to test explicit hypotheses.

1.2.1 Modelling Evolutionary Relationships

The investigations in this thesis use as their base fairly standard modelling and inference methods. It would be impossible to cover all of the techniques in depth here and these techniques are discussed more fully in other texts (e.g., Fisher, 1921,1925; Jukes and Cantor, 1969; Cox and Miller, 1977; Dayhoff *et al*., 1978; Felsenstein, 1981, 2004; Hasegawa *et al*., 1985; Edwards, 1992; Jones *et al*., 1992; Efron and Tibshirani, 1993; Yang, 1993, 1994a, 1994b, 1997; Yang *et al*., 1994; Goldman and Yang, 1994; Goldman and Whelan, 2000; Swofford *et al*., 1996; Massingham, 2002). This chapter introduces key theories, techniques and terminology used in this thesis; more advanced techniques are introduced on a chapter-by-chapter basis alongside novel developments of standard techniques.

There are several ways we can consider the evolution of sites (nucleotides, codons or amino acids) in a multiple alignment, which we assume is provided from another source. The simplest approach would be to ignore the fact that the sequences in the alignment shared a common ancestry and to treat the evolution of each sequence as wholly independent, ignoring the fact that a given sequence may be more closely related to some sequences than others. If we were measuring conservation of a column in the alignment (a single aligned site across all sequences), a different site pattern in one species, say nucleotide ‘A’ in one species where all others have ‘T’, would have the same contribution to a conservation score as if any other single species had an ‘A’ instead of a ‘T’. This approach can, of course, give us some measure of conservation but does not consider that more distantly related species are more likely to have changed since they have been evolving independently for longer.

If we assume that all of our sequences in the alignment have shared a common ancestor we can further assume that sequence evolution has only been independent after the speciation events that cause one ancestral sequence to continue to evolve as two independent new species sequences (i.e., independent evolution after a split in an evolutionary tree). Evolutionary trees, referred to variously as topologies and phylogenies, can describe the statistical dependencies that exist between biological sequences, resulting from the common ancestry of the sequences. For the purposes of this thesis the evolutionary relationships between species or sequences in an alignment are usually modelled with a bifurcating tree. We can add an additional level of description of evolutionary processes by giving meaning to the branch lengths in a phylogeny, such as time or the average number of changes per site in an alignment.

Whilst not the focus of this thesis, many phylogenetic studies are primarily concerned with topology estimation amongst species or sequences. The bifurcating structure and indeed the concept of a tree impose implicit assumptions about the evolutionary process including the sharing of a common ancestor of all sequences in the tree and the independence of the evolutionary process throughout the tree. There are several ways in which these assumptions may be violated, e.g., by (1) the inclusion of paralogous regions which do not have a common ancestor and (2) the occurrence of large scale mutational events that cannot be depicted in a tree structure, such as recombination (Posada and Crandall, 2003), horizontal transfer (Doolittle, 1999) or gene conversion (Whelan and Goldman, 2004). Modelling molecular change by large scale mutational events is difficult and, whilst these events undoubtedly affect molecular evolution, we assume that it has played a minor role in the evolution

of the sequence alignments used throughout this thesis. I also assume that paralogous regions have not been included in the datasets used in this thesis. Indeed, most studies do not consider the possible violation of the assumptions implicit in bifurcating trees because violations are rare and it is difficult to make allowances for their occurrence (Felsenstein, 2004). Thus, a tree structure can be used to accurately describe the evolutionary relationships between our aligned sequences.

An optimal topology is a topology that provides the best explanation for the evolution of our data, given our type of evolutionary analysis. For example, in a parsimony study (see Page and Holmes, 1998), the optimal topology is the one (or ones) that requires the least change across each branch in the tree to proceed from an ancestral sequence to all of the descendent sequences. The optimal topology depends on the criterion of choice; in this thesis it is usually the topology that is most favoured of all possible topologies under the maximum likelihood framework described later in this chapter. It is often not possible to test all possible tree topologies to find the optimal topology because the number of possible topologies increases more than exponentially with a linear increase in the number of species. Thus, we often rely on heuristic methods to find the optimal topology although they are not guaranteed to do so (Press *et al.*, 1992; Felsenstein, 2004). Commonly, heuristic methods search through tree space using a single tree topology and making small rearrangements to find better solutions, progressively improving the tree topology until they settle upon a topology that cannot be improved upon, which we assume is the optimal topology. The effect of using a non-optimal or incorrect tree topology, which are not necessarily the same, depends on the purpose of the study.

It is generally accepted that a reasonable tree topology leads to reasonable estimates of evolutionary parameters (Yang *et al.*, 1994), which are those parts of our evolutionary mathematical models that describe a specific feature of the evolution of our dataset, such as evolutionary rate, and are derived from the dataset itself. Evolutionary parameters are described more fully in section 1.2.3.1.

1.2.2 Models of Nucleotide Substitution

Molecular change underlies our ability to study molecular phylogenetics. Whilst molecular change can be thought of in three broad categories: (1) point mutation, (2) recombination and (3) insertion-deletion events, the models used in this thesis are chiefly concerned with point mutations and do not use the information in sequence gaps in an alignment, due to either insertion or deletion ('indels'). This is a convenience; the locations and sizes of gaps in an alignment are used in some evolutionary models (e.g., McGuire *et al.*, 2001) but the use of these models is not as widespread as the models that do not utilise gap-based information. Point mutations, also called substitutions, have been studied extensively and underpin most currently popular phylogenetic models. Problematic sequence alignments are avoided and the alignments used throughout this thesis are taken to be the correct alignments, as is common practice (e.g., Grassly and Homes, 1997; Pupko *et al.*, 2002; Phillips *et al.*, 2004).

Modelling the processes of evolution requires modelling the changes that occurred between 'hypothetical' ancestor sequences and the descendent sequences that we observe. Estimating the number of changes between sequences is complicated

by the fact that multiple substitutions may occur at a single site and we cannot observe these 'multiple hits' directly. Indeed, the apparent distance between two sequences in an alignment (i.e., the number of differences between the two sequences counted by eye) is a minimum bound for the real distance between them, assuming that the alignment is correct. The greater the apparent distance between two sequences, observed by eye, the more multiple hits are likely to have occurred (Page and Holmes, 1998). With enough multiple hits a sequence can reach saturation, where sites are no longer phylogenetically informative and the probability of each base at a given site is equal to its overall equilibrium frequency within the sequence. An equilibrium frequency may not exist in our biological reality but the commonly used assumption of its existence is a mathematical convenience for the purposes of modelling. Simpler phylogenetic methods, such as parsimony, try to ignore the problem of multiple hits and the phylogeny that results in the least number of mutations required to explain the evolution of a dataset from an ancestral sequence is considered optimal (Page and Holmes, 1998). An alternative approach to parsimony, adopted in this thesis, is the construction of a model of the substitution process, which explicitly takes into account that multiple substitutions may have occurred at a single site. Thus, the models used here describe a series of random mutational events and are variously parameterised to describe the rate at which individual nucleotides replace each other.

Substitutions are commonly modelled as a Markov process (Whelan *et al.*, 2001): substitutions are assumed to occur randomly and have a constant probability across time (time homogeneity). The mutational probability of a site depends only on its current state. Changes at any one site in a sequence are assumed to be independent

of the changes and states of all other sites in the sequence. These independence assumptions are a simplification for modelling purposes and are violated in many real biological examples. Correlations in the changes between sites can be caused by selective pressures as well as structural and functional constraints (Thorne *et al.*, 1996). Modelling evolution between sites that are correlated is complicated but there has been progress within this field, especially with respect to modelling RNA sequence evolution where some sites base pair with others in the sequence (Savill *et al.*, 2001). The independence of sites assumption is necessarily made throughout this thesis. Other widely used assumptions include that the frequency of nucleotides (or amino acids) remains constant over time (stationarity) and that the evolutionary process appears the same going forward as backward (reversibility), which means that the amount of change of any nucleotide *i* to a different nucleotide *j* is equal to the amount of change of *j* mutating to *i*. The assumptions of stationarity and reversibility are computational conveniences and are standard practise within the field of molecular phylogenetics by likelihood (Felsenstein, 2004).

1.2.3 Explicit Models of Character Substitution

In this section I describe the most important areas of probabilistic modelling of substitutions, first describing substitution modelling at the level of nucleotides, then codons and then amino acids. I then proceed to describe a common method for modelling heterogeneity in the substitution rate at different sites along a sequence alignment. Modelling the effects of multiple substitutions with the probabilistic models presented below has been very successful in the field of molecular phylogenetics and the likelihood framework discussed in section 1.3.

1.2.3.1 Explicit Models of Nucleotide Substitution

The instantaneous rates of change between nucleotide *i* and nucleotide *j* can be written as q_{ij}, which is the per-nucleotide probability of *i* mutating to nucleotide *j* per unit time (an 'instant'). The amount of change between nucleotide *i* and nucleotide *j* per unit time is $\pi_i q_{ij}$, and depends on the frequency of nucleotide *i* (π_i) as well as q_{ij}. It is mathematically convenient to represent the changes between all of the characters (e.g., the four nucleotides A, C, G and T) in a matrix where each element represents the instantaneous rate of change between two given characters (e.g., A to C). Thus, an instantaneous rate matrix, or Q-matrix, represents the probability of all different point mutations occurring per unit time (i.e., for all possible states of *i* and *j*). Reading a Q-matrix is simple: the instantaneous probability of nucleotide *i* changing to *j* is represented by the value in the row labelled *i* and column labelled *j*.

Molecular phylogenetics is concerned with the probabilities of nucleotide change over longer periods of time, *t*. For a single nucleotide in state *i*, $P_{ij}(t)$ is the probability of observing state *j* after time *t* has elapsed. If we consider the Q-matrix as a whole, the probabilities of nucleotide change are obtained in the transition probability matrix, $P(t) = e^{Qt}$ (Cox and Miller, 1977). We can calculate this exponential term by decomposition of Q into its eigenvalues and eigenvectors (Felsenstein, 2004). If the time *t* is long enough, the probabilities of each element in the matrix will depend on the equilibrium frequency of the nucleotide being changed to.

There are several interesting features of Q-matrices that should be noted. Firstly, the assumption of reversibility means that $\pi_i q_{ij} = \pi_j q_{ji}$, for all *i* and *j*. Furthermore, the assumption that the instantaneous probability of a nucleotide either staying the same or mutating to another nucleotide remains constant over time means that q_{ij} remains constant for any value of *t*. The diagonals in the instantaneous rate matrix are set to the negative of the sum of the non-diagonal elements in the same row; this is mathematically convenient and facilitates matrix algebra. Where it is convenient I will omit the diagonal elements from the matrices displayed.

The general time reversible model (GTR or REV) (Yang, 1994a) is characterised by six relative rate parameters (a, b, c, d, e, f) and the four nucleotide frequency parameters (π_A, π_C, π_G and π_T) shown in the Q-matrix below:

Q	A	C	G	T
A	–	$\mu a\pi_C$	$\mu b\pi_G$	$\mu c\pi_T$
C	$\mu a\pi_A$	–	$\mu d\pi_G$	$\mu e\pi_T$
G	$\mu b\pi_A$	$\mu d\pi_C$	–	$\mu f\pi_T$
T	$\mu c\pi_A$	$\mu e\pi_C$	$\mu f\pi_G$	–

For the REV matrix above, μ represents a rate factor (see below). Only three of the nucleotide frequencies are 'free parameters', estimated from the data, since one can calculate the frequency of the remaining nucleotide when the three others are known. Similarly, one only need estimate five of the rate parameters (a-f), or 'exchangeabilities', as these can all be set relative to the remaining rate parameter; traditionally, f is set equal to 1.

It is not possible to separate the effects of the substitution rate and the amount of time a sequence has been evolving for because these factors are confounded (Felsenstein, 2004). Therefore, the branch lengths of our trees are generally calculated as the average number of substitutions per site along that branch. To this end, the rate factor μ is scaled to make the mean instantaneous rate of substitution 1, to ensure $\sum_{i,j;\, i\neq j} \pi_i q_{ij} = 1$ or, equivalently, $-\sum_i \pi_i q_{ii} = 1$. This scaling factor depends on our type of substitution matrix. In practise, the mean rate of substitution is always normalised to 1 so there is no need to write μ in further matrices; the numerical value of μ is not important *per se*, and we assume that all matrices have been normalised by the appropriate factor. Thus, we can re-write the REV matrix as follows:

Q	A	C	G	T
A	−	$a\pi_C$	$b\pi_G$	$c\pi_T$
C	$a\pi_A$	−	$d\pi_G$	$e\pi_T$
G	$b\pi_A$	$d\pi_C$	−	π_T
T	$c\pi_A$	$e\pi_C$	π_G	−

All reversible substitution models correspond to the REV matrix with various restrictions placed on the parameters (Swofford, 1996). In the most simple matrix, the Jukes Cantor 'JC69' matrix (Jukes and Cantor, 1969), all of the relative rate parameters are equal to each other (a = b = c = d = e = f) and each of the base frequencies (π_A, π_C, π_G and π_T) is ¼. Thus, having scaled the matrix with the factor μ, the Q-matrix for JC69 can be written:

Q	A	C	G	T
A	−1	1/3	1/3	1/3
C	1/3	−1	1/3	1/3
G	1/3	1/3	−1	1/3
T	1/3	1/3	1/3	−1

There are several other ways of parameterising a Q-matrix. For example, the HKY85 model (Hasegawa *et al.*, 1985) allows base frequencies to differ and allows only different rates of substitution between purine-to-purine or pyrimidine-to-pyrimidine changes (transitions) and pyrimidine-to-purine or purine-to-pyrimidine changes (transversions). HKY85 is often referred to as simply the HKY model. In the HKY model transitions (b and e) are given a rate κ relative to transversions (a, c, d and f); κ is referred to as the transition: transversion (ts: tv) bias. Thus our instantaneous rate matrix for HKY is:

Q	A	C	G	T
A	–	π_C	$\kappa\pi_G$	π_T
C	π_A	–	π_G	$\kappa\pi_T$
G	$\kappa\pi_A$	π_C	–	π_T
T	π_A	$\kappa\pi_C$	π_G	–

1.2.3.2 Explicit Models of Codon Substitution

Transition probability matrices for codon models (Goldman and Yang, 1994) are calculated in the same way as simple DNA matrices. Codon models however, are not restricted to the four states (A, C, G and T) of DNA models. Instead there are 61 possible states for the universal genetic code, representing all of the possible sense codons in a codon alignment; the three stop codons are excluded, as changes to and from these codons are highly likely to destroy protein function (see Chapter 2). Codon models have the added advantage that they can be used to measure the strength of selection on a protein-coding sequence (Muse and Gaut, 1994), which is usually measured by the ratio of non-synonymous to synonymous codon substitutions (d_n:d_s

or ω) relative to the same rate if no selection is occurring. Where ω exceeds one, positive selection is inferred, since the rate at which amino acids are being replaced exceeds the rate at which non-synonymous mutations have been fixed. An ω value equal to one is akin to neutral evolution and an ω value of less than one signifies that our codon sequence is under purifying selection (Yang *et al.*, 2000).

1.2.3.3 Explicit Models of Amino Acid Substitution

The evolution of amino acid sequences can be modelled similarly to nucleotides and codons, insofar as a matrix can be used containing all possible amino acid states and the (instantaneous) transition probabilities between them. Instead of the four states of nucleotides and the 61 states of codons however, there are 20 states for amino acids. Amino acid substitutions are usually between amino acids with similar physicochemical properties and this has proven difficult to model parametrically (e.g., Massingham, 2002). For this reason, the most commonly applied amino acid substitution models are empirical models, discussed in section 1.2.4.

1.2.4 Mechanistic and Empirical Models

In phylogenetic analyses parameters may be estimated anew for each dataset (mechanistic models) or set to values previously estimated from a large dataset that is assumed to be representative for the data under consideration (empirical models). Mechanistic models are particularly valuable when we wish to understand variation between datasets in various parameters, such as nucleotide frequencies or selective pressures (Yang *et al.*, 2000), or to understand features specific to a dataset. In

contrast, empirical models are useful when we would have to estimate a large number of parameters from a dataset, which may have too few characters to do so reliably, or where we expect the evolutionary processes to be similar between datasets and we are more interested, say, in investigating a phylogeny. Most codon models of evolution are mechanistic and because they are far more parameter rich than simpler single nucleotide models parameter estimates may have large standard errors when only a small amount of sequence data is available. Recently, empirical codon models have been developed (Kosiol, 2006), where the parameters have been previously estimated from a large dataset and do not need to be estimated anew for subsequent datasets.

The majority of amino acid modelling uses empirical models. Many popular empirical amino acid matrices have been estimated using large datasets and these matrices can be readily applied to a new dataset without the need to re-estimate all of the state transition rates (e.g., Dayhoff, 1978; Jones *et al.*, 1992; Whelan and Goldman, 2001). Mechanistic additions to protein models allow the relative frequencies of amino acids to be taken into account whilst maintaining the empirically calculated probabilities of instantaneous change between amino acids (Cao *et al.*, 1994; Goldman and Whelan, 2002).

1.2.5 Heterogeneity in Evolutionary Rate across a Sequence

Most coding DNA and amino acid sequences show variation in the substitution rate of each site along a sequence due to differences in selective pressures at different sites. Mutations that change the structure of a protein such that it is less able to perform its function are unlikely to persist. Different parts of the protein are

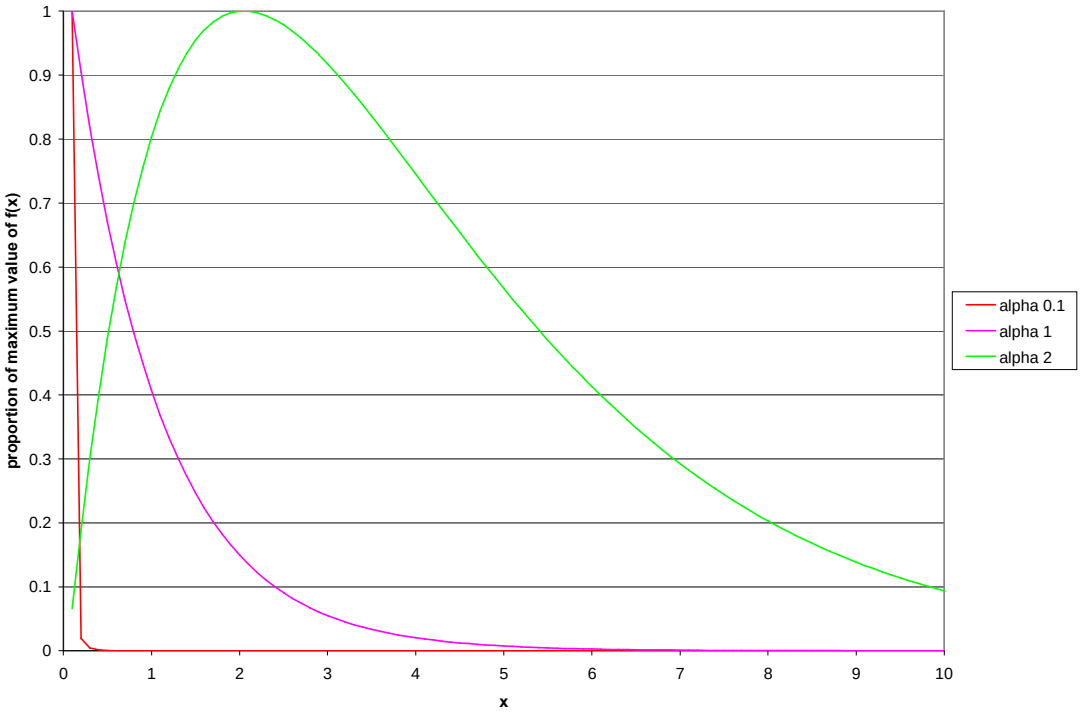
likely to be under different constraint; more substitutions are likely to be deleterious in an active site of a protein versus in a non-functional loop region, for example. Even non-coding DNA, through mutational processes alone, can show significant rate heterogeneity (Mouse Genome Sequencing Consortium, 2002; Hardison *et al.*, 2003). Some mutations are context dependent for example; CG pairs of nucleotides are frequently methylated in vertebrates, leading to mutational deamination (to CA or TG) (Gardiner-Garden and Frommer, 1987). Modelling the rate variation seen in biological sequences can take two forms, categorical and continuous modelling, both of which are used in this thesis.

Categorical modelling of rate variation uses a pre-specified number of rate categories that sites belong to. For example, in this thesis models are used with three pre-specified rate categories, one for each (putative) codon position across a sequence alignment (such that the average rate of evolution for all putative first, second and third codon positions is estimated). Other models with pre-specified rate categories include categories for invariant sites (Hasegawa *et al.*, 1985) or those based on protein secondary structure (Thorne *et al.*, 1996).

Where we are unsure of the number of rate categories to apply to a sequence *a priori* or where it is inappropriate, it is sensible to use a continuous model of rate variation. A γ distribution (gamma distribution) is commonly used; it works well in practise and conveniently, the shape of the γ distribution is governed by a single parameter, α (Yang, 1993). The rate at each site is taken to be a random draw from the distribution. When α is low then there is extreme variation in the rate at which different positions in the alignment have evolved, with most sites having evolved at

very low rates and relatively few having evolved at much higher rates (see Figure 1.1). As α increases the rate variation decreases and the distribution of evolutionary rates becomes bell-shaped and as α tends to infinity (∞) all sites tend towards evolving at the same rate. In practise a continuous γ distribution is implemented very rarely because the computation of α by this method is time consuming and computationally expensive. It is far more practical to use a discretised approximation of the continuous γ distribution, which has a finite number of rate categories (Yang, 1994b).

Figure 1.1 – Examples of the γ distribution, where f(x) is the distribution function for the γ distribution with shape parameter α. For illustrative purposes, f(x) / f_{max}(x) is plotted.



1.3 Likelihood

An evolutionary model can be used by itself to understand sequence evolution, using parameter estimates, or by comparison with another model, to test for the difference in the fit of the models to the evolution of the data (i.e., whether or not a more complex model explains the evolution of the data significantly better than the more simple model). Likelihood is a long established and popular framework with many appealing statistical attributes that is suitable for these analyses (Fisher, 1921, 1925; Edwards, 1992). The likelihood of a hypothesis is related to the probability of observing the data D (a set of aligned nucleotide or amino acid sequences) given the hypothesis (*H*, a specified model of evolution including the phylogenetic tree), such that $L = P$ (D | tree, model). Our data is usually a series of columns of nucleotides in an alignment (D_1, D_2, …D_n, where *n* is the length of the sequence). We use natural logarithms of the likelihood due to the small probabilities involved, which occurs as we multiply the separate probabilities of each column to get a probability for the whole alignment. The log likelihood (or support) for an alignment becomes:

$$\ln L(H \mid D_1, D_2, D_3, \ldots D_n) = \sum_{i=1}^{n} \ln P(D_i \mid H)$$

The multiplication of probabilities (addition of likelihoods) for each site is correct under the assumption that sites evolve independently. The greater the ln *L* value is, corresponding to a higher probability, the better we consider our model to be as an explanation of the evolution of the sequences.

We optimise the parameters (such as tree topology and branch lengths and features such as base frequencies, exchangeabilities and so on if these are included in the model) to obtain those that give the maximum likelihood (ML), i.e., those parameter values that best describe the observed data (Felsenstein, 2004). The values of parameters when the likelihood is maximised are the maximum likelihood estimates (MLE's). The phylogeny describes the evolutionary relationships between the sequences and each branch length describes how much evolutionary change has occurred at that point in the tree. In some cases the parameters we wish to optimise will include the phylogeny, meaning we have to test how well, given the other parts of the model, different trees explain the evolution of the data. In other cases we use a single phylogeny and assume that it is the correct phylogeny that adequately represents the evolutionary history of the data. The model contains the parameters that describe the evolutionary process, such as the JC69 and HKY models etc (Jukes and Cantor, 1969; Hasegawa *et al.*, 1985). All of the likelihood calculations and optimisations in this thesis are made using the PAML suite of programs (Yang, 1997).

Maximising likelihoods can be difficult and all software sometimes fails, for reasons that might be due to features of the data (but not always). Re-running programs using different parameter optimisation options can sometimes lead to more reliable results, but in investigations where a very large number of analyses are performed it is not practical to correct or reattempt all failed analyses. In these cases alternative methods for discarding unreliable results are required. I will consider the methods for discarding results where such failures have occurred in subsequent chapters, as such issues arise.

1.4 Hypothesis Testing

Likelihood analyses allow several levels of complexity of hypothesis testing. Firstly, when we obtain our point estimates of parameters (such as ω or α) we can also calculate confidence intervals around these MLEs. The confidence intervals around the MLE's are those values of the parameter that do not significantly worsen the likelihood. Where our confidence intervals exclude certain values we can exclude certain evolutionary possibilities. For example, if our estimate of the transition: transversion bias parameter were 4.0 +/- 1.5, we know that there is a definite transition: transversion bias as our estimate of the ratio excludes unity (i.e., 1).

Likelihood also allows us to compare competing, mutually exclusive, hypotheses that are represented by models with different restrictions on their model parameters. For example, if we have two evolutionary models, identical except one contains an additional parameter describing some extra feature of sequence evolution, we can test whether the addition of this parameter *significantly* improves our explanation of the evolution of the data. When the simpler model is nested in the complex model, i.e., is a special case of the complex model when the additional parameters of the complex model are fixed at a specific value, the extra free parameters of the complex model necessarily improve that model's fit to the data. We do not test whether the more complex model explains the evolution of the data better than our simple, nested model, but instead whether or not this necessary improvement in model fit is significant. There are also ways of testing between models that are not nested (Goldman, 1993). The three main methods used in this thesis for hypothesis testing are (1) the likelihood ratio test (LRT) (Fisher 1925; Edwards, 1992; Whelan

and Goldman, 1999; Goldman and Whelan, 2000; Felsenstein, 2004), (2) parametric bootstrapping by simulation and (3) non-parametric bootstrapping (Efron and Tibshirani, 1993; Goldman, 1993; Whelan *et al.*, 2001; Felsenstein, 2004).

1.4.1 The Likelihood Ratio Test

The support for a model or hypothesis provided by the data is how well that model explains the evolution of that data (given a fixed topology), given by the log likelihood value. The greater the value, the better the support and the better the model explains the evolution of the data. The LRT statistic is simply the difference (Λ) in the support of our competing hypotheses in the explanation of the evolution of the data:

$$\Lambda = \max [\ln L (H_0 \mid D)] - \max [\ln L (H_1 \mid D)]$$

Here, we denote by H_0 the evolutionary model that corresponds to the null hypothesis, and by H_1 the evolutionary model that corresponds to the alternate hypothesis. The 'max' term corresponds to the likelihood value maximised over all possible parameter values for the evolutionary model (for a fixed tree topology).

In cases where the model corresponding to the null hypothesis (the 'null model') is nested in the alternate model, i.e., is a special case of the alternate model, -2Λ is approximately χ^2 distributed if the null hypothesis is correct, with n degrees of freedom (d.f.) where n is the difference in the number of free parameters between H_0 and H_1 (Wilks, 1938). The χ^2 approximation to the LRT can be used to identify the region of the parameter space about MLEs of the parameters where simple hypotheses

are not rejected at a given significance level when compared to the alternate hypothesis.

The simple χ^2 approximation to the LRT, as described above, is not valid in cases when the null hypothesis model is a special case of the alternate model that has had its relevant parameters fixed to a *boundary* of its possible values (Goldman and Whelan, 2000). In some of these cases we may be able to compare our null and alternate models using an LRT with a known distribution but in some cases this may not be so simple (for example if the nested model is a special case of the alternate model where several parameters have been fixed at their boundaries).

An example of a hypothesis test that causes a complication to the simple χ^2 approximation to the LRT that we can solve would be when the null hypothesis (H_0) is that a sequence alignment had evolved with no rate variation and the alternate hypothesis (H_1) is that the sequence alignment had evolved with rate variation. We could test these hypotheses with, say, the fit of an HKY model to our data versus the fit of an HKY model plus a discretised γ distribution to our data. In this case the null model is a special case of the alternate model, with the α parameter fixed to ∞ (no rate variation). In this situation we can still compare our null and alternate models with a known distribution, half composed of a χ^2 distribution with 1 degree of freedom and half composed of a χ^2 distribution with 0 degrees of freedom (Goldman and Whelan, 2000). Note that this is not frequently performed because, in practise, rate heterogeneity is usually highly significant and a null distribution that is solely χ^2 with one degree of freedom is strongly rejected (meaning that a distribution that is half χ^2 with one degree of freedom and half χ^2 with 0 degrees of freedom would be rejected

too). In cases where models are not nested there is no alternative but to employ a re-sampling or simulation approach (Whelan *et al.*, 2001; Felsenstein, 2004; Goldman, 1993).

1.4.2 Parametric Bootstrapping by Simulation

Parametric bootstrapping and non-parametric bootstrapping are common simulation and re-sampling approaches, respectively, which require the creation of pseudo-replicate datasets (Efron and Tibshirani, 1993). In traditional statistics it is possible to work out our expectations under the null model analytically and independently of inferred parameter values. In phylogenetics there are many test statistics where we cannot do this so the inferred parameter values under the null model are used to generate pseudo-random replicate data, or pseudo-replicates, which necessarily conform to the null model. The null and alternate models are applied to each of the pseudo-replicates. One then obtains a value for the difference in support for the null and alternate models of evolution for each pseudo-replicate dataset (Efron and Tibshirani, 1993). This gives us a distribution of expected values if the null hypothesis model is correct. If the observed test statistic falls outside the confidence boundaries of the distribution of differences in model fit then we reject the null hypothesis.

1.4.3 Non-Parametric Bootstrapping

Non-parametric bootstrapping generates pseudo-replicate datasets using randomisation instead of simulation. Whilst this technique can be used to compare any test statistic to a null distribution, I focus on model fit (log likelihoods) because that is most relevant to this thesis. Recall that the log likelihood is a measure of how well a model fits the observed data and the difference in log likelihoods between two models is a measure of the relative quality of fit of the two models (Whelan *et al*., 2001).

In non-parametric bootstrapping we assume that the data has been generated by the null hypothesis and therefore we re-draw from the data to make pseudo-replicates as a proxy for re-drawing from the null distribution (Efron and Tibshirani, 1993): no simulation is required. The analysis of these pseudo-replicates proceeds under the null and alternate models, in the same way as parametric bootstrapping, such that differences in the fit of the models to the data are calculated for each pseudo-replicate and a distribution of such test statistics is made. If the observed test statistic falls outside the confidence boundaries of the pseudo-replicate derived distribution of test statistics then we reject the null hypothesis.

1.5 Thesis Summary

I have presented a suite of commonly applied techniques for modelling different aspects of nucleotide, codon and amino acid sequence evolution. Furthermore, I have discussed how specific conclusions about sequence evolution can be made by testing between different models of sequence evolution applied to the same dataset. The remainder of this thesis is chiefly concerned with different manipulations of these techniques to address interesting questions of evolutionary biology.

The use of models that have discrete categories of evolutionary rates (and other factors, such as ts:tv bias, rate heterogeneity etc) is introduced in Chapter Two. A set of models is applied to a large dataset in order to tease out the average effects of selection at the different codon positions. The range of models is wider than any previously applied and the effects that selection can have on different evolutionary phenomena is elucidated. Similar techniques are applied to the overlapping reading frames in a hepatitis B virus alignment (Yang *et al*., 1995), giving interesting insights into the effects of overlapping reading frames on viral evolution. Surprisingly an overlapping reading frame does not necessarily lead to an increase in evolutionary constraint.

Inspired by the results of Chapter Two, Chapter Three uses the differences in estimates of model parameters at the three codon positions as an identification tool for protein-coding regions in aligned genomic data, performing well. This chapter introduces the use of windows of DNA across much larger alignments and includes an

extensive study on yeast protein-coding regions following the production of a whole genome yeast alignment based on a previously available dataset (Kellis *et al.*, 2003).

In Chapter Four I continue to use DNA windows and develop a novel ML method for detecting eukaryotic origins of replication by differences in evolutionary processes in windows that have evolved either side of an origin of replication. This method successfully finds a signal for a primate origin of replication that a previous method could not identify with a computational method (Francino and Ochman, 2000). This method can be generalised to use pairs of windows that both progress along an alignment simultaneously to search an entire yeast chromosome. This dataset is not ideal but there is a current paucity of available and annotated data to perform this kind of analysis more successfully.

In Chapter Five the principle of using two DNA windows across a large alignment is generalised and applied to both yeast and mammalian data. I am able to use more data than previous studies, with a methodology not previously applied in this manner. Inferences can be made about the independence of evolution of regions that are specified distances apart on a chromosome and how average estimates of the differences in model parameters between two windows of DNA vary over different scales.

Building upon results from Chapter Five, Chapter Six investigates the subject of phylogenomics and combining different datasets for phylogenetic purposes. I investigate how we might judge the effects of adding different parameters to our models when combining a large number of genes and accounting for differences

between them using a large dataset (Rokas *et al.*, 2003), drawing conclusions on the parameters that are important in such studies with several criteria. I consider the information obtained, observing how the order in which different tree topologies are supported settles to the 'true' order of support, depending on the model used. I conclude that complex models should be used in future studies, which is not common practise, and demonstrate how even large datasets may fail to support a single optimal phylogeny with high bootstrap support.

Chapter 2: Evolutionary Patterns at Different Codon Positions

Contents

2.1 Introduction

In this chapter I introduce the concept of a protein-encoding gene and describe how series of nucleotide triplets, each called a codon, encode functional protein molecules with specific amino acid sequences. I discuss the genetic code that ensures specific codons code for specific amino acids in the protein and how the DNA is used to produce a polypeptide via translation and transcription, and the different functional constraints that are placed upon the three different codon positions by the genetic code. I relate these functional constraints to possible differences in evolution at the codon positions. I discuss the few previous works that have taken place in this field before, such as those by Massingham (2002) and Shapiro *et al*. (2006). Expanding upon the models and techniques discussed in Chapter 1, Sections 1.2.3.1, 1.3 and 1.4.1, I introduce a set of models and tests that are applied to the PANDIT database (Whelan *et al.*, 2006) to determine the differences in evolutionary rates, rate heterogeneity, ts: tv biases and nucleotide frequencies at the different codon positions. Novel observations are made and discussed in the context of the biology of the genetic code. Many of the intra-codon evolutionary differences relate to minimising the effects of mutations on amino acid coding properties of the codon.

Following on from this study I investigate the evolutionary properties at different categories of sites in genes that have overlapping reading frames. I discuss the ways in which different reading frames can overlap and functional constraints that overlapping reading frames may have; I introduce further models and tests that can be used to investigate this topic, similar to those used in the PANDIT database study. I employ a hepatitis B virus (HBV) dataset previously used in a similar but more

limited study (Yang *et al*., 1995) and make novel observations and conclusions, noting that overlapping reading frames do not necessarily cause an increase in evolutionary constraint at those positions.

2.2 The Genetic Code and Protein-Coding Genes

Different proportions of any genome contain sequences that encode lengths of amino acids that fold in a specific manner to produce functional protein molecules. In archaebacteria and eubacteria the proportion of the genome that is protein-coding is usually high (>90%; Audic and Claverie, 1998). In eukaryotes the proportion of the genome that codes for functional protein molecules is often somewhat lower (around 50% in *Saccharomyces cerevisiae*; see Chapter 3), especially in Metazoa (roughly 1.2% in humans; International Human Genome Sequencing Consortium, 2001). In this section I describe how proteins are coded for by specific sequences of nucleotide triplets, codons, in genomic DNA, how these sequences are used as a template to produce messenger RNA (mRNA) during the process of transcription, and how mRNA is used to produce a 'chain' of amino acids, a polypeptide, during translation. In some viruses, such as HIV for example, the genome itself is made of RNA and not DNA and the process by which functional proteins are produced differs from the general system described here. These systems are described elsewhere (e.g., Rambaut *et al*., 2004; Alberts *et al*., 2002). The functional restrictions imposed by the genetic code, discussed in this section, are still relevant to organisms that violate the classic DNA-to-mRNA-to-protein pathway. I discuss methods for investigating different evolutionary constraints at different codon positions, based on the maximum

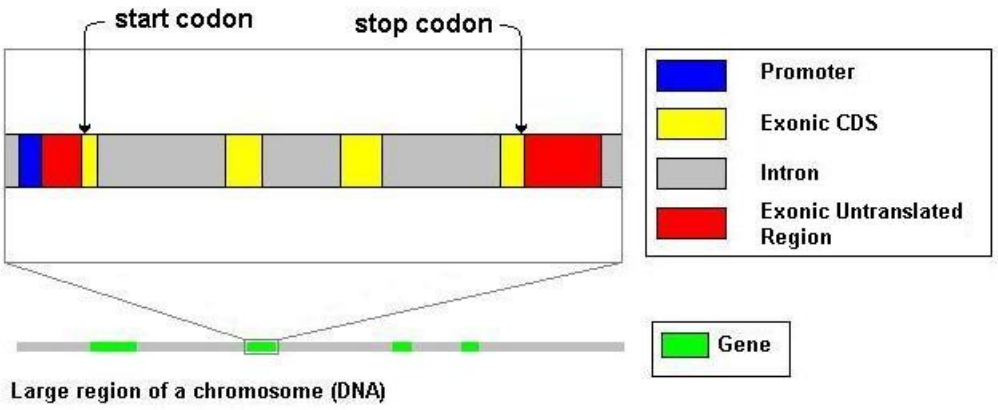
likelihood models discussed in Chapter 1 (Section 1.3), and apply these methods to a large database of aligned codon sequences, PANDIT (Whelan *et al.*, 2006).

The universal genetic code, which represents all codons and which amino acid each one codes for, has been studied extensively and is presented in Table 2.1. The key unit of the genetic code is the codon, a nucleotide triplet that unambiguously codes for a single amino acid or the termination of a polypeptide chain. There are 64 possible nucleotide triplets and only 20 standard amino acids. Each amino acid is coded for by one or more codons but each codon only codes for a single amino acid (or marks the end of a codon sequence). In a simple protein-coding gene, where a single open reading frame codes for the whole protein sequence , the protein-encoding DNA sequence starts with the methionine-coding codon (nucleotides ATG) and progresses through a sequence of other amino-acid-coding codons, finally ending with one of three stop codons (TAA, TAG or TGA). Many genes in higher eukaryotes are coded for by several spatially separated open reading frames; in such cases the first open reading frame will begin with the unique methionine-encoding ATG start codon and the last open reading frame will end with a stop codon. Each of these spatially distinct open reading frames forms an exon and exons are separated by non-protein-encoding DNA sequences termed introns, which are excised from the mRNA sequences before translation (Figure 2.1). Exons can also include untranslated regions that are present in mature mRNA; the protein-encoding part of the exon is the protein coding sequence (CDS).

Table 2.1 – The genetic code. The three nucleotides that compose each codon are shown first, followed by the standard abbreviations of the amino acid that the codon codes for, followed by the single letter notation for that amino acid. The three stop codons, which do not code for an amino acid but terminate protein translation, are highlighted in red.

Second Nucleotide → First Nucleotide ↓	A	C	G	T
A	AAA Lys (K) AAC Asn (N) AAG Lys (K) AAT Asn (N)	ACA Thr (T) ACC Thr (T) ACG Thr (T) ACT Thr (T)	AGA Arg (R) AGC Ser (S) AGG Arg (R) AGT Ser (S)	ATA Ile (I) ATC Ile (I) ATG Met (M) ATT Ile (I)
C	CAA Gln (Q) CAC His (H) CAG Gln (Q) CAT His (H)	CCA Pro (P) CCC Pro (P) CCG Pro (P) CCT Pro (P)	CGA Arg (R) CGC Arg (R) CGG Arg (R) CGT Arg (R)	CTA Leu (L) CTC Leu (L) CTG Leu (L) CTT Leu (L)
G	GAA Glu (E) GAC Asp (D) GAG Glu (E) GAT Asp (D)	GCA Ala (A) GCC Ala (A) GCG Ala (A) GCT Ala (A)	GGA Gly (G) GGC Gly (G) GGG Gly (G) GGT Gly (G)	GTA Val (V) GTC Val (V) GTG Val (V) GTT Val (V)
T	TAA **STOP** TAC Tyr (Y) TAG **STOP** TAT Tyr (Y)	TCA Ser (S) TCC Ser (S) TCG Ser (S) TCT Ser (S)	TGA **STOP** TGC Cys (C) TGG Trp (W) TGT Cys (C)	TTA Leu (L) TTC Phe (F) TTG Leu (L) TTT Phe (F)

Figure 2.1 – The intron-exon structure of many genes in higher eukaryotes.



Transcription is the process of producing mRNA sequences from the protein-coding parts of the genome. The RNA nucleotides that pair with the protein-coding DNA are held against their respective nucleotides in unwound DNA by hydrogen-bonds and these single RNA nucleotides are joined together by various enzymes to produce pre-mRNA. After processing, which includes the addition of a poly-A tail and a 5' modified guanine cap and the splicing of any intron sequences from the RNA by various other components of the cellular molecular machinery, the mRNA is complete (Alberts *et al*., 2002). At this stage in eukaryotes the mRNA is selectively exported from the nucleus to the rough endoplasmic reticulum where there is a dense population of ribosomes. This export does not happen in eubacteria and archaebacteria since they lack a nucleus and other membrane-bound organelles.

Translation is the production of polypeptides from mRNA and occurs in ribosomes. Translation requires specific translation RNAs (tRNAs) that have a specific primed amino acid at one end of the molecule and a specific mRNA codon binding site at the other end of the molecule. The mRNA passes through a subunit of

the ribosome and, codon by codon, the tRNA that binds the particular codon enters another ribosome subunit and the primed amino acid at the non-codon-recognising end of the tRNA is added to the elongating polypeptide chain. The tRNA dissociates and the process is repeated until the stop codon is reached and the full polypeptide chain is complete. No tRNA binds to the any of the three stop codons so the polypeptide chain cannot extend beyond a stop codon; release factors release the mRNA and the polypeptide upon reaching a stop codon (Alberts *et al.*, 2002). Typically, the folding of a nascent polypeptide into a functional conformation also occurs in the endoplasmic reticulum; various pieces of molecular machinery, such as chaperonin proteins, ensure that a polypeptide folds into its correct form.

Different tRNAs bind to the mRNA with different specificity, leading to some redundancy in the genetic code, since 61 codons code for only 20 amino acids. Some amino acids have more than one tRNA molecule, each of which may bind to a different codon, and some tRNA molecules only require the first two nucleotides of the codon to code for a specific amino acid. Thus in some cases, the third nucleotide of the codon may not affect the amino acid that is coded for. In the case of alanine for example (Ala in Table 2.1), the third base does not change the amino acid that is coded as long as the first two bases in the codon are G and C respectively (see Table 2.1). In this case the remaining third codon position is known as a '4D' site, a site that is fourfold degenerate such that any nucleotide in this position codes for the same amino acid. In terms of encoding a specific amino acid, second codon positions are the most functionally constrained and any change to a second codon position nucleotide causes a change in the amino acid that the codon codes for. On average the first codon position is more functionally constrained than the third codon position but

less functionally constrained than the second codon position in terms of encoding a specific amino acid. This chapter is concerned with the effects that the functional constraints caused by tRNA specificity and the genetic code have on differences in evolutionary patterns observed at the different codon positions.

2.3 Selection and Intra-Codon Evolutionary Variation

The optimality of the genetic code has been a source of investigation with regard to the effects of mutation on any given codon (e.g., Freeland and Hurst, 1998; Freeland *et al.*, 2000). The standard genetic code we observe today is likely to have been selected among sets of variants and to have prevailed because it minimises the effects on CDSs of point mutations and mistranslation. A point mutation or an error in how the codon is read by a tRNA is either synonymous (leads to the same amino acid being encoded at that location in the polypeptide) or non-synonymous (results in a substitution by an amino acid with similar physicochemical properties; Freeland and Hurst, 1998).

The relative rates of evolution at the three codon positions have been shown to relate to the level of degeneracy of the genetic code (such that the third codon position evolves faster on average than the first codon position, which evolves faster on average than the second codon position; Massingham, 2002). Massingham demonstrated differences in the average rates of evolution between the codon positions by implementing a maximum likelihood model that optimises all parameters over the three codon positions (although base frequencies were counted) except for evolutionary rates, which were estimated independently for each of the three codon

positions (see Chapter 1.3 for details of optimisation of likelihood models). In the vast majority of cases optimising different evolutionary rates at the three codon positions significantly improved the fit of the model to the dataset (an earlier version of the PANDIT database; Whelan *et al.*, 2006), judged by a LRT compared to a simpler model where a single parameter describing the evolutionary rate was optimised across all codon positions. Shapiro *et al*. (2006) also used nucleotide-based likelihood models where the codon positions of CDSs were considered, finding such models statistically superior to the simple nucleotide models. Compared to Massingham (2002) and the study described in this chapter, the Shapiro *et al*. (2006) dataset was small, consisting of around 300 genes. Massingham (2002) further investigated the properties of a model that incorporated a measure of how different amino acids are to each other and the possible effects this had on how likely new amino acids are to be fixed in a population when a nonsynonymous mutation has occurred in the codon. Indeed, the genetic code and amino acid exchangeabilities were found to explain almost all of the intra-codon rate variation in real data.

Whilst some of the causes of evolutionary variation between the codon positions are well understood, there are large gaps in our knowledge about the effects that functional constraints have on nucleotide-level properties that might be observed as forms of intra-codon variation. We are currently uncertain how the rate heterogeneities, ts: tv biases and base frequencies vary between the different codon positions. It is also not known how these parameters may interact with each other and the estimates of evolutionary rates at the different codon positions.

It is intrinsically interesting to investigate how selection and the structure of the genetic code affect evolution at the different codon positions. Furthermore, features that differ between codon positions could be used to construct simple nucleotide models that identify exon-like evolution in large genome-scale multiple species DNA sequence alignments. The models based on nucleotide-level effects might allow simpler approaches to exon identification in genome-scale studies than full, parameter-rich codon models would permit. Later in this thesis (Chapter 3) I develop models that contain fewer parameters than full codon models but nevertheless identify CDSs reliably in large multiple alignments. These models of codon evolution could also be used by programs that specifically align codon sequences. Furthermore, recognising differences in the evolutionary properties of different codon positions could be used to develop aspects of full codon models, influencing, for example, how we treat differences in base frequencies at different codon positions. The distributions of parameter estimates could also be used as prior distributions in future investigations under the Bayesian framework.

Separately and in certain combinations, I consider the differences between codon positions for the parameters that relate to rate heterogeneity, base frequencies and the ts: tv bias. These factors have not been considered in this context before. For completeness, I also consider, as Massingham (2002) and Shapiro *et al*. (2006) have previously, differences in parameters that describe evolutionary rates at the different codon positions and, unlike previous studies, observe how the parameters describing rates at different codon positions and other parameters interact with the each other.

The different functional constraints at different codon positions imposed by tRNA specificities may be observed as evolutionary constraints; we may expect, for example, that the less mutagenic nucleotides A and T (Nachman and Crowell, 2000; Touchon *et al*., 2003) have higher frequencies at the second codon position than at the first and third codon positions because of selection against too high a rate of mutation at this position. We may also expect the most extreme rate heterogeneity at the second codon position because most second codon sites will evolve very slowly and a few may evolve quickly (at those sites undergoing recurring positive selection for example). Third codon position sites may have the least rate heterogeneity because most of these sites will evolve quite fast and few third codon sites will be particularly constrained because of the degeneracy of the genetic code. The effects of selection on the ts: tv bias at different codon positions may be less predictable. We might expect that the second codon position has a very strong bias because the genetic code is structured such that transitions, which are generally more common than transversions, are more likely to result in a substitution to an amino acid with similar physicochemical properties than transversion mutations. Conversely, an overall reduction in both transitions and transversions at the second codon position, caused by functional constraint, may lead to a smaller ts: tv bias.

2. 4 Examining Intra-Codon Evolutionary Variation

The models used to examine variation in evolutionary processes at different codon sites are similar to the likelihood models discussed in Chapter 1 (Sections 1.2.3.1, 1.2.5 and 1.3). The HKY model of nucleotide substitution is used as a base for the models used here for several reasons. Firstly, it has been successful in modelling nucleotide evolution and one can make useful biological interpretations for each of its parameters. Secondly, it was used by Massingham (2002) in a similar study. Lastly, more complex models, such as REV, have more parameters to estimate and, especially for small datasets, the standard errors of parameter estimates may be high; thus, optimising fewer biologically relevant parameters in the basic model is desirable, especially since the more complex models may require the optimisation across many more parameters than the basic HKY model. The models that were used in this study are described in Table 2.2. A probabilistic model in the maximum likelihood framework was used that optimises all parameters shared over the three codon positions (although base frequencies are counted and not optimised by ML) except for those parameters specifically optimised separately for each of the three codon positions, which depends on the model in question. All model optimisation across datasets were carried out using the program BASEML in the PAML package (Yang, 1997), operated automatically using custom-written Perl scripts.

Table 2.2 – Evolutionary models used to investigate differences in evolutionary processes at different codon positions.

Model Name	Free Parameters	Model description and parameters that are allowed to differ between codon positions
HKY	4 + tree	HKY model (Chapter 1, Section 1.2.3.1).
HKY+G	5 + tree	HKY model plus a single discretised γ distribution.
HKY+R	6 + tree	HKY and estimates an evolutionary rate for each codon position (with the same proportions between branch lengths of the tree).
HKY+R+T	8 + tree	HKY and estimates an evolutionary rate and ts: tv bias for each codon position.
HKY+R+N	12 + tree	HKY and estimates an evolutionary rate and nucleotide frequencies for each codon position (nucleotide frequencies are estimated by counting).
HKY+R+N+T	14 + tree	HKY and estimates an evolutionary rate, nucleotide frequencies and ts: tv bias for each codon position.
HKY+R+G	7 + tree	HKY plus a single discretised γ distribution and estimates an evolutionary rate for each codon position.
HKY+R+T+G	9 + tree	HKY plus a single discretised γ distribution and estimates an evolutionary rate and ts: tv bias for each codon position.
HKY+R+N+G	13 + tree	HKY plus a single discretised γ distribution and estimates an evolutionary rate and nucleotide frequencies for each codon position.
HKY+R+N+T+G	15 + tree	HKY plus a single discretised γ distribution and estimates an evolutionary rate, nucleotide frequencies and ts: tv bias for each codon position.
HKY+R+3G	9 + tree	HKY and estimates an evolutionary rate and the rate heterogeneity for each codon position (i.e., three discretised γ distributions).
HKY+R+T+3G	11 + tree	HKY and estimates an evolutionary rate, ts: tv bias and the rate heterogeneity for each codon position.
HKY+R+N+3G	15 + tree	HKY and estimates an evolutionary rate, nucleotide frequencies and the rate heterogeneity for each codon position.
HKY+R+N+T+3G	17 + tree	HKY and estimates an evolutionary rate, ts: tv bias, nucleotide frequencies and the rate heterogeneity for each codon position.

The χ^2 approximation (with the appropriate degrees of freedom) to the LRT null distribution was used to test between complex models and simpler, nested models to assess the significance of the improvement in the explanation of the evolution of the data by the more complex models (explained in Section 1.4.1). In total, 14 tests were carried out on each dataset. The details of these tests are presented in Table 2.3. We are also interested in the estimated values of the specific parameters that the evolutionary models incorporate and these are presented after the LRT results.

Table 2.3 – Likelihood ratio tests (continued on next page).

Test	Alternate (A) and null models (B) in LRT	Degrees of freedom in LRT	Biological question addressed by the LRT
T-1	A: HKY+R B: HKY	2	Is there a significant difference in the rate of evolution between codon positions?
T-2	A: HKY+R+G B: HKY+G	2	Having considered rate heterogeneity in the null and alternate models, is there a significant difference in the rates of evolution between codon positions?
T-3	A: HKY+R+3G B: HKY+R+G	2	Is there a significant difference in the rate heterogeneity between codon positions?
T-4	A: HKY+R+N+T+3G B: HKY+R+N+T+G	2	Having considered differences in rate, ts: tv bias and nucleotide frequencies at different codon positions in the null and alternate models, is there a significant difference in rate heterogeneity between codon positions?
T-5	A: HKY+R+T B: HKY+R	2	Is there a significant difference in the ts: tv bias between codon positions?
T-6	A: HKY+R+N+T B: HKY+R+N	2	Having considered differences in nucleotide frequencies between codon positions in the null and alternate models, is there a significant difference in the ts: tv bias between codon positions?

Table 2.3 continued.

Test	Alternate (A) and null models (B) in LRT	Degrees of freedom in LRT	Biological question addressed by the LRT
T-7	A: HKY+R+T+G B: HKY+R+G	2	Having considered rate heterogeneity in the null and alternate models, is there a significant difference in the ts: tv bias between codon positions?
T-8	A: HKY+R+N+T+G B: HKY+R+N+G	2	Having considered rate heterogeneity and the nucleotide frequency differences between codon positions in the null and alternate models, is there a significant difference in the ts: tv bias between codon positions?
T-9	A: HKY+R+N+T+3G B: HKY+R+N+3G	2	Having considered differences in rate, rate heterogeneity and nucleotide frequencies at different codon positions in the null and alternate models, is there a significant difference in the ts: tv bias between codon positions?
T-10	A: HKY+R+N B: HKY+R	6	Is there a significant nucleotide frequency difference between codon positions?
T-11	A: HKY+R+N+T B: HKY+R+T	6	Having considered differences in the ts: tv bias between codon positions in the null and alternate models is there a significant difference in nucleotide frequencies between codon positions?
T-12	A: HKY+R+N+G B: HKY+R+G	6	Having considered rate heterogeneity in the null and alternate models, is there a significant difference in nucleotide frequencies between codon positions?
T-13	A: HKY+R+N+T+G B: HKY+R+T+G	6	Having considered rate heterogeneity in the null and alternate models and the ts: tv bias differences between codon positions, is there a significant difference in nucleotide frequencies between codon positions?
T-14	A: HKY+R+N+T+3G B: HKY+R+T+3G	6	Having considered differences in rate, rate heterogeneity and ts: tv bias at different codon positions in the null and alternate models, is there a significant difference in nucleotide frequencies between codon positions?

There are limits to the models that we can apply using PAML (Yang, 1997). It is not possible to model the differences in rate heterogeneity, ts: tv biases or nucleotide frequencies between codon positions without also allowing for differences in evolutionary rates at the different codon positions. However, the results presented later support the inclusion of modelling differences in evolutionary rates between codon positions and the models used are a reasonable set for the purposes of this investigation. We could construct additional LRTs with models that allow non-proportional branch lengths between the trees optimised for each codon position. This would be akin to having a completely separate HKY model or HKY model plus a single discretised γ distribution for each codon position. These LRTs would not be difficult to perform in principle but it would be tedious to do so for an entire database because the number of degrees of freedom in each test and the χ^2 distribution used to approximate the LRT test statistic distribution would depend on the number of sequences in each PANDIT family and the corresponding number of additional branch lengths estimated for each family. Furthermore, these tests are not directly relevant to the factors that we wish to study (such as the differences in the ts: tv bias between codon positions, etc.) and thus, such tests are not pursued here.

It is clear from Table 2.3 that some of the tests address similar biological questions, e.g., T-1 and T-2, T-3 and T-4, T-5 to T-9, and T-10 to T-14. Comparing tests T-1 and T-2, it is clear that they both address whether different codon positions have significantly different evolutionary rates. However, test T-1 and test T-2 address this question in subtly different ways. The null and alternate evolutionary models in test T-2, unlike test T-1, take account of some of the effects of rate heterogeneity across the codon as they have a discretised γ distribution implemented. Thus, if there

is any interaction (non-orthogonality) between rate heterogeneity across the codon and differences in evolutionary rate between codon positions, then this should be revealed by comparisons of test T-1 and test T-2. For example, if the LRT statistic is higher in test T-1 than test T-2 (even though both tests have only one degree of freedom), it means that not accounting for rate heterogeneity across the codon may lead to biased estimates of differences in rate between the codon positions. It will be interesting to note differences in parameter estimates that occur between models testing similar hypotheses.

2. 5 The PANDIT Database

For the results of this study to be both accurate and general, a dataset is required that contains many different multiple-alignments of codon sequences. The PANDIT database (Whelan *et al.*, 2006) is ideal for this study and contains 7775 multiple alignments of codon sequences. Each alignment was produced from amino acid sequences available in the Pfam database (Bateman *et al.*, 2004); the DNA sequences matching the amino acids were extracted from publicly available databases so that the amino acid alignments could be converted into codon alignments. The PANDIT database provides a phylogeny for each family, which is taken to be the correct phylogeny. Whilst the provided phylogeny may not be the true phylogeny in some cases, it is generally accepted that using a reasonable estimate of the phylogeny leads to reasonable parameter estimates (Yang *et al.*, 1994), as mentioned in Chapter 1.2.1. I consider the phylogenies provided for each PANDIT family to be reasonable estimates of the true phylogenies. PANDIT contains alignments of sequences isolated from viruses, prokaryotes and eukaryotes. Considering the large number of PANDIT

families and the broad coverage of the database itself, I assume that the results obtained from this study are applicable to CDSs in general.

All of the evolutionary models listed in Table 2.2 and tests listed in Table 2.3 were applied to each PANDIT family. Those families for which the PAML software failed to optimise correctly for any test (judged by LRT statistic results that are negative, which is not feasible by definition) or optimised too slowly (judged by recording how long the analysis software had been running for each family) were discarded (see Chapter 1). A failure in the optimisation could be caused by the fact that an evolutionary tree is an inadequate description of the evolutionary history for that family, for example. If the software fails to optimise for any single model, we become more suspicious that other models have optimised correctly and thus, it is safest to discard the entire family.

2.6 Results of the Intra-Codon Evolutionary Variation Study on the PANDIT Database

Out of the 7775 PANDIT families, 5727 (74%) optimised fully in an adequate time. The large number of PANDIT families for which results were available makes it likely that the results obtained are a fair representation of the results that would be obtained if all PANDIT families fully optimised. In Table 2.4 I report the number of families that show significant improvements in model fit for each of the LRTs described in Table 2.3 (tests T-1 to T-14), and I then discuss how tree length and number of sequences may affect whether or not an effect is significant. I then present

the average parameter values and their distributions for the significant tests carried out across the PANDIT database and discuss these results in a biological context.

Table 2.4 – Number of significant likelihood ratio test statistics for tests described in Table 2.3.

Test Number	Number of Significant PANDIT families	Percentage of Significant PANDIT families
T-1	5563	97
T-2	5541	97
T-3	2851	50
T-4	1343	23
T-5	3985	70
T-6	4259	74
T-7	4322	75
T-8	4551	79
T-9	4206	73
T-10	5466	95
T-11	5499	96
T-12	5458	95
T-13	5476	96
T-14	5470	96

It is immediately clear that there are significant differences in the patterns of evolution at the three codon positions; many of these have not been described before. I will discuss the results of each series of tests and the values for the corresponding

parameters (i.e., evolutionary rate, rate heterogeneity, ts: tv bias and nucleotide frequencies) in turn in the following sections. For each of the parameters, the values presented have been drawn only from those results where the consideration of the differences in the parameter between codon positions was found to give a significant improvement in model fit.

In the cases where parameters were optimised for each codon position, I present the mean and median values and the distributions of the parameter values for each codon position. Median results are less affected by a small number of extreme parameter values, which may reflect genuine biological phenomena or challenges in optimisation. I also present the mean and median tree lengths for datasets that produced significant test results. The estimated tree lengths for the same families increase as model complexity increases: there are differences in the estimated average tree lengths between tests T-13 and T-14 (Table 2.16) for example, and it is highly likely that almost all of the PANDIT families that gave significant LRT statistics for test T-13 did so for T-14 and *vice versa*. In some cases the median tree lengths may also reflect the strength of the biological effect we are modelling (i.e., a weak effect is only found to be statistically significant where more evolutionary change has taken place). In the display of distributions of parameter values for each PANDIT family I have chosen to present results for a single model where the factor being investigated in the specific test has given a significant LRT statistic. Note that parameter values generally change little between the more complex models, which are preferred (discussed below).

2.6.1 Difference in Evolutionary Rate Between Different Codon Positions

There is a very strong signal in most families for a difference in the rates of evolution between different codon positions. The incorporation of γ into both the null and alternate models makes a very small difference in the number of significant families between LRTs for test T-1 and test T-2 although, as in all cases tested here, there is some interaction between the estimates of parameters in our models: some of the families that give a significant LRT test statistic for test T-1 no longer give a significant LRT test statistic in test T-2 (see Table 2.4). Thus, when our null model considers rate heterogeneity (test T-2), we observe slightly different results. The average observed evolutionary rates of the different codon positions are presented in Tables 2.5 and 2.6. The relative evolutionary rate of codon position 1 is set equal to 1, which means that we only have two parameter values relating to evolutionary rates to estimate from the data, i.e., the evolutionary rates at codon position 2 and codon position 3 (relative to rate = 1 at codon position 1).

Table 2.5 – Results for model HKY+R parameters where test T-1 significant (97% PANDIT families that optimised sucessfully).

	Tree Length	Relative rate Codon pos. 1	Relative rate Codon pos. 2	Relative rate Codon pos. 3
Mean	4.13	1	0.70	2.48
Median	2.84	1	0.69	2.27

Table 2.6 – Results for model HKY+R+G parameters where test T-2 significant (97% PANDIT families that optimised sucessfully).

	Tree Length	Relative rate Codon pos. 1	Relative rate Codon pos. 2	Relative rate Codon pos. 3
Mean	6.01	1	0.65	5.97
Median	4.07	1	0.64	3.02

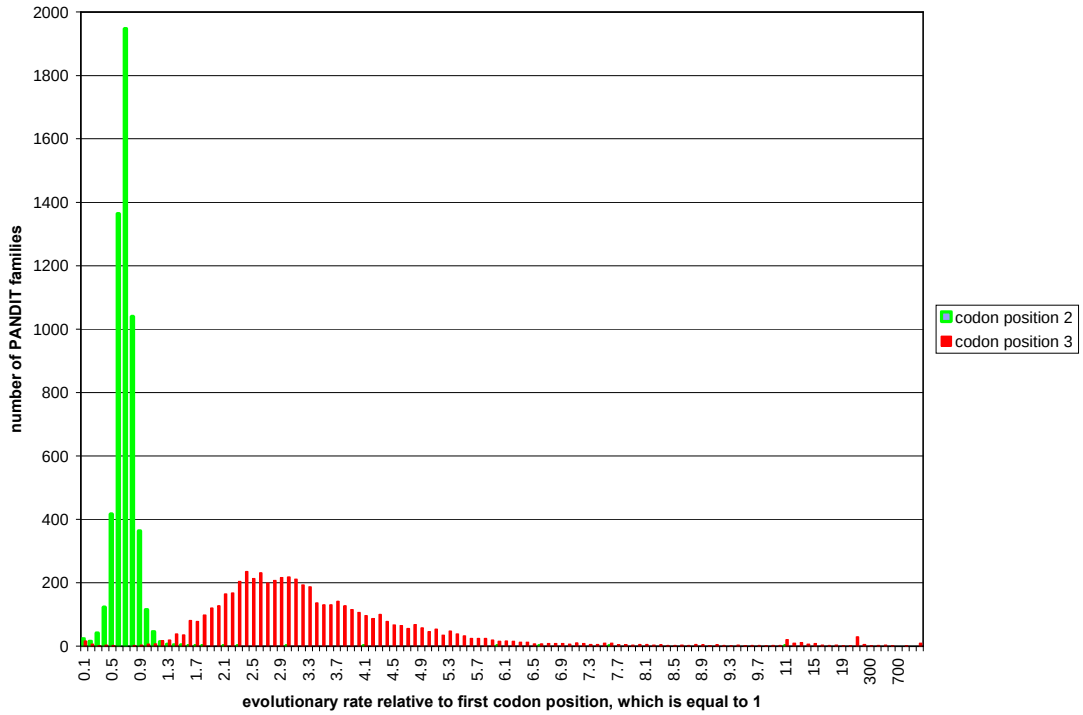
Rate parameter estimates are best taken from test T-2 because it is important to model rate heterogeneity; testing HKY vs. HKY+G is significant for all PANDIT families (results not shown). It is interesting that the median observed evolutionary rate for codon position 2 (but not codon position 3) is very similar for tests T-1 and T-2 (Tables 2.5 and 2.6). The results of Massingham's study of estimating evolutionary rates at the different codon positions using an HKY+R model (normalised so the rate at the first codon position is equal to 1) are presented in Table 2.7. These results are very close to the median results for the estimates of evolutionary rates at the different codon positions presented in Table 2.5. It is better to use the more complex model HKY+G+R, which means the study in this thesis provides more reliable parameter estimates than Massingham (2002).

Table 2.7 – Evolutionary rates across the codon positions from Massingham (2002).

Codon Position	1	2	3
Evolutionary Rate	1	0.71	2.31

The distributions of evolutionary rates of codon positions 2 and 3 (relative to codon position 1) are plotted in Figure 2.2. The distribution of evolutionary rates at the second codon position is particularly tight, with the second codon positions in the vast majority (98.4%) of PANDIT families evolving slower than the first codon position. The distribution of evolutionary rates at the third codon position is much wider across the PANDIT database, indicative of differences in evolutionary constraints at third codon positions in different PANDIT families. The general trend is observed that the evolutionary rate of the third codon position is greater than the first codon position, which in turn is greater than the evolutionary rate of second codon position in concordance with the results of Massingham (2002) (Table 2.7). This can be written $3 > 1 > 2$, which means the parameter estimate of codon position 3 is greater than the parameter estimate of codon position 1, and so on; this notation is used later in this chapter for various parameter estimates of different codon positions, not just evolutionary rates. These results can be easily interpreted considering the functional constraints of the genetic code.

Figure 2.2 – The distributions of evolutionary rates of codon positions 2 and 3 (relative to codon position 1, which has a fixed relative rate = 1 for each PANDIT family). Rate parameters are taken from model HKY+R+G where test T-2 is significant.



2.6.2 Differences in the Heterogeneity of Evolutionary Rate between Different Codon Positions

The second codon position has the most extreme rate heterogeneity (lowest median α value), followed by the first codon position and the third codon position has the least rate heterogeneity (highest median α value). The average observed heterogeneity in evolutionary rates of the different codon positions are presented in Tables 2.8 and 2.9. The mean estimate of parameter α, which governs the shape of the γ distribution, is biased by some exceptionally high values so it is better to use the

median. Extreme rate heterogeneity and very low α values are observed when most sites are conserved but a few change rapidly and the results of measuring rate heterogeneity at the second codon position are consistent with our knowledge of functional constraints. Most second codon positions will be evolutionarily constrained because the amino acid encoded by that codon will be important to the function of the protein and changing the nucleotide at the second codon position changes the encoded amino acid. However, a small number of amino acids may evolve more rapidly in a protein, either due to positive selection or near neutral selection at such sites. The distributions of heterogeneity in evolutionary rates of the codon positions are presented in Figure 2.3.

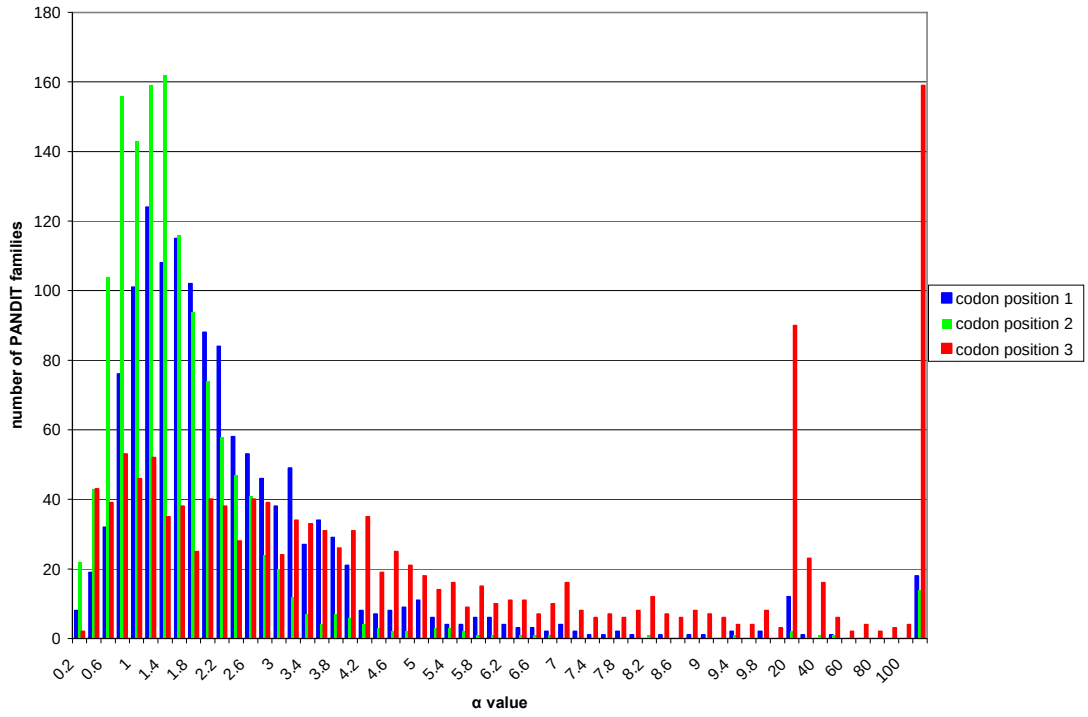
Table 2.8 – Results for model HKY+R+3G parameters where test T-3 significant (50% PANDIT families that optimised sucessfully).

	Tree Length	α at Codon pos. 1	α at Codon pos. 2	α at Codon pos. 3
Mean	10.01	9.03	4.46	214.31
Median	5.50	1.57	1.13	6.92

Table 2.9 – Results for model HKY+R+N+T+3G parameters where test T-4 significant (23% PANDIT families that optimised sucessfully).

	Tree Length	α at Codon pos. 1	α at Codon pos. 2	α at Codon pos. 3
Mean	9.22	12.62	10.00	117.05
Median	6.25	1.76	1.26	6.25

Figure 2.3 – The distributions of rate heterogeneities at the three codon positions, measured by parameter α from model HKY+R+N+T+3G where test T-4 is significant.



The differences in rate heterogeneity between sites are not particularly strong and only 50% of PANDIT families show significant rate heterogeneity variation between codon positions in test T-3. The percentage of PANDIT families that show significant variation in rate heterogeneity between codon positions falls dramatically to 23% when rate, ts: tv bias and nucleotide frequency differences between codon positions are incorporated into both the null and alternate hypothesis models.

Whilst rate heterogeneity has an important effect across whole CDSs, there is less evidence for rate heterogeneity within codon position categories. The estimated values of parameter α for the three within-codon categories are higher than the single estimated value of parameter α across a codon sequence as a whole (results not shown). This is because much of the heterogeneity in evolutionary rate across a

protein can be explained by the different underlying rates of evolution at the different codon positions. Nevertheless, the apparent differences in rate heterogeneity at the different codon position are consistent with our notions of constraint at each position.

2.6.3 Differences in the ts: tv Bias Between Different Codon Positions

The average observed ts: tv biases of the different codon positions are presented in Tables 2.10 to 2.14. Once again, the mean parameter estimates are treated with caution because they may be biased by a small number of extreme values. The median estimates of parameter κ are in better agreement between models than the means and, judging by the very similar κ distributions for codon positions 1 and 2 (Figure 2.4), taken across families in the PANDIT database, it seems that the median estimates are more reliable and conclusions are drawn from the medians and distributions only.

Table 2.10 – Results for model HKY+R+T parameters where test T-5 significant (70% PANDIT families that optimised sucessfully).

	Tree Length	ts: tv at Codon pos. 1	ts: tv at Codon pos. 2	ts: tv at Codon pos. 3
Mean	4.64	1.65	2.35	9.57
Median	3.21	1.59	1.29	3.48

Table 2.11 – Results for model HKY+R+N+T parameters where test T-6 significant (74% PANDIT families that optimised sucessfully).

	Tree Length	ts: tv at Codon pos. 1	ts: tv at Codon pos. 2	ts: tv at Codon pos. 3
Mean	4.17	2.90	5.17	9.59
Median	2.84	1.43	1.46	3.56

Table 2.12 – Results for model HKY+R+T+G parameters where test T-7 significant (75% PANDIT families that optimised sucessfully).

	Tree Length	ts: tv at Codon pos. 1	ts: tv at Codon pos. 2	ts: tv at Codon pos. 3
Mean	7.10	2.01	2.22	45.28
Median	5.02	1.81	1.38	6.33

Table 2.13 – Results for model HKY+R+N+T+G parameters where test T-8 significant (79% PANDIT families that optimised sucessfully).

	Tree Length	ts: tv at Codon pos. 1	ts: tv at Codon pos. 2	ts: tv at Codon pos. 3
Mean	7.23	2.19	2.84	50.24
Median	5.16	1.56	1.55	7.59

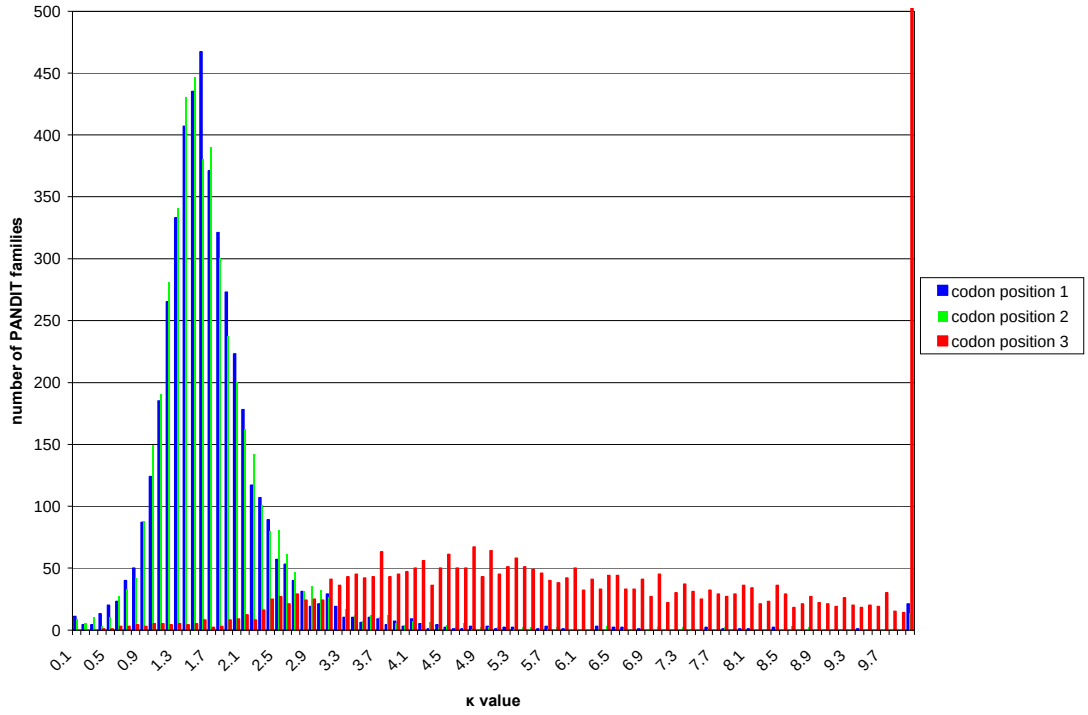
Table 2.14 – Results for model HKY+R+N+T+3G parameters where test T-9 significant (73% PANDIT families that optimised sucessfully).

	Tree Length	ts: tv at Codon pos. 1	ts: tv at Codon pos. 2	ts: tv at Codon pos. 3
Mean	7.55	1.98	2.88	63.05
Median	5.33	1.54	1.55	7.04

When allowing for a different rate and ts: tv bias between codon positions (test T-5) but not accounting for nucleotide frequency differences between codons, it would appear that the third codon position has a greater bias than the first codon position, which has a greater bias that the second codon position, judging by the median κ values at different codon positions across the whole PANDIT database ($3 > 1 > 2$) (Table 2.10). The order of these results is the same for the evolutionary rates. This is in agreement with the results for test T-7, which has the same null and alternate hypothesis models except both the null and alternate models have a single discretised γ distribution over all nucleotide positions (Table 2.12). The mean results are somewhat different but may be affected by extreme values (see Figure 2.4). In some cases there will be very few changes at the second codon position across the whole tree due to constraint; if there is a lack of transversions then the ts: tv bias, unsurprisingly, will be very high.

However, when we also make allowances for differences in the nucleotide frequencies between the different codon positions in the null and alternate hypothesis models and observe the ts: tv biases between the different codon positions (tests T-6 and T-8), the bias is very similar (using the median values) for the first and second codon positions (3 > 1 ≈ 2) (Tables 2.11 and 2.13). Indeed, both the mean and median ts: tv bias are marginally lower in the first codon position than the second codon position. The complexity in estimating the ts: tv bias lies in the interaction with nucleotide frequencies at different codon positions. The interpretation of this result is not simple and it is highly likely that the genetic code may play a role in this effect. The distributions of the ts: tv biases across the PANDIT database are very similar for the first and second codon positions and are presented in Figure 2.4; the distributions of parameter κ are shown for the model HKY+R+N+T+G and not HKY+R+N+T+3G. The results for HKY+R+N+T+3G, are treated with caution because the addition of the extra γ distributions (such that there is one per codon position) did not produce a significant improvement in model fit for many of the PANDIT families (see results of test T-4 in Table 2.4). The interaction of the additional γ distributions may compromise the reliability of our estimates of other parameters such as the ts: tv bias.

Figure 2.4 – The distributions of ts: tv biases at the three codon positions, measured by parameter κ from model HKY+R+N+T+G where test T-8 is significant.



Since nucleotide frequencies are clearly different between the codon positions (see Section 2.6.4 below), our final conclusions regarding the order of the ts: tv biases between the codon positions is 3 > 1 ≈ 2. The ts: tv bias should not be confused with the evolutionary rate: the second codon position evolves more slowly than the first codon position. The high ts: tv bias at the third codon position is likely to result from selection against transversions. Unlike transition mutations, which are usually synonymous at the third codon position, transversions can cause non-synonymous changes at the third codon position. A lower or higher ts: tv bias than the mutation rate could result from selection. At the first and second codon positions, which have very similar ts: tv bias distributions (Figure 2.4), transition mutations are accepted more than transversions but selection acts to reduce the acceptance of both types of mutation, resulting in a lower ts: tv bias than we observe at the third codon

position. I discuss this phenomenon in more detail alongside the results of the differences in nucleotide frequencies at the different codon positions (Section 2.6.4).

2.6.4 Differences in the Nucleotide Frequencies between Different Codon Positions

Average nucleotide frequencies are significantly different between codon positions in roughly 95% of PANDIT families (see Table 2.4). The base frequency results were equal to two decimal places for all tests T-10 to T-14 because the set of families for which results were significant is virtually identical for tests T-10 to T-14. Therefore, a single set of base frequencies is presented in Table 2.15.

Table 2.15 – Nucleotide frequencies at different codon positions for models HKY+R+N, HKY+R+N+T, HKY+R+N+T+G and HKY+R+N+T+3G for tests T-10 to T-14 where LRTs were significant (95-96% PANDIT families).

	A	C	G	T
Mean	1:0.28 2:0.32 3:0.22	1:0.20 2:0.22 3:0.27	1:0.33 2:0.17 3:0.25	1:0.18 2:0.29 3:0.26
Median	1:0.28 2:0.32 3:0.22	1:0.20 2:0.21 3:0.26	1:0.33 2:0.17 3:0.25	1:0.18 2:0.29 3:0.26

The distributions of base frequencies at the three codon positions in cases where test T-13 indicated significant differences between codon positions are shown in Figures 2.5 (distributions of each nucleotide for the different codon positions) and Figure 2.6 (distributions at each codon position for the different nucleotides). Parameter estimates presented are taken from test T-13 in preference to T-14 because the additional γ distributions used in test T-14 do not significantly improve model fit for many PANDIT families (although in practice this has a minimal effect on the parameter estimates presented). Comparing Figures 2.6a-c, it is clear that the different codon positions have different nucleotide frequency distributions. The third codon position has the least nucleotide frequency bias of all codon positions, which can be explained by the weaker effects of selection on this position. Codon position 2 has a strong bias of A and T nucleotides (higher AT content). Codon position 1 has a high frequency of A and G nucleotides and codon position 3 has a relatively low A content. Nucleotides A and T are known to be less mutable than G and C (Nachman and Crowell, 2000; Touchon *et al*., 2003) and the functional constraint at codon position 2 may explain the bias towards less mutagenic nucleotides. The A and T nucleotide bias at the second codon position may also affect our estimates of the differences in the ts: tv bias between different codon positions.

Figure 2.5 (a-d) – The distributions of frequencies of each nucleotide at each codon position taken from model HKY+R+N+T+G where test T-13 is significant. The different nucleotide frequencies are shown in four separate sub-figures in the order A, C, G and T.

Figure 2.5a – Distributions of frequencies of nucleotide A.

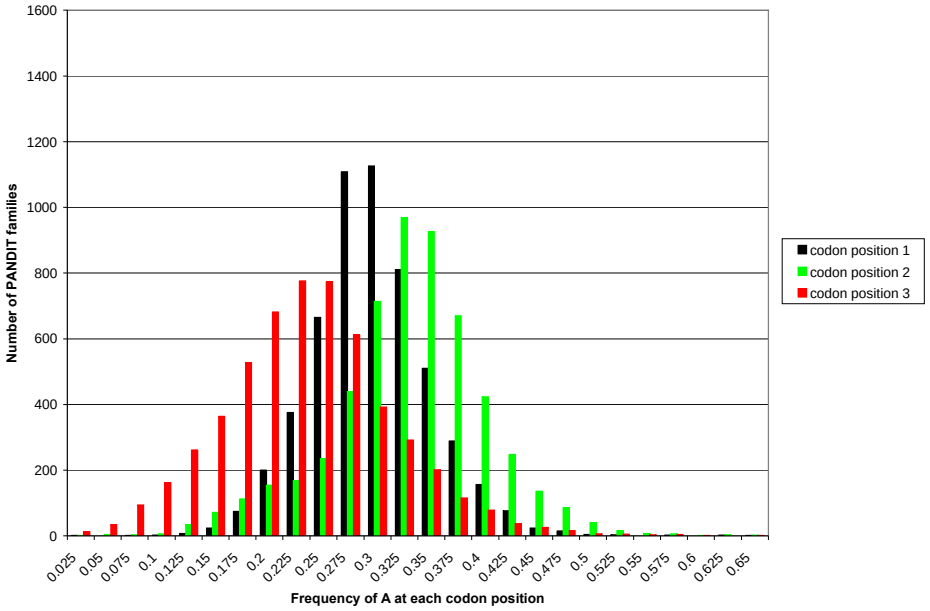


Figure 2.5b – Distributions of frequencies of nucleotide C.

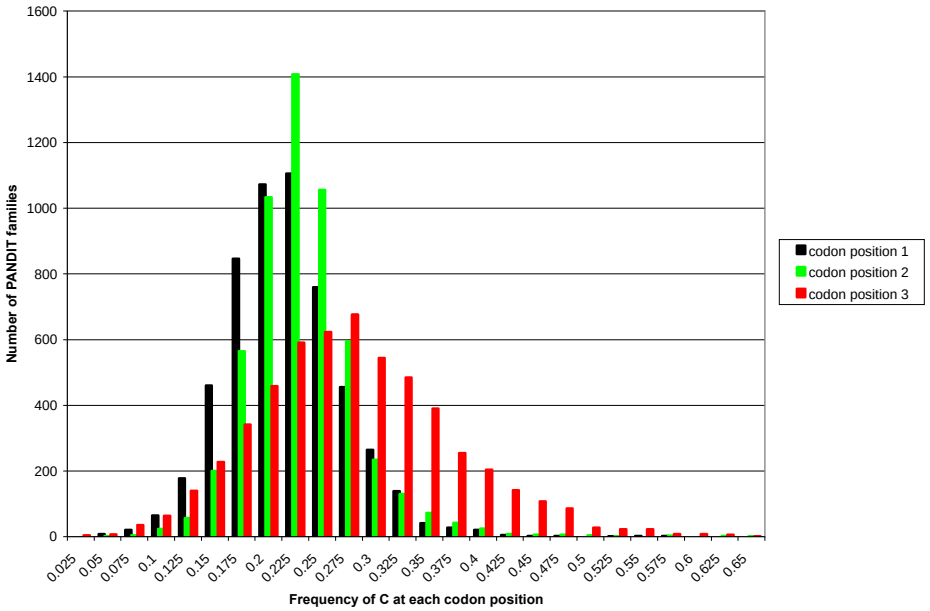


Figure 2.5c – Distributions of frequencies of nucleotide G.

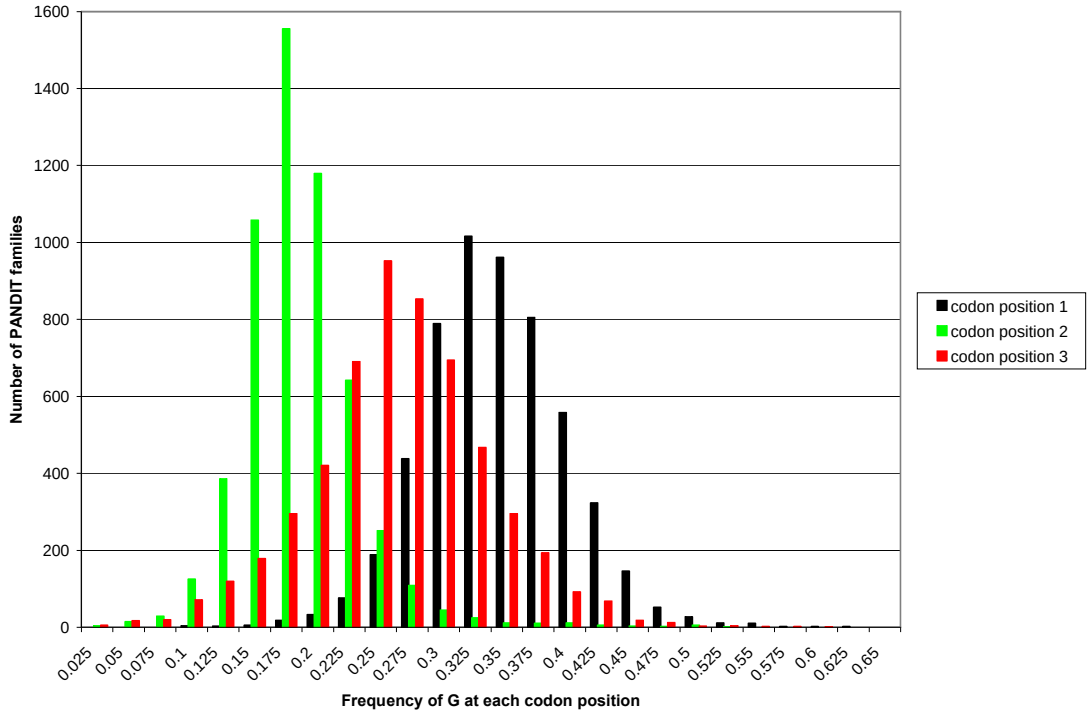


Figure 2.5d – Distributions of frequencies of nucleotide T.

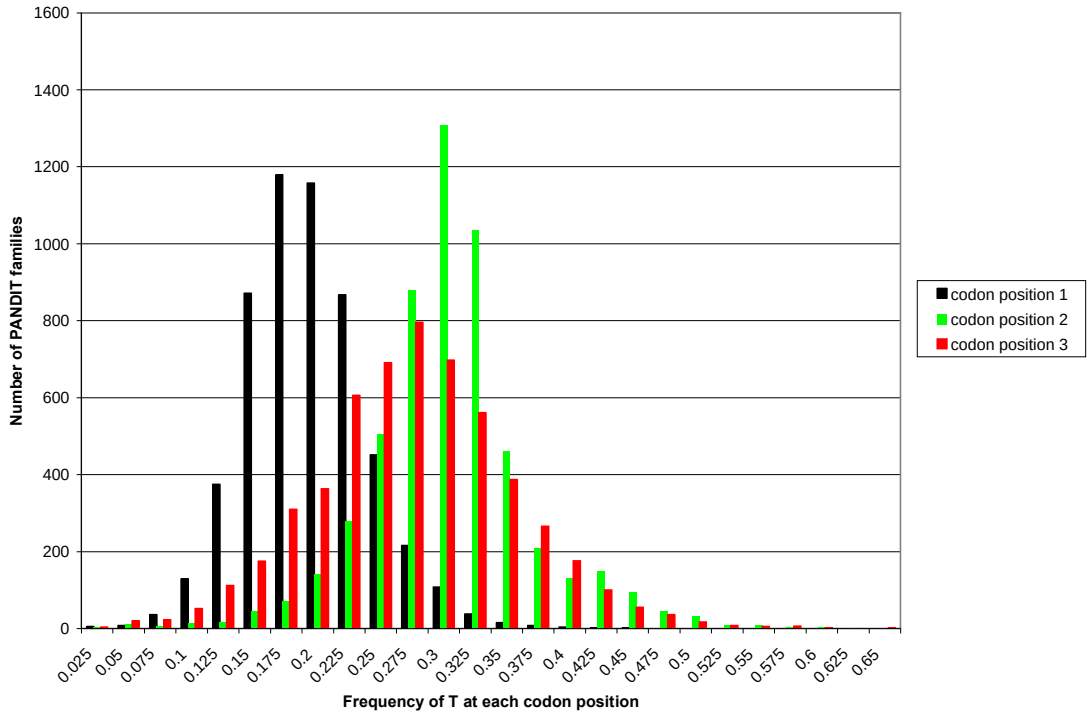


Figure 2.6 (a-c) – The distributions of nucleotide frequencies at the three codon positions taken from model HKY+R+N+T+G where test T-13 is significant. The different nucleotide frequencies are shown for each codon position in three separate sub-figures: codon position 1, codon position 2 and codon position 3.

Figure 2.6a – Codon position 1.

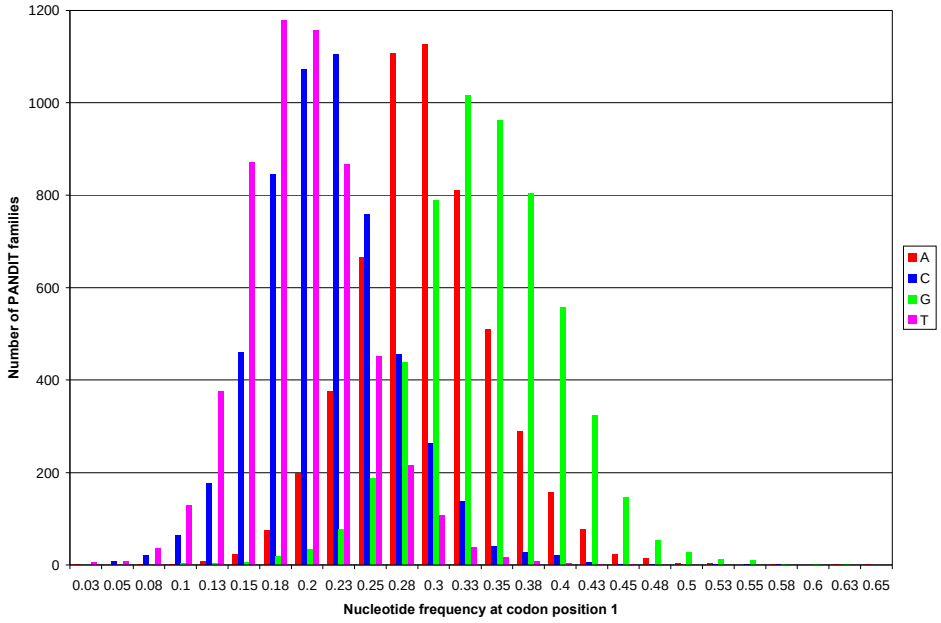


Figure 2.6b – Codon position 2.

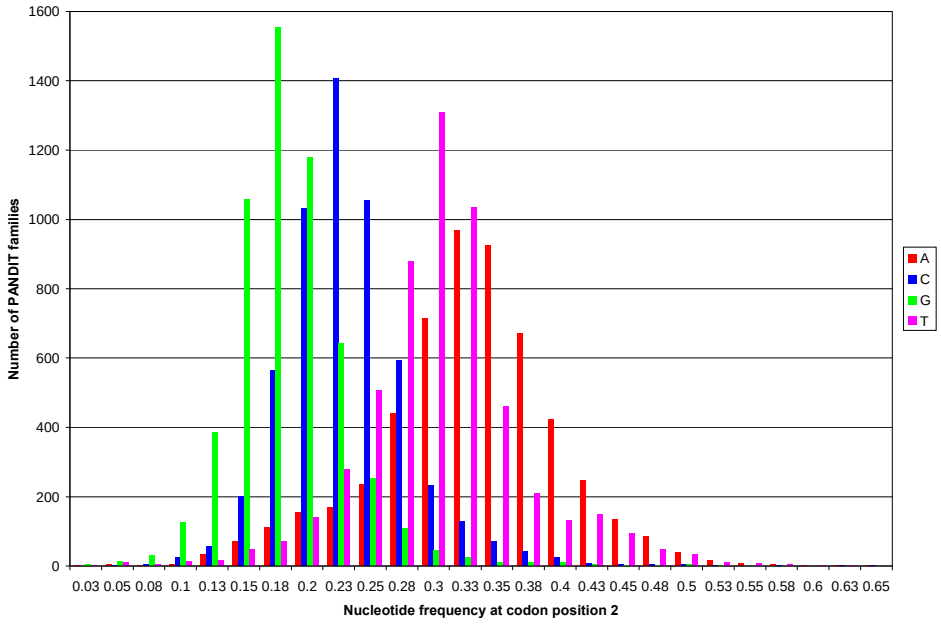
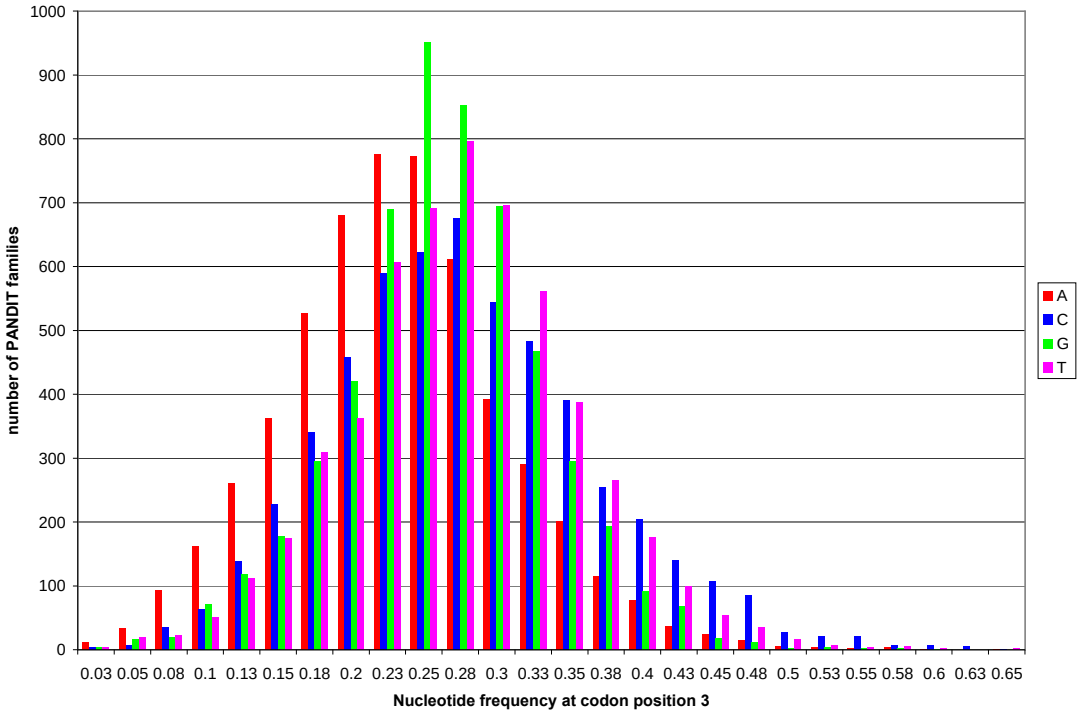


Figure 2.6c – Codon position 3.



There is a general trend with tests T-10 to T-14, that the more complex our null model, the longer the mean or median tree length of families where LRT results are significant. The mean and median tree lengths for the alternate hypothesis model for each of these tests are presented in Table 2.16. It is known that more complex models give longer branch length estimates since they estimate more multiple hits, i.e., hidden substitutions masked by more recent mutations (Page and Holmes, 1998). It may also be the case that longer tree length estimates could result from interactions between the amounts of evolution that the different parameters can explain. When additional factors are incorporated into the null and alternate hypothesis models for each test, less of the evolution of the data can be explained by the individual factor of interest alone, such as nucleotide frequency differences between the codon positions. Thus, a stronger effect of the differences in nucleotide frequency between codon positions may be needed to achieve significance as the null and alternate hypothesis

models get simultaneously more complex but this is unlikely to be relevant in this case, since the sets of PANDIT families that give significant LRT statistics are very similar between tests T-10 to T-14.

Table 2.16 – Tree lengths for significant LRTs from the alternate hypothesis models in tests T-10 to T-14.

Model	HKY+R+N	HKY+R+N +T	HKY+R+N +G	HKY+R+N +T+G	HKY+R+N +T+3G
Mean tree length	4.33	5.49	6.98	7.26	9.51
Median tree length	2.95	3.24	4.51	4.74	4.92

2. 7 Conclusions of Investigating Within Codon Heterogeneity

A large dataset has been used to study previously uncharacterised differences in the evolutionary patterns at different codon positions. I have studied the differences in evolutionary rate, rate heterogeneity, ts: tv biases and nucleotide frequencies between codon positions. Some of the causes for the differences in evolutionary rates between the codon positions are well understood (Massingham, 2002) and the results presented here are consistent with previous studies by Massingham. This study has used more advanced models and a tested a much wider range of hypotheses than any previous study.

The differences in rate heterogeneity between the codon positions are biologically intuitive and relate to the levels of functional constraint that we expect at the different codon positions. The second codon position has the lowest rate of evolution and the most rate heterogeneity. The functional constraints of the genetic code on the second codon position mean that most changes at the second codon position will be selected against. However, some of the nucleotide changes in the second codon position, which are all non-synonymous and lead to a new amino acid at that position in the polypeptide, will occur either because these changes are positively selected or their effects on protein function are close to neutral. Constraints on most second codon positions in a protein-encoding gene with a few second codon positions changing more rapidly leads to an extreme and low value of α, governing the shape of the γ distribution that is used to model rate heterogeneity. The third codon position sites are generally the least constrained and many evolve at similar, higher rates. Very few third codon positions are likely to be very constrained because of the redundancy of the genetic code. Thus, there is less heterogeneity of evolutionary rates at the third codon position, leading to a higher average α value. The first codon position has an intermediate level of functional constraint, again due to the degeneracy in the genetic code, and the rate heterogeneity in codon position one is intermediate between codon positions two and three on average.

Our estimates of the ts: tv bias at each codon position interact with whether or not we have accounted for differences in nucleotide frequencies between the different codon positions. The preferred models, judging by the majority of LRT statistics across the PANDIT database, are the '+N' models and therefore the alternate hypothesis models used in tests T-6 or T-8 give better estimates of the ts: tv biases at

the different codon positions than tests T-5 or T-7 do. Our best estimates of the order of ts: tv biases between the codon positions are 3 > 2 ≈ 1. The biases relate to the changes that are permitted after the average effects of selection and the genetic code. Transitions are much more common in nature than transversions (Topal and Fresco, 1976; Brown *et al*., 1982; Nachman and Crowell, 2000), reflected by the higher bias in the relatively unconstrained third codon position. The genetic code and stronger selection lead to a reduction in the mutational ts: tv evident at the first and second codon positions. It appears that either a relatively large proportion of the non-synonymous amino acid changes that are caused by transitions are selected against or the non-synonymous amino acid changes that are caused by transversions are not so strongly selected against. It is likely that the former explanation is correct because, on the whole, non-synonymous changes of any kind are selected against in protein-encoding sequences (the low ω value of many proteins; Yang and Bielawski, 2000). It is interesting that the ts: tv biases at codon positions 1 and 2 have similar distributions across the families in the PANDIT database. Aside from the interaction between selection and mutation, one might postulate that similar cellular mechanisms cause the similarity but this requires further detailed investigation.

The patterns of the differences in nucleotide frequencies between the codon positions follow my main predictions. The more mutagenic nucleotides G and C have low frequencies at the second codon positions, which is also the most constrained selectively. G and C frequencies are highest in the first and third codon positions, respectively. We should expect that the most mutagenic nucleotide, cytosine, which frequently undergoes deamination (Touchon *et al*., 2003), should be most common at the third codon position and this expectation is confirmed. The relative nucleotide

frequencies at the different codon positions reduce the level of mutations that are likely to cause non-synonymous changes because mutable nucleotides have lower frequencies in non-degenerate positions.

The implications of the results of the studies of the differences in parameter estimates at the three codon positions for families in the PANDIT database are discussed in Section 2.14.

2.8 Variation in Evolutionary Parameters in Overlapping Reading Frames

From the study of the PANDIT database we have a clearer impression of the distributions of parameter estimates of nucleotide-level evolutionary analyses that occur at different sites in the codon. In this section I describe the altered functional constraints found in some protein-encoding DNA sequences in certain viruses where the reading frame of one protein-encoding DNA sequence overlaps the reading frame of another. I analyse the evolutionary constraints of a hepatitis B virus (HBV) dataset that had been previously used to assess certain differences between overlapping reading frame codon positions (Yang *et al.*, 1995). I discuss our naïve expectations of the results of the analysis on overlapping reading frames and find that these are not met for the same sets of evolutionary parameters investigated in the PANDIT database study. I draw novel conclusions about this dataset and overlapping reading frames in general, with the noted caveat that the dataset is small and the results may not be representative of other overlapping reading frames.

Overlapping reading frames occur when the open reading frame of one protein uses some of the same nucleotides that are used in an open reading frame of a different protein. Overlapping reading frames are possible because of the three-nucleotide length of the codon. The functional constraints of both proteins in an overlapping reading frame are expected to affect the evolution of the sequence at a nucleotide level but the effects are not well characterised. One view of overlapping reading frames is: "Considering the double roles performed by sites in these [overlapping] classes, we should expect these three rate parameters to be less than 1…" (Yang *et al.*, 1995: 591), where the '1' refers to the relative rate that the first codon position of non-overlapping reading frames is set to and the 'three rate parameters' are the estimated evolutionary rates of the overlapping reading frame site 'classes'. The site 'classes' that Yang *et al.* refer to are their classes '4', '5' and '6', which are those codon positions where separate genes overlap at their codon positions 1 and 3, 1 and 2 and 2 and 3, respectively. I refer to the set of non-overlapping and overlapping sites as codon classes. The evolutionary constraints at any individual nucleotide will depend on which codon positions of the separate genes overlap.

The opinion that overlapping reading frames should be more constrained than their separate component positions stems from the multiple functions that each nucleotide position in an overlapping reading frame performs. The low ω values observed in many proteins (Yang and Bielawski, 2000) suggest that changes in most sites across a protein will be deleterious. If we consider proteins to be well-adapted to their function then a nucleotide change in an overlapping reading frame is likely to be deleterious overall, since it is unlikely that both proteins will not be functionally impaired by a nucleotide change. If the proteins in the overlapping reading frames are

well adapted then a change in one nucleotide that is non-synonymous and improves the functional ability of one protein may still cause a non-synonymous change in the other protein that may be deleterious. Of course, if many mutations are 'nearly neutral' (Ohta, 1992), then the idea of additional evolutionary constraint in overlapping reading frames may not be justified. The structure of the genetic code (the fact that many single nucleotide mutations that are non-synonymous lead to replacements by amino acids with similar physicochemical properties) may allow open reading frames to evolve at a rate comparable to normal open reading frames.

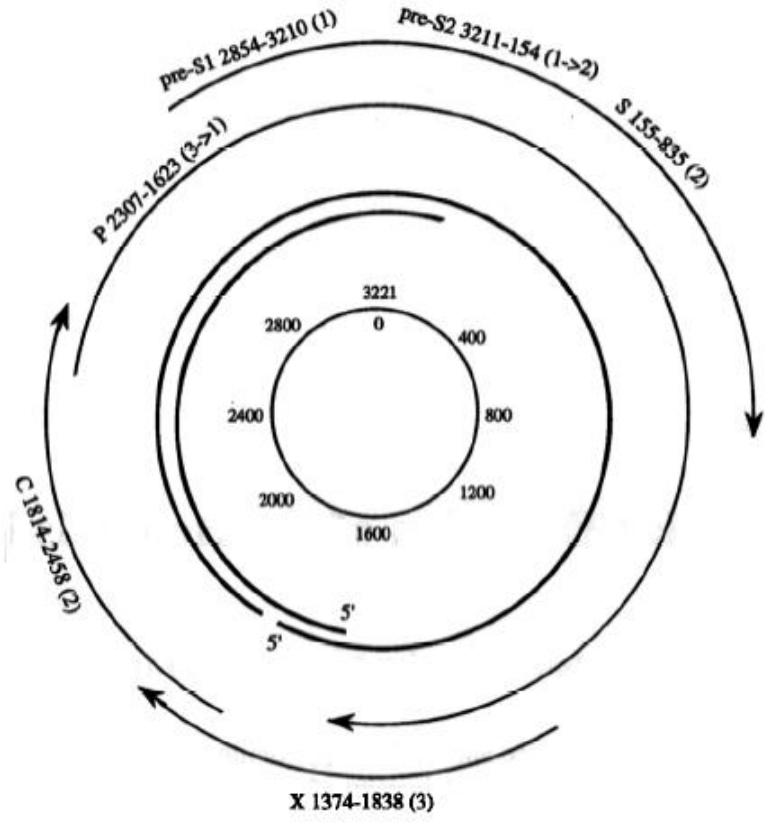
2.9 The Yang *et al.* (1995) Hepatitis B Virus Dataset

Yang *et al.* (1995) published the annotated HBV dataset that is used in this study. HBV is the smallest DNA virus that infects humans. Yang *et al.* aligned 13 complete HBV genomes to produce a 3,176 nucleotide alignment (a single insertion in one sequence was manually edited out by Yang *et al.*) with the start and end positions of each gene annotated. In total the HBV genome encodes six genes: three envelope proteins ('pre-S1', 'pre-S2' and 'S'), a viral capsid protein 'C', a DNA polymerase / reverse transcriptase encoding gene 'P', and gene 'X', which encodes a putative regulatory element. Almost exactly half of the HBV genome consists of genes with overlapping reading frames and the six codon classes are approximately equally represented in the genome.

The HBV genome is not the same as classic double-stranded DNA genomes (such as the human genome). The 'main' approximately 3,200-nucleotide strand of the genome, analysed in this investigation, base pairs with a complementary smaller

strand of 1,700-2,800 nucleotides. The smaller strand base pairs with either end of the main strand, creating a circular genome structure (see Figure 2.7). Much of the main strand remains single stranded. Thus, unlike the human genome, we should not expect the base frequencies of complementary bases (A and T, C and G) to be equal (the so-called parity rule; see Bell, 1999).

Figure 2.7 – The hepatitis B virus genome (reproduced from Yang *et al.*, 1995). Bold lines represent the two DNA strands (one of which is used in this study). Genomic regions encoding proteins S, X, C and P overlap and are differentiated by their reading frames (in parentheses).



Analysis of each gene or overlapping part of any gene is of limited worth because many of the parameters we might wish to estimate in our tests will have very large standard errors (see Yang *et al.*, 1995). Thus, all of the genes are concatenated

into a single dataset and analysed as if they formed one large gene, half of which is composed of overlapping reading frames. The specific evolutionary pressures affecting each gene may make this small set of genes unrepresentative of overlapping reading frame evolution in general but this is the best currently available dataset and these caveats are discussed in more depth in Section 2.12.

2.10 Analysis of the Hepatitis B Virus Dataset

The Yang *et al.* (1995) dataset was analysed in much the same way as the PANDIT families earlier in this chapter; all of the evolutionary models described in Section 2.4 and the subsequent tests T-1 to T-14 were performed on the concatenated HBV dataset. The tests are very similar to those previously applied to the PANDIT database but there are now six different codon classes (three of which relate to the combinations of different separate codon positions overlapping) instead of only the three standard codon positions investigated in the PANDIT database study. Having six codon classes, instead of three, affects the number of parameters to be optimised for each of the models and consequently the number of degrees of freedom that differ between the models used in tests T-1 to T-14. The phylogeny used for the HBV dataset was the main phylogeny used in the original study (Yang *et al.*, 1995).

2. 11 Results of Analysis of Hepatitis B Virus Dataset

I present the results of the LRTs for tests T-1 to T-14 and then discuss the parameter values for the most complex model, HKY+R+N+T+6G, for each of the classes of codon positions, including the three classes of overlapping reading frame

positions. All of the LRTs for tests T-1 to T-4 were significant at the 99% level on this dataset and these results are shown in Table 2.17. The 99% χ^2 significance point for tests T-1 to T-9 is 15.09 and for tests T-10 to T-14 is 30.58.

Table 2.17 – Results of tests T-1 to T-14 preformed on HBV dataset.

Test Number	Degrees of Freedom in LRT	Test Statistic
T-1	5	587.51
T-2	5	316.20
T-3	5	59.75
T-4	5	54.95
T-5	5	29.74
T-6	5	33.10
T-7	5	36.08
T-8	5	40.19
T-9	5	34.44
T-10	15	63.70
T-11	15	67.06
T-12	15	60.76
T-13	15	64.87
T-14	15	65.46

The fit of the evolutionary model to the data improves significantly for this dataset when our alternate model considers any of the following differences between codon site classes: evolutionary rate, rate heterogeneity, ts: tv bias and nucleotide

frequencies. Thus, I present the parameter values for the most complex model used in this study HKY+R+N+T+6G in the following Tables (Tables 2.18).

Tables 2.18 – Parameter estimates at the different codon classes; standard errors in parameter estimates are given where possible.

Codon Position Category	1	2	3	4 (1 + 3)	5 (1 + 2)	6 (2 + 3)
Rate	Set = 1	0.33 ± 0.07	3.92 ± 0.52	1.87 ± 0.28	0.66 ± 0.12	0.94 ± 0.16
α	0.27 ± 0.07	0.05 ± 0.00	1.02 ± 0.18	0.33 ± 0.06	0.14 ± 0.05	0.21 ± 0.05
κ	1.56 ± 0.28	1.97 ± 0.57	4.34 ± 0.45	3.18 ± 0.45	1.73 ± 0.39	2.92 ± 0.55
Nucleotide Frequency	A:0.25 C:0.24 G:0.25 T:0.26	A:0.27 C:0.25 G:0.18 T:0.29	A:0.24 C:0.20 G:0.21 T:0.35	A:0.21 C:0.30 G:0.24 T:0.26	A:0.21 C:0.31 G:0.23 T:0.26	A:0.18 C:0.32 G:0.21 T:0.29

The evolutionary rate results in Table 2.18 are very similar to the Yang *et al.* (1995) "F84 + C + dG" results. I note that the order of evolutionary rates for the non-overlapping reading frames from fastest to slowest is 3 > 1 > 2. This is the same order of rates as we observe across the PANDIT database study. Interestingly though, we do not observe reduced evolutionary rates in the overlapping reading frame sites, i.e., the most conserved of the six codon classes is not one of the three overlapping reading

frame classes. The evolutionary rates for the overlapping codon sites all have values that are intermediate to those of their component positions. For example, class 5 (where reading frames overlap at their first and second codon position) has an evolutionary rate = 0.66, which is less than codon position 1 (rate = 1) and more than codon position 2 (rate = 0.33); this observation hold for codon classes 4 (overlap of reading frames between codon positions one and three) and 6 (overlap of reading frames between codon positions two and three) and their respective 'component' codon positions.

The order of the evolutionary rates within the set of codon positions that constitutes the overlapping reading frames (classes '4', '5' and '6') is as one might expect. Of the positions that overlap a second codon position, classes 5 and 6, the position that also overlaps a third codon position (class 6) evolves faster than the position that overlaps with codon position one (class 5). This trend follows with other combinations of component codon positions in the overlapping reading frame (e.g., a position in an open reading frame that is a third codon position of one gene will evolve slower if it overlaps the second codon position of another gene, as opposed to a first codon position of another gene).

The rate heterogeneity measured by α is given a minimum value bound of 0.05 in the program PAML. Therefore, codon position 2 has extreme rate heterogeneity, with the vast majority of such sites evolving very slowly and only a few sites evolving faster. The observed order of rate heterogeneity parameters (α) in the non-overlapping codon classes of the HBV dataset (3 > 1 > 2) is the same as for the PANDIT database study (Section 2.6.2).

Just as the lowest and highest evolutionary rates were observed in codon positions two and three respectively, it is again the case that the most extreme rate heterogeneities are observed in non-overlapping codon categories; the second codon position has the lowest α value, the third codon position has the highest α value. Once again, as we have observed with the evolutionary rates, the rate heterogeneity parameter α for the overlapping reading frame categories is intermediate to the α values of the component separate codon positions. This holds for all cases and is discussed further in Section 2.12.

The differences in the ts: tv bias between codon positions that are not in overlapping reading frames have the same order as observed in the PANDIT study for the model HKY+R+N+T+3G (3 > 2 > 1), although median the ts: tv bias was only marginally smaller for codon position 1 than codon position 2 in the PANDIT study (Table 2.14). The phenomenon of overlapping reading frame categories having intermediate parameter values relative to their component codon positions that is observed in the analysis of evolutionary rate and rate heterogeneity is again seen for ts: tv biases. For example, codon position one has a ts: tv bias = 1.56 and codon position two has a ts: tv bias = 1.97; overlapping reading frame class 5, which consists of overlaps between the first and second reading frame, has an intermediate ts: tv bias = 1.73.

The HBV nucleotide frequency parameters are unusual (A = 0.227, C = 0.270, G = 0.219, T = 0.284; Yang *et al.*, 1995). Nucleotides C and T, the pyrimidines, are over-represented. This may be a consequence of capsid size limitations since pyrimidines are smaller than purines (Alberts *et al.*, 2002). Even though the HBV genome is CT biased, A and T remain the most prominent nucleotides in the second codon position, as observed in the PANDIT database studies.

From Table 2.18 it is clear that nucleotide A has a reduced frequency in overlapping reading frames compared to all of the codon positions that are not in overlapping reading frames. The comparatively mutagenic nucleotide C has an elevated frequency in overlapping reading frame categories and this may contribute to the fact that the evolutionary rates are not as constrained as Yang *et al*. (1995) expected. There is no such trend with nucleotide G and it appears that there is a small reduction in the frequency of nucleotide T in overlapping reading frames although this is not as marked as the differences in A or C.

2.12 Conclusions of the Hepatitis B Virus Analysis

I first discuss the results of the combined gene dataset in comparison to Yang *et al*.'s (1995) conclusions and then proceed to discuss the validity of combining the separate genes into a single dataset. Although combining separate genes into a single dataset may have an effect, I note that all overlapping reading frame categories are, in effect, combined gene datasets since the overlapping reading frame positions, in one way or another, evolve in a manner that is influenced by both of their component genes. It is a novel observation in this study to note that the evolutionary rates of

overlapping reading frame categories are *intermediate* to their component separate codon positions.

Yang *et al.* (1995) expected that the evolutionary rates in the overlapping site classes should be less than the rates in the non-overlapping codon positions. This was not found and Yang *et al*. (1995) concluded that the unexpected results were due to differences in rate ratios for the three codon positions of different genes and different overall evolutionary rates for each gene. I do not agree with this conclusion and propose that the results observed are a result of HBV biology as it continues to adapt to new hosts and evade the immune systems of the hosts, discussed later in this section.

The differences in the evolutionary rates of different genes will affect overlapping reading frames since they are presumably affected by evolutionary constraints on both of their component genes. If the overlapping reading frames were more constrained than their component non-overlapping codon positions, then the differences in the overall evolutionary rates of the separate genes should not have the observed effect. Furthermore, the differences in rate ratios of different genes should not cause the overlapping reading frame categories to appear as if they are evolving faster than they are (if they are evolving particularly slowly). By expanding the study of evolutionary parameters to observe differences in rate heterogeneity and ts: tv bias, I have been able to demonstrate that the intermediate parameter values of overlapping reading frame categories, relative to their component codon positions, is not restricted to evolutionary rate alone. It is not clear how differences in the evolutionary rates between different genes could cause intermediate values for the parameters related to

rate heterogeneity or ts: tv biases at the overlapping reading frame categories relative to the non-overlapping codon positions in the separate component genes. We are now aware from the study of the PANDIT database that the ts: tv bias of different codon positions is not directly related to the evolutionary rate alone since $3 > 2 \approx 1$ for the ts: tv bias but $3 > 1 > 2$ for the evolutionary rate at different codon positions. Thus, differences in the ts: tv biases between different overlapping reading frame categories and their component codon positions cannot be due to differences in the evolutionary rates between the different genes alone.

It is difficult to ascertain whether combining the separate genes into a single dataset compromises our results. We can observe the patterns of evolution at some of the separate genes in the dataset in order to see if the combined data results appear biased. It would be poor practise to perform this analysis for the more complex models as the improvement in model fit for many of the tests T-1 to T-14 is not significant for the separate genes. This may introduce additional inaccuracies in parameter estimation so it is best that we use simpler models that only observe differences in evolutionary rates at the different codon positions and sites in overlapping reading frames, as in the original Yang *et al.* (1995) paper. However, an additional caveat is introduced by using simpler models; we now know that not accounting for certain differences between codon positions can affect other parameter estimates and the order in which different codon positions rank for those parameters.

I reproduce results originally presented by Yang *et al.* (1995) in Table 2.19, where differences in rates at the different site classes for each gene (separate codon positions and overlapping reading frame sites) were assessed with a REV model that allowed for rate differences between sites and a single discretised γ distribution across all sites to account for some level of rate heterogeneity. It is unlikely that the use of the REV model (instead of HKY; see Chapter 1, Section 1.2.3.1) makes any differences in rank order of rate estimates. The letters in Table 2.19 correspond to the gene name and the subscript numbers correspond to the codon position, e.g., category 'C_3' is all of the non-overlapping third codon position sites for gene 'C' and $P_1(preS)_3$ are the positions where gene 'P' overlaps gene 'preS' at their first and third codon positions respectively. The last three categories ($9C_1X_3 + 48P_3C_1$, $8C_2X_1 + 48P_1C_2$ and $8C_3X_2 + 48P_2C_3$) are made up of more than one class of overlapping sites because there are very few sites in some separate categories.

Table 2.19 – Evolutionary rate differences at different codon classes for each gene. Results reproduced from Yang *et al.* (1995).

Gene and Codon Site Class (Gene $_{\text{codon position}}$)	Number of Sites	Evolutionary Rate and Standard Error
P_1	309	1
P_2	309	0.322 ± 0.075
P_3	309	4.434 ± 0.686
X_1	63	1.213 ± 0.360
X_2	64	0.539 ± 0.199
X_3	63	1.668 ± 0.455
C_1	156	0.677 ± 0.164
C_2	156	0.229 ± 0.081
C_3	156	3.025 ± 0.559
$P_1(preS)_3$	163	3.188 ± 0.579
$P_2(preS)_1$	163	1.316 ± 0.266
$P_3(preS)_2$	163	1.106 ± 0.231
P_1S_3	227	0.903 ± 0.180
P_2S_1	227	0.324 ± 0.085
P_3S_2	227	0.761 ± 0.158
P_3X_1	84	2.135 ± 0.500
P_1X_2	83	0.378 ± 0.143
P_2X_3	83	1.055 ± 0.291
$9C_1X_3 + 48P_3C_1$	57	0.295 ± 0.145
$8C_2X_1 + 48P_1C_2$	57	0.060 ± 0.061
$8C_3X_2 + 48P_2C_3$	57	0.306 ± 0.149

The results presented in Table 2.19 do not agree with the notion that overlapping reading frames are necessarily more constrained than non-overlapping codon positions. I first consider the case of the preS gene, which is completely contained within the P gene. Positions $P_1(preS)_3$ and $P_2(preS)_1$ both evolve significantly more quickly than positions P_1 and P_2 (3.188 > 1 and 1.316 > 0.322 respectively). We may also use P_3X_1 and P_2X_3 as examples; both of these overlapping reading frames evolve at rates that are intermediate to their separate component codon positions i.e., $P_3 > P_3X_1 > X_1$ (4.434 > 2.135 > 1.213) and $X_3 > P_2X_3 > P_2$ (1.668 > 1.055 > 0.322). There are however, examples of overlapping reading frames that are more constrained than their separate component codon positions in non-overlapping reading frames (such as P_1X_2).

The results based on the concatenated dataset lead me to conclude that the HBV overlapping reading frames generally have evolutionary parameters that are intermediate to their component codon positions of separate genes when they are not overlapping. However, whilst I have argued for the validity of the use of the concatenated HBV dataset, we should be even more careful if attempting to draw more general conclusions about the evolution of all overlapping reading frames from this single dataset alone. At the very least, it is clear that increased evolutionary constraint is not a necessary consequence of overlapping reading frames.

There are realistic biological reasons as to why genes with overlapping reading frames should not evolve slower than their separate genes in viruses. Viruses are parasites and their survival depends on evading the immune system of their host. The host immune system has evolved to recognise non-self motifs, antigens, and to

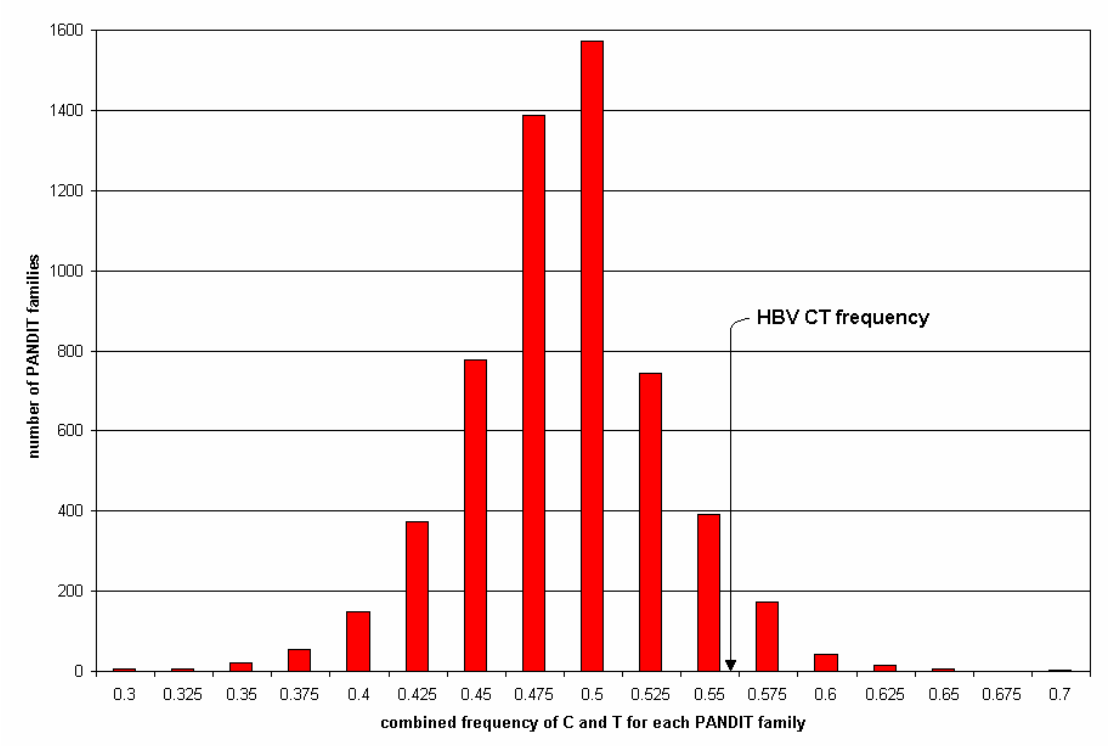
quickly remove particles presenting such motifs. Viral particles are typically presented on MHCI molecules at the surface of infected cells (Alberts *et al*., 2002) and this stimulates an immune response; many viral particles can be antigenic, not just those exposed at the surface of a viral capsid or envelope. Furthermore, a virus must deal with polymorphic genotypes of the host population and the virus may need to evolve rapidly every time a new host is infected. Thus, it is clear that constrained evolution caused by the functional constraints of overlapping reading frames may not benefit the virus in a selective sense. The evolution of overlapping reading frames is clearly influenced by both genes and very slow evolution of either gene may not benefit the virus.

The causes of differences in nucleotide frequencies between the different codon positions and overlapping reading frame positions may be complex. Clearly, there are significant differences in the nucleotide compositions between overlapping reading frames and non-overlapping codon positions. It is likely that the requirements of the genetic code play an important role in the reasons for our observations.

The rank orders of median parameter estimates for the evolutionary rate, rate heterogeneity and ts: tv bias for the first three codon categories of the HBV study are the same as the separate codon positions in the studies across the PANDIT database. Comparing the nucleotide frequencies at the different codon positions in the HBV and PANDIT studies, there are both similarities (A and T are most frequent nucleotides at codon position 2, G is most frequent at codon position 1) and differences (C is more frequent at codon position 3 in the PANDIT study, whereas it is most frequent at codon position 2 in the HBV study). However, the combined C and T frequency of

the pyrimidine-biased HBV dataset falls well within the range of the families used in the PANDIT database study (see Figure 2.8); some PANDIT families have their highest C frequencies at codon position 2 (although this is rare, data not shown).

Figure 2.8 – The distribution of combined C and T frequencies of all families (that successfully optimised for all tests) in the PANDIT database.



2. 13 Future Directions

In the study performed upon the PANDIT database it would be interesting to categorise PANDIT families by their functions (e.g., immune system proteins, house-keeping enzymes etc.) and observe whether different groups of proteins evolve in significantly different ways, according to codon positions. We might expect the *relative* rates of the third codon positions to be much higher in ancient enzymes that are involved in DNA replication, say, than proteins involved in immune recognition, which generally evolve faster at the amino acid level. It would be currently unsatisfactory to perform these analyses on the different PANDIT families because the annotation of many of the families is not adequate. This is likely to change in the near future as protein annotation is a fast-moving field. It would also be interesting to carry out biochemical studies to ascertain if there are any underlying biological reasons why the ts: tv bias distributions are so similar across the PANDIT database for the first and second codon positions.

The results of this study are used in further chapters of this thesis and the differences in evolutionary rates, rate heterogeneity, ts: tv biases and nucleotide frequencies at the codon positions is used as a basis for a simple nucleotide level model that identifies CDSs in genomic-scale multiple alignments. Finally, the distributions of parameters that have been provided can be used as fair prior estimates of parameter values for studies in the Bayesian framework.

It would be very interesting to pursue the study of overlapping reading frames on other datasets, but none are currently available.

2.14 Overall Discussion

I have introduced basic concepts of genes, the genetic code, transcription and translation in order to discuss different functional constraints at the different codon positions and how these may cause heterogeneity in intra-codon evolution. I have discussed how we may use models to analyse several different markers of evolutionary change at the different codon positions and have focused on detecting differences in evolutionary rate, rate heterogeneity, ts: tv biases and nucleotide frequencies at the different codon positions. I then discussed tests between the models and how to determine which model parameters differ significantly between the codon positions. The tests were performed on a very large dataset, the PANDIT database. It is apparent that there is previously unconsidered variation in evolutionary parameters between codon positions and novel observations were commented upon. The differences in the way in which the codon positions evolve has been related to biological phenomena, notably the functional constraints of the genetic code and selection, predominantly for retaining functional residues at specific sites. The different evolutionary parameters have been discussed in terms of how much they vary between codon positions, how many families in the PANDIT database they are significant for and how the order of the parameters at different codon positions may depend on the evolutionary models we use. These results also have a bearing on interactions between evolutionary parameters and which models are most appropriate to use. I have also offered parameter distributions and suggested how these might be usefully applied in future investigations in a Bayesian framework. The knowledge gained in this study is used in further investigations described in this thesis (in particular the design of models in Chapter 3).

The distribution and functional constraints imposed by overlapping reading frames was then considered. The evolutionary properties of overlapping reading frames were investigated using models similar to those used in the PANDIT analysis. An HBV dataset was introduced that was previously used for a similar study (Yang *et al.*, 1995). Analysis of these sequences indicated that factors other than evolutionary rates could differ between overlapping reading frame codon sites, and I used these findings to draw novel conclusions that the values of the evolutionary parameters of overlapping reading frame codon sites seem to be intermediate between the values of their separate codon positions is a novel observation. Overlapping reading frames are not necessarily more conserved than their component codon positions. The adequacy of the HBV dataset was discussed and I noted the problem of obtaining other datasets of the same quality as the Yang *et al*. (1995) dataset for the study of overlapping reading frames.

Chapter 3: Using Tailored Models of Evolution for Protein-Coding Sequence Identification

Contents

3.1 Introduction

In this chapter I use the results from Chapter 2 to devise and test a new method to identify CDSs in multiple alignments, called Periodic Pattern Identification (PPI), which is implemented in the likelihood framework. I start by introducing issues related to genomic organisation and the challenge of identifying protein-encoding genes by computational methods. I discuss various published computational methods that are used to identify candidate CDS regions in both single DNA sequences and multiple alignments. I discuss how fledgling evolutionary methods have been implemented as hidden Markov models (HMMs) and how methods that have used large codon matrices to try and identify CDSs have lacked power because of the large number of parameter values that must be estimated.

I introduce PPI, a nucleotide-level modelling strategy, and apply the method to a genomic multiple alignment of four yeast species that I have produced based on the separate multiple alignments provided by Kellis *et al*. (2003). Significance testing is done on multiple windows at multiple positions in the alignment. I consider the implications of the significance of test scores when multiple tests are carried out and suggest two approaches that can be taken to combat this problem. The first approach is the determination of cut-off scores when the method is applied to a well-annotated dataset. In doing so, I validate the method as a powerful discriminator between CDSs and non-CDSs in multiple alignments. The cut-off scores derived from one aligned yeast chromosome are then applied to tests on another aligned yeast chromosome. The second approach uses the Bonferroni correction to make the point at which we consider an LRT statistic to be significant more stringent. I assess whether transcribed

but unannotated *Saccharomyces cerevisiae* genomic segments identified in a microarray study (David *et al*., manuscript in preparation) have evolutionary properties that are indicative of CDSs and apply the PPI method with the Bonferroni statistical correction for multiple tests. Many of these segments are clearly not CDSs but four candidate regions are identified for further study as potentially novel CDSs.

3.2 Genome Structure and Gene Identification

Protein-coding genes in the genomes of higher eukaryotes often have the intron-exon structure described in Chapter 2.2. In 'lower' eukaryotes, such as the yeast *Saccharomyces cerevisiae*, and in all archaebacteria and eubacteria, protein-coding genes are made of a single open reading frame. Different species have genomes that differ not only in how they encode proteins but also in the percentage of the genome that is composed of protein coding genes. In humans, *Homo sapiens*, less than 1.5% of the genome is composed of protein-coding genes (International Human Genome Sequencing Consortium, 2001). It is estimated that there are about 30000 (non-RNA) genes in the human genome, although different estimation methods affect this figure considerably. In the fruitfly, *Drosophila melanogaster*, roughly 24.1 Megabases of the total ~180 Megabase genome corresponds to CDSs (Adams *et al.*, 2000) (13.4% of the genome). Other model organisms such as *Caenorhabditis elegans* and *Arabidopsis thaliana* are estimated to have similar numbers of protein-encoding genes in their genomes as *D. melanogaster* (in the range of 11000-15000) although the overall sizes of these different genomes may also differ (The Arabidopsis Initiative, 2000). For example, the genome size of *A. thaliana* is only ~125 Megabases, over 50 Megabases smaller than the fruitfly genome. *S. cerevisiae* is

estimated to have between 4700 and 5800 coding open reading frames, depending on which source one uses (Kowalczuk *et al.*, 1999). The *S. cerevisiae* genome is much smaller than the human or fruitfly genome and very roughly, about 50% of the genome is protein-coding. The proportion of eubacterial genomes that are protein coding is usually very high, in the 90-99% range according to some predictions (Audic and Claverie, 1998).

Protein-encoding gene identification is important for the development of new therapies where mutant forms of genes cause disease and in the annotation of sequenced genomes, so we can better understand their structure and evolution. To this end many different computational methods have been developed to try and identify genes (specifically protein-coding genes in the context of this thesis chapter) when scanning across large genomic regions. Before introducing a novel method, based on the insights into codon evolution gained in Chapter 2 of this thesis, I briefly review some of the more popular approaches that have been used recently in the field of gene identification and annotation. I note that whilst the computational methods are very important in the identification and annotation of genes, the annotation of most human genes (and other species) is based on cDNA sequence data (Zhang, 2002). The cDNA sequences are produced by reverse transcription from mRNA sequences that have been isolated directly from cells that are actively undergoing the transcription process. These cDNA sequences are then used to identify their corresponding gene sequences in the genome. Thus, most genes have been identified through laboratory-based techniques and not through strict computational methods. Furthermore, computational methods are complementary to laboratory research and sequences that are classified as genes by computational methods still require validation through laboratory studies.

Many computational methods have been published that identify putative CDSs with varying degrees of success. Whilst some gene-identification programs are based on finding splice sites of introns, which necessarily occur very near to the end of exons, and gene promoters, I will focus the discussion on those computational methods, like the novel method I will introduce, which identify CDSs. Scanning sequences for several features at the same time, such as splice sites, exons and promoters, can improve the accuracy and power of gene identification methods. The protein-encoding sequence identifiers that I will discuss include those that are based on single sequences, alignments and comparative genomics (Zhang, 2002; Kellis *et al.*, 2004). The method that I introduce in this chapter is based on comparative genomics. It is beyond the scope of this chapter to even begin to attempt a fully comprehensive review of all the gene identification tools that are available; for a more complete review see Zhang (2002). Note that there are additional tools available, such as QRNA (Rivas and Eddy, 2001), for the identification of RNA genes that do not produce functional protein molecules.

There are many ways to identify either small regions containing CDSs in a large genomic region or the CDSs themselves because of the features that are associated with genes for biological reasons. In vertebrates for example, genes are commonly found in regions with high numbers of CG dinucleotides, so called CpG islands. These are uncommon outside of gene encoding regions because the CG dinucleotides are methylated and undergo rapid deamination, which does not occur in the so-called CpG islands (Gardiner-Garden and Frommer, 1987; Nachman and Crowell, 2000). Additionally, promoters and enhancer sequences can be used

alongside intron splice sites, the sites where introns are cleaved out of the early RNA transcript in the nucleolus, to identify gene regions and CDS boundaries. Content measures are commonly used to discriminate CDSs from other sequence. Frame specific hexamer frequencies are able to detect the majority of long protein-encoding genomic regions (Zhang, 2002). Content measures work because adjacent nucleotides do not occur independently of each other in CDSs, but tend to occur as 'words'. Some words are especially enriched in CDSs and hexamers, sequences of six nucleotides, are commonly used to identify such words. Scores for different 'words' can be used in combination to identify longer sequences that are enriched in CDSs.

Start codon and stop codons can also be used to identify CDS boundaries. Although this is of limited use for identifying internal exons in species that have genes with an intron-exon structure (Figure 1.1, Chapter 1), it can help to identify entire CDSs in species such as *Saccharomyces cerevisiae* that have genes mostley without introns. Some apparent start and stop codons may just be random sequences present in non-coding DNA so using these codons alone may identify spurious exons. Some methods recognise protein-coding regions typically by finding lengths of nucleotides that lack stop codons that are too long to have occurred by chance (Kellis *et al.*, 2004). Other features, such as the poly-A tract found at the 3' end of genes in many species, can be used to identify the ends of genes. Combining these features with each other and also with cDNA and other non-computational gene-identification tools enables us to identify many genes accurately with a single genomic sequence.

Alignment methods attempt to produce pairwise or multiple alignments from the sequences of homologous regions in different species and pre-existing annotations

from one species are then transferred to the other species. Comparative genomics methods often use conservation within a region of the alignment to identify a CDS (Kellis *et al.*, 2004). The use of the distribution of gaps in an alignment, caused by insertion or deletion events, can provide a powerful tool to identify CDSs. If CDSs are conserved across species it is quite rare that there will be gaps at these positions in a high quality multiple alignment, especially if these gaps are not in multiples of three due to selection for preservation of the correct reading frame in the CDS within each species (Kellis *et al.*, 2004).

Many of the more recent gene-finding models, including some that function on a single sequence and almost all that identify genes using comparative genomics, use fully probabilistic state models, HMMs. In an HMM the DNA sequence is partitioned into different states, such as exon and non-exon. HMMs rely on probabilistic models that use information in the patterns of sequence states along sequences. Our 'observations', the nucleotides in a single sequence or the pattern of nucleotides in an alignment column, are assumed to be caused by each state (e.g., a region that has evolved as an exon vs. 'junk' DNA). As an HMM that analyses a single sequence 'reads' along the sequence it assigns a probability that the sequence is in a particular state, such as an exon, based on the nucleotide at that site and the probability of being in an exon at the previous site (or sites, as any number can be theoretically taken into account). An HMM that analyses multiple sequences bases the probability of being in a certain state on the pattern of nucleotides in the alignment column plus the probability of states at the previous site, as before. Thus, an HMM-based method can identify regions that have a high probability of being an exon and regions that have a low probability of being an exon. An HMM is not limited to having just two states: it

can have many, which might correspond to untranslated regions, promoters, CDSs, introns and intergenic regions, for example.

For HMMs to be able to identify the patterns that different states of a sequence exhibit (and the patterns that occur as one state ends and another begins) they require training. A training set of representative and annotated data is used to assign probabilities to the different states (i.e., which patterns of nucleotides are indicative of which states and of moving between states). These probabilities, which relate to different states in the training set, are then applied to a test dataset. The success of an HMM in accurately identifying regions in different states depends on how well the model can be trained. If the training dataset is very different from the test datasets that the model is later applied to, the HMM may perform poorly. There are several popular HMM-based programs for gene identification including TWINSCAN (Korf *et al.*, 2001), DOUBLESCAN (Meyer and Durbin, 2002) and SHADOWER (McAuliffe *et al*., 2004).

TWINSCAN analyses a pair of aligned sequences and categorises alignment columns into one of seven categories, only one of which corresponds to CDSs (Korf *et al.*, 2001). DOUBLESCAN also functions on pairs of sequences and performs a pairwise alignment and annotation at the same time. The DOUBLESCAN HMM incorporates 54 states, including states for start and stop codons and states that consider whether aligned codons code for chemically similar amino acids or not (Meyer and Durbin, 2002). SHADOWER incorporates the information contained in phylogenetic trees to analyse sequences from multiple, closely related organisms. SHADOWER considers the phase of the first and last three nucleotides in a CDS such

that adjacent (non-terminal) CDSs start at the correct phase. Internal nucleotides of a CDS are all considered as a single state (McAuliffe *et al*., 2004). Other HMMs used to identify genes may use different states to describe CDSs. The gene-finding model GENIE (Kulp *et al.*, 1996) uses codon frequencies conditioned on the frequency of the three nucleotides in a window and the preceding codon.

HMMs are particularly suited to comparative genomic studies because the patterns that CDSs display in multiple alignments are even more striking than those seen in single sequences. This is predominantly because of the high degree of conservation shown by most exons, relative to non-exonic sequences, which is useful information in terms of modelling different model states. However, a recent HMM, which incorporates a full 64 state codon model (see Chapter 1.2.5) for CDS states and is capable of handling any number of genomic sequences did not perform particularly well in reliable CDS identification on the dataset used because of the very large number of parameters that need to be estimated from the training dataset (Pedersen and Hein, 2003).

There are many challenges that our computational methods must deal with as they continue to develop. Pseudogenes, regions of a genome that correspond to ancient genes that have lost their function and no longer encode functional genes, pose a problem for many gene finding methods since many pseudogenes retain features that make them appear as if they were still functional genes. Processed pseudogenes correspond to mRNA sequences that have been reverse transcribed to DNA and integrated into the genome. They appear as if they were normal genes but lack introns (in those species that this is relevant to) and they also lack promoters and

enhancers and are often found far from the CpG islands that are normally associated with complete genes. Processed pseudogenes may often be identified incorrectly as long internal exons of a complete gene that has several exons, where the terminal exons are usually associated with more of the common gene-associated traits (such as the poly-A tract and promoters). Nonsense mutations, mutations in the non-functional sequence that appear as a stop codon in the middle of a CDS, are most commonly used to identify pseudogenes but many pseudogenes lack them and cannot be identified in this way. Current gene-prediction programs are biased towards intron-containing genes and it is thought that they may miss many intronless genes (Zhang, 2002).

3.3 Novel Models to Identify CDSs

The comparative method I now introduce that identifies CDSs in multiple alignments is not biased towards genes with introns and does not require a training set *per se*. The models I introduce are implemented in the likelihood framework and require far fewer parameters than many HMM methods. There are very good reasons to continue using HMMs that incorporate the evolutionary information of a phylogeny (evolutionary HMMs) in gene identification and the model that I present below, as well as being used in its own right, could be incorporated into the CDS state models that form part of HMMs that use comparative sequence data. Just like other comparative genomic methods, the method I introduce here, called Periodic Pattern Identification (PPI), requires a multiple alignment of reasonable quality where genes are preserved across species and the CDSs align reasonably well with each other. All comparative genomic methods, by their very nature, require feature conservation

across a reasonable number of species in order to identify any feature. A poor multiple sequence alignment will act to reduce the signal that we are able to detect.

In Chapter 2 I found that there are strong patterns of variation in properties such base frequencies and the ts: tv bias, which had not been previously considered in this context, and parameters describing evolutionary rate (also described by Massingham, 2002) at different codon positions; this will only occur in protein-coding DNA. No-one has previously tested whether there is enough information in comparative sequences to identify coding DNA using this signal alone. Indeed, none of the HMM methods described earlier have used the periodicity in evolutionary rates due to different codon positions to describe CDS states. This is perhaps unsurprising since phylogenetic trees, which help to calculate evolutionary rates, have only very recently been incorporated into HMMs with the aim of detecting CDSs (McAuliffe *et al*., 2004). By testing for the periodicity of these evolutionary signals at the nucleotide level, instead of at the codon level, we can use models with relatively few parameters to identify CDSs with more efficient use of the data.

Let us suppose that our analysis takes a small window of DNA sequence in a multiple alignment, say of 100 nucleotides in length, and asks whether this window demonstrates CDS-like evolution. We form two competing hypotheses: the null hypothesis states that the window does not evolve with CDS-like properties and the alternate hypothesis states that it does. We apply two evolutionary models to that sequence: (1) a null hypothesis model (null model) that assumes all nucleotides in that window evolve according to the same evolutionary pattern and (2) an alternate hypothesis model (alternate model) that assumes codon-like evolution. The alternate

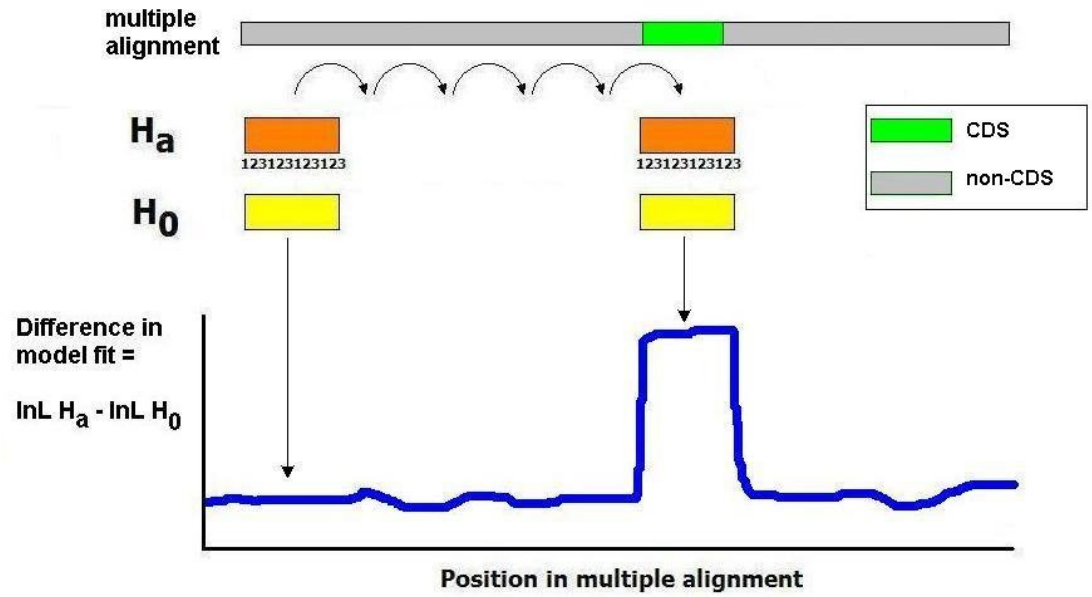
model postulates that features like the evolutionary rate are similar at all sites in the window that are three nucleotides apart, which is the same as the periodicity we would expect to see in a series of codons (and in fact we do see, according to the results of the research across the PANDIT database in Chapter 2). The alternate model has three categories of sites, which would correspond to the three different codon positions if the window is fully embedded in a CDS. Thus, three rates are allowed in the window, one for each of three categories of site. It is not important which actual category of sites in our model falls at which codon position in a real CDS: the method will make no assumption about which site category corresponds to the first, second and third positions of codons. The method does assume that every third alignment position is in the same category and that the categories have different evolutionary dynamics. The models need not just consider rate differences between the site categories; they can also consider all of the other factors that have been previously considered in Chapter 2, such as rate heterogeneity, ts: tv biases and nucleotide frequencies.

For our 100 nucleotide window of the DNA multiple alignment, let us consider two scenarios. Let us consider first the behaviour of the null and alternate models when the window is entirely in a non-CDS region and second, when the window is entirely within a CDS. When the window is contained entirely in a non-CDS region the null model explains the evolution of the data reasonably well and adding the extra categories that, for example, estimate the evolutionary rate of every third site in the sequence with a single parameter (for each of the three site categories) does little to improve our fit of the model to the data. However, if our window is entirely contained in a CDS, the null model gives a poor explanation of the evolution

of the data whereas the alternate model provides a much better explanation of the evolution of the data. Thus, only when our window is in a CDS will the difference in model fit between the alternate and null models be high (see Figure 3.1). As we proceed along a multiple alignment with a sliding window (subsequent non-overlapping windows of DNA data), we should expect to find regions of high and low differences in model fit, i.e., high and low test scores, corresponding to CDSs and non-CDSs respectively. Sliding window analyses are popular methods for analysing small regions of large alignments at a time and have previously been used to detect recombination between regions of an alignment (Grassly and Holmes, 1997).

Where a window crosses a CDS boundary we should expect a test statistic value that is intermediate between the expected scores for when a window is in a CDS and when it is not in a CDS. This is because the alternate model describes the evolution of some of the window better than the null model but it does not describe the evolution of the data well for the entire window. The actual test score will depend on how much of the window corresponds to a CDS. Thus, to ensure a lower number of windows that cross CDS boundaries it is desirable to use small windows. However, the benefits of using small windows must be weighed against the size of window that is needed to obtain decent resolution between CDS-like and non-CDS-like evolution. The step size of the window is commonly equal to the window size, such that all windows of data are non-overlapping and any single alignment column is used only once. Overlapping windows may also be used although this may affect how we approach the issue of multiple tests, since the higher number of tests does not use any extra data.

Figure 3.1 – The behaviour of the null and alternate models as our DNA window progresses along a multiple alignment. H_a and H_0 correspond to the alternate and null models applied to the data respectively. As the window passes through a CDS the difference in model fit between H_a and H_0 should increase significantly.



In the case where our alternate model has three categories for evolutionary rates, our null model is nested in the alternate model and there are two degrees of freedom between the models. Thus, since the models satisfy the necessary conditions, we can perform a likelihood ratio test (LRT) between the models where twice the difference between the log-likelihoods of the models is distributed according to a χ^2 distribution if the null model is in fact the true model (Chapter 1.4.1). This is true for any one randomly-chosen comparison but issues of multiple testing must be addressed. Table 3.1 presents the models that will be applied to our data in a similar format to Table 2.2 in Chapter 2. Table 3.2 presents the *tests* that will be performed between the models in a similar format to Table 2.3 in Chapter 2.

Table 3.1 – The evolutionary models that will be applied to each window of DNA data in the multiple alignment.

Model Name	Used for H_0 or H_a	Free Parameters	Model description and parameters that are allowed to differ between the three site categories (1, 2, 3, 1, 2, 3, 1, 2, 3 etc)
HKY	H_0	4 + tree	HKY model (Chapter 1, Section 1.2.3.1)
HKY+G	H_0	5 + tree	HKY model plus a single discretised γ distribution
HKY+R	H_a	6 + tree	HKY and estimates an evolutionary rate for each codon position (with the same proportions between branch lengths of the tree)
HKY+R+N	H_a	12 + tree	HKY and estimates an evolutionary rate and nucleotide frequencies for each codon position (nucleotide frequencies are estimated by counting)
HKY+R+N+T	H_a	14 + tree	HKY and estimates an evolutionary rate, nucleotide frequencies and ts: tv bias for each codon position
HKY+R+G	H_a	7 + tree	HKY plus a single discretised γ distribution and estimates an evolutionary rate for each codon position
HKY+R+N+G	H_a	13 + tree	HKY plus a single discretised γ distribution and estimates an evolutionary rate and nucleotide frequencies for each codon position
HKY+R+N+T+G	H_a	15 + tree	HKY plus a single discretised γ distribution and estimates an evolutionary rate, nucleotide frequencies and ts: tv bias for each codon position

I decided not to implement any model that uses more than one γ distribution because rate heterogeneity between codon positions was found only to be significant in between 23% and 50% of PANDIT families depending on the test employed (tests T-4 and T-3, respectively; Chapter 2.6). Additionally, since the tests for non-equality of nucleotide frequencies were significant between codon positions in the study across the PANDIT database in most cases (≥ 95% of families in tests T-10 to T-14, Chapter 2.6), any test applied here that considers the ts: tv differences between site categories also considers nucleotide frequency differences. Models were tested that considered

the ts: tv bias between site categories and not nucleotide frequencies but these performed relatively poorly (data not shown).

Table 3.2 – The statistical tests that will be applied to each window of DNA data in the multiple alignment. The 'E' in the test name, standing for 'Exonic CDS', enables these tests to be distinguished from the tests used in Chapter 2.

Test	Alternate (A) and null models (B) in LRT	Degrees of freedom in LRT	Biological question addressed by the LRT
T-E1	A: HKY+R B: HKY	2	Is there a significant difference in the rate of evolution between site categories?
T-E2	A: HKY+R+N B: HKY	8	Is there a significant difference in the rate of evolution and the nucleotide frequencies between site categories?
T-E3	A: HKY+R+N+T B: HKY	10	Is there a significant difference in the rate of evolution, the nucleotide frequencies and the ts: tv bias between site categories?
T-E4	A: HKY+R+G B: HKY+G	2	Having considered rate heterogeneity across all sites in the null and alternate models, is there a significant difference in the rate of evolution between site categories?
T-E5	A: HKY+R+N+G B: HKY+G	8	Having considered rate heterogeneity across all sites in the null and alternate models, is there a significant difference in the rate of evolution and the nucleotide frequencies between site categories?
T-E6	A: HKY+R+N+T+G B: HKY+G	10	Having considered rate heterogeneity across all sites in the null and alternate models, is there a significant difference in the rate of evolution, the nucleotide frequencies and the ts: tv bias between site categories?

Note that tests T-E4 – T-E6 are the same as tests T-E1 – T-E3, respectively, except that both the alternate and null models have a single discretised γ distribution applied across all sites in models T-E4 – T-E6. Note also that the test statistics for each of these tests is taken to be twice the log likelihood difference between the alternate and null hypothesis models (as described in Chapter 1).

The models that we are choosing to apply to the data are conceptually straightforward. It is perhaps strange that such models have not been used before on multiple species alignments to try and locate CDSs. Several reasons may explain this. Firstly, until some of the research performed in this thesis was carried out (Chapter 2), some of the differences between codon positions were not characterised, in terms of their evolutionary parameters at the nucleotide level. Secondly, only very recently have we had adequate datasets on which we could test such models, because producing alignments of multiple genomes is a complex and time consuming task. There still remains a paucity of very large multiple alignments. Thirdly, current gene prediction methods perform well when we consider many of the methods together. Finally, some simple counting measures that score putative CDSs by their rate differences at different site categories do exist but remain unpublished and are not implemented in the likelihood framework with consideration of the evolutionary history between sequences (Damian Keefe, personal communication). Only very recently have HMMs incorporated phylogenies in gene-finding programs and been applied to multiple species genomic-scale alignments (McAuliffe *et al*., 2004). The application of gene-finders to multiple species genomic alignments is likely to continue and evolutionary HMMs will play an increasing role in these studies.

3.4 Significance of Test Scores

The results of any one LRT across a large dataset is valid in its own right. However, when we perform sliding window analyses across large multiple alignments, we must consider that a 5% error rate for any given window will translate to a large number of false positives when we perform tests on a large number of windows. One simple correction that we can make for a large number of tests is the Bonferroni correction (Simes, 1986; Felsenstein, 2004; Wong *et al*., 2004), which ensures that the overall false positive rate remains low. The Bonferroni correction is the most widely used correction for multiple testing but it has been shown to be conservative in some cases and consequently, we may dismiss certain results as not being significant when in fact they are significant (Simes, 1986; Felsenstein, 2004; Wong *et al*., 2004). Thus, where I use the standard Bonferroni correction (section 3.10.2) I also employ the seldom-used Simes' improved Bonferroni procedure (Simes, 1986; Wong *et al*., 2004) and report the effects that this has on our conclusions.

The standard Bonferroni correction involves dividing the type I error rate (often 5%) by the number of tests performed. The Simes' correction involves first ranking all of the P-values from the lowest to highest. If any site has a P-value smaller than its designated type I error rate divided by its *rank*, the result is significant (Simes, 1986; Wong *et al*., 2004). Other measures of statistical significance for genomewide studies have also been developed, such as the 'q' value principle (Storey and Tibshirani, 2003), which is a measure of the false discovery rate and not a false positive rate, but this is not yet widely applied.

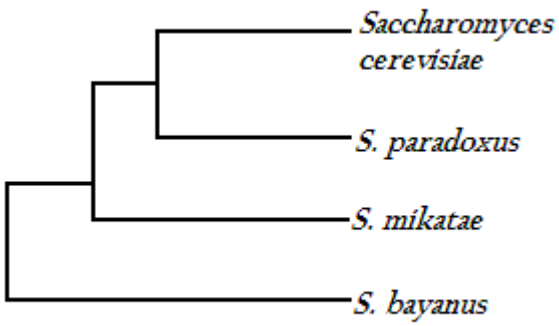
Another approach that we may pursue requires the training of our model on an appropriate dataset to determine sensible cut-off scores for the LRT statistic for each test (test T-E1 to T-E6) that we consider to have biological validity. We can then use these cut-off scores for future experiments on similar datasets. The cut-off scores will depend on factors that are specific to the multiple alignment and specific to the tests used. Alignment-specific factors will include the number of species in the multiple alignment, the evolutionary distance between them and the levels of conservation between CDSs and non-CDSs. Test-specific factors will include the size of the window and which test (T-E1 to T-E6, Table 3.2) is being performed on each window. Determining cut-off scores should not depend on the number of tests performed but this may affect the accuracy of the cut-off scores. Whether we choose to use a cut-off score or the Bonferroni correction will depend on the nature of dataset we are using and the questions we wish to ask. Once more annotated data is available, it will be increasingly possible to run trial studies on annotated datasets that are similar to unannotated datasets we might wish to study to determine sensible cut-off scores.

3.5 An Approximate Four Species Yeast Genome Alignment

In order to determine the discriminatory power of the method and suitable cut-off scores for whether or not we consider a window to be in a CDS, we need a suitable dataset. This dataset must be well-annotated and cover a large genomic region. The ENCODE datasets (see Chapter 1) do not have comprehensive coverage for most of the regions for the majority of species provided and the high levels of absent data for some species (ENCODE project consortium, 2004) is discouraging.

The most suitable dataset that I have located is a multiple alignment of four yeast species: *Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus* (Kellis *et al.*, 2003). This large dataset consists of separate multiple alignments for each CDS and inter-CDS region for all four yeast species, where available. The data were presented in the order that the genes occur in *S. cerevisiae* and the gene orders are not necessarily the same in the other yeast species (although much is preserved). This will be an issue that all multiple species genomic alignments face and thus, we can only use alignments that have the correct gene order for one species in the overall genomic multiple alignment. The 9650 separate small multiple alignments were concatenated to produce 17 whole-chromosome alignments using my own custom written Perl scripts, creating an approximate 4-species complete genome multiple alignment. Hereafter, I refer to this dataset as the yeast genome alignment (these data are also used in Chapters 4 and 5). The phylogeny of the four yeast species in the multiple alignments is taken from (Kellis *et al.*, 2003) and is presented in Figure 3.2. Further details and the separate sequence files can be found at http://www.broad.mit.edu/ftp/pub/annotation/fungi/comp_yeasts/.

Figure 3.2 – The phylogeny of the four yeast species presented in Kellis *et al.* (2003). Branch lengths have no meaning, only the branch order matters.



3.6 Testing Periodic Pattern Identification – A Well-Annotated Chromosome

The locations of all of the CDSs and non-CDS regions in the yeast genome alignment are known since CDS sequences were indicated in the separate alignments. Kellis *et al*. (2003) identified all putative ORFs using the Yeast Genome Database (Cherry *et al*., 1997) and then refined this set using reading frame conservation of the putative ORFs across the three remaining yeast species. I have used the ~1.6 Megabase multiple alignment of *S. cerevisiae* chromosome 4 to test the power of the PPI method. Chromosome 4 was chosen because it is the largest single chromosomal alignment. Non-overlapping sliding windows of different sizes were taken across this multiple alignment, such that the window length was the same as the distance between start points of consecutive windows, and tests T-E1 – T-E6 were performed on each window. Window sizes used were 100, 200, 300, 400, 500, 750 and 1000 nucleotides. With windows of larger sizes there were too few data points for each test that did not overlap a CDS boundary within the window to produce reliable averages. By ensuring windows were non-overlapping later comparisons could be made between the use of a cut-off score and the Bonferroni correction. Using non-overlapping windows means that some CDSs may be missed (since a test window may never be fully embedded in the CDS); this is unlikely to be the case for small window sizes but missing some CDSs should not be too problematic due to the large size of the dataset. Large windows may always overlap a CDS boundary, regardless of the step size, since genes are quite close together in the yeast genome alignment. The annotation of the chromosome 4 alignment allowed windows to be classified into those that were

entirely contained within a CDS, entirely contained within a non-CDS region or overlapped a CDS boundary.

Test statistics were removed where there had been a clear failure in the optimisation of parameter values for that window, for example when the tree length was too high (≥ 3 changes per site) or when the LRT statistic for any test took a negative value (see Chapter 1). Very long tree lengths (far above the dataset average) occur when BASEML cannot optimise the true tree length correctly, due either to saturation of mutations or another feature of the data in the window. When this occurs we should not expect the corresponding log likelihoods to be accurate. A negative test statistic means that the alternate hypothesis model apparently explains the evolution of the data worse than the nested null model that has fewer parameters, which can only occur when the alternate hypothesis model has failed to optimise (see Chapter 1). Additionally, regardless of the window length used, results were retained only where each of the four yeast species had more than 50 nucleotides in the test window, to ensure that a reasonable amount of data was used to calculate the test statistic.

3.7 Results of the *S. cerevisiae* Chromosome 4 Analysis

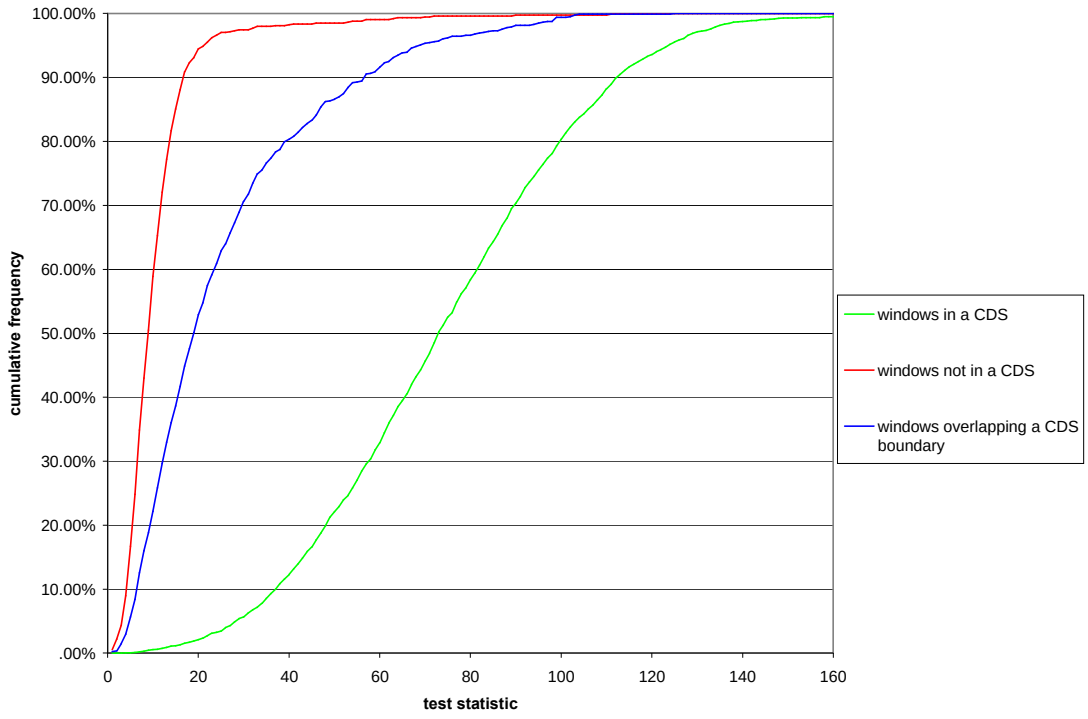
In order to demonstrate the power of the tests, Table 3.3 presents the percentage of 200-nucleotide windows contained in sequence annotated as a CDS that have test statistics above specific cut-off scores. The cut-off scores were taken as the 95% and 99% points of the distribution of test statistics where windows were in sequences annotated as non-CDS. Overall, all of the tests performed well and there were clear test statistic differences between windows that are entirely contained in a

non-CDS region and windows that are entirely contained within a CDS. Even with small window sizes there were large differences in the test statistic scores for those windows in CDSs and those in non-CDSs for each test. The cumulative frequencies of the test statistic scores for test T-E3 for 200-nucleotide windows in a CDS, not in a CDS and overlapping a CDS boundary are presented for the three distributions are shown in Figure 3.4. The results for windows of different sizes are discussed later.

Table 3.3 – The percentage of windows in a CDS that have test statistic scores above the 95% or 99% points of the distribution for windows that are entirely not in an CDS. Windows described as 'overlapping' are those overlapping an CDS boundary, containing both CDSs and non-CDSs.

Test	Using 95% Non-CDS Score		Using 99% Non-CDS Score	
	% CDSs Above Score	% Overlapping Windows	% CDSs Above Score	% Overlapping Windows
T-E1	97	51	69	9
T-E2	98	48	70	9
T-E3	98	45	71	10
T-E4	96	46	68	7
T-E5	98	41	65	7
T-E6	97	41	67	8

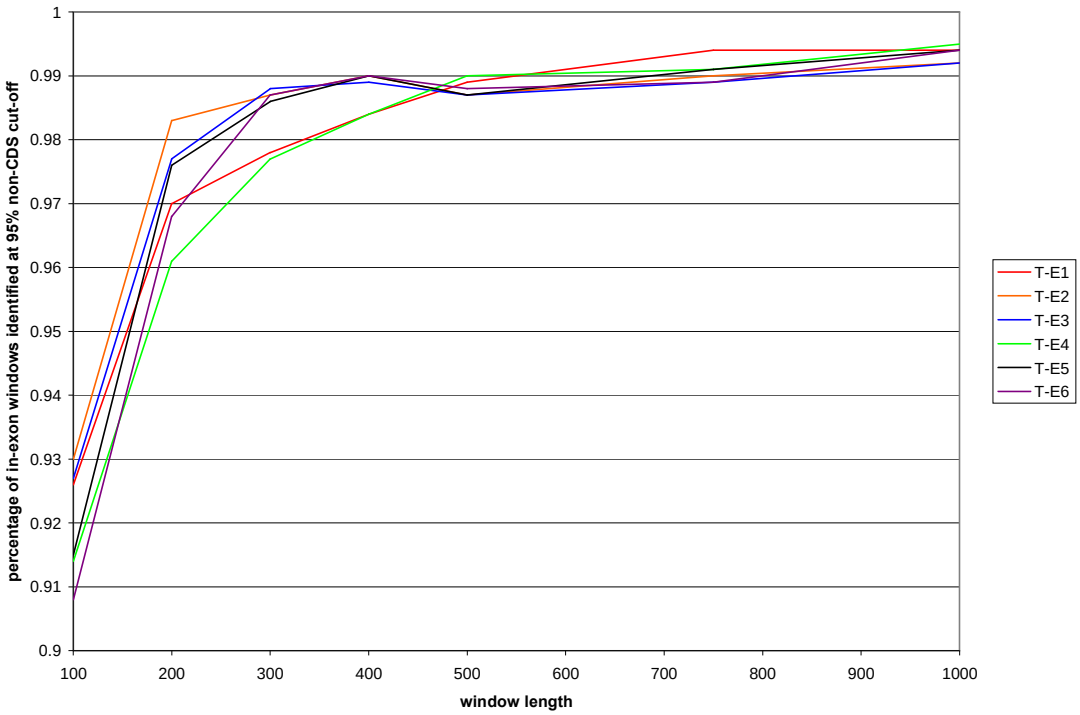
Figure 3.4 – The cumulative frequency of test statistics for test T-E3 for 200-nucleotide windows in a CDS, not in a CDS and overlapping a CDS boundary.



From Figure 3.4 it is clear that most of the windows that are not in CDSs have the lowest test scores; over 90% of such non-CDS 200-nucleotide windows have scores of 17 or less for test T-E3. In stark contrast, for the same test conditions, ~1.5% of windows contained in a CDS have test scores of 17 or less. Unsurprisingly, windows that overlap a CDS boundary have a cumulative frequency curve that is intermediate for the windows in CDSs and those not in CDSs. This is due to the fact that some of the data in windows overlapping CDS boundaries will have evolved in a CDS-like manner but the remainder of the window should not show the expected three-nucleotide periodicity in factors such as evolutionary rates, leading to an intermediate signal in the PPI method.

Models that do not incorporate γ distributions (T-E1 to T-E3) perform consistently better the equivalent models that use a γ distribution (tests T-E4 to T-E6 respectively) with a 5% and 1% false positive rate (see Table 3.3). The improvement in the power of the tests by not incorporating γ distributions (T-E1 to T-E3) is small but marked and is seen across most window lengths. This is likely caused by the interaction of our estimates of parameters governing the shape of the γ distribution and evolutionary rate. When we use models with a γ distribution, our estimates of the different rates at different site categories may not be as marked. When very long windows are used the power of all tests is virtually identical (eliminating 95% of non-CDS windows whilst retaining over 99% of windows that are contained entirely within a CDS). Additionally, tests that do not use models incorporating a γ distribution classify more windows that overlap a CDS boundary as positive results. This is a desirable property; I consider it better to classify windows that overlap a CDS as CDSs rather than as non-CDS DNA, since they would warrant further investigation. The discriminatory power with a 5% false positive rate is shown for all window sizes tested for tests T-E1 to T-E6 in Figure 3.5. Clearly, then, PPI is a powerful method for discriminating between non-CDS and CDS DNA.

Figure 3.5 – The discriminatory power of the PPI method for tests T-E1 to T-E6 for all window sizes (using the 95% point of the non-CDS distribution as a cut-off score). Note that the y-axis starts at 90% power.



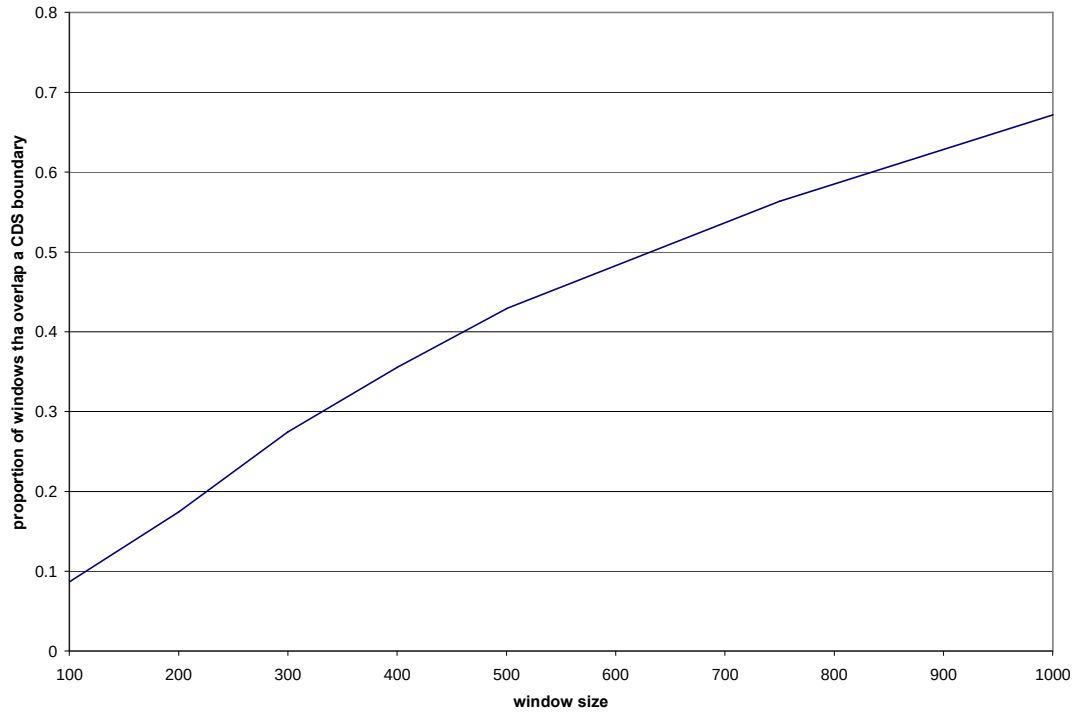
It is clear from Figure 3.5 that for any window size the discriminatory powers of all six tests are very similar. I pursue the three tests T-E1 to T-E3 on future datasets in this investigation because of their small improvement in power above the tests using models incorporating γ distributions (tests T-E4 to T-E6). There is little difference in the discriminatory power between tests T-E1 to T-E3; for completeness, I will implement all three of these tests.

For other datasets we may find that tests T-E4 to T-E6 are more powerful than tests T-E1 to T-E3. However, there are further reasons why it may be preferable to use tests T-E1 to T-E3, even if tests T-E4 to T-E6 are found to be more powerful. If we are to implement models incorporating a γ distribution in multiple alignment

datasets of only two sequences then there may be problems estimating parameters for the γ distribution and other factors at the same time since our dataset contains too little information. For example, when estimating parameters such as α at the same time as estimating the evolutionary rate, there is a problem of identifiability; any given rate could explain the evolution of the data adequately given an appropriate level of rate heterogeneity (Nick Goldman, personal communication).

As well as considering the models that will be implemented in the remainder of this investigation, it is also important to determine the window sizes that will be used. With larger window sizes there is improved discrimination between windows that are entirely contained in a CDS and windows that are not in a CDS. However, whether a window is contained in a CDS or not would not be known (or assumed) for a novel dataset; windows are more likely to overlap a CDS boundary as the window size increases. The proportions of all windows for which tests were carried out that overlap a CDS boundary are presented in Figure 3.6. The optimal window size may be genome-specific; the results discussed here are only necessarily valid in the context of the yeast data.

Figure 3.6 – Proportions of all windows that overlap a CDS boundary for a given window size (with non-overlapping increments).



Clearly then, there is a trade-off between the discriminatory power of our models to more reliably identify windows in CDSs and the probability that any single window overlaps a CDS boundary. In order to non-arbitrarily choose the best window size one can calculate the product of the proportion of windows that do not overlap CDS boundaries and the discriminatory power of the best models. This value should be roughly equivalent to the proportion of windows in CDSs that are correctly classified as being wholly in CDSs. If the performance of the top model from T-E1 to T-E3 is used or if the average performance of models T-E1 to T-E3 is used then 100-nucleotide windows are favoured. Thus, in future tests, 100 nucleotide windows will be used with tests T-E1 to T-E3. Furthermore, since small windows can reliably distinguish between CDSs and non-CDS regions, the resolution of the precise location of a CDS is improved by using smaller windows.

3.8 Testing Periodic Pattern Identification – Using the Cut-Off Scores

PPI has been shown to be a powerful discriminator of windows in CDSs and windows not in CDSs for the well-annotated yeast chromosome 4 alignment. The use of a well-annotated chromosome alignment has allowed the models to be trained, such that suitable cut-off scores have been determined to discriminate between windows that are in a CDS and those not in a CDS. Tests T-E1 to T-E3 are applied to another chromosome alignment, the *S. cerevisiae* chromosome 7 alignment, with non-overlapping 100-nucleotide windows and using the cut-off scores from the chromosome 4 alignment. The chromosome 7 multiple alignment is the largest of the separate *S. cerevisiae* chromosome alignments after chromosome 4, at just under 1.1 Mb in length. Since all of the yeast chromosome alignments are similarly annotated, one can test the suitability of the cut-off scores derived from chromosome 4 and the performance of the PPI method overall.

Out of a total of 7708 non-overlapping 100-nucleotide windows for which there was adequate sequence (> 50 nucleotides) for all four yeast species and which correctly optimised for tests T-E1 to T-E3, there were 5502, 1565 and 641 windows in CDSs, non-CDS regions and overlapping a CDS boundary, respectively. The results of using cut-off test statistic scores obtained from the chromosome 4 alignment study on chromosome 7 data are presented in Table 3.4.

Table 3.4 – The percentage of each type of window correctly classified using the cut-off scores from the chromosome 4 analysis on the chromosome 7 alignment. The numbers in brackets are the actual number of correctly classified windows of each type. A window overlapping a CDS boundary is classified correctly if it is classified as a CDS.

Test	Cut-off Score	CDSs	Non-CDS	Overlapping
T-E1	6.4	93% (5141)	96% (1507)	37% (235)
T-E2	15.9	93% (5131)	96% (1504)	33% (214)
T-E3	19.1	93% (5104)	96% (1500)	33% (209)

Each test classifies approximately 94% of windows that are entirely in an non-CDS region or in a CDS correctly. Overall, since we consider that windows overlapping a CDS boundary should be classified as CDSs, each test classifies approximately 89% of all windows correctly using the cut-off test statistic scores from the chromosome 4 analysis. Test T-E1 performs the best overall, classifying the most windows correctly. A low proportion of windows overlapping CDS boundaries are classified correctly and this is because, with only 100-nucleotide windows overall, there is often not enough signal in the small part of the window that corresponds to CDS data to produce a high enough test statistic. The tests produce very few false positives and using the cut-off scores from the chromosome 4 analysis, fewer than 5% of the true negative windows produce false positive results. In summary, of all the windows that exceed the cut-off scores for tests T-E1 to T-E3 roughly 99% are embedded within CDSs or overlap CDS boundaries.

To compare the use of cut-off scores with the Bonferroni correction, the results of the latter approach applied to the 100-nucleotide window chromosome 4 and chromosome 7 analyses are presented in Table 3.5. Using the Simes' correction (Simes, 1986) made little difference although this was slightly less conservative (results not shown). The large number of windows in these analyses has made the Bonferroni correction very conservative and a much higher percentage of CDSs are not classified as such than if one were to use the cut-off score method.

Table 3.5 – The percentage of each type of window correctly classified using the Bonferroni correction in the chromosome 4 and chromosome 7 analyses. The numbers in brackets are the actual number of correctly classified windows of each type.

S. cerevisiae chromosome	Test	CDSs	Non-CDS	Overlapping
4	T-E1	59.7% (3948)	99.7% (1953)	4.8% (39)
4	T-E2	50.8% (3354)	99.6% (1950)	3.8% (31)
4	T-E3	48.8% (3226)	99.7% (1951)	3.8% (31)
7	T-E1	62.2% (3420)	99.9% (1563)	6.9% (44)
7	T-E2	54.2% (2980)	99.9% (1563)	4.8% (31)
7	T-E3	52.0% (2863)	99.9% (1563)	4.8% (31)

Compared to Table 3.4, Table 3.5 suggests that the significant test statistic scores after Bonferroni correction are very conservative for tests T-E1 to T-E3. Much lower test statistic scores classify a much higher percentage of windows in CDSs correctly whilst still classifying a very high percentage of windows in non-CDS regions correctly (Table 3.4). The results of the Bonferroni correction method are

similar for chromosome 4 and chromosome 7 but a lower percentage of CDSs are correctly classified for chromosome 4 (Table 3.5). The Bonferroni correction is more conservative for chromosome 4 because of the greater number of tests that the significance scores are adjusted for (9380 for chromosome 4 compared to 7708 for chromosome 7).

3.9 Discussion

There are clear and strong differences in the evolutionary signals between CDSs and non-CDS regions. CDSs tend to have high PPI LRT scores and show a clear periodicity of three nucleotides in their patterns of evolution. Non-CDS regions do not show this periodicity strongly and tend to have low PPI LRT scores. The tests are powerful enough to be used on multiple alignments with a small number of species, four in the example presented, and for small windows of DNA. We can use well-annotated datasets to determine LRT cut-off scores that may later be applied to similar test datasets. In some cases the use of cut-off scores will not be possible because we require a well-annotated dataset in the first place, which may not be available. In other cases the details of the investigation may make the use of cut-off scores inappropriate, for example when our study forces us to test windows of different sizes. In such a case we can only use statistical corrections for multiple tests, which I have shown to be conservative (Table 3.5). I discuss the power and utility of the PPI method in section 3.12. Under the preferred conditions, the PPI method is a powerful tool for distinguishing between CDS and non-CDS data, demonstrated by the results of the chromosome 7 analysis.

3.10 Analysis of Unannotated Transcribed Regions in the *S. cerevisiae* Genome

Having demonstrated the power of the PPI method in a well annotated dataset, I now apply the PPI method (tests T-E1 to T-E3) in a test of whether previously unannotated regions of the *S. cerevisiae* genome that are known to be transcribed show the evolutionary signals that we consider to be indicative of CDS evolution to suggest candidate sequences for further functional investigation (David *et al.*, manuscript in preparation).

Recently, a very high resolution transcriptional analysis was carried out on *S. cerevisiae*. Using microarray techniques involving partially overlapping 25-mer oligonucleotides, the level of transcription of blocks of yeast sequence was recorded at four nucleotide resolution across the whole genome (David *et al.*, manuscript in preparation). Over 8000 distinct genome segments were identified that had levels of expression above the background level; of these, 125 were not associated with known annotated genes. I was asked by the experimentalists involved in this study to assess whether the transcribed unannotated segments had any kind of 'coding signal', which provided a good opportunity to use the PPI method on an unannotated dataset. Thus, using the PPI method, I sought to test whether any of these segments demonstrated CDS-like evolution, or whether these transcribed segments were unlikely to be CDSs.

3.10.1 Identifying Unannotated Transcribed Segments in the Yeast Genome and Testing Them for CDS-Like Evolution

The transcription dataset consisted of three subsets, representing the test dataset, the positive control dataset and the negative control dataset; these datasets consisted of genome segments that were unannotated transcribed segments, annotated and transcribed CDS segments, and untranscribed segments, respectively. Where possible, the segments were extracted from the concatenated Kellis *et al.* (2003) multiple alignment that represented the entire *S. cerevisiae* genome (see Section 3.5 above). Multiple alignments of the extracted segments were rejected if there was not an adequate amount of sequence (> 50 nucleotides) for each of the four yeast species for that segment.

The numbers of segments for each of the three categories of data that were identified in the Kellis *et al.* (2003) genome multiple alignment and had an adequate amount of sequence for each species are presented in Table 3.6. Note that in this application this is not a windowing approach: the lengths of the genomic segments of interest are pre-defined by the transcription level study. The lengths of the different segments range from 145-1880, 57-11269 and 63-3137 nucleotides in the test dataset, positive controls and negatives controls, respectively. The lengths of the segments preclude using the cut-off scores detailed earlier in this chapter because cut-off scores are determined for DNA alignment windows of a specific size. Whilst we could use a calibration curve of cut-off scores for windows of different sizes in theory, it is not possible in this case to determine the cut-off scores that are representative of very large window sizes since larger windows are more likely to overlap CDS boundaries.

Therefore, I will use LRTs comparing twice the log likelihood difference between the alternate and null hypothesis models with the appropriate χ^2 distribution and significance values will be adjusted for multiple testing using the standard Bonferroni correction and Simes' improvement to the Bonferroni correction (Wong *et al*., 2004).

Table 3.6 – Sequence segments identified in the Kellis *et al*. (2003) multiple alignment and those for which there was an adequate amount of sequence.

Dataset	Total Number of Segments	Number found in Kellis *et al.* (2003) concatenated dataset	Number with adequate sequence in all yeast species
Test set (unknown transcribed)	125	101 (81%)	57 (46%)
Positive control (transcribed CDSs)	4964	4187 (84%)	3539 (71%)
Negative control (untranscribed segments)	1255	1074 (86%)	877 (70%)

The percentage of sequences that were located in the Kellis *et al.* (2003) dataset is lower than one would have hoped. This is probably because the Kellis *et al.* (2003) dataset is incomplete in some respects; there are regions of the chromosomes that are not covered by the smaller multiple alignment files provided by Kellis *et al.*, in particular towards the ends of the *S. cerevisiae* chromosomes. Furthermore, either for biological reasons such as genomic insertions or deletions or because certain areas were sequenced in the other three yeast species more thoroughly than others, a fair number of segments fail to reach the lower limit of adequate sequence in all yeast species in the multiple alignments. The percentage of sequences with adequate

sequence in all four yeast species is lower in the test set than in the positive and negative control sequences (46%, 71% and 70% respectively). This suggests that the test set sequences are in regions that are not as comprehensively sequenced in the other yeast species as the positive or negative control sequences, or that the test set is enriched in sequences that are present only in *S. cerevisiae*.

3.10.2 Results and Conclusions

The test dataset (unannotated transcribed segments), positive control dataset (transcribed CDSs) and negative control dataset (untranscribed segments) were analysed using tests T-E1 to T-E3, and also by consideration of the inferred evolutionary rates between the codon sites (from fastest to slowest) and phylogenetic tree lengths. The evolutionary rates between the three site categories show marked differences in CDSs (as per Chapter 2) and we can compare the distributions or differences in rates between the site categories for the transcribed unannotated segments and transcribed CDSs. Furthermore, CDSs tend to be relatively conserved because of selection to retain their functions. The phylogenetic tree lengths can be used to assess whether or not the transcribed unannotated segments are as conserved as the transcribed CDSs.

The cumulative frequencies for the LRT statistic scores for tests T-E1 to T-E3 appear almost identical and thus I present the results for T-E3 only, in Figure 3.6. The tree lengths and relative rates of evolution (from highest to lowest) between the site categories were similar for most of the segments across the models HKY, HKY+R, HKY+R+N and HKY+R+N+T and the differences between the three datasets were

evident regardless of which model the relative rates or tree lengths were derived from. Figure 3.7 shows the relative rates of the three site categories (from highest to lowest) for each segment in each of the three datasets from the model HKY+R. Furthermore, Figure 3.8 shows the distribution of the tree lengths (as percentages of each dataset) from the model HKY+R.

Figure 3.6 – The cumulative frequencies (%) of the LRT statistic for test T-E3.

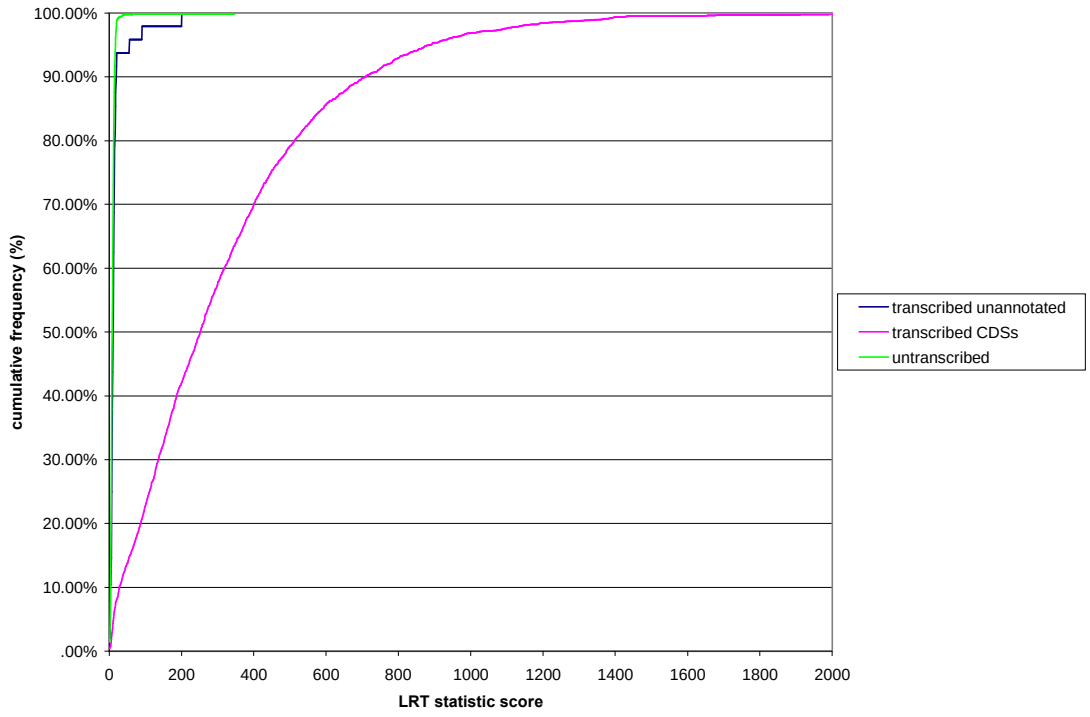


Figure 3.7 – The relative rates of the three site categories (from highest to lowest) derived from model HKY+R. For display purposes only, the data points have been normalised such that the sum of the relative rates for the three site categories for each data point = 1. No assumptions are made about the reading frame of the segments so the relative rates are ranked, which is why all data points fall in 1/6 of this plot. Non-CDSs should have roughly equal rates for the three 'putative codon positions'; hence such points fall near the middle of the triangle. Where the rates at the three 'putative codon positions' differ more, as we expect to find in CDSs, the data points fall further from the middle of the graph.

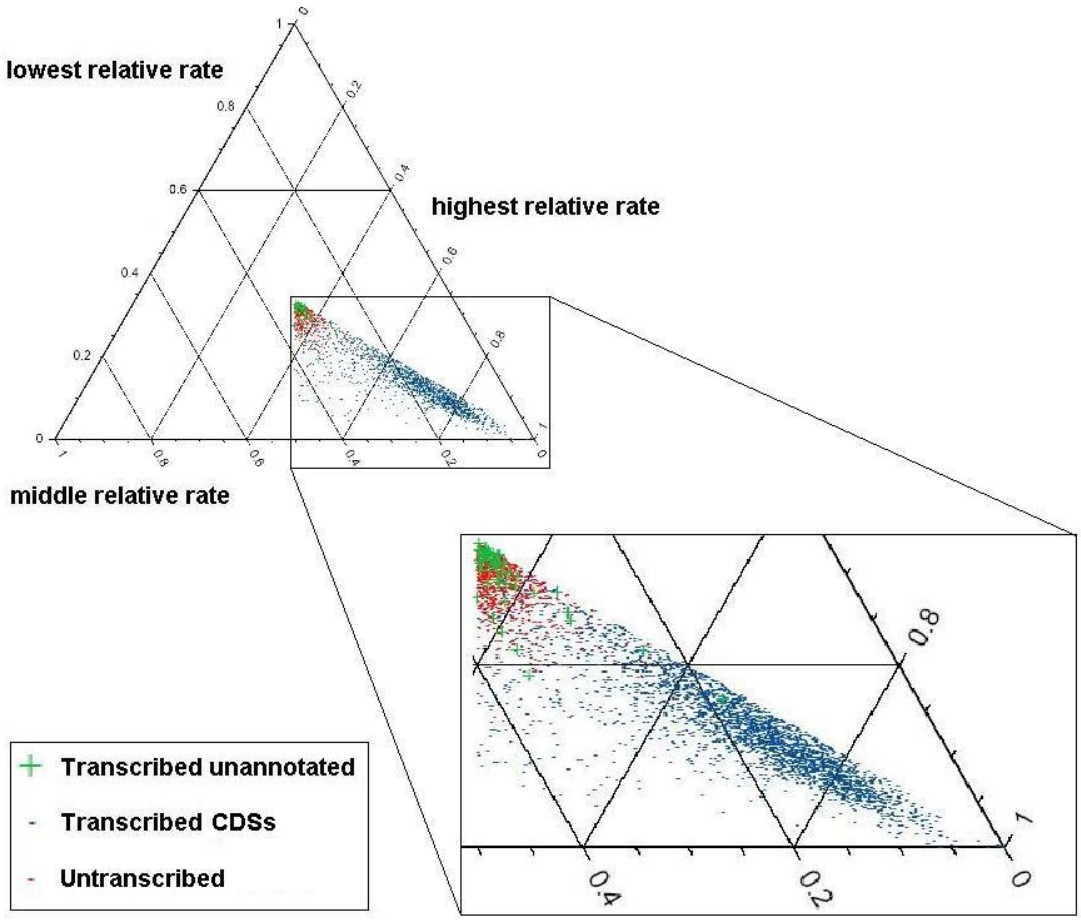


Figure 3.8 – The tree lengths of the segments, derived from model HKY+R.

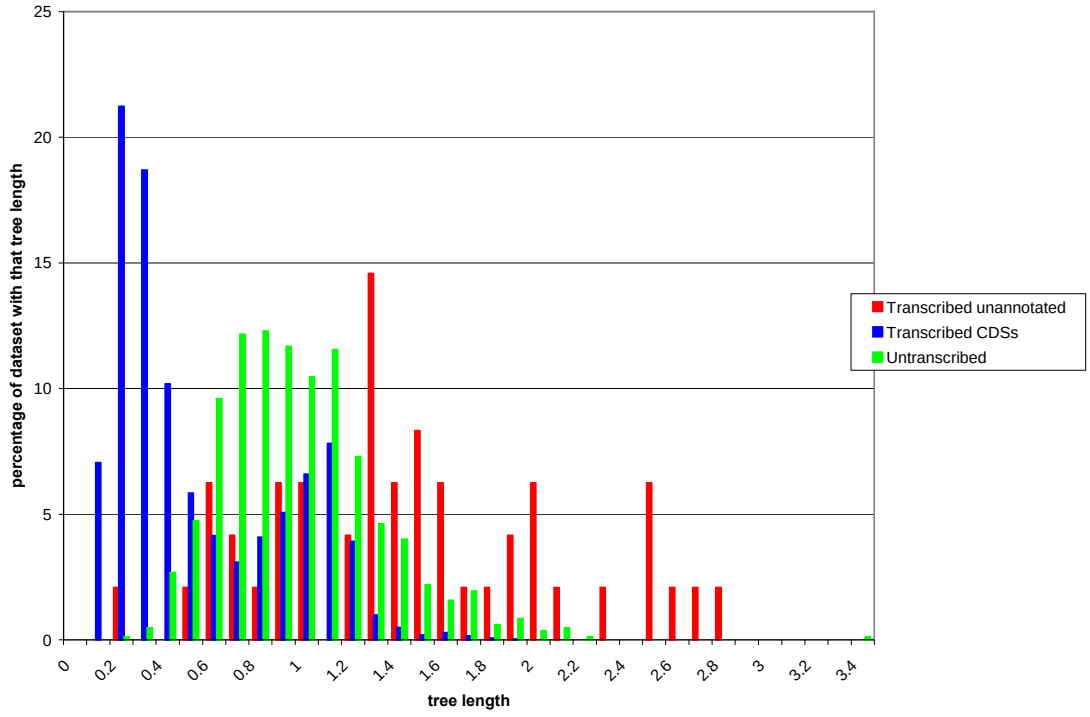


Figure 3.6 shows that the vast majority of the unannotated transcribed regions have very low PPI scores. The cumulative frequency curve of the transcribed unannotated dataset is very similar to the curve for the untranscribed segments, which we assume are not CDSs. Furthermore, Figure 3.7 shows that most of the transcribed unannotated segments do not have large rate differences between the site categories. The blue points in Figure 3.7 are more spread out and this indicates that the annotated CDSs tend to have greater differences in the relative rates of evolution at the three site categories. The distribution of relative rates at the three site categories is similar for the untranscribed and unannotated transcribed segments. Additionally, the unannotated transcribed regions are noticeably less conserved than even the untranscribed regions, evident from their long tree lengths in Figure 3.8. The transcribed CDSs are relatively conserved and tend to have short tree lengths. Thus,

from Figures 3.6, 3.7 and 3.8, we can already conclude that the bulk of the transcribed unannotated segments discovered in yeast are unlikely to be CDSs.

Whilst it seems clear that most of the transcribed segments are not CDSs, some of the unannotated transcribed segments may still have CDS-like evolutionary properties. To more closely examine whether any of the unannotated transcribed segments may be part of protein-encoding genes, the number of segments tested that produce significant LRT test scores for tests T-E1 to T-E3 are presented in Table 3.7.

Table 3.7 – The number of significant segments from each dataset for tests T-E1 to T-E3 after Bonferroni correction.

Dataset (total no. of segments)	Positive control, transcribed CDSs (3539)	Negative control, untranscribed (877)	Test set, unknown transcribed (57)
T-E1	3064 (86.6%)	2 (0.2%)	4 (7.0%)
T-E2	3049 (86.1%)	2 (0.2%)	3 (5.3%)
T-E3	3047 (86.1%)	2 (0.2%)	3 (5.3%)

The results in Table 3.7 for the positive and negative control datasets are consistent with our expectations, considering that the Bonferroni test is conservative. If the Bonferroni test were not conservative we should expect a false positive rate of roughly 5% for the negative controls. As we decrease the Type I error rate, the power (rate of acceptance of genuine positives) of the test also necessarily decreases. A power of ~86% in the positive control dataset is encouraging. This is higher than the whole yeast chromosome studies (Table 3.5, 48-62% power) because we are correcting for fewer multiple tests and test sequences are longer in this study.

The numbers of families that are significant are very similar for each test T-E1 to T-E2. I analyse the four families that have significant LRT scores for any of the tests T-E1 to T-E3 for the test dataset in more detail. The families that are chosen by tests T-E2 and T-E3 are subsets of those chosen by T-E1.

If the Simes' improved correction to the Bonferroni procedure (Simes, 1986; Wong *et al*., 2004) is used instead of the standard Bonferroni procedure the results are identical, except for a few extra true positives in the positive control set (3118 (88.1%), 3097 (87.5%) and 3094 (87.4%) for tests T-E1 to T-E3, respectively); therefore, this technique is not further discussed.

From Table 3.7 it is clear that the test dataset is slightly enriched with sequences that have high PPI LRT test scores relative to the negative control dataset. Four unannotated transcribed segments have significant PPI LRT scores after Bonferroni correction: two segments on *S. cerevisiae* chromosome 2, one segment on chromosome 4 and one segment on chromosome 12. The four segments have lower tree lengths than the median of the test set (data not shown). The individual segments were investigated using visualisations of the *S. cerevisiae* genome that includes the locations of all known genes and other features (Wolfgang Huber, personal communication). The order of the test scores from strongest to weakest was the same for all four segments regardless of the test used (T-E1 to T-E3). The segment with the weakest significant PPI LRT score, which was significant for test T-E1 but not T-E2 or T-E3, falls on the minus strand on chromosome 2 after the gene DER1, which is on the plus strand. Functional studies would be required to determine conclusively

whether a CDS was present. The next transcribed unannotated segment with a significant PPI LRT statistic is on chromosome 4 on the sense strand just before the BCS1 gene, which is on the minus strand. This segment is an excellent candidate for a short ORF. The segment with the second highest PPI LRT statistic of all unannotated transcribed segments is on the plus strand of chromosome 2, between the minus strand genes ORC2 and TRM7. This segment is also an excellent candidate for further studies. Most interesting of all is the unannotated transcribed segment with the highest PPI LRT statistic, which falls on chromosome 12. This segment is in the tail end of the pseudogene SDC25. It may be the case that this CDS is still active in the three other yeast species in the multiple alignment and is only a pseudogene in *S. cerevisiae* and hence its evolution appears very CDS-like; this could be tested with functional studies. It is interesting that the end of this pseudogene is still transcribed at a significant level in *S. cerevisiae*.

3.10.3 Discussion

The PPI method was applied with the Bonferroni statistical correction for multiple tests because it was not practical to use PPI LRT cut-off scores to ascertain whether or not a CDS-like evolutionary signal was present in each segment; candidate regions for further functional studies as CDSs were identified. The dataset of transcribed unannotated regions is slightly enriched with sequences that have high PPI LRT scores relative to the negative control dataset, but the percentage of sequences that have high PPI LRT scores in the transcribed unannotated dataset is still very low compared to the positive control dataset. The enrichment of sequences with high PPI

LRT scores in the transcribed unannotated dataset is represented by four sequences with significant test scores.

The unannotated transcribed segment with the highest PPI LRT score across tests TE-1 to T-E3 corresponds to the tail end of a pseudogene in *S. cerevisiae*. The tail end of the pseudogene may have no function but still be transcribed, or may have some residual or novel function. It is unclear what mechanisms might be responsible for the transcription of only the tail end of the pseudogene; transcription factor binding sites internal to the gene may make this possible. We might expect high PPI test scores for this segment especially if the gene did not lose its function in the three other yeast species in the multiple alignment.

Having demonstrated the power of the PPI method earlier in this chapter, it is clear from the low test scores that most unannotated transcribed segments are very unlikely to be CDSs. The relatively low level of conservation of many of these segments across the four yeast species is consistent with this notion. Thus, such segments are being transcribed for other reasons. Some segments could be genomic 'noise' transcribed by accident because small nearby sequences happen to act as transcription promoters and / or some segments may serve functional roles as RNA molecules (in particular as anti-sense transcripts that regulate the expression of other genes). This requires further elucidation by detailed functional studies of the transcribed unannotated segments. With a small number of exceptions, the results have shown that it is not worth testing many of the transcribed unannotated segments for CDS functions.

3.11 Future Directions

The PPI method is powerful enough to be of use in real-life studies (as above) in its own right and it can be performed using existing software alongside simple additional scripts. The method could be applied to assess properties of the genes predicted by other methods. Continued use of the PPI method on well-annotated datasets, such as the remainder of the *S. cerevisiae* genome alignments, could address questions regarding the differences in CDS evolution on different chromosomes. The PPI method has been developed to a stage where it could be usefully incorporated into an HMM for gene identification using additional features (such as splice site sequences, etc.). The integration of scanning for CDSs using nucleotide-level properties would require far fewer parameters than using a codon model, which may allow more reliable parameter estimation and greater power. Ultimately, all gene verifications will require functional studies, but I have shown that this method can be used to propose candidate sequences for further investigation.

3.12 Overall Discussion

The insights gained into features of codon evolution in Chapter 2 have been used to devise models to search for CDSs in multiple alignments (or to identify candidate sequences for further functional studies). The method has been shown to be a powerful discriminator of CDSs and non-CDSs in a well-annotated dataset. I considered the implications of multiple tests and have shown that dataset-derived cut-off scores or LRTs incorporating the Bonferroni correction or Simes' correction may

be of use, depending on the details of the study. Using cut-off scores, when possible, seems to be less conservative.

The method was applied to microarray-identified unannotated transcribed regions of *S. cerevisiae* to determine whether they displayed CDS-like evolutionary properties. Four regions were identified that showed significant PPI LRT scores and these regions are candidates for further functional studies. Interestingly, most of the unannotated transcribed regions did not display CDS-like evolutionary properties.

Chapter 4: Identifying Origins of Replication

Contents

4.1 Introduction

There are many features of genomes that we might wish to identify and annotate aside from genes. If different genomic features cause differences in how the genomic region encoding them evolves, then evolutionary models can be tailored to test for these different evolutionary signals. In this chapter, models are tailored to test for origins of replication initiation (ORIs) in double stranded DNA genomes, which are the locations of the start points of genome replication. I begin this chapter with a discussion of the process of DNA replication and ORIs in different species and then discuss the specific evolutionary signal that may be used to test for ORIs in a multiple species DNA sequence alignment.

I develop the complementary mutation (CM) model in the likelihood framework and apply the CM model to the Francino and Ochman (2000) dataset, an example of a known ORI location where the authors had concluded there was not a significant signal for 'in silico' detection. In contrast, the CM model does find a significant signal for the ORI in this dataset.

I then demonstrate how the CM model may be more generally applied to larger datasets, using an annotated alignment of *Saccharomyces cerevisiae* chromosome 3 as an example. A pair-sliding-windows method is developed, where a pair of windows of alignment data simultaneously progress along the chromosomal alignment separated by a constant gap size between the windows. Whilst I demonstrate how the method may be generally applied, this specific dataset yields negative results, which I discuss in terms of the high gene density in the alignment

that may confound the evolutionary signal, the small number of species in the alignment, the strength of the true evolutionary signal and whether or not there is in fact any signal to detect because the locations of the ORIs may not be homologous in the different species in the alignment. I conclude that the further use of this method will rely on the production of whole genome alignments that more closely resemble the Francino and Ochman (2000) multiple species sequence alignment.

4.2 DNA Replication

Replication of the genome is a requirement that must be fulfilled in order to replicate the organism, ensuring that each of the progeny carries genetic information inherited from the parent. For organisms that have a double stranded DNA genome, the specificity of base pairing between complementary nucleotides on opposite DNA strands (A with T, G with C) allows each strand to be used as a template for the replication of a novel DNA strand. The semi-conservative hypothesis postulates that a double stranded DNA molecule is separated into the two component strands and a new DNA strand is formed against each of the two original strands because the exposed bases of the original strands bind specifically to their complementary nucleotides (Watson and Crick, 1953). This process is complex and requires many enzymes to unwind and separate the original double stranded DNA molecule, join together separate nucleotides as they hydrogen bond with one of the original DNA template strands, and excise and replace incorrectly paired nucleotides. Points of the genome at which DNA replication is initiated are known as origins of replication initiation (ORIs). Replication of DNA usually progresses bi-directionally from an ORI. The size of the actual ORI is not well characterised, and it is not certain exactly

how specific the site of an ORI is or whether the sizes of ORIs differ in different species. The location of the vast majority of known ORIs have been determined empirically (although Breier *et al.*, 2004 have identified novel putative yeast origins based on features of genomic sequence) and the size of the region that an ORI can be localised to depends on the experimental procedure. Some procedures have localised ORIs in yeast to within a few hundred nucleotides, others to just under a couple of thousand nucleotides (Cherry *et al*., 1997).

Identifying ORIs would allow us to relate them to other evolutionary phenomena, such as chromosome break-points and hot spots for recombination or mutation. Eubacteria contain only a single ORI but eukaryotes have larger genomes and multiple chromosomes, and typically have many ORIs on each chromosome. Bacterial replication origins vary in size (from around 200-1000 nucleotides in length) and typically contain promoter sites of replication (DnaA boxes) and an AT-rich region (Mackiewicz *et al.*, 2004). Computational prediction of ORIs in some 'lower' eukaryotes, such as fission yeast (*S. pombe*), is quite simple because the ORIs are always associated with specific short DNA sequences. However, predicting which of these ORI-associated sequences is fully functional may not be so simple because initiation at one ORI sequence may inhibit the initiation of replication at other nearby ORI-associated sequences. This problem may occur in any species where specific sequences are associated with ORIs. In the brewer's yeast *S. cerevisiae* each ORI is located in a region termed an autonomously replicating sequence (ARS), which is associated with a short ARS consensus sequence; however, the ARS consensus is insufficient to form a functional ARS by itself. There are many ARS consensus sequences within the *S. cerevisiae* genome that do not function as part of an ARS

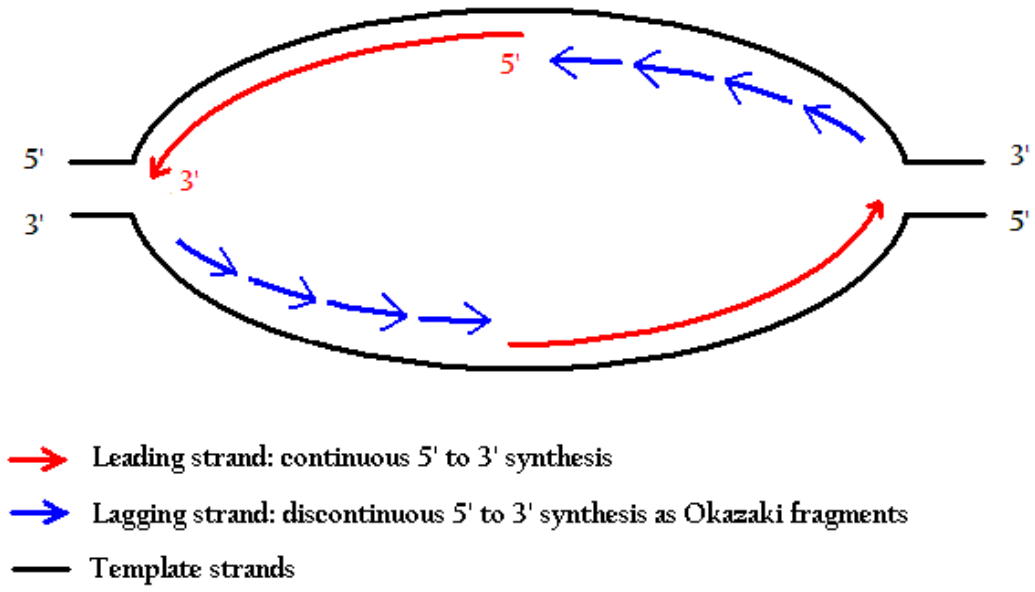
because the additional accessory sequences that are required are absent (Breier *et al*., 2004). Thus, given a genomic sequence, we can identify ARS consensus sequences but this is insufficient to predict the locations of ORIs. However, since yeast ARSs are also associated with two or three other domains, including A-rich regions and DNA unwinding elements (DUEs) they may be readily identified by properties of the sequence alongside the ARS consensus (see below).

The combination of these associated consensus sequences and other factors has led Breier *et al*. (2004) to develop ORISCAN, a program that identifies ORIs in *S. cerevisiae*. ORISCAN was trained on 26 known yeast ORIs. Applied to the remainder of the genome (i.e., excluding the training dataset), 84 of the top 100 ORISCAN ORI predictions matched known ARSs. In addition to the performance of ORISCAN, Breier *et al*. (2004) also investigated conservation of ARSs in different yeast species and noted that ORIs tended to be more conserved among different yeast species than classes of neutrally evolving 'junk' DNA. Note that the conservation of ORIs was not used in ORI detection. ORISCAN was designed for use in yeast and such a program will be of less use in Metazoa because they lack any known consensus sequence associated with ORIs. Thus, identifying ORIs in Metazoa is more challenging computationally and evolutionary studies may be of use.

The physical process of DNA replication only occurs in a single direction (5' to 3'), which obviously poses a functional challenge because an ORI expands bi-directionally (Francino and Ochman, 2000); each template DNA strand opens up in the 5' to 3' direction and in the 3' to 5' direction. If we consider the process of replication on a single nascent DNA strand then continuous DNA replication can only

occur on one side of the ORI (the 5' side of the template strand); this is the 'leading' side of the ORI. However, since the ORI opens up bi-directionally a different process occurs on the nascent strand as the template strand opens in the 3' direction; this is the 'lagging' side of the ORI. DNA replication at the lagging template strand proceeds by the production of small (~200 nucleotide) DNA fragments called Okazaki fragments, which are produced discontinuously in the 5' to 3' direction of the nascent DNA. Thus, as the replication 'bubble' opens in the template strand 3' direction, periodically an Okazaki fragment is started and the DNA replication occurs in the 5' to 3' direction until it reaches the start of the preceding Okazaki fragment. These Okazaki fragments are later joined together to produce a single new DNA molecule on the lagging strand side of the origin. This process is presented in Figure 4.1.

Figure 4.1 – An ORI progressing bi-directionally along the DNA molecule.



Leading strand and lagging strand DNA replication use different polymerases, which have different rates of error (Francino and Ochman, 2000). Furthermore, the

DNA template strand is exposed to the cellular environment as a single stranded region for long periods of time on the lagging strand relative to the leading strand (Francino and Ochman, 2000). The combination of these factors means that error rates and therefore mutation rates on the leading and lagging strands may be different and this could cause a mutational asymmetry around an ORI. Experimental systems have reported incidences of asymmetric mutation between the leading and lagging strand and it is often the case that mutations are reported as higher on the lagging strand (Francino and Ochman, 2000). Furthermore, asymmetries in base compositions either side of ORIs have been identified in many bacteria (Lobry and Sueoka, 2002).

The paired strands of DNA must have the same amount of G and C bases, due to base pairing, but for a single DNA strand there may be a much higher G than C content or *vice versa*. Observing a single DNA strand, if the leading and lagging sides of an ORI have different mutational biases then we may observe deviations from the overall average nucleotide composition either side of the ORI. There have been several studies that use nucleotide compositional biases either side of known replication origins to detect asymmetries (including Lobry and Sueoka, 2002; Mackiewicz *et al.*, 2004). These studies have used single genomic sequences to consider base composition biases. Evolutionary rate differences either side of an ORI cannot be detected using single genomic sequences.

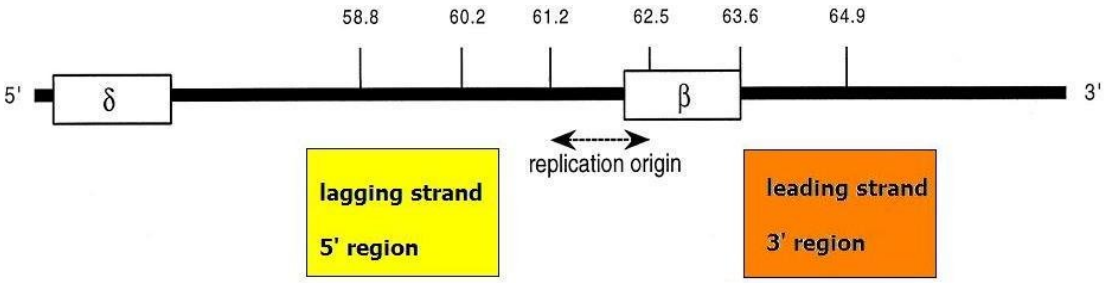
Methodologically, detecting potential mutational biases either side of an ORI is more complicated than detecting base composition biases because it requires a multiple alignment with adequate sequence either side of a known ORI across a number of species to infer the pattern of past mutations. A challenge to any

evolutionary method may be inherent to the data; many ORI locations may not be particularly conserved across species so closely related species that share the same ORI location are required in the multiple alignment (Francino and Ochman, 2000). Any evolutionary method for detecting ORIs implicitly assumes that at least some fixed mutations derive from replication errors (in the germline). Any mutational biases that ORIs cause may be hard to detect because of the effects of selection and amino acid compositions of protein-encoding sequences and their surrounding regions. Strand-specific mutational biases may also result from transcription and preferential coding strand deamination (Francino and Ochman, 2001; Touchon *et al.*, 2003). Whilst a handful of ORIs have been studied in detail in different species, the conservation of ORIs across species is not well characterised and although there is some evidence that sequences corresponding to ORI locations are more conserved than 'junk' DNA (Breier *et al*. 2004), empirical studies have not yet focused on ORI conservation across species. Investigation into the conservation of ORIs has only begun fairly recently and consequently there is a paucity of datasets that have data either side of a known ORI for a reasonable number of species. Additionally, where data is available for a reasonable number of species (such as different species of yeast), the gene density is high and it is likely that any evolutionary signal of an ORI is masked or confounded by the strong effects of selection on genes.

Francino and Ochman (2000) produced a dataset for 11 primate species that had sequence blocks either side of a known human ORI that occurs very near the β-globin gene (Figure 4.2); the location of this ORI is thought to be conserved across the species in the dataset (Francino and Ochman, 2000). Sequences were available on both sides of the ORI for only 8 species, which is the data I use. No sequence in the

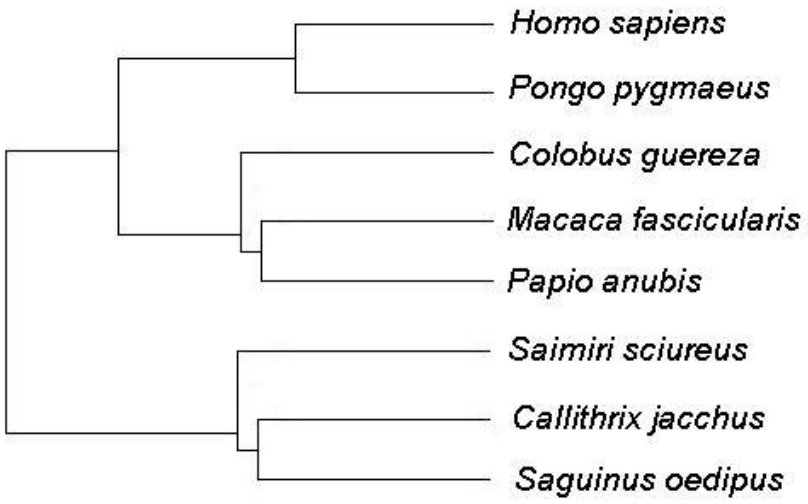
dataset overlaps any CDS and I assume that the DNA in the Francino and Ochman (2000) multiple alignment has evolved neutrally or with levels of selection that are too low to have affected biases in nucleotide substitution. Genes and other genomic regions, (such as promoters, enhancers, histone-binding elements etc) may be strongly selected, which may bias the rates and types of substitution, potentially masking any evolutionary signal that an ORI might otherwise cause in neutrally evolving DNA.

Figure 4.2 – The primate β-globin ORI region (reproduced from Francino and Ochman, 2000). Numbers are in kilobases and correspond to the human sequence with GenBank accession number U01317. Symbols δ and β correspond to the δ-globin and β-globin genes respectively. Note that the arrow marked 'replication origin' is the experimentally determined range for the actual ORI location, not the size of the ORI itself.



The Francino and Ochman (2000) dataset consisted of an approximately 5 Kb leading strand multiple alignment and an approximately 2 Kb lagging strand multiple alignment, either side of the ORI. The phylogeny of the 8 primate species for which sequence was available for both sides of the ORI is presented in Figure 4.3.

Figure 4.3 – The phylogeny presented in Francino and Ochman (2000) for the 8 primate species that had sequences in the multiple alignment either side of the ORI in the β-globin region.



Francino and Ochman used a parsimony-based approach to assess whether mutational biases existed in their dataset due to asymmetries caused by the ORI and concluded that their tests "do not support the existence of a mutational bias between the leading and lagging strands of chromosomal DNA replication in primates." (Francino and Ochman, 2000: 416). The parsimony-based approach they used employs counts of observed changes and does not attempt to reconstruct ancestral sequences. The essence of their method is to assess the ratios of the numbers of inferred complementary substitutions either side of the ORI. For example, the base pairing in a nascent DNA double helix will mean that a mutation on the leading strand from an A to G, say, will lead to a change from T to C on the opposite strand, which will happen to have been replicated as a lagging strand (see Figure 4.1). This will occur because the nucleotide A pairs with T and when the A mutates to a G, the T will be replaced by a C (such that the base pairing between the new nucleotides G-C is

maintained). If the leading strand had a particularly strong A to G mutation bias, then the lagging strand would have a T to C bias. This can be presented as follows:

$$\text{Bias value for } A \rightarrow G\ (B_{AG}) = \frac{(A \rightarrow G_{LD}) / (T \rightarrow C_{LD})}{(A \rightarrow G_{LG}) / (T \rightarrow C_{LG})}$$

where 'LD' and 'LG' stand for the processes on the leading and lagging strands, respectively. Where no bias exists the bias value = 1. This calculation was repeated by Francino and Ochman (2000) for all possible mutations (e.g., A → C, A → G, A → T, C → A, C → G, etc.), counted by parsimony, and only two pairs of complementary transversions behaved in a manner that was similar to strand bias predictions. Francino and Ochman (2000: 419) concluded that there was "no general pattern of asymmetry between the leading and lagging strands replicated from the β-globin origin." Base compositions were not significantly different either side of the ORI.

4.3 Developing a Likelihood-Based Model to Detect an ORI

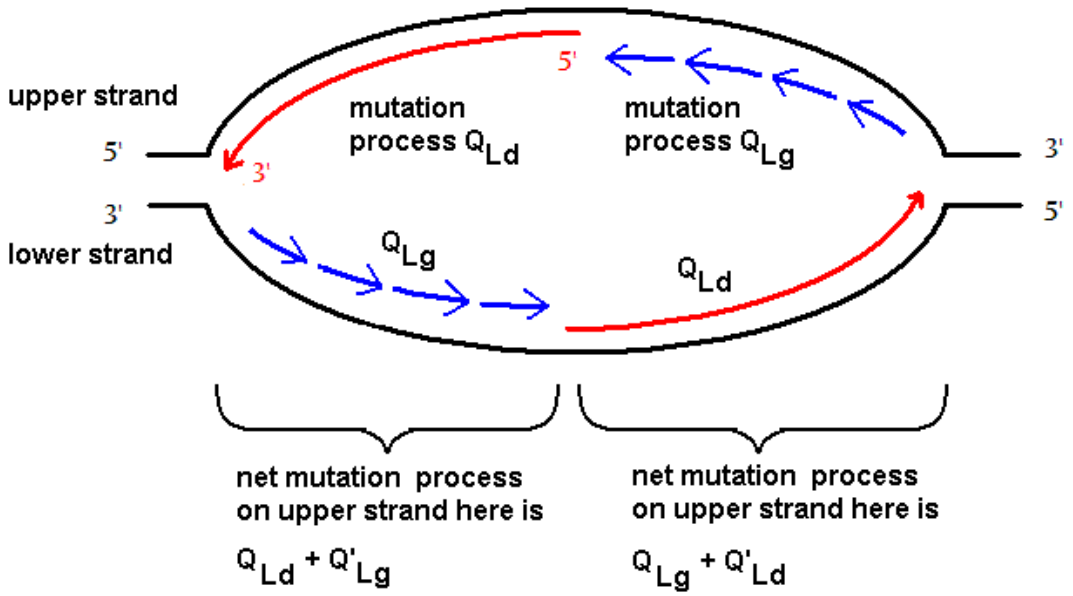
I revisit the Francino and Ochman (2000) alignments of 8 primate species, totalling approximately 13.8 Kb in length, with a novel model, the complementary mutation model (CM model), which is implemented in the likelihood framework and considers all possible nucleotide substitution biases either side of an ORI *simultaneously*. In contrast to the method of Francino and Ochman (2000), the CM model does find a significant asymmetry in mutational processes either side of the β-globin ORI.

Consider the implicit assumptions of the CM model with regard to the rates of complementary mutations either side of an ORI. The expectations of the rates of different mutations in the CM model are illustrated in Figure 4.4 and are formally described in the series of Q-matrices after Figure 4.4. Let us first consider that there may be mutational asymmetries between the leading and lagging strand. In one round of DNA replication any mutations in the lagging strand will have no effect on the leading strand sequence (and *vice versa*). However, over many generations there will be an accumulation of mutations, which are a combination of the mutations that have occurred on both strands. Thus, when we observe a single strand, as recorded in a sequence database say, one side of the ORI will have mutated with the mutational processes of the leading strand and complementary processes of the lagging strand whereas the other side of the ORI will have mutated with the mutational process of the lagging strand and the complementary process of the leading strand. If we make the reasonable further assumption that the leading strand mutation process is the same on 5' side of the upper strand on one side of an ORI as the leading strand mutation process on the 5' side of the lower strand on the other side of the ORI and likewise with lagging strand processes (see Figure 4.4), then we can assume the process on the 5' side of the ORI on the upper strand is complementary to the process on the 3' side of the ORI on the upper strand. This forms the basis to the CM model.

If our assumptions regarding mutational asymmetries are correct then if we have two windows of DNA, one either side of a known ORI, the mutational processes one side of the ORI should be complementary to the mutational processes on the other side of the ORI. In practice, it is difficult to use existing software to analyse sequence data using complementary mutational rates in different regions of our data. We can

however, consider the complement of the nucleotides on one side of the ORI and in doing so we would expect that the mutational processes in the complemented data to one side of the ORI would appear similar to the mutational processes in uncomplemented data on the other side of the ORI. This forms the basis of our null and alternate hypothesis models (see matrices below). Our null hypothesis is that the mutational processes are the same either side of the ORI. Our alternate hypothesis is that the mutational processes either side of the ORI are the same except complementary and we will complement the data to one side of the (putative) ORI in the alternate model to perform the likelihood calculations necessary to test this assertion.

Figure 4.4 – The rates of mutation on each strand, either side of an ORI. Leading, lagging and template strands are as Figure 4.1. Q_{Ld} is the mutation process on the leading strand; Q_{Lg} is the mutation process on the lagging strand. Q' is the complementary process to Q.



I now give the matrices that describe the processes on the upper strand in Figure 4.4, which show that the mutational processes are complementary either side of an ORI.

If Q =

	A	C	G	T
A	–	a	b	c
C	g	–	d	e
G	h	i	–	f
T	j	k	l	–

then Q' =

	A	C	G	T
A	–	l	k	j
C	f	–	i	h
G	e	d	–	g
T	c	b	a	–

Thus let us suppose that Q_{Ld} =

	A	C	G	T
A	–	a_1	b_1	c_1
C	g_1	–	d_1	e_1
G	h_1	i_1	–	f_1
T	j_1	k_1	l_1	–

and Q_{Lg} =

	A	C	G	T
A	–	a_2	b_2	c_2
C	g_2	–	d_2	e_2
G	h_2	i_2	–	f_2
T	j_2	k_2	l_2	–

Then $Q_{Ld} + Q'_{Lg}$ =

	A	C	G	T
A	–	$a_1 + l_2$	$b_1 + k_2$	$c_1 + j_2$
C	$g_1 + f_2$	–	$d_1 + i_2$	$e_1 + h_2$
G	$h_1 + e_2$	$i_1 + d_2$	–	$f_1 + g_2$
T	$j_1 + c_2$	$k_1 + b_2$	$l_1 + a_2$	–

and $Q_{Lg} + Q'_{Ld}$ =

	A	C	G	T
A	–	$a_2 + l_1$	$b_2 + k_1$	$c_2 + j_1$
C	$g_2 + f_1$	–	$d_2 + i_1$	$e_2 + h_1$
G	$h_2 + e_1$	$i_2 + d_1$	–	$f_2 + g_1$
T	$j_2 + c_1$	$k_2 + b_1$	$l_2 + a_1$	–

$\equiv (Q_{Ld} + Q'_{Lg})'$

The exchangeability parameters are particularly important in this investigation so a REV model is employed, as opposed to an HKY model with only one exchangeability parameter (the ts: tv bias). Our null hypothesis model, or null model, will apply a single evolutionary model, REV plus a γ distribution, to both DNA multiple alignments that are either side of the putative ORI. Our alternate hypothesis model, or alternate model, will apply the exact same evolutionary model to the two

multiple alignments except all of the bases that are on one side of the ORI will be complemented, such that nucleotides A, C, G and T are substituted in this window with T, G, C and A, respectively. Note that whether or not the 5' or 3' window is complemented makes no difference to the test statistic.

Unlike the models used previously in this thesis, the null model is not nested in the alternate model; the null and alternate models have the same number of parameters to be estimated. This means that we cannot use a χ^2 approximation to the LRT to assess the significance of the CM test statistic, which is the difference in model fit between the alternate and null models. Indeed, the CM model test statistic may be positive or negative and the alternate model may provide a worse explanation of the evolution of our data than the null model. Non-nested models can still be compared using an LRT, providing an adequate distribution of test statistics under the null model can be attained (Cox, 1961, 1962). Thus, in order to assess the significance of our test statistic, a parametric bootstrapping technique is used (Chapter 1, section 1.4.2).

In the parametric bootstrapping technique, many datasets are simulated with the null model evolutionary parameters; in this investigation 1000 pseudo-replicate datasets are simulated. Gaps that are present in the original dataset will be introduced at the same locations in the pseudo-replicate datasets. The null and alternate models will be applied to all pseudo-replicate datasets and each such test statistic will be recorded. The observed CM test statistic will be compared to the distribution of the pseudo-replicate CM model test statistics. If the observed CM test statistic is significantly more positive than the 95^{th} percentile mark of the pseudo-replicate CM

test statistic distribution, then we reject the null hypothesis and we have identified a significant evolutionary signal based on mutational exchangeabilities from the ORI. Note that the CM test statistic is: lnL (H_a) – lnL (H_0); since we are not using a χ^2 approximation to the LRT we do not need to consider twice this value. The models were implemented using programs in the PAML package (Yang, 1997) and using custom written Perl scripts.

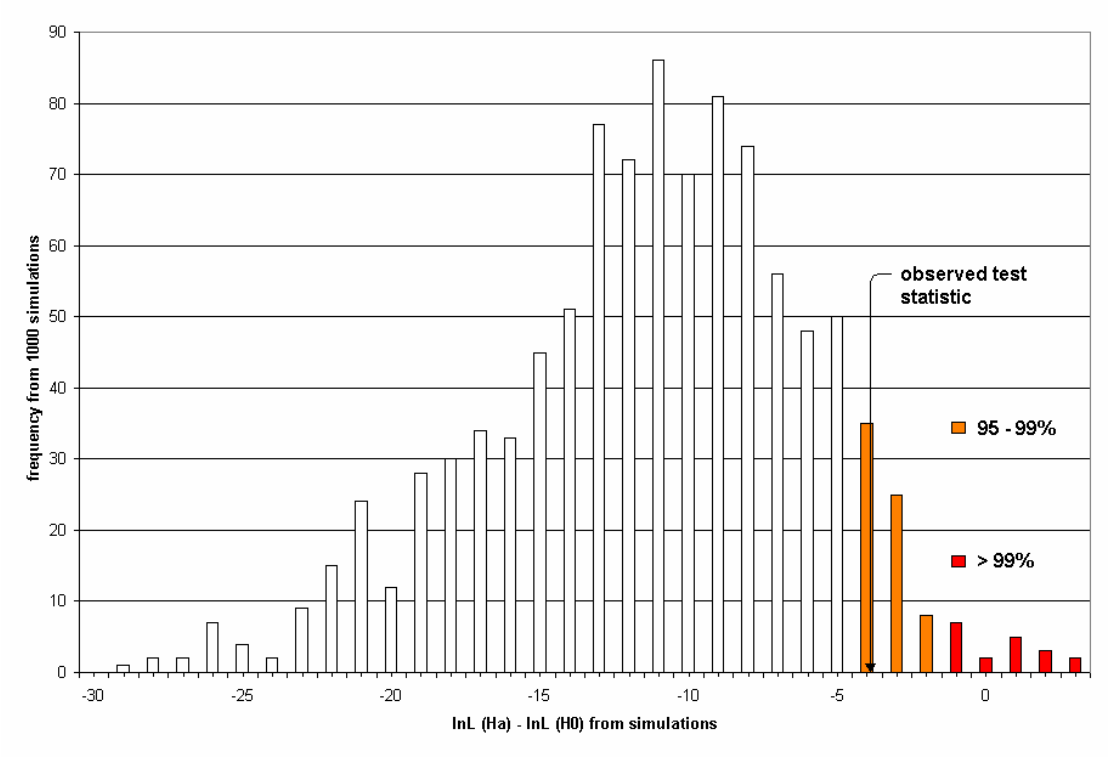
4.4 Results and Conclusions of the CM model

The results of the observed CM test statistic and the CM test statistic distribution for the 1000 pseudo-replicate datasets are presented in Table 4.1. The distribution of pseudo-replicate test statistics and the observed test statistic are shown in Figure 4.5.

Table 4.1 – Test statistic values for the real and pseudo-replicate data.

	Pseudo-replicate 50% distribution	Pseudo-replicate 95% distribution	Pseudo-replicate 99% distribution	Test statistic for real data
Test statistic	-11.39	-3.96	-0.36	-3.49

Figure 4.5 – The distribution of 1000 bootstrap pseudo-replicate test statistic results and the observed test statistic from the study of the β-globin ORI region.



The observed CM test statistic is greater than the 95% mark of the pseudo-replicate distribution of test statistics, which means the test statistic is significant at this level. The observed CM test statistic is not significant at the more stringent 99% level and the test statistic is negative. Thus, although the result is significant, the evolutionary signal detected by the CM method is fairly weak, even given the number of species and size of the alignment windows.

For windows of DNA that are not either side of an ORI it is unlikely that the rates of each mutation in one window would be similar to the rates of the complementary mutations in the other window, which is the assertion of the alternate model. It is interesting that the for the observed test statistic the alternate model still provides a lower likelihood for the data than our null model (the test statistic is

negative). This is perhaps not surprising if the effects of the ORI on the relative rates of each type of mutation are weak or if the evolutionary signal of complementarity of mutation rates has been degraded by other processes that may not be strand-specific (such as selection on genes, discussed later). However, if the ORI were not present we would expect the alternate model to perform significantly worse than it already performs, according to the distribution of pseudo-replicate test statistics; hence the significance of the CM model test statistic.

From one perspective the LRT is conservative because it takes the state of non-complementarity as the null model and tries to reject this model at its 95% point; the non-complementarity model is favoured. One could take the alternative approach and choose complementarity as the null model. Only if we rejected the complementarity null could we say that a dataset definitely lacks the expected evolutionary signal of an ORI.

4.5 Discussion

It is clear that an ORI may leave a significant evolutionary signal and now that this signal has been demonstrated, one would hope to use such a signal to identify putative ORIs at the sequence level. Whilst the evolutionary signal that affects complementary mutations is significant and the ORI is detected by this method, the signal is weak even though the multiple alignment windows are fairly long, as both the 3' and 5' windows are 2 and 5 Kb respectively, and the multiple alignments are composed of sequences from eight species. The phylogenetic distribution of species is fairly small, which is confirmed by the overall tree lengths of each window, which are

each around 0.8 (changes per site across the whole tree). However, using wider phylogenetic distributions of species may be inappropriate because it is uncertain how conserved ORIs are across species with a wider phylogenetic distribution. I have been unable to find other datasets that have a fairly high number of closely related species upon which further studies could be carried out.

4.6 Sliding Window Application of the CM Model

Experimental determination of the location of ORIs is difficult and expensive, so a method that could scan large alignments would be valuable. To test the performance of the method however, a dataset with well-annotated ORI positions is best. The ENCODE datasets lack this annotation. Thus, to demonstrate how the method could be generally applied, the four species multiple alignment of *S. cerevisiae* chromosome 3 that uses data from Kellis *et al.* (2003) is used; other chromosomes in this dataset were used in Chapter 3. The locations of all of the known autonomously replicating sequences (ARSs), which contain ORIs, were obtained for chromosome 3 using information from the Yeast Genome Database (YGD) (Cherry *et al.*, 1997). There are a total of 19 listed ARSs on *S. cerevisiae* chromosome 3. The sequences for all of the ARSs were extracted from a reference *S. cerevisiae* chromosome 3 sequence and, where possible, these sequences were identified in the multiple alignment of chromosome 3. Out of the 19 ARSs, 12 were fully identified in the chromosome 3 multiple alignment; the largest ARS was just over 1800 nucleotides. The remaining 7 ARSs were in sequence gaps in the multiple alignment (particularly those ARSs near the ends of the chromosome) and therefore we should not expect them to affect our ability to find the other 12. Considering the extensive *a*

priori knowledge of the dataset, this investigation chiefly concerns the potential utility of a sliding-window method and is not a final trial of the fully developed method.

It is noteworthy that the Kellis *et al*. (2003) data contains sequences for only four yeast species and the average tree length for any window exceeds one substitution per site (data not shown). The Francino and Ochman (2000) dataset has a smaller tree length (~0.8 substitutions per site, data not shown) and more species (8 instead of 4), making the evolutionary information more rich in the Francino and Ochman (2000) dataset. It is not clear which tree length is 'best' since we want enough evolutionary distance between the species for many substitutions to have occurred (generating a strong evolutionary signal) but not so great a distance that there is saturation of substitutions or changes to the location of ORIs across species. Additionally, the gene density of yeast is very high; the sliding windows application of the CM model would be better suited to a genomic alignment of a large number of primate species, which have a lower gene density. Furthermore, programs such as ORISCAN (Breier *et al*., 2004) are available for identifying ORIs in non-metazoan species. Nevertheless, The Kellis *et al*. (2003) dataset is the best dataset currently available for the purposes of this investigation and it is important to test the method, even if one does not expect excellent results, to demonstrate how the method may be applied to other datasets as they become available.

Within each *S. cerevisiae* ARS there is at least one ARS consensus sequence ([A/T]TTTA[C/T][A/G]TTT[A/T]), an 11 nucleotide element required in all ARSs. This is the minimum ARS length and there is an elongated consensus that is less well conserved of up to 17 nucleotides (Breier *et al*., 2004). Some ARSs have more than

one ARS consensus sequence, however, and many ARS consensus sequences within the genome are not associated with an ORI. Accessory sequences are required to produce a functional ARS and where the DNA actually opens up within the ARS and replication begins is not completely understood. Therefore, when using a sliding window analysis across the chromosome 3 alignment, the central gap between the two windows must be at least as wide as the largest ARS so we can be sure that one window is on the leading side of the ORI and the other window is on the lagging side of the ORI. Thus the gap between the two test windows was chosen to be 2000 nucleotides.

A sliding window analysis of pairs of windows each of size 100, 200 or 500, with a central gap of 2000 nucleotides and a 500 nucleotide step size, was performed on the yeast chromosome 3 alignment. Whilst these window sizes are very small compared to the two windows in the Francino and Ochman (2000) dataset, which gave a significant result at the 95% but not the 99% level, the gene density in the yeast alignment is so high that using longer windows would overlap genes more often and this could confound the evolutionary signal even further. Unfortunately, even with windows of this size, many of the tests will have at least one of the windows partially overlapping exonic sequence. The gene density in the β-globin ORI region is far lower than in the Kellis *et al.* (2003) multiple alignments and the Francino and Ochman (2000) dataset did not suffer the same problems of confounding evolutionary signals that we might expect to find in this yeast dataset (i.e., the two test windows did not overlap any coding sequence).

The tests were performed using custom written Perl scripts to manipulate sequences, which were analysed using BASEML in the PAML package (Yang, 1997). The evolutionary model REV plus a γ distribution was used, as in my previous study on the Francino and Ochman 8-species dataset (2000). Issues of performing multiple tests on the dataset are discussed in the next section.

4.6.1 Results of the Sliding Window Application of the CM Model

Test windows were rejected where there was inadequate sequence (<50 nucleotides) available for all four species in either of the sliding windows. Due to this restriction, there were adequate test statistic results for only 5 of the 12 ARSs identified in the chromosome 3 multiple alignment (as well as many results where windows were not either side of an ORI).

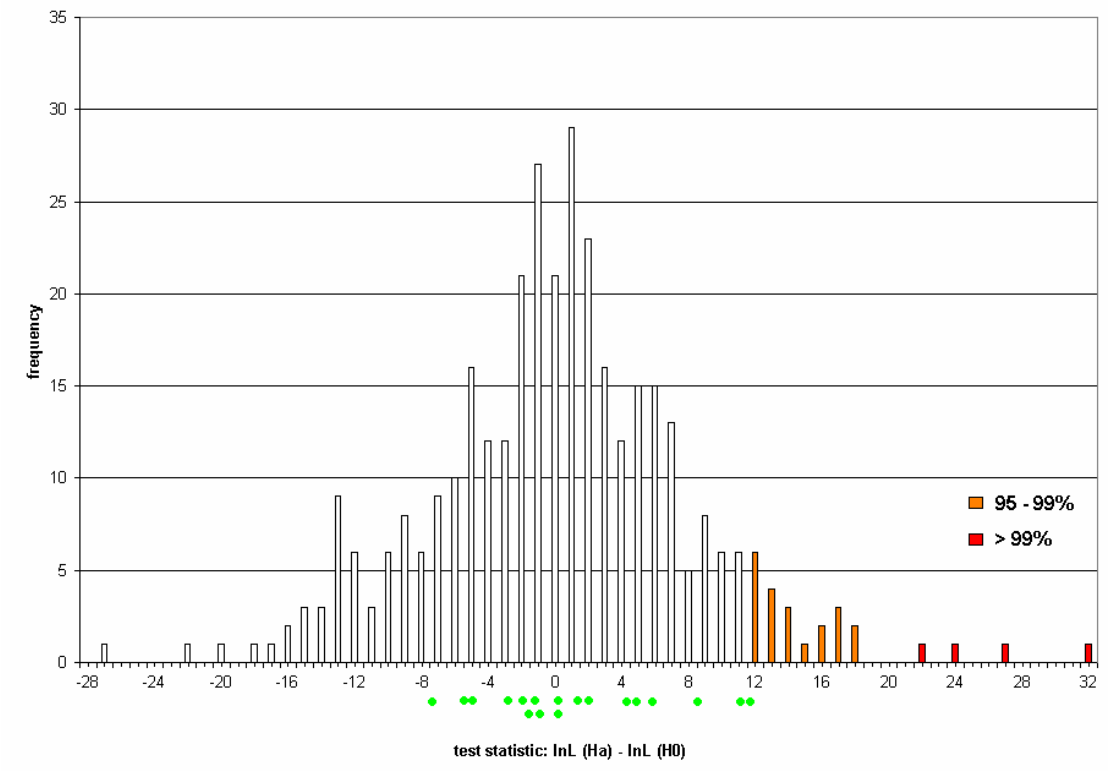
I now discuss the issues of multiple testing that relate to this study and some basic approaches that I employ to ascertain whether the identified ARSs give an evolutionary signal that is deemed significant and indicative of an ORI. The competing hypotheses use models that are not nested and thus, there is no simple statistical distribution that can be used to compare the fit of the two models. Thus, there is no known *a priori* distribution that can be used to approximate the null and we cannot use techniques such as the Bonferroni correction to address the issue of multiple tests. Furthermore, because there are pairs of sliding windows, the position of the second window may overlap the position of a (future) first window once the pair of windows has progressed along the alignment enough. Thus, some of the data

may be used in more than one test, as part of the first window of one test and part of the second window of another test. For smaller sized datasets this could be overcome by bootstrapping, because the pseudo-replicate datasets would use data in a similar manner. However, unlike in the study of the Francino and Ochman (2000) dataset, parametric bootstrapping is not a viable option because of the size of the dataset. When we produce a pseudo-replicate the parameters need to be representative of the dataset as a whole. Producing a pseudo-replicate of a chromosomal alignment based on a single set of evolutionary parameters does not encompass the variation in the multiple alignment in different parts of the chromosome. One could produce a series of parametric bootstrap distributions for sub-regions of the larger multiple alignment. Furthermore, one could use a well-annotated dataset to produce cut-off scores, in a similar fashion to Chapter 3. Unfortunately, the difficulties in detecting any signal for an ORI with this yeast dataset precludes doing so, as discussed below.

There are far fewer windows that contain real ORIs than do not across the whole chromosome and the windows containing origins of replication should have little effect on the overall distribution of test statistic scores. Thus, for each window size we can view the entire set of test statistic scores as a null distribution. We can then compare, *en masse*, all of the test statistics for pairs of windows either side of an ARS to this proxy for a null distribution. Whilst we may not easily control for multiple tests, we can at least observe whether any genuine positive (i.e., where the pair of windows are either side of a known ARS) produces a test statistic score that is greater than a fairly stringent mark in the null distribution. I have chosen the 99% point of the null distribution for this end and have found that, for all window sizes (100, 200 and 500 nucleotides), no CM test statistic produced by windows that are

either side of a known ARS exceeds this mark. Thus, using this scheme, no ARS produces a significant result regardless of the window size used. A histogram of the test statistic values over all windows is presented in Figure 4.6, marked with the scores for tests where the pair of windows were either side of a known ARS location.

Figure 4.6 – The test statistic distribution obtained from all pairs of windows (of size 500 nucleotides). Each of green dots below the x-axis represents a test score where the pair of windows was either side of a known ORI. No such test scores are significant.



The negative results may be due to biological factors or the way in which the method has been applied to this dataset. Either way, the lack of success of the CM method for this dataset means that one cannot use the cut-off score method applied in Chapter 3. It is possible that we are unable to detect significant evolutionary signals

for genuine ORIs because even the largest window size used is simply too small to detect a weak signal. However, larger windows are more likely to overlap a CDS, confounding the evolutionary signal. The gene density is far higher in the yeast alignment than in primates in general and it may be inappropriate to apply a sliding window method such as this to non-intergenic regions. I repeated the entire sliding window study retaining only the CM test statistics where both windows, either side of a central gap, fell entirely in intergenic regions but this provided too few CM test statistics over the entire chromosome to be worthy of further pursuit (data not shown). The four species yeast multiple alignment dataset presents many challenges to the application of the CM method.

The weak evolutionary signal detected in the 8-species Francino and Ochman (2000) dataset suggests that the four species yeast alignment may not contain enough evolutionary information to detect ORIs conserved across the species, regardless of the window size. If the evolutionary signal is weak then we may simply need more species, which are each more closely related. Although the ORIs seem to be more conserved across different yeast than neutrally evolving DNA (Breier *et al*., 2004), it is not known if these ORIs are used at the same frequency in different species; different ORIs may be recruited with different frequencies in different species. Thus, whilst ORI sequences may be fairly conserved across yeast species there may still be little or no evolutionary signal to detect at any one ORI. Additionally, this method may detect a signal where the blocks of transcribed sequences are close to each other but on different strands because of the opposite direction of transcription-associated deamination (Green *et al*., 2003). In theory, the CM method could be purposely used

to identify where transcription blocks switched from the positive to the negative strand.

Our inability to detect a signal for locations of known ARSs in *S. cerevisiae* may be due to interference by the evolutionary forces in genes (and other non-strand specific forces), which may mask an ORI-like evolutionary signal, having too few species in the multiple alignment to adequately detect a signal or, more simply, the location of the most frequently recruited ORIs may not be conserved across the yeast species used in this study.

4.7 Future Directions

The CM model has been applied successfully to the Francino and Ochman (2000) dataset. Whilst this constitutes proof of method for a single well-annotated dataset where we consider the ORI location likely to be conserved across species and shows that ORIs may leave a significant evolutionary signal, it does not show how the method might be used on larger datasets. To this end, the sliding windows method was developed. Whilst I have demonstrated how a pair of sliding windows can be applied to a genomic scale multiple alignment I have been unable to produce positive results in this manner (and consequently determine cut-off scores that could be used on novel datasets). The very high gene density of yeast that may confound the evolutionary signal, small number of species in the multiple alignment and unknown conservation of ORIs across yeast species makes this dataset less than ideal. As the number of sequenced genomes in Metazoa increases, particularly across primates, more multiple alignments will be produced that have a high number of closely related

species that are more likely to have ORI locations conserved across the species and the evolutionary signal is less likely to be confounded by genes since gene density is much lower. The further application of the CM method is likely to require better data although this is because of the weak nature of the evolutionary signal itself and not through a failing of the method. The CM model is best applied to species with low gene densities, which are coincidentally the same species that have the least number of sequence-level features associated with ORIs (i.e., Metazoa). Future use of the CM model may address to what extent ORI locations are conserved across metazoan species.

4.8 Discussion

I have introduced DNA replication and discussed the architecture of ORIs and a predicted evolutionary signal caused by differences in mutation on the leading strand and lagging strand. The CM model was developed in the likelihood framework and applied successfully to the Francino and Ochman (2000) dataset, identifying a fairly weak but significant signal for a known ORI, presumed to be conserved across 8 primate species. I have shown how the CM model might be more generally applied as a tool to detect novel ORIs but have been unable to show any power to detect known ORIs (where the location is only known for certain in one of the species, *S. cerevisiae*) when applying the method in the manner of sliding windows. Finally, I have discussed how the further development of the CM model will require additional datasets. Whilst this is frustrating at present, it is likely that such datasets will become available within the next few years.

Chapter 5: Measuring Changes in Evolutionary Dynamics Across Large Regions

Contents

5.1 Introduction

I begin this chapter with a discussion of the previous research into the spatial scales over which mutation rates change and propose that it is valuable to investigate not only evolutionary rate differences over different spatial scales but also how local among-site rate variation itself varies over different genomic scales, as well as spatial variation in the transition: transversion bias and nucleotide frequencies. I discuss how one might use a pair-sliding window analysis, in a similar manner to Chapter 4, with novel applications. Here I develop a method to use many window analyses to observe general trends in parameter changes across a region. The average results and distributions of results of many different pair-window analyses may be of interest in themselves, particularly if analyses are run with different sizes of the central gap between the two windows. I investigate these phenomena in both yeasts and mammals and find that regions that are physically closer to each other tend to evolve more similarly than regions that are more distant and different parameters may vary over different scales. I use large multiple alignments to reliably investigate these phenomena. Results are discussed in detail, considering differences and similarities in the evolutionary properties between two mammalian datasets and between two yeast datasets and also between yeasts and mammals. Finally, I consider the biological reasons that may explain the results observed and suggest further investigations that could be pursued.

5.2 Background

There has been considerable research into the spatial scales over which mutation rates change (Lercher *et al*., 2001 and 2004; Webster *et al*., 2003; Keightley *et al*., 2005a), which has been of particular relevance to the study of conserved non-genic regions (CNGs). CNGs are genomic regions of various sizes that are highly conserved across species but are not thought to encode genes (Dermitzakis *et al*., 2003; Keightley *et al*., 2005b). It is likely that many CNGs have important functions and selection has conserved their sequences across species; they may serve as enhancers of genes, for example, or have roles in chromatin remodelling. Knowing the scales over which mutation rates vary may affect our interpretation of CNG evolution because some CNGs may be conserved due to low regional mutational rates (although this is thought not to be the case). Furthermore, our ability to detect CNGs depends on how different their mutation rates are to those of surrounding sequences. Thus, large regions with low mutation rates present a challenge to detecting additionally-conserved regions within them.

Previous studies of spatial variation in mutation rates have usually focused on fourfold degenerate synonymous sites in codon sequences to measure substitution rates. Keightley *et al.* (2005a) randomly chose 1000 genes aligned in human and chimpanzee and compared the synonymous substitution rates to the evolutionary rates of linked, upstream sequences (within 6Kb). The evolutionary rates of the synonymous sites in the genes were found to be more similar to those of the linked upstream regions than the upstream regions of other genes. Keightley *et al*. (2005a) say nothing about trends in mutation rates across large regions. Lercher *et al*. (2001)

take a mean evolutionary rate of genes in human-mouse and mouse-rat comparisons and compare the rates of evolution of given genes with other genes within certain genetic linkage distances (not absolute distances). Closely linked genes have more similar nonsynonymous and synonymous substitution rates than expected by chance. Lercher *et al*. (2001) also showed that genes on any given chromosome have significantly more similar rates of nonsynonymous and synonymous site evolution than genes on other chromosomes in both the human-mouse and mouse-rat comparisons, and found that this did not relate to the GC content of the region. In contrast, Webster *et al*. (2003) found little or no regional variation in the mutation rate of chimpanzee and human in 1.8 Mb of genomic alignments of human, chimp and baboon. Thus, apparent regional differences in mutation rates depends on the species being investigated (as well as the genomic region studied); more divergent species tend to show more evidence of regional variation, which may be because there is more change between homologous positions (akin to having more data).

The majority of studies have focused on the mutation rates of genes and genetic distances, instead of physical distances, between regions. Such restrictions are no longer necessary in the genomic era and we can observe regional variation in evolutionary phenomena other than evolutionary rates.

In Chapter 3 I introduced the use of single sliding windows to test for an exon-like evolutionary signal across yeast chromosome alignments. I introduced the use of pairs of sliding windows to try and identify an evolutionary signal for ORIs in Chapter 4. Pairs of sliding windows can be applied more generally in order to test for a suite of evolutionary differences between the data in the two windows. The key

issue of modelling is that we choose appropriate models to represent our null and alternate hypotheses; the difference between the models in fit to observed data allows us to conclude whether there are significant and specific differences in the patterns of evolution between the pair of windows. As our pair of windows proceeds incrementally along a large multiple alignment, such as a chromosomal alignment, a large number of test statistics are generated.

There are two main ways in which the test statistics could be used. Firstly, the specific 'trace' of test statistic values as the two windows progress along a multiple alignment can inform us about interesting features at specific locations in the multiple alignment. The CM model, presented in Chapter 4, is one example of where the trace of test statistics across a large multiple alignment could be used to identify a specific genomic feature. These traces might be applied most usefully to search for boundaries of genomic features, such as genes or perhaps isochores, which differ from each other by way of nucleotide composition in certain metazoan species (Eyre-Walker and Hurst, 2001; Montoya-Burgos *et al*., 2003; Belle *et al*., 2004). When using a pair of windows to search for specific features that evolve differently from their neighbouring genomic regions one must consider how large each window should be and the size of the gap between them.

Generating a very large number of test statistics for windows of a given size, with a given gap size between the windows, allows us to use pairs of windows in a second way. Comparing test statistics for a large number of windows and, in particular, varying the size of the central gap between the windows, allows us to ask questions regarding the spatial scales over which our estimates of different

evolutionary parameters vary. Using pairs of windows in this way is the focus of this chapter. This allows us to ask interesting biological questions that have been difficult to address previously comparing the differences in spatial scales over which our estimates of evolutionary parameters become significantly different for different multiple alignments (Lercher *et al*., 2001 and 2004; Webster *et al*., 2003; Keightley *et al*., 2005a). For example, it is not known if rate heterogeneity changes over a shorter spatial scale than nucleotide frequencies. It is also not known whether spatial variation in the different evolutionary processes is qualitatively the same in different multiple alignments. This chapter has important implications for phylogenomic studies, which may combine data or analyse data separately and then combine results to estimate phylogenetic trees and estimate evolutionary parameters. If one combines the data we need to know how best to perform an analysis. Indeed, there has been research regarding how best to analyse multiple datasets (e.g., Yang, 1996; Pupko *et al*., 2002). I now describe the models and LRTs used in this investigation.

5.3 Models

In Chapter 2 I investigated the differences in the evolutionary rate, rate heterogeneity, transition: transversion bias and nucleotide frequencies between different codon positions, observing variation in evolutionary dynamics only within short sequences and not considering systematic variation from one alignment to the next. I investigate these same factors in this chapter in terms of the spatial scales over which our estimates vary significantly for the parameters that correspond to these evolutionary phenomena.

In this thesis, I have documented cases where model complexity may affect the log likelihood differences between alternate hypothesis and null hypothesis models (see Chapter 2 for example). Estimates of evolutionary rate differences between two segments of a multiple alignment may be affected by whether or not both the null and alternate models also estimate the amount of evolutionary rate heterogeneity in each of the two segments. Thus, we might find that the complexity of null and alternate models used can affect our estimates of the spatial scales over which the parameter estimates, which we are specifically interested in, vary significantly. The simplest solution, therefore, is to use a comprehensive suite of models with different levels of complexity to thoroughly investigate spatial parameter variation. The models used in this investigation are presented in Table 5.1.

Table 5.1 – The models used in this investigation.

Model Name	Free Parameters	Model description and parameters that are allowed to differ between windows
HKY	4 + tree	HKY model (Chapter 1, Section 1.2.3.1).
HKY+G	5 + tree	HKY model plus a single discretised γ distribution.
HKY+R	5 + tree	HKY and estimates an independent evolutionary rate for each window (with the same proportions between branch lengths of the tree).
HKY+R+T	6 + tree	HKY and estimates an independent evolutionary rate and ts: tv bias for each window.
HKY+R+N	8 + tree	HKY and estimates an independent evolutionary rate and nucleotide frequencies for each window (nucleotide frequencies are estimated by counting).
HKY+R+N+T	9 + tree	HKY and estimates an independent evolutionary rate, nucleotide frequencies and ts: tv bias for each window.
HKY+R+G	6 + tree	HKY plus a single discretised γ distribution and estimates an independent evolutionary rate for each window.
HKY+R+T+G	7 + tree	HKY plus a single discretised γ distribution and estimates an independent evolutionary rate and ts: tv bias for each window.
HKY+R+N+G	9 + tree	HKY plus a single discretised γ distribution and estimates an independent evolutionary rate and nucleotide frequencies for each window.
HKY+R+N+T+G	10 + tree	HKY plus a single discretised γ distribution and estimates an independent evolutionary rate, nucleotide frequencies and ts: tv bias for each window.
HKY+R+2G	7 + tree	HKY and estimates an independent evolutionary rate and the rate heterogeneity for each window (i.e., 3 discretised γ distributions).
HKY+R+T+2G	8 + tree	HKY and estimates an independent evolutionary rate, ts: tv bias and the rate heterogeneity for each window.
HKY+R+N+2G	10 + tree	HKY and estimates an independent evolutionary rate, nucleotide frequencies and the rate heterogeneity for each window.
HKY+R+N+T+2G	11 + tree	HKY and estimates an independent evolutionary rate, ts: tv bias, nucleotide frequencies and the rate heterogeneity for each window.

Specific tests can be designed by comparing the models in Table 5.1. The tests can be designed to satisfy the conditions required to use the χ^2 approximation to the LRT (detailed in Chapter 1, section 1.41) and thus the test statistics are twice the log likelihood differences between the alternate hypothesis models and null hypothesis models. The statistical tests between nested null hypothesis models and the more complex alternate models are presented in Table 5.2. Note that further comparisons between different models are also possible but are not carried out because the tests presented in Table 5.2 already cover a wide range of hypotheses. Certain combinations, such as HKY+2G vs. HKY+G, are not carried out because of the limits of PAML software; HKY+2G is not possible without a comparison of evolutionary rate between the pair of windows at the same time for example (i.e., HKY+R+2G is possible but HKY+2G is not).

Table 5.2 – The evolutionary models that will be applied to each pair of windows of DNA data in the multiple alignment. The 'W' in the test name, standing for 'Windows', enables these tests to be distinguished from the tests used in Chapter 2 and Chapter 3 (continued on next page).

Test	Alternate (A) and null models (B) in LRT	Degrees of freedom in LRT	Biological question addressed by the LRT
T-W1	A: HKY+R B: HKY	1	Is there a significant difference in the rate of evolution between windows?
T-W2	A: HKY+R+G B: HKY+G	1	Having considered rate heterogeneity in the null and alternate models, is there a significant difference in the rates of evolution between windows?

Table 5.2 continued.

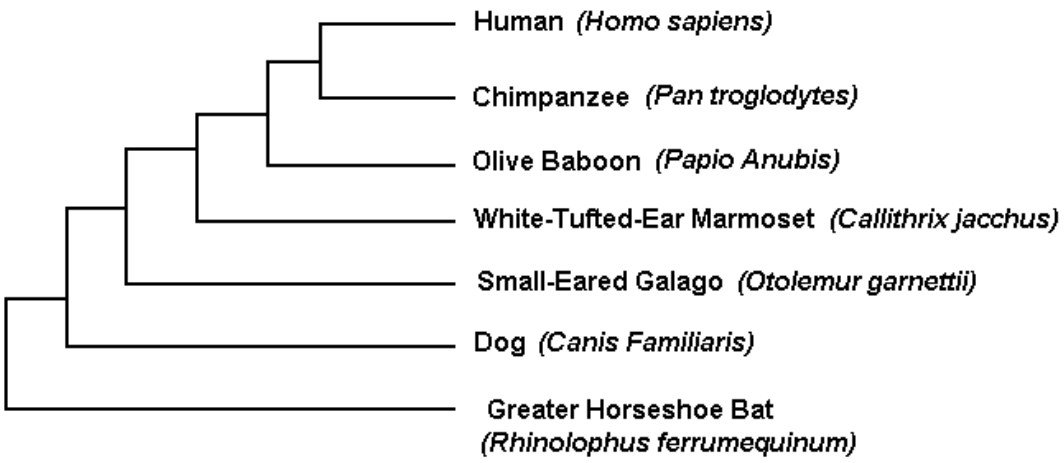
Test	Alternate (A) and null models (B) in LRT	Degrees of freedom in LRT	Biological question addressed by the LRT
T-W3	A: HKY+R+2G B: HKY+R+G	1	Is there a significant difference in the rate heterogeneity between windows?
T-W4	A: HKY+R+N+T+2G B: HKY+R+N+T+G	1	Having considered differences in rate, ts: tv bias and nucleotide frequencies at different windows in the null and alternate models, is there a significant difference in rate heterogeneity between windows?
T-W5	A: HKY+R+T B: HKY+R	1	Is there a significant difference in the ts: tv bias between windows?
T-W6	A: HKY+R+N+T B: HKY+R+N	1	Having considered differences in nucleotide frequencies between windows in the null and alternate models, is there a significant difference in the ts: tv bias between windows?
T-W7	A: HKY+R+T+G B: HKY+R+G	1	Having considered rate heterogeneity in the null and alternate models, is there a significant difference in the ts: tv bias between windows?
T-W8	A: HKY+R+N+T+G B: HKY+R+N+G	1	Having considered rate heterogeneity in the null and alternate models and the nucleotide frequency differences between windows, is there a significant difference in the ts: tv bias between windows?
T-W9	A: HKY+R+N+T+2G B: HKY+R+N+2G	1	Having considered differences in rate, rate heterogeneity and nucleotide frequencies at different windows in the null and alternate models, is there a significant difference in the ts: tv bias between windows?
T-W10	A: HKY+R+N B: HKY+R	3	Is there a significant nucleotide frequency difference between windows?
T-W11	A: HKY+R+N+T B: HKY+R+T	3	Having considered differences in the ts: tv bias between windows in the null and alternate models is there a significant difference in nucleotide frequencies between windows?
T-W12	A: HKY+R+N+G B: HKY+R+G	3	Having considered rate heterogeneity in the null and alternate models, is there a significant difference in nucleotide frequencies between windows?
T-W13	A: HKY+R+N+T+G B: HKY+R+T+G	3	Having considered rate heterogeneity in the null and alternate models and the ts: tv bias differences between windows, is there a significant difference in nucleotide frequencies between windows?
T-W14	A: HKY+R+N+T+2G B: HKY+R+T+2G	3	Having considered differences in rate, rate heterogeneity and ts: tv bias at different windows in the null and alternate models, is there a significant difference in nucleotide frequencies between windows?

5.4 Method

I explore the spatial scales over which our estimates of evolutionary parameters vary in both yeasts and mammals using the two largest *Saccharomyces cerevisiae* chromosome alignments (chromosomes 4 and 7), made using the smaller alignments from the data of Kellis *et al*. (2003) (see Chapter 3), and using a subset of species from the first two target regions in the September 2005 freeze of the ENCODE dataset (The ENCODE Project Consortium, 2004). ENCODE target region 1 spans the Cystic Fibrosis Transporter Gene (CFTR) on human chromosome 7; ENCODE target region 2 spans the Interleukin genes on human chromosome 5. Some sequence is available for many different species for regions homologous to both human ENCODE target regions but the coverage varies extensively for different species. Thus, I have chosen to study seven species for which a reasonable amount of data is available for both targets (roughly 50% or more of the original ENCODE multiple alignment length is represented by each species). Sequences for other species were removed from the multiple alignment of each target region and then alignment columns that contained only gaps in the remaining seven species were removed. This provided a rapid means to a multiple alignment for each target region and I posit that removing additional species will have little effect on the quality of the remaining alignment. The yeast chromosome multiple alignments were ~1.6 Mb and ~1.1 Mb for chromosomes 4 and 7, respectively. The ENCODE target multiple alignments were ~2.0 Mb and ~1.1 Mb for targets 1 and 2, respectively, for the seven remaining species.

The phylogeny of the four yeast species in the *S. cerevisiae* chromosome alignments has already been presented in Figure 3.2; the phylogeny of the mammal species from the ENCODE regions is shown in Figure 5.1 (The ENCODE Project Consortium, 2004).

Figure 5.1 – The phylogeny of the mammal species used to compile adjusted datasets from ENCODE regions 1 and 2.



Each of the models detailed in Table 5.1 was applied to all of the pairs of windows applied to each of the four datasets, allowing the application of all the tests in Table 5.2. Each of the pair of windows was set to size 100, 200, 500 or 1000 nucleotides, which are representative of the range of window sizes commonly used for single gene biological analyses. The size of the central gap between the windows was varied between each dataset analysis from a very small size to a very large size; the sizes used were 0, 100, 200, 300, 400, 500, 750, 1000, 2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000, 25000, 30000, 40000 and 50000 nucleotides. A central gap of 0 nucleotides means that the second window starts at the nucleotide after the

final nucleotide of the first window in the alignment. The increment size between the start points of successive analyses along each multiple alignment was the same as the size of a single window itself (such that the first window in one analysis did not overlap the position of the first window from the previous analysis). Analyses were performed using the program BASEML in the PAML package (Yang, 1997) and custom written Perl scripts for the preparation of data files and extraction of results from the BASEML output files for each separate pair-window analysis.

To ensure that test statistics were derived from analyses that successfully optimised log likelihoods and parameter values for each model (see Chapter 1), several safeguards were put in place. Firstly, the results from pairs of windows were rejected unless each species had at least 50 nucleotides of sequence in the alignment for each of the pair of windows, regardless of the size of the window. Secondly, all results for a pair of windows were rejected if any of the test statistics for tests T-W1 to T-W14 were negative, which only occurs when the model that represents the alternate hypothesis fails to optimise.

When the null hypothesis model fails to optimise we obtain a test that is more positive than expected; this is harder to correct for because some analyses may produce test statistics that are highly positive for biological reasons (i.e., there have been very different patterns of evolution in each window of a window-pair). It is often the case that when a model fails to optimise properly the tree length obtained is very high, much higher than most of the analyses that did optimise successfully. Thus, results for a pair-window analysis were rejected where any model gave a tree length greater than 4 for any of the four datasets. It is coincidental that a cut-off tree length

of 4 was appropriate for both yeast and mammals. This value was chosen based on the histograms of tree lengths obtained for each of the four datasets (data not shown). One does not wish to choose a tree length that is too conservative and eliminate a pair-window analysis result unnecessarily. Furthermore, unlike the log likelihoods for each test T-W1 to T-W14, which may depend on the size of each window in a pair-window analysis (100, 200, 500 or 1000 nucleotides), the median tree length should not be particularly affected by the window sizes because tree lengths are given as the average number of changes *per nucleotide*. It is difficult to conceive of any biological reason why tree lengths for any of the separate datasets, which all have median values of less than 1, would be as high as 4. Such high tree lengths may occur in regions where alignment has been problematic or where evolution has been extraordinarily fast because of changes in the number of microsatellite repeats, say, but these evolutionary changes are not the point mutations we are interested in modelling.

Missing data in the multiple alignments may affect test statistic values and this is unavoidable for genomic alignments; thus, a requirement for a minimum amount of data in each window is imposed and it is assumed that missing data will not introduce any systematic bias to the results and only noise (since both the null and alternate models will be affected).

The necessary model optimisations are carried out for each test (Table 5.2), for each size of window pairs and for each different size of gap between the windows. A single test statistic value is used as a measure of how strong the signal is for a difference in evolutionary parameters between two windows. All test statistics are collected together for a given test, window size and gap size and the median and

distribution of the test statistics are calculated. Median log likelihoods ratios were calculated instead of mean values to avoid biasing our average results for each dataset analysis by a small number of extreme values (consisting of all of the pair-window analyses for one dataset with a given size of window and size of central gap between each pair of windows). Even if our stringent scheme for whether or not test results for any single pair-window analysis should be retained does not eliminate some test statistics where some models have failed to optimise, one hopes that the median test statistic values for each test are only affected very slightly and in the same way for each combination of test, dataset, window size and central gap size.

There are no issues that arise due to multiple testing *per se* because we are observing the median test statistic values over a large number of tests and not looking at each test statistic separately. Furthermore, we are more interested in observing trends in the median test statistics as the gap size between the pair of windows changes than in whether or not the median test statistics are themselves significant (although it is of interest to note whether the median test statistic becomes significant only after a given distance). In other words, test statistics are used as a measure of evolutionary heterogeneity and not for performing significance tests.

5.5 Results and Conclusions

5.5.1 Do different models affect our interpretation of changes in the same evolutionary factor over different spatial scales?

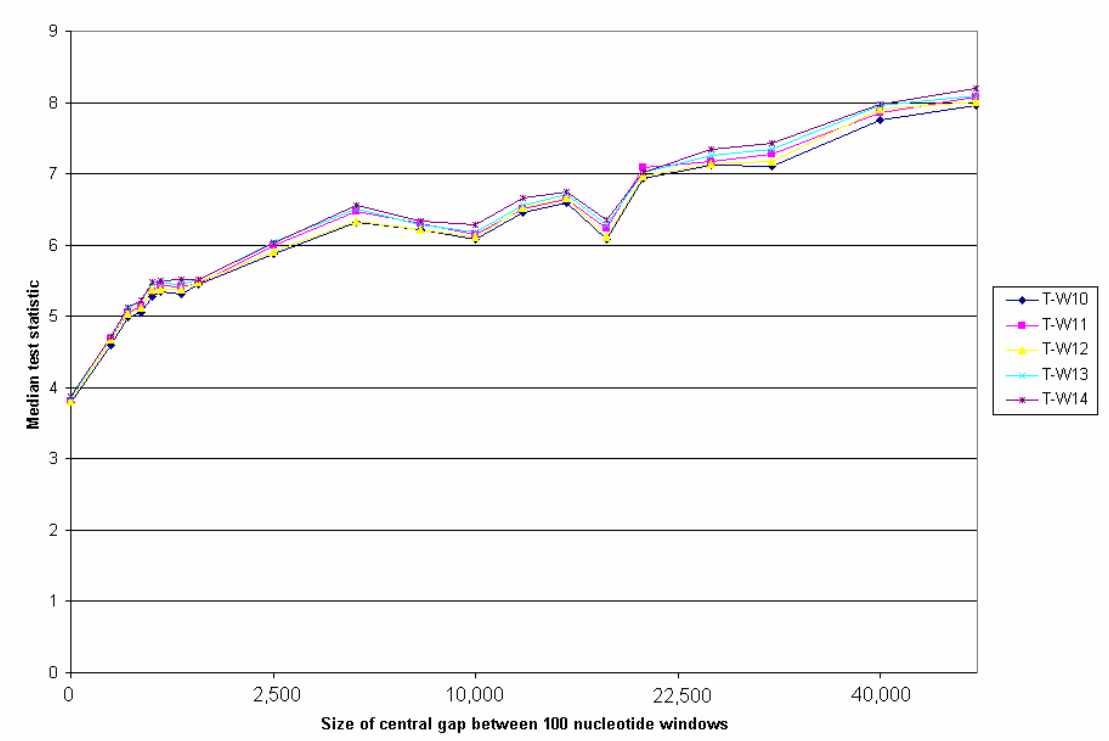
Before examining the spatial trends in evolutionary phenomena, I consider whether the complexity of the model can make qualitative differences to the results because there are different tests that study each of the spatial differences in evolutionary phenomena (e.g., T-W10 to T-W14 all study nucleotide frequency differences between windows), while accounting for various combinations of other parameters. Amongst the different tests for the same given phenomenon, the actual test one uses makes little difference to the results and conclusions we draw. I justify these claims in subsequent paragraphs. Whilst the values of the median test statistics may differ slightly between models, this makes very little difference to the distance over which different median parameter estimates may be significant between tests (according to the χ^2 approximation to the LRT).

If one concludes that there are spatial trends in the median test statistic values for any given type of test (e.g., T-W5 to T-W9 all observe difference in the transition: transversion bias between windows) then, theoretically, different models may be more appropriate than others at different spatial scales. One might expect a more complex model to produce significant results only above certain distances and this could affect our model choice. This is not the case, however, and I show that the most complex model for each phenomenon (rate, rate heterogeneity, transition: transversion bias and nucleotide frequencies) is generally the most powerful for the four datasets

investigated here; for this reason I favour the use of the complex models. Furthermore, even if the complex models were not the most powerful, this investigation uses an averaging process over a large number of window pairs; consequently one does not expect to lose much power when estimating more parameters. Nevertheless, using simple or complex models in the tests makes little difference when investigating the same evolutionary factor and the model one opts for is unlikely to dramatically affect the outcome of this investigation.

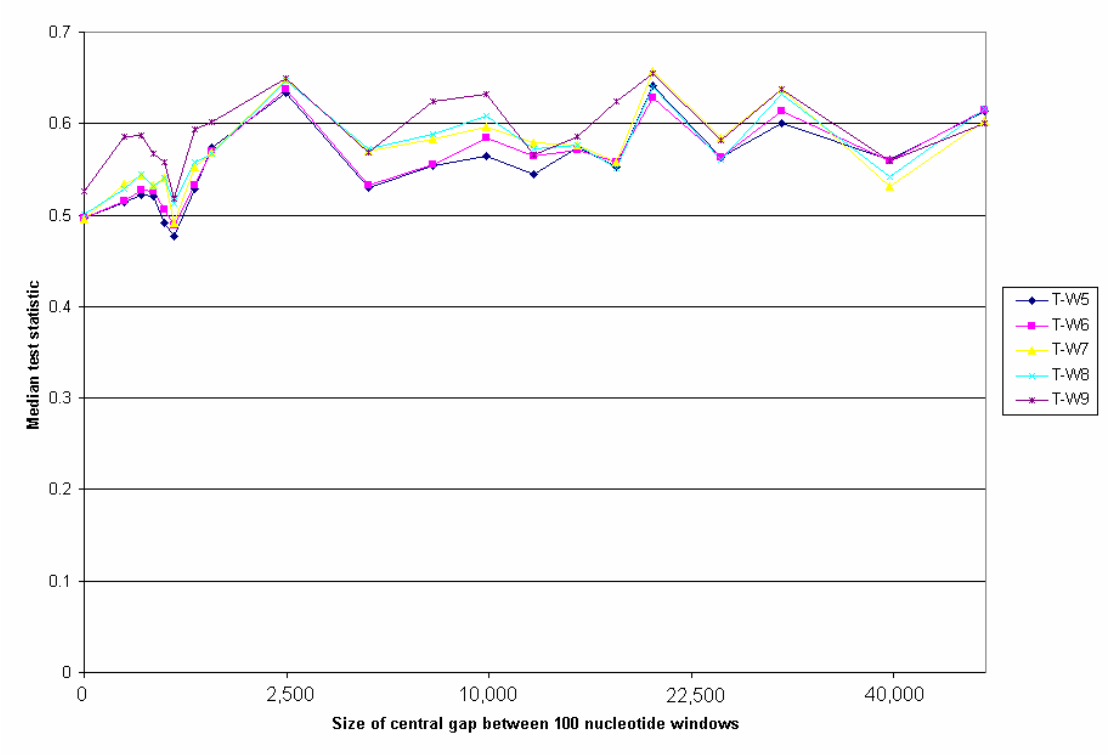
I focus the analysis of the results on tests where 100 nucleotide windows were used (results were similar for other window sizes; results not shown). In all datasets (yeasts and ENCODE) whether test T-W10, T-W11, T-W12, T-W13 or T-W14 is used makes little difference to the median estimate of the test statistic and the nucleotide frequency differences between windows. This is illustrated for ENCODE region 2 in Figure 5.2. On all plots in this chapter the *x*-axis is plotted using a square root transform with the *x*-axis labels corresponding to untransformed values; this increases the spacing between small gap size data points relative to the space between larger gap size data points, making results clearer at lower distances.

Figure 5.2 – Median test statistics for tests T-W10 to T-W14 for 100 nucleotide test windows for ENCODE target region 2.



Tests T-W5 to T-W9 all test for transition: transversion bias differences between windows; they differ in the other factors accounted for at the same time. Figure 5.3 shows the differences in the performance of the different models T-W5 to T-W9 for the ENCODE region 2 alignment. Again, the different tests give very similar results.

Figure 5.3 – Median test statistics for tests T-W5 to T-W9 for 100 nucleotide test windows for the ENCODE region 2 alignment.



For some tests applied to certain datasets, the choice of tests of certain biological variations is more significant. Tests T-W1 and T-W2 test for differences in evolutionary rates between pairs of windows. Whether test T-W1 or T-W2 is used seems to make little difference to the median test statistic for the yeast datasets but does make a more sizeable difference for the ENCODE datasets (Figure 5.4). Tests T-W3 and T-W4 test for differences in evolutionary rate heterogeneity between pairs of windows. Whether test T-W3 or T-W4 is used seems to make little difference to the median test statistic for the ENCODE datasets but does make a more sizeable difference for the yeast datasets (Figure 5.5).

Figure 5.4 – The median test statistics for tests T-W1 and T-W2 for the *S. cerevisiae* chromosome 4 alignment and ENCODE target region 1 alignment for 100 nucleotide test windows.

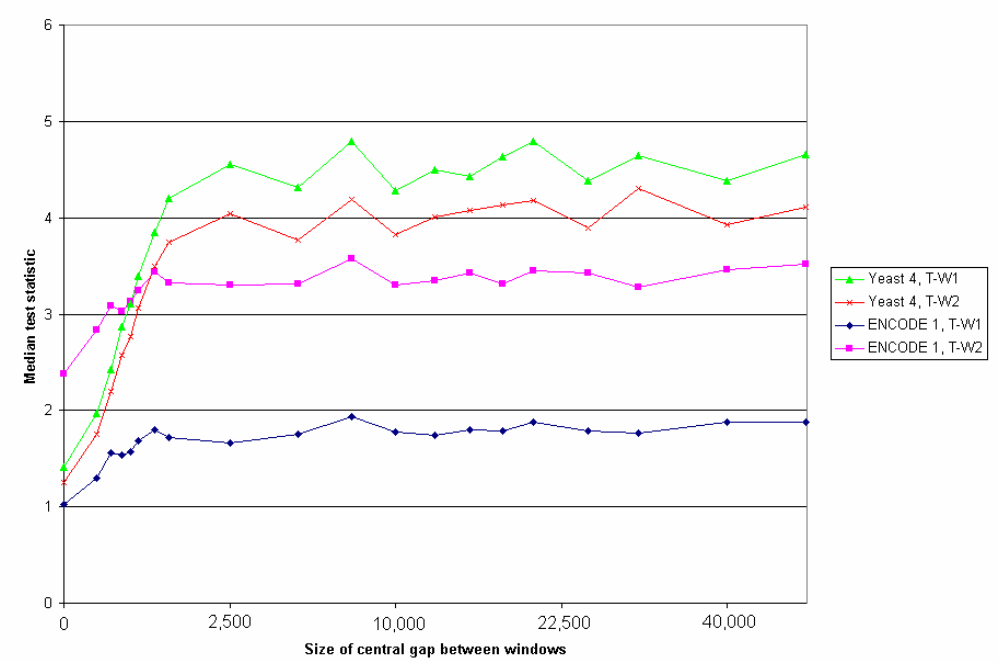
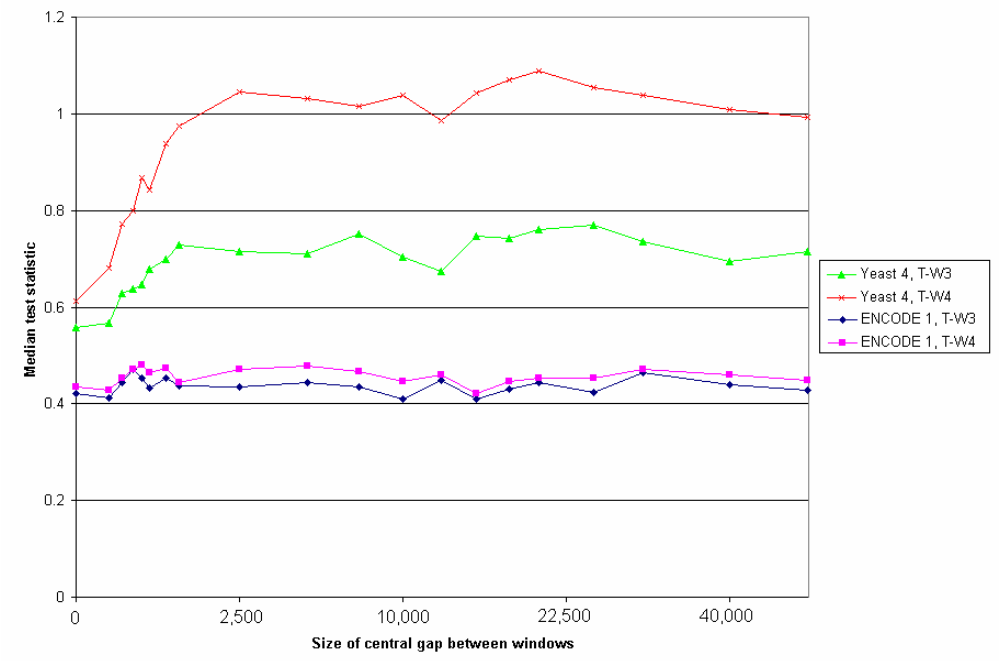


Figure 5.5 – The median test statistics for tests T-W3 and T-W4 for the *S. cerevisiae* chromosome 4 alignment and ENCODE target region 1 alignment for 100 nucleotide test windows.



The more complex models shown in Figures 5.2-5.5 give higher median test statistics than the more simple models (although this is less obvious in Figure 5.2 because of the scaling of the *y*-axis). This suggests that the complex models are gaining more power from more accurate modelling and less confounding of the modelled and un-modelled effects than they are losing from increased estimation uncertainty caused by greater modelling complexity. The more complex models also give consistently higher median test statistics for randomised datasets, which are used (below) to deliberately eliminate spatial trends (see Figures 5.12a-d). In light of these results, only the tests using the most complex models to study each biological factor will be used here (i.e., tests T-W2, T-W4, T-W9 and T-W14). However, all model approaches are valid and in future studies one may find examples where the less complex models are more powerful.

5.5.2 Are there spatial trends in the dynamics of sequence evolution?

The results support the notion that for evolutionary rate, rate heterogeneity, transition: transversion bias and nucleotide frequencies, windows of DNA that are closer together tend to evolve more similarly, i.e., tend to have more similar parameter estimates. The results are observed for most models, for most window sizes and for most datasets. The very stringent criteria that needed to be fulfilled for results of a single pair-window analysis to be retained meant that for the *S. cerevisiae* chromosome 7 alignment, there were too few results for some combinations of window size and gap size for the median value to be considered without suspicion. Since the step size between windows was the same as the size of one of the windows of a pair, there were most frequently too few results for the large window sizes. The

results of four models (T-W2, T-W4, T-W9 and T-14), which are the most complex models investigating differences in evolutionary rate, rate heterogeneity, transition: transversion bias and nucleotide frequencies respectively between pairs of windows are presented in Figures 5.6 and 5.7 (a, b, c and d for tests T-W2, T-W4, T-W9 and T-W14 respectively) for the *S. cerevisiae* chromosome 4 alignment and ENCODE target region 1 alignment, respectively.

Figure 5.6 (a-d) – Median test statistic results for tests T-W2 (5.6a), T-W4 (5.6b), T-W9 (5.6c) and T-W14 (5.6d) for all window sizes and central gap sizes for the *S. cerevisiae* chromosome 4 multiple alignment. The horizontal line that marks a significant evolutionary difference between one pair of windows (according to the χ^2 approximation to the LRT with the appropriate degrees of freedom for each test) is also shown (blue).

Fig. 5.6a – Test of spatial variation in evolutionary rate (T-W2) for yeast chromosome 4.

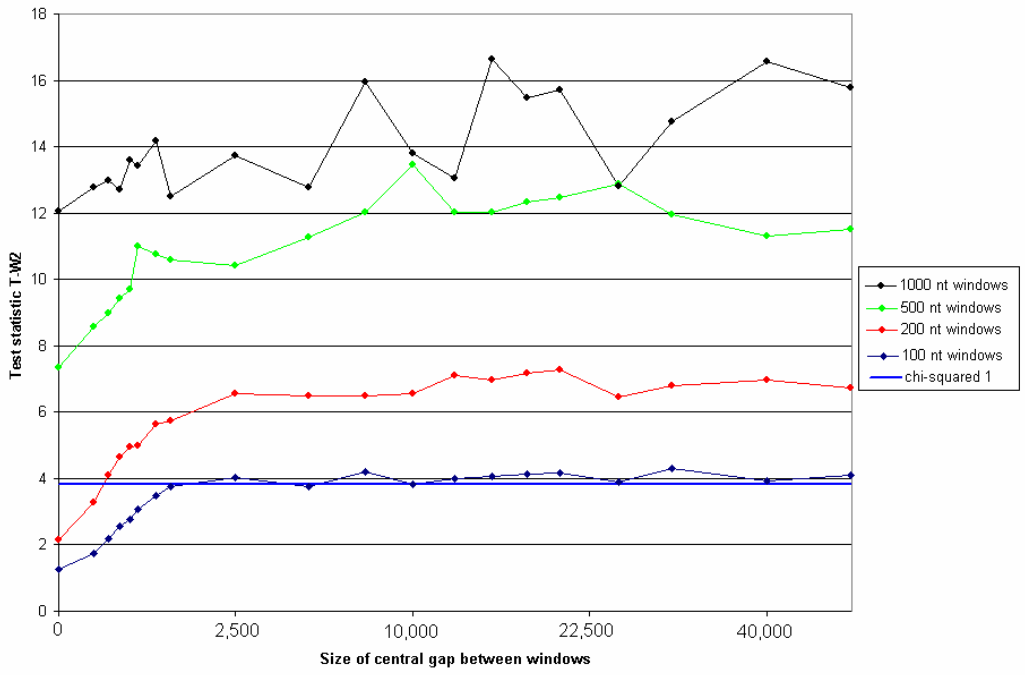


Fig. 5.6b – Test of spatial variation in evolutionary rate heterogeneity (T-W4) for yeast chromosome 4.

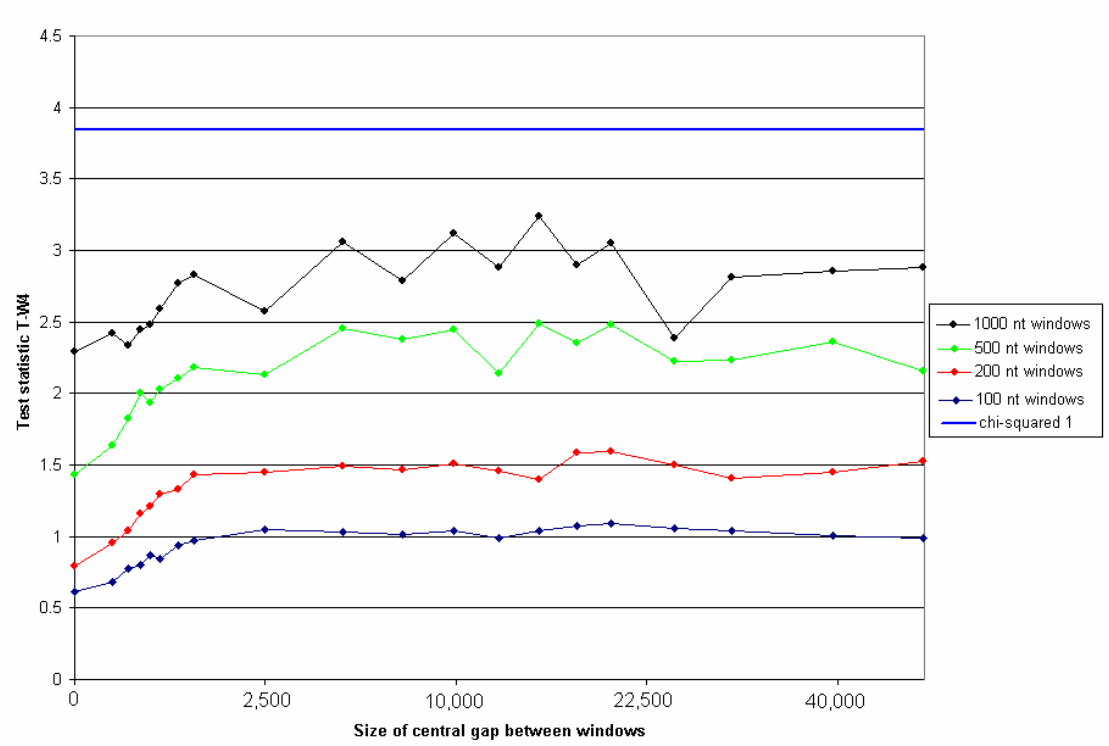


Fig. 5.6c – Test of spatial variation in ts: tv bias (T-W9) for yeast chromosome 4.

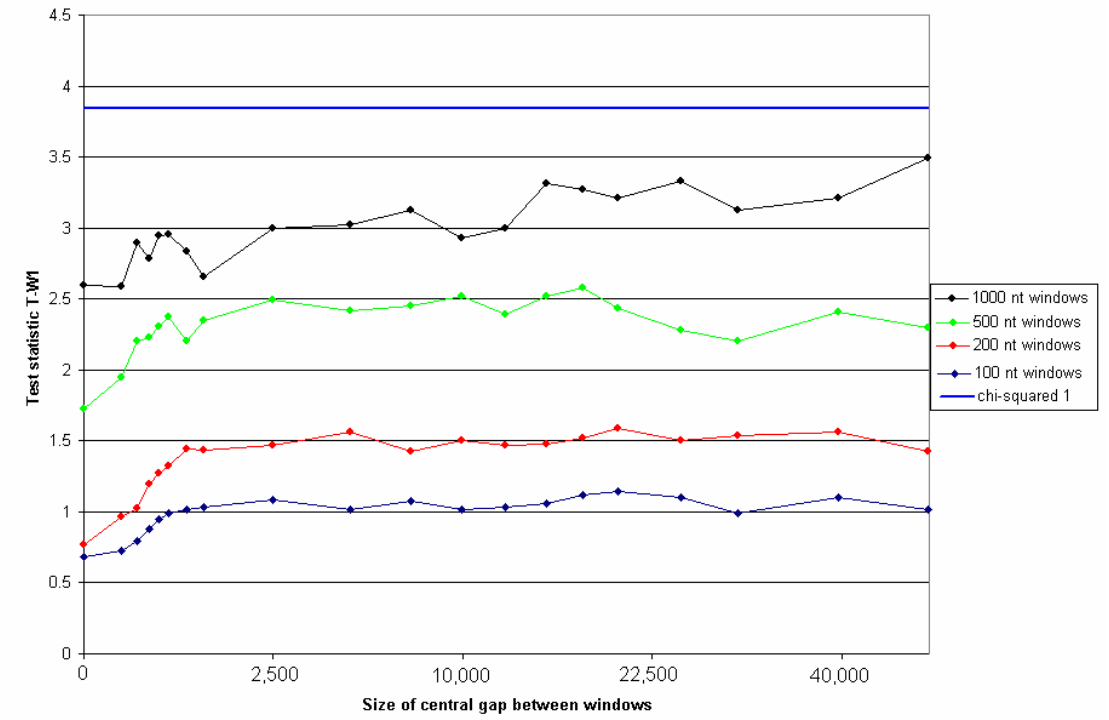


Fig. 5.6d – Test of spatial variation in nucleotide frequencies (T-W14) for yeast chromosome 4.

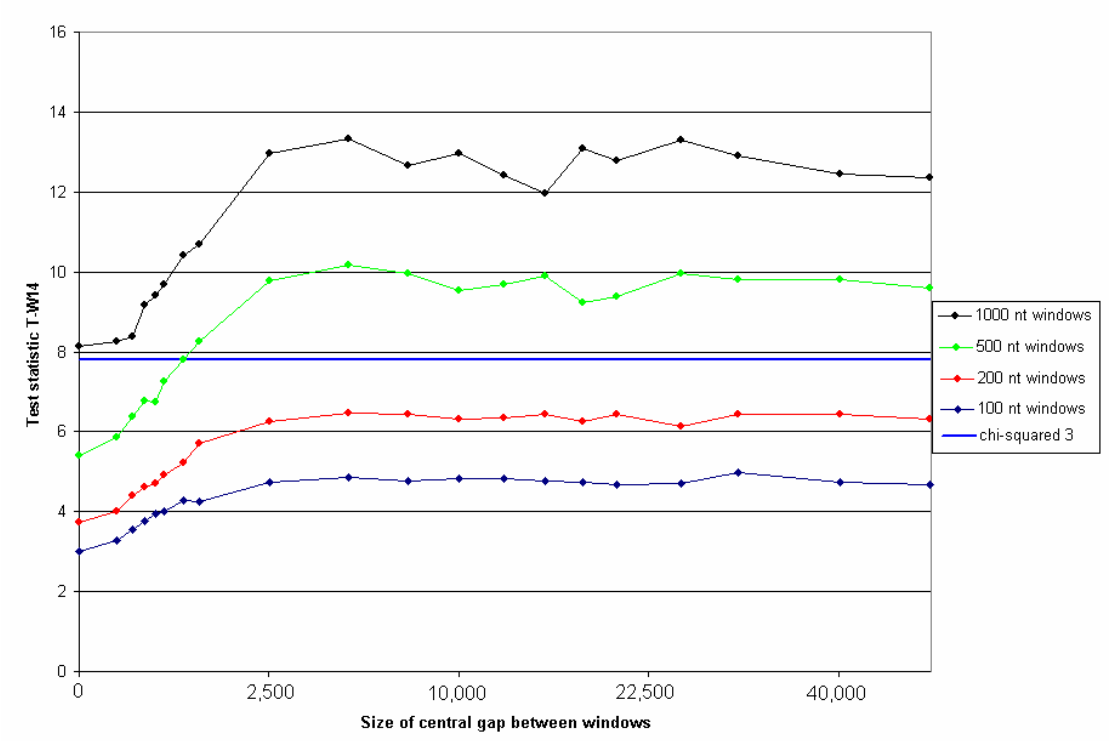


Figure 5.7 (a-d) – Median test statistic results for tests T-W2 (5.7a), T-W4 (5.7b), T-W9 (5.7c) and T-W14 (5.7d) for all window sizes and central gap sizes for the ENCODE target region 1 multiple alignment. The horizontal line that marks a significant evolutionary difference between the pair of windows (according to the χ^2 approximation to the LRT with the appropriate degrees of freedom for each test) is also shown (blue).

Fig. 5.7a – Test of spatial variation in evolutionary rate (T-W2) for ENCODE region 1.

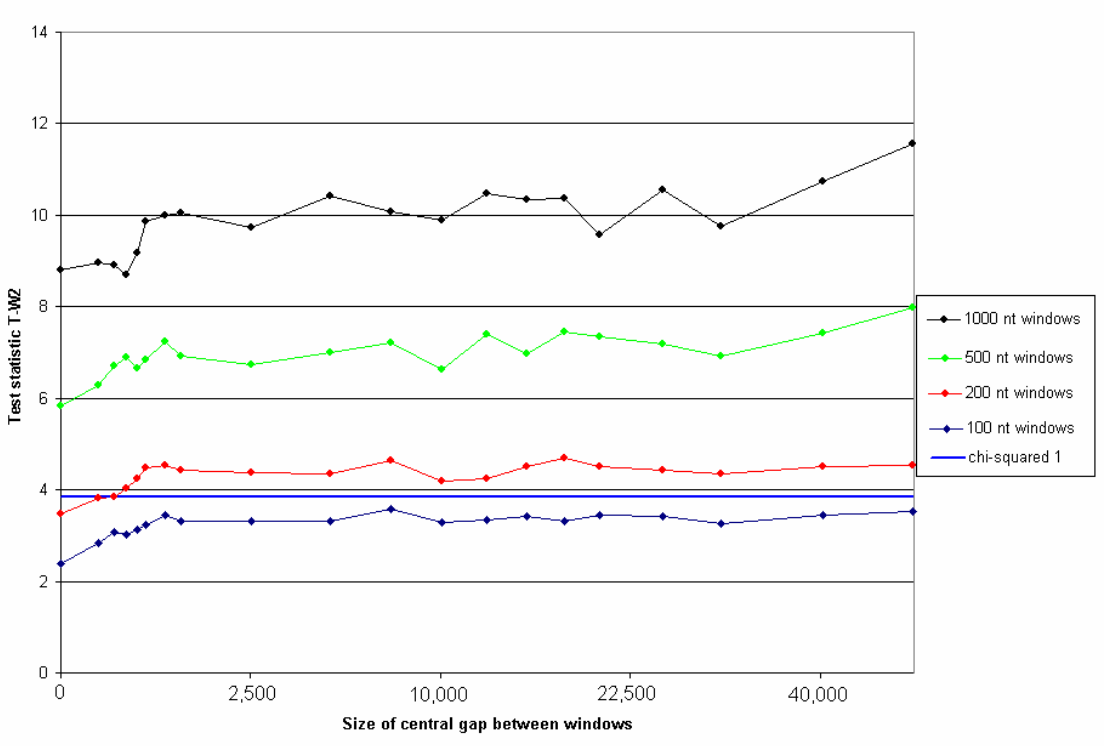


Fig. 5.7b – Test of spatial variation in evolutionary rate heterogeneity (T-W4) for ENCODE region 1. The test statistics are very small so the χ^2 significance line (at the value 3.84 on the *y*-axis) is not shown.

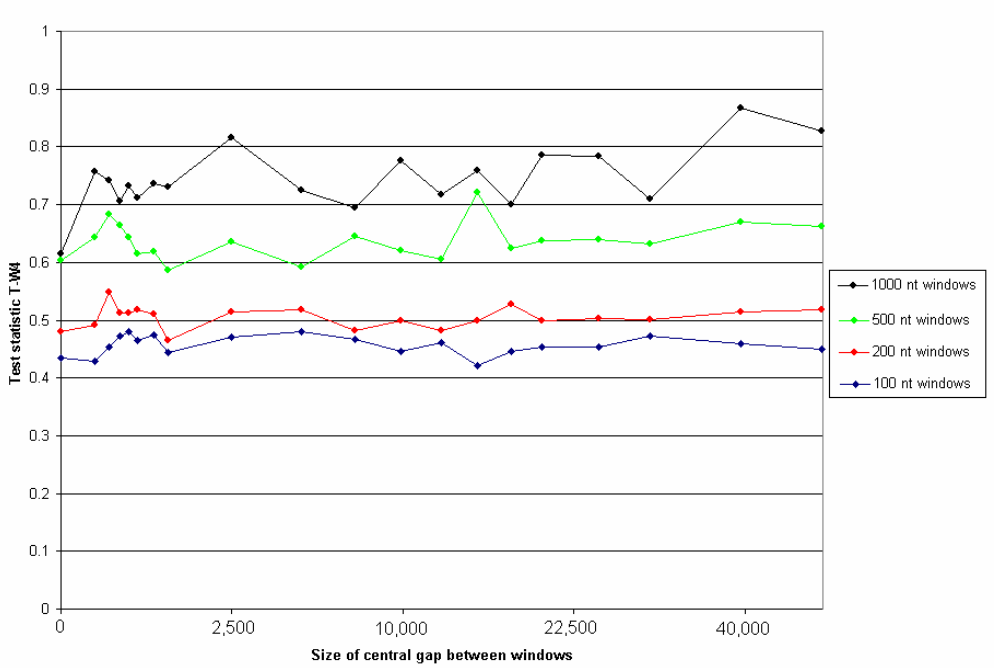


Fig. 5.7c – Test of spatial variation in ts: tv bias (T-W9) for ENCODE region 1. The test statistics are very small so the χ^2 significance line (at the value 3.84 on the *y*-axis) is not shown.

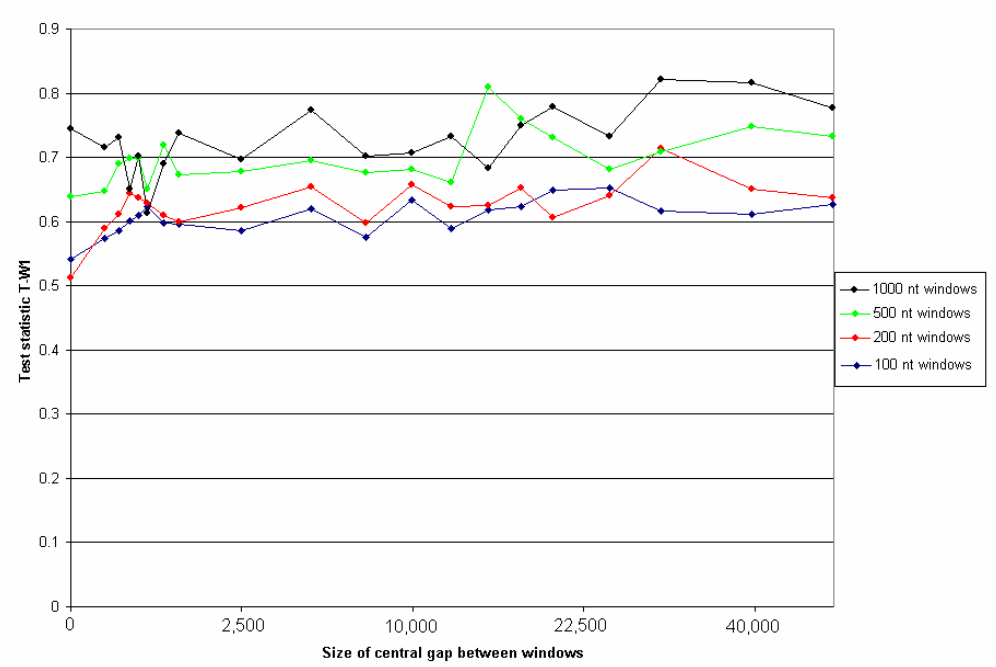
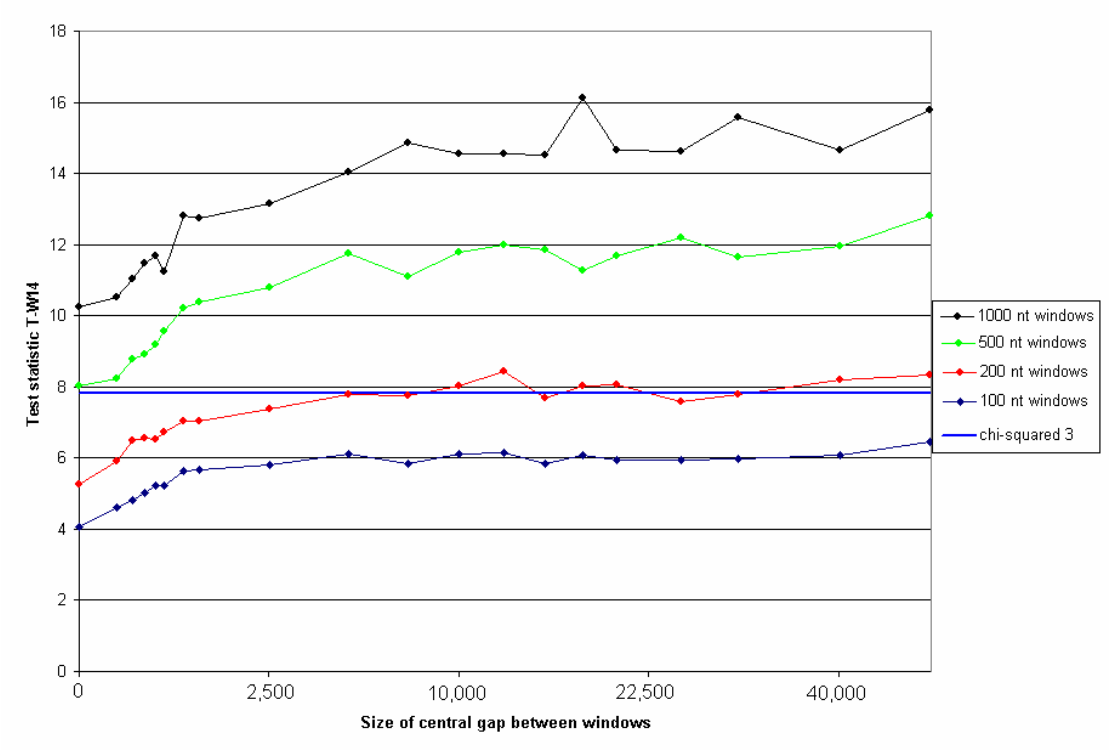


Fig. 5.7d – Test of spatial variation in nucleotide frequencies (T-W14) for ENCODE region 1.



There are dramatic differences in the way different evolutionary parameters change over different spatial scales. In Figure 5.6 (*S. cerevisiae* chromosome 4 results) each graph (a-d) shows the trend that there is an increase in the median test statistic value as the distance between pairs of windows increases from 0 to approximately 500-1000 nucleotides, for each window size. These trends are also observed in Figures 5.7a, c and d for the ENCODE regions for each window size. Beyond certain distances the plots become approximately flat (containing a 'plateau'), indicating that there is no greater level of heterogeneity at even greater distances. If one then observes Figure 5.7b the median test statistics (for tests T-4, testing rate heterogeneity) change very little at different distances between the pair of windows.

This suggests that there are no detectable spatial dependencies of the average differences in rate heterogeneity for this dataset.

If a trend is apparent only for very small distances and small window sizes, it may be that large window sizes are already too large to observe any spatial effect, i.e., the parameters change at different distances between windows but they have reached a plateau at a distance less than a single large window. This becomes important in the choice of the optimum window size for future investigations and is discussed in the next section. Additionally, there are fewer data points when larger window sizes are used (because the step size between analyses is the same as the windows size), which may affect our ability to detect spatial trends in parameter values for large window sizes; this effect is probably minor.

Thus, with respect to all phenomena investigated in yeast and all *but* rate heterogeneity in mammals, neighbouring regions tend to evolve more similarly than other regions on a chromosome that are more distant. The question of whether the effect is 'significant' is somewhat contrived since we are observing an average effect, which is a summary of a distribution of many LRT results. There are both significant and non-significant examples at all gap and window sizes; the graphs are showing the size of the average effect. Window sizes may affect the median test statistic quantitatively but do not tend to affect the trends observed in a qualitative manner, which is discussed more below. At some stage, for most window sizes, parameters and datasets, there is a distance at which the median test statistic value appears to plateau. At distances greater than the 'plateau-distance', two windows of DNA evolve, on average, with no greater difference than the difference at the plateau, with

respect to the feature that is being investigated by the given test. Not all model parameters appear to reach a plateau-distance within the very large distances (up to 50 Kb) studied in this investigation for all datasets. I discuss this in more detail in section 5.5.4.

5.5.3 What are the effects of using different window sizes?

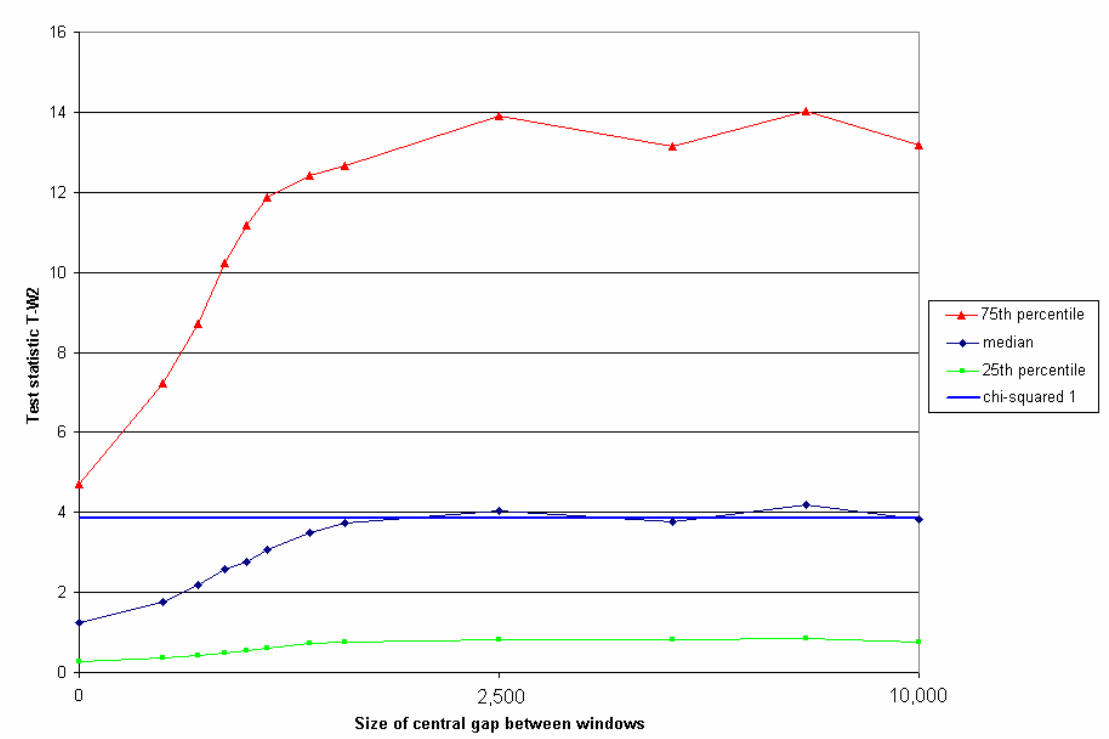
There is a clear effect of using different sizes of windows for each window in the pair-window analyses (see Figures 5.6 and 5.7). Larger window sizes result in larger median test statistic results in almost all cases. This indicates that larger windows lead to a greater ability to detect differences, which is expected, except in those cases where non-independence in evolutionary processes only occur over small scales. In such cases a single large window of a pair may already cover a distance that exceeds the plateau-distance (see Figure 5.7c, showing a spatial trend in transition: transversion biases that may only be detected over small distances with small window sizes in mammals). In most cases the effects of window sizes are generally not qualitative, only quantitative. Figure 5.7b is more complex because the line for each window size is almost flat, suggesting there is little spatial variation, but the lines do get higher as window sizes get bigger, which suggests an increased ability to detect variation. The increase in the median test statistic size is very small, however (and consider the different scales of the *y*-axis used in the different figures).

There is some evidence that using larger window sizes may allow us to detect different effects than small windows; evolutionary differences that only occur over different scales may be more easily detected by large windows (see below, e.g.,

Figure 5.11), discussed in detail later. I favour the use of small window sizes because there is a requirement for a large number of data points for the median to accurately reflect the underlying differences in evolutionary parameters at different spatial distances. However, if one made the increment size small between start points of subsequent pair-window analyses, instead of making the increment size the same as the size of a single window, this would not be an issue. However, this would require far more analyses and would be computationally expensive. Nevertheless, to ensure that we can observe effects that reach a plateau-distance over short distances, I focus most of the remainder of my analyses on results from small windows. I note that no single window size is best and all window sizes can be informative.

I present the interquartile ranges of the distribution of test statistics for the *S. cerevisiae* chromosome 4 multiple alignment in Figure 5.8 for test T-W2 for distances of less than or equal to 10000 nucleotides between a pair of windows each of size 100 nucleotides (cf. Figure 5.6a). It is clear from Figure 5.8 that there is a very wide spread in the test statistic results for any single set of pair-window analyses at a given central gap size and this illustrates the requirement of a large number of data points to reliably estimate the median test statistic value. This is typical of the results of any test for any other parameter combination for any dataset in this study, which would not have been possible in the pre-genomic era.

Figure 5.8 – The median and interquartile ranges for test T-W2, window sizes 100 nucleotides, central gap sizes less than or equal to 10000 nucleotides for the *S. cerevisiae* chromosome 4 multiple alignment. The line that marks a significant evolutionary difference between the pair of windows according to the χ^2 approximation to the LRT with the appropriate degrees of freedom for each test is also shown.



5.5.4 Are there differences in the spatial scales over which different factors change?

Considering the existence of spatial differences in the factors studied (rate, rate heterogeneity, ts: tv bias and nucleotide frequencies), I now ask whether these factors differ over different distances: do the factors have different 'plateau-distances' (the distance between two windows after which the evolutionary processes for given parameters do not become any more different)? I focus the analysis of results on

window sizes of 100 nucleotides and the most complex tests examining each factor. Figure 5.9 presents the median test statistics for 100-nucleotide windows for the *S. cerevisiae* chromosome 4 multiple alignment for tests T-W2 (evolutionary rate), T-W4 (rate heterogeneity), T-W9 (transition: transversion bias) and T-W14 (nucleotide frequency) for distances of up to 10000 nucleotides. The corresponding results for the ENCODE region 1 alignments are shown in Figure 5.10. (Note that the details in Figures 5.9 and 5.10 have been shown previously within Figures 5.6a-d and 5.7a-d, respectively.)

Figure 5.9 - Median test statistics for 100-nucleotide windows for the *S. cerevisiae* chromosome 4 multiple alignment for tests T-W2 (evolutionary rate), T-W4 (rate heterogeneity), T-W9 (transition: transversion bias) and T-W14 (nucleotide frequency) for distances of up to 10000 nucleotides. Note that test T-W14 involves three degrees of freedom, the other tests involve one degree of freedom. We are interested in observing the distances where the test statistics seem to plateau and not the values themselves.

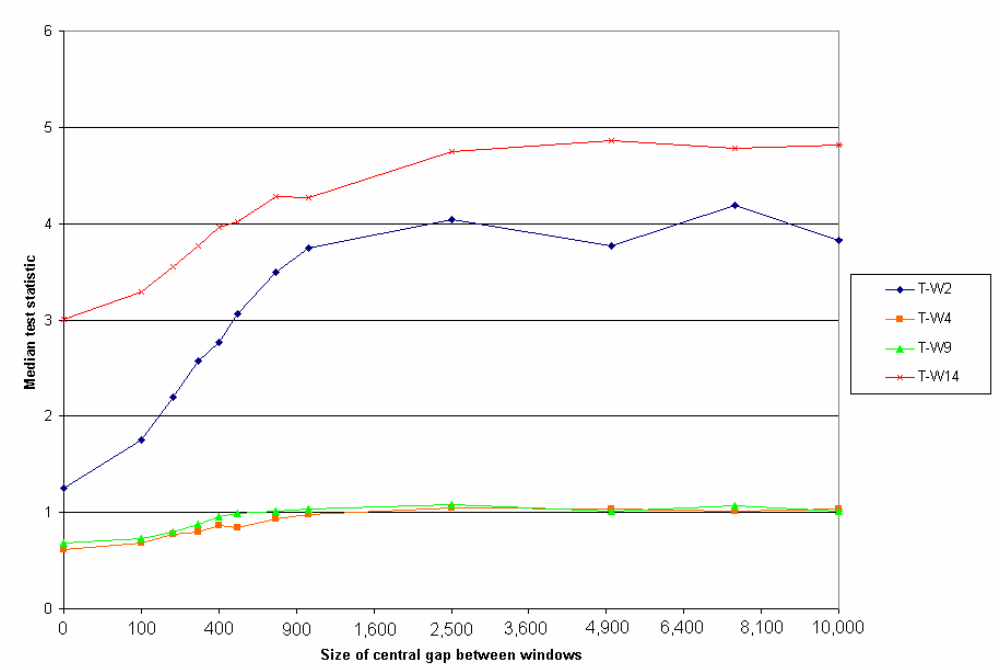
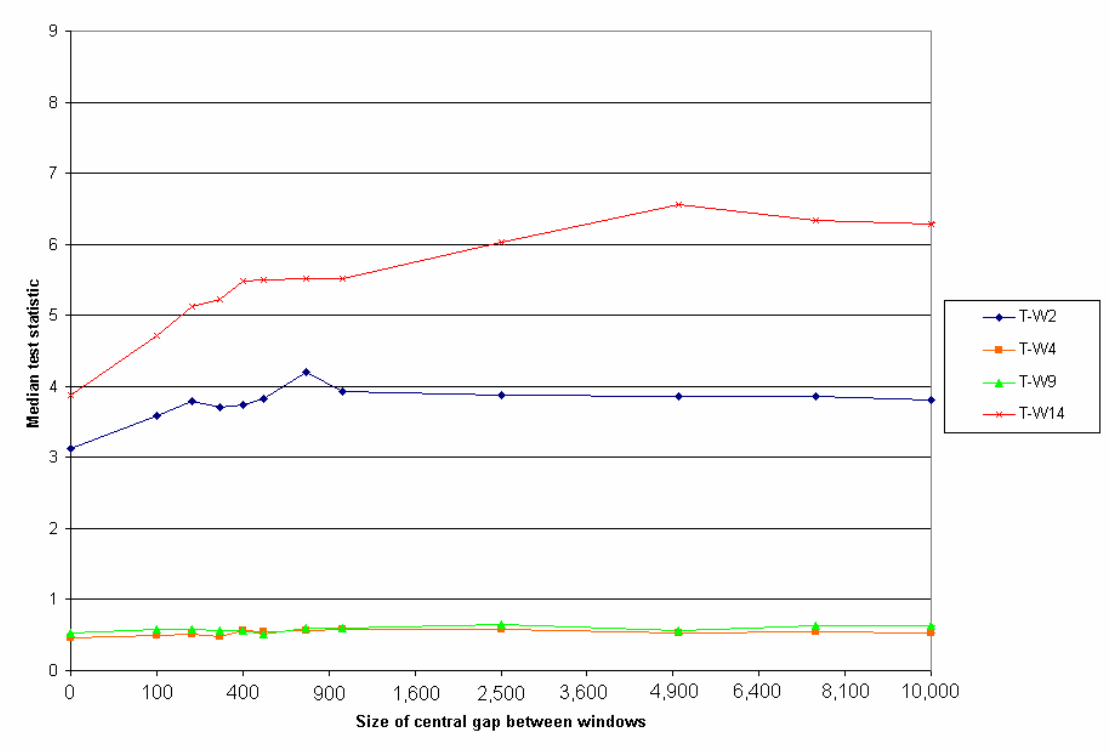


Figure 5.10 - Median test statistics for 100 nucleotide windows for the ENCODE target region 1 multiple alignment for tests T-W2 (evolutionary rate), T-W4 (rate heterogeneity), T-W9 (transition: transversion bias) and T-W14 (nucleotide frequency) for distances of up to 10000 nucleotides.



From Figures 5.9 and 5.10, there are clear differences in the plateau-distances of different factors (and these are not affected by the window size used). The plateau-distances of each factor are presented in Table 5.3 for yeast chromosome 4 and ENCODE region 1. In terms of the genomic ranges over which parameters vary, nucleotide frequencies are correlated over greater distances than evolutionary rates, which are correlated over greater distances than rate heterogeneity, which are correlated over similar distances as the transition: transversion bias (for both yeast chromosome 4 and ENCODE region 1). Even regions that are fairly close together evolve more-or-less independently with respect rate heterogeneity and transition: transversion biases. It appears that those factors that produce high median test

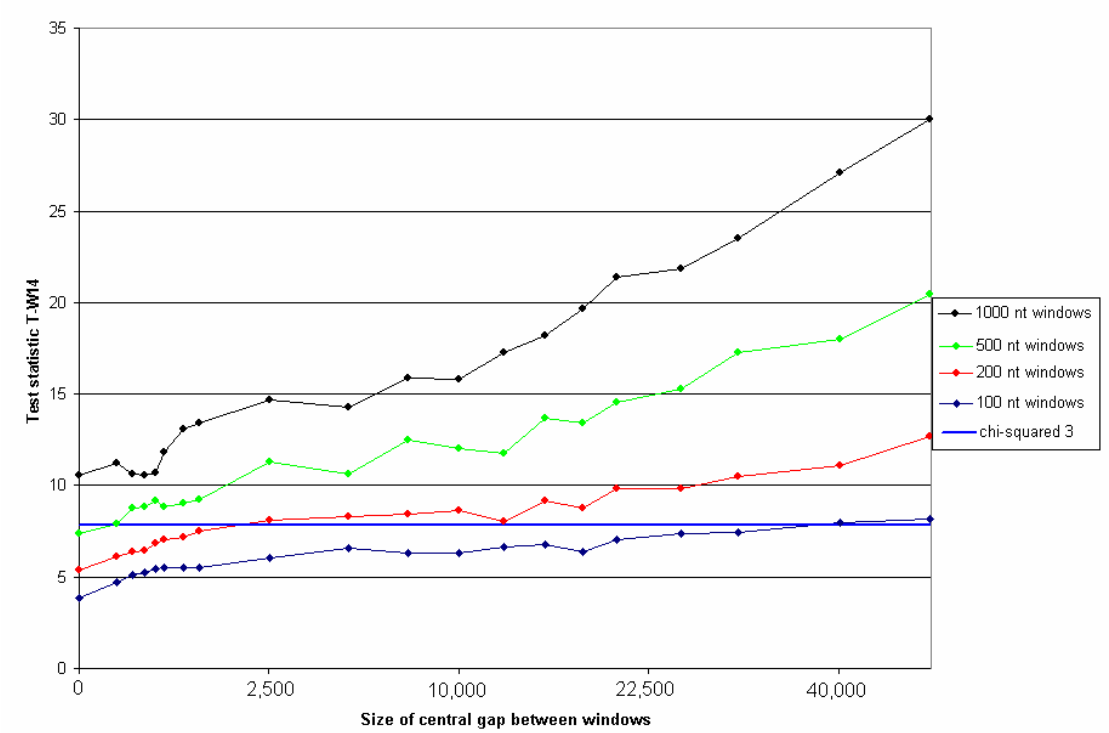
statistics (e.g., nucleotide frequencies) are also those where the spatial trend is more evident and there tends to be larger plateau-distances. It is also interesting to note that the same parameters differ less than evolutionary rate and nucleotide frequencies (in terms of allowing for differences between sites and subsequent improvement in model fit) between different sites in a codon too (see Chapter 2).

Table 5.3 – Approximate plateau-distances for each factor for yeast chromosome 4 and ENCODE region 1.

Factor Modelled	Test	Yeast Chromosome 4	ENCODE region 1
Rate	T-W2	2.5 Kb	1 Kb
Rate heterogeneity	T-W4	1 Kb	0.5 Kb
Transition: transversion bias	T-W9	1 Kb	0.5 Kb
Nucleotide frequencies	T-W14	5 Kb	10 Kb

Figure 5.11 shows the median test statistics for modelling differences in nucleotide frequencies between a pair of windows for test T-W14, for all window sizes for ENCODE target region 2. Unusually, in relation to the other results, there seems to be no plateau within the range of central gap sizes investigated. This indicates that regional variation at a much greater scale is possible. It is not clear from the results of ENCODE region 1 whether the median test statistic for T-W14 continues to rise with the distance between the pair of windows because, if there is a trend, it is much more subtle (see Figure 5.7d).

Figure 5.11 – Test statistics for test T-W14 for all window sizes (100, 200, 500 and 1000 nucleotides) for ENCODE target region 2. Note that there is no plateau in the median test statistic value up to a distance of 50000 nucleotides between the pair of windows.



5.5.5 Might the observed trends occur by random effects?

We wish to assess whether random variation could cause the trends in median test statistics observed for all tests (T-W1 to T-W14) for the two ENCODE datasets and for the yeast datasets (although in yeast the median test statistics for spatial differences in rate heterogeneity and transition: transversion biases only show apparent non-independence over very short distances and the effects are very small). We could construct many non-parametric bootstraps of each dataset and repeat the analyses; this would be very computationally expensive and is therefore impractical.

Alternatively, we could randomise the distance categories to which the real test statistics are assigned. If the size of the central gap between windows does not affect the median test statistic values, the apparent trends between the windows would be maintained. Different gap sizes produced different numbers of test statistic results; this could theoretically cause the apparent trends we observe. I produce 100 randomisations for pairs of windows of 100 nucleotides, which I then compare to the trends seen in the observed data, for each dataset. The results of the reassignment of each test statistic to a random central-gap distance category for ENCODE region 1 are presented in Figures 5.12 a-d. None of the trends observed in Figure 5.6 a-d and Figure 5.7 a-d are evident in Figure 5.12 a-d. Thus, the different number of pair-window analyses used to produce the different median test statistics does not introduce bias. The spatial non-independence of evolutionary forces (evident in Figures 5.6 and 5.7) is non-random, which means these patterns must be caused by biological factors.

Figure 5.12 (a-d) – The median test statistic values for the randomised ENCODE region 1 results for all tests pertaining to (a) evolutionary rate (T-W1 and T-W2), (b) rate heterogeneity (T-W3 and T-W4), (c) transition: transversion bias (T-W5 to T-W9) and (d) nucleotide frequencies (T-W-10 to T-W14) for 100 nucleotide windows.

Fig. 5.12a – randomised T-W1 and T-W2 for ENCODE region 1.

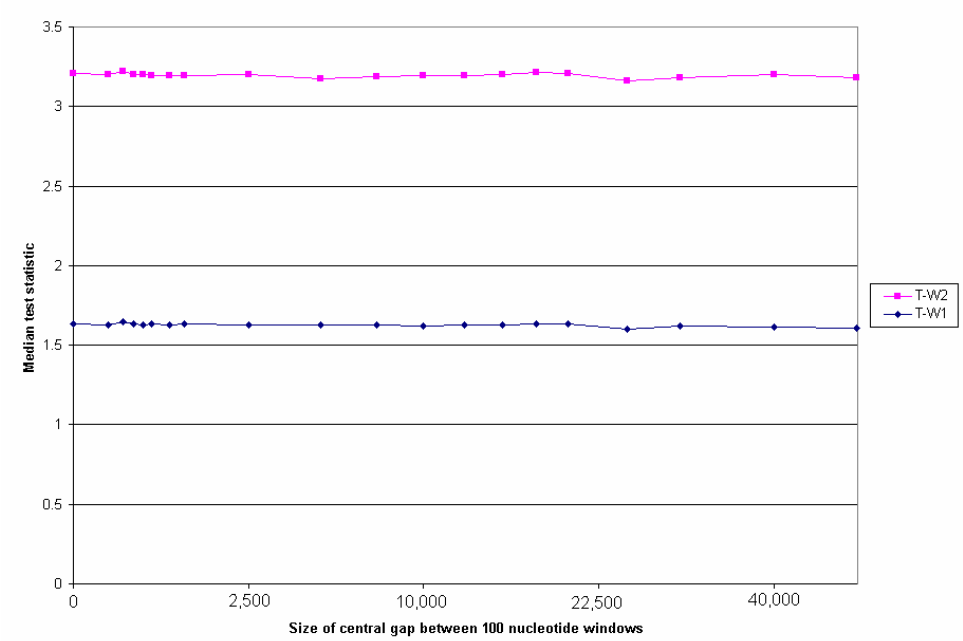


Fig. 5.12b – randomised T-W3 and T-W4 for ENCODE region 1.

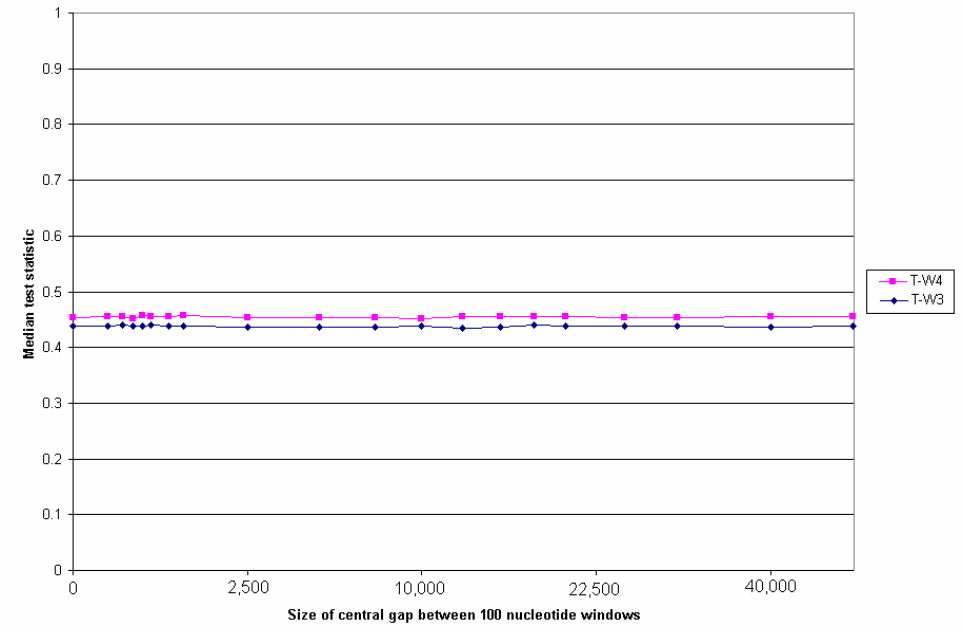


Fig. 5.12c – randomised T-W5 to T-W9 for ENCODE region 1.

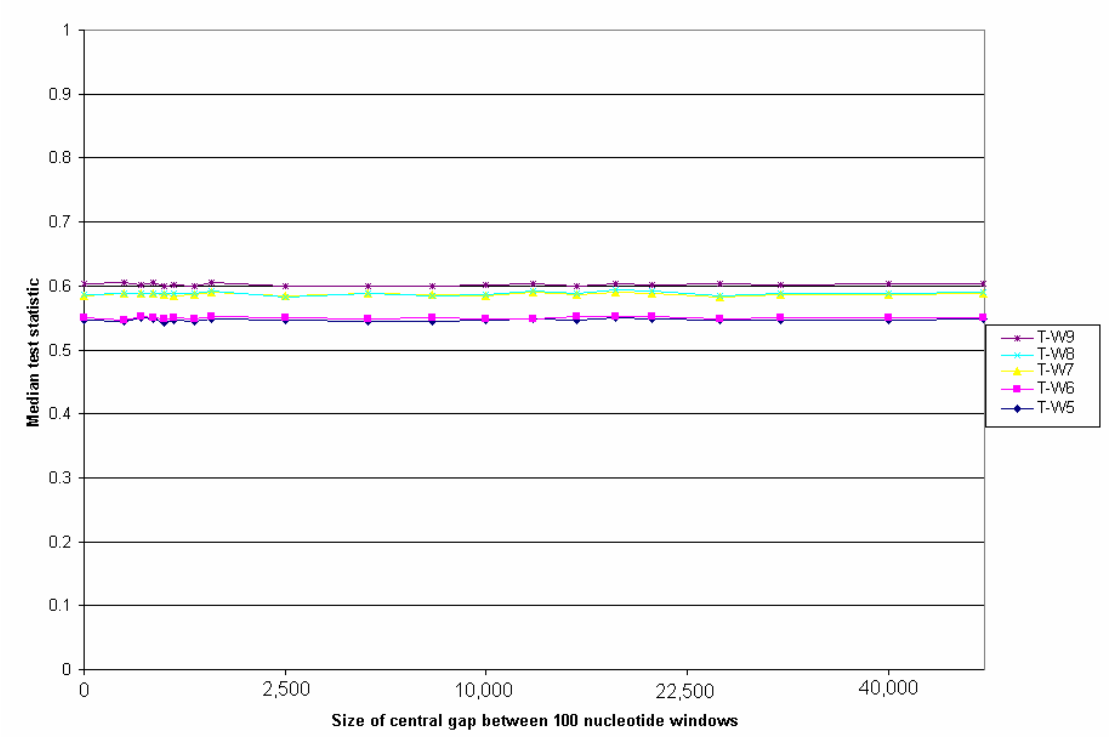
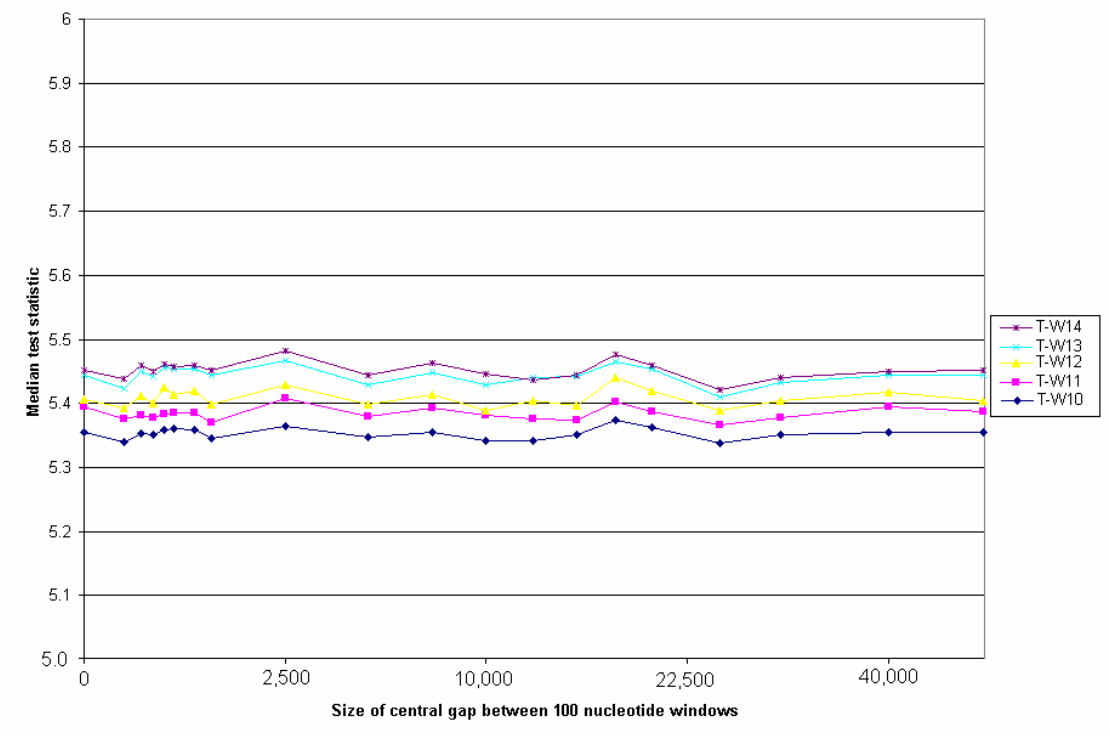


Fig. 5.12d – randomised T-W10 to T-W14 for ENCODE region 1. Note the scale on the *y*-axis starts at 5.



5.5.7 Are there differences within and between the yeast and mammalian datasets?

There is a strong general agreement amongst most median test statistics between the two yeast datasets and between the two mammalian datasets. Within the ENCODE dataset the size of the test statistics and the effects of distance between pairs of windows on the behaviour of the test statistics is very similar for tests T-W1 to T-W9, which include the tests for evolutionary rate, rate heterogeneity and transition: transversion bias. The only differences in the behaviour of median test statistics between the ENCODE datasets seems to be for tests T-W10 to T-W14 (concerning differences in nucleotide frequencies between windows). ENCODE target region 2 shows a much stronger trend for a continued increase in the median test statistics for tests T-W10 to T-W14 than ENCODE target region 1, for all distances investigated (see Figures 5.11 and 5.7d). ENCODE target region 1 may still show an increase in the median test statistic values for tests T-W10 to T-W14 but the effect is much more subtle. The difference between the two ENCODE datasets may be because ENCODE target region 2 contains data from more than one isochore (regions with characteristic nucleotide contents detectably different from other regions; Belle *et al*., 2004; Montoya-Burgos *et al*., 2003; Eyre-Walker and Hurst, 2001). Whilst the definition of isochores is somewhat hazy, the GC percent tracks on the UCSC genome browser clearly show regions of contrasting GC contents within ENCODE target region 2, which is not seen for ENCODE target region 1 (data not shown).

In contrast, all median test statistics follow very similar trends between the two yeast datasets. Between the yeast and ENCODE datasets, tests T-W1 and T-W2 (evolutionary rate) have similar patterns of change with distance between windows. Differences in rate heterogeneity and transition: transversion biases at different distances are more marked in the yeast datasets than in the ENCODE datasets and this is most likely due to the high gene density in yeast. If a very small DNA window is non-coding in yeast then the next very small window is also likely to be non-coding. As the distance between windows increases the probability of the second window being in a gene-encoding region changes. Over distances of a few thousand nucleotides these probabilities change until we reach distances beyond which regions are essentially independent and spatial constraints between genes are no longer a factor. The differences in the factors investigated here are likely to be high when one window is in non-coding DNA and the other window is in coding DNA. Aside from gene density, the yeast and mammalian datasets also differ in the properties of tests T-W10 to T-W14, probably because the isochore structure of large genomic regions does not exist in yeast. The yeast test statistics tend to be larger than the same test for the ENCODE region because of the greater evolutionary distance between the yeast species than the mammalian species in the different datasets (see Figures 5.4 and 5.5, for example).

5.6 Discussion

This study has investigated the spatial trends in evolutionary phenomena (evolutionary rate, rate heterogeneity, transition: transversion biases and nucleotide frequencies) measured by maximum likelihood models. In both mammals and yeasts we observe that regions that are physically closer to each other tend to evolve more

similarly than regions that are more distant. This study has been possible because of the large genomic-scale multiple alignments that have been made available recently. Reliable estimation of the median test statistics requires a large number of tests because separate pair-window test statistics depend on the particular evolutionary histories of each window of data. I have demonstrated that most median test statistics are smaller for short distances between windows, and rise to a plateau. At distances above a loosely defined 'plateau-distance', windows evolve independently with respect to the parameters we are investigating for a specific test (e.g., evolutionary rate with test T-W2). Plateau-distances are generally less than 5 Kb. However, not all evolutionary phenomena produce median test statistics that plateau within the range of distances between pair-windows investigated here (nucleotide frequencies for ENCODE target region 2 and possibly ENCODE target region 1). Furthermore, some evolutionary phenomena do not change much at all for any given distance and these phenomena tend to produce low median test statistic values even for very short distances (e.g., tests for rate heterogeneity and transition: transversion biases in the ENCODE regions). In yeasts, all of the phenomena investigated produce median test statistics that increase with distances between windows and then plateau. It is quite likely that some previous attempts to determine the nature of spatial variation in certain evolutionary forces have yielded negative results because there has been too little data, species have been phylogenetically close and / or trends were looked for only over large distances (beyond plateau-distances) (see Webster *et al*., 2003).

Differences in the results between yeasts and mammals are most likely due to the much higher gene density in yeasts and the isochore structure of mammalian genomes. There is a good general agreement in the behaviour of all median test

statistics between the two yeast datasets as distances increase between windows. The two ENCODE datasets produce median test statistics that behave in a similar way for all tests except those where nucleotide frequencies are estimated separately for the two windows in the alternate model (T-W10 to T-W14). It is interesting to note that nucleotide frequencies tend to vary more than evolutionary rates, which vary more than transition: transversion biases and rate heterogeneity (in terms of the effects of accounting for differences between windows on model fit). In Chapter 2 it was noted that differences in nucleotide frequencies and evolutionary rates also vary extensively between codon positions, compared to transition: transversion biases and rate heterogeneity. I explore differences in these factors between genes (instead of codon positions or regions) in Chapter 6.

Investigation of the biological causes of the spatial variation in evolutionary phenomena is beyond the scope of this chapter. Whatever the reasons behind the effects observed in this investigation, it is perhaps unsurprising that regions close to each other evolve more similarly than regions that are further away. Whatever specific factors affect one region seem intuitively more likely to affect a nearby region. Lercher *et al.* (2001) note that mutations induced by recombination are more likely to be similar in regions that are close together than those that are further apart. Other factors, such as replication of DNA by the same ORI or transcription-associated mutations of nucleotides in a transcription block may also cause nearby regions to evolve in a similar fashion (Green *et al*., 2003). Many other factors could be involved that are somewhat reciprocal; for example, a region that evolves in a particular manner may develop a characteristic GC content, which may further result in

windows of DNA in this region appearing to evolve more similarly than windows of DNA from more distant regions.

The results presented in this chapter have important implications for other active areas of research. Clearly, the non-independent evolution of neighbouring genomic regions may affect how we analyse large contiguous blocks of data for large-scale phylogenetic analyses. Furthermore, the differences in the patterns of evolution between different genomic regions may also affect how we analyse a large dataset of sequences from different genomic regions for phylogenetic analyses. I consider this problem in more detail in Chapter 6.

5.7 Future Directions

When more data becomes available this investigation could be pursued further for other species and also for other regions in the genomes of the same species investigated here. To mask the effects of the high gene density in yeast, it would be possible to repeat the entire study but retain only those pair-window analyses when both windows were entirely contained in intergenic regions. It would be of particular interest to analyse more ENCODE regions to determine whether isochores affect the results as I have suggested. I have deliberately avoided consideration of the specific trace of test statistics along a genomic alignment in the chapter but this could be pursued in future.

Chapter 6: Combining Data for Phylogenomic Studies

Contents

6.1 Introduction

This chapter considers how the statistical fit of a model to the evolution of a multiple-gene dataset is affected by model complexity and how using different models can affect the conclusions we draw about the evolution of a dataset. I begin this chapter with a discussion of the effects of the large scale sequencing programs of the genomic era on phylogenetic methods, which have caused an expansion in the field of 'phylogenomics'. I discuss several phylogenomic methods, which harness the data in different ways, and explore questions that can be asked with a class of methods called the super-matrix methods. I use a well-characterised dataset of 106 yeast genes (Rokas *et al*., 2003) to determine the factors that are important to model in phylogenomic studies for accurate inferences. I determine how estimating model parameters independently for different genes affects model fit and could affect the optimal inferred tree topology for the dataset. Using a variety of criteria, I find that complex models are generally favoured with large datasets and that different factors can affect model fit dramatically. I discuss the implications in such studies of the non-proportional branch lengths of different genes in the dataset (heterotachy) and model bias. I also consider how the order of sub-optimal trees is affected by model complexity by introducing the idea of 'settling'. Throughout this chapter I consider the ways in which model choice may affect our conclusions in phylogenomic studies.

6.2 Phylogenomics

Many fields of study entered a new era with the advent of datasets of genomic proportions. We can infer phylogenies and reconstruct the evolutionary histories of different species using more information and one reason to use more information is an improved signal to noise ratio. The evolutionary peculiarities of small datasets may make them prone to supporting incorrect phylogenies whereas large datasets may buffer against the small scale fluctuations in evolutionary processes. The use of very large datasets for the estimation of phylogenies, part of the expanding field of phylogenomics, has spawned new methods of tree-estimation and has raised questions about how large datasets should be handled in order to make accurate phylogeny estimations. Chapter 5 showed that small neighbouring regions of data, which we may wish to combine, may evolve significantly differently even if they are in the same genomic region. The accuracy of phylogenetic studies may depend on how we treat the variation inherent in large datasets.

Using different evolutionary models can make differences to inferred phylogenies (e.g., Phillips *et al*., 2004). It has been shown that when comparing a suite of evolutionary models, all within the likelihood framework, larger differences in model fit are more likely to lead to a different optimal topology (N. Goldman and S. Whelan, personal communication). Thus, it is important to consider the differences in model fit caused by estimating different parameters of evolutionary models (such as evolutionary rates and so on) separately for different units of our larger datasets. This completes a theme that has been explored throughout this thesis: factors that vary

between codon positions and genomic regions have already been studied, and this chapter investigates how these same factors vary between genes.

Within recent years there has been a decisive change in the emphasis of phylogenomic studies from the search for the types of smaller markers that are likely to be particularly informative about the phylogenetic history of species, such as individual genes, to the methods that can be applied to much larger datasets. This change has been fuelled by the advent of large genomic datasets, which mean that classic likelihood methods can be applied on a larger scale, and novel methods that estimate changes in genomic structure (e.g., gene loss and gain, gene order rearrangement, etc.) have been developed. Phylogenomic methods, much like simpler phylogenetic studies, tend to be two-stage processes, involving the identification of homologous characters between different species first and tree reconstruction second. Broadly speaking, there are two categories of phylogenomic methods: sequence-based methods and methods based on whole-genome features. The investigations in this chapter focus on sequence-based methods.

Sequence-based methods are currently the most popular phylogenomic methods, broadly consisting of super-tree and super-matrix methods. The super-tree approach involves combining optimal trees from the separate analyses of individual genes, each of which contains data from partially overlapping sets of species (Delsuc *et al*., 2005). The super-matrix approach, pursued in this chapter with partitioned likelihood models, involves the analysis of large datasets of concatenated individual genes (e.g. Bapteste *et al.*, 2002). Any gene missing for a given species is considered as missing data. Interestingly, when super-matrix datasets are large, the optimal

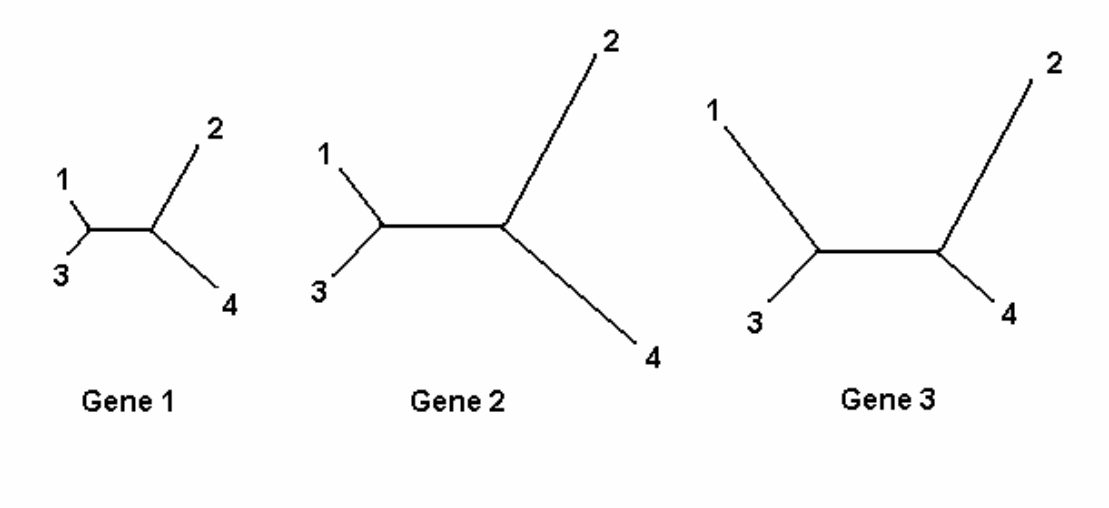
estimated phylogeny tends not to be sensitive to large amounts of missing data for some species (Philippe *et al*., 2004).

Gene order methods generally consider the minimum number of breakpoints that would have had to occur between genomes (and reconstruct an ancestral gene order) or, more simply, score the presence or absence of consecutive pairs of orthologous genes between genomes (Delsuc *et al*., 2005). Gene content methods generally reconstruct phylogenies based on the distances between different genomes, which are in turn based on the genes that are shared or absent between species (Dopazo *et al*., 2004; Gu and Zhang, 2004; Huson and Steel, 2004). Rare genomic changes (such as retrotransposition events and gene fusion or fission events) are more difficult to use in a quantitative manner because they are essentially morphological changes, albeit at a very small scale; they can be scored in terms of presence or absence between genomes (Delsuc *et al*., 2005). I do not consider these methods further because they are beyond the scope of the research interests of this chapter; they are however, exciting and active areas of research that use large amounts of data for phylogenetic purposes and are reasonable alternative approaches to the super-matrix methods that are explored in more detail in this chapter.

Figure 6.1 illustrates the issue of phylogenies with the same topology but non-proportional branch lengths. This phenomenon is termed ‘heterotachy’, meaning ‘different speeds’, and occurs when there are changes in site-specific evolutionary rates. Where heterotachy occurs and is not considered in modelling then the assumption that the data is drawn from an identical distribution is violated; likelihood methods may then suffer from systematic biases and may mis-estimate the phylogeny

of the dataset (Kolaczkowski and Thornton, 2004). Indeed, for a multiple-gene dataset, even when the topologies of the phylogenies of the different genes are identical, the estimated phylogeny may have a topology different to that of the genes. Although heterotachy between genes can be addressed fairly trivially by modelling the evolution of the different genes separately (and then combining likelihood scores) it may be harder to identify sites within genes that have evolved so differently, since we have few or no *a priori* clues as to which sites differ from which other sites.

Figure 6.1 – Phylogenies that share the same overall topology but have either proportional or non-proportional branch lengths relative to the central tree. The phylogeny for gene 2 has all branch lengths proportional to gene 1 (all branch lengths are twice as long in gene 2). Gene 3 has non-proportional branch lengths compared to genes 1 or 2, i.e., there is no single scaling factor by which all branches of gene 1 can be multiplied to attain the phylogeny of gene 3. The branching order (topology) is the same for all three genes.



Recently, there has been research that demonstrates that heterotachy may be less important than originally thought and the systematic biases in tree estimation may not be so severe (e.g., Gaucher and Miyamoto, 2005; Steel, 2005). Indeed, if the taxa that have long branches are different between genes in a larger multi-gene dataset, the effect of such branches on phylogenetic resolution can actually be minimised in analyses using the multi-gene dataset (Gontcharov *et al*., 2004). The extent to which heterotachy may exist between genes is not fully understood and I discuss its relevance to the Rokas *et al*. (2003) dataset, presented later, in terms of comparing models that estimate a scaling factor for the same tree topology between genes and models that allow non-proportional branch lengths between the trees for different genes. I show that the separate phylogenies of the 106 yeast genes have significantly non-proportional branch lengths between the genes but this does not affect estimates of the shape of the optimal phylogeny for the concatenated dataset.

When many genes are combined into a single dataset for a super-matrix analysis there are considerable issues regarding model complexity and the phylogenies that are supported by the data. One must consider how the evolution of the separate units of the dataset (e.g., individual gene multiple alignments) are modelled and how the units of data are combined into a larger dataset. With a large number of species and / or a large amount of data it may not be practical to perform a complete analysis of tree space and we may use heuristic tree searches. The criterion on which the optimal topology is selected is simple using likelihood methods although it is more challenging to select the most appropriate model to analyse the data with and to decide whether the data should be analysed as nucleotides or amino acids, both of which have been used in phylogenomic studies previously (e.g., Pupko *et al*., 2002;

Rokas *et al*., 2003). Later in this chapter I discuss the evolutionary factors that may vary between the genes and various means to decide which factors are most important in evolutionary modelling, including the Akaike Information Criterion (AIC) (Posada and Crandall, 2001).

Pupko *et al*. (2002) analysed two large nuclear datasets and one large mitochondrial dataset of amino acid sequences and suggested ways in which large amino acid datasets should be considered in future. Within each dataset, each consisting of sequences from at least 28 species, genes were combined into one large dataset and analysed with three models. In the "concatenated model" all genes were assumed to have the same branch lengths; in the "proportional model" branch lengths of trees were assumed to be proportional between genes; in the "separate model" the branch lengths for each gene are assumed to be non-proportional. Furthermore, among-site rate variation was either modelled using one γ distribution for all genes (the 'homogeneous' model) or using one γ distribution for each gene (the 'heterogeneous' model). Other model parameters, such as amino acid frequencies, were assumed to be the same across all genes. Pupko *et al*. (2002) concluded that one using the heterogeneous model represents the most appropriate model for all three datasets to describe among-site rate variation. Furthermore, using the "separate model" or "proportional model" was found to be the most appropriate for the branch length analyses of each dataset (for one and two datasets respectively). Additionally, whether the "concatenated model", "proportional model" or "separate model" is used to analyse the datasets can affect which tree is chosen as the optimal phylogeny.

Whilst it might be methodologically appropriate to analyse large gene datasets using codon models this has not been pursued because searching tree space adequately using such complex models is computationally expensive (Ren *et al.*, 2005). Very large gene datasets have been analysed largely as nucleotides (Rokas *et al*., 2003), which I pursue here, but future investigations may use the types of models that allow for differences in evolution between the different codon positions (Chapters 2 and 3).

As super-matrix studies become increasingly more popular there remain important questions about how to handle large datasets. Clearly, if large differences in model fit are more likely to lead to differences in the optimal tree topology (S. Whelan and N. Goldman, personal communication), the factors that cause the most change should be assessed quantitatively. Additionally, it will be important to observe how the order in which the models support different phylogenies changes with model complexity. Various phylogenetic approaches, including studies in the Bayesian framework, may generate confidence sets of trees. The confidence set of trees may change with our modelling approaches. Finally, even if different models choose the same tree topology as optimal, using different models may affect our confidence in this topology. Yang *et al*. (1995b) noted that overly simple models tend to give inflated levels of confidence in a given tree topology being optimal (e.g., too small confidence sets or too low measures of variance). These phenomena are addressed using different models to analyse the Rokas *et al*. (2003) dataset.

6.3 Model Fit with a Large Phylogenomic Dataset

Rokas *et al.* (2003) used a phylogenomic yeast dataset that has become a standard dataset for subsequent phylogenomic methodology studies. The dataset consists of 106 orthologous genes that are distributed across all 16 chromosomes in the *Saccharomyces cerevisiae* genome, with sequences for each gene also available for seven other yeast species (*S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, *S. kluyveri* and *Candida albicans*). Rokas *et al*. (2003) performed phylogenetic reconstructions using maximum likelihood analyses on the nucleotide data, and maximum parsimony on both the nucleotide and amino acid data. For the maximum likelihood analyses, Rokas *et al*. (2003) used a model no more complex than REV with a one γ distribution for all genes (plus an allowance for invariant sites, those that are identical at a given multiple alignment position across all eight species). A total of 24 different tree topologies were found to be optimal for the analysis of the 106 separate genes and these form the 'tree set' further considered in that study and this thesis. The tree set consists of a fair number of topologies that can be compared for the entire dataset since it is not practical to compare all possible tree topologies.

In contrast to the studies of individual genes, the concatenated dataset produced the same inferred optimal tree for all three methods (ML on nucleotide data and MP on nucleotide and amino acid data) with 100% bootstrap support for each node. This tree topology is shown in Figure 6.2; the complete set of tree topologies used in this investigation is shown in Figure 6.3.

Figure 6.2 – The inferred optimal phylogeny for the Rokas *et al*. (2003) dataset.

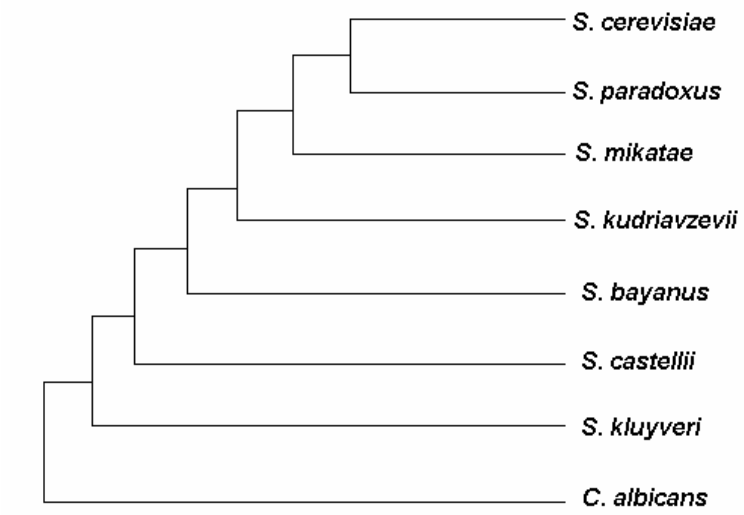
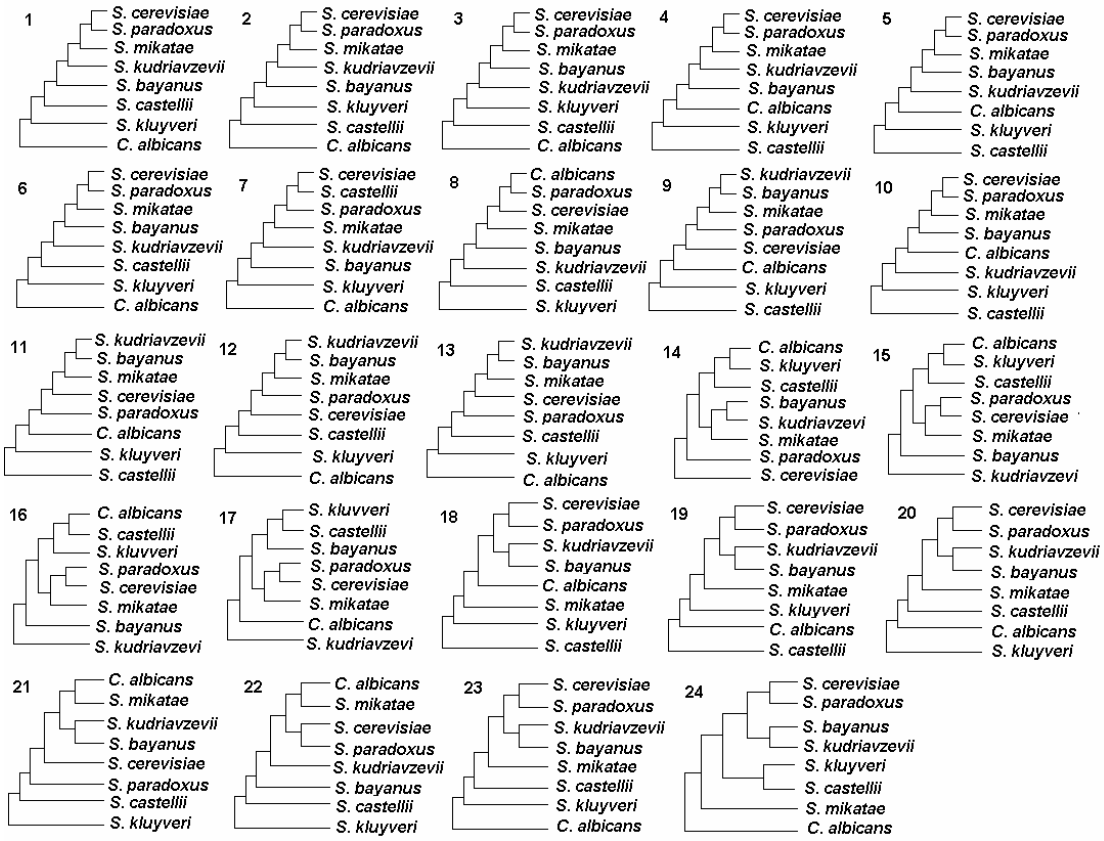


Figure 6.3 – The complete set of tree topologies used in this investigation, which is the 'tree set' as defined by Rokas *et al*. (2003). Tree 1 in this figure is the same as Figure 6.2, and is the inferred optimal tree identified by Rokas *et al*. (2003) and by all of the models used in this investigation.



Rokas *et al*. (2003) concluded that using large datasets may improve the resolution of a phylogenetic signal that is difficult to detect using small datasets, and that their studies may have important implications for the resolution of the 'tree of life'. Even if small datasets could not be used to resolve the 'tree of life', much larger datasets may be used to do so with their improved signal to noise ratio. However, the resolution of the inferred phylogeny for this dataset was not too difficult compared to other phylogenetic problems that may have smaller internal branch lengths, such as the 'tree of life'. Furthermore, the 'tree of life' contains events such as genome fusions and horizontal gene transfer (Rivera and Lake, 2004), which cannot be adequately displayed in a simple tree-like phylogeny. Other factors, such as non-stationarity of the substitutional processes, will also add to the complexity of inferring distant relationships (Yang and Roberts, 1995).

Subsequent studies using the Rokas *et al*. (2003) dataset have shown that, depending on the method used, other tree topologies may indeed be optimal and retain 100% bootstrap support for each node of the tree when using the entire concatenated dataset (e.g., Phillips *et al*., 2004, using a minimum evolution method). I revisit the Rokas *et al*. (2003) dataset and investigate how estimating model parameters separately for the different genes in the concatenated dataset affects the fit of the model to the dataset (judged by the likelihood values obtained). The inferred optimal tree is consistent regardless of the likelihood model that I implement in this chapter (see Figure 6.2), and is not the same as the tree found as optimal by Phillips *et al*. (2004) because, by the authors' observation, the method they used to determine an optimal phylogeny is biased. However, the order in which sub-optimal trees are favoured in my analyses varies between models. As the models get more complex and

capture more information about the evolution of the data, the disagreement about the ordering of the sub-optimal topologies tends to decrease. This phenomenon, which I term 'settling', reflects the extent to which we think we have described the important evolutionary features within the dataset; this is justified later.

Whilst Rokas *et al.*'s (2003) study employed the REV model for analysis, I choose to employ both the HKY and REV models of analysis. One would usually favour the more complex model (REV, which has five exchangeabilities instead of a single transition: transversion bias parameter in the HKY model) but I also use the HKY model because it has been used previously in Chapters 2, 3 and 5; this means that the results of this investigation will be easier to compare to other results presented in this thesis than if the REV model is used alone. I use the REV model of analysis because, considering the amount of data, it does provide a significant improvement in model fit over the HKY model. The conclusions drawn by this investigation are shown not to depend greatly on whether HKY or REV is used as a basis for most models.

6.4 Methods of Model Fit Study

The models that are used in this investigation to analyse the 106-gene concatenated Rokas *et al*. (2003) dataset are presented in Table 6.1. An unrooted tree topology for n species has $2n - 3$ branches (Felsenstein, 2004) and thus the unrooted phylogeny of the eight yeast species had 13 branch length parameters, which form part of the model parameters. Note that each model in Table 6.1 was applied to the 106-gene concatenated dataset for each of the 24 trees in the tree set (see Figure 6.3).

Table 6.1 – Models used to analyse the Rokas *et al*. (2003) dataset of 106 genes for eight yeast species. Notation G means that a single γ distribution was used to model rate variation across all genes; notation 106G means that a separate γ distribution was used to model rate variation for each of the 106 genes. Notations R, T and N (rate, transition: transversion bias and nucleotide frequencies respectively) mean that separate parameter estimates for these factors were made for each gene. The parts of the table that concern models with non-proportional branch lengths are shaded in grey. The AIC scores for each of the models are also shown (see text for details).

Model	Number of Model Parameters	AIC Score	Model	Number of Model Parameters	AIC Score
HKY	17	1418851	REV	21	1413906
HKY+G	18	1369629	REV+G	22	1365778
HKY+R	122	1412003	REV+R	126	1406970
HKY+R+T	227	1411605	REV+R+T	651	1405810
HKY+R+N	437	1411374	REV+R+N	441	1406319
HKY+R+N+T	542	1410989	REV+R+N+T	966	1404895
HKY+R+G	123	1366715	REV+R+G	127	1362666
HKY+R+T+G	228	1366323	REV+R+T+G	652	1361885
HKY+R+N+G	438	1366098	REV+R+N+G	442	1361978
HKY+R+N+T+G	543	1365691	REV+R+N+T+G	967	1361028
HKY+R+106G	228	1365050	REV+R+106G	232	1361138
HKY+R+T+106G	333	1364247	REV+R+T+106G	757	1360043
HKY+R+N+106G	543	1364370	REV+R+N+106G	547	1360393
HKY+R+N+T+106G	648	1363570	REV+R+N+T+106G	1072	1359159
Separate ‘HKY’ for each gene	1802	1409658	Separate ‘REV’ for each gene	2226	1403606
Separate ‘HKY+G’ for each gene	1908	1361824	Separate ‘REV+G’ for each gene	2332	1357462

All of the tests between nested models in Table 6.1 have been performed and in every case the more complex model was favoured using the χ^2 approximation to the LRT (results not shown). However, this investigation is not primarily concerned with which model is best but how the choice of models affects our inferences. Of course, it is still of some interest to confirm that very complex models do best using large amounts of data, which further confirms that whilst our most complex models may be performing well, the more simple models are not fully describing the evolutionary process.

In order to compare different models that may not be nested and to assess the relative importance of different factors introduced at each stage of increasing model complexity, I use two techniques that can be more widely applied than the LRT between nested models: the Akaike Information Criterion (AIC) and the likelihood improvement in nested models per extra degree of freedom. The AIC is a technique that can be used to identify the most appropriate model amongst a set of models, regardless of whether they are nested or not (Posada and Crandall, 2001).The formula defining the AIC is as follows:

$$AIC = -2l(\theta_{est}) + 2K$$

where $l(\theta_{est})$ is the log likelihood given estimated model parameters, θ_{est}, and K is the number of parameters in that model. The AIC is calculated at the maximum likelihood ($\max_\theta l$), which is achieved at the maximum likelihood parameter values ($\theta_{est} = \text{argmax}_\theta\, l$). The model that provides the lowest AIC score is the optimal model under AIC criteria.

The log-likelihood improvement per degree of freedom (added to a model) gives a measure of how much different biological factors are contributing relative to each other (whereas the AIC is an overall model selection criterion). Every degree of freedom that we add to a model requires an additional parameter to be estimated from the data and I consider, per degree of freedom, which evolutionary factors are most important to estimate as distinct parameters between genes. I then observe how modelling different factors affects the order in which different topologies are favoured. These methods may be of use in determining whether our models may be biased or whether they have failed to capture essential features of the evolutionary process.

6.5 Results and Conclusions of Model Fit Study

The same tree topology is optimal for all models I apply to the 106-gene concatenated Rokas *et al*. (2003) dataset (see Figure 6.2). The fact that the optimal tree is the same between models that do and do not allow non-proportional branch lengths between the separate genes suggests that heterotachy has not greatly affected our ability to estimate the tree for this dataset.

The AIC results for the analysis of the Rokas *et al*. (2003) dataset are presented in Table 6.1. Under the AIC, the best model is a separate REV+G model for each gene. This suggests that there may be considerable heterotachy and that using proportional trees between genes is a sub-optimal modelling method (although it has not affected the optimal tree topology in this case). I now consider the relative modelling importance of different biological effects. Starting with the basic HKY or

REV model, the single factor that provides the greatest change in the AIC score is a single γ distribution, accounting for rate heterogeneity across sites. It is then difficult to decide how to proceed through a series of more complex models using the AIC because it becomes a challenge to distinguish between the different factors that can be added to the +R models and the +R model itself. Thus, I consider a different technique to address this question further. If we consider the likelihood improvement from the most basic model (HKY or REV) per extra degree of freedom added to the model, the order in which different factors should be added to an HKY or REV model are shown in Table 6.2 and Table 6.3 respectively.

Table 6.2 – The order in which parameters are added to the basic HKY model based on the per-degree of freedom additional log likelihood improvement in model fit.

Rank	Model Comparison	Likelihood Improvement per Parameter Added	Description of Extra Parameters
1	HKY → HKY+G	24611.85	γ-distributed rates over sites (one distribution for all genes)
2	HKY+G → HKY+R+G	14.88	Different mean rate for each gene
3	HKY+R+G → HKY+R+106G	8.93	γ-distributed rates over sites (one distribution per gene)
4	HKY+R+106G → HKY+R+T+106G	4.82	Different transition: transversion bias for each gene
5	HKY+R+T+106G → HKY+R+T+N+106G	2.08	Different nucleotide frequencies for each gene
6	HKY+R+T+N+106G → separate 'HKY+G' for each gene	1.69	Non-proportional branch lengths for each gene

Table 6.3 – The order in which parameters are added to the basic REV model based on the per-degree of freedom additional log likelihood improvement in model fit.

Rank	Model Comparison	Likelihood Improvement per Parameter Added	Description of Extra Parameters
1	REV → REV+G	24065.01	γ-distributed rates over sites (one distribution for all genes)
2	REV+G → REV+R+G	15.82	Different mean rate for each gene
3	REV+R+G → REV+R+106G	8.28	γ-distributed rates over sites (one distribution per gene)
4	REV+R+106G → REV+R+N+106G	2.18	Different nucleotide frequencies for each gene
5	REV+R+N+106G → REV+R+T+N+106G	2.18	Different exchangeabilities for each gene
6	REV+R+T+N+106G → separate 'REV+G' for each gene	1.67	Non-proportional branch lengths for each gene

Comparing Tables 6.2 and 6.3, the order in which different factors are most important to add to the basic model (under the per degree of freedom likelihood increase approach) is not the same for HKY and REV analyses. For the REV analyses the nucleotide frequencies provide a greater per degree of freedom log likelihood increase than the exchangeability parameters, whereas the single transition: transversion bias parameter in the HKY model provides a greater per degree of freedom log likelihood increase than considering the nucleotide frequency differences between genes. The difference in order is due to the difference in the number of exchangeability parameters between the HKY and REV models (one and five parameters respectively). In all cases of the REV models, models that consider

nucleotide frequency differences between genes provide a better explanation of the evolution of the data than the models that consider exchangeability differences between genes (REV+R+N vs. REV+R+T, REV+R+N+G vs. REV+R+T+G, REV+R+N+106G vs. REV+R+T+106G). However, for the HKY models, all models that consider exchangeability parameter differences between genes provide a better explanation of the evolution of the data than the models that consider nucleotide frequency differences between genes (HKY+R+N vs. HKY+R+T, HKY+R+N+G vs. HKY+R+T+G and HKY+R+N+106G vs. HKY+R+T+106G). Having a single transition: transversion bias parameter is clearly important (as in the HKY models); the addition of subsequent exchangeability parameters is less important, although still statistically significant.

Unlike in Chapter 2, where nucleotide frequency differences were more important to consider than transition: transversion bias differences between codon positions, and Chapter 5, where nucleotide frequencies varied more over large alignments than transition: transversion biases, modelling differences between the transition: transversion bias between *genes* may be more important than nucleotide frequency differences. The AIC scores suggest that for REV-based models, it is more important to consider exchangeability differences between the genes than nucleotide frequency differences between genes in all relevant model comparisons (see Table 6.1). AIC scores for HKY-based models are not consistent: in some cases it is more important to consider nucleotide frequency differences between genes but in one case (for the most complex models HKY+R+T+106G vs. HKY+R+N+106G) it is more important to consider exchangeability differences between genes before nucleotide frequency differences. We are considering the evolution of the yeast genes in this

chapter as strings of nucleotides and we are not considering codon position effects; differences in constraint of different genes may affect the average transition: transversion biases for each gene and the overall nucleotide frequency differences may be less affected by differences in constraint between genes; this would require further investigation.

It is not perhaps surprising that for both HKY and REV models, the most important factors to consider are overall rate heterogeneity and differences in evolutionary rate between genes. The fact that we consider each gene without respect for different codon positions within each gene makes it less surprising that we find it is important to consider differences in rate heterogeneity between each gene. Interactions are observed between parameter estimates as in previous studies in this thesis; the log likelihood differences caused by adding a certain parameter to a model depend on the other factors that are already considered in that model. The use of sub-optimal tree topologies does not change the order in which either HKY or REV model factors (rate heterogeneity, rate differences between genes, etc) should be added using the per degree of freedom log likelihood increase approach. This may be important in future studies if one is uncertain if the topology being used is the correct topology.

In this investigation, all of the factors that have been considered in the models are significant; future phylogenomic analyses should use the most complex models available to analyse large datasets. The use of the most complex models available is currently not common practice. The datasets used by Pupko *et al*. (2002) did not always favour the most complex model but were considerably smaller than the Rokas *et al*. (2003) dataset. With respect to tree inference, which remains the main aim of

most current phylogenomic studies, the choice of model that I used did not make a difference to the topology of the optimal tree, although other studies have shown that different models can select different topologies as optimal (Phillips *et al*., 2004). Rokas *et al*. (2003) suggest that with enough data phylogenetic problems can be solved. In order to test this fully one must consider the confidence in the results, which is discussed in the next section.

6.6 Phylogenetic Settling

The simplest way to assess confidence in the results is simple bootstrapping, which Rokas *et al*. (2003) have already performed. They found that the phylogeny shown in Figure 6.2 was the optimal phylogeny for all 100 pseudo-replicates of the concatenated dataset produced. It is known that using simple models tends to lead to over-confidence in bootstrap results (Yang *et al*., 1995b). However, when I repeated the bootstrapping procedure using the most complex model (a separate 'REV+G' model for each of the 106 genes), the phylogeny in Figure 6.2 remained the optimal phylogeny for all 100 pseudo-replicates (data not shown). This suggests that, unless one uses a model that is known to be biased, such as in Phillips *et al*. (2004), the Rokas *et al*. (2003) dataset has a clear phylogenetic signal supporting the phylogeny presented in Figure 6.2, which is not necessarily supported by all individual genes. However, other very large datasets, where the evolutionary history of the species may have involved rapid radiation for example, may fail to support a single phylogeny with 100% bootstrap support.

In this investigation it was noticed that there is conflicting support for the order in which the 23 sub-optimal tree topologies are favoured, depending on which evolutionary model one uses in the analysis. As the complexity of the models increases, there appears to be a tendency for the order of the phylogenies to 'settle'. Each successive change to the complexity of the model appears to capture less additional information than previous models and there is less change in the order of the sub-optimal topologies; I term this 'phylogenetic settling'. In difficult phylogenetic problems the optimal topology may not settle with 100% bootstrap support for even the largest datasets available. I present the order in which different tree topologies are supported for the HKY and REV models in Figures 6.4 and 6.5; each tree topology is coded for by a single colour; the actual topology is unimportant. The important point to note is the differences in the order of colours that can be seen for different models. Note that a settled order of tree topologies means that the model has converged in terms of its assessment of the order of optimal and sub-optimal topologies, which does not necessarily mean that the model (or the optimal tree) is correct.

Figure 6.4 – Phylogenetic settling of the HKY-based models. The models are ordered according to their AIC score from highest to lowest (see Table 6.1). The most complex model used in this investigation (a separate 'REV+G' for each gene) has the tree order against which all other tree orders are compared (see Figure 6.5). The rank order of trees is shown in the first column.

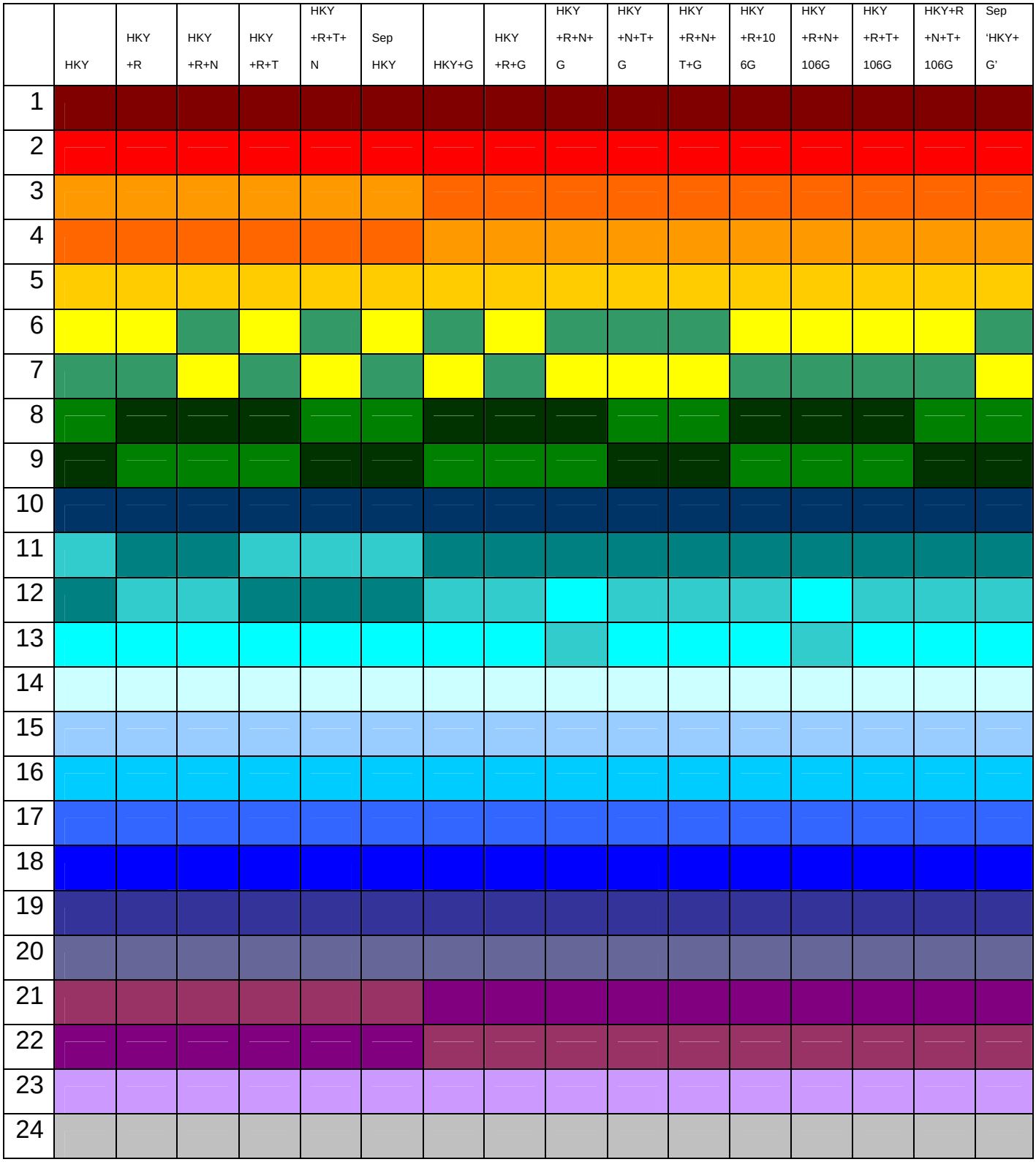
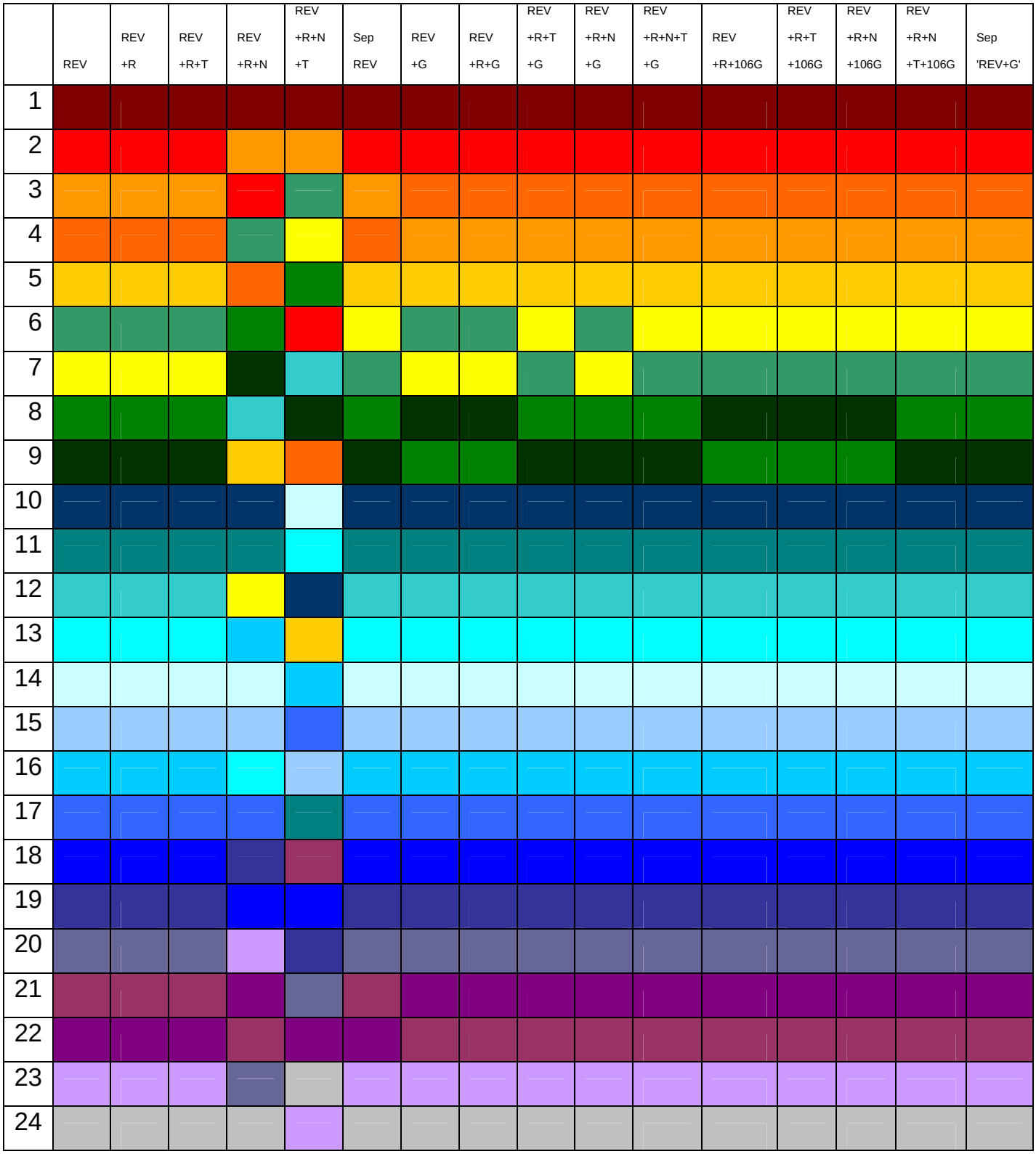


Figure 6.5 – Phylogenetic settling of the REV-based models. The models are ordered according to their AIC score from highest to lowest (see Table 6.1). The most complex model used in this investigation (a separate 'REV+G' for each gene) has the tree order against which all other tree orders are compared. The rank order of trees is shown in the first column.



Comparing Figures 6.4 and 6.5, it becomes apparent that models similar apart from the fact that they have either HKY or REV as a base do not necessarily agree on the same order of phylogenies. Models REV+R+N and REV+R+N+T in Figure 6.5 are particularly 'unsettled' and the rank orders of phylogenies are markedly different from the final order obtained using a separate REV+G model for each gene. We also see the effects of the modelling of rate heterogeneity on phylogenetic settling. As the REV models incorporate a γ distribution (Figure 6.5, to right starting from model REV+G), the top five tree topologies are the same for each more complex model, whereas the non-γ distribution models only agree on the top phylogeny. In Figure 6.4 the HKY models that incorporate a γ distribution also share the same top five phylogenies; the non-γ distribution HKY models all share the same top five phylogenies but the order differs from the models with a γ distribution. This demonstrates how failing to account for certain aspects important in the evolution of a dataset could bias the results. Even if the top tree topology does not change across the models used in this investigation it is conceivable that trees that are not significantly different from the top tree specified by some criterion, a 'confidence set', could change depending on the model. It is also easy to imagine another case where the top tree does change across different models. Note that confidence sets of trees are used in analyses in the Bayesian framework, for example, and changing the confidence set of trees could affect results.

It has now been established that the rank order of phylogenies in the tree set depends on the model used to analyse the data. Over 100 bootstrap replicates for any model choice, the rank ordering of the 24 trees was not consistent (data not shown). Thus, even though the dataset is very large, it is not large enough to resolve the rank

order of tree topologies for any given model. The rank order of phylogenies is not consistent between models or within all bootstrap pseudo-replicates for a single model, amongst those models used in this investigation.

6.7 Discussion

I began this chapter with a brief review of phylogenomic methods and discussed challenges that super-matrix methods face; these methods are likely to remain the most prevalent of phylogenomic methods for some time to come. Considering differences between the evolutionary dynamics of the units in a super-matrix, i.e., the genes in the Rokas *et al*. (2003) dataset, which were not addressed in the original study can make a considerable difference to model fit (and the parameter estimates, such as branch lengths) in a study. Although all of the models used in this study supported the same optimal topology, other studies have shown that using a different model may affect the optimal topology even for this dataset (Phillips *et al*., 2004). I have demonstrated how using different models (e.g., HKY vs. REV) can affect the order of importance of the factors that we consider most important to model and estimate separately for each super-matrix unit. I have investigated the properties of models that make the most difference to model fit when analysing a large dataset for phylogenomic purposes. Even relatively minor effects are significant with a large dataset and I suggest, based on these results, that more complex models should be used in phylogenomic studies in future; this is not currently common practice. Furthermore, I have discussed the differences between genes when analysed as nucleotide data and suggested why the results are different from those that studied differences between codon positions or genomic regions.

The results obtained in the study of phylogenetic settling have demonstrated that consistency within the data can be investigated beyond the single optimal tree topology, which can affect the confidence set of trees used in certain types of study. These results suggest it may be premature to assume that we may be able to resolve difficult phylogenetic problems simply by having large datasets. Clearly, as Phillips *et al*. (2004) have previously suggested, the choice of model in a phylogenomic study is paramount and inappropriate models can introduce bias. I have shown how similar but different models could affect the results of a phylogenomic study, and introduced a new way in which one can assess model bias. A good example of this was that all HKY models that incorporate a γ distribution share the same top five phylogenies as each other and the non-γ distribution HKY models all share the same top five phylogenies as each other, but the order differs between the γ distribution and non-γ distribution models (Figure 6.4).

Interestingly, the most complex model is favoured by the AIC, which suggests that significant heterotachy may exist in this dataset and serves as a warning that super-matrix analyses may be prone to any biases in phylogeny estimation that this may cause. However, the fact that this heterotachy clearly exists within the data suggests that it may not be the gross problem for phylogenetic methods suggested by Kolaczkowski and Thornton (2004) because the models with and without non-proportional branch lengths give the same optimal phylogeny (Figure 6.2). The results of this nucleotide-type analysis are broadly consistent with the amino acid datasets analysis results of Pupko *et al*. (2002) in the sense that the more complex models tend to be favoured.

6.8 Future Directions

Most importantly, this type of study could be applied to more datasets, especially those where the phylogeny is more difficult to resolve. The base of the mammalian phylogeny is notoriously difficult to resolve due to rapid radiation of lineages, and the publication of more genomes will make it possible to assemble a large dataset of aligned genes for mammals. One might expect different trees to be optimal depending on the models used in such an analysis and, for this notoriously difficult problem, one may not observe such high bootstrap support as for the Rokas *et al*. (2003) dataset. Furthermore, it would be interesting to analyse phylogenomic datasets making allowances for differences in evolutionary properties of the different codon positions (see Chapter 2). This is advisable since it captures more evolutionary information than nucleotide models but there are not so many parameters as we must estimate in a codon model, which carries a large computational burden. This would address whether the differences in constraint of different genes affects the average transition: transversion biases or nucleotide frequency parameters more in a nucleotide-level analysis. Finally, it would be interesting to determine how often, for real multiple-gene datasets, models that allow non-proportional branch lengths and models that allow only proportional branch lengths between the phylogenies of aligned genes infer different phylogenies as optimal. This would assess how large a problem heterotachy is likely to be when making tree inferences with large datasets in future studies.

References

Adams M, Celniker S, Holt R, Evans C, Gocayne J, Amanatides P, Scherer S, Li P, Hoskins R, Galle R, George R, Lewis S, Richards S, Ashburner M, Henderson S, Sutton G, Wortman J, Yandell M, Zhang O, Chen L, Brandon R, Rogers Y, Blazej R, Champe M, Pfeiffer B, Wan K, Doyle C, Baxter E, Helt G, Nelson C, Miklos G, Abril J, Agbayani A, An H, Andrews-Pfannkoch C, Baldwin D, Ballew R, Basu A, Baxendale J, Bayraktaroglu L, Beasley E, Beeson K, Benos P, Berman B, Bhandari D, Bolshakov S, Borkova D, Botchan M, Bouck J, Brokstein P, Brottier P, Burtis K, Busam D, Butler H, Cadieu E, Center A, Chandra I, Cherry J, Cawley S, Dahlke C, Davenport L, Davies P, de Pablos B, Delcher A, Deng Z, Mays A, Dew I, Dietz S, Dodson K, Doup L, Downes M, Dugan-Rocha S, Dunkov B, Dunn P, Durbin K, Evangelista C, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian A, Garg N, Gelbart W, Glasser K, Glodek A, Gong F, Gorrell J, Gu Z, Guan P, Harris M, Harris N, Harvey D, Heiman T, Hernandez J, Houck J, Hostin D, Houston K, Howland T, Wei M, Ibegwam C, Jalali M, Kalush F, Karpen G, Ke Z, Kennison J, Ketchum K, Kimmel B, Kodira C, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky A, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh T, McLeod M, McPherson D, Merkulov G, Milshina N, Mobarry C, Morris J, Moshrefi A, Mount S, Moy M, Murphy B, Murphy L, Muzny D, Nelson D, Nelson D, Nelson K, Nixon K, Nusskern D, Pacleb J, Palazzolo M, Pittman G, Pan S, Pollard J, Puri V, Reese M, Reinert K, Remington K, Saunders R, Scheeler F, Shen H, Shue B, Sidén-Kiamos I, Simpson M, Skupski M, Smith T, Spier E, Spradling A, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang A, Wang X, Wang Z, Wassarman D, Weinstock G, Weissenbach J, Williams S, Woodage T, Worley K, Wu D, Yang S, Yao Q, Ye J, Yeh R, Zaveri J, Zhan M, Zhang G, Zhao O, Zheng L, Zheng X, Zhong F, Zhong W, Zhou X, Zhu S, Zhu X, Smith H, Gibbs R, Myers E, Rubin G and Venter C, 2000. The genome sequence of *Drosophila melanogaster*. Science 287: 2185-95

Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P, 2002. *Molecular Biology of the Cell*, 4th Edition. Garland Science Press, New York

The Arabidopsis Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815

Audic S and Claverie JM, 1998. Self-identification of protein-coding regions in microbial genomes. Proceedings of the National Academy of Sciences USA, 95: 10026-10031

Bapteste E, Brinkmann H, Lee J, Moore D, Sensen C, Gordon P, Durufle L, Gaasterland T, Lopez P, Muller M and Philippe H, 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proceedings of the National Academy of Sciences USA 99: 1414-1419

Bateman A, Coin L, Durbin R, Finn R, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E, Studholme D, Yeats C and Sean R. Eddy, 2004. The Pfam protein families database. Nucleic Acids Research, Database Issue 32:D138-D141

Belle E, Duret L, Galtier N and Eyre-Walker A, 2004. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. Journal of Molecular Evolution, online publication 10.1007/s00239-004-2487-x

Boffelli D, McAuliffe J, Ovcharenko D, Lewis K, Ovcharenko I, Pachter L and Rubin E, 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391-1394

Birney E, Andrews D, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez X, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp G, Meidl P, Mongin E, Pettett R, Potter R, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Cameron G, Durbin R, Cox A, Hubbard T and Clamp M, 2004. An overview of Ensembl. Genome Research 14: 925-928

Breier A, Chatterji S and Cozzarelli N, 2004. Prediction of *Saccharomyces cerevisiae* replication origins. Genome Biology 2004, 5:R22

Brown W, Prager E, Wang A and Wilson A, 1982. Mitochondrial sequences of primates: tempo and mode of evolution. Journal of Molecular Evolution 18: 225-239

Cao Y, Adachi J, Janke A, Pääbo S and Hasegawa M, 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. Journal of Molecular Evolution 39:519-527

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer R and Botstein D, 1997. Genetic and physical maps of *Saccharomyces cerevisiae*. Nature 387 (6632 Supplement): 67-73

Cox D and Miller H, 1977. *The Theory of Stochastic Processes*. Chapman and Hall, New York

Cox D, 1962. Further results on tests of separate families of hypotheses. Journal of the Royal Statistical Society B 24: 406-424

Cox D, 1961. Tests of separate families of hypotheses. Proceedings of the 4th Berkeley Symposium 1: 105-123. University of California Press

David L, Huber W, Granovskaia M, Palm C, Toedling J, Bofkin L, Jones T, Davis R and Steinmetz L, 2005. High resolution transcriptional architecture of yeast. Manuscript in preparation

Dayhoff M, Schwartz R and Orcutt B, 1978. A model of evolutionary change in proteins (345-352 in Atlas of Protein Sequence Structure, vol. 5, National Biomedical Research Foundation, Washington DC, edited by Dayhoff M)

Delsuc F, Brinkmann H and Philippe H, 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6: 361-375

Dermitzakis E, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C and Antonarakis S, 2003. Evolutionary Discrimination of Mammalian Conserved Non-Genic Sequences (CNGs). Science 302: 1033-1035

Doolittle W, 1999. Lateral genomics. Trends in Cell Biology 9:M5-8

Dopazo H, Santoyo J and Dopazo J, 2004. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. Bioinformatics 20 (supplement 1): i116-i121

Edwards A, 1992. *Likelihood, Extended Edition*. John Hopkins University Press, Baltimore

Efron B and Tibshirani R, 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York

ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640

Eyre-Walker A and Hurst L, 2001. The evolution of isochores. Nature Reviews Genetics 2: 549-555
Doolittle W, 1999. Phylogenetic classification and the universal tree. Science 284: 2124-2129

Felsenstein J, 2004. *Inferring Phylogenies*. Sinauer Associates Inc, Sunderland MA

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

Fisher R, 1925. Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22:700-725

Fisher R, 1921. On the 'Probable Error' of a coefficient of correlation deduced from a small sample. Metron 1: 3-32

Francino M and Ochman H, 2001. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Molecular Biology and Evolution 18: 1147-1150

Francino M and Ochman H., 2000. Strand symmetry around the β-globin origin of replication in primates. Molecular Biology and Evolution 17: 416-422

Freeland S, Knight R, Landweber L and Hurst L, 2000. Early fixation of an optimal genetic code. Molecular Biology and Evolution 17: 511-518

Freeland S and Hurst L, 1998. The genetic code is one in a million. Journal of Molecular Evolution 47: 238-248

Gaucher E and Miyamoto M, 2005. A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Molecular Phylogenetics and Evolution 37: 928-931

Gardiner-Garden M and Frommer M, 1987. CpG islands in vertebrate genomes. Journal of Molecular Evolution 196: 261-282

Goldman N, 1993. Statistical tests of models of DNA substitution. Journal of Molecular Evolution 37: 650-661

Goldman N, Thorne J and Jones D, 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analysis. Journal of Molecular Biology 263: 196-208

Goldman N and Whelan S, 2000. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Molecular Biology and Evolution 17: 975-978

Goldman N and Whelan S, 2002. A novel use of equilibrium frequencies in models of sequence evolution. Molecular Biology Evolution 19: 1821-1831

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

Gontcharov A, Marin B and Melkonian M, 2004. Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and *rbc*L sequence comparisons in the Zygnematophycaea (Streptophyta). Molecular Biology and Evolution 21: 612-624

Grassly N and Holmes E, 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Molecular Biology and Evolution 14: 239-247

Green P, Ewing P, Miller W, Thomas P, NISC Comparative Sequencing Program and Green E, 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nature Genetics 33: 514-517

Gu X and Zhang H, 2004. Genome phylogenetic analysis based on extended gene contents. Molecular Biology and Evolution 21: 1401-1408

Hardison R, Roskin K, Yang S, Diekhans M, Kent W, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, Schwartz S, Furey T, Whelan S, Goldman N, Smit A, Miller W, Chiaromonte F and Haussler D, 2003. Covariation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. Genome Research 13: 13-26

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

Hefferon T, Groman J, Yurk C and Cutting G, 2004. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. Proceedings of the National Academy of Sciences USA 101: 3504-3509

Huson D and Steel M, 2004. Phylogenetic trees based on gene content. Bioinformatics 20: 2044-2049

International Chicken Genome Sequencing Consortium, 2004. Sequence and comparative analysis of the chicken provide unique perspectives on vertebrate evolution. Nature 432: 695-716

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921

Jones D, Taylor W and Thornton J, 1992. The rapid generation of mutation data matrices from protein sequence. CABIOS 8: 275-282

Jukes T and Cantor C, 1969. Evolution of protein molecules (21-132 in *Mammalian Protein Metabolism*, Academic Press, New York, edited by Munro H)

Keightley P, Lercher P and Eyre-Walker A, 2005a. Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biology 3: 282-288

Keightley P, Kryukov G, Sunyaev S, Halligan S and Gaffney D, 2005b. Evolutionary constraints in conserved nongenic sequences of mammals. Genome Research 15: 1373-1378

Kellis M, Patterson N, Birren R, Berger B and Lander E. 2004. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. Journal of Computational Biology 11: 319-355

Kellis M, Patterson N, Endrizzi M, Birren and Lander E. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241-254

Kolaczkowski B and Thornton J, 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431: 980-984

Korf I, Flicek P, Duan D and Brent M, 2001. Integrating genomic homology into gene structure prediction. Bioinformatics 17:S140-S148

Kosiol C, 2006. Thesis submitted.

Kowalczuk M, Mackiewicz P, Gierlik A, Dudek M and Cebrat S, 1999. Total number of coding open reading frames in the yeast genome. Yeast 15: 1031-1034

Kulp D, Haussler D, Reese M and Eeckman F, 1996. A generalized hidden markov model for the recognition of human genes in DNA. Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology 134-141. AAAI Press, CITY.

Lercher M, Chamary J and Hurst L, 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Research 14: 1002-1013

Lercher M, Williams E and Hurst L, 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. Molecular Biology and Evolution 18: 2032-2039

Lobry J and Sueoka N, 2002. Asymmetric directional mutation pressures in bacteria. Genome Biology 3: 0058.1-0058.14

Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek M and Cebrat S, 2004. Where does bacterial replication start? Rules for predicting the *oriC* region. Nucleic Acids Research 32: 3781-3791

Massingham T, 2002. Detecting positive selection in proteins: models of evolution and statistical tests. PhD Thesis, University of Cambridge

McAuliffe J, Pachter L and Jordan M, 2004. Multiple-sequence functional annotation and the generalised hidden markov phylogeny. Bioinformatics 20: 1850-1860

McGuire G, Denham M and Balding D, 2001. Models of evolution for DNA sequences including gaps. Molecular Biology and Evolution 18: 481-490

Meyer I and Durbin R, 2002. Comparative *ab initio* prediction of gene structures using pair HMMs. Bioinformatics 18: 1309-1318

Montoya-Burgos J, Boursot P and Galtier N, 2003. Recombination explains isochores in mammalian genomes. Trends in Genetics 19: 128-130

Mouse Genome Sequencing Consortium, 2002. Initial sequencing of the mouse genome. Nature 420:520-562

Muse SV and Gaut B, 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. Molecular Biology and Evolution 11: 715-724

Nachman M and Crowell S, 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304

Ohta T, 1992. The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics 23: 263-286

Page R and Holmes E, 1998. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Press, Oxford

Pedersen J and Hein J, 2003. Gene finding with a hidden markov model of genome structure and evolution. Bioinformatics 19: 219-227

Philippe H, Snell E, Bapteste E, Lopez P, Holland P and Casane D, 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Molecular Biology and Evolution 21: 1740-1752

Phillips M, Delsuc F and Penny D, 2004. Genome-scale phylogeny and the detection of systematic biases. Molecular Biology and Evolution 21: 1455-1458

Posada D and Crandall K, 2001. Selecting the best-fit model of nucleotide substitution. Systematic Biology 50: 580-601

Posada D and Crandall K, 2002. The effect of recombination on the accuracy of phylogeny estimation. Journal of Molecular Evolution 54: 396-402

Press H, Teukolsky S, Vetterling W and Flannery B, 1992. *Numerical Recipes in C*. Cambridge University Press, Cambridge

Pupko T, Huchon D, Cao Y, Okada N, and Hasegawa M, 2002. Combining multiple datasets in a likelihood analysis: which models are best. Molecular Biology and Evolution 19: 2294-2307

Qian B and Goldstein R, 2003. Detecting distant homologues using phylogenetic tree-based HMMs. Proteins 52: 446-453

Rambaut A, Posada D, Crandall K and Holmes E, 2004. The causes and consequences of HIV evolution. Nature Reviews Genetics: 52-61

Rat Genome Sequencing Consortium, 2004. Genome sequence of the brown Norway rat yields insights into mammalian evolution. Nature 428: 493-521

Ren F, Tanaka H and Yang Z, 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Systematic Biology 54: 808-818

Rivas E and Eddy S, 2001. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2:8

Rivera M and Lake J, 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431: 152-155

Rokas A, Williams B, King N and Carroll S, 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798-804

Savill N, Hoyle D, and Higgs P, 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics 157: 399-411

Shapiro B, Rambaut A and Drummond A, 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Molecular Biology and Evolution 23: 7-9

Simes R, 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751-754

Steel M, 2005. Should phylogenetic models be trying to 'fit an elephant'? Trends in Genetics 21:307-309

Storey J and Tibshirani R, 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences USA 100: 9440-9445

Swofford D, Olsen G, Wadell P and Hillis D, 1996. Phylogenetic Inference (407-543 in *Molecular Systematics* 2nd Edition, Sinaeur Associates Inc, Sunderland MA, by Hillis D, Moritz C and Mable K)

Suzuki Y and Gojobori T, 1999. A method for detecting positive selection at single amino acid sites. Molecular Biology and Evolution 16:1315-1328

Thorne J, Goldman N and Jones D, 1996. Combining protein evolution and secondary structure. Molecular Biology and Evolution 13:666-673

Topal M and Fresco J, 1976. Complementary base pairing and the origin of substitution mutations. Nature 263: 289-293

Touchon M, Nicolay S, Arneodo A, d'Aubenton-Carafa Y and Thermes C, 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Letters 555: 579-582

Watson J and Crick F, 1953. A structure for deoxyribose nucleic acid. Nature 4356: 737-738.

Webster M, Smith N and Ellergen H, 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Molecular Biology and Evolution 20: 278-286

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

Whelan S and Goldman N, 1999. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Molecular Biology and Evolution 16: 1292-1299

Whelan S and Goldman N, 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution 18:691-699

Whelan S and Goldman N, 2004. Estimating the frequency of events that cause multiple-nucleotide changes. Genetics 167: 2027-2043

Whelan S, Lió P, Goldman N, 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends in Genetics 17:261-272

Wilks S, 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics 9:60-62

Wong W, Yang Z, Goldman N and Nielsen R, 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168: 1041-1051

Yang Z, 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396-1401

Yang Z, 1994a. Estimating the pattern of nucleotide substitution. Journal of Molecular Evolution 39:105-111

Yang Z, 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306-314

Yang Z, 1996. maximum-likelihood models for combined analyses of multiple sequence data. Journal of Molecular Evolution 42: 587-596

Yang Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555-556

Yang Z and Bielawski B, 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution 15: 496-503

Yang Z, Goldman N and Friday A, 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Molecular Biology and Evolution 1: 725-736

Yang Z, Goldman N, and Friday A, 1995b. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Systematic Biology 44: 384-399

Yang Z, Lauder IJ, Lin HJ, 1995. Molecular evolution of the Hepatitis B Virus genome. Journal of Molecular Evolution. 41:587-596

Yang Z, Nielsen R, Goldman N and Pedersen A, 2000. Codon-substitution models for the heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449

Yang Z and Roberts D, 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Molecular Biology and Evolution 12: 451-458

Zhang M, 2002. Computational prediction of eukaryotic protein-coding genes. Nature Reviews Genetics 3:698-709

Zhu G, Golding G and Dean AM, 2005. The selective cause of an ancient adaptation. Science 307: 1279-1282