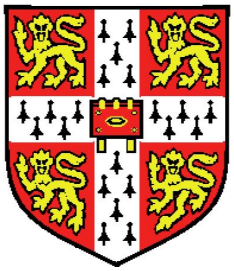
Mathematical Models and Statistics for Evolutionary Inference



Sarah L Parks

European Bioinformatics Institute

Sidney Sussex College, University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

August 2014

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text and acknowledgements.

This dissertation 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 60,000 words as defined by the Biology Degree Committee.

Sarah Parks

Acknowledgements

Firstly, I would like to thank my supervisor Nick Goldman for introducing me to the fascinating world of phylogenetics, fostering my love of maths, and along the way teaching me about English sentence structure. His support and guidance have been invaluable. I would also like to thank him for allowing me to visit the ONS for three months, and to take part in a range of teaching activities.

Working in the Goldman Group has been a real pleasure. I would like to thank Botond Sipos, Tim Massingham, Christophe Dessimoz, Kevin Gori, Greg Slodkowicz, Roland Schwarz, Asif Tamuri and visiting members of the Goldman Group for inspiring and entertaining chats, both scientific and non-scientific in nature.

Whilst working on my PhD I have had the pleasure of talking to and getting advice from my Thesis Advisory Committee: Simon Tavaré, John Marioni, and Jan Korbel, who have helped guide both my thesis, and my future career plans. I would like to thank Sergei Kovakovsky Pond for extremely responsive help with HyPhy, both fixing bugs and introducing me to its batch language; Botond Sipos, Tjaart de Beer, Janet Thornton and Nick Goldman for interesting discussions about modelling evolution in humans; and Greg Slodkowicz for long discussions about multiple testing correction.

My PhD was funded by both EMBL and the BBSRC, to whom I am very grateful. I would particularly like to thank the BBSRC for giving me the opportunity to work at the ONS, and for funding my trip.

I would also like to thank Aidan Budd and Julius Ross for inviting me to teach, either at the University or further afield, and all of the people I have taught with over the years for their encouragement and support.

I am grateful to Asif Tamuri, Laura Emery, Camille Terfve, Steve Wilder, Greg Slodkowicz and Nick Goldman for reading and correcting draft chapters of my thesis.

I would like to thank my fantastic network of friends, both in Cambridge and further afield, who have always been there to support me when I needed it, or to help me celebrate when something went well. Thanks for cold mornings on the river, silly gym sessions, help with house-moving, visits to Cambridge, cocktail parties, tea parties, meals out and late night Skype calls. A special mention to Camille Terfve, with whom I shared this journey, for being the greatest friend anyone could hope for.

Finally I would like to thank my family: Mum, Dad, Nana, Grandad, Granny, Gill, Simon, Robert, Anna, and Dan for being there at all times and always believing in me. A special final thank you to Dan, for his patience and love, coping with my regular travelling and always being there to put a smile on my face.

Abstract

Molecular phylogenetics aims to use genomic sequence data to reconstruct the evolutionary history of species and understand evolutionary processes. It has proven to be widely useful with statistical methods using mathematical models of the process of evolution being extensively developed. Despite this, there are still a number of fundamental questions and assumptions in phylogeny reconstruction and statistical modelling of evolution that are not well understood. The focus of this thesis is exploring, and improving understanding of, three of these issues.

The first of these is long branch attraction (LBA), a regularly cited but poorly understood phenomenon that has been claimed to cause inaccurate phylogenetic tree topology reconstruction. LBA has been reported to affect all tree reconstruction methods, including maximum likelihood, a state-of-the-art statistical method that forms the basis of all of the data analysis methods studied in this thesis. I carry out an analysis of the simplest possible case, that of one long branch on a three-species tree, showing that even this causes counterintuitive results that I am able to explain. I then move to four-species trees and show that the LBA phenomenon does exist, but is not caused by an attraction.

Secondly, I study the impact of the assumption of time-reversibility that is made by the majority of models used for phylogeny reconstruction. Models of the process of substitution of characters (nucleotides or amino acids) in molecular sequences are time-reversible if, at equilibrium, the amount of change between characters i and j is the same as the amount of change between characters j and i. This assumption has no biological basis; it has been maintained solely for computational and mathematical reasons. Relaxing it could produce models that describe the process of evolution better. I analyse whether non-reversible models fit data better for nucleotides and, for the first time, amino acids, and show that they are often a significant improvement. Analysis of non-reversibility also requires measures of quantification. I develop measures for non-reversibility and show that they help distinguish effects different from the total strength of evidence of reversibility.

Finally, I develop an improved test for positive selection. Detecting site-wise positive selection has long been a key subject in evolutionary biology. Unfortunately many methods are statistically very conservative and lack power. I develop a new method using site-wise likelihood ratio tests that achieves greatly improved power whilst retaining control of the false positive rate. It involves altering the null hypothesis of a previous site-wise likelihood ratio test so that the null hypothesis better fits the data. I test this method using simulations over a wide range of realistic conditions, showing that the power can be doubled or tripled while the false positive rate is kept under control.

Contents

Chapter 1

Introduction

Molecular phylogenetics has proven to be widely useful for improving our understanding of biological processes. This ranges from illustrating evolutionary relationships (Meredith et al. 2011), understanding evolutionary mechanisms by comparing models (Whelan et al. 2001), and identifying sites under positive and purifying selection (Nielsen and Yang 1998; Massingham and Goldman 2005) to understanding host-pathogen relationships (Tamuri et al. 2012), dating speciation events (Rutschmann 2006), and helping locate the source of disease outbreaks in hospitals (Köser et al. 2013).

Despite the widespread use of phylogenies and related parameters for understanding evolutionary processes, and the development of increasingly complicated models, there are still numerous fundamental questions and assumptions in statistical modelling of evolution which are not well understood. The focus of this thesis is exploring, and improving understanding of, some of these issues, specifically: long branch attraction, a poorly understood phenomenon claimed to cause inaccurate topology inference; reversibility, an assumption made by almost all evolutionary models for computational rather than biological reasons; and site-wise positive selection, which many current tests lack power to find.

1.1 Phylogenetics

The genetic information of an organism is contained in its genome. This information resides in DNA sequences that are passed on from parent to child by replication. A genome is made up of coding and non-coding DNA sequence. Coding sequences are transcribed into messenger RNAs and then translated into proteins. Non-coding genetic material is not translated, but can nonetheless have a functional role at the DNA or RNA level. Proteins are biological molecules that

perform a variety of functions in the cell such as DNA replication, catalysing metabolic reactions, and transporting molecules around the cell. They consist of chains of amino acids that are encoded in genomic sequences using the genetic code, a code based on triplets of DNA called codons. There are 20 amino acids and $4^3 = 64$ codons, only three of which are stop codons, meaning the genetic code is degenerate, with some amino acids encoded by more than one codon.

The process of replication is imperfect, and so is that of DNA repair. Errors, called mutations, can therefore be made and transmitted to the next generation. For both coding and non-coding sequences, as it is the nucleotides that mutate, evolution is often modelled on the nucleotide level. Because functional consequences of mutations in protein-coding sequences mostly happen at the protein level, it is also popular to model mutations on the amino acid and codon level for such sequences (Yang 2014). Over generations mutations are either lost, or spread within a population until they reach equilibrium. If all individuals have them then they are fixed. Fixed mutations, often called substitutions, leave a molecular record of the evolutionary relationships between species that molecular phylogenetics aims to reconstruct. These relationships are generally represented by a phylogenetic tree, which can either be rooted, so that it indicates the ancestor-descendent relationships between the organisms studied, or unrooted, and hence just displays branching patterns. We assume that all sites within the sequence being studied evolve on the same tree, and that once a branching event has occurred lineages evolve independently. In this thesis branch lengths on a tree indicate the amount of evolution, measured as the number of substitutions per site. If additional information is available it can also be possible to calibrate evolutionary trees so that branch lengths represent calendar time (Yang 2014).

Mutations can take the form of substitutions at single sites, short sequence insertions or deletions (indels), or larger-scale events such as recombination or horizontal gene transfer. Substitutions at single sites can be modelled using Markov processes (Yang 2014). These substitution models, and phylogenies estimated using them, are the focus of this thesis. Although substitutions are generally assumed to act on one site at a time, substitution models can also be extended to model mutations on fixed length, non-overlapping, DNA segments, such as codons (Goldman and Yang 1994; Muse and Gaut 1994), or potentially the entire sequence (Robinson et al. 2003; Rodrigue et al. 2005; Yu and Thorne 2006). In general the larger the segment, the more computationally intractable the problem, so models for the entire sequence are not widely used.

Insertions and deletions are of variable length and can affect multiple sites at a time. Generally, the process of alignment aims to identify homologous sites and

the positions of any indels; gaps in sequence alignments are then usually ignored for the aims of mathematical modelling. It is also possible to treat gaps as another state in a substitution model (McGuire et al. 2001), meaning that each gap is treated independently of other gaps, but this is uncommon (Yang 2006). Ideally, we would use models which take into account both substitutions and indels. Attempts have been made to develop these (e.g. Bishop and Thompson 1986; Thorne et al. 1991; Thorne and Kishino 1992; Hein et al. 2000; Hein et al. 2003); unfortunately these models are generally still not computationally tractable (Yang 2014). Progress is being made on methods for the joint inference of alignment and phylogeny (Miklós et al. 2004; Lunter et al. 2005; Fleissner et al. 2005; Redelings and Suchard 2005), and on estimates of evolutionary distances between sequences which take into account indels as well as substitutions (Schwarz et al. 2010).

Large-scale events are also not specifically modelled by substitution models, however their presence can affect evolutionary analyses. Processes that lead to genetic material being passed amongst organisms in a manner not reflecting the usual tree-like structure of inter-species relationships, such as recombination or horizontal gene transfer, lead to violation of the assumption that all sites evolve along the same tree. Recombination has been shown to adversely affect tree reconstruction (Posada and Crandall 2002), molecular clock inference (Schierup and Hein 2000), the detection of positively selected sites (Anisimova et al. 2003; Shriner et al. 2003) and ancestral reconstruction (Arenas and Posada 2010). Therefore, methods for identifying and handling recombination events have been developed (Grassly and Holmes 1997; Husmeier and Wright 2001; Kosakovsky Pond et al. 2006b). A variety of methods also exist for detecting horizontal gene transfer (Dessimoz et al. 2008; Abby et al. 2010). In theory, an alternative way to model these events is to relax the assumption that all sites evolve under the same tree, and represent evolutionary relationships as a network where nodes can represent large-scale events as well as speciation. In practice, the calculation of rooted phylogenetic networks is difficult and there are not yet any widely used tools (Huson and Scornavacca 2010). For a detailed mathematical introduction to phylogenetic networks and methods for building them see Huson et al. (2011); for a biological viewpoint see Morrison (2011). In this thesis I study mostly theoretical problems, and assume that users will check for the presence of large-scale events before using methodology presented here.

Evolutionary relationships between species were originally inferred by comparing morphological features. These phenotypical changes are caused by molecular changes, so as sequence data slowly became available sequence alignments became an alternative, and often preferable, data set to use for building phylo-

genetic trees. Initially, the changes between sequences were counted and used to infer the divergence between certain sequences (Zuckerkandl and Pauling 1962; Zuckerkandl and Pauling 1965), or to develop probabilistic models of amino-acid substitutions (Eck and Dayhoff 1966; Dayhoff and Eck 1968; Dayhoff et al. 1972, 1978). These first studies used the principle of parsimony, that the simplest explanation is the best explanation, interpreted to mean that the tree with the fewest changes on it is the best tree. This principle was already the basis for building trees from morphological data. Use of parsimony for studying molecular data was then promoted as it was believed to be 'assumption-free' (Wiley 1981). In fact, although it may not make explicit assumptions, there are implicit assumptions, although it can be difficult to identify them (e.g. Goldman 1990; Yang 2006).

Contemporaneously, statistical models of evolution were also being developed. Starting with a model describing nucleotide evolution (Jukes and Cantor 1969), Neyman (1971) showed how substitutions over a tree could be described by a Markov process, and, with the work of Felsenstein (1981), inference of these models on trees was made computationally feasible. During the 1980s, while both 'assumption-free' methods and statistical model-based methods were being developed and used on molecular data, there was heated debate over which approach was better (Felsenstein 2001, 2004). However, from the mid 1990s onwards statistical methods became the methods of choice as they allowed for the fit of the model to the data to be evaluated and improved (Goldman 1993). They were also shown to be statistically consistent under the true model (Chang 1996; Rogers 1997), meaning that the probability of obtaining the correct tree tends to one as the amount of data tends to infinity, whereas parsimony was shown to be inconsistent in some circumstances (Felsenstein 1978). Since then, with the increase in both quantity of sequence data and computational power, a large number of evolutionary models and statistical methods for constructing phylogenies and understanding evolutionary processes have been developed.

Statistical methods can be split into two categories: frequentist methods, such as maximum likelihood (ML), and Bayesian methods. Both methodologies are widely used within phylogenetics (Yang 2014); both have advantages and disadvantages. ML is a standard statistical framework for model inference, and comes with well-studied techniques for comparing and testing models (Fisher 1921, 1925; Wilks 1938; Edwards 1972). It yields a best point estimate of the parameter(s) of interest. Bayesian inference on the other hand outputs the posterior distribution of the parameter(s) given a prior distribution and the data (Yang 2014). This technique requires the investigator to specify a prior probability of the hypothesis. In this thesis I only study ML inference.

Another set of methods for building phylogenetic trees, distance matrix methods, uses pairwise distances, estimates of the evolutionary distance between pairs of sequences, generally in units of the average number of substitutions per site which have occurred between the two sequences. Pairwise distances are calculated between all sequences of interest, and fitted to a tree. These methods do not make full use of the information in a sequence alignment, and are therefore not expected to be optimal (Yang 2014). However, they have been shown to work reasonably (Felsenstein 2004), and are generally faster than ML and Bayesian approaches. Distance methods are not specifically studied in this thesis but they are described in Chapter 2, and found to be useful when considering the effects of ML on long branches in Chapter 3.

Statistical methods require a model of evolution specifying the rates of change between states. The first, and most simplistic, models made a large number of assumptions about the biological process of evolution. Over time these assumptions have been relaxed in order to find models which fit data better, but there are a number of common assumptions that most models still hold. Some key assumptions are that: sites evolve independently of each other; sites evolve under the same process and same rate; rates of change are constant over time (time-homogeneity); the process is at equilibrium (stationary); and the process is reversible (the direction of change cannot be determined at equilibrium). In most real data sets some of these assumptions will be violated. Methods which relax these assumptions are under development, and I will now briefly review the progress so far.

Independence between sites is a difficult assumption to relax computationally. Even if only immediate neighbours are considered, 'contagious dependence' means that each neighbour is affected by its own neighbours so that all the nucleotides are actually dependent on each other (Lunter and Hein 2004). However, there is significance evidence that sites are not independent, and that their context is important. A common example of this is the hypermutability of CpGs. CpGs are a C nucleotide directly followed by a G nucleotide in the genome. Mutation of the C in CpGs to T is the most common substitution in mammalian genomes, and occurs around 10–50 times more frequently than any other substitution (Duncan and Miller 1980; Duret et al. 2006; Walser and Furano 2010). To take this into account, models which allow dinucleotide substitution rates to be estimated have been developed (von Haeseler and Schoniger 1998; Jensen and Pedersen 2000; Lunter and Hein 2004; Hwang and Green 2004; Siepel and Haussler 2004). Work is ongoing on context-dependent models; for a detailed review of the development of context-dependent non-coding models and current issues in the field see Baele

(2011).

The assumption that sites evolve at the same rate and with the same process is often violated, for example by positions in codons evolving at different rates. This is often dealt with by modelling codons themselves instead of nucleotides or amino acids (Muse and Gaut 1994; Goldman and Yang 1994), allowing independence between sites, which are now whole codons, to be kept, while taking into account codon structure. However, due to variation in selective constraints over protein sequences, rates may vary over the entire protein, and not just between codon positions. The most common way to incorporate this is to describe the rate as a random draw from a statistical distribution for each site (Yang 1993, 1994b). Further details are given in Section 2.1.1. Variation in process can also be modelled in a similar way, for example variation in selective pressure over sites (Nielsen and Yang 1998). More complex heterogeneity, such as variation in the whole substitution matrix, has also been explored (Lartillot and Philippe 2004).

Time-homogeneity means that the process itself does not change over the tree. This assumption can be violated in a number of ways: for example, selective pressure may vary over time, or pathogens may have different evolutionary processes in different hosts. The first of these examples is generally dealt with by using branch-site models, which allow selective pressures to be different on predetermined branches so that episodic selection can be located (Yang and Nielsen 2002). In the second example different processes can be estimated on different branches of the phylogeny, significantly improving the fit of the data (dos Reis et al. 2009; Tamuri et al. 2009). Work is ongoing on further models for taking time-non-homogeneity into account (Goode et al. 2008; Blanquart and Lartillot 2008; Bielejec et al. 2014).

A process is at equilibrium (stationary) if the probability of finding a site in a particular state does not vary over the tree. It therefore has an equilibrium distribution consisting of these probabilities for each state. The assumption of stationarity has been tested and found to be violated both for the evolution of GC content (Arndt et al. 2005), and for general evolutionary models (Squartini and Arndt 2008). Non-stationary models have been suggested (Barry and Hartigan 1987) and shown to give a better fit to the data (Jayaswal et al. 2010). However, relaxing the assumption of stationarity increases the number of possible models; choosing between this set of models is a difficult problem (Jayaswal et al. 2010, 2011). The majority of models used currently are stationary ones.

A process is reversible if, at equilibrium, the amount of change from nucleotide i to nucleotide j is the same as the amount of change from nucleotide j to nucleotide i (Norris 1998). For a process to be reversible it must have an equilibrium

(stationary) distribution. However just because a process is not at equilibrium does not mean that it is not reversible. For example, if a reversible process is acting on a sequence which is not at its equilibrium distribution, then this process will not be stationary until equilibrium is reached. Reversibility is an assumption held by the majority of models for computational, not biological, reasons. In Chapter 4, I assess whether non-reversible models describe the evolutionary process better than reversible models.

1.2 Thesis Outline

This thesis consists of three largely separate projects, all of which aim at better understanding or improved methodology for phylogenetic inference. They are all based on ML estimation of phylogenies from sequence alignments, focusing on different parameters of interest. Each chapter contains its own introduction; however, as the mathematical background to these projects is similar, Chapter 2 contains an introduction to the necessary mathematics and assumptions. Throughout this thesis I assume evolution proceeds according to a Markov process which can be parameterised in a variety of biologically informative ways, such as on a nucleotide, amino acid, or codon level. Many methodologies exist for inferring phylogenies, as well as parameters of these processes, from sequence alignments. In this thesis I focus on ML estimation, but I also use distance methods to gain insight on ML processes and therefore I briefly describe both methods of inferring phylogenies.

Chapter 3 describes work on ML inference of small trees in the presence of long branches. It has long been known that long branches can cause problems for topology estimation (Felsenstein 1978). Intuitively, this is not surprising, as the longer the branch the more changes that will have occurred which need to be inferred. Discussion of this has previously focused on long branch attraction (LBA), a regularly cited term generally used to describe a propensity for long branches to be joined together in estimated trees (Hendy and Penny 1989). LBA has been claimed to affect all major phylogenetic reconstruction methods, including ML (Huelsenbeck 1995). Despite the widespread use of this term in the literature, exactly what LBA is and what may be causing it is poorly understood, even for simple evolutionary models and small model trees. Until now the focus has always been on two long branches, and no-one has considered the extent to which even one long branch may be problematic. I look, for the first time, at the effect of just one long branch, in particular the placement of one long branch on a three species tree, and show that it is placed unexpectedly. I am able to explain

this with the use of both ML and distance method equations. I go on to look at the placement of two long branches on four-taxon trees, showing that there is no attraction between long branches, but that for extreme branch lengths long branches are joined together disproportionally often. These results illustrate that even small model trees are still interesting to help understand how ML phylogenetic reconstruction works, and that LBA is a complicated phenomenon that deserves further study. This work has been published in *Systematic Biology* and is presented here largely unchanged (Parks and Goldman 2014).

Chapter 4 investigates the assumption of time-reversibility on phylogeny inference. Almost all evolutionary models commonly used in phylogenetic analysis are a subset of the general time reversible (GTR, REV) model (Lanave et al. 1984; Tavaré 1986), and therefore assume the mathematical condition of time reversibility. This assumption reduces computational effort and eases mathematical complexity when calculating likelihoods (Yang 2014); it has no biological basis. Relaxing this assumption by using a non-reversible model may fit data significantly better, potentially giving a more accurate description of evolution, better trees and other benefits. A consensus has not been reached on whether non-reversible models are significantly better than reversible models (Yang 1994c; Squartini and Arndt 2008; Jayaswal et al. 2010). Often studies have only looked at a few trees and alignments and made decisions on the utility of non-reversible models based on this. Further, most of this work has been done on nucleotide data sets: no previous research has investigated non-reversible models for amino acid data sets, even though many phylogenies are built using amino acid data. Analysis of non-reversibility also requires measures of quantification, which so far have not been developed. In Chapter 4 I explore the use of non-reversible models for nucleotide data sets, and, for the first time, amino acid data sets. I devise a number of new measures of non-reversibility and explored their relationship to measures of the strength or evidence for reversibility, such as likelihood ratio tests, a common method for choosing between nested models. I show that these effects are different and can be distinguished. I perform these assessments on both amino acid and nucleotide data, and cover a much larger range of data sets than previous papers, allowing me to draw conclusions about the applicability of non-reversible models for nucleotide and amino acid alignment data sets.

Chapter 5 moves to codon models that allow for measurement of selection. Finding positively selected genes, or sites in genes, is a key question in biology (Kosiol et al. 2008). A variety of maximum likelihood and Bayesian methods for testing for positive selection using a parameter ω, the ratio of the fixation rates of non-synonymous and synonymous mutations, have been developed (Nielsen

and Yang 1998; Massingham and Goldman 2005; Kosakovsky Pond and Muse 2005; Murrell et al. 2013). These can be statistically conservative when applied to real genes, and hence achieve a lower power than desired. One of these tests, embodied in the SLR method (Massingham and Goldman 2005), estimates the maximum likelihood value of ω for each site, and then tests whether this value is greater than 1. The test as originally published by Massingham and Goldman (2005) is conservative because the null hypothesis assumes all sites are neutral ($\omega = 1$), whereas in reality the majority of sites in genes are under purifying selection ($\omega < 1$). I present a new method to test for positive selection that has greatly improved power whilst retaining control of the false positive rate. This involves a new site-wise likelihood ratio test, designed to have power and control when many of the sites in the gene are under purifying selection (as is typically the case), and a diagnostic for detecting certain situations in which the original SLR test should be preferred. The new test achieves improved power by fitting the null hypothesis to the data and then performing parametric bootstraps (Efron and Tibshirani 1993). The method is tested using simulations over a wide range of realistic conditions, including standard comparisons used in previous studies and larger and more realistic examples modelled on real-life studies. I show that, for those rare cases where all sites are either strictly neutral ($\omega = 1$) or positively selected ($\omega > 1$), the new method performs as well as SLR. More importantly, for genes where many of the sites are conserved, this method has much better power than SLR and a controlled false positive rate.

In Chapter 6 I summarise the work and tie together the themes within the separate chapters.

Chapter 2

Mathematical Background

This chapter covers the mathematical background to this thesis. It is not possible to cover the entire field of phylogenetic theory, so I will concentrate on areas where I make contributions, giving an overview of the main assumptions made when modelling evolution and introducing relevant mathematical terminology and techniques. A useful reference text is Yang (2014).

In this thesis I assume that sequence alignments are given and correct. This is a standard assumption, as it is very difficult to use knowledge, or inferred knowledge, of alignment error, in phylogeny inference. Methods that take into account alignment uncertainty by integrating over all possible alignments while inferring the phylogeny and model parameters are being developed (Lunter et al. 2005; Redelings and Suchard 2005; Redelings 2014); however, they are still very computationally expensive and cannot be applied to large alignments. Although it is clear that the assumption that alignments are correct may often be invalidated, the inference of alignments is a different but related problem which is not assessed in this thesis.

2.1 Markov Models

Evolutionary models describe the rates of change between states. An assumption of almost all evolutionary models is that future events depend only on the current state and not on past events. In probability theory this 'lack of memory' assumption is called the Markov property, and processes where it holds are Markov processes (Norris 1998). This assumption is reasonable biologically because at any instant in time mutations, and natural selection, can only occur on the sequence present, and will not know what has occurred before. Some experimental protein studies have shown evidence of non-Markovian amino acid

processes (Benner et al. 1994; Mitchison and Durbin 1995); however, it has since been shown that data produced by a nucleotide or codon Markov process can be non-Markovian when analysed on the amino acid level (Kosiol and Goldman 2011).

It is generally also assumed that sites evolve independently. This assumption is unlikely to hold in practice due to selective and functional constraints; however, it is made because it significantly reduces the computational complexity of model estimation. Work has been carried out on context-dependent models which relax this assumption; details of some of these models are given in Section 1.1 (also see von Haeseler and Schoniger 1998; Jensen and Pedersen 2000; Lunter and Hein 2004; Hwang and Green 2004; Siepel and Haussler 2004; Baele 2011 etc.). These models are not yet in general use and are not considered in this thesis.

Throughout this thesis I also assume time homogeneity, meaning that instantaneous rates of change are constant (and hence there is one substitution model for the whole tree). Progress is being made in the development of time-inhomogeneous models (Blanquart and Lartillot 2008; Bielejec et al. 2014); however, these models are still experimental, and it is computationally and mathematically easier to keep this assumption.

2.1.1 Instantaneous Rate and Probability Matrices

The simplifying assumptions detailed above allow evolution to be described by a continuous-time Markov process. Sequences evolve by a series of independent substitutions that each change a character at one site into another character. This is generally modelled as a process over a tree, where branching events often signify speciations, and after branching events the two daughters evolve independently.

The rate at which substitutions between states occur is described by an $N \times N$ instantaneous rate matrix Q, with elements q_{ij} describing the instantaneous rate of change from character i to character j, $i \neq j$, and N the size of the character alphabet. The states of the Markov process are the characters in a sequence. In this thesis three different character alphabets are considered: nucleotides, amino acids and codons (without stop codons). N is therefore 4, 20 and 61 (for the universal genetic code), respectively. Whilst in a state i, the rate of leaving is the sum of the rates of moving from i to any other state j. The diagonals q_{ii} are set to be equal to -1 times the total rate of leaving a site, so that the rows of the matrix sum to 0 (Norris 1998). This matrix acts independently at each site.

In general when using these models to describe evolution or make inferences about evolution from data we are interested in the probability of moving between

two states in a given time period. In an arbitrarily small time interval Δt, the probability that a nucleotide i will change to a nucleotide j ($j \neq i$) is $q_{ij}\Delta t$. Over a larger time period, t, the probability matrix, $P(t)$, is calculated from

$$P(t) = \mathrm{e}^{tQ} = I + tQ + \frac{(tQ)^2}{2!} + \frac{(tQ)^3}{3!} + ...$$

where I is the appropriate size identity matrix (Norris 1998). $p_{ij}(t)$ is then the probability of change from i to j after a time t for all i and j. The matrix exponential can be calculated in a variety of ways (Moler and Loan 2003). In phylogenetics, for reversible models (see below), this is often carried out using matrix decomposition of Q into eigenvalues and eigenvectors (Yang 2006). For non-reversible models, this does not work well, so instead a function of the matrix is squared repeatedly. This is more computationally expensive than matrix decomposition (Yang 2006).

If a Markov process is irreducible, meaning that $p_{i,j}(t) > 0 \; \forall \; i \neq j$ and $t > 0$, and recurrent, meaning that the probability of returning to each state unboundedly many times is 1 (Norris 1998), then it has a unique equilibrium distribution, π. This means that in the long run the probability of finding a Markov process in a particular state converges to a value which is independent of the starting state. The equilibrium distribution π can be found by solving either

$$\pi P(t) = \pi \text{ for any } t > 0$$

or

$$\pi Q = 0$$

(Norris 1998). In general, phylogenetic Markov processes satisfy these conditions and therefore have unique equilibrium distributions. We generally also assume that processes are stationary, meaning that they are at equilibrium throughout the time period (evolutionary time) under consideration and the probability of finding any site in a particular state does not change over time.

The majority of phylogenetic Markov processes are also time reversible, meaning that, at equilibrium, the amount of change from nucleotide i to nucleotide j is the same as the amount of change from nucleotide j to nucleotide i (Norris 1998). Equivalently, a time reversible process (at equilibrium) observed forwards in time is indistinguishable from the same process observed backwards. Reversible models satisfy the detailed balance equations

$$\pi_i q_{ij} = \pi_j q_{ji} \; \forall i, j$$

where $\pi_i q_{ij}$, the amount of change from i to j, is often called the probability flux (Kelly 1979). If a model is reversible then this relationship also holds for all probability matrices: $\pi_i p(t)_{ij} = \pi_j p(t)_{ji}$. Although a reversible model at equilibrium is stationary, it is possible for a process to be reversible but non-stationary if it initially acts on a distribution which is not its equilibrium distribution.

Reversibility is not biologically necessary, but is assumed for the majority of phylogenetic Markov processes as it eases computational and mathematical issues, including matrix exponentiation (Yang 2006). Chapter 4 of this thesis is devoted to the study of the effect of this assumption on the inference of nucleotide and amino acid models, and calculation of measures for assessing the level of non-reversibility of models.

For all Markov processes time and rate are confounded, so it is not possible to estimate one without knowing extra information about the other (Felsenstein 1981). Any scalar product of Q is therefore indistinguishable from Q with a different unit of time. In general, and throughout this thesis, matrices are normalised so that the expected substitution rate at equilibrium, $\sum_i \sum_{j \neq i} \pi_i q_{ij}$, is 1. Time, and branch lengths on a tree, are therefore measured as the expected number of substitutions per site.

Markov processes are often parameterised to take into account biological features. For a reversible model the Q matrix can be written as the product of a diagonal matrix of the equilibrium distribution, and a symmetric matrix S representing the rates of substitution between two states (often called the exchangeabilities), so that $q_{ij} = \pi_j s_{ij}$ $\forall i \neq j$ and $s_{ij} = s_{ji}$. The equilibrium distribution is specifically parameterised within the Q matrix. The maximum number of parameters is:

$$\underbrace{\frac{N(N-1)}{2} - 1}_{\text{exchangeabilities accounting for matrix normalisation}} + \underbrace{(N-1)}_{\text{equilibrium distribution}} = \frac{N(N+1)}{2} - 2$$

This equals 8, 208, and 1889 for nucleotide, amino acid, and codon models, respectively. Non-reversible models are under less constraint than reversible models, and can have up to $N(N-1) - 1$ parameters (again accounting for matrix normalisation), giving 11, 379 and 3659 parameters for nucleotide, amino acid, and codon models, respectively.

The Markov processes described so far assume that each site evolves under the same process, but in real life this assumption is often violated. A common example is rate variation among sites, potentially caused by varying selective pressure across sites or codon structure. Another example is selective pressure

itself, where we are interested in measuring the variation between sites and in finding sites with particularly high or low selective pressure. A common way to introduce process variation across sites, but keep site independence, is to describe the variable of interest at each site as a random draw from a distribution. This allows each site to have a defined probability of having a certain value of this variable, independent of its neighbours. Again, this variable is assumed to stay constant throughout the tree (time homogeneity). For among-site rate variation, the most commonly used distribution is a gamma distribution (Yang 1993, 1994b). Work is also being carried out on models which do not assume that sites are independent. Details of these are given in Section 1.1.

2.1.2 Nucleotide Models

Nucleotide models have an alphabet of four characters and have been extensively developed and used. They are generally parameterised to take into account biological features; these parameters are then estimated for each data set individually. The first, and simplest, nucleotide model developed assumes that the rates of moving from one state to any other state are equal (Jukes and Cantor 1969); the probabilities of being in each state are also all equal. In this model, typically abbreviated as JC or JC69, there are no free parameters.

This was a reasonable starting point, but mutation is a biochemical process acting on each nucleotide differently, making it likely that some changes are more probable than others. As the availability of sequence data increased this was shown to be true, as it was noticed that transitions — substitutions within purines (nucleotides A and G) or pyrimidines (nucleotides C and T) — occurred more often than transversions — substitutions between purines and pyrimidines. The K80 model developed by Kimura (1980), therefore took into account this bias by incorporating a parameter, κ, representing the transition/transversion ratio. The probabilities of being in each state are still all equal. κ is generally estimated for each specific data set, often by ML, with values greater than one indicating that transitions occur at a higher rate than transversions, which is a better fit for most biological data sets (e.g. Brown et al. 1982).

The JC69 and K80 models described so far have equilibrium distributions with equal base frequencies. In reality many biological sequences have unequal nucleotide frequencies due to selective or mutational biases, so a model that incorporates this may better describe evolution. The first model to incorporate this was F81, developed by Felsenstein (1981), which had three free parameters to represent the probability of three of the four nucleotides. (A parameter for

the fourth nucleotide is not required as the probabilities of the four nucleotides needs to sum to one.) Further models which take into account both unequal base frequencies and transition/transversion bias were also developed: F84 (included in PHYLIP from 1984, and later published by Kishino and Hasegawa 1989), HKY (Hasegawa et al. 1985) and TN93 (Tamura and Nei 1993).

The most general time-reversible model, called the GTR or REV model, has three parameters to describe the equilibrium frequencies and five exchangeability parameters to describe the rates of change between characters (Lanave et al. 1984; Tavaré 1986). This model is often found to be the best model for describing nucleotide sequences, and is very regularly used (Kelchner and Thomas 2007). Computationally, of the models described in this section, this is the hardest to estimate. For $N = 4$ this is not a problem, neither for speed or robustness of parameter estimates, but for larger state spaces accurately estimating a general non-reversible model can become an issue.

All of the models described so far are time-reversible. Non-reversible models can also be parameterised in a biologically relevant way. For example, the reverse complement symmetric model accounts for the pairing between DNA strands in double-stranded organisms by setting the rate of substitution from one base to another equal to the rate of substitution between the conjugate of those two bases (Wu and Maeda 1987; Lobry 1995). This model has 8 free parameters, the same number as GTR, but is non-reversible. The most general non-reversible nucleotide model has 11 parameters. In my opinion, non-reversible models have not seen much use as they are computationally and mathematically more difficult to calculate. In Chapter 4 I discuss how useful they are and whether they should be used preferentially over reversible models.

2.1.3 Amino Acid Models

Amino acid models may be preferred over nucleotide models as amino acids have a larger character alphabet (20) and, due to the degeneracy of the genetic code, there are fewer amino acid changes than nucleotide changes, making amino acid sequences easier to align. In contrast to the parametric nucleotide models, amino acid models are generally empirical. That is, they are derived from a large number of sequence alignments and then assumed to be representative for the data under consideration, and hence are applied with no, or only a few, new parameters being estimated. The first models of amino acid change were calculated by using parsimony to count the number of changes between amino acids in a large number of sequence alignments (Dayhoff et al. 1978; Jones et al. 1992). If there have

been multiple changes at any sites then parsimony-based counting methods will underestimate the number of changes between amino acids. To take into account multiple hits, ML methods were introduced, and are now the most common way to estimate models (Adachi and Hasegawa 1996; Whelan et al. 2001; Le and Gascuel 2008; Dang et al. 2010, 2011).

To improve applicability, models specific to certain protein functions or locations were built, such as mitochondrial (Adachi and Hasegawa 1996; Yang et al. 1998), chloroplast (Adachi et al. 2000; Cox and Foster 2013) and secondary structure models (Goldman et al. 1998). Additionally a method for customising empirical models to fit a data set of interest was developed, which uses the exchangeabilities (s_{ij}) from the empirical model but replaces the equilibrium distribution (π) with the estimated distribution of amino acids from the data set in question. This can be applied to any of the empirical models, and is generally denoted by adding '+F' to a model's name or acronym (Cao et al. 1994). Recently an even more customised 'semi-empirical' model has been developed which uses principal component analysis to find the substitution rates which covary the most among different protein families, and then fits the top principal components to the data to find the best model for the particular data set (Zoller and Schneider 2013).

All amino acid models described and used so far are reversible. In Chapter 4 I study non-reversible amino acid models for the first time, and assess whether using non-reversible models significantly improves the fit of amino acid models to data.

2.1.4 Codon Models

Codons are the nucleotide triplets that encode amino acids. There are $4^3 = 64$ possible codons, but only 20 amino acids, so codons are degenerate, with some amino acids encoded by six codons, whilst others are encoded by just one or two. The degeneracy is mainly concentrated in the third codon position, with the majority of third codon position substitutions being synonymous (not changing the amino acid). Some first codon position substitutions are also synonymous, whereas all second codon position substitutions are nonsynonymous.

Modelling evolution at the codon level allows for the genetic code to be taken into account. This means that different codon positions no longer have to evolve at the same rate or with the same process. It also means that nonsynonymous and synonymous changes can be modelled differently. An important consequence of this, and the initial motivation for the development of codon models, is that rates

of synonymous and nonsynonymous substitutions can be measured and compared to determine the level of selection acting on a protein or site. A common assumption is that, as synonymous substitutions do not affect the amino acid, they evolve neutrally. In this thesis 'neutral' is used to mean 'under no selective pressure', and is not related to distributions of selection effects being described as neutral or nearly neutral (Ohta and Gillespie 1996). If a sequence is evolving neutrally then the same will be true of nonsynonymous substitutions so the probability of fixation of nonsynonymous substitutions is the same as that of synonymous substitutions. If ω is the relative probability of fixation of a nonsynonymous substitution to a synonymous substitution, then $\omega = 1$ for neutrally evolving coding sequence. An excess of nonsynonymous substitutions compared to synonymous substitutions means that the amino acid is changed more often than would be expected under neutrality ($\omega > 1$). This is generally taken to indicate positive selection. On the other hand a dearth of nonsynonymous substitutions compared to synonymous substitutions means that the amino acid changes less often than would be expected, and so is under purifying selection ($\omega < 1$).

Positive selection itself has been divided into different types, including a) diversifying selection that causes frequent amino acid changes, e.g. an arms race between a host and a pathogen, b) directional selection that causes a particular set of amino acid changes to be rapidly fixed within a population, and c) balancing or frequency-dependent selection that causes an increase in variability within a population if there is a fitness advantage in maintaining a polymorphism (Yang 2006). Such processes of selection can also be categorised by whether they occur for a short time period, such as directional selection, or whether they are long-term and occur over the whole of the evolution of the sequences under study. Searching for $\omega > 1$ over a whole phylogeny is particularly useful for detecting long-term diversifying selection; it is not as useful for detecting short-term processes as these may not have $\omega > 1$ over the whole tree.

Codon models were originally developed by Goldman and Yang (1994) and Muse and Gaut (1994). Codons are modelled using a 61×61 substitution rate matrix which describes the rate of change between all codons except stop codons. Stop codons are not included as substitutions to or from them are likely to be highly deleterious and are therefore unlikely to be fixed. It is assumed that multiple substitutions do not occur instantaneously; rates of change between codons which differ by more than one nucleotide are all 0. This is a standard assumption for modelling codon evolution. There is however evidence that incorporating instantaneous multiple nucleotide changes improves the fit of models (Whelan and Goldman 2004; Kosiol et al. 2007). In this thesis, following the approach

of Nielsen and Yang (1998), I assume both synonymous and nonsynonymous mutations follow an HKY model, with the rate of substitutions proportional to the equilibrium frequency of each codon π_i, and the rate of transitions κ times higher than the rate of transversions (Hasegawa et al. 1985). It is also possible to use a GTR model instead (Kosakovsky Pond and Frost 2005; Murrell et al. 2013). Nonsynonymous substitutions occur ω times more often than synonymous substitutions. The Q matrix is then given by :

$$q_{ij} = \begin{cases} 0 & \text{if more than 1 nucleotide substitution required} \\ \pi_j & \text{if } i \to j \text{ is a synonymous transversion} \\ \pi_j\kappa & \text{if } i \to j \text{ is a synonymous transition} \\ \pi_j\omega & \text{if } i \to j \text{ is a nonsynonymous transversion} \\ \pi_j\kappa\omega & \text{if } i \to j \text{ is a nonsynonymous transition} \end{cases}$$

where κ and ω are estimated from the data, i and j are codons, and $i \neq j$. The equilibrium frequency parameters π_i can be calculated in a variety of ways. One way is to estimate the four nucleotide frequencies and then calculate the expected frequency of each codon based on these (F1×4); alternatively, nucleotide frequencies can be estimated for each codon position separately, with codon frequencies again estimated from these nucleotide frequencies (F3×4); another way is to directly estimate the frequency of each codon within a data set (F61). These methods have 3, 9 and 60 free parameters respectively. In this thesis F61 is used as it has been shown to give the best fit to real data (Goldman and Yang 1994; Yang and Nielsen 1998). This model is reversible.

Initially all parameters, including ω, were assumed to be constant for all sites of a protein, meaning that the estimated value of ω represented an average over all sites. Studies using this approach generally found that adaptive evolution ($\omega > 1$) was rare, and most proteins are under strong purifying selection ($\omega \ll 1$) (Endo et al. 1996). The low number of genes found using these models to be under positive selection was probably an underestimate (Sharp 1997). In general, if positive selection acts on a gene, it will not be expected to act on every site. Many sites will be highly adapted to their function, and hence will be under purifying selection. The average ω value therefore may be less than 1 even if some sites are under positive selection. To deal with this, methods were developed which allow ω to vary over sites, and to be estimated for each site so that sites under positive selection could be identified. In Chapter 5, I develop a new method for detecting sites under positive selection. The introduction to that chapter describes the history of different methods for detecting positive selection, and gives further

details on different methodologies. Further details on the uses of codon models can be found in Anisimova and Kosiol (2009).

2.2 Distance Methods for Inferring Phylogenies

Distance matrix (DM) methods for inferring phylogenetic trees are based on computing pairwise distances and using some criterion to fit these distances to a tree (Yang 2006). The simplest method to calculate pairwise distances is to count the differences between two aligned sequences. This does not take into account the possibility of multiple changes having occurred at the same site and therefore the distance will always be underestimated. An improvement on this is to use a distance derived from a Markov model, such as those described above, which intrinsically takes into account the possibility of multiple substitutions. For the simplest model, Jukes Cantor, the distance, which can be derived from the probability matrix, is

$$d = \frac{3}{4}\log(1 - \frac{4}{3}p)$$

where p is the proportion of sites that differ between two sequences. If $p \geq 3/4$ this distance is infinite. Derivations of distances for more complicated models can be found in Yang (2006).

Distance methods aim to find a tree and branch lengths which best fit the pairwise distances. A large number of methods have been developed: I will briefly describe a few of these, but further details can be found in Felsenstein (2004).

Least squares is a common method which aims to find the tree and branch lengths which minimise the sum of squared differences between the distance on the tree and the pairwise distances (Cavalli-Sforza and Edwards 1967). The sum is often weighted by the inverse of the square of the distances to take into account the fact that larger distances have a higher variance (Fitch and Margoliash 1967). An alternative method is minimum evolution, which uses least squares to find branch lengths for each topology but chooses the shortest tree instead of the least squares tree as the best tree (Rzhetsky and Nei 1992, 1993). For both of these methods, to guarantee finding the best tree, all possible topologies must be checked. This is generally not feasible, so heuristic searches are used instead. It should be noted that, even if the best topology is found, there is no guarantee that it is also the correct topology.

These methods can be time consuming, and therefore algorithmic approximations to them were developed. Neighbor-joining (Saitou and Nei 1987) is one very

popular algorithm that is also often used to find a starting tree for ML analyses (Yang 2007; Guindon et al. 2010). It is essentially a clustering algorithm which also calculates branch lengths. Neighbor-joining is surprisingly accurate and very fast as no tree search is performed; therefore, it is a popular method for quickly building trees. Comparisons of neighbour-joining and Fitch-Margoliash can be found in Kuhner and Felsenstein (1994).

In general, distance methods are not as accurate as ML as they only use information about pairwise relationships. They are, however, much faster and can be useful for gaining insights into ML as the equations are generally more tractable. I have used this approach in Chapter 3, to study the placement of long branches on three taxon trees.

2.3 Maximum Likelihood for Inferring Phylogenetic Trees and Model Parameter Values

ML is a long-established method for statistical inference which has been extensively tested and successfully applied to a variety of problems (Fisher 1921, 1925; Edwards 1972). It is consistent, meaning that as the amount of data tends to infinity the probability of obtaining the true parameter value tends to one (Wald 1949; Chang 1996; Rogers 1997). This is a desirable property, however it only holds when the model is true, and does not tell us about what may happen under model misspecification.

The likelihood value, L, used in phylogenetic inference problems is the probability of data, usually an alignment, given a tree with branch lengths and the parameters of a model:

$$L = P(Data|tree, model)$$

In phylogenetics, sites are generally assumed to be independent so the likelihood of an alignment is the product of the likelihoods of each site. Due to the very small probabilities involved in these calculations the natural logarithm of the likelihood is usually used. The log likelihood is:

$$\log L = \log(P(Data|tree, model)) = \sum_{i=1}^{n} \log(P(x_i|tree, model))$$

where x_i is column i of an alignment and n is the length of the alignment. The likelihood of data at each site given a tree with branch lengths and a Markov process model of evolution, described in Section 2.1, can be found using Felsen-

stein's pruning algorithm (Felsenstein 1981). The algorithm moves through the tree from the tips to the root calculating the likelihood as it goes, and significantly reduces the computational effort required to calculate the likelihood. For reversible models, as there is no direction on the tree, the root can be placed wherever is most convenient.

The aim of ML methods is to find the parameters that optimise the likelihood for a given data set. These are called the ML estimates (MLEs). Parameters can include the tree topology, branch lengths, and model parameters, but do not necessarily include all of these things. For example, when looking for positive selection the tree topology is often assumed to be known, as it has been found that, provided the tree topology is a decent estimate of the tree, the exact topology does not affect parameter estimates greatly (Yang et al. 1994). In general, a Markov model is chosen and the parameters of the model are found using numerical optimisation (Yang 2006). Markov models are not generally themselves searched over, although methods to do this are being developed (Huelsenbeck and Dyer 2004); instead, model parameter and tree optimisation are performed for each model of interest, and then hypothesis tests are performed to find the best model. Topology search is carried out by finding the optimal branch lengths of a given topology using numerical optimisation, and then rearranging the topology, finding the optimal branch lengths on the new topology, and calculating whether the new topology has a better likelihood than the old topology. This is repeated until no improvement can be made. Topology rearrangement is usually performed either by exchanging nearest neighbours (nearest neighbour interchange) or pruning a subtree and re-grafting it back onto the tree in a new place (subtree prune and re-graft). These methods do not necessarily cover all possible trees and therefore there is no guarantee that the optimal topology will be found (Morrison 2007; Whelan and Money 2010; Money and Whelan 2012).

Branch length optimisation is much faster for reversible models as we can use the 'pulley principle', the ability to calculate the likelihood from any position on the tree and the likelihood be unchanged (Felsenstein 1981). For non-reversible models the direction of evolution along each branch on the tree affects the likelihood, so it is not possible to move the root around arbitrarily. Another implication of the pulley principle is that for a reversible model only an unrooted topology can be inferred, whereas for a nonreversible model inferred topologies are rooted.

There are a variety of programs available to find the ML tree and parameters. In this thesis I use PAML (Yang 2007), PhyML (Guindon et al. 2010), and HyPhy (Kosakovsky Pond et al. 2005).

2.3.1 Information Content

For the majority of phylogenetic inference problems the number of parameters to be inferred depends on the model of interest and the number of sequences. As sequence length increases, the amount of information about the parameters of interest also increases, until, as the sequences tend to infinite length, the probability of obtaining the true parameter values tends to one (in other words, ML is consistent).

In some circumstances, we may wish to infer a parameter for each site. For example, we may be interested in the level of selection, which we believe will vary between sites (see Chapter 5). Increasing sequence length still improves the ability to infer parameters that are shared over all sites, but it does not improve the ability to make the site-wise estimates. For these, a larger tree is required to obtain more information, instead of an increase in sequence length. This tree needs to have either more sequences or a longer total branch length, or potentially both, such that enough substitutions have occurred to accurately estimate the parameter of interest if these substitutions were known, and the individual branch lengths are short enough so that the number of substitutions that have occurred can be inferred sufficiently accurately. If a tree grows in this fashion then, as the number of sequences tends to infinity, the probability of obtaining the true site-wise parameter values will tend to one.

2.4 Hypothesis Testing

Likelihood inference can be used to provide information about the process of evolution by examining the parameter values and comparing models. As well as obtaining point estimates of parameters, ML allows us to obtain confidence intervals around these estimates which describe how sure we are about the value of the estimate (Yang 2006). For a 95% confidence interval, there is a 95% probability that the interval encompasses the real value; large intervals indicate low confidence in the estimate, and small intervals indicate high confidence. These intervals can also be used to test hypotheses. For example if we have a site-wise estimate of ω of 3 ± 1 then we can be confident that that site has $\omega > 1$ and hence is under positive selection. Confidence intervals are calculated by finding the values either side of the ML estimate for the parameter of interest that correspond to the ML value minus half of the likelihood ratio test (see below) value used for determining significance (Silvey 1975; Yang 2006).

Likelihood inference can also be used to compare hypotheses (Wilks 1938). If

we have a null hypothesis H_0, which is nested inside an alternative hypothesis, H_1, such that if certain parameters within H_1 are restricted H_0 is reached, then we can use a likelihood ratio test (LRT) to compare them. The LRT statistic is twice the difference between the maximum log-likelihood of the alternative model and the maximum log-likelihood of the null model:

$$\Lambda = 2(\log(L(H_1)) - \log(L(H_0)))$$

LRTs allow us to assess whether the more complex alternative model, with fewer restricted parameters, significantly improves the model of evolution explaining the data. The more complex model will always give the same or better likelihood, so we are not just looking for whether the model is better, but for whether it is significantly better. Under the null, Λ is χ^2_f distributed, where f is the difference in degrees of freedom of the null and the alternative models (Wilks 1938). The null hypothesis is rejected if the p-value, the probability that a value greater than or equal to Λ would be produced by the χ^2_f distribution, is less than or equal to the significance level, α, of the test.

The significance level, or size, α is also the false positive rate (FPR), the proportion of times the null is rejected when the null is in fact the true model, expected for the test if it is repeated multiple times; α is therefore chosen to be the desired FPR. Another quantity of interest is the power, or the true positive rate (TPR), the proportion of times the null is rejected correctly if the test is repeated multiple times. These two quantities are linked, so that reducing the size also reduces the power of a test. For LRTs, the power is not controlled; often the aim is to develop a test with the most power for a given size. LRTs are used in Chapter 4 to compare reversible and non-reversible models.

The χ^2_f distribution is applicable to LRTs provided the null hypothesis does not place a parameter on the boundary of the possible region in the alternative hypothesis. If parameters are on boundaries then an alternative distribution must be derived (Self and Liang 1987). One example of this is a test for positive selection, comparing $H_0 : \omega = 1$ against $H_1 : \omega > 1$; the null hypothesis clearly has ω on the boundary of the parameter space of the alternative. If the true value conforms to the null then estimates will be distributed around the true value, meaning that if there were no restrictions on ω, half of the estimated values would be less than one, and half would be greater than 1. Due to the restrictions, the first half will have $\omega = 1$ and hence $\Lambda = 0$, whilst the second half will have $\omega > 1$ and Λ will be χ^2_1 distributed. LRT values can therefore be compared with a 50:50 mixture of a point mass at 0 and a χ^2_1 distribution

(Massingham and Goldman 2005). A similar argument and distribution can be used for testing whether there is among site rate variation (Self and Liang 1987; Whelan and Goldman 1999; Goldman and Whelan 2000).

Use of the χ^2 distribution for LRTs is only applicable to nested hypotheses, so if we wish to test non-nested hypotheses then either a distribution needs to be derived for the particular test, or parametric or non-parametric bootstrapping can be carried out (Efron and Tibshirani 1993). Bootstrapping is a method for producing pseudo-replicate data sets which are assumed to follow the distribution of the null model. The technique was originally introduced to phylogenetics by Felsenstein (1985), for assessing the uncertainty in a tree topology. Generally 100 data sets are produced and then analysed in exactly the same way as the original data set. The LRT values for these bootstrap data sets give a distribution of expected values if the null is correct. LRT values of the original data set are compared to this distribution to determine significance. Pseudo-replicates can be produced either by inferring parameters under the null and simulating data with these parameters (parametric bootstrap) or by re-sampling the data with replacement assuming that it conforms to the null (non-parametric bootstrap). In phylogenetics nonparametric bootstraps have been used to test phylogenies (Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999) and compare the fit of amino acid models (Whelan and Goldman 2001), and parametric bootstraps have been used to assess the adequacy of models of DNA sequence evolution (Goldman 1993). In Chapter 5 I develop a new test for positive selection which uses a parametric bootstrap.

Chapter 3

Maximum Likelihood Inference of Long Branches

A paper based on the work presented in this chapter has been accepted for publication in *Systematic Biology* (Parks and Goldman 2014). The version presented here is largely the same as the second revised version of that paper, but with an extended conclusions section.

3.1 Introduction

Amongst the methods for phylogenetic tree reconstruction from molecular sequence data, maximum likelihood (ML) is one of the most popular due to its statistical basis, robustness and the fact that it appears to suffer less from biases. Additionally, ML is known to be a consistent method if the assumed model is correct (Chang 1996; Rogers 1997), meaning that as the amount of data tends to infinity the probability of obtaining the correct tree tends to one. Consistency, however, is not informative about performance of a method with finite data, and with finite data ML can struggle, particularly if long branches are present on the tree. The reasons for this are unknown. ML with the correct model should be able to deal with parallel substitutions and multiple substitutions at sites (Chang 1996), phenomena that occur when branches are long, but despite this it has been reported to be biased towards trees with long branches placed together (Huelsenbeck 1995).

One of the reasons that biases in ML reconstruction (for example, issues caused by long branches) are not well understood is that very few analytical solutions for ML exist, and the solutions that do exist are for small trees and simple models. This means that ML tree reconstruction is generally carried out

using numerical maximisation and heuristics. Yang (2000) derived a set of analytic solutions for a three-taxon tree using two-state characters. Since then further analytic solutions for three-taxon trees with two-state or four-state characters, and four-taxon trees with two-state characters have been derived (Chor et al. 2001; Chor and Snir 2004; Chor et al. 2006a,b; Chor and Snir 2007). All of these studies consider trees with a molecular clock, meaning that biases caused by long tip branches cannot be studied, as it is not possible to have short tip branches joined to long tip branches. Further analytical solutions are required to fully understand long branch biases.

Long branches represent a large amount of evolutionary change for which there are only a few observations. Various effects of long branches on tree reconstruction have been reported, starting with Felsenstein (1978). Felsenstein studied a four-taxon tree with two long branches (P) and three short branches (Q) (Fig. 3.1). He proved that with two-state characters there are combinations of P and Q for which parsimony reconstruction is inconsistent. This region of branch length space is now widely called the Felsenstein zone (Huelsenbeck and Hillis 1993). Since Felsenstein's paper, conditions for inconsistency of parsimony have been extended to any number of character states and five different parameters for branch lengths instead of two (Zharkikh and Li 1992; Schulmeister 2004). Larger trees have also been examined, with further inconsistency conditions found (Kim 1996).

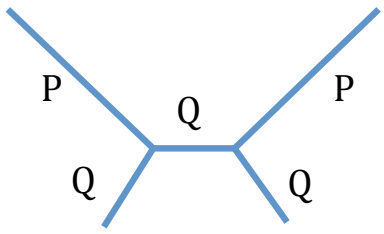


Figure 3.1: Tree used by Felsenstein to show that parsimony could be inconsistent. The short branch length is Q and the long branch length is P.

Following Felsenstein's early work on inconsistency it became widely accepted that such problems were due to 'attraction' amongst long branches. It also became clear that these problems may not be restricted to parsimony only. Numerous simulation studies tested whether the accuracy of other tree reconstruction methods is affected by the presence of two long branches (Huelsenbeck and Hillis 1993; Kuhner and Felsenstein 1994; Gaut and Lewis 1995; Huelsenbeck 1995).

One of the most thorough was carried out by Huelsenbeck (1995). Using the same tree as Felsenstein, but with four-state characters, he tested the consistency, efficiency and robustness of 26 reconstruction methods. This showed that under model misspecification all methods could suffer from inconsistency, and that long branch effects seem to be more of a problem with shorter sequences. It also showed that the presence of long branches does seem to affect ML, although the effects were not as strong as for the other methods investigated.

The term 'Long Branch Attraction' (LBA) has become widely used to describe long branches being incorrectly placed together on a phylogenetic tree. However, LBA is not well-defined and statistical inconsistency, model violation and claims that certain methods are unable to deal with parallelism and convergence have been variously cited as both definitions and explanations (Philippe and Laurent 1998; Sanderson et al. 2000; Anderson and Swofford 2004). Initial studies on LBA were theoretical, with data obtained by simulation. However, after the coining of the term LBA by Hendy and Penny (1989), there was interest in whether it could affect real data. Conclusive biological evidence has been difficult to find because the true tree is never known for real data. However, the publication of a number of papers proposing that LBA can affect real data (Huelsenbeck 1997, 1998) led to LBA being frequently cited as the reason for unexpected phylogenetic results (e.g. Stiller and Hall 1999; Sanderson et al. 2000; Philippe and Germot 2000; Wiens and Hollingsworth 2000; Qiu et al. 2001; Omilian and Taylor 2001; Dacks et al. 2002; Stefanović et al. 2004; Wilcox et al. 2004; Inagaki et al. 2004; Fares et al. 2006; Barros et al. 2008; Dabert et al. 2010; Bodilis et al. 2011; Li et al. 2014). Methods to detect LBA have also been widely discussed and include: finding two long branches together; showing a 'better' method doesn't place the long branches together; showing the branches are long enough to attract by simulation; breaking up a long branch; and removing one of the long branches and reconstructing the tree to see if the other long branch moves (Huelsenbeck 1997; Bergsten 2005; Connor et al. 2010). There is, however, no method that can guarantee a particular topology has been caused by LBA.

In addition to being poorly defined and difficult to locate, the reasons for assuming problems to arise from interactions between multiple long branches, or for naming LBA an 'attraction', are not clear. 'Attraction' implies that there is an interaction between long branches and that this interaction causes them to be placed closer together. However this has never been proven and indeed our knowledge of the problems engendered by long branches is incomplete. In this chapter I aim for a greater understanding of the behaviour of ML tree inference in the presence of individual long branches. I then extend the analysis to the case of

two long branches, looking for any additional effects related to their interaction. To do this I need to distinguish between difficulty in placing long branches and attraction between long branches. If an attraction were to exist then its effects could be interpreted, and hence measured, in different ways. I will define two such ways as 'long branch joining' (LBJ) where long branches are incorrectly joined together on a tree, and 'long branch closeness' (LBC) where long branches are closer together on the reconstructed topology than on the true topology. Knowledge of whether either of these two phenomena occur will lead to a greater understanding of the effects of long branches on tree reconstruction. I will focus on ML with the correct model, which is consistent. I find this more approachable than looking at model-misspecification; with the wrong model anything could happen, but under the correct model ML is expected to perform well.

In this chapter I start by looking at the placement of one long branch by ML. This is important because correct placement of a branch between two nodes is necessary for all tree reconstruction. I use a three-taxon tree as it is the simplest possible tree for reconstruction yet gives interesting and counterintuitive results. Placement of long branches is assessed by simulations followed by ML tree reconstruction for the simulated data sets. The distribution of placement of long branches is then studied using analyses of both ML and distance matrix equations for three-taxon trees. This gives insight into why long branches may cause problems for tree reconstruction, and allows for partial analytical solutions of the four-state character, three-taxon tree without a molecular clock. I then use knowledge about the placement of one long branch to look at the effect of two long branches. Four-taxon trees are used, as the three possible topologies are the simplest that allow me to investigate both LBC and LBJ phenomena. I test for the existence of both LBC and LBJ, allowing me to split any potential 'attraction' into two parts and see which occur. This reveals the complexity of the problem and highlights that further work will be necessary to fully understand it.

3.2 Methods

3.2.1 Evolutionary Models and Trees

This chapter considers nucleotide sequences evolved under Jukes Cantor (JC) evolution (Jukes and Cantor 1969; Yang 2006). This is both the simplest model and shows the properties of ML estimation on which I wish to concentrate. Sequences are simulated without insertions or deletions so alignment is not neces-

sary. Data at each site are assumed to be independent and identically distributed; the order of the sites therefore does not matter, just the counts of each possible nucleotide pattern. Unrooted trees are used as JC is reversible and no molecular clock is assumed; hence a rooted tree cannot be found.

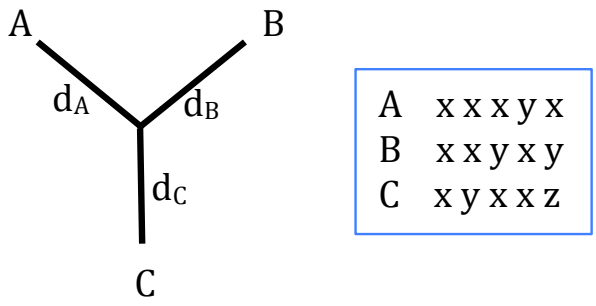


Figure 3.2: Unrooted three-taxon tree with the five possible site patterns when considering Jukes Cantor evolution, where x, y, and z are any three different nucleotides.

For an unrooted three-taxon tree (Fig. 3.2) there are $4^3 = 64$ possible combinations of the nucleotides at a site over the three taxa. These combinations are called site patterns. In the JC model each nucleotide has equal base frequency and mutation rate, meaning that many of these site patterns have the same probability of occurring. In fact, it does not matter which nucleotides are present for different taxa, just whether the nucleotides are different for the different taxa. This means that the site patterns can be reduced to just five patterns of interest, $P = \{xxx, xxy, xyx, yxx, xyz\}$, where x, y and z are any three different nucleotides. The pattern xxx thus represents four possible nucleotide combinations (AAA, CCC, GGG, and TTT), and the remaining patterns represent 12, 12, 12 and 24 nucleotide combinations, respectively. Data can then be represented as counts of these five different patterns from a sequence alignment. For an alignment of length n, these counts will be written as n_r for each pattern $r \in P$, and $\sum_{r \in P} n_r = n$. For a four-taxon tree there are 256 possible site-patterns, which can be reduced to 15 patterns of interest for JC evolution.

3.2.2 Maximum Likelihood

In order to look for analytical solutions, the likelihood function was derived for a three-taxon tree using standard methods (Yang 2006). This derivation is shown in Appendix A.1.

ML tree reconstruction was also conducted using the baseml program from the PAML package (Yang 2007). As small trees are investigated a heuristic search for the ML branch lengths can be performed for each topology individually, and then compared to find the ML tree. Use of a heuristic search means that results may be dependent on the starting values used for branch lengths. Additionally the presence of long branches makes the search more difficult. To improve my ability to find ML values, baseml was run from five different starting points for each analysis, and the ML tree was chosen as the tree with the highest likelihood from these runs. To check that five runs were enough I assessed how often the results would change if only four runs were carried out. The changes were minimal, even for long branch lengths. Baseml was modified to help it find the ML tree when the likelihood was very flat, and to make sure restrictions on branch lengths did not stop it from finding the ML tree. If runs of baseml found trees with different long branch lengths but a very similar likelihood, I hypothesised that the ML tree in fact had an infinite branch length. This was then tested by analytically calculating the likelihood of the tree with an infinite branch length and comparing it with the likelihoods from baseml. A higher analytical likelihood was taken as confirmation that the branch was infinitely long. In this case there is no information about where the branch should be placed on the tree, so any placement made by baseml would be artifactual. Therefore for these trees the branch in question was recorded as being of infinite length and having no meaningful position on the tree.

To test procedures for artefacts, phylogenetic inferences were repeated using PhyML (Guindon et al. 2010). Baseml invariably found either the same tree as PhyML or a tree with a higher likelihood, increasing my confidence in baseml's ML estimates for the analyses needed in this paper. Since baseml and PhyML are optimised for different tasks in phylogenetic inference, no broader conclusions about the merits of the two programs are drawn.

3.2.3 Distance Matrix Equations

Although I do not study performance of DM methods in this paper I find it useful to draw on some of these ideas to help understand the performance of ML methods. Under the JC model, the pairwise distance is $D_{ij} = -\frac{3}{4}\log(1 - \frac{4}{3}U_{ij})$ where U_{ij} is the fraction of bases that differ between the two taxa i and j (Yang 2006). For each pair of taxa, U_{ij} can be written as a sum of pattern counts divided by the sequence length; for example, between taxa A and B of Figure 3.2, $U_{AB} = (n_{xyz} + n_{xyx} + n_{yxx})/n$. If $U_{ij} \geq 0.75$ then the distance between the

two taxa is infinite, so for a finite data set there is a maximum distance between two taxa that can be measured before the two taxa are estimated to be infinitely far apart.

On an unrooted three-taxon tree minimum evolution, neighbor-joining and both weighted and unweighted least squares methods result in the same branch lengths, as the distances can be exactly fit to the tree. The branch lengths are:

$$d_A = \frac{(D_{AB} + D_{AC} - D_{BC})}{2} \quad d_B = \frac{(D_{AB} + D_{BC} - D_{AC})}{2} \quad d_C = \frac{(D_{AC} + D_{BC} - D_{AB})}{2} \tag{3.1}$$

These calculations can result in negative branch lengths which are not biologically meaningful. Some software therefore require a positivity constraint in order to guarantee results that are meaningful in a phylogenetic context.

3.2.4 Simulations

For three-taxon trees simulations were run under JC evolution producing 5000 data sets of 300bp sequences, unless otherwise stated. This is a realistic sequence length for a small protein, and allows me to look at how ML works for limited data. For four-taxon trees sequence length was increased to 1000bp due to the use of two long branches. All simulations were conducted using evolver from the PAML package (Yang 2007).

3.3 One Long Branch on Three-Species Trees

3.3.1 ML Inference

To explore the placement of one long branch on a tree I simulated data from a three-taxon unrooted tree (Fig. 3.2) with a long branch, and constructed and examined trees inferred from this simulated data. The three-taxon case is used as it is the simplest possible; there is only one topology so the only inference question is the branch lengths. Six different branch lengths were used for d_C ($d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$). d_A and d_B were set to 0.1 to make the distance from A to B easy to estimate (Fig. A.1) so that I could concentrate on placement of the long branch. Estimation of d_C also behaves as expected, getting harder as d_C increases (Fig. A.2). Unexpected results come from looking at the position of where the branch to C joins the A–B path (Fig. 3.3). The placement of C is measured as a fraction along the A–B path. If C is placed on one end of the A–B path, so that the branch to A has length 0 ($d_A = 0$), then C is measured as

being at 0 on the A–B path; if C is placed on the other end, and $d_B = 0$, then C is measured at 1. Trees with inferred infinite branch lengths are not included in these plots.

When d_C is of the same length as the other branches ($d_C = 0.1$) then tree reconstruction is accurate and C is distributed around its original position. As d_C increases the distribution spreads over the A–B path and, counterintuitively, starts to accumulate at the edges of the A–B path and in the centre. For long d_C, I expected the placement of C to be uniform over the A–B path, reflecting the lack of information about the relationship between C and the other taxa, and that if there was a peak it would be gradual and centered. This was not seen here.

Note that for these simulations d_A and d_B were kept constant. The same effect is seen for other values of d_A and d_B, although the precise values of d_C needed for the effect to become apparent depends on d_A and d_B (results not shown). The effect is also present for all finite values of n; as n increases the effect is less for any given combination of d_A, d_B, and d_C but it can again be made to appear by increasing d_C. Figure 3.4 shows the proportion of data sets giving trees with branch lengths of zero for increasing d_C lengths and different sequence lengths. For a longer sequence length ($n = 1000$) the proportion of data sets giving trees with branch lengths of zero for a given value of d_C is lower than for $n = 300$; for a shorter sequence length ($n = 100$) it is higher. ML is however consistent under the correct model so for any finite d_A, d_B, and d_C, as $n \to \infty$ the estimates will tend towards the correct values and the effect will disappear.

Faced with the counterintuitive results of Figure 3.3, my next goal is to explain these distributions. First I will concentrate on the feature of many of the reconstructed trees having $d_A = 0$ or $d_B = 0$ when d_C is large. To understand this I need to know the features of data sets that cause trees with zero branch lengths. I use DM methods as an initial approach, followed by an analysis of the ML equations. Combining these two approaches allows me to find maxima for the ML equations with zero or infinite branch lengths, and predict quite accurately when these will be global maxima. This means that for a given data set I can predict if the tree will have a zero or infinite branch length; for trees where I predict this I can also derive the branch lengths of the other branches.

3.3.2 Distance Matrix Analysis

The simulated data sets were analysed using DM methods because DM equations can be easy to interpret and may give intuition into the behaviour shown in Figure

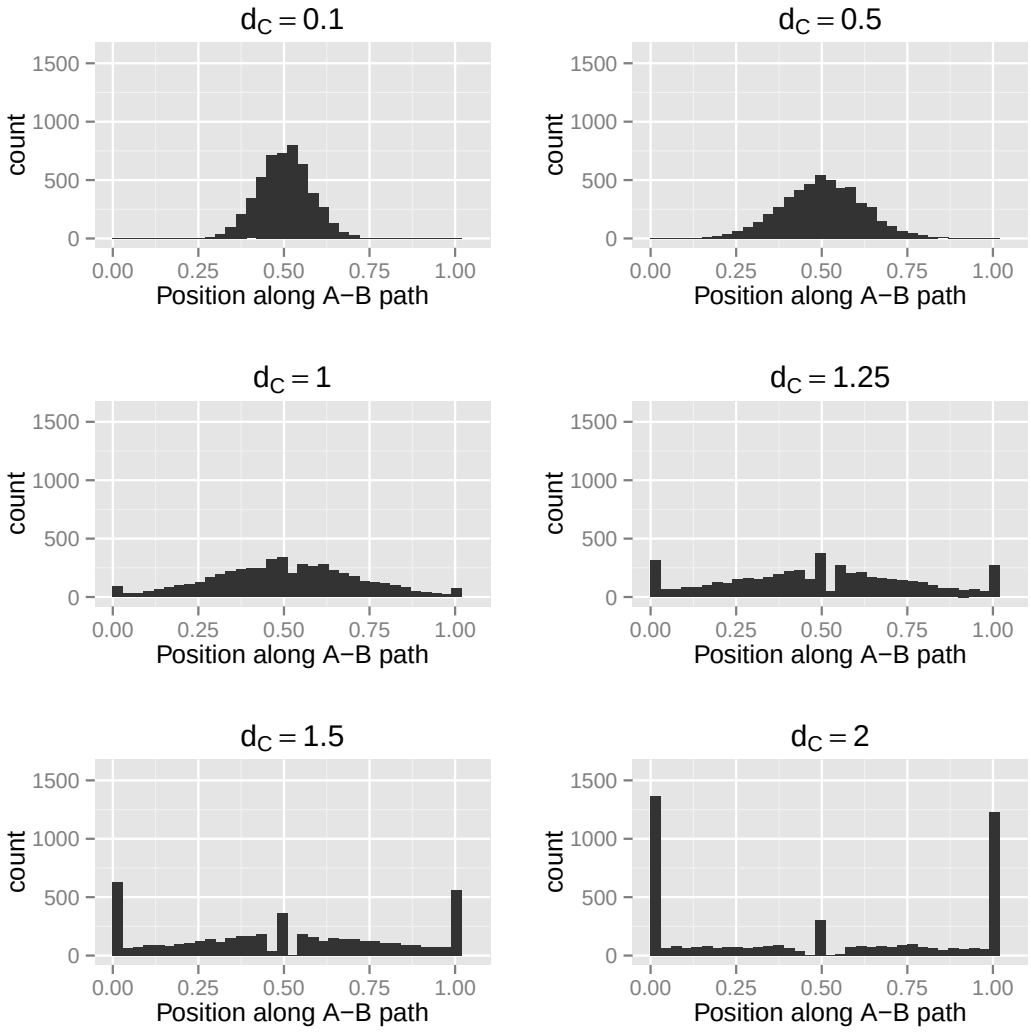


Figure 3.3: Distributions of the location of the branch leading to C on the A–B path for trees simulated with $d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$. For each value of d_C, 5000 data sets were run; those that produced a tree with a predicted infinite branch length are not plotted: this corresponds to 0, 0, 0, 0, 1, and 92 data sets, respectively. The distributions of d_C and $d_A + d_B$ along with plots of the position of C against d_C and $d_A + d_B$ are shown in Figures A.1:A.4.

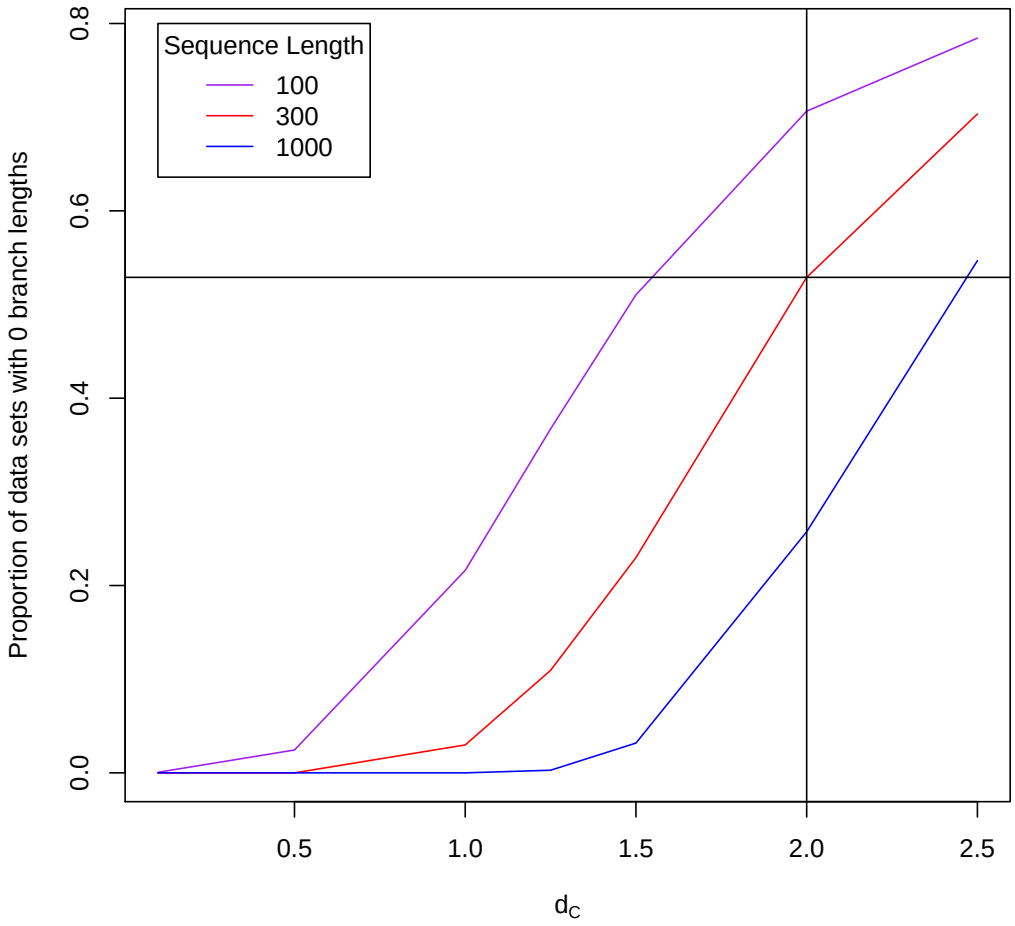


Figure 3.4: The proportion of data sets giving trees with branches of length 0 for increasing d_C lengths and different sequence lengths. As sequence length increases the proportion of data sets giving trees with branches of length 0 decreases for a given length of d_C. In the main text I consider sequence lengths of 300 and d_C lengths up to 2. For the longer sequence length of 1000, at $d_C = 2$, the proportion of data sets giving trees with branches of length 0 is much lower than for the sequence length of 300; increasing d_C to 2.5 however increases y back to the level it is at with $n = 300$ and $d_C = 2$. For the shorter sequence length of 100, the same proportion of data sets giving trees with branches of length 0 as for $n = 300$ and $d_C = 2$ is reached by the time $d_C = 1.6$.

3.3. Equation 3.1 gives the branch lengths of the three-taxon tree obtained using DM methods. One of branch lengths is zero or negative if the triangle inequality is violated and one of the following conditions holds:

$$D_{BC} \geq D_{AB} + D_{AC} \qquad D_{AC} \geq D_{AB} + D_{BC} \qquad D_{AB} \geq D_{AC} + D_{BC} \tag{3.2}$$

These conditions can be used as predictors for ML results by calculating pairwise distances for each data set from its pattern count data (as explained in Methods) and checking if the inequalities given above hold. If one of the inequalities holds then one of the branch lengths is less than or equal to zero for the DM method and I predict that the branch length will be zero for ML. Figure 3.5 shows a version of Figure 3.3 where the data sets with predicted zero branch lengths are plotted in blue and the remaining data sets are in red. This shows that the accuracy of the conditions is high. Accuracy will be more fully examined later.

Some inferred trees have infinite branch lengths, making placement of taxon C impossible. Therefore I am also interested in identifying trees with infinite branch lengths from DM analyses. Pairwise distances are infinite if $U_{ij} \geq 0.75$ (see Methods). If exactly one pairwise distance is infinite then one of the conditions shown above (Equation 3.2) holds. This means that with DM methods there will be one negative branch length and two infinite branches (Equation 3.1). By comparing this with ML results I find that this corresponds to cases where the ML tree has one zero branch length, and finite lengths for the other branches. This can therefore be included as a case where a zero branch length is predicted if one of the conditions above (Equation 3.2) holds.

If two pairwise distances are infinite, for example D_{AC} and D_{BC}, then there can be no knowledge about the placement of one of the taxa, here C, so the length of its branch will be infinite. For any taxon X, if the other two taxa are Y and Z, then I would expect the branch to X to be infinite if D_{YX} and D_{ZX} are infinite. If three pairwise distances are infinite then there can be no knowledge of the relationship of any of the taxa so at least two of the branch lengths should be infinite. This gives conditions for infinite branches, which again can be used as predictors for ML results. All predictors are shown in Table 3.1.

The accuracy of these DM-based predictors of ML behaviour was tested using simulation, comparing ML results with predictions made from the count data. I simulated 5000 data sets from the tree in Figure 3.2 with $d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$ and $d_A = d_B = 0.05, 0.1, 0.2, 0.3$. The values for d_A and d_B were again chosen to exhibit a range of lengths where estimation would be relatively easy. In these

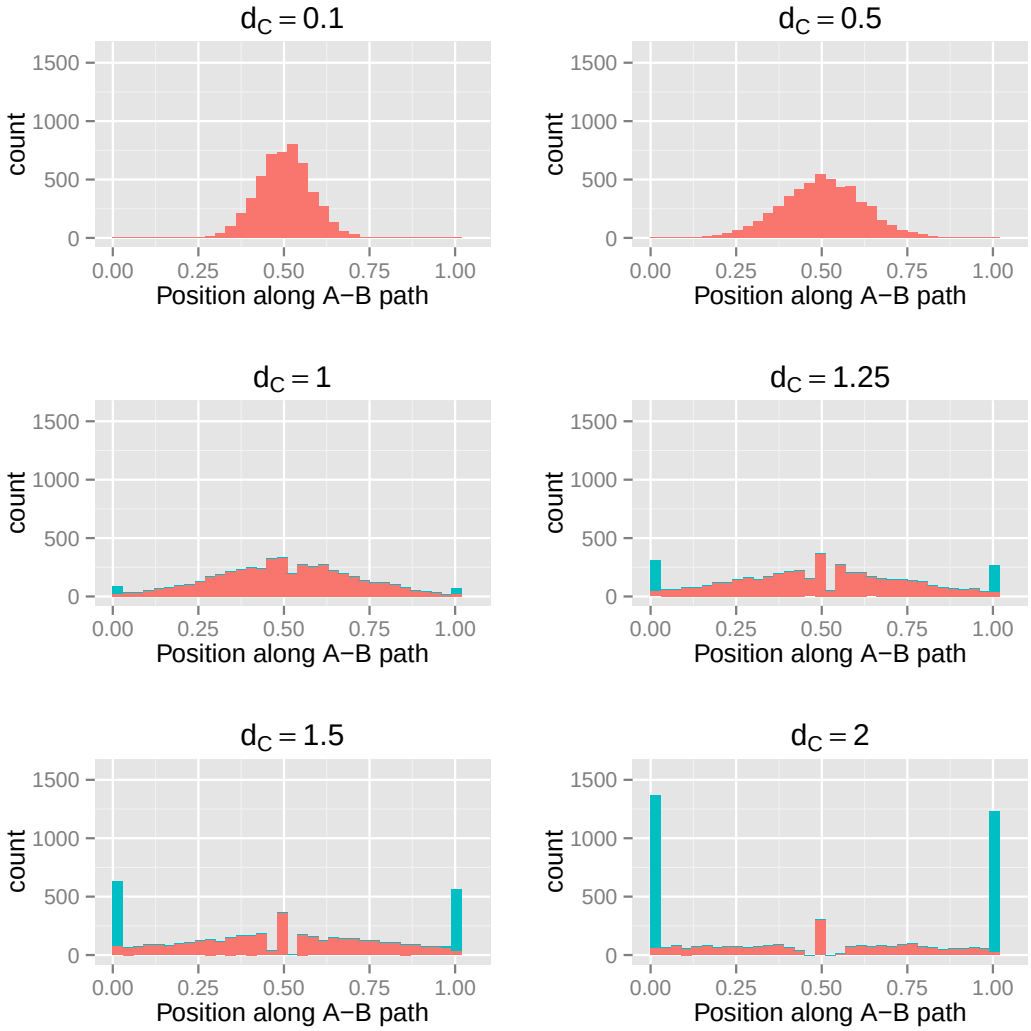


Figure 3.5: Stacked histogram showing distributions of the location of the branch leading to C on the A–B path for trees with $d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$. The distributions are the same as in Figure 3.3, but have been split so trees predicted to have zero branch lengths are coloured in blue, and the remaining trees are in red. Incorrect predictions are those that are blue but not located at 0 or 1 on the x-axis, or red and located at 0 or 1.

Table 3.1: Predictions for branch lengths of the ML tree using pairwise distances.

Conditions	Prediction
$D_{BC} \geq D_{AB} + D_{AC}$ (incl. $D_{BC} = \infty$)	$d_A = 0$
$D_{AC} \geq D_{AB} + D_{BC}$ (incl. $D_{AC} = \infty$)	$d_B = 0$
$D_{AB} \geq D_{AC} + D_{BC}$ (incl. $D_{AB} = \infty$)	$d_C = 0$
$D_{AB} = \infty$ & $D_{AC} = \infty$	$d_A = \infty$
$D_{AB} = \infty$ & $D_{BC} = \infty$	$d_B = \infty$
$D_{AC} = \infty$ & $D_{BC} = \infty$	$d_C = \infty$
$D_{AB} = \infty$ & $D_{AC} = \infty$ & $D_{BC} = \infty$	At least two of the branch lengths are infinite

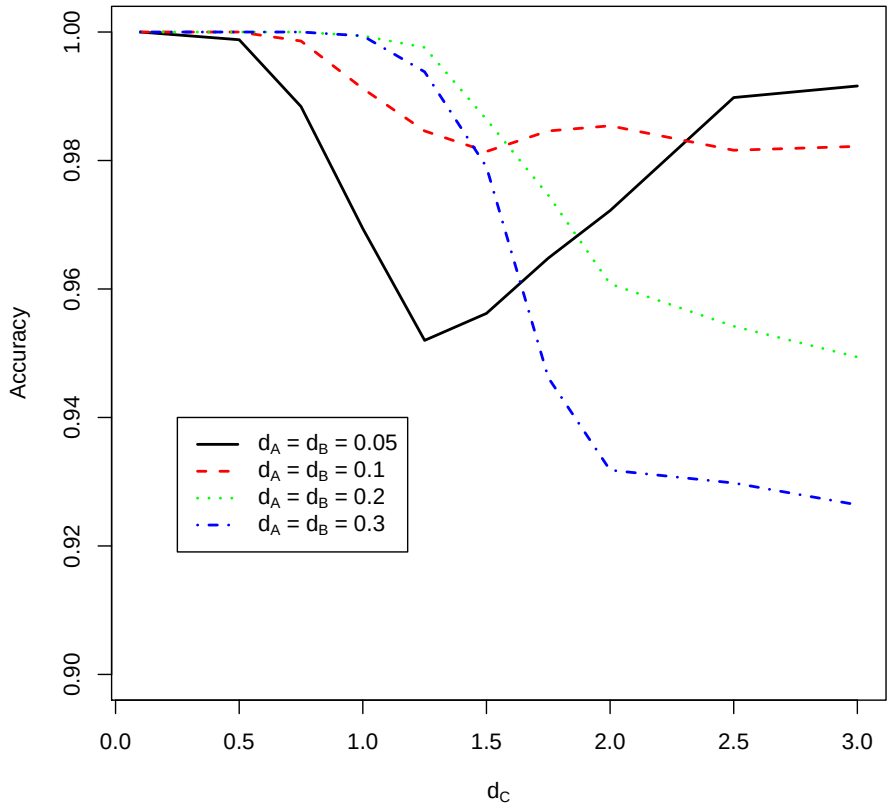


Figure 3.6: The accuracy of DM conditions for predicting zero branch lengths on ML trees for different long branch lengths. Four different lengths of A–B have been used, with $d_A = d_B$ throughout. Accuracy is defined as the proportion of true results, i.e. the number of true positives and true negatives divided by the total number of results.

simulations the DM conditions for infinite branch lengths matched ML with 100% accuracy. The accuracy for the zero branch length DM conditions is shown in Figure 3.6. These conditions are at least 95% accurate for all simulations apart from $d_A = d_B = 0.3$ where they remain more than 90% accurate.

Zero-length branches can be explained by noting that with long branch lengths data frequently has $|D_{BC} - D_{AC}| \geq D_{AB}$. This occurs because estimates of D_{BC} and D_{AC} have high variance if d_C is large. This then leads to inference of a zero branch length.

The good prediction accuracy suggests that the DM conditions are closely related to ML inference. The next section attempts to derive analytic ML solutions that would give perfect understanding of our counterintuitive findings.

3.3.3 Maximum Likelihood Analysis

Branch lengths can be derived by finding the global maximum of the likelihood equation. One approach to do this is to find all of the local maxima and compare their values to find the greatest. I have not been able to achieve this due to the complexity of the ML equations. However, I have been able to find all the local maxima with zero or infinite branch lengths. I can then compare the likelihoods to find the greatest, and using the DM results I can then predict when this result is the global maximum. This allows me to predict not only if there is a zero or infinite branch length, but also the other branch lengths on the tree.

The ML equation for a three-taxon tree is a function of the five pattern counts and the three branch lengths (see Appendix A.1, Equation A.2). My aim is to find the three optimal branch lengths for a given set of pattern counts. The solution space of the ML equation is therefore a three-dimensional region with each dimension representing a branch length. Branch lengths are restricted to be non-negative, so the boundaries of the region occur when one or more of the branches are either zero or infinite. The space representing all solutions with any zero or infinite branch lengths is therefore the surface of a convex polyhedron, which has been made compact (i.e. closed and bounded) by the addition of points at infinity, from now on described as a cube, giving 26 regions (8 points, 12 lines and 6 planes) to investigate. Figure 3.7 illustrates this as a cube where finite boundaries have been drawn to represent ∞ for ease of understanding. The interior of the region represents all cases where each of d_A, d_B and d_C is positive and finite.

To solve for local maxima of the likelihood function at the boundaries, I restrict the ML equations to each of the points, lines or planes on the surface of the cube and solve for maxima in each region. Standard methods were used to solve for maxima (Luenberger 1984); the derivations of all of the possible maxima on boundaries are shown in Appendix A.1. Because I have not found a solution for all maxima in the interior of the cube I cannot in general determine whether each maximum will be a local or global maximum; to do this I would have to compare the likelihood values of all the maxima, including any in the interior. However, in some special cases I am able to determine the global maximum, and these are detailed in Table 3.2. The rest of the local maxima are detailed in Table 3.3.

These results correspond to the peaks at the edge of the distributions shown in Figure 3.3, but they do not account for the peak in the middle of the distribution, or the gap around it (clearest when $d_C = 1.5$ or 2). To explain this we need to

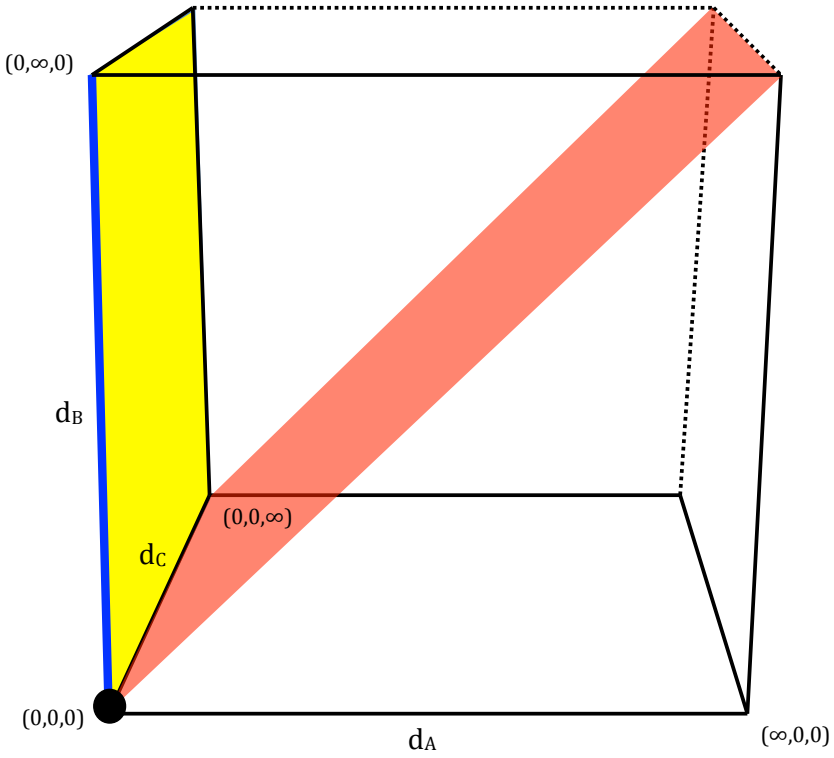


Figure 3.7: The solution space of the ML equation is an infinitely bounded convex polyhedron. One point (black), one line (blue), one surface plane (yellow), the interior plane $d_A = d_B$ (red), and three lines where two variables are at ∞ (dotted line) are highlighted; when the ML equation is restricted to regions such as these analytical solutions can be found for local maxima.

consider the red plane, $d_A = d_B$, in Figure 3.7. If we require $d_A = d_B$ then it is possible to find an optimum which corresponds to $n_{xyx} = n_{yxx}$. As two of the branch lengths are equal this tree is now equivalent to a three-taxon tree with a molecular clock, so the branch lengths can be derived from the solution given in Chor et al. (2006a). Examining the ML simulations shows that all of the data sets in the peak in the middle of the plots have $n_{xyx} = n_{yxx}$, and that if $n_{xyx} = n_{yxx}$ then the branch to C either falls exactly in the middle or on the edges of the A–B path (Fig. A.5). This corresponds to the optimum at $d_A = d_B$ being either a maximum or a minimum. In comparison, if n_{yxx} and n_{xyx} differ then there are a variety of places where this branch can be placed. From this it can be deduced that the gap seen on the distribution is due to the fact that if the data are symmetric then C can either be placed in the middle or on the edge, whereas when data are not symmetric there are many more options for placement of C.

Table 3.2: Global maxima of the ML equations on the boundaries of the solution space.

Conditions		(d_A, d_B, d_C)	Likelihood Value
$n_{xxx} = n$		(0,0,0)	$-n\log(4)$
$n_{xyz} = n_{xyx} = n_{yxx} = 0$	$n_{xxx} \leq n/4$	$(0,0,\infty)$	$-n\log(16)$
	$n_{xxx} > n/4$	$(0,0,-\frac{3}{4}\log(\frac{4n_{xxx}-n}{3n}))$	$n_{xxx}\log(\frac{n_{xxx}}{4n}) + (n-n_{xxx})\log(\frac{n-n_{xxx}}{12n})$
$n_{xyz} = n_{yxx} = n_{xxy} = 0$	$n_{xxx} \leq n/4$	$(0,\infty,0)$	$-n\log(16)$
	$n_{xxx} > n/4$	$(0,-\frac{3}{4}\log(\frac{4n_{xxx}-n}{3n}),0)$	$n_{xxx}\log(\frac{n_{xxx}}{4n}) + (n-n_{xxx})\log(\frac{n-n_{xxx}}{12n})$
$n_{xyz} = n_{xxy} = n_{xyx} = 0$	$n_{xxx} \leq n/4$	$(\infty,0,0)$	$-n\log(16)$
	$n_{xxx} > n/4$	$(\frac{3}{4}\log(\frac{4n_{xxx}-n}{3n}),0,0)$	$n_{xxx}\log(\frac{n_{xxx}}{4n}) + (n-n_{xxx})\log(\frac{n-n_{xxx}}{12n})$

All results shown so far are for the JC model. Studies on real data generally use a more complicated model such as the GTR model (Tavaré 1986). The simulations and tree reconstructions described above have been repeated using the GTR model with realistic parameters (Murphy et al. 2001) (Fig. A.6). Again for long branch lengths many trees have zero branch lengths. However, there is no sharp peak and gap in the middle of the A–B path; I conclude that this is caused by the symmetric nature of the JC model, which is not present in the GTR model.

3.3.4 Combined ML and DM Analysis

Combining the ML and DM analyses allows for a more complete understanding of the distributions in Figure 3.3. DM analysis has allowed me to predict whether the tree will have an infinite or zero branch length; in these cases, ML analysis can be used to derive the other branch lengths of the tree. Therefore a possible workflow is as follows (Fig. 3.8): first, check for the known global maxima. If none of these is found then DM analysis can be used to predict whether the tree has a zero or infinite branch length (to the described accuracy in Fig. 3.6). If a zero or infinite branch length is predicted then the relevant ML solution can be used to find it. Otherwise a numerical optimisation program must be used to find the global maximum.

Table 3.3: Local maxima of the ML equations on the boundaries of the solution space.

Conditions	Optimum	(d_A, d_B, d_C)	Likelihood Value
$n_{xxx}+n_{xxy} > n/4$, $n_{xxx}+n_{xyx} > n/4$ $\frac{\partial L}{\partial d_A} \leq 0$	Local Max	$(0, -\frac{3}{4}\log(1-\frac{4(n-n_{xxx}-n_{xxy})}{3n}), -\frac{3}{4}\log(1-\frac{4(n-n_{xxx}-n_{xyx})}{3n}))$	$n_{xxx}\log(\frac{(n_{xxx}+n_{xxy})(n_{xxx}+n_{xyx})}{4n^2})+$ $n_{xxy}\log(\frac{(n_{xxx}+n_{xxy})(n-n_{xxx}-n_{xyx})}{12n^2})+$ $n_{xyx}\log(\frac{(n_{xxx}+n_{xyx})(n-n_{xxx}-n_{xxy})}{12n^2})+$ $(n-n_{xxx}-n_{xxy}-n_{xyx})$ $\log(\frac{(n-n_{xxx}-n_{xyx})(n-n_{xxx}-n_{xxy})}{36n^2})$
$n_{xxx}+n_{xxy} > n/4$, $n_{xxx}+n_{yxx} > n/4$ $\frac{\partial L}{\partial d_B} \leq 0$	Local Max	$(-\frac{3}{4}\log(1-\frac{4(n-n_{xxx}-n_{xxy})}{3n}), 0, -\frac{3}{4}\log(1-\frac{4(n-n_{xxx}-n_{yxx})}{3n}))$	$n_{xxx}\log(\frac{(n_{xxx}+n_{xxy})(n_{xxx}+n_{yxx})}{4n^2})+$ $n_{xxy}\log(\frac{(n_{xxx}+n_{xxy})(n-n_{xxx}-n_{yxx})}{12n^2})+$ $n_{yxx}\log(\frac{(n_{xxx}+n_{yxx})(n-n_{xxx}-n_{xxy})}{12n^2})+$ $(n-n_{xxx}-n_{yxx}-n_{xxy})$ $\log(\frac{(n-n_{xxx}-n_{yxx})(n-n_{xxx}-n_{xxy})}{36n^2})$
$n_{xxx}+n_{xyx} > n/4$, $n_{xxx}+n_{yxx} > n/4$ $\frac{\partial L}{\partial d_C} \leq 0$	Local Max	$(-\frac{3}{4}\log(1-\frac{4(n-n_{xxx}-n_{xyx})}{3n}), -\frac{3}{4}\log(1-\frac{4(n-n_{xxx}-n_{yxx})}{3n}), 0)$	$n_{xxx}\log(\frac{(n_{xxx}+n_{yxx})(n_{xxx}+n_{xyx})}{4n^2})+$ $n_{xyx}\log(\frac{(n_{xxx}+n_{xyx})(n-n_{xxx}-n_{yxx})}{12n^2})+$ $n_{yxx}\log(\frac{(n_{xxx}+n_{yxx})(n-n_{xxx}-n_{xyx})}{12n^2})+$ $(n-n_{xxx}-n_{yxx}-n_{xyx})$ $\log(\frac{(n-n_{xxx}-n_{yxx})(n-n_{xxx}+n_{xyx})}{36n^2})$
$n_{xxx}+n_{yxx} > n/4$	Local Max or Local Min	(∞, a, b) where $a+b=-\frac{3}{4}\log(\frac{4(n_{xxx}+n_{yxx})-n}{3n})$	$(n-n_{xxx}-n_{yxx})\log(\frac{n-n_{xxx}-n_{yxx}}{48n})+$ $(n_{xxx}+n_{yxx})\log(\frac{n_{xxx}+n_{yxx}}{16n})$
$n_{xxx}+n_{xyx} > n/4$	Local Max or Local Min	(a, ∞, b) where $a+b=-\frac{3}{4}\log(\frac{4(n_{xxx}+n_{xyx})-n}{3n})$	$(n-n_{xxx}-n_{xyx})\log(\frac{n-n_{xxx}-n_{xyx}}{48n})+$ $(n_{xxx}+n_{xyx})\log(\frac{n_{xxx}+n_{xyx}}{16n})$
$n_{xxx}+n_{xxy} > n/4$	Local Max or Local Min	(a, b, ∞) where $a+b=-\frac{3}{4}\log(\frac{4(n_{xxx}+n_{xxy})-n}{3n})$	$(n-n_{xxx}-n_{xxy})\log(\frac{n-n_{xxx}-n_{xxy}}{48n})+$ $(n_{xxx}+n_{xxy})\log(\frac{n_{xxx}+n_{xxy}}{16n})$
-	Local Max or Local Min	(∞,∞,∞)	$-n\log(64)$

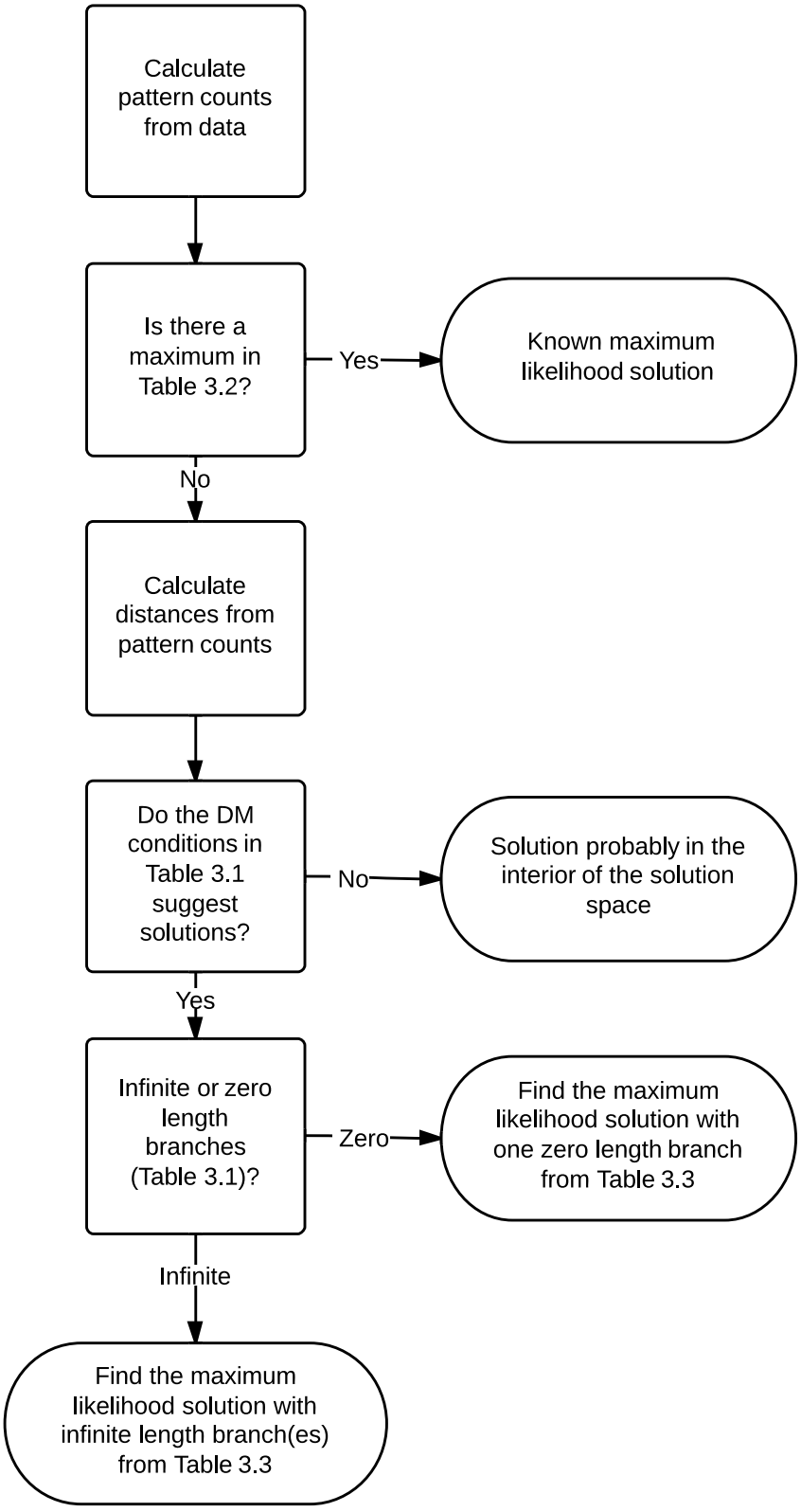


Figure 3.8: Workflow for using the ML and DM results to find the maximum likelihood solution for a three-taxon tree.

3.3.5 Conclusions

Combining the analyses, Figure 3.3 can now largely be explained. This explanation can be used to split the results into separate subsets, as in Figure 3.5. The conditions given can be used to predict which subset a new data set will belong to (Fig. 3.8). An intuitive explanation can also be constructed for the trees with zero-length branches. By comparison with DM methods it can be seen that trees would be reconstructed with negative branch lengths. However, ML tree reconstruction does not permit negative branch lengths and hence trees are instead given zero branch lengths in these cases. These negative branch lengths are obtained because of the high variance involved in estimating long branch lengths.

I further analysed whether the variance involved in estimating long branches could explain this phenomenon. For distance methods it is possible to estimate the variance of the estimates of d_A, d_B, and d_C as a function of the sequence length and the three branch lengths (see Appendix A.2). I am most interested in the first two of these, as these are the ones most often inferred as zero.

If d_A is assumed to be normally distributed then it is possible to estimate the proportion of times that d_A is inferred to be less than or equal to zero. The same analysis can be repeated for d_B, comparing the estimated proportions with the proportion of times that either DM or ML methods inferred that d_A or d_B was zero (Table 3.4). These predictions are close to the values for both DM and ML, and are slightly closer to the DM values. This is expected as they are derived from the variance of the distance estimates. The predictions tend to be slightly smaller than the proportions found in the simulations. This could be because of the approximations in the derivation of the variance (see Appendix A.2), or alternatively it could indicate that the distribution is not quite normal. This would not be surprising as, although the counts of differences between sequences may well be normally distributed, the Jukes Cantor distance involves a subsequent logarithmic transformation.

Table 3.4: Proportion of trees with zero branch lengths for different methods

d_C	Predicted	Found using DM	Found Using ML
0.1	0	0	0
0.5	0.0002	0	0
1	0.0224	0.0262	0.0264
1.25	0.0842	0.0998	0.1034
1.5	0.1996	0.2192	0.2202
2	0.4930	0.5064	0.5220

In summary, analysis of the variance of individual branch length estimates

is able to give a good prediction of the frequency of occurrence of zero-length branches, suggesting that this could be an important explanatory factor.

3.4 Four species tree

Long branch attraction (LBA) is normally discussed when an (unexpected) topology with two long branches grouped together is obtained following tree reconstruction. This means LBA is generally only considered for trees with two long branches where there are multiple different possible topologies. To allow analysis of these situations, I now focus on four-taxon trees with two long branches. Two different forms of LBA have already been defined: long branch closeness (LBC) and long branch joining (LBJ). These will now be investigated to gain an insight into what any 'attraction' might be.

3.4.1 LBC

LBC is defined as long branches being closer together on the constructed topology than on the true topology. To investigate this I simulated four-taxon data sets from the tree in Figure 3.9a and applied ML to reconstruct the two three-taxon trees in Figure 3.9b, and the best four-taxon tree (one of Fig. 3.9c–f). This allows me to assess how the placement of a long branch is affected by the presence of another long branch. On the three-taxon trees only one long branch is present so no attraction could have occurred.

If there were an attraction then I would expect the long branches (Y and Z) to be closer on the four-taxon tree than on the three-taxon tree. To investigate this the relative position of Y and Z on the inferred trees has been calculated. To find the relative position on the three-taxon trees the position of the branches to Y and Z are calculated as fractions along the W–X path of their respective trees, as previously; the relative position, x, is then the difference between these two fractions (Fig. 3.9b). For each four-taxon tree the positions are again calculated for Y and Z as fractions for each topology and the relative position y is recorded (Fig. 3.9c–f). For topologies 3.9d and 3.9f, $y = 0$ is recorded as the branches to Y and Z fall in the same place on the W–X path. All simulations were performed as described in Section 3.2.4. The length of the W–X path is kept constant at 0.1 with Y and Z evenly spaced between W and X.

Figure 3.9g shows distributions of the relative position of Y and Z for the three-taxon trees (x-axis) against that for the four-taxon tree (y-axis) when the length of the branches to Y and Z is 1.5. The points are coloured according to

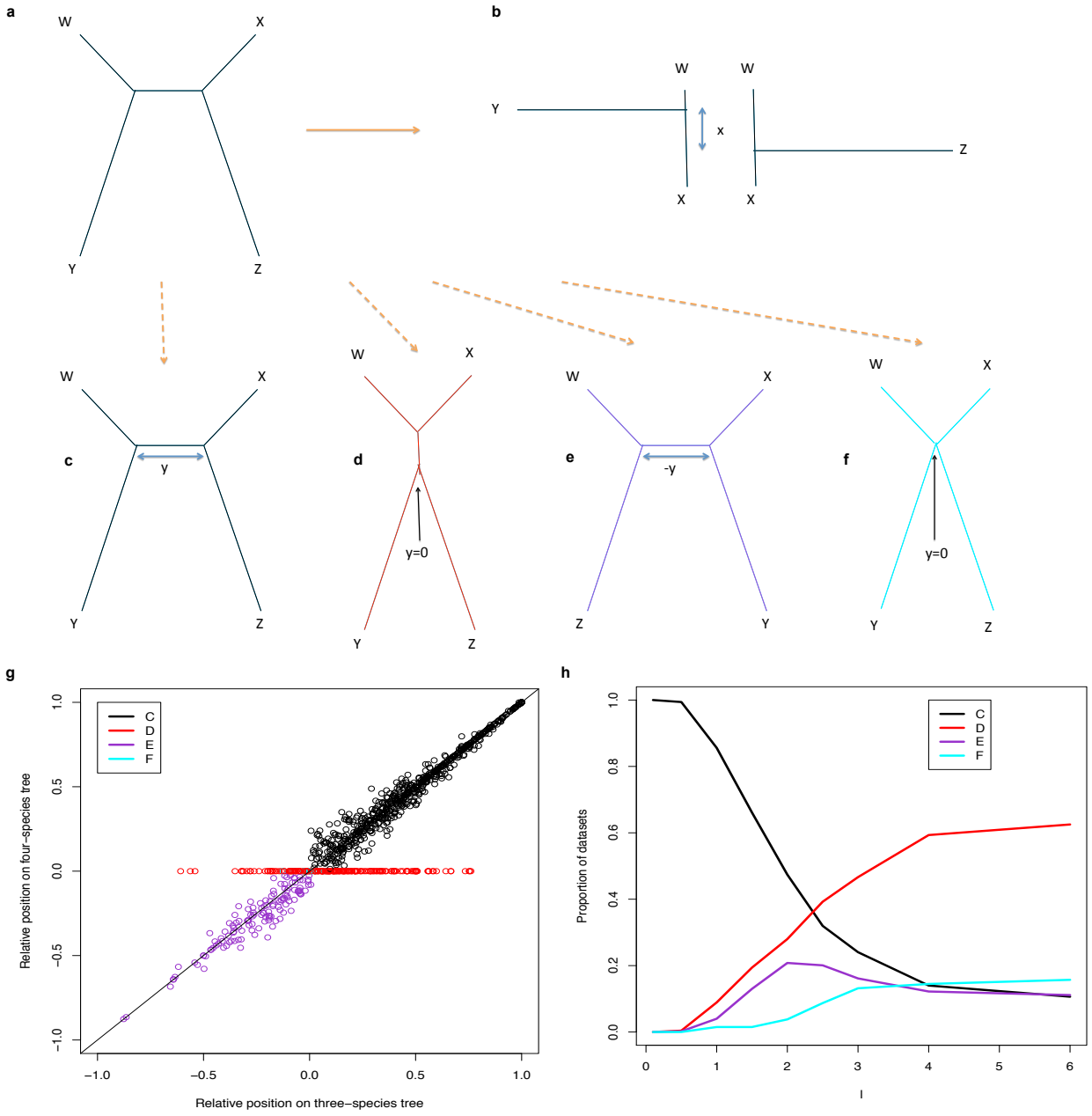


Figure 3.9: **a** The four-taxon tree used for simulations. The path between W and X is always of length 0.1 with Y and Z evenly spaced along it. The simulated data are used to construct the ML three-taxon trees (W,X,Y) and (W,X,Z), **b**, and the ML four-taxon tree (one of **c**–**f**). Distances x and y, as indicated in **b**–**f**, measure the inferred distance between the branches to taxa Y and Z. **g** The relative position of Y and Z on the W–X path on the three-taxon trees (x-axis) versus that on the optimal four-taxon tree (y-axis). Lengths of 1.5 are used for branches to Y and Z; equivalent results are seen for other lengths. **h** The proportions of different topologies obtained for different lengths of Y and Z.

the topology of the inferred ML four-taxon tree. Also indicated is the line $x = y$; points on this line have the same relative position on the three and four-taxon trees. If topology 3.9c, the correct topology, underwent LBC then the black points would lie below this line. Similarly, the points for topology 3.9e, a wrong topology with the long branches not joined to one another, would lie above this line. As can be seen these points are not distributed as would be expected for LBC; in fact there is a small asymmetry in the opposite direction to that which would be expected under LBC. This shows that the branches do not get closer together; if anything they get slightly further apart. This asymmetry becomes significant (Binomial, $p < 0.05$) for topology 3.9c once the long branches are of length 1.5. For topology 3.9e this asymmetry is significant (Binomial, $p < 0.05$) earlier, at a branch length of 0.75. For topologies 3.9c and 3.9e the positions of Y on the three-taxon tree are also very similar to their positions on the four-taxon tree (Fig. A.7; correlations of 0.96 and 0.97, respectively). The equivalent can be shown for the position of Z (results not shown). These results clearly show that for topologies 3.9c and 3.9e there is no attraction and no LBC occurs. This is the case for any Y and Z lengths (results not shown). I have also explored the possibility that instead of long branches becoming closer together, short branches become closer together. This can be analysed analogously to LBC, and it can be shown that there is also no Short Branch Closeness (results not shown).

3.4.2 LBJ

LBJ is defined as long branches being incorrectly joined to one another on a tree. To investigate this I measured the proportion of different ML topologies for different long branch lengths (Fig. 3.9h). For short branch lengths the results are as expected with the majority of the data sets having the correct topology. As the long branch length increases the proportion of the correct topology (3.9c) decreases, and the proportions of the other topologies increase, with the topology with the long branches placed together (3.9d) increasing in proportion more than topology 3.9e. For branch lengths longer than 2, topology 3.9d continues to increase whereas topology 3.9e starts to decrease. Finally topology 3.9d levels off at ∼60% of the trees with all the other topologies levelling off at ∼13%. This shows that for very long branch lengths there is a strong bias towards placing the long branches together, and that for infinite branch lengths instead of getting each topology chosen randomly, topology 3.9d would be chosen over half of the time. This shows that LBJ is occurring.

The details of these results are dependent on both sequence length and the

length of the W–X path. If sequence length is increased then longer branch lengths are required to see the patterns shown here: however, with long-enough branch lengths they will still occur. However, for any length of branch to Y and Z, if sequence length is increased enough then the correct topology will be reached 100% of the time, as ML phylogenetic inference is consistent. The final proportions of the topologies are dependent on the length of the W–X path; however, the existence of the bias is not removed by changing the W–X path length.

As with the three-taxon tree problem, the simulations and tree reconstructions shown above have been repeated using the GTR model with realistic parameters (Murphy et al. 2001) (Fig. A.8). Again LBC does not occur (results not shown) but for long branch lengths LBJ does occur. However, longer branch lengths are required for LBJ to occur with GTR than with Jukes Cantor. This is probably because, although on average the bases are mutating at the same rates, in the GTR model some rates will be slower than average, and some faster. This means that saturation will not be reached by all sites at the same time, so at long branch lengths there will still be information about the tree in some of the sites. Connecting this with the concept of effective sequence length (Nasrallah et al. 2011), the length of an 'ideal' sequence required to get the same behaviour as a real sequence, indicates that effective sequence length may be model dependent. It is important to note that this does not tell us which model would perform better if there were any model misspecification, as would likely be the case in the majority of empirical studies.

I find the extent of the phenomenon of LBJ surprising. It is important to note here that when two quantities can tend to infinity, the order in which limits to infinity are taken can be important. The extent of LBJ is affected by both the sequence length and the long branch length, and the outcome is controlled by the order in which these approach infinity. If $P_{n,L}(T')$ is the probability that ML recovers tree T' (any tree, including T) from n sites generated on T, where L is the long branch length, then if sequence length is taken to infinity first then:

$$\lim_{L\to\infty}\lim_{n\to\infty} P_{n,L}(T) = 1$$

If instead the order of the limits is reversed then:

$$\lim_{n\to\infty}\lim_{L\to\infty} P_{n,L}(T) = c < 1$$

(If limits are taken simultaneously then $P_{n,L}(T)$ converges to $c < 1$ unless n grows

exponentially faster than L, in which case $P_{n,L}(T)$ converges to 1 (Martyn and Steel 2012). This convergence to a value less than 1 is what is seen in Figure 3.9h, where for long branch lengths the correct tree is only obtained about 13% of the time. In order to understand this phenomenon it would be useful to obtain bounds on c. It is possible to show that, in the limits, the probability of obtaining topology 3.9c and topology 3.9e is the same, and hence $c \leq 1/2$ (see Parks and Goldman (2014), Supplementary Methods). This is still much larger than the 13% seen in our simulation. If tighter bounds could be obtained on c then it could significantly improve our understanding of LBJ.

3.4.3 Conclusions

The addition of an extra taxon to a tree increases the number of possible wrong trees which could be inferred, and stochastic error means that they will be inferred sometimes. I have shown that when long branches are not joined to one another they do not appear to attract, so there is no LBC. However the proportion of time long branches join is dependent on branch length, and biases towards trees with long branches placed together get worse as branch lengths increase. These results show that LBJ does happen and is related to the existence of long branches, but it is caused neither by inconsistency or attraction. 'Long branch joining' may be a better term than 'long branch attraction'.

3.5 Conclusions and Future Work

In this chapter I have shown that placing one long branch is difficult for ML, even with the correct model. Counterintuitively, there is a bias towards the tips of the three-taxon tree. Consideration of DM and ML methods has led to insights as to why this bias exists, as well as predictions and ML solutions for trees with zero and infinite branch lengths.

The phenomenon previously denoted by LBA has been analysed for small trees and two distinct analysable phenomena distinguished: LBC and LBJ. LBC is defined as long branches being closer together on the constructed topology than on the true topology. LBJ is defined as long branches being incorrectly joined together on a tree. It has been shown that LBC does not exist on four-taxon trees, and that the long branches do not interact with each other when they are not placed together on a tree. However LBJ does exist and is the same effect as found previously (Huelsenbeck and Hillis 1993). As LBC does not exist, the phrase LBA, which has come to be used for this effect, does not seem appropriate.

The reason for LBJ is still an open question.

An explanation for LBJ would be a significant step forward in understanding the problems caused by multiple long branches, and devising techniques to overcome the problems. One possible avenue of research would be to use a similar approach to that used to understand three species trees, by comparing the results of DM and ML analyses. For three-species trees this is simplified by the fact that there is only one topology. Unfortunately this does not seem to work as well on four-species trees because as branch length increases the best distance-based topology increasingly becomes different to the ML topology.

An alternative approach is to consider what happens when long branch lengths are infinite. If the true branch lengths are infinite then there is approximately a 50% chance that the ML estimate will be finite, and a 50% chance that it will be infinite. It may then be possible to predict the proportions of different topologies based on whether branch lengths are estimated as finite or infinite. This approach appears to be promising, as some branch length combinations do regularly yield the same topology, and deserves further study.

The results shown here have been obtained with long branch lengths and limited amounts of data, which raises the question of whether it is likely that any of these effects will be seen in real data. It is difficult to make direct comparisons from the results shown here to papers citing LBA because real data will not conform to a specific evolutionary model, and is likely to be significantly more complicated than the models examined here. Additionally, empirical studies all use more than three taxa. The effects described in this chapter were seen for single long branches as short as 1 (expected substitution per site), well within the bounds of many existing studies. For the cases with two long branches, LBJ only becomes a real problem when the long branches are of length 2 or greater. For these lengths it would be difficult to align the sequences. However, real sequences have much more complicated evolution than that assumed here, and there is no way of dismissing LBJ as a possible problem for real data.

Previously a large number of tests for LBA have been suggested. My results indicate that these tests may not all be appropriate. For example, one such method is based on removing one of the long branches and then repeating the reconstruction. If the long branch maintains its original position then this was taken to indicate LBA had not taken place (Pol and Siddall 2001). However, I have shown that even one long branch is not necessarily expected to be placed correctly, suggesting this test may not be adequate. Another method proposes detection of LBA by comparing results using a phylogenetic inference method that suffers less from LBA (Huelsenbeck 1997), but my finding that even ML

with a correctly specified model can suffer from LBA (in fact, LBJ) indicates that care should be taken to ensure methods shown to be robust to this problem are used.

My study shows that even one long branch may be placed incorrectly and in an unexpected way by ML on problems as simple as three or four-taxon trees with a correctly specified substitution model. Although not in itself informative about behaviour on larger trees, this gives cause for concern when analysing trees with even one very long branch, and highlights the fact that investigations involving larger trees are needed, as currently not much is known (Kuhner and Felsenstein 1994; Pol and Siddall 2001; Kück et al. 2012). It would be interesting to see how the results of this chapter scale up and whether LBJ and LBC occur on larger topologies.

Chapter 4

Reversibility

A paper based on the work presented in this chapter is under review at *Systematic Biology*. The version presented here is largely the same, with a slightly altered Methods section, a case study using non-reversible models (de Beer et al. 2013), and an expanded Discussion and Future Work section.

4.1 Introduction

The first evolutionary model described for nucleotide substitutions assumed that the rates of change between all nucleotides were equal, and therefore had no parameters to estimate (Jukes and Cantor 1969). This assumption was progressively relaxed, with parameters introduced to describe biological features of the data (e.g. Kimura 1980; Felsenstein 1981; Hasegawa et al. 1985; Tamura 1992; Tamura and Nei 1993). One assumption that is generally still retained is that of time reversibility, with almost all evolutionary models commonly used in phylogenetic analysis a subset of the general time reversible model (GTR, sometimes also denoted REV) (Lanave et al. 1984; Tavaré 1986; Yang and Goldman 1994). A model is time reversible if, at equilibrium, the amount of change from character i to character j is the same as the amount of change from character j to character i (Norris 1998). This means that it is not possible to determine the direction of a process. In phylogenetics, the assumption of reversibility is made for the sake of convenience. It makes the estimation of probability matrices mathematically easier (Kelly 1979; Golub and Van Loan 1996), speeding up likelihood calculations. It also reduces computational effort as, when using likelihood-based inference (such as maximum likelihood or Bayesian inference) under a reversible model, it is possible to calculate the likelihood from any place on the tree (the "pulley principle" of Felsenstein 1981). Both of these things make it easier and

faster to estimate models, trees and branch lengths under a reversible model. When computational power is limited this is a distinct advantage.

While the assumption of reversibility makes likelihood calculations easier, it is without biological justification. Furthermore, there is evidence that the evolutionary process is not reversible: for example, the increased rate of change of C in CpG sites (Coulondre et al. 1978). If a non-reversible model fits the data better then, aside from a greater understanding of evolution, its use may improve the accuracy of inferred tree topologies and branch lengths and of downstream analyses such as searches for positive selection. In theory non-reversible models can also be used for tree-rooting, whereas reversible models cannot as the direction of the process is unknown. It has, however, been shown that in general alternative methods for tree-rooting perform better (Huelsenbeck et al. 2002; Yap and Speed 2005).

Yang (1994a) was the first to explore the use of non-reversible models and test whether they fit data better than reversible models, concluding that "The use of the [non-reversible] model does not appear to be worthwhile" (Yang 1994a, p. 105). However, that study analysed only two small alignments of closely related nucleotide sequences; since then, further studies have indicated that non-reversible models may be a significant improvement over reversible models (Squartini and Arndt 2008; Baele et al. 2010; Jayaswal et al. 2010; De Maio et al. 2013a,b). These studies still only cover either a small number of data sets (Baele et al. 2010; Jayaswal et al. 2010) or a small number of species (Squartini and Arndt 2008; De Maio et al. 2013a,b) and leave open the question of whether non-reversible models should be preferred over reversible models. In practice, almost everyone still uses reversible models due to their simplicity and speed, and the fact that very few of the widely-used tree reconstruction programs implement non-reversible models; in short, for reasons of convenience.

The majority of the studies so far have considered only nucleotide data, despite the fact that amino acid data are often preferred for building trees. No non-reversible amino acid models have been explored and codon models, commonly used to study selective pressures in proteins, also tend to be reversible (Anisimova and Kosiol 2009), with only one paper estimating non-reversible codon models (De Maio et al. 2013a). Non-reversible amino acid and codon models have been avoided probably because fitting such a large number of parameters is challenging both computationally (Boussau and Gouy 2006) and mathematically (Golub and Van Loan 1996). In this chapter I look for the first time at non-reversible amino acid models.

Analysis of non-reversibility requires measures that quantify it, as well as

statistical tests of whether data are significantly better fit by a model that is reversible or not. Quantification of non-reversibility has not been greatly discussed before, despite the fact that a measure of non-reversibility would be very useful. The need for quantification is illustrated by a suggestion that "highly non-reversible" models may be useful for tree rooting (Huelsenbeck et al. 2002, p. 39) despite the fact that it is unclear what 'highly non-reversible' means. Quantification may also be useful for downstream analyses using non-reversible models. I develop a number of measures of reversibility and look at how they relate to each other and to tests for non-reversibility.

Testing for non-reversibility has, on the other hand, been discussed extensively. This can be split into two main ideas; testing whether a non-reversible model fits the data better than a reversible model, for example with an LRT (Yang 1994a), or testing whether a model itself is non-reversible. The former approach is a standard mathematical method for testing nested models, commonly used in phylogenetics for deciding whether a more complicated model should be preferred over a simpler one (Yang 2006) (see Section 2.4). The statistic used in this test, the likelihood ratio, is a measure of the strength of evidence for non-reversibility in the data. Figure 4.1 shows reversible and non-reversible models, inferred by maximum likelihood (ML), for two alignments: one where the non-reversible model is significantly better than the reversible model, and one where the non-reversible model is not significantly better.

The latter approach, testing a single model for non-reversibility, has been developed on models derived from pairs of sequences (Saccone et al. 1990; Rzhetsky and Nei 1995; Eyre-Walker 1999; Ababneh et al. 2006) and on a single phylogenetic branch connecting an ancestral node with a more recent one (Squartini and Arndt 2008). The measures used in these tests could in principle be used for quantification of non-reversibility, but currently this is not possible as either the measures only apply to pairwise comparisons (Saccone et al. 1990; Rzhetsky and Nei 1995; Eyre-Walker 1999; Ababneh et al. 2006) or there are multiple indices which cannot be easily combined into one measure (Squartini and Arndt 2008).

I develop, compare and discuss measures for quantifying non-reversibility, and suggest the most useful. Contrasting them to LRTs allows me to explore the relationship between the strength of evidence for non-reversibility and levels of non-reversibility. These assessments are performed on both nucleotide and, for the first time, amino acid data, and cover a much larger range of data sets than previous studies. This allows me to draw conclusions about the applicability of non-reversible models for nucleotide and amino acid alignment data sets.

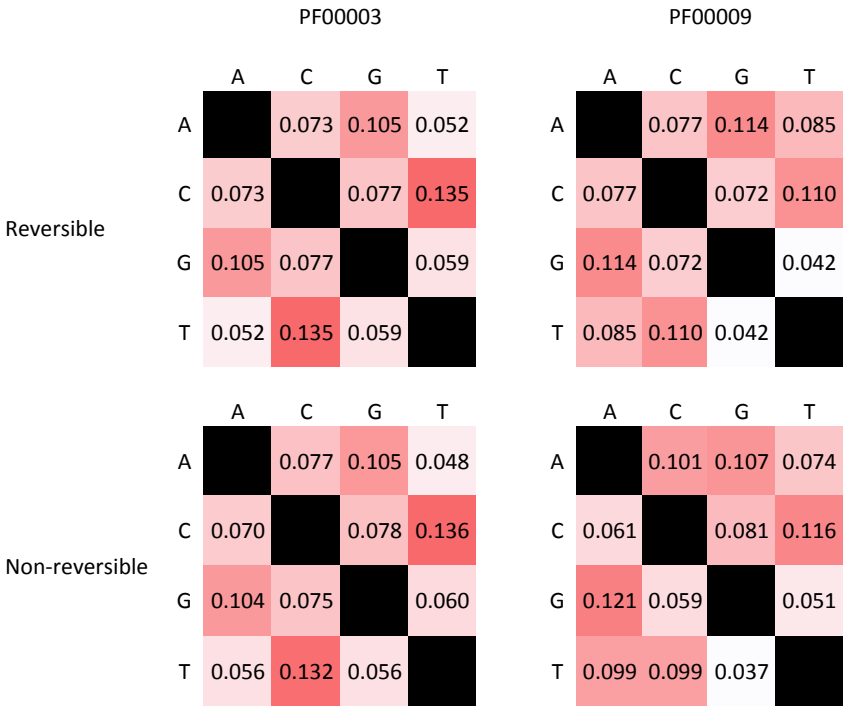


Figure 4.1: Probability flux ($\pi_i Q_{ij}$) ML estimates of reversible and non-reversible models for two alignments: PF00003, where the non-reversible model was not significantly better than the reversible model; and PF00009, where it was. The fluxes of the reversible models are symmetric, by design. The non-reversible model for PF00003 is close to being symmetric, with values very similar to the reversible model. The non-reversible model for PF00009, on the other hand, is much less symmetric: compare (e.g.) the A-C and C-A substitutions.

4.2 Methods

4.2.1 Alignments

To create a data set of nucleotide multiple sequence alignments (MSAs) I selected the first 400 MSAs with more than four sequences, greater than 100 nucleotide sites, and greater than 60% of reliably aligned sites from Pandit, a database of MSAs for protein domains over a wide range of species along with associated trees (Whelan et al. 2006). This filtering was carried out to remove small or low-quality MSAs where it would not be possible to estimate substitution rates accurately, using similar criteria to Zoller and Schneider (2013). I also used 100 gene sequence alignments spanning the range from slow- to fast-evolving genes from 38 mammals, sampled from a genome-wide set of mammalian gene alignments with an associated rooted tree derived from the Mammalian Genome Project (Lindblad-Toh et al. 2011) and augmented using other mammalian genomes from release 63 of the Ensembl database (Flicek et al. 2011) by Jordan (2011).

To create a data set of amino acid MSAs I selected the first 200 MSAs in Pandit with more than four sequences, greater than 100 amino acid sites, and greater than 60% of reliably aligned sites. Amino acid models are much more difficult to build computationally, both due to the amount of time they take to estimate and the large number of parameters required to be estimated. For these reasons fewer amino acid alignments were used than nucleotide alignments. Amino acid versions of the 100 alignments of genes across 38 mammals used for the nucleotide data set were also selected. Summary information about the MSAs used can be found in Table 4.1, and a full list of data sets used is provided in Appendix B.

4.2.2 Trees

Trees for each Pandit alignment were taken from the Pandit database. These were produced by building trees with a variety of different methods including PhyML, BioNJ and FastME, and then choosing the tree with the best likelihood (Whelan et al. 2006). For the mammal data sets I use the tree in Jordan (2011), pruned to the species present in each alignment.

Table 4.1: Information about MSAs

			Minimum	Median	Mean	Maximum
Nucleotide Data Sets	Pandit (400)[a]	Number of Sequences	5	22	50	1627
		Sequence Length (nt)	102	542	708	2868
		Tree Length[b]	1·4	9·1	21	475
	Mammal (100)[a]	Number of Sequences	5	32	30	38
		Sequence Length (nt)	120	720	1074	5400
		Tree Length[b]	0·2	1·9	2·2	7·3
Amino Acid Data Sets	Pandit (184)[a]	Number of Sequences	6	25	53	352
		Sequence Length (aa)	103	271	300	956
		Tree Length[b]	1·4	9·7	22	211
	Mammal (71)[a]	Number of Sequences	5	32	30	38
		Sequence Length (aa)	40	280	403	1800
		Tree Length[b]	0·5	2·2	2·3	6·4

[a] In parentheses are the number of MSAs used, including only those for which both a non-reversible and a reversible model could be found using HyPhy (Kosakovsky Pond et al. 2005).

[b] Tree lengths shown were calculated for the tree topology associated with each alignment, using a reversible model. Tree lengths derived using non-reversible models were very similar (results not shown).

For non-reversible models rooted trees are required. The trees in Pandit are rooted arbitrarily. While ideally the optimal root position would be found by trying every possible position of the root and picking the position with the highest likelihood, this is computationally very expensive. Instead, I tried two possible rootings, midpoint rooting and the arbitrary rooting in Pandit, and chose the one with the higher likelihood. Comparing the results from these two different rootings shows that position of the root does not make a big difference to any of the conclusions (results not shown). The mammal data were included in the study in order to study some trees where the root is known with confidence, although it should be noted that the mammal trees tend to be shorter and cover fewer species than the Pandit data sets. Further information about the trees used can be found in Table 4.1.

4.2.3 Substitution Models

In this chapter I use general reversible and non-reversible models, represented by instantaneous rate matrices Q and stationary distributions π, as described in section 2.1.1. A model is reversible if it satisfies the detailed balance equations, $\pi_i Q_{ij} = \pi_j Q_{ji}$, for all i and j (Norris 1998). For reversible models, for $i \neq j$, the elements Q_{ij} can be written as $\pi_j S_{ij}$ where the S_{ij} satisfy $S_{ij} = S_{ji}$. The general reversible and non-reversible nucleotide models have 8 and 11 parameters, respectively, and the general reversible and non-reversible amino acid models have 208 and 379 parameters, respectively. All models are normalised so that the expected number of substitutions per site at equilibrium, $\sum_i \sum_{j \neq i} \pi_i Q_{ij}$, is equal to 1.

4.2.4 Maximum Likelihood Estimation

For each MSA, I find the models that are the best fit, under the constraint of reversibility and without that constraint. I use ML to do this, by optimising the likelihood on the corresponding topology (assumed to be correct), with choice of rootings (where appropriate) and maximizing over nuisance parameters such as branch lengths. This yields a ML reversible model Q^{GTR} (with equilibrium distribution π^{GTR}) and a ML non-reversible model Q^{NR} (with equilibrium distribution π^{NR}). These calculations are carried out independently for each MSA. I used HyPhy for all optimisations (Kosakovsky Pond et al. 2005), with nucleotide models verified using baseml in PAML (Yang 2007). For each MSA, HyPhy was run from five different starting points to confirm the inference of the best model. If, for any MSA, HyPhy could not converge on both a reversible and non-reversible model

then that MSA was not used (see Tab. 4.1). This usually corresponded to MSAs where the alignment had many gaps. HyPhy was chosen as it is the only program I know of that can calculate stationary non-reversible models for nucleotide and amino acid alignments. I also tried using Xrate (Klosterman et al. 2006), which can calculate non-reversible models for both alignment types; however, it turned out that it cannot calculate stationary non-reversible models unless the desired stationary distribution is known in advance.

4.2.4.1 Computational Issues

Obtaining models that could be reliably confirmed to be optimal if a program was re-run proved to be challenging, particularly for amino acid models. To try to get the best model, as well as trying many starting positions, I adjusted parameters within baseml and HyPhy that are used to decide when the best model has been found. During this process I also encountered a number of bugs in HyPhy which caused it to either not finish optimising or to output clearly wrong answers. Although the bugs I have encountered have been fixed (with very helpful assistance provided by Sergei Kovakovsky-Pond), this illustrates that optimising non-reversible models, particularly on small amino acid data sets, is a very difficult problem.

4.2.5 Quantification of Non-reversibility

Four different measures of non-reversibility are described below. These are calculated using the elements of one or both of the ML reversible and ML non-reversible models Q^{GTR} and Q^{NR}.

4.2.5.1 Deviation from Detailed Balance

$$\nabla_A = \sum_i \sum_{\substack{j \\ j<i}} |\pi_i^{NR} Q_{ij}^{NR} - \pi_j^{NR} Q_{ji}^{NR}|$$

Here, non-reversibility is being quantified by a measure of how far the ML non-reversible model is from satisfying the detailed balance equations. If the model is reversible, then $\nabla_A = 0$; otherwise $0 < \nabla_A \leq 1$. This was previously suggested as a measure of non-reversibility by Huelsenbeck et al. (2002).

4.2.5.2 Distance Between Q^{NR} and the Closest Reversible Model by Probability Flux

$$\nabla_B = \min_{\pi_i, S_{ij}} \sum_i \sum_{\substack{j \\ j \neq i}} |\pi_i^{NR} Q_{ij}^{NR} - \pi_i \pi_j S_{ij}|$$

Here, the values of the variables π_i and S_{ij} (constrained such that $S_{ij} = S_{ji}$) that minimise ∇_B define the reversible model that is closest to Q^{NR}. If the distance is zero then the model is reversible. Despite being motivated differently, it can be shown that this measure is the same as the deviation from detailed balance: $\nabla_B = \nabla_A$.

To see this we can expand the summation giving

$$\begin{aligned}
\nabla_B &= \min_{\pi_i, S_{ij}} \left[\sum_i \sum_{\substack{j \\ j<i}} (|\pi_i^{NR} Q_{ij}^{NR} - \pi_i \pi_j S_{ij}| + |\pi_j^{NR} Q_{ji}^{NR} - \pi_j \pi_i S_{ji}|) \right] \\
&= \min_{\pi_i, S_{ij}} \left[\sum_i \sum_{\substack{j \\ j<i}} (|\pi_i^{NR} Q_{ij}^{NR} - \pi_i \pi_j S_{ij}| + |\pi_j^{NR} Q_{ji}^{NR} - \pi_i \pi_j S_{ij}|) \right]
\end{aligned}$$

because $S_{ij} = S_{ji}$.

By the triangle equality,

$$\sum_i \sum_{\substack{j \\ j<i}} (|\pi_i^{NR} Q_{ij}^{NR} - \pi_i \pi_j S_{ij}| + |\pi_j^{NR} Q_{ji}^{NR} - \pi_i \pi_j S_{ij}|) \geq \sum_i \sum_{\substack{j \\ j<i}} |\pi_i^{NR} Q_{ij}^{NR} - \pi_j^{NR} Q_{ji}^{NR}|$$

The quantity on the left is bounded below by ∇_A, proving that $\nabla_B \geq \nabla_A$.

To prove $\nabla_A = \nabla_B$, it remains to show that there always exists a reversible rate matrix for which this lower bound is attained. Such a matrix is given by $Q_{ij} = \pi_j S_{ij}$, where $\pi_i = 1/n$ for all i and each $S_{ij} = S_{ji}$ is chosen to satisfy either $\pi_i^{NR} Q_{ij}^{NR} \leq \pi_i \pi_j S_{ij} \leq \pi_j^{NR} Q_{ji}^{NR}$ or $\pi_i^{NR} Q_{ij}^{NR} \geq \pi_i \pi_j S_{ij} \geq \pi_j^{NR} Q_{ji}^{NR}$. Then for each pair i,j:

$$|\pi_i^{NR} Q_{ij}^{NR} - \pi_i \pi_j S_{ij}| + |\pi_j^{NR} Q_{ji}^{NR} - \pi_i \pi_j S_{ij}| = |\pi_i^{NR} Q_{ij}^{NR} - \pi_j^{NR} Q_{ji}^{NR}|$$

and hence $\nabla_A = \nabla_B$. We therefore focus on ∇_A.

4.2.5.3 Distance Between Q^{NR} and Q^{GTR}

$$\nabla_C = \sum_i \sum_{\substack{j \\ j \neq i}} |Q_{ij}^{NR} - Q_{ij}^{GTR}|$$

Non-reversibility can also be measured as the distance between the ML non-reversible and reversible models. If the data set is best described by a reversible model, then the ML reversible and non-reversible models will be the same and $\nabla_C = 0$. If the evolution of the data set is better described by a non-reversible model then $\nabla_C > 0$.

4.2.5.4 Distance Between Q^{NR} and Q^{GTR} by probability flux

$$\nabla_D = \sum_i \sum_{\substack{j \\ j \neq i}} |\pi_i^{NR} Q_{ij}^{NR} - \pi_i^{GTR} Q_{ij}^{GTR}|$$

The distance between the ML reversible and non-reversible models can also be calculated by comparing the probability fluxes $\pi_i Q_{ij}$. As above, this measure is 0 if the evolution of a data set is best described by a reversible model, and greater than 0 otherwise. Under some conditions, $\nabla_D = \nabla_A$, but this is not always true.

Again, this can be seen by expanding the terms in the summation

$$\begin{aligned}
\nabla_D &= \sum_i \sum_{\substack{j \\ j<i}} (|\pi_i^{NR} Q_{ij}^{NR} - \pi_i^{GTR} Q_{ij}^{GTR}| + |\pi_j^{NR} Q_{ji}^{NR} - \pi_j^{GTR} Q_{ji}^{GTR}|) \\
&= \sum_i \sum_{\substack{j \\ j<i}} (|\pi_i^{NR} Q_{ij}^{NR} - \pi_i^{GTR} Q_{ij}^{GTR}| + |\pi_j^{NR} Q_{ji}^{NR} - \pi_i^{GTR} Q_{ij}^{GTR}|) \\
&\begin{cases}
= \sum_i \sum_{\substack{j \\ j<i}} |\pi_i^{NR} Q_{ij}^{NR} - \pi_j^{NR} Q_{ji}^{NR}| & \text{if } (\pi_i^{NR} Q_{ij}^{NR} > \pi_i^{GTR} Q_{ij}^{GTR} > \pi_j^{NR} Q_{ji}^{NR}) \\
& \text{or } (\pi_j^{NR} Q_{ji}^{NR} > \pi_j^{GTR} Q_{ji}^{GTR} > \pi_i^{NR} Q_{ij}^{NR}), \forall ij \\
> \sum_i \sum_{\substack{j \\ j<i}} |\pi_i^{NR} Q_{ij}^{NR} - \pi_j^{NR} Q_{ji}^{NR}| & \text{otherwise}
\end{cases}
\end{aligned}$$

If for each $i < j$ there exist values x_{ij} such that $\pi_i^{NR} Q_{ij}^{NR} \leq x_{ij} \leq \pi_j^{NR} Q_{ji}^{NR}$ or $\pi_i^{NR} Q_{ij}^{NR} \geq x_{ij} \geq \pi_j^{NR} Q_{ji}^{NR}$, then $\nabla_D = \nabla_A$. Otherwise, $\nabla_D > \nabla_A$. For normalised matrices, this measure ranges from 0 to 2.

4.2.6 Significance Testing

LRTs were carried out to compare the non-reversible and reversible models. The LRT statistic for non-reversible evolution for an alignment, Λ, is twice the difference in log-likelihood between the ML non-reversible and ML reversible models:

$$\Lambda = 2(\log(L^{NR}) - \log(L^{GTR}))$$

This statistic is then compared to the χ_d^2 distribution, where the degrees of freedom d is given by the difference between the number of free parameters in the

two models. For nucleotide data, the non-reversible model has 11 parameters and the reversible model has 8 parameters, so attained values of Λ are compared to χ^2_3. For amino acids, the non-reversible and reversible models have 379 and 208 parameters, respectively, so Λ is compared to χ^2_{171}. For brevity, I refer to MSAs for which the non-reversible model is significantly better than the reversible model as 'non-reversible MSAs'; otherwise, I refer to them as 'reversible MSAs'.

4.3 Results

4.3.1 What is the Relationship Between the ∇ Measures?

The three measures ∇_A, ∇_C and ∇_D were calculated for each MSA. The measures show high correlation (Fig. 4.2), with the correlation between ∇_A and ∇_D the greatest. This is expected as ∇_A and ∇_D are often identical (see Section 4.2.5.4). These correlations show that, despite being derived from different matrices and parameter combinations, the measures are capturing similar effects. The correlations are lower for the amino acid MSAs, probably due to the higher variance associated with inferring a greater number of parameters in each amino acid model.

4.3.2 What is the Relationship Between the Measures of Non-reversibility and the LRT?

Next I explored how these measures of reversibility relate to the LRT statistic Λ (Tab. 4.2 and Fig. 4.3). As the three measures are highly correlated, I focus on ∇_A. Non-reversible MSAs are shown in red and yellow (for Pandit and mammal data, respectively), and reversible MSAs are shown in black and blue. Histograms on the right of the plots show the distribution of ∇_A for non-reversible and reversible MSAs. Similar plots for the other measures can be found in Appendix B.

Table 4.2: Correlation between the ∇ measures and the LRT statistic Λ for nucleotide and amino acid data.

	Nucleotides		Amino Acids	
	Λ	Scaled Λ	Λ	Scaled Λ
∇_A	0.259	0.887	0.711	0.766
∇_C	0.259	0.877	0.700	0.647
∇_D	0.259	0.887	0.700	0.773

The measures defined in this paper are estimates of the strength of non-reversibility for a certain data set, whereas Λ quantifies the evidence for non-

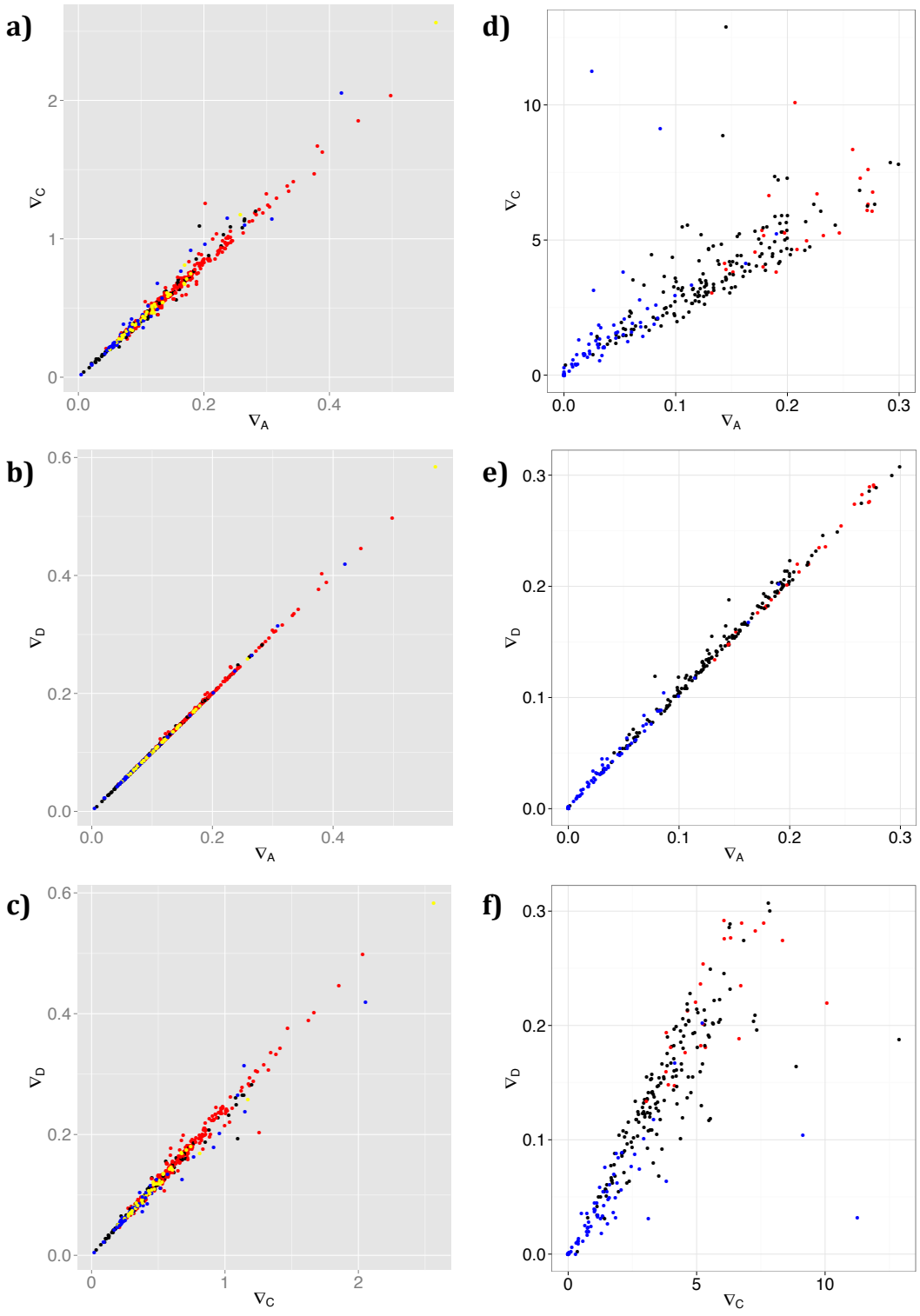


Figure 4.2: Relationships between ∇_A, ∇_C and ∇_D for nucleotide MSAs (**a-c**; grey backgrounds) and amino acid MSAs (**d-f**; white backgrounds). Each point corresponds to one MSA. Pandit MSAs are red if they are significantly better described by a non-reversible model and black otherwise. Mammal MSAs are yellow if they are significantly better described by a non-reversible model and blue otherwise. Correlations for **a-f** are 0.990, 0.999, 0.990, 0.821, 0.998 and 0.839, respectively.

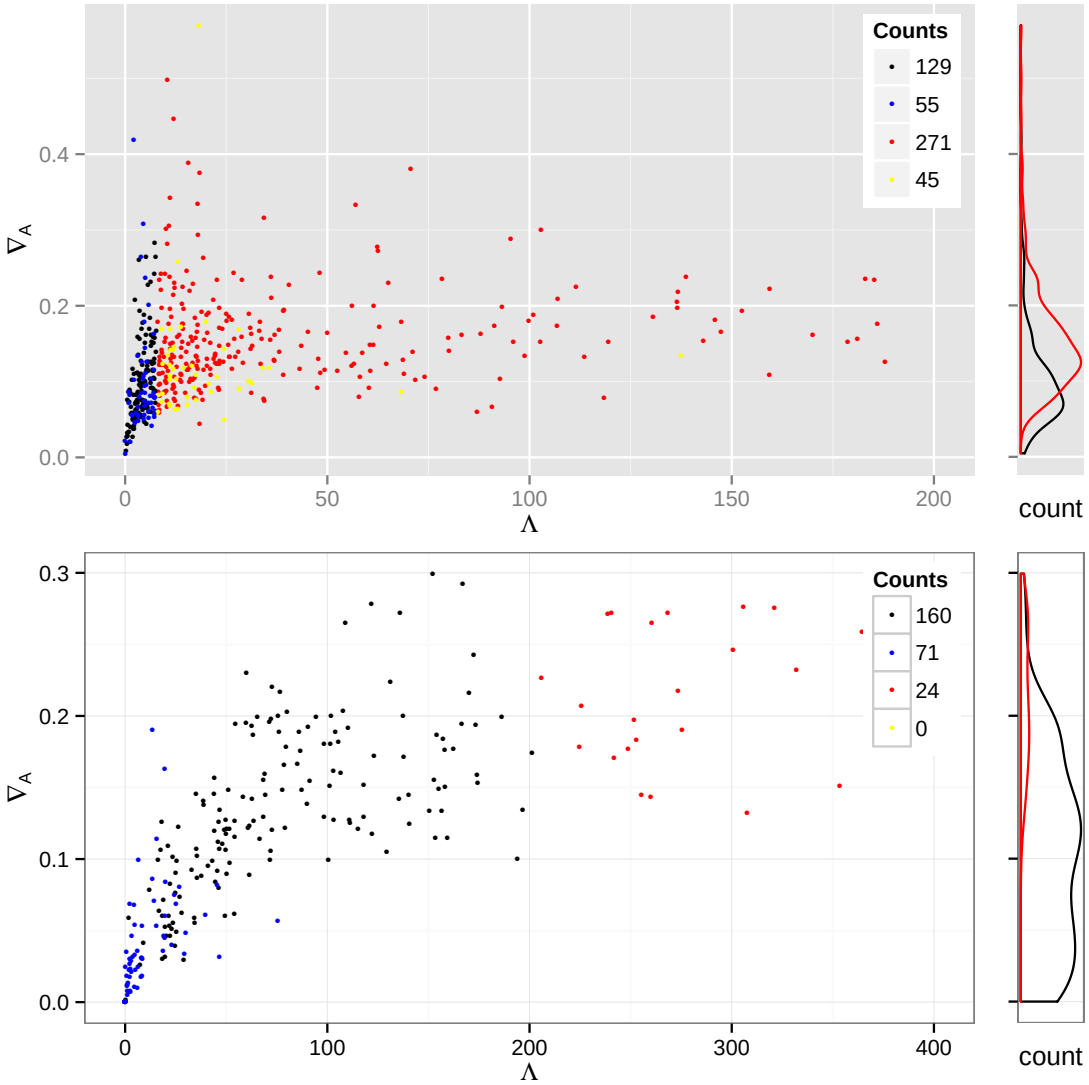


Figure 4.3: Relationship between the LRT statistic Λ and ∇_A for nucleotides (top, grey background) and amino acids (bottom, white background). Each point corresponds to a MSA: Pandit MSAs are red if they are significantly better described by a non-reversible model and black otherwise; mammal MSAs are yellow if they are significantly better described by a non-reversible model and blue otherwise. Histograms on the right show ∇_A for MSAs found to be significant (red) or non-significant (black). Corresponding plots with ∇_C and ∇_D are shown in Figures B.1 and B.2.

reversibility. These are not expected to be the same, and indeed this expectation is confirmed by my results (Tab. 4.2 and Fig. 4.3). For both the nucleotide and amino acid data sets, the distributions of ∇_A for non-reversible and reversible MSAs overlap greatly. For nucleotides there is also a low correlation (0.259) between Λ and ∇_A, with the range in ∇_A values large for low Λ values, but decreasing as Λ increases (Fig. 4.3). For the amino acid data sets the correlation is higher (0.711), but with fewer non-reversible MSAs, and again there are many reversible MSAs with similar or higher ∇_A values than non-reversible MSAs. For both nucleotide and amino acid alignments there are a number of MSAs that have low Λ, and thus are not significant using a χ^2 test, but high ∇_A, meaning that the optimal non-reversible model is far from reversible but there is not significant evidence in the data to prefer the non-reversible model. This could occur when the alignment and/or tree are very short, making the estimation of the models less accurate.

I expect the ability to detect non-reversibility using an LRT to depend on the strength of non-reversibility (∇_A) and the information content of the sequence alignment. The quantity of information in a sequence alignment is proportional to the length of the alignment, and also depends on the tree (topology and branch lengths) relating the sequences. The information content of a tree is not a well-understood quantity (Geuten et al. 2007), but the number of sequences and total tree length are possible proxies. I tested both of these, and found that my intuitive idea of the information content of trees was best-represented by the number of sequences (results not shown). Therefore, I used it as a proxy in this study.

To explore how the relationship between measures of non-reversibility and the LRT is affected by variation in these factors, I simulated reversible sequences for a variety of trees and sequence lengths and then calculated non-reversibility measures and LRT statistics for these data sets. I illustrate these simulations using two trees taken from the Pandit data set (32 species and 154 species) and using the reversible nucleotide model inferred from Pandit MSA PF00003. For each topology I simulated 100 data sets using sequence lengths of 100, 1000 and 5000. Figure 4.4a shows the relationship of the inferred ∇_A and Λ for these simulations. Since the simulation substitution model was reversible, the true value of ∇_A should be 0. As sequence length increases, the inferred values of ∇_A get smaller, getting closer to 0. ∇_A is also smaller for the larger tree (more sequences; also greater tree length) . This is as expected as adding more sequences, or longer sequences, increases the accuracy of model inference.

These simulations illustrate that the signal for non-reversibility contains components attributable to both the inferred level of non-reversibility and the infor-

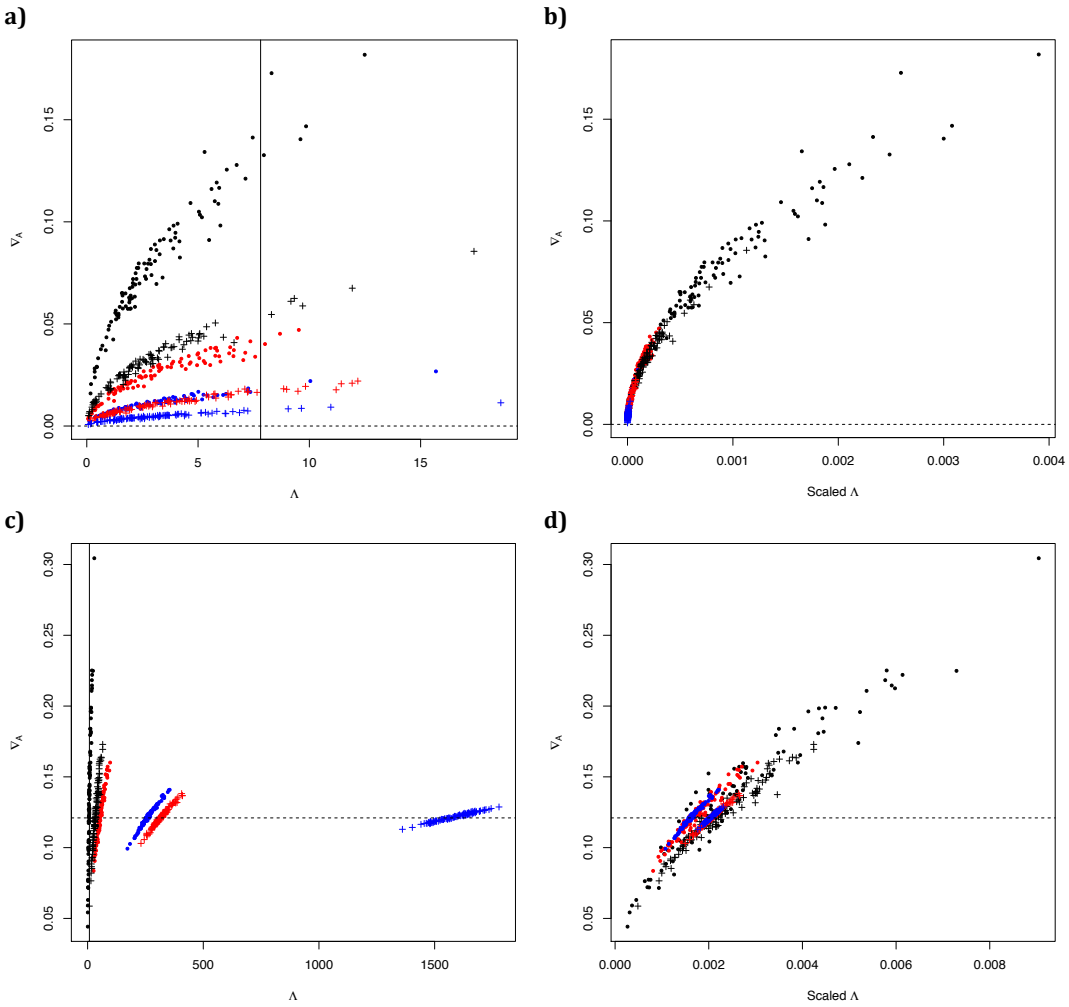


Figure 4.4: **a** The LRT statistic Λ against inferred ∇_A for data simulated under a reversible model for a 32 species tree (dots) and a 154 species tree (plus signs) with sequence lengths of 100 (black), 1000 (red) and 5000 (blue). The vertical line shows the 5% cut-off point for χ^2_3; only 5% of the points should be above this line. This is approximately true for our simulations. The horizontal dashed line shows ∇_A for the matrix used for simulation; as the matrix is reversible $\nabla_A = 0$. **b** Scaled Λ against ∇_A for the same data as in **a**. Λ is scaled by the inverse of information content, estimated as the product of sequence length and sequence number. **c**–**d** show the equivalent of **a**–**b**, using a non-reversible model for simulation. As sequence length and tree size increase, so does Λ; the LRT finds all MSAs to be non-reversible, apart from 54% of the least informative (32 species tree, sequence length 100) data sets (**c**). The value of ∇_A for the matrix used for simulation is 0.121 (horizontal dashed line). As sequence and tree length increase, inferred values of ∇_A converge to the correct value.

mation content of the MSA. To clarify the distinction further, I scaled Λ by the inverse of information content (measured as the product of sequence length and sequence number, as described above). The scaled Λ represents the signal of non-reversibility per sequence site, and shows a simple monotonic relationship with inferred ∇_A value across all of our simulation conditions (Fig. 4.4b). I repeated the simulations using the non-reversible nucleotide model inferred from Pandit MSA PF00009 (see Fig. 4.1) and similar features are seen (Fig. 4.4c–d).

The aim of scaling the LRT statistic is to capture the evidence of non-reversibility per nucleotide or amino acid, which I hope is relatively independent of the effects of the tree. Indeed, when looking at real data, the scaled Λ does correlate well with the measures (see Tab. 4.2, Fig. 4.5, B.1 and B.2). This confirms that it is possible to separate the effects of the strength of the signal (∇_A) from the total amount of evidence for non-reversibility, and reinforces my belief that there is a place for both tests of the presence of non-reversibility, and measures of non-reversibility itself.

4.3.3 Do Non-reversible Models Describe the Evolutionary Process Better than Reversible Models?

For nucleotide MSAs, 68% of the Pandit MSAs and 45% of the mammal MSAs are significantly better described by a non-reversible model (Fig. 4.3). Fewer of the mammal MSAs are significant, probably because on average both the trees and the alignments are smaller (Tab. 4.1). Analysing the Pandit MSA codon positions independently shows that 53% are significant for the first codon position, 52% for the second position and 63% for the third position. This corresponds to more MSAs being significantly better described by a non-reversible model when more change is expected. These results conflict with previous recommendations that non-reversible models are not important (Yang 1994a).

This is the first time non-reversible amino acid models have been studied. 13% of the Pandit amino acid MSAs and none of the mammal amino acid MSAs are significantly better described by a non-reversible model (Fig. 4.3). The fact that fewer MSAs are significant for amino acids than for nucleotides can in part be attributed to the facts that the amino acid alignments are on average 1/3 of the length of the nucleotide alignments, and the amino acid alphabet is much larger meaning that there are many more parameters to estimate. All of the MSAs that are detected as non-reversible for amino acids are also non-reversible for nucleotides. My results indicate that for large alignments or more divergent amino acid data sets, non-reversible models may be valuable.

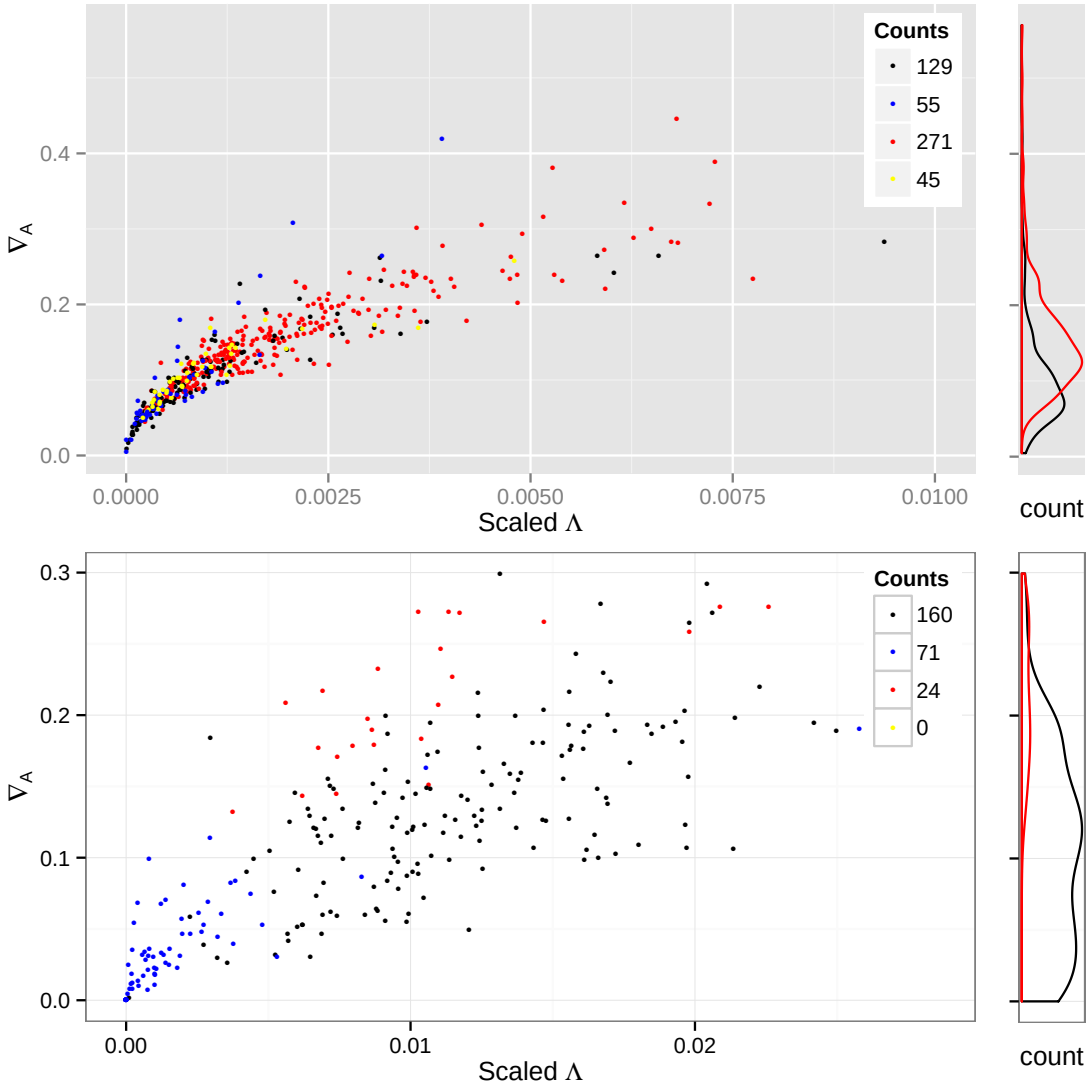


Figure 4.5: Relationship between scaled Λ and ∇_A for nucleotides (top, grey background) and amino acids (bottom, white background). Plots are as in Figure 4.3, except Λ is now scaled by the inverse of the information content, as approximated by the product of sequence length and sequence number, minus the number of gaps. Analogous comparisons with ∇_C and ∇_D are shown in Figures B.3 and B.4.

4.4 A Case Study Using Reversible and Non-reversible Models

Whilst I was working on the use of non-reversible models I got involved in a project with Tjaart de Beer on comparing mutations in the healthy human population, taken from the 1000 Genomes (1kG) Project (The 1000 Genomes Project Consortium 2010), with mutations known to cause disease (from OMIM, Amberger et al. 2009). This project has been published (de Beer et al. 2013). Below I present my work on the project, which focuses on building models of evolution in humans using the 1kG data and comparing these with other empirical models. The full paper can be seen in Appendix B.3.

The 1kG Project Consortium aimed to catalogue at least 95% of human DNA variants (with a frequency of occurrence of $> 1\%$) found worldwide, providing a rich set of single nucleotide polymorphisms (SNPs) known to occur in healthy individuals. To understand the features of this set of SNPs in humans I built instantaneous rate matrices for both amino acids and codons, and compared them with previous empirical models (such as those described in Sections 2.1.3 and 2.1.4). Generally, instantaneous rate matrices are built from between-species data, and hence it is assumed that they are modelling the result of mutation and selection. The 1kG data is within-species data, so I expect that, due to the time-scale and the relatively weak selection in human populations (Lohmueller et al. 2008), selection will not have had a great effect, meaning that the model will mainly represent the mutation process. There will however be some effect of selection, as SNPs which are lethal, or have been fixed (or lost) very quickly, will not be present in the data.

Rate matrices were built by counting SNPs and then converting the matrix of counts into an instantaneous rate matrix using methods described in Kosiol and Goldman (2005). As I am interested in human evolution I want to capture only SNPs occurring within the human lineage (i.e. since divergence from the most recent common ancestor of the 1kG sample). Since the relationship structure within each sub-population is not clear, a conservative method is to count SNPs that occur only in one sub-population. Each SNP is counted only once. This method also allows me to determine the direction of each SNP, as the allele present in all of the other sub-populations, which is very likely to be the ancestral allele, is known. The use of directed changes means that the inferred instantaneous rate matrix can be non-reversible.

4.4.1 Amino Acid Models

For amino acids this procedure produces a non-reversible model with $\nabla_A = 0.32$. This value of ∇_A is higher than that of any of the amino acid models built in this chapter, indicating that the 1kG model is strongly non-reversible. To investigate properties of the 1kG model I compared it with matrices calculated for nuclear genes (Dayhoff et al. 1978; Jones et al. 1992; Liò and Goldman 1999; Whelan et al. 2001; Le and Gascuel 2008; Zoller and Schneider 2013), mitochondrial genes (Adachi and Hasegawa 1996; Yang et al. 1998; Abascal et al. 2007; Rota-Stabelli et al. 2009), chloroplast genes (Adachi et al. 2000; Cox and Foster 2013), and separately for the exposed and buried residues of globular proteins (Goldman et al. 1998). As these models can have up to 379 parameters, they can each be considered to be a point in 379 dimensional space. This is difficult to visualise and therefore I used principal component analysis (PCA) to reduce the dimensionality and display the matrices in a lower-dimensional space that captures the main variation between the matrices.

Model comparison can be carried out using: (a) the Q_{ij} values themselves, (b) the exchangeability parameters S_{ij} (recall $S_{ij} = Q_{ij}/\pi_j$ see Section 2.1.1), or (c) the probability flux $\pi_i Q_{ij}$. I tested these three possible parameterisations and found that they yield similar results; therefore, I only show results using Q_{ij} (Figure 4.6). The 1kG model is the only model which is non-reversible. To check that any differences between the 1kG model and other models are not caused by this fact, I also calculated and included a reversible version of the 1kG model.

The first principal component clearly separates the two 1kG models (reversible and non-reversible), which are placed very close together, from all of the previously calculated models. The second principal component then spreads models out based on whether the alignments used to build them are made up mainly of exposed or buried domains, with the mitochondrial models on one end built from nearly all membrane proteins, and models built from only exposed regions of proteins at the other end. I have also added the Pandit reversible and non-reversible models, discussed earlier in this chapter, to the plot. The paired reversible and non-reversible points fall close together, showing that including direction does not make as big a difference as the characteristics of the protein families. Models from individual families form a cloud around the models derived from collections of nuclear proteins, further underlining the difference between the 1kG models and the between-species models.

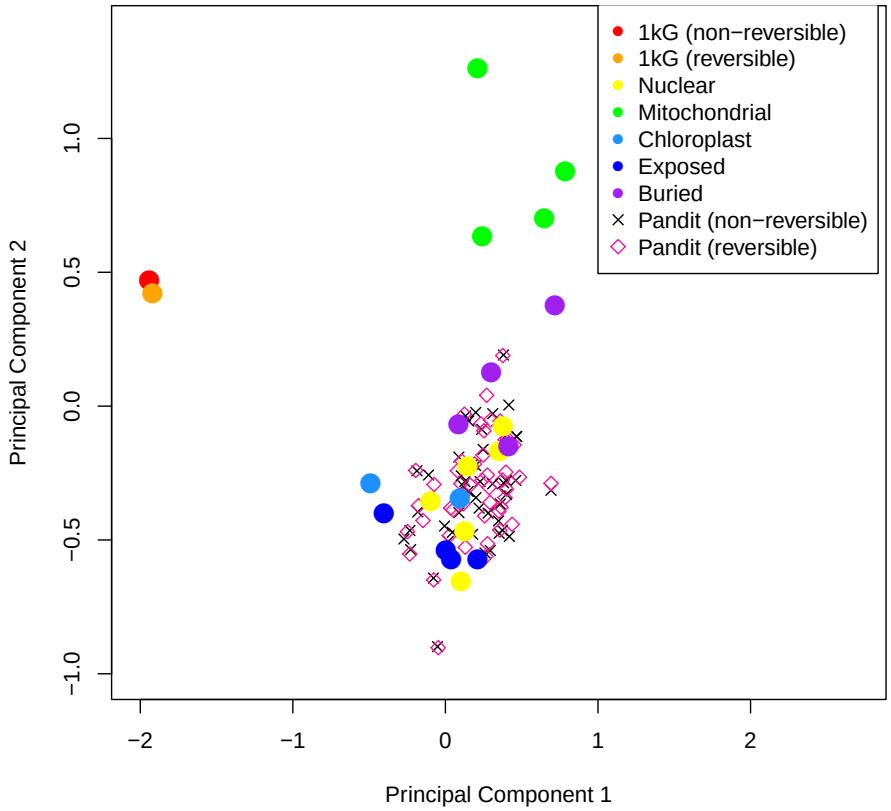


Figure 4.6: The first two components of a PCA of instantaneous rate matrices. Matrices included are 1kG (with and without assuming direction), nuclear (WAG (Whelan et al. 2001), JTT (Jones et al. 1992), LG (Le and Gascuel 2008), PAM (Dayhoff et al. 1978), tm126 (Liò and Goldman 1999), PCMA (Zoller and Schneider 2013)), mitochondrial (mtREV24 (Adachi and Hasegawa 1996), mtMam (Yang et al. 1998), mtArt (Abascal et al. 2007), mtZoa (Rota-Stabelli et al. 2009)), chloroplast (cpREV (Adachi et al. 2000), cpREV64 (Cox and Foster 2013)), exposed (alpha helix, beta sheet, coil, turn) (Goldman et al. 1998), buried (alpha helix, beta sheet, coil, turn) (Goldman et al. 1998), and Pandit (50 protein domains, with or without assuming direction) (Whelan et al. 2006). Principal components 1 and 2 represent 34% and 20% of the variance respectively. All other principal components represent 9% or less of the variance each.

4.4.2 Codon Models

I expect that one difference between the 1kG models and the other models is the amount of selection which can have occurred. To explore this idea I calculated a codon model for the 1kG data, using the same counting method as for the amino acid data, however including both synonymous and non-synonymous changes. This allows me to calculate a value of ω for the 1kG data, which is 0.8. This agrees with my expectation that the model should be close to selection-free.

Figure 4.7 shows the rates of change from one codon to another, dependent on whether (a) the change is in a CpG, (b) the change could be to a CpG but that CpG crosses two codons so it is not possible to tell from this data, or (c) the change does not affect a CpG. CpG changes have a much higher rate of change, with potential CpG changes also having a higher average rate than changes which do not affect CpGs. This is as expected as it is known that CpG dinucleotides in DNA tend to mutate at rates 10–50 times higher than other nucleotides. Interestingly this effect cannot be seen in other empirical codon matrices such as CodonPAM (Schneider et al. 2005), calculated by counting changes in a large number of alignments, or ECM (Kosiol et al. 2007), calculated using ML on a large number of alignments (results not shown).

4.4.3 Conclusions

In this subsection I have estimated a non-reversible model for human evolution from 1kG data and shown that this model is different to previously calculated between-species models. It is striking that a simple PCA analysis distinguishes the 1kG data so well, and that the second dimension is also easily interpreted. This suggests that there is scope in the future to extend this sort of analysis to better understand the effects of selection and of protein environment on evolutionary trends. While non-reversibility is not as big a difference as these other effects, the model has a higher ∇_A value than any of the pandit data sets analysed.

By calculating codon models I have investigated the possibility that the differences between the models are due to the lack of selective pressure in the 1kG model. This confirmed that the model does not contain much selective pressure. It also indicated that the CpG mutation rate is much higher than other mutation rates. This is a directional effect, present for both synonymous and non-synonymous SNPs, and so explains the non-reversibility of the amino acid model. This CpG effect is not seen in other empirical codon models, possibly indicating that CpG mutations occur at a very high rate and then are selected out, so that the effect is not seen as strongly when looking across multiple species.

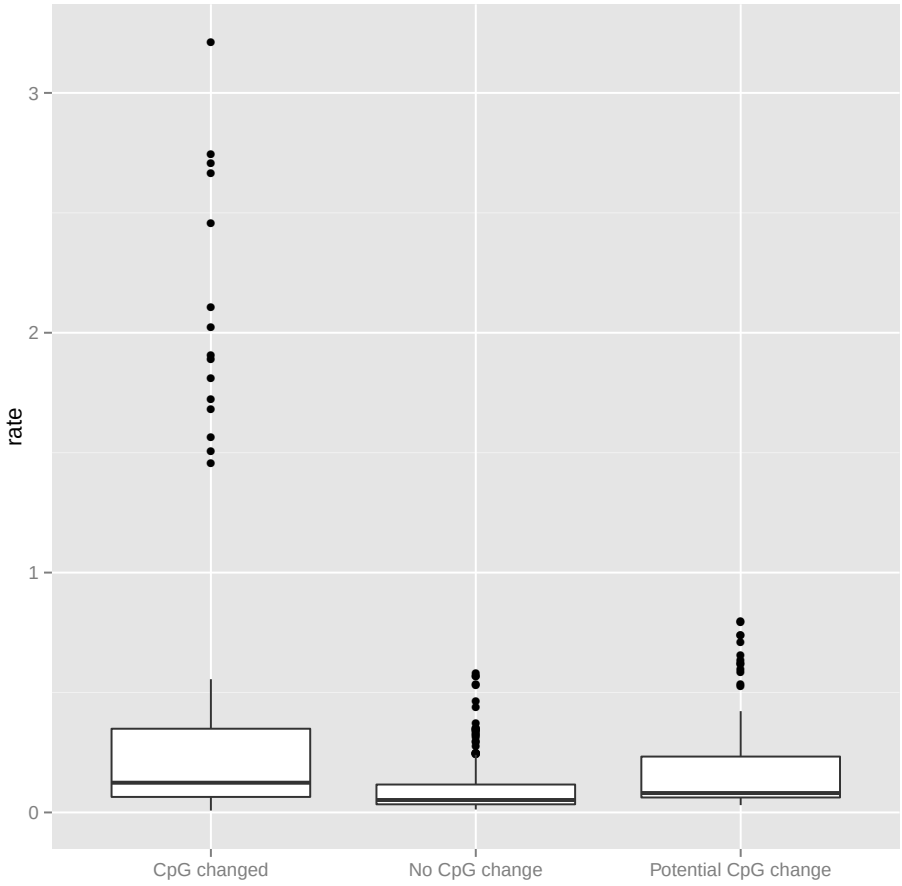


Figure 4.7: Dependence of mutation rates on the change in CpG status.

This could also explain why the ∇_A value for the within-species 1kG data is higher than those for the between-species Pandit data sets.

4.5 Discussion and Future Work

I have shown that non-reversible models may well be useful to provide a better model for evolution, particularly for nucleotide sequences. These better models of evolution improve our understanding of the processes of molecular evolution and could improve downstream analyses. Over 60% of the nucleotide MSAs in this study were significantly better described by a non-reversible model. Considering the facts that the power to detect non-reversibility depends on sequence length and that the MSAs used in this study were in general fairly short, this is a strong indication that non-reversible models may well be worth considering for nucleotide studies.

In general, for amino acid analyses, a new model is not built for each MSA; instead, empirical models built from large sets of sequence alignments are used (Whelan et al. 2001). This is because accurate reversible models are difficult to build for small amino acid MSAs due to the large number of parameters that need to be fitted. Non-reversible models are even more difficult to fit, due to computational issues in addition to the fact that they have nearly twice as many parameters. Currently, all of the standard empirical amino acid models are reversible. In this study, despite only using small MSAs, LRTs significantly favoured non-reversible models for some amino acid MSAs. For this reason it may be worth considering investigating non-reversible amino acid models derived from large numbers of sequence alignments.

Using the 1kG data I have derived non-reversible amino acid and codon models. These models are built from within-species data instead of between-species data, and therefore are expected to model mainly the mutation process, with only a small effect of selection. PCA showed that the 1kG amino add model is indeed very different from other instantaneous rate matrices, and the 1kG codon model confirmed that the model is close to being neutral. The codon model revealed a very strong CpG effect in the mutation data, which may explain the high level of non-reversibility in the amino-acid model.

I have compared three different measures of non-reversibility with the LRT, itself a measure of the significance of a non-reversible model over a reversible model. The measures of non-reversibility correlate strongly with one-another, showing that they capture similar information on non-reversibility. They correlate less strongly with the LRT statistic Λ, as expected, showing that the amount of non-

reversibility can be distinguished from our ability to detect non-reversibility. My preferred measure is the deviation from detailed balance, ∇_A, as this captures the most intuition about reversibility. It quantifies the deviation from the equations that define reversibility, as well as being interpretable as a measure of the distance between a non-reversible model and the closest reversible model. In addition, its calculation only requires the ML non-reversible model to be estimated.

In this chapter I have presented the evaluated value of the measure of non-reversibility as the 'true' amount of non-reversibility in a data set, whereas in fact it is an estimate, as the ML model used to calculate it will typically be an estimate. To assess the accuracy of the value of the measure, and to determine the variance, a bootstrap could be performed on the data, with the measure calculated for each bootstrap sample. A confidence interval could then be constructed, perhaps allowing a new way to test for whether the data set is non-reversible. In a traditional setting this would be done by testing whether the value indicating a reversible model, here 0, is included in the confidence interval. As 0 is the lower bound of the measure, it is unlikely that it will be reached, even if the model is truly reversible, so an alternate method would need to be developed to perform this test.

As well as improving the model of evolution, one reason for using non-reversible models would be in hope of improving the branch lengths and topology of inferred trees. The focus of this chapter has been to assess and analyse the quantification and significance of non-reversibility. Because of this, and due to the computational expense of finding the optimal topology, only one topology was analysed for each alignment. This means I have not explored the extent to which non-reversible models improve tree estimation. I have compared the branch lengths estimated using the two models, and found that the branch lengths of inferred trees are not greatly changed by the use of a non-reversible model (results not shown). These results indicate that it is unlikely that tree estimation will be greatly affected by using non-reversible models although it is not possible to say what would happen on larger data sets.

These results may indicate that for biologists interested mainly in divergence patterns, non-reversible models are unnecessary. However, often the process itself, and hence the parameters within the model, are of interest. For example, in the case study in Section 4.4, a motivation for building models of mutations that occur in healthy individuals, and those that are known to cause disease, is to be able to compare these models in order to help predict whether newly found mutations are disease-causing. Having a more accurate model of both of these processes is important to help with predictions. Another example is within

cancer cells, where modelling the mutational process may help us understand the development of different cancers. Alexandrov et al. (2013) built models for different cancer processes, and then inferred patterns of mutations which were attributed to different mutagenic processes. The specific parameters of interest in these models are the rates of change between bases; as there is no biological reason for expecting them to be reversible they are modelled non-reversibly. Another area where we are interested in the process itself is within healthy somatic cell lineages, where modelling the processes of genome change between successive generations of normal cells can help us reconstruct and understand development and ageing (Behjati et al. 2014).

One subject raised in the Introduction of this chapter was that of not knowing what "highly non-reversible" means. The adoption of a standard measure such as ∇_A would make it possible to define cut-offs for different levels of reversibility using the distribution of the measures for significant data sets found in this paper (Fig. 4.3) and knowledge of what might be observed by chance (Fig. 4.4). It should be noted, however, that LRTs should still be used to check for significant evidence of non-reversibility, as the size of the effect and the significance are not the same thing.

I have looked at the most general formulation for non-reversible models, where there are the maximum number of parameters. Analogous to the study of special cases of the nucleotide GTR model, such as those of Jukes and Cantor (1969), Kimura (1980), Felsenstein (1981), Hasegawa et al. (1985), Tamura (1992) and Tamura and Nei (1993), it may of course be possible to find non-reversible models that are significantly better than reversible models, but which have fewer free parameters. An example of a non-reversible model for nucleotide sequences with fewer parameters is the reverse complement symmetric model which accounts for the pairing between DNA strands in double-stranded organisms by setting the rate of substitution from one base to another equal to the rate of substitution between the conjugate of those two bases (Wu and Maeda 1987; Lobry 1995). This model has fewer parameters than the general non-reversible model (and, indeed, the same number as the GTR model). This may be enough to model non-reversible evolution in organisms with double-stranded DNA.

One issue in the use of non-reversible models is the availability of software that can estimate them. For nucleotide sequences, baseml in PAML (Yang 2007), HyPhy (Kosakovsky Pond et al. 2005) and nhPhyML (Boussau and Gouy 2006) can estimate non-reversible models. Of these, only nhPhyML is designed to perform tree search. For amino acid sequences, only HyPhy can estimate non-reversible models and there are no programs that can estimate non-reversible

models and perform tree search. This problem does not just affect estimation programs: jModelTest2 (Darriba et al. 2012), a popular program used to test models and choose the best model for each MSA, also does not include a non-reversible model. My results show that, to understand the process of evolution, non-reversible models are often better. Further work needs to be carried out to improve our ability to do this.

Chapter 5

Positive Selection

At the time of writing this thesis, the work presented in this chapter is being prepared for submission to *Molecular Biology and Evolution*.

5.1 Introduction

Detecting protein sites under positive selection is a key question in biology, as positively-selected sites are good candidates for the causes of species differences, or for functional, or medical, significance. For example, tests have been used to find sites which may be in an evolutionary arms race (Yang et al. 2000); to find sites under positive selection in mammals, helping to improve our understanding of mammal evolution (Lindblad-Toh et al. 2011); in combination with other methodology to help select sites of possible phenotypic effect (Conde et al. 2006); or to find sites in HIV which have evolved differently in different populations (Kosakovsky Pond et al. 2006a).

In the study of protein-coding DNA sequence, selective pressure is frequently inferred by consideration of the ratio of the rates of fixation of nonsynonymous and synonymous mutations. Synonymous substitutions do not change the amino acid, and are often assumed not to affect the fitness of the protein and to be selectively neutral. If nonsynonymous changes are also selectively neutral, they will be fixed at the same rate, so the nonsynonymous/synonymous ratio (ω, otherwise known as d_N/d_S) will be 1. If nonsynonymous substitutions improve the fitness of the protein, meaning that they occur more often than synonymous substitutions, then $\omega > 1$, suggesting positive selection has occurred. If nonsynonymous substitutions reduce the fitness of the protein then $\omega < 1$, suggesting purifying selection.

A number of methods have been proposed for estimating ω in an aligned set of inter-species protein-coding sequences. Initial methods focussed on counting

nonsynonymous and synonymous changes between pairs of sequences (e.g. Li et al. 1985; Nei and Gojobori 1986). These had the limitation that only two sequences could be analysed at a time, and that ω was assumed to be the same for all the sites of an entire gene sequence. These limitations were overcome by the development of methods that took into account the phylogeny (Goldman and Yang 1994; Muse and Gaut 1994) and allowed ω to vary between sites ('site models') (Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yang et al. 2000). These methods also allow users to perform statistical tests of the level of selection. Further methods have been developed that allow for ω to be different on certain branches of the phylogeny, either having one ω value for a whole gene ('branch models') (Yang 1998) or allowing for site-wise variation as well as lineage variation ('branch-site models') (Yang and Nielsen 2002).

In this chapter I focus on using site models to test for sites evolving under positive selection throughout a gene phylogeny. These models are useful for detecting diversifying selection, the fixation of a succession of amino acid changes (see Section 2.1.4). They assume that codons evolve independently without any context-dependent effects, and often also that synonymous substitutions are neutral and all occur at the same constant rate. The first assumption is unrealistic; however, as with nucleotide and amino acid models, it is very difficult to relax (see Section 1.1). The second assumption is held in the majority of models used to detect positive selection. This assumption is known to be violated in many cases, for example due to codon usage bias (Sharp and Li 1987), but this violation may not always be large. For example, in mammals, codon bias is known to be weak (Plotkin and Kudla 2011). It is also known that certain regions of genes will be under different levels of bias (Plotkin and Kudla 2011); these regions can therefore be analysed separately. While the relaxation of some of the more restrictive assumptions could be a fruitful area for further research, the detection of site-specific selection over phylogenies has been successful and useful based on existing models and methods, and the work I will describe increases their utility by improving the statistical power of existing hypothesis testing approaches. Two main modelling approaches have been developed for site models: random effects likelihood (REL) models, and fixed effects likelihood (FEL) models.

5.1.1 REL Models

REL models infer a gene-specific distribution for ω, from which each site is assumed to be drawn independently (Nielsen and Yang 1998; Kosakovsky Pond and Muse 2005; Murrell et al. 2013). A popular example of this is the method

introduced by Nielsen and Yang (1998), implemented in the codeml program in PAML (Yang 2007). In a two-part procedure, the appropriateness of a model incorporating positive selection is tested by performing an LRT between two nested models, only one of which allows for positive selection. One possible such pair are the models M8A and M8, which describe the distribution of ω over sites as a mixture between a beta distribution (allowing only $\omega \in [0, 1]$) and a point mass either at $\omega = 1$ (neutral evolution, model M8A) or at any value ≥ 1 (positive selection, M8; Swanson et al. 2003). If the LRT is significant in favour of the model permitting positively selected sites, Bayes Empirical Bayes posterior probabilities of each site being positively selected are then checked and can be used to infer individual sites evolving under positive selection (Yang et al. 2005).

A fast approximate Bayesian REL method, FUBAR, has been developed by Murrell et al. (2013). This method does not assume that the synonymous substitution rate is constant over sites, and instead estimates the rates of nonsynonymous and synonymous change at each site and assesses whether the nonsynonymous rate is greater than the synonymous rate.

In both of these methods specific sites are determined as being under positive selection if their posterior probability is higher than a given threshold. There is no formal control of the FPR from using posterior probabilities, although for any chosen threshold there will of course be some FPR. By analogy to the terminology used in statistical hypothesis testing, it is convenient in what follows to describe methods based on posterior probabilities as conservative if they appear to sacrifice power because of having a lower false positive rate than might be acceptable.

5.1.2 FEL Models

In FEL models, ω is inferred independently for each site (Kosakovsky Pond and Frost 2005; Massingham and Goldman 2005). LRTs are then carried out on each site, comparing the hypotheses $H_0 : \omega = 1$ and $H_1 : \omega > 1$. One such method, SLR (Massingham and Goldman 2005), assumes codons evolve independently under a continuous-time Markov process of single-nucleotide substitutions, as described in Section 2.1.4. The tree topology, branch lengths, equilibrium codon frequencies and the transition-transversion rate ratio are assumed to be constant over all sites and are estimated using all of the sites under the null model of neutrality (i.e. all sites have $\omega = 1$). These parameters are then held constant while site-wise likelihoods are calculated for the null model ($\omega = 1$) and the alternative model ($\omega > 1$).

To test for positive selection, a site-wise LRT of the null distribution, $H_0 : \omega =$

1 against the alternative $H_1 : \omega > 1$ is performed. The site-wise LRT statistic for this test is

$$\Lambda = 2 \log \left[\frac{l(\hat{\omega})}{l(1)} \right]$$

where $l(\omega)$ is the site-wise likelihood, and $\hat{\omega}$ is the $\omega > 1$ which maximises $l(\omega)$. This is compared to a 50:50 mixture of a χ^2_1 distribution and a point mass at 0 (Self and Liang 1987) (denoted $\bar{\chi}^2_1$ as in Goldman and Whelan 2000). Further details can be found in Massingham and Goldman (2005).

SLR assumes synonymous substitution rates are constant over sites; however it is possible to let them vary, as in Kosakovsky Pond and Frost (2005).

5.1.3 Comparing REL and FEL Models

FEL models have the advantage over REL models that they do not make assumptions about how the level of selection varies along sequences, but they have the disadvantage that knowledge about the whole ω distribution cannot be used to determine whether each site is under positive selection. Studies to date have shown existing REL and FEL methods perform fairly similarly (Massingham and Goldman 2005; Kosakovsky Pond and Frost 2005). However, they are all statistically conservative, having a lower false positive rate (FPR) than expected (Anisimova et al. 2002; Massingham and Goldman 2005; Yang et al. 2005; Murrell et al. 2013). Consequently the power, or true positive rate (TPR), is lower than could be achieved, meaning that many positively selected sites may be missed at any given nominal FPR. A less conservative method could allow us to find these sites.

5.1.4 Why is SLR Conservative?

In this chapter I develop a new site-wise test, the gSLR test, inspired by the SLR method of Massingham and Goldman (2005), which is less conservative and achieves higher power. To motivate this new test we need to understand why SLR is conservative. In SLR, hypotheses $H_0 : \omega = 1$ and $H_1 : \omega > 1$ are assessed for each site. This is done by comparing the LRT statistic with $\bar{\chi}^2_1$. When data that conform to the null hypothesis H_0 are considered, this test has the appropriate FPR (Massingham and Goldman 2005). Typical proteins, however, do not consist of all neutral sites, but have many sites under purifying selection. Figure 5.1a illustrates this, showing the distribution of ω estimated from a genome-wide scan of 38 mammals (Jordan 2011). More than 90% of these sites have estimated $\omega < 1$. This is clearly different from the null ($H_0 : \omega = 1$) assumed by the SLR

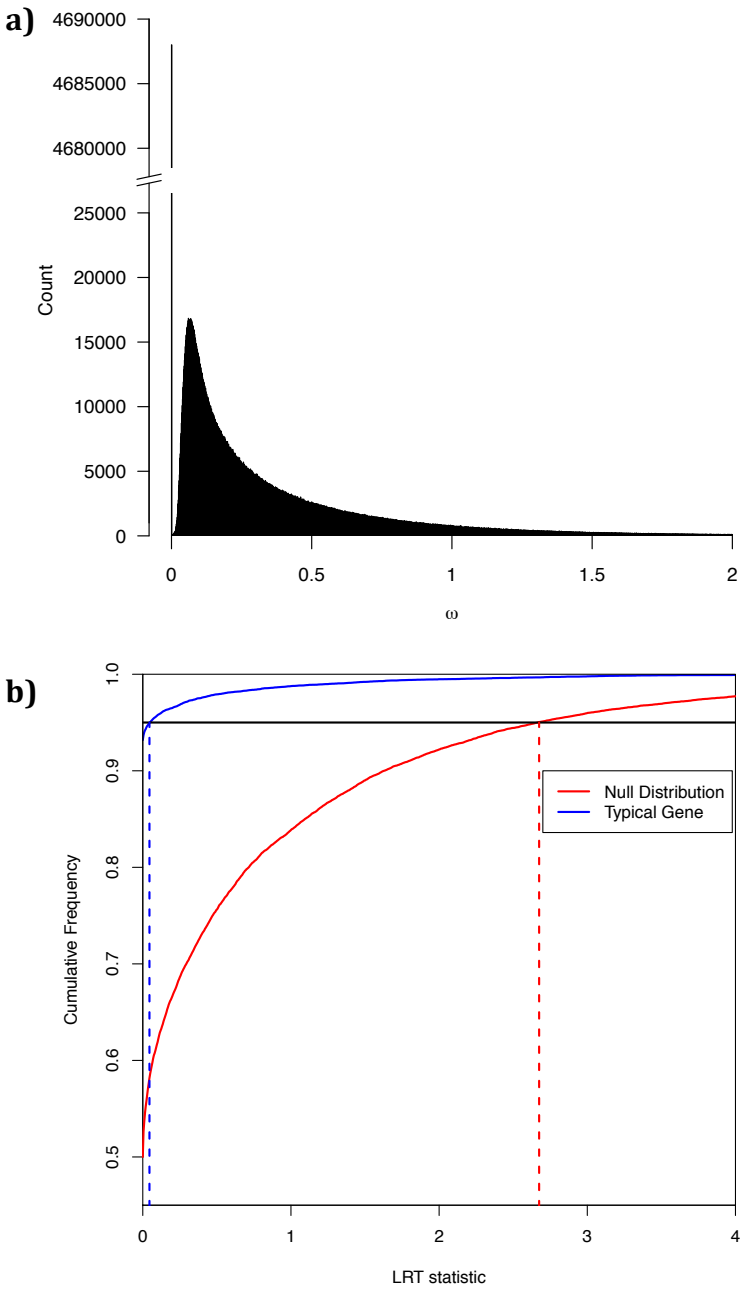


Figure 5.1: **a** The distribution of ω estimated from a genome wide scan of 38 mammals (Jordan 2011). **b** Cumulative frequencies of the LRT statistic for the null distribution of SLR (red) and for simulated data sets with the ω-distribution shown in Figure 5.1**a**, but without any positive selection (no sites with $\omega > 1$; blue). The horizontal line crosses the curves at the 5% FPR, corresponding to LRT statistics of 2.7 for the SLR null (red dashed line) and 0.065 for a typical gene data set (blue dashed line).

test.

How does the difference between the null distribution of the test and the ω-distribution of the data affect the test? Figure 5.1b shows the cumulative frequency of LRT statistics for data simulated under the null hypothesis and data simulated with the ω-distribution of the mammal data set shown in Figure 5.1a, but without any positive selection (no sites with $\omega > 1$). This latter distribution is representative of the sites of a typical gene not under positive selection. The large difference between these two distributions clearly indicates that for a typical gene data set, the $\bar{\chi}^2_1$ distribution is not optimal for performing the site-wise LRTs. For the null distribution, a critical value of 2.7 for the LRT statistic corresponds to an FPR of 0.05. For a typical gene data set this LRT statistic corresponds to a much lower FPR of 0.005, and to achieve an FPR of 0.05 the LRT statistic threshold can be reduced to 0.065. Stated another way, if a null hypothesis threshold of 2.7 is used on a typical gene data set then all sites with an LRT statistic between 0.065 and 2.7, which could be inferred as positively selected without raising the FPR above 0.05, will not be. Consequently the test will be conservative, with any positively selected sites with an LRT statistic between 0.065 and 2.7 remaining undetected. Another way of looking at this is shown in Figure C.1.

In short, SLR is conservative because its null distribution does not fit typical data. In this chapter I develop a new method, the gSLR test, in which I have altered the null distribution so that it may fit the observed data. Point estimates of ω are unaltered, but statistical significance is determined using either a parametric bootstrap, or a χ^2-mixture. In order to control the FPR even in unfavourable situations, I also develop diagnostics to decide when this new test is appropriate. I show that it is applicable to a large proportion of protein sequence alignments, and often doubles the statistical power available with SLR.

5.2 gSLR Test

5.2.1 gSLR Test Statistic

SLR is conservative because the null distribution does not account for the large proportion of sites in most genes that are under purifying selection. To improve the power of the test, I change the null hypothesis to recognise that site-wise ω values may be less than 1. In other words I alter the test so that, instead of comparing hypotheses $H_0 : \omega = 1$ and $H_1 : \omega > 1$, I compare $H_0 : \omega \leq 1$ and $H_1 : \omega > 1$. This change means that instead of comparing a simple hypothesis (the distribution of the data is fully specified) with a composite hypothesis (the

distribution of the data is not completely specified), I am comparing two composite hypotheses. I therefore use the generalised likelihood ratio (gLR; Wilks 1938) instead of the likelihood ratio used in the SLR method (see Methods). Wilks's gLR statistic, δ_g, compares the maximised likelihood of the null model to the likelihood maximised over both the alternative and null models:

$$\delta_g = 2 \log \left[\frac{\sup_{\omega \in [0,\infty)} l(\omega)}{\sup_{\omega \in [0,1]} l(\omega)} \right] = 2 \log \left[\frac{l(\hat{\omega})}{l(\hat{\omega}_0)} \right]$$

where $\hat{\omega}$ is the $\omega \in [0, \infty)$ that maximises the site-wise likelihood $l(\omega)$, and $\hat{\omega}_0$ maximises $l(\omega)$ over $\omega \in [0, 1]$. If $\hat{\omega} < 1$ then necessarily $\hat{\omega}_0 = \hat{\omega}$ and $\delta_g = 0$.

5.2.2 Parametric Bootstrap Test of Significance

In order to decide whether a certain value of δ_g is significant I need to know how probable that value is under the null hypothesis. Because for my new test statistic the null distribution is dependent on the ω-distribution of the specific data set being analysed, it must be determined independently for each data set. Cox (1961) proposed that, because under the null hypothesis the best explanation of the data is that parameters of interest take their maximum likelihood (ML) values, these values can be used to calculate the distribution of δ_g when a more-general result is not available. This method was tested and found to be satisfactory by Cox (1962) and Lindsay (1974a,b). In practice it can be performed using Monte Carlo methods as shown by Goldman (1993); in a phylogenetic context this is generally called a 'parametric bootstrap' (Felsenstein 1988, also see Section 2.4).

Following the usual procedures for SLR (Massingham and Goldman 2005), an ML value of ω for each site is estimated under the null hypothesis (the values $\hat{\omega}_0$). To determine the distribution of gLR statistics if this distribution were true, I discretise the $\hat{\omega}_0$-distribution into 10 equal width bins and simulate 100 data sets with this $\hat{\omega}_0$-distribution using EvolverNSsites (Yang 2007). Simulations are performed using the tree, κ value and equilibrium codon frequencies inferred from the original data (see Section 5.1.2). Site-wise gLR statistics are estimated for all bootstrap samples, and the value of δ_g from the original data set compared to these to determine significance. If there were gaps in the original alignment then the same pattern of gaps is mapped onto each of the bootstrap alignments, so that the bootstrap samples do not contain more data than the original data set (Goldman et al. 1998).

This method can be time-consuming as it involves simulating 100 new data sets, and estimating branch lengths and site-wise ω values for each of these data

sets independently. Therefore, I have developed an alternative approach which does not require bootstrapping.

5.2.3 χ^2-Mixture Distribution Test of Significance

Note that the SLR distribution of a 50:50 mixture of χ^2_1 and point mass at 0 is the correct distribution for the special case of the new test that the data do in fact conform to the null of SLR ($\omega = 1$). This $\bar{\chi}^2_1$ distribution arises because in this case half of the data have $\hat{\omega} \leq 1$ and hence $\delta_g = 0$, and the other half will have $\hat{\omega} > 1$ and $\delta_g > 0$. The distribution of δ_g in the first half is a point mass at zero; for the second half, δ_g will be χ^2_1-distributed because δ_g in the new test is the same as the LRT statistic in SLR (see Section 5.1.2). The 50:50 mixture $\bar{\chi}^2_1$ is therefore appropriate to give the probability of a certain value of δ_g under the null model when the data conforms to the null of SLR (Massingham and Goldman 2005).

In general in the new test, data do not conform to the null of SLR, but have many sites under purifying selection. Therefore I no longer expect half of the distribution to have $\delta_g = 0$ and half to be χ^2_1-distributed: I now expect more than half the sites to have $\delta_g = 0$. I can therefore construct a mixture distribution that is $x\%$ a point mass at 0 and $(100 - x)\%$ a χ^2_1 distribution, where x is estimated from the data as the percentage of samples which have $\hat{\omega} \leq 1$. The attained value of δ_g can then be compared with this distribution to determine significance.

5.2.4 Confidence Intervals

An alternative way of thinking about significance tests is to look at the confidence interval (CI) around the estimated ω value for a site (see Section 2.4). For example, for tests at the 5% FPR level, I wish to estimate a CI so that there is a 95% probability that the interval encompasses the real ω value at that site. When testing for positive selection, if the confidence interval includes $\omega = 1$ then the test is not significant; if the lower bound of the confidence interval is greater than 1 then the test is significant. Both the parametric bootstrap and χ^2-mixture variants of the gSLR test give disciplined ways of altering the critical LRT value used to determine significance, and it is possible to alter the confidence intervals of ω estimates to reflect this.

Application of standard likelihood theory (see Silvey 1975; Yang 2006, and Section 2.4) to the original SLR test means that CIs are calculated by finding the values either side of $\hat{\omega}$ that correspond to the ML value minus half of the LRT value used for determining significance, which for a 95% confidence interval

is 1.35 (Massingham and Goldman 2005). The region between these contains those ω values that could not be distinguished from the ML estimate at the 5% significance level. Analogously, for the gSLR test I use the same procedure but replace 1.35 with half of the 95% point of the distribution under the null in the new test being used. Note that, as I am testing for $\omega > 1$, the lowest possible lower bound of the confidence intervals is 1.

5.2.5 Limitations and Diagnostics

If the ω-distribution estimated from the data is different from the true ω-distribution, the new test may not have a controlled FPR. Fortunately, two situations in which this can occur can be identified from the data; any data sets so identified can be analysed using a test known to have a controlled FPR, such as SLR.

The first case where FPR can become inflated occurs when trees are short. In this case there will have been limited time for substitutions to occur, meaning that many sites will have no nonsynonymous changes. For these sites $\hat{\omega} = 0$, even though the real value of ω may be higher. The $\hat{\omega}$-distribution will therefore be shifted to the left. When parametric bootstrap samples are simulated using this $\hat{\omega}$-distribution there will be further sites where $\omega \neq 0$ but $\hat{\omega} = 0$ (see Fig. 5.2a). This shifting of the $\hat{\omega}$-distributions to the left (Fig. 5.2a) causes the FPR of the bootstrap variant of the new test to be higher than intended.

Short trees also cause problems for the χ^2-mixture variant of the gSLR test. When most sites have no changes, then any sites with nonsynonymous changes tend to have a very high $\hat{\omega}$ (see Fig. 5.2a). The site-wise $\hat{\omega}$-distribution will have a large peak at 0, a few sites with $0 < \hat{\omega} < 1$, and some sites with $\hat{\omega} \gg 1$. $2\delta_g$ for the sites with $\hat{\omega} \gg 1$ is not χ^2_1-distributed; in fact, the χ^2_1 distribution is too far to the left and the FPR again is higher than desired.

To identify data sets where trees are too short I can use the fact that many sites will then incorrectly have $\hat{\omega} = 0$. I propose that data sets with too high a proportion of sites with $\hat{\omega} = 0$ may be analysed using SLR instead of the new test. Discussion of how a threshold can be chosen for this is given in Section 5.3.1.

The second case of a potentially inflated FPR only affects the parametric bootstrap variant of my test and occurs if the data are close to the SLR null distribution ($\omega = 1$ for all sites). This is very rare in real data, but if many sites in a data set have $\omega = 1$ then approximately half will have $\hat{\omega} > 1$, and half $\hat{\omega} < 1$. However, under the null hypothesis that $\omega \leq 1$, half of the sites will have $\hat{\omega} < 1$ and half $\hat{\omega} = 1$. This is very different from the real distribution, being

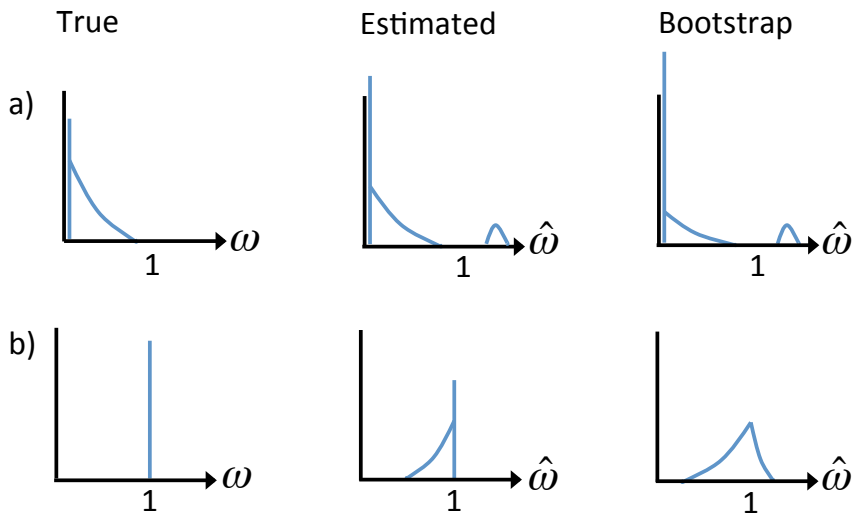


Figure 5.2: Two examples of possible true ω-distributions, along with their estimates under the null and the estimated distribution of the resulting bootstrap, showing pathological conditions in which the three distributions can differ. **a** When the tree is short many more of the sites will have $\hat{\omega} = 0$ in the estimated distribution (middle) than in the real distribution (left); this will be further exaggerated for the bootstrap (right). Any sites with nonsynonymous changes will tend to have inflated $\hat{\omega}$, and hence there may be some sites with $\hat{\omega} > 1$ in the estimated (middle) and bootstrap (right) distributions, even if there are none with true $\omega > 1$ (left). **b** If all sites are neutral (left) then the estimated distribution under the null will have approximately half of the sites with $\hat{\omega} = 1$ and half with $\hat{\omega} < 1$ (middle). When the bootstrap is then carried out there will be more sites with $\hat{\omega} < 1$ than with $\hat{\omega} > 1$ (right).

shifted to the left (Fig. 5.2b). When bootstrap samples are produced based on this $\hat{\omega}$-distribution, they will have more sites with $\hat{\omega} < 1$ than with $\hat{\omega} > 1$; there will therefore be more than half of the sites with gLR statistics $\delta_g = 0$, and fewer than half with $\delta_g > 0$. Again this shift to the left causes the FPR to be inflated. This effect will decrease as more species are added to the tree and tree length is increased, but the perfect result of a 5% FPR will only occur when site-wise $\hat{\omega}$ estimates are perfect (see Section 2.3.1).

The proportion of sites with $\hat{\omega} > 1$ can be used to determine whether a given data set is similar to the null of the SLR test. The χ^2-mixture variant can then be used as an alternative to the bootstrap variant, as its FPR is unaffected by data being close to the SLR null distribution ($\omega = 1$ for all sites) and it is very similar to the original SLR test in these circumstances. Again, discussion of how a threshold can be picked for this is given in Section 5.3.1.

5.2.6 Workflow for the gSLR Test

The workflow for the new test for positive selection, along with diagnostics to decide whether the new test or the original SLR method should be used, is shown in Figure 5.3. ω is estimated for each site in the data set and the value of the short tree diagnostic is determined. If the diagnostic is satisfied then, depending on time constraints and other preferences, the χ^2-mixture or the bootstrap variant is chosen. If the bootstrap variant is desired then another diagnostic must be checked to ensure that the $\hat{\omega}$-distribution is appropriate for the test. If so, the bootstrap variant can be performed. If the first diagnostic is not satisfied then the original SLR test can be used; if the second diagnostic is not satisfied then the χ^2-mixture variant can be used instead of the bootstrap variant.

5.2.7 False Discovery Rate Correction

In this chapter I concentrate on creating a test which is both powerful, having a high TPR, and controlled, meaning that the user-defined FPR does not exceed a pre-specified value. As with previous papers (Nielsen and Yang 1998; Anisimova et al. 2001, 2002; Wong et al. 2004; Massingham and Goldman 2005; Murrell et al. 2013), performance is assessed by focussing on the FPR and TPR.

In practice, detection of sites with a user-defined expected false discovery rate (FDR), rather than a user-defined FPR, may be desired, e.g. if performing a genome-wide scan. A number of FDR correction methods have been developed (e.g. Benjamini and Hochberg 1995; Storey 2003; Efron 2010). I have tested these and find that on our simulations they yield largely the same results (results not

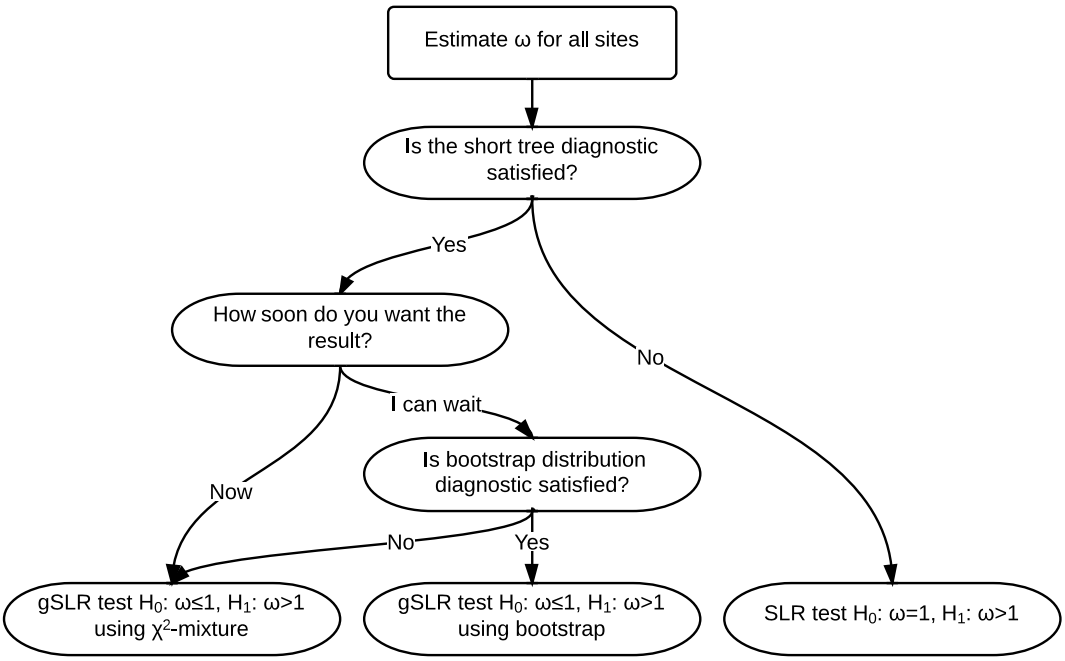


Figure 5.3: Workflow for the new test

shown). Benjamini-Hochberg FDR correction (Benjamini and Hochberg 1995) is used in this chapter to illustrate different methods' performance under FDR constraints.

5.3 Results

I assessed the new gSLR test in a range of simulations, aiming both to investigate how it performs on ω-distributions similar to those in real proteins and to check the FPR is controlled, including in difficult situations. The distributions used were: (A) the M8 model from codeml with $p_0 = 0.9432$, $\omega = 2.081$, $p = 0.572$, and $q = 2.172$ (as in Anisimova et al. 2001, 2002; Massingham and Goldman 2005); (B) the distribution of Figure 5.1a, derived empirically by fitting a large number of categories to ω values of mammal alignments used in Jordan (2011); (C) a mixture distribution taking the value $\omega = 0.5$ with probability 0.75 or else $\omega = 1.5$ (as in Massingham and Goldman 2005); (D) a mixture distribution taking the value $\omega = 1$ with probability 0.9432 or else $\omega = 2.081$ (as in Massingham and Goldman 2005, cf. case A); and (E) the null model for SLR of a point mass at 1 (as in Massingham and Goldman 2005).

Cases A and B were chosen to represent realistic examples of positive selection. Codeml is expected to perform particularly well on case A as a beta distribution is the assumed model for models M8A and M8. Case C was chosen as a problem

known to be difficult for all tests; cases D and E were chosen to represent the most difficult cases for the bootstrap variant of the new test, which are when the data are close to neutral (see Section 5.2.5). In case D, there are positively selected sites that I hope to be able to distinguish from a more conserved background; in case E, representing pure neutral evolution, there are none.

Twelve different trees were used for simulations, comprising four different topologies, each with three different branch length scalings. Three of these topologies were taken from previous studies of positive selection as useful examples of a range of possible situations; a 44-species tree was also included so that power on large trees could be tested. On each topology, branch lengths were scaled to make a small, medium and large version of the tree, with scalings taken from previous papers where available. Figure 5.4 shows the four different topologies with the medium version of branch lengths for each topology indicated; Table 5.1 shows the tree length scalings used for the different topologies.

Simulations were performed using evolverNSsites (Yang 2007) to produce sequences with a length of 200 codons. For all simulations, a transition-transversion rate ratio of 2 was used. Provided sequences are sufficiently long to accurately estimate topology, branch lengths, equilibrium codon frequencies and the transition-transversion rate ratio, sequence length does not affect accuracy of SLR, so it was not varied (Massingham and Goldman 2005).

Table 5.1: Tree lengths in units of substitutions per site for the three different sizes of the four topologies.

	Small	Medium	Large
6-species	0.11	1.1	11
12-species	0.202	2.02	20.2
17-species	2.10	8.43	16.86
44-species	2.85	11.42	22.84

I compare the performance of the new gSLR test with the original SLR test (Massingham and Goldman 2005), codeml using comparisons of model M8A and model M8 (Swanson et al. 2003; Yang 2007), and FUBAR (Murrell et al. 2013). The new test and the original SLR test are designed to identify sites under positive selection with an FPR chosen by the user: sites are inferred to be under positive selection based on p-values of LRTs. Codeml and FUBAR calculate the posterior probability that a site is under positive selection; a posterior probability threshold is then used to infer positively selected sites. Posterior probabilities and p-values are not equivalent; I follow standard procedures for each approach and have selected commonly used cut-offs of 0.05 for p-values, and 0.95 for posterior

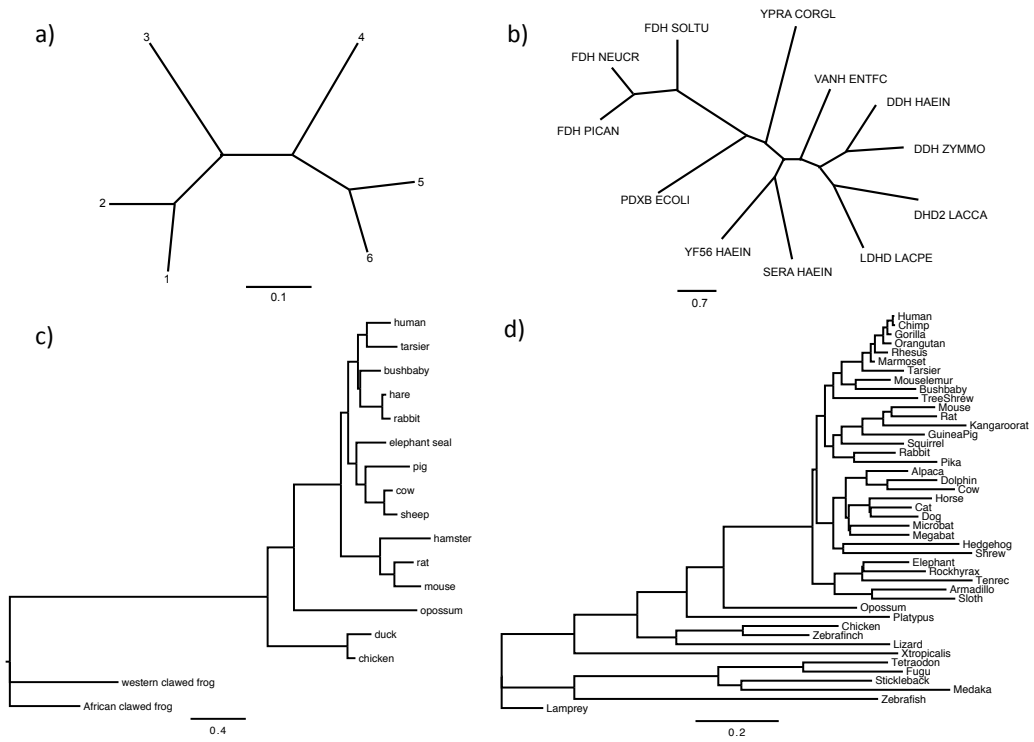


Figure 5.4: Trees used in analyses: **a** Artificial 6-species tree (Anisimova et al. 2001; Massingham and Goldman 2005), **b** 12-species tree of 2-hydroxyacid dehydrogenase (Massingham and Goldman 2005), **c** β-globin tree for 17 vertebrate sequences (Yang et al. 2000; Anisimova et al. 2001), **d** 44-species tree used by the ENCODE project (Birney et al. 2007; Nikolaev et al. 2007). Branch lengths were scaled to make a small, a medium and a large version of each tree: tree lengths are shown in Table 5.1. Branch lengths shown here are the medium version of each tree.

probabilities.

Figures 5.5–5.7 show the FPRs, TPRs, and the number of data sets which passed the diagnostics for the new tests, for the 12 trees under each of the five different selection scenarios. Focusing first on simulation cases A and B (Fig. 5.5) we note that FPRs are higher for both variants of the new test than for any other methods, as expected, but remain below 5% as desired. The power (TPR) is greatly improved for all trees where the new test is used, often doubling. Power is lower for case B than for case A because many of the sites are under weak positive selection in case B, whereas in case A all positively selected sites are under fairly strong positive selection. The χ^2-mixture and bootstrap variants of the new test give similar increases in power, with the χ^2-mixture tending to be slightly more conservative than the bootstrap. Both variants perform better than SLR, which performs better than codeml; all of these methods perform better than FUBAR.

Cases C and D were chosen as difficult cases; they are not, however, very realistic. In case C both variants of the new test again have a controlled FPR below 5% but higher than other tests (Fig. 5.6). There is an increase in power over other tests, although the increase is less than before. In case D, both variants of the new test and SLR perform equally well, having a controlled FPR and good power (Fig. 5.6). Again all are better than codeml and FUBAR. Case E (Fig. 5.7) shows that, when there are no positively selected sites to find, the new test is still controlled.

With only a few slight exceptions in the most difficult cases, the new gSLR test's FPR is controlled in all cases studied. The suggested diagnostic tests successfully guard against cases where the new tests could fail: for the two very short trees (6-small and 12-small) they almost always indicate that SLR should be used in preference to either variant of the new test. For all other trees the χ^2-mixture variant can be used all the time. The bootstrap variant could always be used for the realistic simulations (cases A and B), provided the tree is not too short. It is excluded more as data sets become more similar to the null of the SLR test; for

		FPR					TPR					Number of datasets that passed the diagnostics	
Case A		SLR	gSLR + bootstrap	gSLR + χ^2-mixture	Codeml M8a vs M8	FUBAR	SLR	gSLR + bootstrap	gSLR + χ^2-mixture	Codeml M8a vs M8	FUBAR	gSLR + bootstrap	gSLR + χ^2-mixture
	Small	0.0093	0.0093	0.0093	0.0000	0.0000	0.0852	0.0852	0.0852	0.0014	0.0021	0	0
6-species tree	Medium	0.0024	0.0234	0.0183	0.0000	0.0002	0.2394	0.5905	0.5494	0.0410	0.0620	100	100
	Large	0.0001	0.0261	0.0235	0.0000	0.0000	0.0052	0.5075	0.4731	0.0007	0.0000	100	100
	Small	0.0040	0.0040	0.0040	0.0000	0.0001	0.1343	0.1343	0.1343	0.0050	0.0128	0	0
12-species tree	Medium	0.0014	0.0206	0.0165	0.0001	0.0001	0.4053	0.7709	0.7444	0.1514	0.1136	100	100
	Large	0.0006	0.0183	0.0160	0.0002	0.0002	0.0931	0.6924	0.6706	0.0210	0.0044	100	100
	Small	0.0021	0.0200	0.0168	0.0001	0.0002	0.3804	0.7628	0.7389	0.1256	0.1131	100	100
17-species tree	Medium	0.0006	0.0156	0.0108	0.0003	0.0004	0.5082	0.9230	0.8879	0.3328	0.2342	100	100
	Large	0.0004	0.0138	0.0090	0.0002	0.0004	0.5090	0.9328	0.8957	0.3632	0.2147	100	100
	Small	0.0024	0.0205	0.0163	0.0004	0.0001	0.5375	0.8799	0.8581	0.3055	0.1595	100	100
44-species tree	Medium	0.0009	0.0103	0.0075	0.0001	0.0004	0.8413	0.9813	0.9772	0.7480	0.3488	100	100
	Large	0.0002	0.0079	0.0053	0.0001	0.0006	0.8867	0.9943	0.9898	0.8318	0.4077	100	100
Case B													
	Small	0.0115	0.0115	0.0115	0.0000	0.0002	0.0595	0.0595	0.0595	0.0000	0.0011	0	0
6-species tree	Medium	0.0042	0.0283	0.0204	0.0001	0.0003	0.1432	0.3767	0.3268	0.0042	0.0226	100	100
	Large	0.0003	0.0351	0.0210	0.0000	0.0000	0.0099	0.4432	0.3118	0.0000	0.0000	100	100
	Small	0.0073	0.0073	0.0073	0.0001	0.0002	0.0783	0.0783	0.0783	0.0029	0.0051	0	0
12-species tree	Medium	0.0034	0.0227	0.0185	0.0001	0.0003	0.1780	0.4426	0.3940	0.0268	0.0486	100	100
	Large	0.0010	0.0254	0.0179	0.0002	0.0003	0.0782	0.5378	0.4467	0.0044	0.0022	100	100
	Small	0.0033	0.0239	0.0188	0.0001	0.0003	0.1900	0.4473	0.4019	0.0337	0.0507	100	100
17-species tree	Medium	0.0017	0.0206	0.0144	0.0004	0.0005	0.2437	0.5975	0.5406	0.1012	0.0873	100	100
	Large	0.0009	0.0184	0.0126	0.0001	0.0006	0.2647	0.6334	0.5754	0.1134	0.1027	100	100
	Small	0.0028	0.0233	0.0173	0.0002	0.0002	0.2503	0.5369	0.4932	0.0820	0.0719	100	100
44-species tree	Medium	0.0022	0.0189	0.0131	0.0009	0.0008	0.3891	0.7181	0.6607	0.2654	0.1605	100	100
	Large	0.0009	0.0152	0.0084	0.0007	0.0009	0.4146	0.7483	0.6796	0.3450	0.1373	100	100

Figure 5.5: FPRs and TPRs for the five tests compared in this paper, and the number of data sets that passed the diagnostics for the new tests, for simulation cases A and B and 12 different trees. No FPRs exceed 0.05. TPRs are shaded from white to dark green as values increase from 0 to 1.

		FPR					TPR					Number of datasets that passed the diagnostics	
Case C		SLR	gSLR + bootstrap	gSLR + χ^2-mixture	Codeml M8a vs M8	FUBAR	SLR	gSLR + bootstrap	gSLR + χ^2-mixture	Codeml M8a vs M8	FUBAR	gSLR + bootstrap	gSLR + χ^2-mixture
6-species tree	Medium	0.0094	0.0231	0.0150	0.0000	0.0002	0.1081	0.1782	0.1463	0.0006	0.0094	58	100
	Large	0.0029	0.0215	0.0090	0.0000	0.0000	0.0746	0.2826	0.1635	0.0004	0.0007	98	100
	Small	0.0124	0.0124	0.0124	0.0000	0.0001	0.0562	0.0562	0.0562	0.0002	0.0018	0	0
12-species tree	Medium	0.0063	0.0173	0.0121	0.0001	0.0002	0.1441	0.2469	0.2062	0.0060	0.0162	77	100
	Large	0.0004	0.0049	0.0021	0.0000	0.0004	0.1686	0.4094	0.3066	0.0159	0.0082	100	100
	Small	0.0052	0.0160	0.0100	0.0000	0.0000	0.1419	0.2517	0.2061	0.0031	0.0140	86	100
17-species tree	Medium	0.0007	0.0059	0.0031	0.0000	0.0000	0.2251	0.4026	0.3275	0.0308	0.0383	98	100
	Large	0.0006	0.0033	0.0019	0.0000	0.0001	0.2359	0.4572	0.3673	0.0477	0.0430	99	100
	Small	0.0029	0.0115	0.0074	0.0000	0.0000	0.1769	0.2930	0.2518	0.0120	0.0175	88	100
44-species tree	Medium	0.0001	0.0021	0.0010	0.0000	0.0001	0.3507	0.5653	0.4851	0.1835	0.0561	100	100
	Large	0.0001	0.0005	0.0001	0.0000	0.0005	0.4500	0.7078	0.6009	0.3870	0.0982	100	100
Case D													
	Small	0.0566	0.0566	0.0566	0.0013	0.0002	0.1075	0.1212	0.1075	0.0025	0.0023	0	0
6-species tree	Medium	0.0446	0.0446	0.0447	0.0021	0.0024	0.1832	0.1886	0.1832	0.0063	0.0217	1	100
	Large	0.0392	0.0392	0.0390	0.0070	0.0009	0.2128	0.1801	0.2128	0.0110	0.0055	2	100
	Small	0.0438	0.0438	0.0438	0.0027	0.0006	0.0895	0.1117	0.0895	0.0073	0.0018	0	0
12-species tree	Medium	0.0468	0.0473	0.0496	0.0043	0.0023	0.2856	0.2783	0.2878	0.0278	0.0397	2	100
	Large	0.0492	0.0494	0.0475	0.0009	0.0053	0.4508	0.4482	0.4531	0.0079	0.0304	0	100
	Small	0.0483	0.0492	0.0518	0.0101	0.0025	0.3073	0.2779	0.3104	0.0272	0.0309	0	100
17-species tree	Medium	0.0532	0.0535	0.0511	0.0002	0.0015	0.4492	0.4715	0.4492	0.0267	0.0580	3	100
	Large	0.0523	0.0525	0.0497	0.0005	0.0032	0.5872	0.5548	0.5880	0.0457	0.0903	0	100
	Small	0.0523	0.0526	0.0527	0.0019	0.0011	0.3383	0.3617	0.3403	0.0316	0.0374	2	100
44-species tree	Medium	0.0569	0.0570	0.0555	0.0028	0.0051	0.6795	0.7329	0.6806	0.2053	0.1417	1	100
	Large	0.0542	0.0544	0.0524	0.0014	0.0096	0.8400	0.8645	0.8415	0.3311	0.1625	1	100

Figure 5.6: FPRs and TPRs for the five tests compared in this paper, and the number of data sets that passed the diagnostics for the new tests, for simulation cases C and D and 12 different trees. FPRs are shaded in white below 0.05, and then from white at 0.05 to dark red at 0.1 as values increase. Shading of TPRs is as in Figure 5.5

		FPR					TPR					Number of datasets that passed the diagnostics	
Case E		SLR	gSLR + bootstrap	gSLR + χ^2-mixture	Codeml M8a vs M8	FUBAR	SLR	gSLR + bootstrap	gSLR + χ^2-mixture	Codeml M8a vs M8	FUBAR	gSLR + bootstrap	gSLR + χ^2-mixture
	Small	0.0592	0.0592	0.0592	0.0020	0.0004	NA	NA	NA	NA	NA	0	0
6-species tree	Medium	0.0421	0.0421	0.0423	0.0017	0.0021	NA	NA	NA	NA	NA	1	100
	Large	0.0369	0.0369	0.0395	0.0011	0.0006	NA	NA	NA	NA	NA	6	100
	Small	0.0383	0.0383	0.0383	0.0014	0.0005	NA	NA	NA	NA	NA	1	1
12-species tree	Medium	0.0469	0.0478	0.0520	0.0070	0.0030	NA	NA	NA	NA	NA	3	100
	Large	0.0510	0.0515	0.0506	0.0002	0.0051	NA	NA	NA	NA	NA	0	100
	Small	0.0459	0.0483	0.0510	0.0056	0.0022	NA	NA	NA	NA	NA	2	100
17-species tree	Medium	0.0547	0.0554	0.0546	0.0006	0.0032	NA	NA	NA	NA	NA	0	100
	Large	0.0501	0.0509	0.0507	0.0002	0.0044	NA	NA	NA	NA	NA	2	100
	Small	0.0508	0.0522	0.0523	0.0097	0.0014	NA	NA	NA	NA	NA	2	100
44-species tree	Medium	0.0547	0.0557	0.0569	0.0005	0.0050	NA	NA	NA	NA	NA	4	100
	Large	0.0507	0.0516	0.0517	0.0013	0.0085	NA	NA	NA	NA	NA	1	100

Figure 5.7: FPRs and TPRs for the five tests compared in this paper, and the number of data sets that passed the diagnostics for the new tests, for simulation case E and 12 different trees. Shading is as in Figure 5.6. NA indicates undefined TPRs in cases where there are no true positives to find.

cases D and E, the bootstrap variant is almost never used.

5.3.1 Diagnostics

As described earlier, my new method comprises a modified test for positive selection along with diagnostics to distinguish when the chosen variant of the new test is appropriate for use with a data set, and when an alternative should be used instead. There are two situations when the FPR of the test might be higher than desired: when the tree is very short and hence the alignment contains insufficient information to estimate site-wise ω accurately, or when data are close to neutral. The former affects both variants for choosing significance, and the latter affects only the bootstrap variant. My aim is to choose diagnostics and cut-offs that return the FPR of the test to the correct value, whilst not removing an excessive number of data sets that already have an appropriate FPR. In this way, the new test improves power on as many data sets as possible whilst still having a controlled FPR.

To identify data sets with trees that are too short I used the data from simulations A, B and C on all of the 12 trees, as I knew these data were not close to neutral, and hence high FPRs would not be caused by the underlying ω-distribution in the data sets. I looked at different ways of summarising the data that might be informative about the tree being too small. Examples include the proportion of sites with $\hat{\omega} > 0$, the ratio of the proportion of sites with $\hat{\omega} > 0$ in the bootstrap and original samples, and the ratio of the proportion of sites with $\hat{\omega} = 1$ in the bootstrap and original samples. The summary statistic that best separated data sets on short trees from those on larger trees, and could be carried out without time-consuming bootstrap sampling, was the proportion of sites with $\hat{\omega} > 0$ (results not shown). The cut-off chosen was 0.2, meaning that at least 20% of the sites need to have $\hat{\omega} > 0$ for the new test to be used (see Fig. 5.8a). This cut-off applies to both the bootstrap and the χ^2-mixture variants of the new test.

To identify data close to neutral I used data from all of my simulations, but removed any trees identified to be too short by the first diagnostic. Again I looked at a variety of ways of summarising the data that would be informative about whether the data was close to neutral (data not shown). The most useful statistic was the proportion of sites with $\hat{\omega} \geq 1$, where a cut-off of 0.35 gave best results (see Fig. 5.8b).

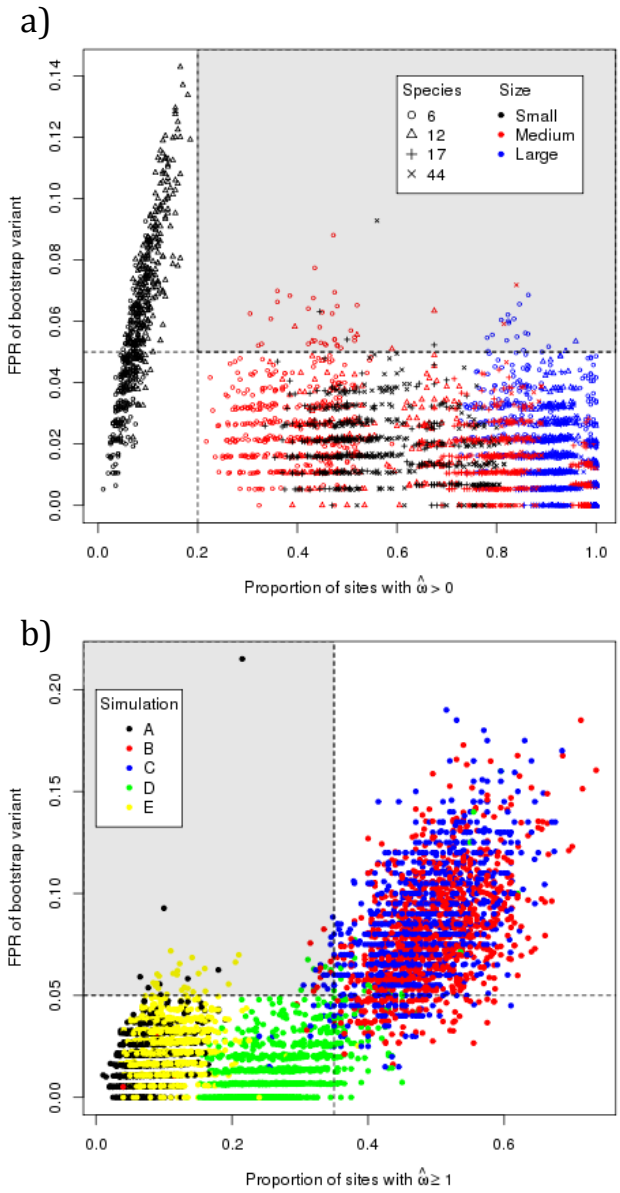


Figure 5.8: **a** Proportion of sites with $\hat{\omega} > 0$ plotted against the FPR of the bootstrap variant of the new test, for ω-distributions A, B and C. Each point corresponds to a simulated data set, with colour and shape indicating the tree that data set was simulated on. The small trees (6-species small and 12-species small) have very low proportion of sites with $\hat{\omega} > 0$ and can also have very high FPRs. Based on this plot, 0.2 was chosen as a cut-off which splits the small trees where the tests can have too high FPRs from larger trees. The shaded area contains the few remaining data sets not eliminated at this threshold where the FPR is too high. An equivalent plot can be produced for the χ^2-mixture variant; the same threshold is appropriate. **b** Proportion of sites with $\hat{\omega} \geq 1$ plotted against the FPR of the bootstrap variant, for all data sets that passed the short tree diagnostic. Each point corresponds to a simulated data set, with colour indicating the ω-distribution used in that simulation. As the proportion of sites with $\hat{\omega} \geq 1$ increases so does the FPR. A threshold of 0.35 was chosen. Again, the shaded area contains data sets not eliminated by the cut-off where the FPR is too high.

5.3.2 False Discovery Rate Correction

Figure 5.9 shows results from my simulations after FDR correction is applied to SLR and the new gSLR test, at an FDR threshold of 5%. FDR correction is not required for codeml and FUBAR, as a posterior probability threshold of z guarantees that the FDR is less than or equal to $1-z$, provided that the posterior probabilities are correct (Wong et al. 2004). By this criterion, both variants of the new test perform roughly similarly to codeml. On large trees these methods all perform better than the original SLR method or FUBAR.

5.3.3 Real Data

Although I have taken care to devise diagnostic tests to identify difficult cases where the gSLR test could fail, my approach is designed to be applicable to many data sets. To illustrate this I have checked how many of the mammal protein alignments from the data set described earlier (Jordan 2011) could use the new test. 10% of the proteins would not pass the short tree diagnostic, and a further 0.5% have an $\hat{\omega}$-distribution that is inappropriate for the bootstrap variant. This means that the bootstrap variant of the new test can be used on 89.5% of the proteins (14042 out of 15712), and the χ^2-mixture variant can be used on 90% of the proteins (14111 out of 15712), potentially increasing power and allowing more positively selected sites to be found. Recall that when the new test cannot be applied it is still possible to use SLR.

As a further example, I show the application of both the new gSLR and old SLR tests to a mammalian alignment of CD22, a regulatory molecule that prevents the over-action of the immune system and the development of autoimmune diseases (Hatta et al. 1999). Using the original SLR test 63 sites are found to be under positive selection. With the new test, using either the χ^2-mixture or the bootstrap variants, this is increased to 89 sites. (After Benjamini-Hochberg FDR correction, these figures are reduced to 16 and 23, respectively.) Figure 5.10 shows 100 amino acids of the alignment along with the corresponding phylogenetic tree. The track under the alignment shows $\hat{\omega}$ for each site when no restrictions are put on ω (black bar), along with the confidence intervals around $\hat{\omega}$ derived using SLR and the new test using the χ^2-mixture distribution to determine significance. This illustrates that the new test finds more sites to be under positive selection, and has tighter confidence intervals.

			TPR			
Case A		SLR	gSLR + bootstrap	gSLR + χ^2-mixture	Codeml M8a vs M8	FUBAR
	Small	0.0000	0.0000	0.0000	0.0014	0.0021
6-species tree	Medium	0.0083	0.0268	0.0174	0.0410	0.0620
	Large	0.0000	0.0072	0.0000	0.0007	0.0000
	Small	0.0039	0.0039	0.0039	0.0050	0.0128
12-species tree	Medium	0.0270	0.1274	0.1071	0.1514	0.1136
	Large	0.0000	0.0241	0.0000	0.0210	0.0044
	Small	0.0185	0.1209	0.0911	0.1256	0.1131
17-species tree	Medium	0.0582	0.2663	0.1880	0.3328	0.2342
	Large	0.0216	0.2665	0.1509	0.3632	0.2147
	Small	0.0763	0.2561	0.2311	0.3055	0.1595
44-species tree	Medium	0.3881	0.7519	0.7058	0.7480	0.3488
	Large	0.4768	0.8318	0.7772	0.8318	0.4077
Case B						
	Small	0.0006	0.0006	0.0006	0.0000	0.0011
6-species tree	Medium	0.0016	0.0096	0.0061	0.0042	0.0226
	Large	0.0000	0.0041	0.0000	0.0000	0.0000
	Small	0.0027	0.0027	0.0027	0.0029	0.0051
12-species tree	Medium	0.0134	0.0340	0.0341	0.0268	0.0486
	Large	0.0000	0.0116	0.0000	0.0044	0.0022
	Small	0.0179	0.0400	0.0350	0.0337	0.0507
17-species tree	Medium	0.0211	0.0920	0.0631	0.1012	0.0873
	Large	0.0225	0.0998	0.0637	0.1134	0.1027
	Small	0.0339	0.0676	0.0708	0.0820	0.0719
44-species tree	Medium	0.1202	0.2186	0.2025	0.2654	0.1605
	Large	0.1250	0.2562	0.2192	0.3450	0.1373
Case C						
	Small	0.0016	0.0016	0.0016	0.0002	0.0014
6-species tree	Medium	0.0016	0.0034	0.0020	0.0006	0.0094
	Large	0.0000	0.0076	0.0000	0.0004	0.0007
	Small	0.0008	0.0008	0.0008	0.0002	0.0018
12-species tree	Medium	0.0044	0.0056	0.0048	0.0060	0.0162
	Large	0.0008	0.0134	0.0018	0.0159	0.0082
	Small	0.0028	0.0092	0.0055	0.0031	0.0140
17-species tree	Medium	0.0070	0.0218	0.0145	0.0308	0.0383
	Large	0.0060	0.0338	0.0192	0.0477	0.0430
	Small	0.0089	0.0125	0.0122	0.0120	0.0175
44-species tree	Medium	0.0395	0.0900	0.0748	0.1835	0.0561
	Large	0.0843	0.2224	0.1580	0.3870	0.0982
Case D						
	Small	0.0043	0.0043	0.0043	0.0025	0.0023
6-species tree	Medium	0.0042	0.0051	0.0042	0.0063	0.0217
	Large	0.0000	0.0000	0.0000	0.0110	0.0055
	Small	0.0062	0.0062	0.0062	0.0073	0.0018
12-species tree	Medium	0.0150	0.0150	0.0156	0.0278	0.0397
	Large	0.0082	0.0082	0.0115	0.0079	0.0304
	Small	0.0101	0.0102	0.0087	0.0272	0.0309
17-species tree	Medium	0.0426	0.0424	0.0384	0.0267	0.0580
	Large	0.0815	0.0824	0.0826	0.0457	0.0903
	Small	0.0419	0.0419	0.0419	0.0316	0.0374
44-species tree	Medium	0.2502	0.2502	0.2489	0.2053	0.1417
	Large	0.3995	0.4010	0.4018	0.3311	0.1625

Figure 5.9: TPR after FDR correction has been applied to the site-wise tests. Shading is as in Figure 5.5. FPRs are not shown; they are all below 0.0015.

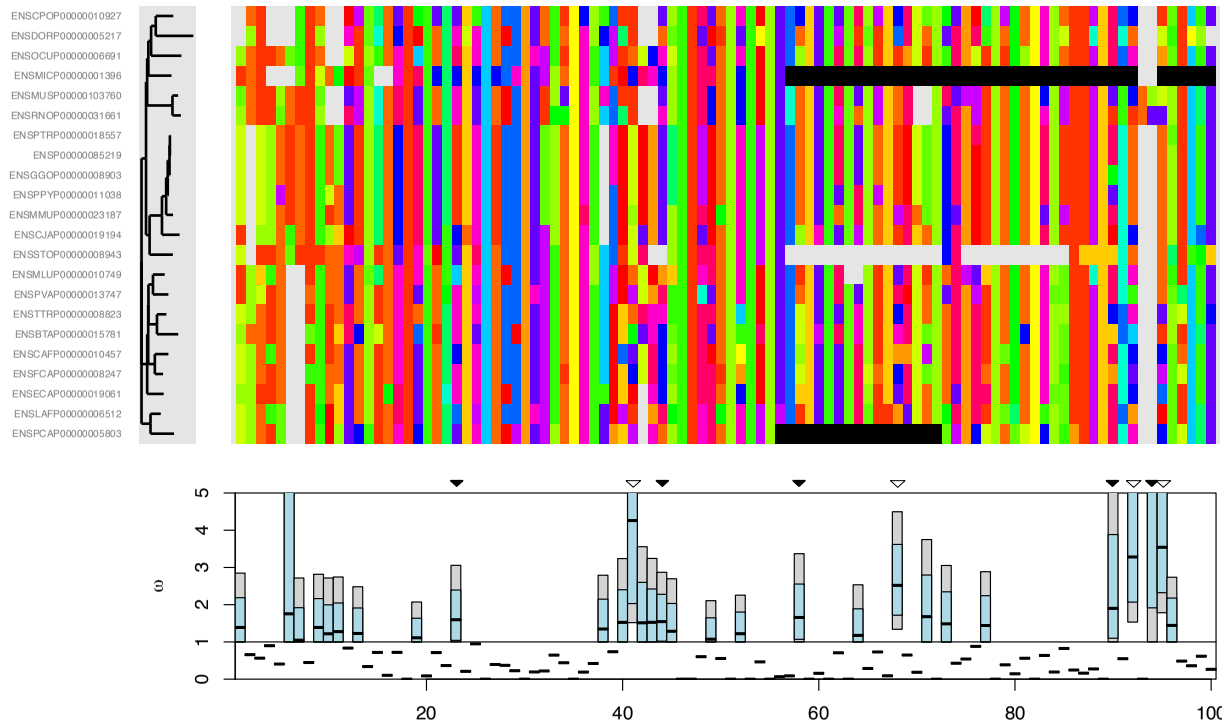


Figure 5.10: 100 bases of a mammalian alignment of CD22 (with sites only present in one species discarded), along with the phylogenetic tree. The track under the alignment shows $\hat{\omega}$ for each site when no restrictions are put on ω (black bar), and the confidence intervals around $\hat{\omega}$ derived using SLR (in grey) and the new gSLR test using the χ^2-mixture distribution (in blue) to determine significance. (As both of these tests are for $\omega > 1$, the lower bound on the confidence intervals cannot be lower than 1. To improve clarity, confidence intervals for sites with $\hat{\omega} < 1$ are not shown.) Sites under positive selection are indicated by arrows above the track. An empty arrow indicates that both tests find that site to be under positive selection (4 sites in this alignment region); a solid arrow indicates that only the new test finds this site to be under positive selection (5 sites).

5.4 Discussion

In this chapter I have shown that some current tests for positive selection lack power, and therefore positively selected sites may be going undetected. I have presented a new test for positive selection, which has higher power than SLR but maintains a controlled FPR. This method alters the null hypothesis of SLR to take into account the fact that most sites in a protein are negatively selected. It also uses diagnostic tests to check when it is inappropriate for proteins, and there is a good alternative test, SLR, that can be used in these situations. This method therefore still has the advantages of a FEL method, in that assumptions do not need to be made about how the level of selection varies along the sequences, whilst also having the advantage of the REL method in that knowledge about the whole distribution can allow us to decide whether each site is likely to be under positive selection.

In practice, particularly if wishing to apply this method to a genome-wide scan, FDR correction may be desired. I have shown that after straightforward FDR correction the new method presented here performs well, and is a great improvement over SLR with FDR correction. FUBAR does not require any correction, but tends to have lower power. On a per-gene basis, codeml performs similarly to the method introduced here; however, it is not clear how FDR correction should be performed for (e.g.) genome-wide analysis with codeml, due to its combination of both gene-wise LRTs and Bayesian posterior probabilities.

It is known that, in general, methods for detecting positive selection struggle with power on short trees (e.g. Anisimova et al. 2001). There has not been a good way to detect when trees are too short, and so it has been difficult to know whether an absence of inferred positively selected sites was because there were none to find, or because the tree was too short and uninformative. My method includes a diagnostic test to decide when trees are too short. This diagnostic is not applicable only to my new method; it could be used in conjunction with other tests to identify when the tree may be too short to permit enough statistical power to find positive selection.

Both variants of the new test often double statistical power compared to SLR. The χ^2-mixture variant tends to be more conservative than the bootstrap, but is significantly faster to apply as simulations are not required. This may make it more readily applicable, particularly for large alignments. As the new method takes into account the distribution of ω over sites in order to determine whether sites are under positive selection, it will have the correct FPR, and good power, for any possible real ω-distributions.

Chapter 6

Conclusions

In this thesis I have presented three projects focused on better understanding or improving methodology for maximum likelihood phylogenetic inference. A link between these projects is that they are all fundamentally related to the fact that, whilst maximum likelihood will accurately reconstruct the tree and parameters of interest given the correct model and infinite data, in real life studies we always have finite data. This can manifest as a problem in a number of ways.

The issue of finite data can be clearly seen when long branches are present on a tree. In Chapter 3 I showed that, for a given number of sequence sites, on a three species tree, as the length of one branch is increased, that branch is inferred with increasing probability to join at the tips of the other branches. If sequence length is increased this feature decreases in prevalence, increasing again if branch length is further increased. If both sequence length and branch length are increasing, the relative rate becomes important in order to infer whether this feature will occur.

Distance matrix methods also infer the location of long branches to be at the tips of other branches, often agreeing with maximum likelihood, allowing equations describing when this occurs for distance matrix methods to be used to make predictions for maximum likelihood. The inference of long branches at the tips appears to be at least partly explained by the high variance of branch length estimates when one branch is long.

On a four species tree, with two long branches, if sequences were of infinite length then placement would again be accurate. With realistic length sequences, however, I showed that, although there is no attraction between long branches, long branches are placed together preferentially. Therefore, I suggest that the phenomenon previously called Long Branch Attraction should be renamed to the more appropriate Long Branch Joining. Unfortunately, this joining has not been explained. Future work could focus on finding an explanation for this, as this

could allow for the development of methods to detect Long Branch Joining using quartets.

The results presented here on long branches hold for small trees with simple models that are known to be correct. This is a gross simplification of real life, making it difficult to infer how these phenomena might manifest themselves in real studies. One way of getting closer to this understanding would be to analyse the placement of one long branch, and investigate the existence of Long Branch Joining and Long Branch Closeness, on larger trees.

The theme of issues caused by finite data carries over to the work of Chapter 4, where non-reversible models were investigated as an alternative to reversible models. If we had infinite data we would be able to accurately build complicated models; with finite data, however, we need to worry about the accuracy of parameters being estimated, and whether new parameters significantly improve model fit. Chapter 4 also changes focus, from thinking about what can be done with existing models (a theme picked up again in Chapter 5) to improving the models themselves. This is important as, in real life studies, while we know models will always be incorrect, we want them to describe the most important features of the data.

I showed that nucleotide data sets are often better described by a non-reversible model, whereas the corresponding amino acid data sets are not. This is probably because, for amino acid models, switching from a reversible model to a non-reversible model involves the introduction of 171 new parameters. This does not necessarily mean that the true amino acid model is reversible, just that more data is needed to get a sufficient increase in fit of the model for the non-reversible model to be chosen over a reversible model. In practice, the majority of amino acid data sets will be of a similar size to the ones used here, so these results indicate that non-reversible models may not be justified for modelling amino acid evolution.

To be able to take advantage of the improvement in fit given by non-reversible models it needs to be practical to build them. Currently this is not the case, as only a few of the common phylogeny inference software include them as options and they are also not present in model testing software. In the future the focus should be on making non-reversible model inference more practical and available, so that better models can be inferred.

Chapter 5 moves on to the problem of accurately finding sites under positive selection. Again, a finite data issue arises: instead of sequence length, problems are now caused by tree size (although sequences long enough to accurately infer a tree and other model parameters are required). As described in Section 2.3.1,

if the number of sequences is taken to infinity in a suitable fashion, then it would be possible to accurately reconstruct ω, the parameter describing the relative fixation probabilities of non-synonymous and synonymous mutations, for each site and test whether it is greater than 1 with high power. However with small trees, power is reduced. The method presented here attempts to improve power by taking into account the fact that many sites are under purifying selection, and hence the null distribution of current site-wise tests is not correct for most real alignments. The modified site-wise generalised likelihood (gSLR) test I propose is effective, often doubling or tripling power whilst keeping the false positive rate under control.

These three projects highlight the complexities of maximum likelihood estimation of trees and parameters with finite data, and the need for continued work on the fundamentals of phylogenetic inference.

Appendix A

A.1 Finding Solutions to the Likelihood Equation

The log-likelihood of a tree is given as

$$LogL = \sum_{r \in P} n_r \log\left(\frac{p_r}{v_r}\right)$$

where P is the set of pattern counts, p_r is the probability, n_r the observed number of occurrences of pattern $r \in P$, and v_r is the number of different nucleotide combinations which are in the form of the pattern $r \in P$. The probabilities p_r are functions of the branch lengths and the topology. The term, v_r, is present because within the likelihood function we require the probability of a specific nucleotide combination, not the probability of a group of combinations. On a three-taxon tree there is one topology and five possible patterns, $P = \{xxx, xxy, xyx, yxx, xyz\}$, where x, y and z represent any three different nucleotides. For a tree with branch lengths d_A, d_B, and d_C (Fig. 3.2), the pattern probabilities are:

$$\begin{aligned}
p_{xxx} &= \tfrac{1}{16}\left(1 + 3e^{-\frac{4}{3}(d_A+d_B)} + 3e^{-\frac{4}{3}(d_A+d_C)} + 3e^{-\frac{4}{3}(d_B+d_C)} + 6e^{-\frac{4}{3}(d_A+d_B+d_C)}\right) \\
p_{xxy} &= \tfrac{3}{16}\left(1 + 3e^{-\frac{4}{3}(d_A+d_B)} - e^{-\frac{4}{3}(d_A+d_C)} - e^{-\frac{4}{3}(d_B+d_C)} - 2e^{-\frac{4}{3}(d_A+d_B+d_C)}\right) \\
p_{xyx} &= \tfrac{3}{16}\left(1 - e^{-\frac{4}{3}(d_A+d_B)} + 3e^{-\frac{4}{3}(d_A+d_C)} - e^{-\frac{4}{3}(d_B+d_C)} - 2e^{-\frac{4}{3}(d_A+d_B+d_C)}\right) \\
p_{yxx} &= \tfrac{3}{16}\left(1 - e^{-\frac{4}{3}(d_A+d_B)} - e^{-\frac{4}{3}(d_A+d_C)} + 3e^{-\frac{4}{3}(d_B+d_C)} - 2e^{-\frac{4}{3}(d_A+d_B+d_C)}\right) \\
p_{xyz} &= \tfrac{3}{8}\left(1 - e^{-\frac{4}{3}(d_A+d_B)} - e^{-\frac{4}{3}(d_A+d_C)} - e^{-\frac{4}{3}(d_B+d_C)} + 2e^{-\frac{4}{3}(d_A+d_B+d_C)}\right)
\end{aligned} \quad \text{(A.1)}$$

The log-likelihood function is then:

$$\begin{aligned}
LogL =& n_{xxx} \log\left(\tfrac{1}{64}\left(1+3e^{-\frac{4}{3}(d_A+d_B)}+3e^{-\frac{4}{3}(d_A+d_C)}+3e^{-\frac{4}{3}(d_B+d_C)}+6e^{-\frac{4}{3}(d_A+d_B+d_C)}\right)\right) \\
&+ n_{xxy} \log\left(\tfrac{1}{64}\left(1+3e^{-\frac{4}{3}(d_A+d_B)}-e^{-\frac{4}{3}(d_A+d_C)}-e^{-\frac{4}{3}(d_B+d_C)}-2e^{-\frac{4}{3}(d_A+d_B+d_C)}\right)\right) \\
&+ n_{xyx} \log\left(\tfrac{1}{64}\left(1-e^{-\frac{4}{3}(d_A+d_B)}+3e^{-\frac{4}{3}(d_A+d_C)}-e^{-\frac{4}{3}(d_B+d_C)}-2e^{-\frac{4}{3}(d_A+d_B+d_C)}\right)\right) \\
&+ n_{yxx} \log\left(\tfrac{1}{64}\left(1-e^{-\frac{4}{3}(d_A+d_B)}-e^{-\frac{4}{3}(d_A+d_C)}+3e^{-\frac{4}{3}(d_B+d_C)}-2e^{-\frac{4}{3}(d_A+d_B+d_C)}\right)\right) \\
&+ n_{xyz} \log\left(\tfrac{1}{64}\left(1-e^{-\frac{4}{3}(d_A+d_B)}-e^{-\frac{4}{3}(d_A+d_C)}-e^{-\frac{4}{3}(d_B+d_C)}+2e^{-\frac{4}{3}(d_A+d_B+d_C)}\right)\right)
\end{aligned} \tag{A.2}$$

I have not been able to solve this equation over the whole solution space. Instead I have solved it on all of the boundaries (Fig. 3.7). To solve for maxima on the boundaries I restrict the likelihood equation to the boundary of interest and solve for optima. I then classify the optima to find when they are maxima. In the following section all optima on boundaries are determined.

Optima with Zero Branch Lengths

All Branch Lengths Zero

Firstly consider the point $(d_A, d_B, d_C) = (0, 0, 0)$ (Fig. 3.7). This has two values depending on the pattern counts, either $L = -\infty$ or $L = -n\log(4)$. The former is clearly a minimum and occurs when there is any difference between the sequences. The latter occurs when

$$n_{xxx} = n \qquad\qquad n_{xxy} = n_{xyx} = n_{yxx} = n_{xyz} = 0 \tag{A.3}$$

In this case the only probability that affects the likelihood value is p_{xxx}. The maximal likelihood value occurs when $p_{xxx} = 1$, which occurs when all of the branch lengths are zero. This point is therefore the global maximum when all the sequences are the same.

In summary, $(d_A, d_B, d_C) = (0, 0, 0)$ is the global ML solution in the case $n_{xxx} = n$. Otherwise it is not a maximum.

Two Branch Lengths Zero

The next possibility is that two branch lengths are zero. Assume those two branch lengths are d_A and d_C; this corresponds to the blue line on Figure 3.7. If $n_{xyz} = n_{yxx} = n_{xxy} = 0$ there is an optimum at $(d_A, d_B, d_C) = (0, \infty, 0)$ with $L = -n\log(16)$. This is a local maximum if $n_{xxx} \leq n/4$. This can be seen by looking at the gradient of the likelihood function on the line $d_1 = d_2 = 0$ as it approaches ∞.

If instead $n_{xxx} > n/4$ then there is a critical point at

$$d_A = 0 \qquad d_B = -\frac{3}{4}\log\left(\frac{4n_{xxx} - n}{3n}\right) \qquad d_C = 0$$

with

$$L = n_{xxx}\log\left(\frac{n_{xxx}}{4n}\right) + (n - n_{xxx})\log\left(\frac{n - n_{xxx}}{12n}\right)$$

This is a local maximum as both gradients on the d_A and d_C boundaries are negative and the Hessian in d_B is

$$\frac{\partial^2 L}{\partial d_B^2} = -\frac{n(n - 4n_{xxx})^2}{9n_{xxx}(n - n_{xxx})}$$

which is always negative when $n_{xxx} > n/4$.

These are also the global maxima if $n_{xyz} = n_{yxx} = n_{xxy} = 0$. This can be seen by going through all the other possible optima for this set of pattern counts and comparing the likelihood values. In this case it is possible to solve for optima inside the solution space (i.e. $d_i > 0$ for $i = A, B, C$) and find that there are no real maxima when either $n_{xxx} > 0$ or $n_{xyx} > 0$. This means it is possible to find all of the optima and compare the likelihood values to find out which one(s) are the global maxima. If the above conditions on the pattern counts are not met then $L = -\infty$ for any length of d_C.

Using the symmetry of the likelihood equations, similar optima to those above can be found for the cases $n_{xyz} = n_{xxy} = n_{xyx} = 0$ and $n_{xyz} = n_{yxx} = n_{xyx} = 0$.

In summary, $(0, 0, \infty)$ is the global ML solution if $n_{xyz} = n_{yxx} = n_{xxy} = 0$ and $n_{xxx} \leq n/4$. If $n_{xyz} = n_{yxx} = n_{xxy} = 0$ and $n_{xxx} > n/4$, $(0, 0, -\frac{3}{4}\log(\frac{4n_{xxx} - n}{3n}))$ is the global ML solution. Otherwise there is no maximum. By symmetry there are three possible maxima, one on each of the three lines ($n_{xyz} = n_{xxy} = n_{xyx} = 0$, $n_{xyz} = n_{yxx} = n_{xxy} = 0$, $n_{xyz} = n_{yxx} = n_{xxy} = 0$), but only one of them can occur

at a time.

One Branch Length Zero

Next take d_A to be a zero branch length and the other distances to be greater than zero. This corresponds to the yellow shaded plane on Figure 3.7. If $n_{xxx}+n_{xxy} > n/4$ and $n_{xxx}+n_{xyx} > n/4$ and $\frac{\partial L}{\partial d_A} \leq 0$ then there is an optimum at

$$d_A = 0 \quad d_B = -\frac{3}{4}\log\left(1 - \frac{4(n - n_{xxx} - n_{xxy})}{3n}\right) \quad d_C = -\frac{3}{4}\log\left(1 - \frac{4(n - n_{xxx} - n_{xyx})}{3n}\right)$$

In this case,

$$L = n_{xxx}\log\left(\frac{(n_{xxx}+n_{xxy})(n_{xxx}+n_{xyx})}{4n^2}\right) + n_{xxy}\log\left(\frac{(n_{xxx}+n_{xxy})(n - n_{xxx} - n_{xyx})}{12n^2}\right)$$

$$+n_{xyx}\log\left(\frac{(n_{xxx}+n_{xyx})(n - n_{xxx} - n_{xxy})}{12n^2}\right)$$

$$+(n - n_{xxx} - n_{xxy} - n_{xyx})\log\left(\frac{(n - n_{xxx} - n_{xyx})(n - n_{xxx} + n_{xxy})}{36n^2}\right)$$

As this optimum is on a boundary ($d_A = 0$) there is no requirement for the gradient in the d_A direction to be zero; it just needs to be non-positive. This is then the highest local point in this direction as we move away from the boundary. The gradients in the d_B and d_C directions do however still need to be zero. The Hessian can then be used in just these two variables to classify the optimum. By checking the Hessian matrix with respect to d_B and d_C we can see that the eigenvalues are both negative so this critical point is always a local maximum:

$$\lambda_1 = -\frac{n(n - 4n_{xxx} - 4n_{xxy})^2}{9(n_{xxx}+n_{xxy})(n - n_{xxx} - n_{xxy})}$$

$$\lambda_2 = -\frac{n(n - 4n_{xxx} - 4n_{xyx})^2}{9(n_{xxx}+n_{xyx})(n - n_{xxx} - n_{xyx})}$$

By the symmetry of the likelihood equations there are also two more optima; one where $d_B = 0$, and one where $d_C = 0$.

In summary $(0, -\frac{3}{4}\log(1-\frac{4(n-n_{xxx}-n_{xxy})}{3n}), -\frac{3}{4}\log(1-\frac{4(n-n_{xxx}-n_{xyx})}{3n}))$ is a local maximum if $n_{xxx}+n_{xxy} > n/4$ and $n_{xxx}+n_{xyx} > n/4$ and $\frac{\partial L}{\partial d_A} \leq 0$. Otherwise there is no maximum here.

From the conditions given above it can be seen that if we consider the three quantities $\alpha = n_{xxx}+n_{xxy}$, $\beta = n_{xxx}+n_{xyx}$, and $\gamma = n_{xxx}+n_{yxx}$, then if one of these is less than or equal to $n/4$, there can only be one maximum with one of the three branch lengths equal to zero. However if $\alpha, \beta, \gamma > n/4$ it is possible for there to be a maxima on each of the three planes where one of the branch lengths is 0. There is one particular case where there are two maxima each with a different branch being equal to zero. This can occur when two of the three pattern counts which compare y in one sequence against x in the other two, (e.g. $n_{xxy}, n_{xyx}, n_{yxx}$) are the same. There is then an optimum that has two branch lengths the same. So if $n_{xxy} = n_{xyx}$ then there is an optimum where $d_A = d_B = a$, where a is defined by the particular pattern counts. This optimum can be either a maximum or a minimum. In the case that it is a minimum then, due to the symmetry of the likelihood equation, there are two local maxima, each with one branch length zero and with the same likelihood value: $(d_A, d_B, d_C) = (0, 2a, d_C)$ and $(d_A, d_B, d_C) = (2a, 0, d_C)$.

Optima with Infinite Branch Lengths

Two or Three Branch Lengths Infinite

Now consider the 'lines' of the solution space where two branch lengths are infinity (dotted lines on Fig. 3.7), and the 'point' connecting these lines where all of the branch lengths are infinity. These are all always extrema (either maxima or minima) of the likelihood function and all have the same likelihood value, $L = -n\log(64)$, depending only on the length of the sequences, but they are not all always local maxima. It is not possible to determine whether these points are maxima or minima, so it is necessary to consider them as possible maxima until another local maximum with a larger likelihood value is found.

One Branch Length Infinite

The final optimum is when there is just one infinite branch. If $n_{xxx} + n_{yxx} > n/4$ and $n_{xxy} + n_{xyx} + n_{xyz} > 0$ then there are extrema when

$$d_A = \infty \qquad\qquad d_B + d_C = -\frac{3}{4}\log\left(\frac{4(n_{xxx} + n_{yxx}) - n}{3n}\right)$$

with

$$L = (n - n_{xxx} - n_{yxx})\log\left(\frac{n - n_{xxx} - n_{yxx}}{48n}\right) + (n_{xxx} + n_{yxx})\log\left(\frac{n_{xxx} + n_{yxx}}{16n}\right)$$

In this case it is not possible to derive a solution for d_B or d_C separately, meaning that there is a line of extrema here. Again it is not possible to determine whether these are maxima, so they must be regarded as possible maxima until an optimum with a higher likelihood is found. Due to the symmetry of the three-taxon tree it is easy to work out a similar equation for when $d_B = \infty$ or $d_C = \infty$.

In summary $(\infty, -\frac{3}{4}\log(\frac{4(n_{xxx}+n_{yxx})-n}{3n}) - \mu, \mu)$, where $0 \leq \mu \leq -\frac{3}{4}\log(\frac{4(n_{xxx}+n_{yxx})-n}{3n})$, is a line of extrema if $n_{xxx} + n_{yxx} > n/4$.

A.2 Calculating the Variance of the Distance Estimate of a Branch Length

Expected value of $\hat{D}_{ij}$

The estimated JC distance between sequence i and j is

$$\hat{D}_{ij} = -\frac{3}{4}ln(1-\frac{4}{3}\hat{p}_{ij}) \tag{A.4}$$

$$\begin{aligned} E(\hat{D}_{ij}) =& E(-\frac{3}{4}ln(1-\frac{4}{3}\hat{p}_{ij})) \\ \simeq& -\frac{3}{4}ln(1-\frac{4}{3}p_{ij}) \end{aligned} \tag{A.5}$$

(Bulmer 1991; Tajima 1993). As the JC model is assumed to be the true model A.4 can be rearranged to obtain

$$p_{ij} = \frac{3}{4}(1-e^{-\frac{4}{3}D_{ij}}) \tag{A.6}$$

From equations A.5 and A.6 the expectation of $\hat{D}_{ij}$ is

$$E(\hat{D}_{ij}) \simeq D_{ij} \tag{A.7}$$

Expected value of $\hat{d}_i$

The expected value of any branch length, $\hat{d}_i$, on the tree in Figure 3.2 is approximately equal to the branch length itself. For example:

$$\begin{aligned} E(\hat{d}_A) =& E(\frac{1}{2}(\hat{D}_{AB}+\hat{D}_{AC}-\hat{D}_{BC})) \\ =& \frac{1}{2}(E(\hat{D}_{AB})+E(\hat{D}_{AC})-E(\hat{D}_{BC})) \\ \simeq& \frac{1}{2}(D_{AB}+D_{AC}-D_{BC}) \\ \simeq& d_A \end{aligned}$$

Variance of $\hat{d}_i$

The variance of any branch length, $\hat{d}_i$, is then a function of the variance and covariances of the D_{ij} terms. For example:

$$\begin{aligned}
Var(\hat{d}_A) =& Var(\frac{1}{2}(\hat{D}_{AB} + \hat{D}_{AC} - \hat{D}_{BC})) \\
=& \frac{1}{4} Var(\hat{D}_{AB} + \hat{D}_{AC} - \hat{D}_{BC}) \\
=& \frac{1}{4}(Var(\hat{D}_{AB} + \hat{D}_{AC}) + Var(\hat{D}_{BC}) - 2Cov(\hat{D}_{AB} + \hat{D}_{AC}, \hat{D}_{BC})) \\
=& \frac{1}{4}(Var(\hat{D}_{AB}) + Var(\hat{D}_{AC}) + Var(\hat{D}_{BC}) \\
& + 2Cov(\hat{D}_{AB}, \hat{D}_{AC}) - 2Cov(\hat{D}_{AB}, \hat{D}_{BC}) - 2Cov(\hat{D}_{AC}, \hat{D}_{BC}))
\end{aligned}$$

To calculate this both the variance of the D_{ij} terms and the covariances between them are needed. The variance of a JC distance, when the sequence is of length n, is

$$Var(\hat{D}_{ij}) \simeq \frac{p_{ij}(1 - p_{ij})}{n(1 - \frac{4}{3}p_{ij})^2}$$

(Bulmer 1991; Tajima 1993). Substituting in equation A.6 we obtain

$$Var(\hat{D}_{ij}) \simeq \frac{3}{16n}(e^{\frac{4}{3}D_{ij}} - 1)(e^{\frac{4}{3}D_{ij}} + 3)$$

The covariance between $\hat{D}_{ij}$ and $\hat{D}_{ik}$ is the variance of the length of the branch shared by the paths between both sets of sequences (Nei and Jin 1989; Rzhetsky and Nei 1992). Substituting these into the expression for the variance gives:

$$\begin{aligned}
Var(\hat{d}_A) \simeq& \frac{3}{64n}((e^{\frac{4}{3}(d_A+d_B)} - 1)(e^{\frac{4}{3}(d_A+d_B)} + 3) + (e^{\frac{4}{3}(d_A+d_C)} - 1)(e^{\frac{4}{3}(d_A+d_C)} + 3) \\
& + (e^{\frac{4}{3}(d_B+d_C)} - 1)(e^{\frac{4}{3}(d_B+d_C)} + 3)) + 2(e^{\frac{4}{3}(d_A} - 1)(e^{\frac{4}{3}(d_A)} + 3) \\
& - 2(e^{\frac{4}{3}(d_B)} - 1)(e^{\frac{4}{3}(d_B)} + 3) - 2(e^{\frac{4}{3}(d_C)} - 1)(e^{\frac{4}{3}(d_C)} + 3))
\end{aligned}$$

Equivalent formulae can be derived for the other two branch lengths on the three-species tree. The derivations of these formulae involve approximations. The accuracy of these approximations increases with sequence length and reduces as branch length increases. As $d_A = 0.1$, which is short, $E(\hat{d}_A)$ should be accurate. However, the formula for the variance may contain slightly more error.

A.3 Supplementary Figures

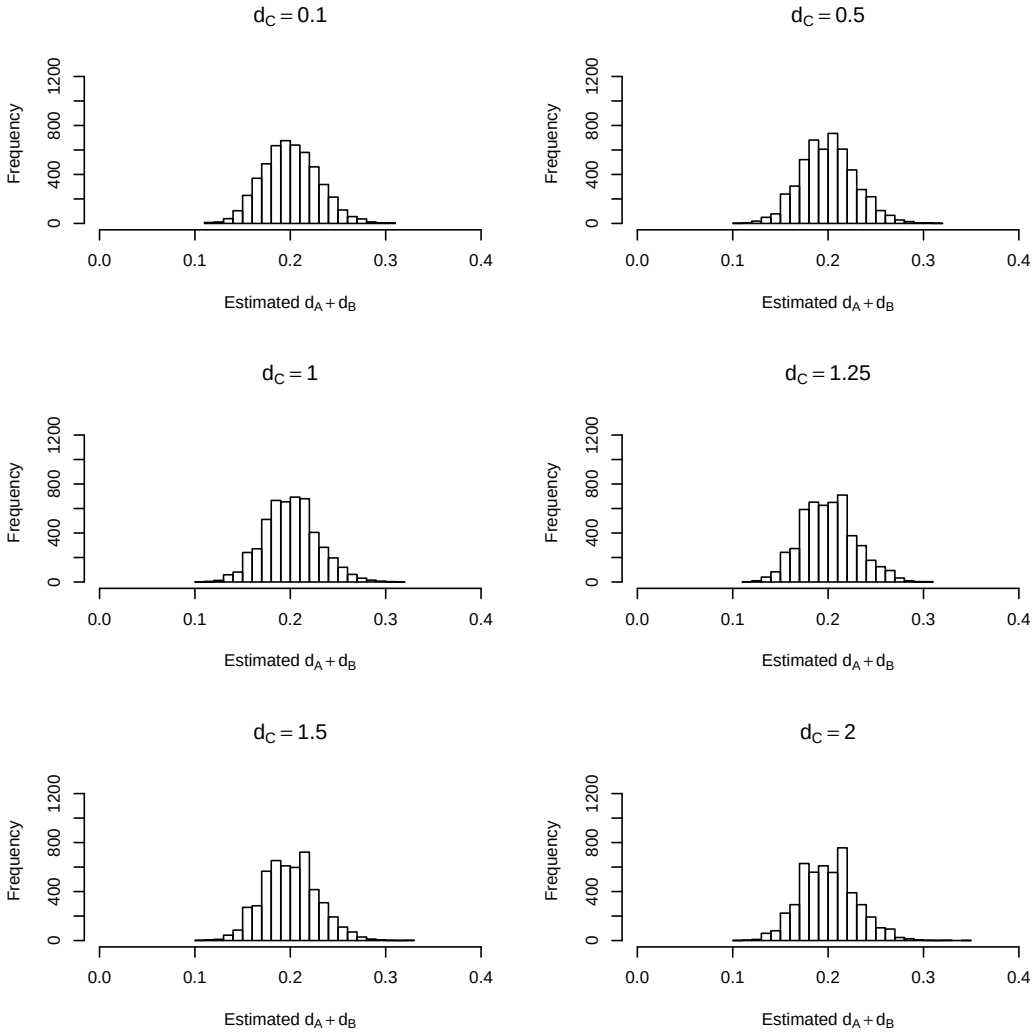


Figure A.1: Distributions of estimated values of $d_A + d_B$ for trees simulated with $d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$. Kolmogorov-Smirnoff tests with multiple testing correction indicate no significant differences between these distributions.

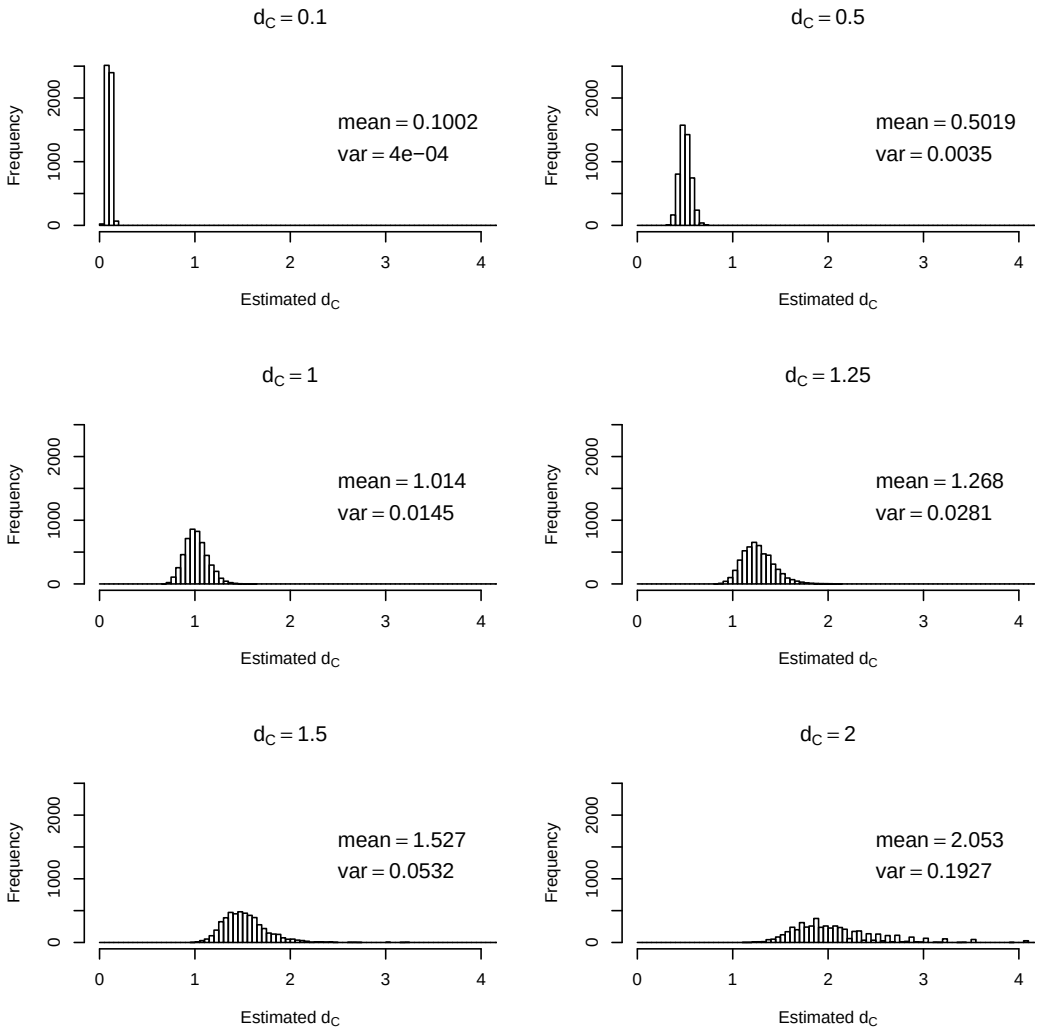


Figure A.2: Distributions of estimated values of d_C for trees simulated with $d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$. Note that the estimates appear unbiased, with variance increasing as the value of d_C increases.

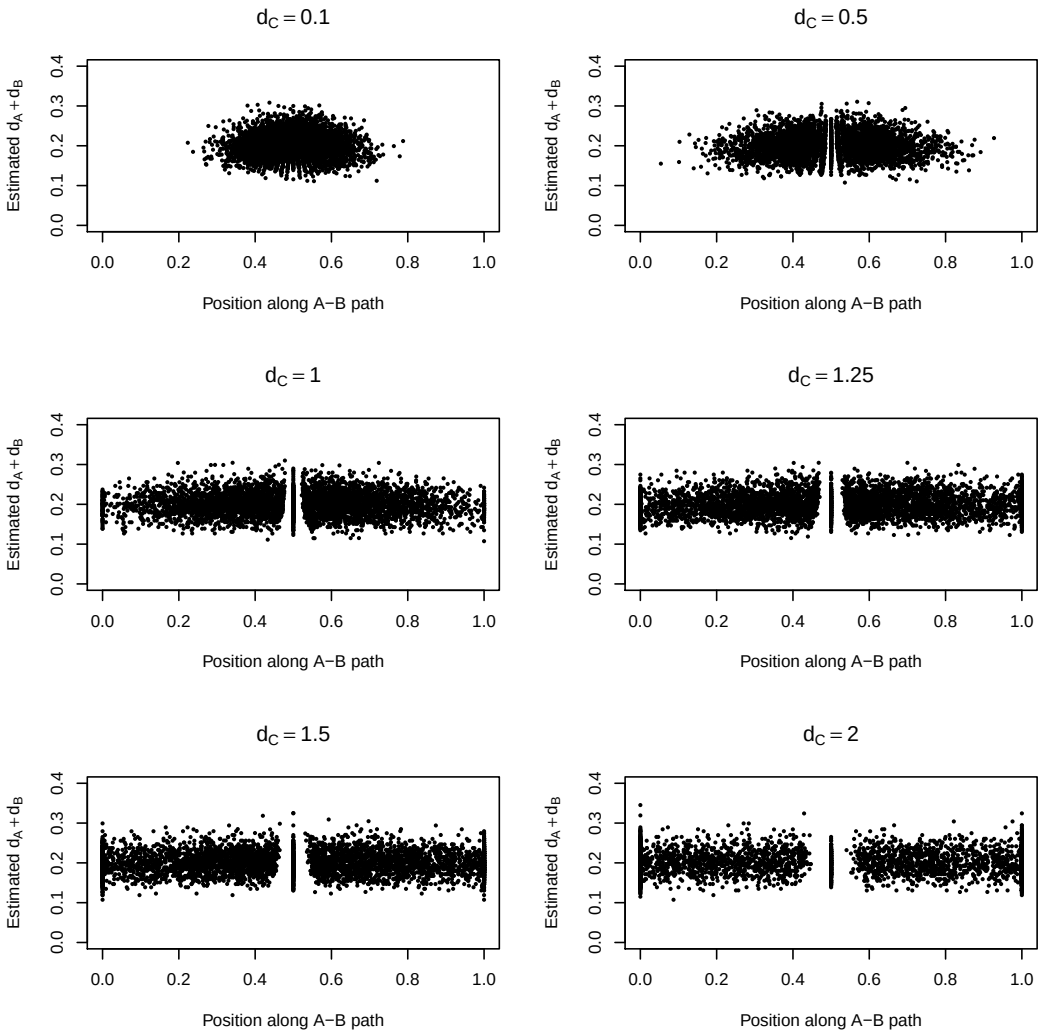


Figure A.3: The location of the branch leading to C on the A–B path against estimated values of $d_A + d_B$ for trees simulated with $d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$. There is no relationship between the two variables.

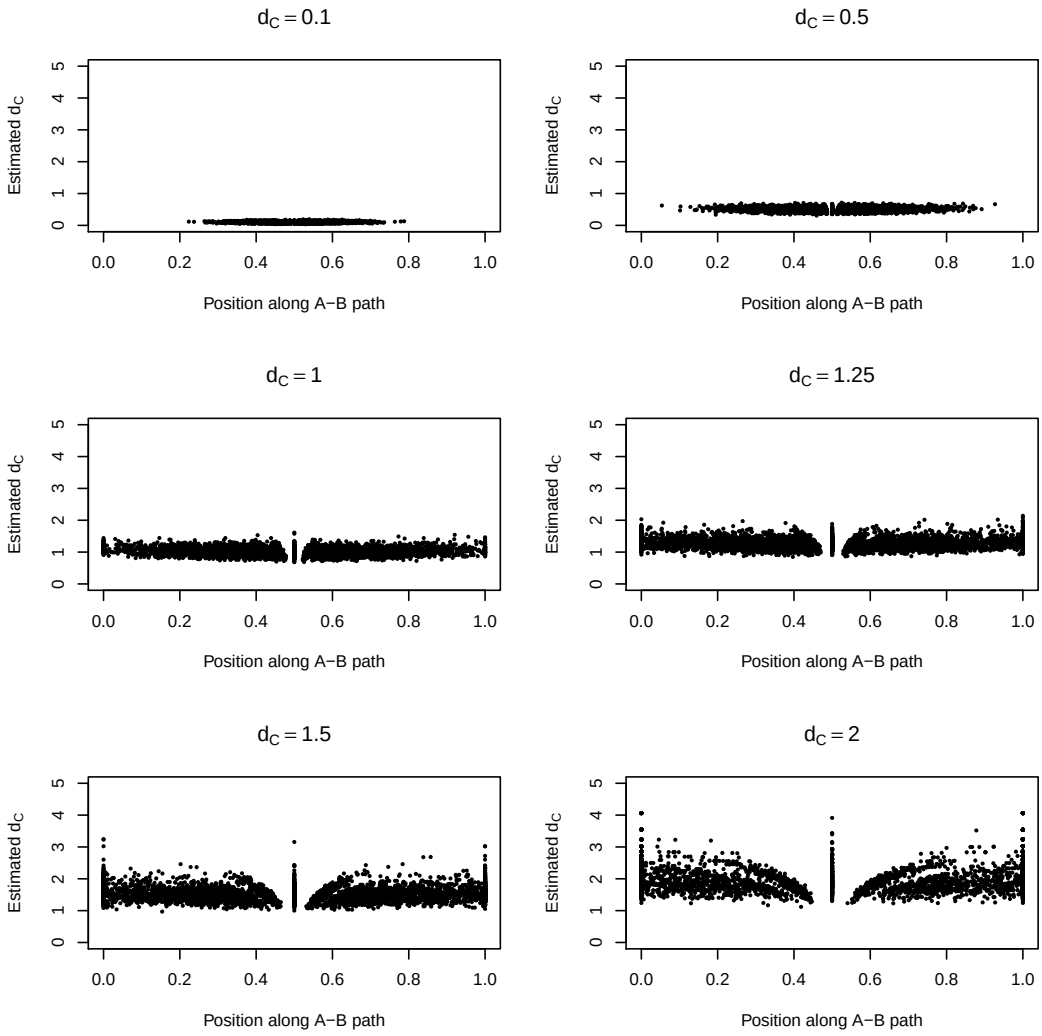


Figure A.4: The location of the branch leading to C on the A–B path against estimated values of d_C for trees simulated with $d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$. There is no relationship between the two variables.

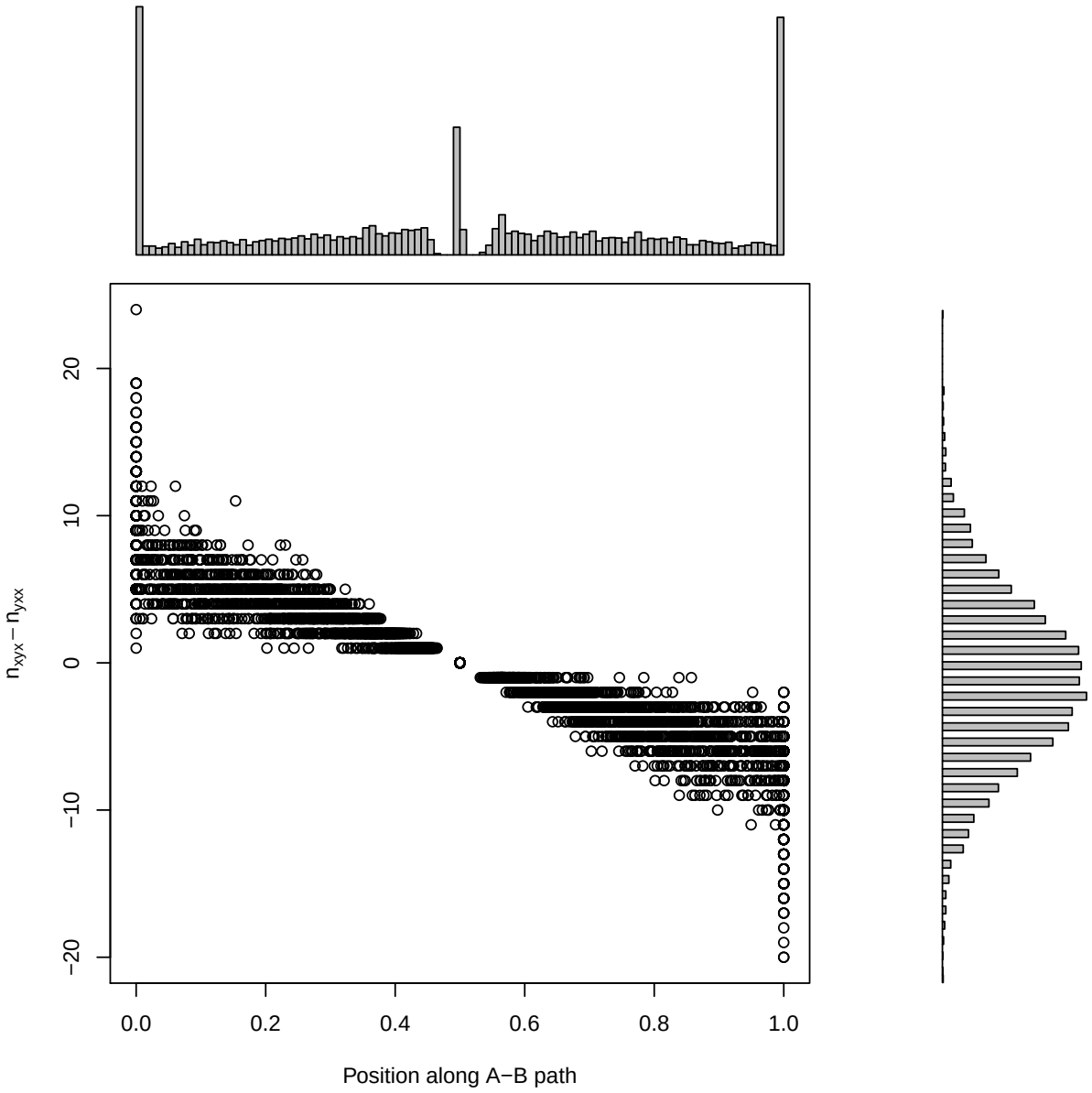


Figure A.5: $n_{xyx} - n_{yxx}$ against location of the branch leading to C on the A–B path.

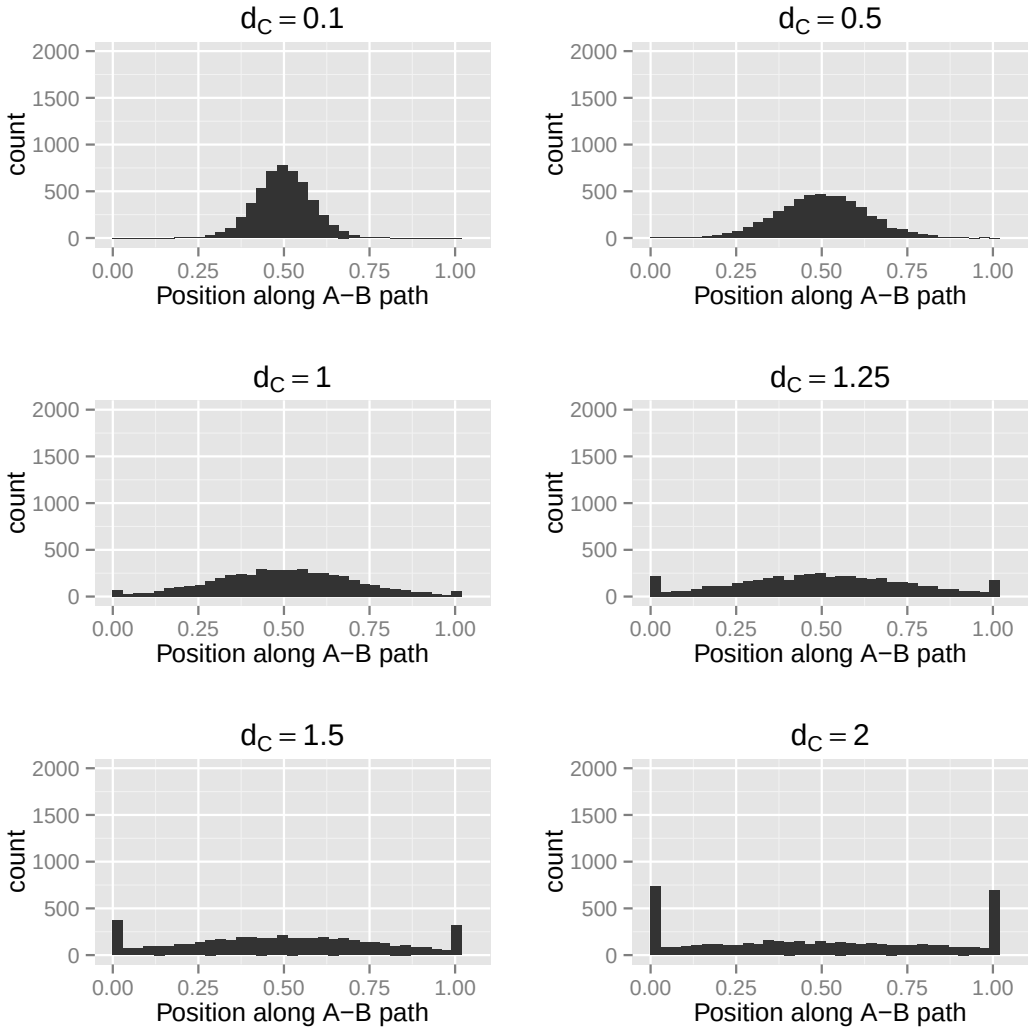


Figure A.6: Distributions of the location of the branch leading to C on the A–B path for trees with $d_C = 0.1, 0.5, 1, 1.25, 1.5, 2$. These simulations were run using the GTR model for simulation and tree reconstruction with parameters taken from Murphy et al. 2001.

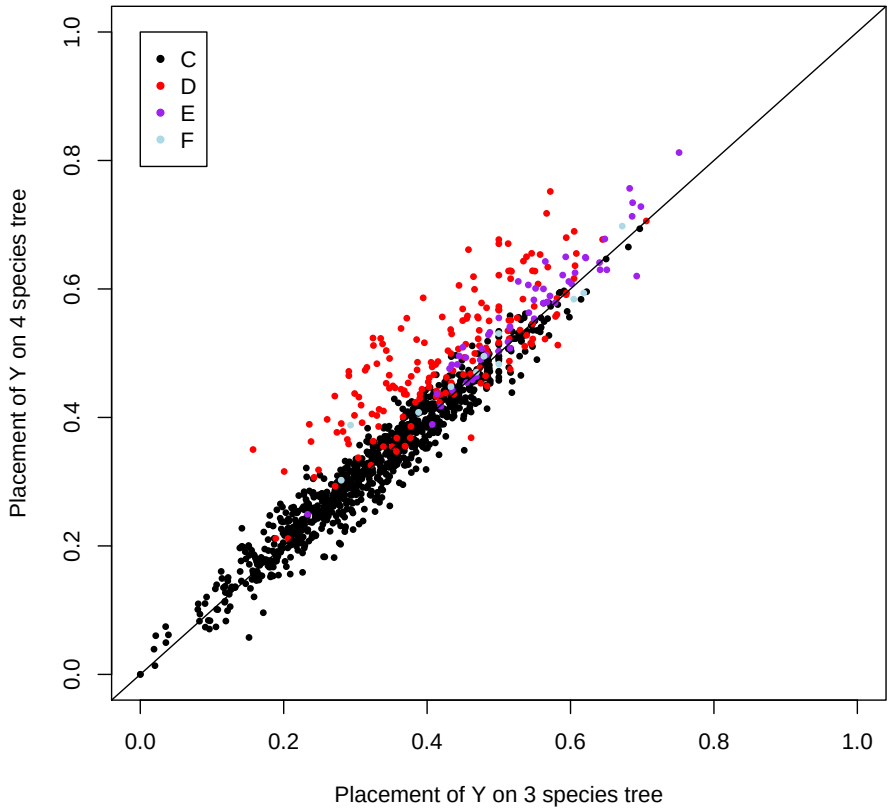


Figure A.7: Position of Y on inferred three-species trees versus position of Y on inferred four-species trees for data simulated on tree 3.9a with long branch length 1.5. Points coloured according to the ML reconstructed topology shown in Figure 3.9. The position of Y was measured as a fraction along the W–X path. In the case of topology 3.9d, as the branch to Y does not directly meet this path, the position along the W–X path where the middle branch joins is used.

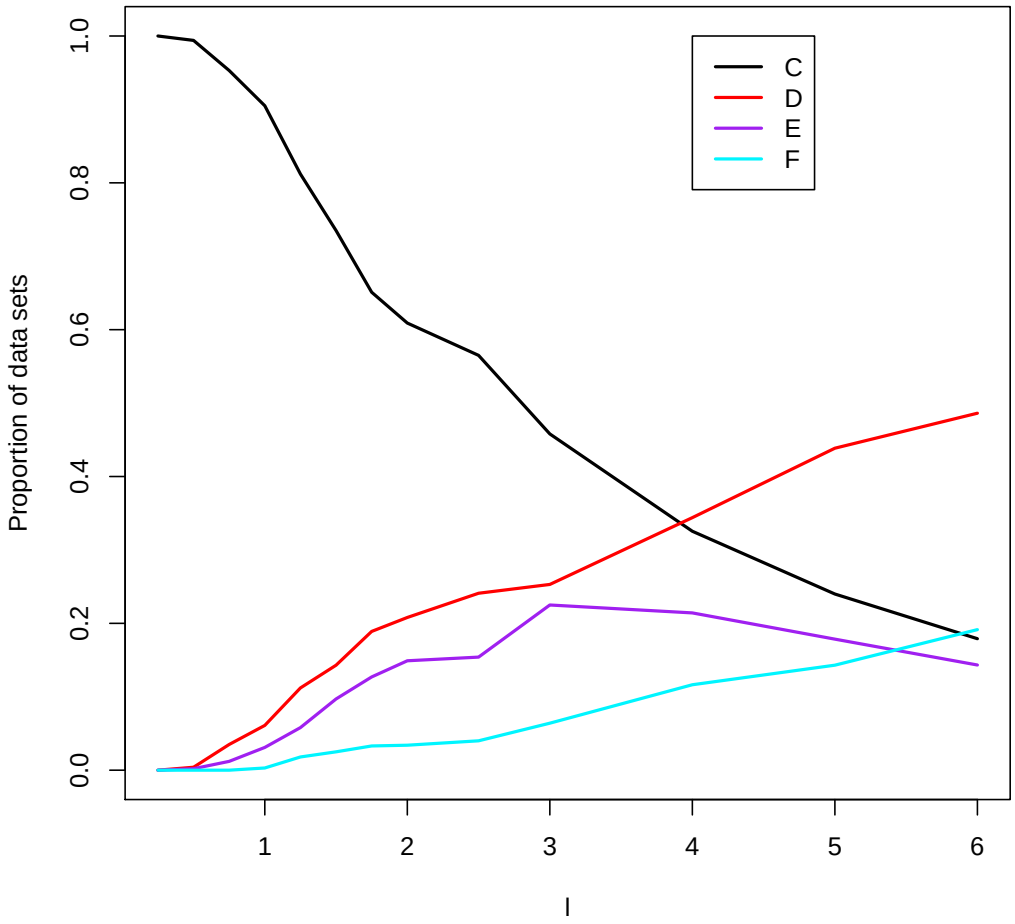


Figure A.8: The proportions of different topologies obtained for different lengths of Y and Z. These simulations were run using the GTR model for simulation and tree reconstruction with parameters taken from Murphy et al. 2001.

Appendix B

B.1 Supplementary Figures

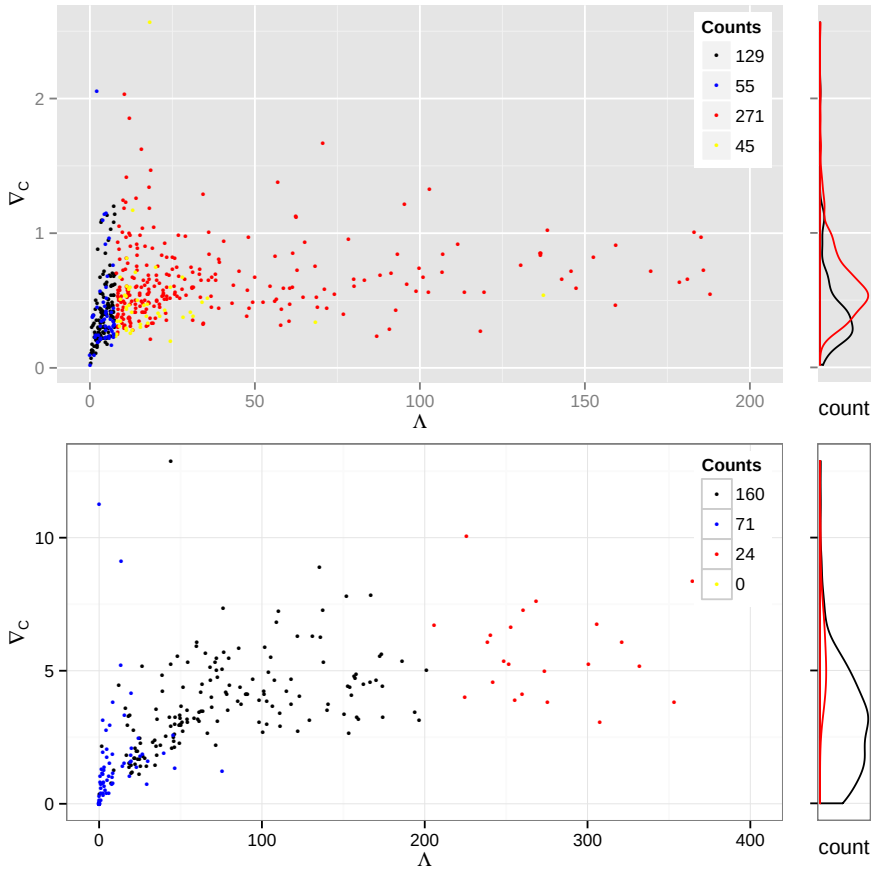


Figure B.1: Relationship between the LRT statistic Λ and ∇_C for nucleotides (top, grey backgound) and amino acids (bottom, white background). Each point corresponds to a MSA. Pandit MSAs are red if they are significantly better described by a non-reversible model and black otherwise. Mammal MSAs are yellow if they are significantly better described by a non-reversible model and blue otherwise. Histograms on the right show ∇_C for MSAs found to be significant (red) or non-significant (black).

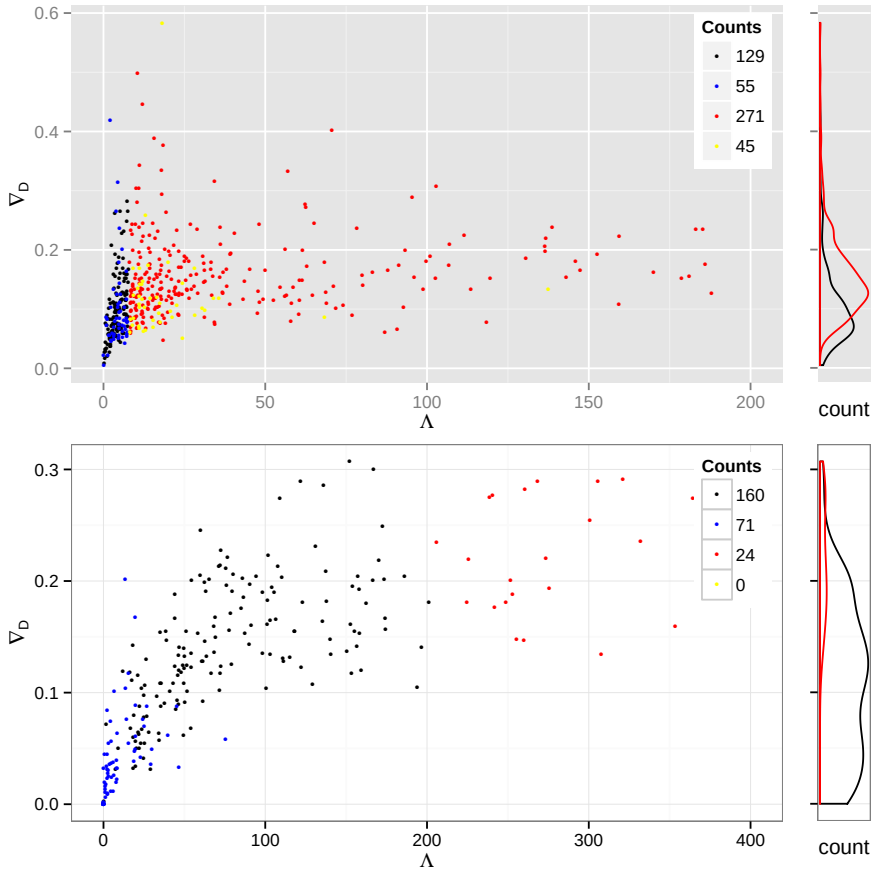


Figure B.2: Relationship between the LRT statistic Λ and ∇_D for nucleotides (top, grey backgound) and amino acids (bottom, white background). Each point corresponds to a MSA. Pandit MSAs are red if they are significantly better described by a non-reversible model and black otherwise. Mammal MSAs are yellow if they are significantly better described by a non-reversible model and blue otherwise. Histograms on the right show ∇_D for MSAs found to be significant (red) or non-significant (black).

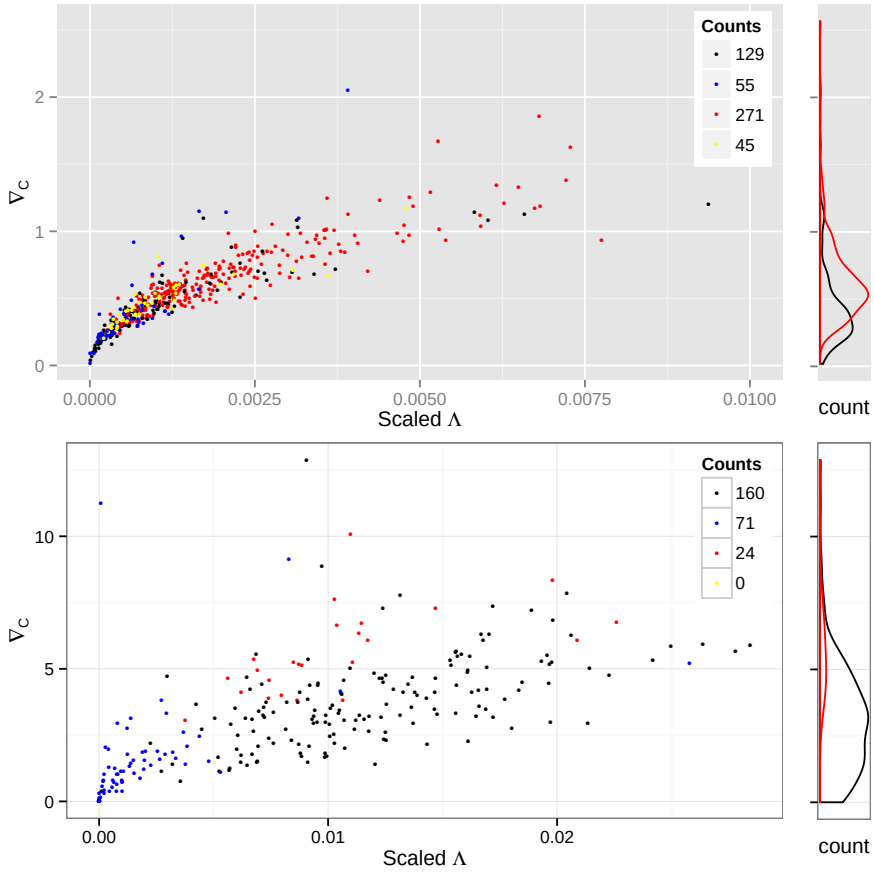


Figure B.3: Relationship between scaled Λ and ∇_C for nucleotides (top, grey backgound) and amino acids (bottom, white background). Plots are as in B.1, except Λ is now scaled by the inverse of the information content, as approximated by the product of sequence length and sequence number, minus the number of gaps.

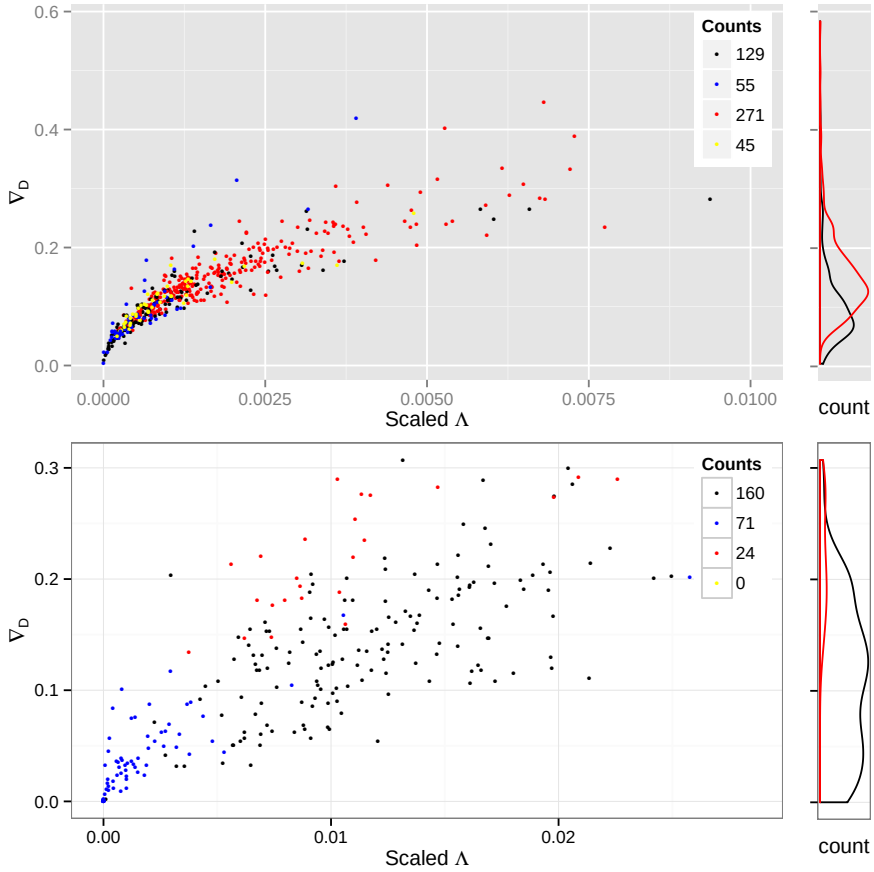


Figure B.4: Relationship between scaled Λ and ∇_D for nucleotides (top, grey backgound) and amino acids (bottom, white background). Plots are as in B.2, except Λ is now scaled by the inverse of the information content, as approximated by the product of sequence length and sequence number, minus the number of gaps.

B.2 Dataset Lists

B.2.1 Pandit Data

A list of the Pandit data sets used in this study. This data is available from http://www.ebi.ac.uk/research/goldman/software/pandit

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00001	Yes	Yes	1299	64
PF00002	Yes	Yes	1128	32
PF00003	Yes	No	876	29
PF00005	Yes	Yes	999	60
PF00007	Yes	No	336	20
PF00008	Yes	No	153	55
PF00009	Yes	Yes	1290	208
PF00011	Yes	Yes	345	34
PF00012	Yes	Yes	2064	33
PF00013	Yes	No	279	421
PF00014	Yes	No	240	126
PF00015	Yes	Yes	855	10
PF00016	Yes	No	960	16
PF00017	Yes	Yes	327	54
PF00018	Yes	No	195	60
PF00019	Yes	Yes	354	21
PF00020	Yes	No	150	35
PF00021	Yes	Yes	327	42
PF00022	Yes	Yes	1167	20
PF00024	Yes	Yes	456	101
PF00025	Yes	Yes	603	19
PF00026	Yes	Yes	1548	21
PF00027	Yes	Yes	426	349
PF00028	Yes	Yes	387	57
PF00029	Yes	Yes	777	15
PF00030	Yes	No	285	33
PF00031	Yes	Yes	327	40
PF00032	Yes	No	348	9
PF00033	Yes	Yes	636	8
PF00034	Yes	Yes	393	26
PF00035	Yes	No	231	76
PF00038	Yes	Yes	1131	30
PF00039	Yes	No	126	7
PF00040	Yes	No	126	15
PF00041	Yes	Yes	360	89

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00042	Yes	Yes	501	34
PF00043	Yes	Yes	471	58
PF00044	Yes	Yes	630	108
PF00045	Yes	No	174	68
PF00046	Yes	No	213	179
PF00047	Yes	No	279	79
PF00048	Yes	No	225	103
PF00049	Yes	Yes	528	26
PF00050	Yes	No	198	33
PF00051	Yes	No	255	23
PF00052	Yes	Yes	450	9
PF00053	Yes	No	216	67
PF00054	Yes	Yes	519	16
PF00055	Yes	No	792	7
PF00056	Yes	Yes	525	28
PF00057	Yes	No	186	50
PF00058	Yes	No	171	27
PF00059	Yes	Yes	408	42
PF00060	Yes	Yes	1275	43
PF00061	Yes	Yes	558	132
PF00062	Yes	Yes	384	11
PF00063	Yes	Yes	2868	23
PF00064	Yes	Yes	1458	7
PF00066	Yes	No	126	10
PF00067	Yes	Yes	1677	49
PF00068	Yes	Yes	414	14
PF00069	Yes	Yes	1419	51
PF00070	Yes	Yes	390	127
PF00071	Yes	No	687	60
PF00072	Yes	Yes	474	53
PF00073	Yes	Yes	939	44
PF00074	Yes	Yes	426	11
PF00075	Yes	Yes	738	51
PF00076	Yes	No	240	75
PF00077	Yes	Yes	381	37
PF00078	Yes	Yes	1191	145
PF00079	Yes	Yes	1269	39
PF00080	Yes	No	609	23
PF00081	Yes	No	273	22
PF00083	Yes	Yes	1851	48
PF00084	Yes	No	276	78
PF00085	Yes	Yes	366	47

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00086	Yes	No	252	19
PF00087	Yes	No	243	7
PF00088	Yes	No	150	26
PF00089	Yes	Yes	1065	65
PF00090	Yes	No	186	30
PF00091	Yes	Yes	855	119
PF00092	Yes	Yes	888	179
PF00093	Yes	No	297	19
PF00094	Yes	Yes	645	68
PF00095	Yes	No	204	76
PF00096	Yes	No	105	168
PF00097	Yes	No	222	63
PF00100	Yes	Yes	1209	82
PF00101	Yes	Yes	384	45
PF00102	Yes	Yes	1353	111
PF00103	Yes	Yes	711	17
PF00105	Yes	No	240	25
PF00106	Yes	No	654	261
PF00107	Yes	No	648	107
PF00108	Yes	Yes	873	19
PF00109	Yes	Yes	1041	168
PF00110	Yes	No	1182	15
PF00111	Yes	Yes	369	188
PF00112	Yes	Yes	1101	154
PF00113	Yes	Yes	894	11
PF00114	Yes	Yes	465	20
PF00115	Yes	Yes	1572	23
PF00116	Yes	Yes	408	31
PF00117	Yes	Yes	876	131
PF00118	Yes	Yes	1875	40
PF00119	Yes	Yes	594	27
PF00120	Yes	Yes	1020	103
PF00121	Yes	Yes	849	54
PF00122	Yes	Yes	1011	102
PF00124	Yes	Yes	930	12
PF00125	Yes	No	255	69
PF00126	Yes	No	255	1627
PF00127	Yes	No	387	19
PF00128	Yes	Yes	2019	51
PF00129	Yes	Yes	543	24
PF00130	Yes	No	225	43
PF00131	Yes	No	240	14

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00133	Yes	Yes	2442	20
PF00134	Yes	Yes	582	149
PF00135	Yes	Yes	2667	118
PF00137	Yes	No	210	35
PF00138	Yes	No	171	43
PF00139	Yes	Yes	687	13
PF00140	Yes	No	117	62
PF00141	Yes	Yes	1797	293
PF00142	Yes	Yes	846	15
PF00143	Yes	Yes	579	17
PF00145	Yes	Yes	1638	27
PF00146	Yes	Yes	1032	24
PF00147	Yes	Yes	816	10
PF00148	Yes	Yes	1704	31
PF00150	Yes	Yes	1539	63
PF00151	Yes	Yes	1137	14
PF00153	Yes	Yes	456	202
PF00154	Yes	Yes	987	26
PF00155	Yes	No	1578	47
PF00156	Yes	Yes	723	126
PF00157	Yes	No	234	10
PF00158	Yes	Yes	852	289
PF00160	Yes	Yes	573	20
PF00161	Yes	Yes	1053	17
PF00162	Yes	Yes	1590	25
PF00163	Yes	Yes	339	19
PF00164	Yes	Yes	441	15
PF00165	Yes	No	150	88
PF00166	Yes	Yes	381	48
PF00167	Yes	Yes	498	27
PF00170	Yes	No	195	21
PF00171	Yes	No	1848	118
PF00172	Yes	No	156	27
PF00174	Yes	Yes	783	13
PF00175	Yes	Yes	447	71
PF00176	Yes	Yes	1530	22
PF00177	Yes	Yes	555	18
PF00178	Yes	No	276	14
PF00179	Yes	Yes	717	72
PF00180	Yes	Yes	1392	23
PF00181	Yes	No	282	60
PF00182	Yes	Yes	711	15

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00183	Yes	Yes	1713	10
PF00184	Yes	No	240	10
PF00185	Yes	Yes	561	19
PF00186	Yes	Yes	771	16
PF00187	Yes	No	132	17
PF00188	Yes	No	849	289
PF00189	Yes	No	282	24
PF00191	Yes	No	243	170
PF00193	Yes	Yes	315	8
PF00194	Yes	Yes	858	11
PF00195	Yes	Yes	696	20
PF00196	Yes	No	174	30
PF00197	Yes	Yes	618	9
PF00198	Yes	Yes	870	78
PF00199	Yes	Yes	1224	18
PF00200	Yes	Yes	333	85
PF00201	Yes	Yes	1548	14
PF00202	Yes	Yes	1236	17
PF00203	Yes	No	246	20
PF00204	Yes	Yes	705	63
PF00205	Yes	Yes	561	22
PF00206	Yes	Yes	1035	19
PF00207	Yes	Yes	333	58
PF00208	Yes	Yes	990	126
PF00209	Yes	Yes	1926	12
PF00210	Yes	Yes	612	92
PF00211	Yes	Yes	744	20
PF00212	Yes	Yes	315	12
PF00213	Yes	Yes	540	20
PF00214	Yes	No	126	9
PF00215	Yes	Yes	1206	87
PF00216	Yes	Yes	318	76
PF00217	Yes	Yes	825	6
PF00218	Yes	Yes	846	20
PF00219	Yes	No	288	11
PF00221	Yes	Yes	1551	8
PF00224	Yes	Yes	1161	11
PF00225	Yes	Yes	1944	92
PF00226	Yes	Yes	309	264
PF00227	Yes	Yes	807	178
PF00229	Yes	Yes	513	52
PF00230	Yes	Yes	1005	17

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00231	Yes	Yes	1029	18
PF00232	Yes	Yes	1707	14
PF00233	Yes	Yes	765	7
PF00235	Yes	Yes	480	14
PF00236	Yes	No	291	8
PF00237	Yes	Yes	420	20
PF00238	Yes	Yes	426	18
PF00239	Yes	Yes	501	16
PF00240	Yes	No	267	85
PF00241	Yes	Yes	495	15
PF00242	Yes	Yes	1269	6
PF00243	Yes	Yes	381	7
PF00244	Yes	Yes	726	15
PF00245	Yes	Yes	1833	18
PF00248	Yes	Yes	1206	63
PF00249	Yes	No	300	196
PF00250	Yes	Yes	420	21
PF00251	Yes	No	1647	19
PF00252	Yes	Yes	444	16
PF00253	Yes	No	171	20
PF00255	Yes	Yes	354	12
PF00256	Yes	No	102	88
PF00257	Yes	Yes	1188	24
PF00258	Yes	Yes	720	91
PF00260	Yes	No	216	18
PF00261	Yes	No	711	8
PF00262	Yes	Yes	1167	11
PF00263	Yes	Yes	1044	54
PF00264	Yes	Yes	1224	13
PF00265	Yes	Yes	573	13
PF00266	Yes	Yes	1152	14
PF00267	Yes	Yes	1269	15
PF00268	Yes	Yes	897	13
PF00269	Yes	No	186	7
PF00272	Yes	No	102	13
PF00273	Yes	Yes	558	43
PF00274	Yes	Yes	1059	11
PF00275	Yes	Yes	1518	33
PF00276	Yes	No	291	12
PF00277	Yes	Yes	336	6
PF00278	Yes	Yes	660	55
PF00280	Yes	No	195	7

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00281	Yes	No	183	16
PF00282	Yes	Yes	1221	12
PF00284	Yes	No	120	14
PF00285	Yes	Yes	1233	20
PF00286	Yes	Yes	429	15
PF00287	Yes	Yes	1032	11
PF00289	Yes	Yes	405	18
PF00290	Yes	Yes	849	21
PF00291	Yes	Yes	1308	96
PF00292	Yes	Yes	378	6
PF00294	Yes	Yes	1212	77
PF00295	Yes	Yes	1170	15
PF00296	Yes	Yes	1119	15
PF00297	Yes	Yes	918	14
PF00298	Yes	No	213	13
PF00301	Yes	No	162	33
PF00302	Yes	Yes	618	11
PF00303	Yes	Yes	1152	18
PF00304	Yes	No	177	39
PF00305	Yes	Yes	2094	17
PF00306	Yes	Yes	633	149
PF00307	Yes	Yes	453	196
PF00308	Yes	No	1017	12
PF00309	Yes	No	147	29
PF00310	Yes	No	1737	59
PF00312	Yes	No	264	7
PF00313	Yes	No	225	38
PF00314	Yes	No	645	5
PF00316	Yes	No	1140	14
PF00318	Yes	No	666	20
PF00319	Yes	No	153	10
PF00320	Yes	No	135	75
PF00324	Yes	No	1863	28
PF00325	Yes	No	102	12
PF00326	Yes	No	840	72
PF00327	Yes	No	165	12
PF00329	Yes	No	237	55
PF00330	Yes	No	1593	11
PF00331	Yes	No	1533	53
PF00332	Yes	No	999	13
PF00333	Yes	No	204	11
PF00334	Yes	No	405	11

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00335	Yes	No	1086	67
PF00338	Yes	No	291	15
PF00339	Yes	No	615	51
PF00340	Yes	No	462	16
PF00342	Yes	No	1566	9
PF00343	Yes	No	2241	8
PF00344	Yes	No	1188	17
PF00345	Yes	No	447	20
PF00346	Yes	No	825	6
PF00347	Yes	No	285	131
PF00348	Yes	No	867	16
PF00349	Yes	No	717	15
PF00351	Yes	No	996	6
PF00352	Yes	No	318	54
PF00354	Yes	No	627	9
PF00355	Yes	No	399	157
PF00358	Yes	No	399	14
PF00359	Yes	No	465	13
PF00360	Yes	No	546	9
PF00361	Yes	No	1056	31
PF00362	Yes	No	1332	17
PF00363	Yes	No	291	13
PF00364	Yes	No	228	47
PF00365	Yes	No	939	12
PF00366	Yes	No	213	14
PF00367	Yes	No	105	18
PF00368	Yes	No	1299	21
PF00370	Yes	No	801	12
PF00372	Yes	No	1044	9
PF00373	Yes	No	630	8
PF00374	Yes	No	1746	11
PF00375	Yes	No	1419	12
PF00376	Yes	No	150	361
PF00377	Yes	No	396	7
PF00378	Yes	No	528	12
PF00379	Yes	No	219	68
PF00380	Yes	No	408	9
PF00381	Yes	No	261	11
PF00382	Yes	No	222	10
PF00383	Yes	No	618	61
PF00384	Yes	No	2424	30
PF00385	Yes	No	267	181

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00386	Yes	No	483	49
PF00387	Yes	No	384	8
PF00388	Yes	No	462	13
PF00389	Yes	No	315	175
PF00390	Yes	No	597	55
PF00391	Yes	No	429	90
PF00392	Yes	No	192	26
PF00393	Yes	No	990	56
PF00394	Yes	No	657	71
PF00395	Yes	No	177	85
PF00396	Yes	No	129	14
PF00398	Yes	No	1020	21
PF00401	Yes	No	153	20
PF00403	Yes	No	219	126
PF00405	Yes	No	1050	8
PF00406	Yes	No	573	20
PF00407	Yes	No	486	35
PF00408	Yes	No	342	16
PF00410	Yes	No	516	20
PF00411	Yes	No	372	11
PF00412	Yes	No	222	40
PF00415	Yes	No	204	7
PF00416	Yes	No	396	13
PF00417	Yes	No	291	24
PF00418	Yes	No	105	11
PF00419	Yes	No	630	24
PF00420	Yes	No	417	194
PF00421	Yes	No	1575	17
PF00423	Yes	No	1884	13
PF00424	Yes	No	405	12
PF00425	Yes	No	978	27
PF00426	Yes	No	2370	7
PF00427	Yes	No	450	23
PF00428	Yes	No	357	24
PF00429	Yes	No	2343	15
PF00430	Yes	No	396	25
PF00431	Yes	No	474	30
PF00432	Yes	No	177	70
PF00434	Yes	No	987	8
PF00435	Yes	No	348	79
PF00436	Yes	No	390	74
PF00437	Yes	No	1053	23

Pandit ID	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
PF00438	Yes	No	303	12
PF00439	Yes	No	282	66
PF00440	Yes	No	147	108
PF00441	Yes	No	537	61
PF00444	Yes	No	114	7
PF00445	Yes	No	795	62
PF00447	Yes	No	624	7
PF00448	Yes	No	666	56
PF00449	Yes	No	378	11
PF00450	Yes	No	2052	81
PF00452	Yes	No	336	14
PF00453	Yes	No	333	12
PF00454	Yes	No	1449	45
PF00455	Yes	No	705	14
PF00456	Yes	No	1029	12
PF00457	Yes	No	612	11
PF00458	Yes	No	171	14
PF00462	Yes	No	396	36
PF00463	Yes	No	1620	7
PF00464	Yes	No	1206	12
PF00465	Yes	No	1257	14
PF00466	Yes	No	318	26
PF00467	Yes	No	132	225
PF00468	Yes	No	132	5
PF00469	Yes	No	753	18
PF00471	Yes	No	174	8
PF00472	Yes	No	420	35
PF00474	Yes	No	1341	8
PF00475	Yes	No	450	11

B.2.2 Mammal Data

A list of the mammal data sets used in this study. Alignments were created from a genome-wide set of mammalian gene alignments with an associated rooted tree derived from the Mammalian Genome Project (Lindblad-Toh et al. 2011) augmented using other mammalian genomes from release 63 of the Ensembl database (Flicek et al. 2011) by Jordan (2011). Data sets are available from http://www.ebi.ac.uk/goldman-srv/Reversibility/

Gene ID	Gene Name	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
ENSG00000164898	C7orf55	Yes	Yes	240	27
ENSG00000136518	ACTL6A	Yes	No	600	38
ENSG00000038295	TLL1	Yes	Yes	1800	37
ENSG00000165219	GAPVD1	Yes	Yes	1920	36
ENSG00000163710	PCOLCE2	Yes	Yes	720	36
ENSG00000184903	IMMP2L	Yes	Yes	240	29
ENSG00000157554	ERG	Yes	Yes	720	34
ENSMUSG00000004268	Emg1	Yes	No	360	32
ENSG00000168772	CXXC4	Yes	No	480	6
ENSG00000154945	ANKRD40	Yes	No	480	29
ENSG00000050748	MAPK9	Yes	Yes	600	34
ENSG00000177565	TBL1XR1	Yes	No	720	34
ENSG00000206579	XKR4	Yes	Yes	1440	25
ENSG00000147099	HDAC8	Yes	Yes	480	33
ENSG00000161594	KLHL10	Yes	No	960	36
ENSG00000140750	ARHGAP17	Yes	No	1320	36
ENSG00000071189	SNX13	Yes	Yes	1200	36
ENSG00000103460	TOX3	Yes	Yes	840	38
ENSG00000161835	GRASP	Yes	Yes	720	26
ENSCAFG00000023168	HMGN2Ω_CANFA	Yes	No	120	5
ENSG00000171227	TMEM37	Yes	Yes	360	25
ENSG00000169856	ONECUT1	Yes	Yes	600	26
ENSG00000136367	ZFHX2	Yes	Yes	3840	20
ENSG00000166526	ZNF3	Yes	Yes	720	30
ENSG00000134490	C18orf45	Yes	Yes	480	35
ENSG00000204952	FBXO47	Yes	Yes	600	35
ENSG00000143032	BARHL2	Yes	No	480	36
ENSG00000110514	MADD	Yes	Yes	1800	37
ENSG00000197956	S100A6	Yes	Yes	120	24
ENSG00000099326	MZF1	Yes	Yes	1200	25
ENSG00000136738	STAM	Yes	Yes	720	38
ENSG00000103091	WDR59	Yes	Yes	1440	37
ENSG00000109911	ELP4	Yes	No	1080	32

Gene ID	Gene Name	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
ENSG00000198517	MAFK	Yes	Yes	240	28
ENSG00000064835	POU1F1	Yes	Yes	600	36
ENSG00000196104	SPOCK3	Yes	Yes	600	31
ENSG00000169375	SIN3A	Yes	No	1440	38
ENSG00000198937	C6orf129	Yes	No	360	32
ENSG00000143702	CEP170	Yes	Yes	2160	29
ENSG00000112246	SIM1	Yes	Yes	960	38
ENSG00000175514	GPR152	Yes	Yes	840	19
ENSG00000163902	RPN1	Yes	Yes	720	30
ENSG00000163214	DHX57	Yes	Yes	1560	35
ENSG00000147408	CSGALNACT1	Yes	No	840	38
ENSG00000173597	SULT1B1	Yes	Yes	360	27
ENSG00000087301	TXNDC16	Yes	Yes	1080	35
ENSG00000151062	CACNA2D4	Yes	Yes	1560	30
ENSG00000154065	ANKRD29	Yes	Yes	1080	35
ENSG00000079393	DUSP13	Yes	Yes	600	24
ENSG00000116353	MECR	Yes	Yes	600	38
ENSG00000117090	SLAMF1	Yes	No	720	33
ENSG00000185689	C6orf201	Yes	Yes	480	31
ENSG00000162618	ELTD1	Yes	Yes	1200	37
ENSG00000237521	OR7E24	Yes	Yes	480	6
ENSG00000137825	ITPKA	Yes	No	720	22
ENSG00000064313	TAF2	Yes	No	1440	36
ENSG00000067533	RRP15	Yes	Yes	600	32
ENSG00000167037	SGSM1	Yes	Yes	2760	35
ENSG00000117069	ST6GALNAC5	Yes	Yes	480	29
ENSG00000168818	STX18	Yes	No	600	32
ENSG00000143479	DYRK3	Yes	Yes	720	33
ENSG00000134463	ECHDC3	Yes	Yes	480	32
ENSG00000005238	KIAA1539	Yes	No	600	31
ENSG00000157343	C6orf81	Yes	Yes	600	32
ENSG00000144730	IL17RD	Yes	Yes	1080	35
ENSG00000148019	CEP78	Yes	Yes	1560	36
ENSG00000168758	SEMA4C	Yes	Yes	1080	34
ENSG00000225190	PLEKHM1	Yes	Yes	1680	26
ENSG00000151835	SACS	Yes	Yes	4800	35
ENSG00000167751	KLK2	Yes	Yes	480	5
ENSG00000131459	GFPT2	Yes	Yes	960	37
ENSG00000196678	ERI2	Yes	No	960	35
ENSG00000167880	EVPL	Yes	Yes	2280	33
ENSG00000136717	BIN1	Yes	Yes	1200	31
ENSG00000196878	LAMB3	Yes	Yes	1440	32

Gene ID	Gene Name	Nucleotides?	Amino Acids?	Sequence Length	Sequence Number
ENSG00000213123	TCTEX1D2	Yes	Yes	360	20
ENSG00000107611	CUBN	Yes	Yes	4560	37
ENSG00000152049	KCNE4	Yes	No	360	36
ENSG00000162066	AMDHD2	Yes	No	1080	34
ENSG00000130021	HDHD1	Yes	Yes	480	20
ENSG00000121895	TMEM156	Yes	Yes	600	37
ENSG00000156127	BATF	Yes	No	240	33
ENSG00000184207	PGP	Yes	No	480	12
ENSG00000105429	MEGF8	Yes	Yes	3720	30
ENSG00000172464	OR5AP2	Yes	No	360	15
ENSG00000131398	KCNC3	Yes	Yes	1560	12
ENSG00000165389	C14orf147	Yes	No	120	23
ENSG00000079557	AFM	Yes	No	1080	27
ENSG00000126259	KIRREL2	Yes	Yes	1080	31
ENSG00000175920	DOK7	Yes	Yes	1200	28
ENSG00000144712	CAND2	Yes	Yes	1920	28
ENSG00000167077	MEI1	Yes	Yes	2040	35
ENSG00000132768	DPH2	Yes	No	600	31
ENSG00000188582	PAQR9	Yes	No	480	24
ENSG00000105254	TBCB	Yes	Yes	360	34
ENSG00000160505	NLRP4	Yes	Yes	1440	19
ENSG00000105518	TMEM205	Yes	Yes	240	26
ENSG00000133401	PDZD2	Yes	Yes	5400	35
ENSG00000196943	C14orf21	Yes	Yes	840	35
ENSG00000130827	PLXNA3	Yes	No	2400	32

B.3 Paper

Discerning the difference between disease-associated variants and variants observed in healthy individuals is a key step in predicting whether new variants could be harmful. In this paper we compared disease-associated variants from OMIM (Amberger et al. 2009) with variants found in the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010), looking both at the overall distribution of variants, and their placement on proteins. I worked on producing a model of evolution in humans using the 1000 Genomes data, and comparing it with other models.

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset

Tjaart A. P. de Beer*, Roman A. Laskowski, Sarah L. Parks, Botond Sipos, Nick Goldman, Janet M. Thornton

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genomes Campus, Cambridge, Cambridgeshire, United Kingdom

Abstract

The 1000 Genomes Project data provides a natural background dataset for amino acid germline mutations in humans. Since the direction of mutation is known, the amino acid exchange matrix generated from the observed nucleotide variants is asymmetric and the mutabilities of the different amino acids are very different. These differences predominantly reflect preferences for nucleotide mutations in the DNA (especially the high mutation rate of the CpG dinucleotide, which makes arginine mutability very much higher than other amino acids) rather than selection imposed by protein structure constraints, although there is evidence for the latter as well. The variants occur predominantly on the surface of proteins (82%), with a slight preference for sites which are more exposed and less well conserved than random. Mutations to functional residues occur about half as often as expected by chance. The disease-associated amino acid variant distributions in OMIM are radically different from those expected on the basis of the 1000 Genomes dataset. The disease-associated variants preferentially occur in more conserved sites, compared to 1000 Genomes mutations. Many of the amino acid exchange profiles appear to exhibit an anti-correlation, with common exchanges in one dataset being rare in the other. Disease-associated variants exhibit more extreme differences in amino acid size and hydrophobicity. More modelling of the mutational processes at the nucleotide level is needed, but these observations should contribute to an improved prediction of the effects of specific variants in humans.

Citation: de Beer TAP, Laskowski RA, Parks SL, Sipos B, Goldman N, et al. (2013) Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset. PLoS Comput Biol 9(12): e1003382. doi:10.1371/journal.pcbi.1003382

Editor: Yana Bromberg, Rutgers University, United States of America

Received April 29, 2013; **Accepted** October 22, 2013; **Published** December 12, 2013

Funding: This work was supported in part by the National Institutes of Health grant GM094585, by the U. S. Department of Energy, Office of Biological and Environmental Research, under contract DE-AC02-06CH11357 (Midwest Center for Structural Genomics) as well as EMBL-EBI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: tjaart@ebi.ac.uk

Introduction

With the release of the 1000 Genomes Project (1 kG) data [1], it has become feasible to study human protein variation on a large scale. The main aim of the 1 kG project was to discover and characterize at least 95% of human DNA variants (with a frequency of occurrence of >1%) found in multiple human populations across the world. Five main populations were sampled with ancestry in Europe, West Africa, the Americas, East Asia and South Asia. The project has provided a rich set of synonymous (sSNPs) and non-synonymous (nsSNPs) variants for 1092 individuals from diverse populations. It is estimated from the 1 kG data that each individual will, on average, differ from the reference human genome sequence at 10,000–12,000 synonymous sites in addition to 10,000–11,000 non-synonymous sites [1]. As these nsSNPs change the amino acid sequence of the protein, the changes have the potential to affect the structure and function of the corresponding proteins. The 1000 Genomes Project data set is valuable in that it is large and not derived from a disease cohort but rather seeks to capture variants found in a disparate set of healthy individuals. This can be used to characterise differences on average between disease-associated and benign mutations (or at least mutations not known to be associated with disease) as well as exploring their structural characteristics and preferences. The reports from the 1000 Genomes Consortium [1,2] have focused on genome and nucleotide variation, and other papers consider mutations in association with a specific disease (e.g. cancer) [3].

Various databases such as the Online database of Mendelian Inheritance in Man (OMIM, [4]), the UniProtKB human polymorphism set (Humsavar, [5]) and the Human Gene Mutation Database (HGMD, [6]) collect information on inherited diseases associated with variants. The Humsavar database contains disease-associated variants from the literature and OMIM. OMIM currently contains information on approximately 10,200 nsSNPs associated with diseases (December 2011) and Humsavar about 23,500 disease-associated nsSNPs. Most of the phenotypical effects and their molecular origins are not well established, so predicting the functional effect of a single amino acid variant is of great medical interest. The main methods assume that mutations in highly conserved residues cause disease and thus, by using alignments to homologous sequences and residue similarity, the severity of the variant can be gauged. More advanced methods include information derived from protein structures (such as solvent accessibility, free energy changes, environment specific substitution tables and functional annotations) to improve the accuracy (see review by [7]). The advantage

Author Summary

In this paper we compare the differences between 'natural' and disease-associated amino acid variants at both sequence as well as structural levels. We used data from the 1000 Genomes Project (1 kG), the OMIM database and UniProtKB Humsavar. The results highlight the complex interplay of features from the level of the DNA up to protein sequence and structure. The codon CpG dinucleotide content plays a large role in determining which amino acids mutate. This in turn affects the mutability of amino acids and a clear difference was seen between non-disease and disease variants where amino acids that are naturally very mutable show the opposite trend in the disease-associated data. The current results show evidence for some selection, mainly in that the variants occur slightly more often on the surface of the protein and are much less likely to be annotated as functional than expected by chance. However we should note that even the best definition of functional, taken from structural data, is limited. Even with these caveats, it is clear that the 1 kG variants eschew functional residues as defined here, a trend which is surprisingly even stronger in the OMIM data.

of using a 3D approach for prediction is that the consequence and characteristics of the variant can be studied in its specific environment in the protein. This provides a level of information beyond a sequence or a sequence alignment [8]. If there are ligands present, the interaction between the mutated amino acid and the ligand can be studied. This has been successfully applied to various individual proteins on a case-by-case basis [9,10]. In total over 30 different programs to predict the effects of these variants have been published, including Condel [11], SNAP [12], SDM [13], PolyPhen [14], VEP [15], SIFT [16,17] and SNP&GO [18]. Most of these algorithms can only predict whether a specific variant will be neutral or deleterious for the protein with various degrees of accuracy, although measuring accuracy is challenging in the absence of a good benchmark.

To allow the accurate prediction of functional effects of SNPs, we need a thorough understanding of why amino acids mutate in humans. Various groups have worked on the effect of the mutations and numerous studies have been done on small specific sets of proteins [8,19–22]. Blundell and co-workers have found that the local environment around an amino acid plays a large role in the effect that selection has on a mutation in a specific position [21]. This has led to the development of environment specific substitution matrices [23,24] that incorporate structural constraints. Subramanian and Kumar [25] did a detailed analysis on a set of 8,627 disease-associated mutations and found that disease-associated mutations tend to occur on inter-species conserved residues. The common factor between these studies is that they try to understand the effect that selection and structural constraints have on disease vs non-disease states in selected sets of proteins. Very few studies have tried to unravel the underlying cause for mutation patterns seen in human proteins. With this work we aim to elucidate why certain amino acids mutate more and try to understand the underlying mechanisms present in the mutation process. We gather the data for all the amino acid mutations found in the 1000 Genomes Project to characterise their sequence and structural properties, providing a benchmark background against which to compare the disease-associated nsSNPs in OMIM and Humsavar.

Results

The 1000 Genomes Project data were queried to retrieve all the nsSNPs, which were filtered to include only those that occurred in a single population (see methods). This ensures that only the more recent mutational events in human evolution are included and simplifies counting. In addition variants at a single site were only counted once even if they occur in multiple individuals, since such clusters are assumed to represent a single variation event that has been inherited in the other individuals. For 3D analysis only human proteins, for which complete structures are available, were included to ensure accurate analysis of 3D features. For solvent accessibility calculations, a monomer subset was also generated to avoid problems with uncertain multimeric states and validate our findings on the larger dataset. Homology models based on close relatives were used to extend the data set and see if the trends observed in the experimental structures were preserved. Table 1 summarizes the five data sets created and used in this study.

The amino acid exchange matrix derived from the 1000 Genomes Project dataset

Figure 1 shows the amino acid exchange matrix generated from the ~106,000 nsSNPs found in the 1 kG data. Amino acid mutations requiring two or three base changes are not defined in this dataset due to technical reasons. The 1 kG matrix exhibits several interesting features, most of which reflect the genetic code and the differential mutability of various codons. All possible single base changes are observed. The matrix is not symmetrical as a result of the differences in frequency of occurrence of amino acids as well as differences in their mutabilities [26,27]. As expected

Table 1. The different datasets constructed and used in this study and their composition.

Data set	Protein chains	nsSNPs	Description
1 kG	19,058	106,311	A data set containing all the 1 kG variants filtered by population.
OMIM	19,058	10,151	A protein sequence based set containing OMIM variants for all reviewed UniProt human proteins.
Humsavar	19,058	23,846	A set based on human disease polymorphisms from UniProt.
3D	2,139	10,628	A protein 3D structure based set consisting of 1 kG variants for proteins that have a complete structure in the PDB.
Monomer	325	1,461	A subset of the 3D set containing only proteins identified as being monomeric.
Model	2,630	13,037	A set based on human ModBase homology models where sequence coverage and identity are between 90–100%.

doi:10.1371/journal.pcbi.1003382.t001

	To																					
From	R	K	D	E	N	Q	S	G	H	T	A	P	Y	V	M	C	L	F	I	W	Tot.	Mut.
R		738 39	0 0	0 2	0 0	4630 81	553 59	761 372	4125 83	247 30	0 0	216 71	0 2	0 0	71 39	3420 125	467 112	0 0	137 7	2695 82	18060 1104	0.0308 0.0019
K	1151 23		0 1	843 303	689 75	346 27	0 1	0 0	0 0	300 17	0 2	0 0	0 0	0 0	87 29	0 0	0 0	0 1	84 12	0 1	3500 492	0.0058 0.0008
D	0 0	0 1		855 45	1994 52	0 0	0 0	780 206	419 26	0 1	153 66	0 0	378 30	379 34	0 0	0 0	0 0	0 0	0 1	0 0	4958 462	0.0100 0.0009
E	0 6	2682 93	983 32		0 0	631 28	0 0	628 132	0 1	0 1	319 48	0 0	0 0	241 35	0 0	0 0	0 0	0 0	0 0	0 0	5484 376	0.0074 0.0005
N	0 1	546 66	516 167	0 1		0 0	2030 37	0 1	236 17	170 17	0 1	0 0	133 18	0 1	0 2	0 0	0 0	0 0	158 32	0 0	3789 361	0.0101 0.0010
Q	875 363	351 28	0 0	521 39	0 0		0 0	0 4	764 37	0 0	0 0	243 29	0 0	0 1	0 0	0 0	192 37	0 0	0 0	0 0	2946 538	0.0059 0.0011
S	648 74	0 1	0 0	0 1	1103 110	0 1		659 190	0 0	637 13	273 47	541 101	284 18	0 0	0 0	673 57	1386 27	816 68	242 17	44 22	7306 747	0.0084 0.0009
G	1957 107	0 1	852 101	804 61	0 0	0 0	1598 24		0 0	0 0	469 22	0 0	0 0	630 48	0 1	197 32	0 1	0 0	0 0	84 21	6591 419	0.0097 0.0006
H	863 378	0 0	130 57	0 1	166 17	486 35	0 0	0 0		0 0	0 0	145 23	869 51	0 0	0 0	0 0	188 20	0 0	0 0	0 0	2847 582	0.0104 0.0021
T	206 24	200 21	0 0	0 0	389 17	0 1	740 17	0 0	0 0		1420 195	186 40	0 0	0 0	2272 79	0 1	0 2	0 0	1781 105	0 0	7194 502	0.0130 0.0009
A	0 4	0 1	309 25	205 28	0 0	0 4	892 6	509 40	0 1	3529 51		357 33	0 0	3224 53	0 0	0 3	0 0	0 1	0 0	0 0	9025 250	0.0124 0.0003
P	502 113	0 1	0 0	0 1	0 0	198 42	2086 91	0 0	261 29	650 37	666 90		0 0	0 0	0 0	0 0	3309 291	0 0	0 0	0 1	7672 696	0.0118 0.0011
Y	0 0	0 0	62 65	0 0	81 18	0 1	105 19	0 2	490 59	0 0	0 1	0 0		0 0	0 1	1125 124	0 1	198 8	0 0	0 0	2061 299	0.0074 0.0011
V	0 3	0 1	95 63	98 38	0 0	0 0	0 1	191 134	0 0	0 1	932 168	0 0	0 1		2319 74	0 1	989 50	286 20	3057 42	0 0	7967 597	0.0129 0.0010
M	104 5	107 18	0 0	0 0	0 0	0 0	0 0	0 0	0 0	705 119	0 0	0 0	0 0	1050 156		0 0	335 20	0 0	826 29	0 0	3127 347	0.0139 0.0015
C	350 396	0 0	0 0	0 0	0 0	0 0	260 39	116 65	0 0	0 0	0 0	0 0	594 139	0 1	0 0		0 0	169 35	0 0	82 39	1571 714	0.0066 0.0030
L	227 90	0 1	0 0	0 0	0 0	145 10	279 93	0 0	83 21	0 0	0 0	692 225	0 0	1078 60	304 20	0 0		1366 92	387 13	95 6	4656 631	0.0045 0.0006
F	0 0	0 0	0 0	0 0	0 0	0 0	291 90	0 0	0 0	0 0	0 1	0 0	124 6	159 44	0 0	144 38	880 80		117 16	0 0	1715 275	0.0045 0.0007
I	35 4	30 11	0 0	0 0	180 19	0 0	131 18	0 0	0 0	1468 108	0 2	0 0	0 2	2189 55	690 62	0 0	337 8	242 15		0 1	5302 305	0.0117 0.0007
W	225 384	0 0	0 0	0 0	0 0	0 2	38 8	51 22	0 0	0 0	0 0	0 0	0 0	0 0	0 0	169 27	57 11	0 0	0 0		540 454	0.0043 0.0036
1kG OMIM	7143 1975	4654 283	2947 511	3326 520	4602 308	6436 232	9003 503	3695 1168	6378 274	7706 395	4232 643	2380 522	2382 267	8950 488	5743 307	5728 408	8140 660	3077 240	6789 274	3000 173	106311 10151	

Figure 1. The amino acid exchanges observed in human protein variants. The 1 kG data set is the top row of each cell and OMIM the bottom row of each cell*. Amino acids are arranged by 1 letter code according to increasing hydrophobicity (least hydrophobic is left and most hydrophobic is right) using the Fauchère and Pliska scale [58]. Yellow blocks indicate mutations where there are statistically significant differences between 1 kG and OMIM. Blue blocks indicate where no mutations were present in the 1 kG data set. White blocks show where there are no statistically significant differences. Green blocks show where there are proportionally more 1 kG mutations compared to OMIM. Orange blocks show where there are proportionally more OMIM mutations than 1 kG. The mutability scores (see methods) for the 1 kG and OMIM sets are shown in the last column. *Note that these matrices are fundamentally different. The 1 kG data set gathers all the observed mutations in the 1 kG project, counting each only once; the OMIM data set combines information gathered from potentially many individuals but filtered to identify those mutations associated with a disease.
doi:10.1371/journal.pcbi.1003382.g001

there is a strong correlation ($r = 0.786$) between the frequency of occurrence of amino acids in the human proteome and the number of associated codons. Figure 2 shows that, excluding Arg and Leu which are extreme outliers, there is a strong trend for amino acids with a higher frequency of occurrence to have more mutations ($r = 0.836$). Taken together this leads to a relatively strong correlation ($r = 0.741$) between the number of codons and the number of mutations. In contrast, the frequency of the gained amino acids, resulting from the mutation, shows little correlation between frequency of occurrence and number of mutations ($r = 0.349$).

Amino acid mutabilities

The mutabilities of the amino acids (see methods) in the 1 kG dataset are shown in the last column of Figure 1. Arg (0.031) is the most mutable, whilst the more chemically complex amino acids, Trp (0.004) and Phe (0.005) have the lowest mutabilities. There is no correlation in the 1000 Genomes data between mutability and frequency of occurrence ($r = -0.003$ excluding Arg) nor between mutability and the number of codons (Figure 3). It is well known that CpG dinucleotides in DNA tend to mutate at rates 10–50 times higher than other dinucleotides [28,29] and thus amino acids with a CpG present in their codons will mutate with a higher probability (see Figure 4). Four out of the six codons for Arg include CpG sequences, and Arg mutates more frequently than any other residue, with a mutability (0.031) which is over twice as high as its nearest rival. This high mutability also reflects the fact that the CpG in the Arg codons occur in the non-wobble positions so nucleotide mutations give rise to non-synonymous SNPs. In contrast Leu which also has six codons, none of which contain

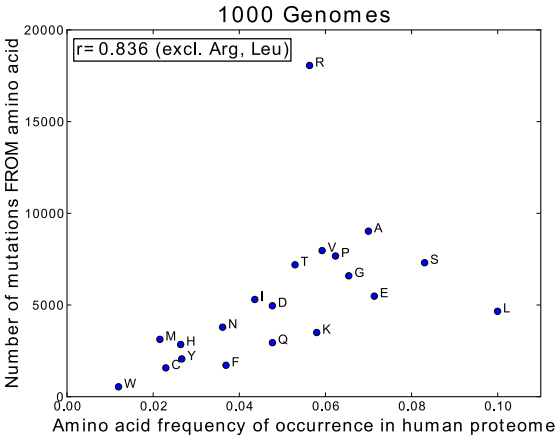


Figure 2. Comparison of the number of mutating residues vs the amino acid frequency of occurrence.
doi:10.1371/journal.pcbi.1003382.g002

CpG, has a low mutability (0.005) and mutates six times less frequently than Arg. However the correlation with CpG is far from perfect and other factors must have an effect. For example, Met, which has only one codon with no CpG dinucleotide, is the second most mutable amino acid (0.014).

Figure 4 shows the clear pattern of amino acid gain and loss in the human proteome. Jordan [26] and Zuckerkandl [30] long since identified that Cys, Met, His, Ser and Phe are being accrued significantly in the human proteome. Our data confirm a net gain of these five amino acids, and Val, Asn, Ile and Thr were also confirmed as weak gainers. Jordan and co-workers also identified strong losers and our data again confirm that Pro, Ala, Gly and Glu are strong losers. Lys was identified as a weak loser but our larger dataset suggests that lysine should be considered a weak gainer in humans. Arg is the strongest loser in the human genome (similar to the human set in [26] but not other considered species).

We calculated the mutability for every amino acid on a population specific basis. None of the populations showed a

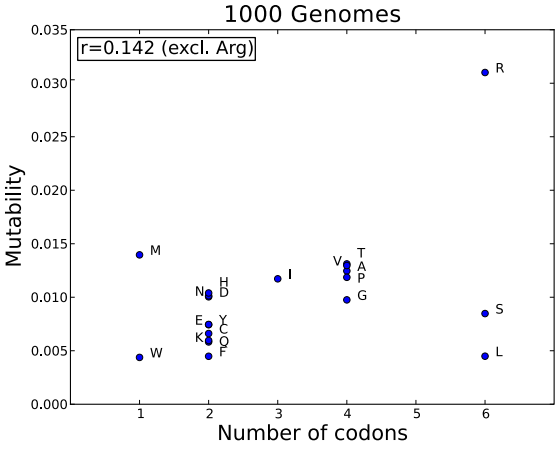


Figure 3. Amino acid mutability vs the number of codons in the 1 kG data.
doi:10.1371/journal.pcbi.1003382.g003

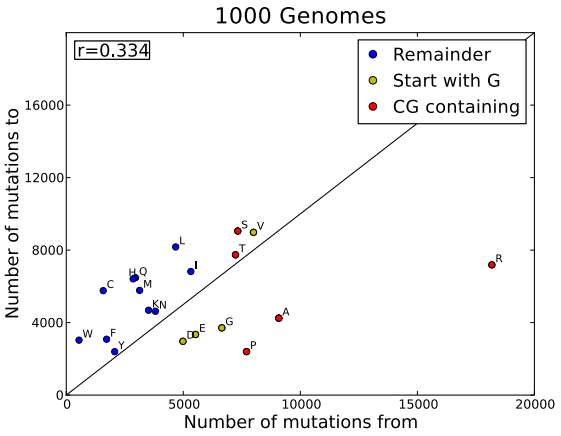


Figure 4. A visual representation of the asymmetry of the 1 kG data. The plot shows the difference between how often an amino acid mutates vs how often it is mutated to. These are raw counts and also reflect the frequency of occurrence. Each amino acid is coloured according to CpG content. Red: a CpG dinucleotide occurs in its codons; yellow: if one of its codons start with a G (with a C possibly preceding it); blue: no CpG in its codons. The black line indicates the diagonal where 'mutations to' equals 'mutations from'.
doi:10.1371/journal.pcbi.1003382.g004

different pattern of amino acid mutabilities, compared to the overall trend with correlation coefficients equal to 1.0 (Figure S1). Using the individual amino acid mutabilites, we looked at aggregate protein mutability differences by adding up the individual mutabilities for every amino acid in each protein in the data set and normalising by protein length. This was compared to the aggregate mutabilities of proteins involved in disease as classified by OMIM and Humsavar. The average score for disease-associated proteins was 0.0103 and for non-disease proteins 0.0102 with a median of 0.01022 ($\sigma = 0.0006$) and 0.01018 ($\sigma = 0.0005$), respectively, indicating that protein aggregate mutability has no bearing on disease-association (Figure S2).

The effects of physicochemical characteristics of the amino acids on their mutability

As well as constraints on the mutational process at the DNA level, the consequence of a variant on the protein structure and function will also have an impact on the number of observed mutations. If a variant interferes with the structure and function of a protein and that protein is essential, then this variant is less likely to be seen. However comparison of mutability with the size and hydrophobicity of the amino acid shows very little correlation in the 1 kG dataset. There is a moderate anti-correlation between higher mutability and size ($r = -0.474$), with the smaller amino acids mutating more frequently, but no correlation at all between mutability and hydrophobicity ($r = -0.082$) although the large hydrophobic amino acids (Leu, Phe and Trp) have the lowest mutability scores. Trp has the fewest mutations (544, even though all SNPs in Trp codons result in a change of amino acid) and also the lowest mutability score (0.004) together with Phe. In addition to their complexity and low abundance, Phe and Trp often occur in specialized roles such as the interior of proteins, π-π stacking or ring interactions and this might add to their low mutability. The mutability of Cys is also low, perhaps reflecting its role in disulphide bridges, which help to stabilise extracellular proteins.

The structural properties of 1000 Genomes variants

To investigate the structural characteristics of these variants, three sets of protein structures were compiled, namely the 3D set, the monomer set and the model set (Table 1). The 3D and monomer set were constructed from data in the PDB (see methods) while the model set and the subsequent variant modelling was created and performed using Modbase [31] and Modeller [32], built into an in-house homology modelling pipeline. The 3D set contains 2,139 protein chains. A total of 10,628 1 kG nsSNPs were found in these chains, of which protein models, based on the known structures of human proteins could be built for 5,524. The monomer set contains 325 protein chains identified as monomers and a total of 1,461 1 kG nsSNPs were found, of which 897 could be modelled. The model set, including models based on homologues from the PDB, contained 2,630 protein chains and 12,432 out of 13,037 nsSNPs could be modelled. For the Humsavar set we found 5,592 nsSNPs of which 3,942 could be modelled.

Figure 5A shows a comparison of the solvent accessibility distribution for all residues compared to that for the variants. On average the variants in the 1 kG are slightly more exposed. An analysis of the solvent exposed residues found that, for the most accurate monomer set, 79% of nsSNPs are solvent exposed compared to 73% of all residues ($p = 0.001$). For the structures in the model set, 81.9% of nsSNPs were solvent exposed. For all three datasets, the 1 kG variants have a slight preference to occur on the surface of proteins compared to all residues. Figure 5B shows that there were no appreciable differences in secondary structure preferences between variants and other residues.

Do natural mutations occur in functionally annotated residues?

Functional annotation for each human protein was derived using SAS (Sequence Annotated by Structure, [33]). Table 2 shows the different functional annotations for each set. The vast majority of functional annotations identified, make contacts to ligands (using PDBsum data, [34]) or site interactions in the proteins (as defined in the PDB). Only 15.5% of the mutations (1,648 of 10,628) in the 3D set were annotated with a function compared to 29.1% of all residues in the set of human structures (Figure 5C). These data show that the observed mutations in the 1000 Genomes occur less frequently in the functionally annotated residues compared to all residues.

Residue conservation

Residue conservation scores, defined as the variation of the residues at a given site in the protein across multiple species, were obtained for all sites in the human proteome (where sufficient data are available) from the Evolutionary Trace server [35]. These scores are distributed across the whole range of conservation (Figure 6) with a mean score of 0.48. The scores for all the sites with mutations in the 1000 Genomes data show a slightly different distribution from all residues, with a small but significant shift ($p<2.2\times10^{-16}$) towards the less conserved sites and a reduced mean conservation score of 0.43. Clearly natural variation occurs across all conservation levels and is not limited to non-conserved residues.

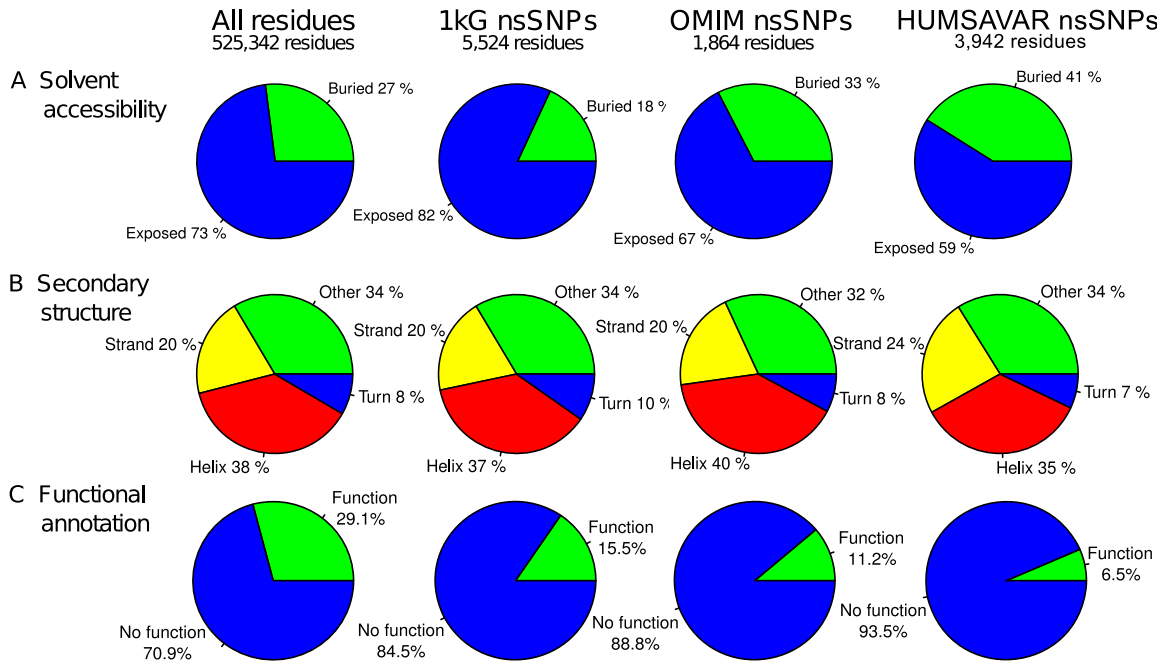


Figure 5. Site properties for all residues, 1 kG nsSNPs, OMIM nsSNPs and Humsavar nsSNPs in the structure 3D set. (A) the solvent accessibility for the variants in the four datasets, (B) the secondary structure in which each of the variants occurs, (C) the functional annotation of every variant in the four datasets.
doi:10.1371/journal.pcbi.1003382.g005

Table 2. The various functions assigned to nsSNPs in each set.

Set	Site	Ligand	Site/ligand overlap	Metal	Catalytic	Overall (non-redundant)
3D	1,414	1,432	1,220	334	17	1,648 (15.5%)
Monomer	281	273	245	83	4	312 (21.4%)
OMIM	163	184	147	17	17	209 (2.1%)
Humsavar	305	285	252	58	41	355 (51.2%)
Models	1,538	1,443	1,304	376	36	1,676 (12.9%)

'Site' refers to residue specific annotations made by depositors of PDB structures, 'Ligand' refers to residues involved in binding a ligand, 'Metal' refers to residues coordinating with metals and 'Catalytic' to residues involved in the catalytic activity of the protein. The % of non-redundant assigned residues that are 'functional' is also shown.

doi:10.1371/journal.pcbi.1003382.t002

Amino acid exchange characteristics in 1000 Genome data

For each amino acid the mutation profile can be calculated showing the preference for specific X = >Y mutations in the 1000 Genomes data. These profiles, given for all the amino acids in Figure 7, show that there are striking differences in frequency of occurrence for the different exchanges. For example, in the 1 kG set Arg shows a strong preference to mutate to Gln and His, whilst mutations to Ser, Gly and Pro are much less frequent. All the amino acids show these differential exchange rates. Figure 8A shows the distribution of changes in energy of the whole protein caused by each mutation, evaluated as the statistical potential energy DOPE score (Discrete Optimised Protein Energy) in Modeller. 68.1% of the 1 kG variants increase the DOPE score (i.e. make the protein less stable). This implies that most natural variants decrease the stability of the protein, albeit by a very small amount. The distribution of changes in size and hydrophobicity for all observed mutations (Figure 8B and 8C) show that 59.4% of mutations increase the hydrophobicity of the amino acid and 52.4% of mutations increase the size. Over 84% of variants

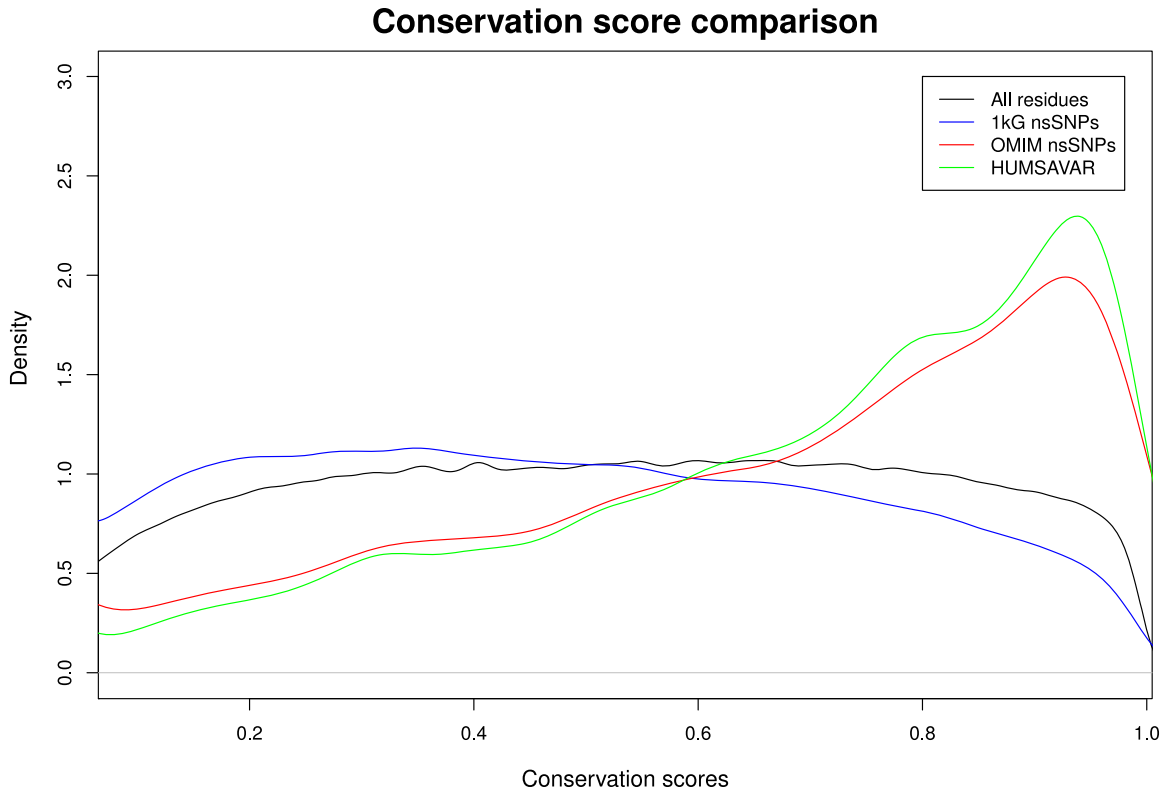


Figure 6. Comparison of the conservation scores in the four sets used. The density distribution of residue conservation scores for all the amino acid positions in UniProt (9,532,474 residues, black), 1 kG (185,428 residues, blue), OMIM (8,099 residues, red) and Humsavar (21,446 residues, green). The conservation scores range from 0 for non-conserved residues to 1 for highly conserved residues.

doi:10.1371/journal.pcbi.1003382.g006

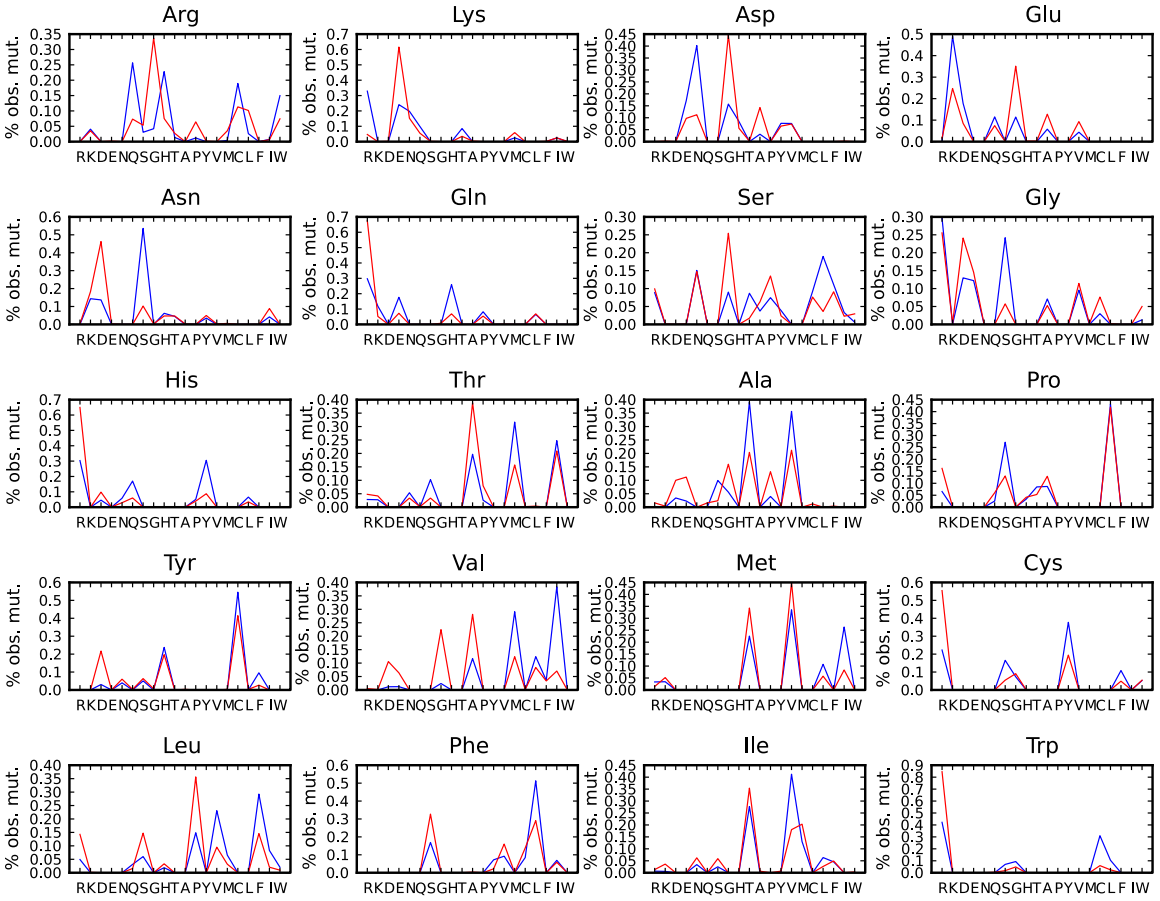


Figure 7. Comparison of the differences in observed mutations in the various sets. Comparison of the differences in the % of observed mutations in the 1 kG (blue) and OMIM (red) sets for one amino acid mutating to all others e.g. proportionally, more mutations from Lys to Glu are recorded in OMIM than in the 1 kG set. Each plot shows the results of mutation from a specific amino acid (e.g. Arg at top left) to every other amino acid.
doi:10.1371/journal.pcbi.1003382.g007

change their size by less than 50 Da. 72% of variants change their hydrophobicity by less than 1 unit. Extreme changes are rare. At this stage these observations provide empirical expectation rates for amino acid exchanges in humans and result from the genetic code, the nucleotide exchange rates and also some selection at the protein level. However without a good random model it is difficult to be confident about the importance of the different contributions to such variation.

Comparison of 1000 Genome variants with those predicted by the PAM and WAG mutation matrices

The 1 kG counts matrix is a snapshot of mutations that have occurred in humans in a short period of time. To understand this process the count matrix can be converted into an instantaneous rate matrix describing the rates of change of each amino acid in humans in a time-independent manner [36]. Instantaneous rate matrices have previously been built from a wide selection of protein alignments across many species including nuclear proteins, mitochondrial proteins, chloroplast proteins, buried protein domains and exposed protein domains. PCA can be used to compare these inter-species matrices with the 1 kG intra-species matrix (Figure 9A–C). The 1 kG matrix was built using data where the direction of the mutations is known whereas all other matrices were calculated assuming direction is unknown. This was compared to the WAG [37] and PAM matrix [38]. To check that any differences between the 1 kG matrix and the other matrices are not caused by using direction, a directionless matrix has also been included in the plot (Figure 9D). In this plot, principal component one clearly separates the 1 kG matrices, which are placed very close together, from all of the previously calculated matrices. Principal component two then spreads matrices out based on whether the alignments used to build them are made up mainly of exposed or buried domains, with the mitochondrial matrices at the one extreme built from nearly all membrane proteins, and matrices built from only exposed regions of proteins at the other.

A difference between the intra-species data and the inter-species matrices is the amount of selection which has occurred. Due to the time-scale for the 1 kG data and the relatively weak selection in human populations [39,40] the only mutations which are not observed are lethal mutations. This means that there should be a

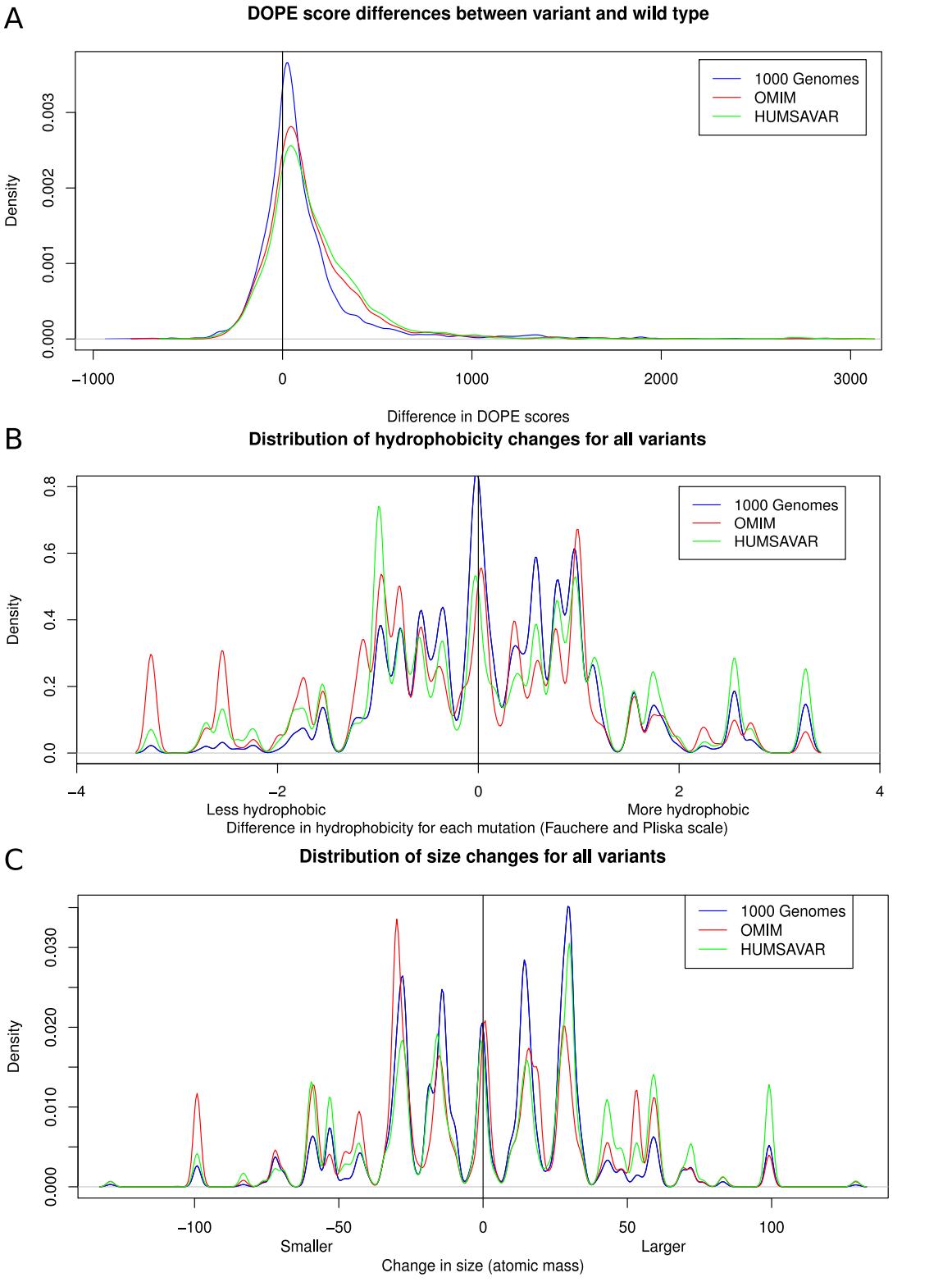

A
DOPE score differences between variant and wild type
1000 Genomes
OMIM
HUMSAVAR
Density
0.000
0.001
0.002
0.003
−1000
0
1000
2000
3000
Difference in DOPE scores
B
Distribution of hydrophobicity changes for all variants
1000 Genomes
OMIM
HUMSAVAR
Density
0.0
0.2
0.4
0.6
0.8
−4
−2
0
2
4
Less hydrophobic
More hydrophobic
Difference in hydrophobicity for each mutation (Fauchere and Pliska scale)
C
Distribution of size changes for all variants
1000 Genomes
OMIM
HUMSAVAR
Density
0.000
0.010
0.020
0.030
−100
−50
0
50
100
Smaller
Larger
Change in size (atomic mass)

Figure 8. Comparison between the physicochemical properties of the wildtype and the mutant models for each of the data sets. Plots showing the differences between (A) Modeller DOPE scores for the wild type and mutant model (based on 3D, 10,628 mutations, and Humsavar sets, 21,446 residues), (B) changes in hydrophobicity between wild type and mutant in both sets and (C) changes in size between wild type and mutation in both sets.
doi:10.1371/journal.pcbi.1003382.g008

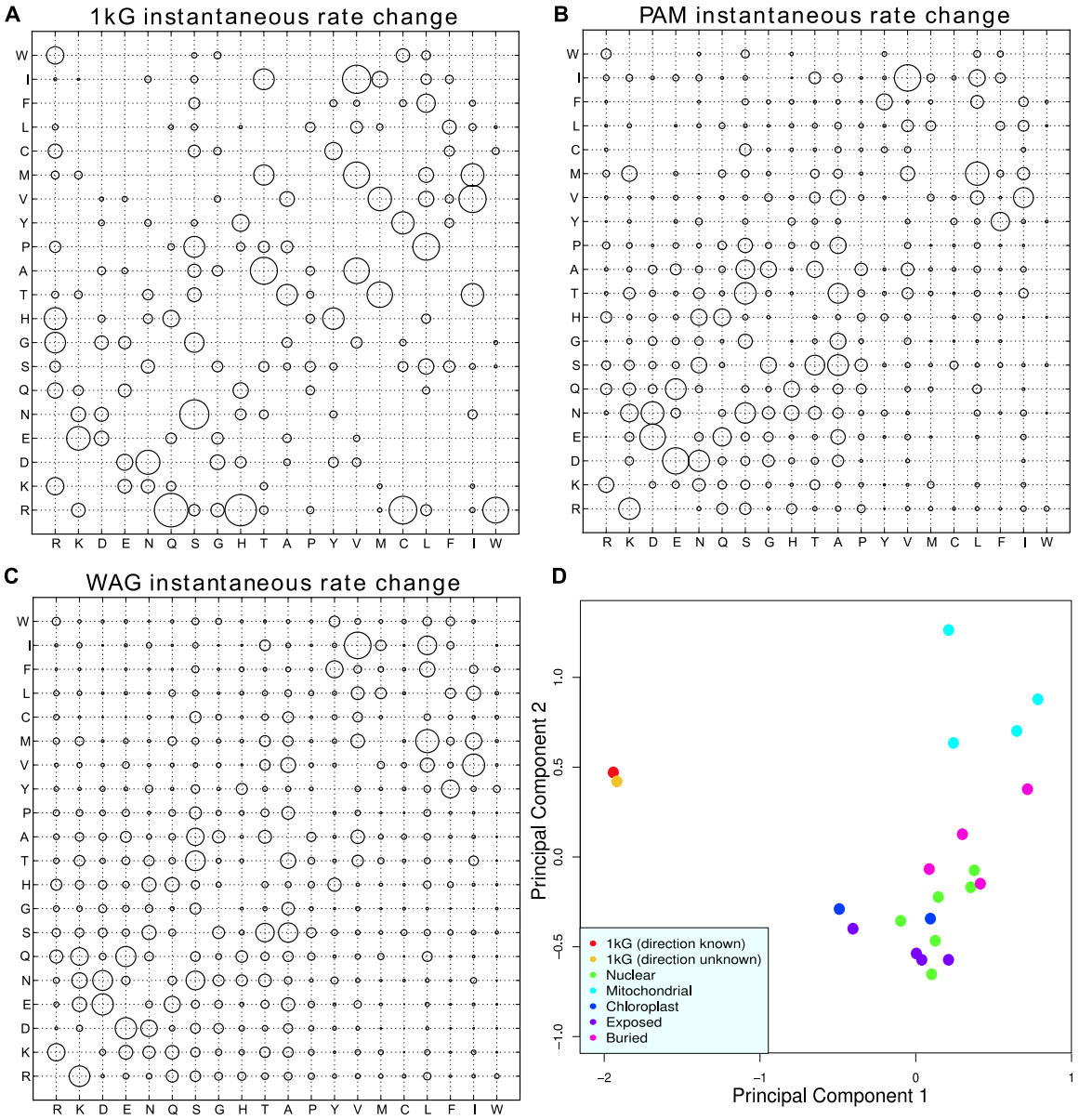


Figure 9. Bubble plots comparing the relative differences between the instantaneous rate change matrices of the data sets. (A) 1 kG data, (B) PAM matrix and (C) WAG matrix. (D) A PCA (first two components) plot showing the separation of the 1 kG matrices from other matrices. Matrices included are 1 kG (with and without assuming direction), nuclear (WAG, JTT, LG, PAM, tm126, PCMA), mitochondrial (mtREV24, mtMam, mtArt, mtZoa), chloroplast (cpREV, cpREV64), exposed (alpha helix, beta sheet, coil, turn) and buried (alpha helix, beta sheet, coil, turn). Principal components one and two represent 34% and 20% of the variance, respectively. All other principal components represent 9% or less of the variance each. Amino acids are arranged according to increasing hydrophobicity.
doi:10.1371/journal.pcbi.1003382.g009

limited effect of selection on the 1 kG matrix. By using no allele frequency cutoff for the minor alleles when building the count matrix, we gather the maximum amount of information about the mutation process. The counts are necessarily shaped by mutation and selection but will mostly reflect the mutation process. The inter-species matrices (e.g. PAM and WAG in Figure 9B,C) on the other hand are subject to selection pressures. This could explain why the 1 kG matrix is so different from the other matrices. One clear factor is CpG hypermutability: for example, changes from Arg, an amino acid with four of six codons containing a CpG, have a very high rate in the 1 kG data, and not in WAG (Figure 9A,B). In fact only codons containing a CpG have high rates overall (Figure 10). The most plausible explanation is that these CpG mutations are occurring at a very high rate and then are selected out so that the effect is not seen as strongly when looking across multiple species.

Comparison between the 1000 Genomes variants and the disease-associated variants

For comparison, we have constructed the amino acid exchange counts matrix for data from the OMIM database and the associated plots for these mutations (Figures 1–8). Disease variants from the UniProtKB/Swiss-Prot Human polymorphisms and disease mutations index (Humsavar) were also included with plots available in the supplement (Figures S3, S4, S5). Our focus however is on the OMIM set. In contrast to the 1 kG data, various double and triple base mutations are observed in the OMIM set, however the three triple base changes (Phe-Lys, Met-Tyr and Trp-Ile) were checked back to the publications and all were found to be errors either in the paper or in OMIM and were removed. 82 two base changes were found in OMIM and a few (10%) randomly selected changes were manually checked with no errors found. Clearly the OMIM data are radically different from the 1000 Genome data, in that they are all independent observations of variable confidence and manually determined by individual scientists. They only represent a small fraction of disease-associated nsSNPs and the number of mutations

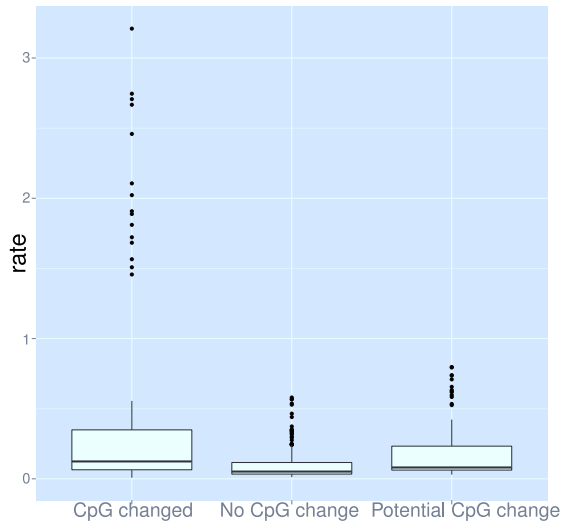


Figure 10. Dependence of mutation rates on the change in CpG status. Rates of change from codons were calculated similarly to the amino acid rate matrix [36], but on a 61 by 61 codon matrix.
doi:10.1371/journal.pcbi.1003382.g010

(~10,000), is approximately ten times smaller than the number of 1000 Genomes mutations. The normalised OMIM counts that differ from the 1 kG dataset are coloured in Figure 1. Considering just the residue type, if we exclude Arg, the overall correlation between the normalised frequencies of occurrence of the mutated residues in the two datasets is only 0.14 and between 1 kG and Humsavar it is 0.48. If we compare all 148 observed X = >Y frequencies, the correlation between 1 kG and OMIM is 0.51 and 1 kG and Humsavar is 0.79.

Previous studies have found that mutations from Arg and Gly are the major contributors to human genetic disease and have been shown to make up about 30% of the mutations involved in disease [41]. In this updated and much expanded set, variants from Arg and Gly only make up 15% of the disease causing mutations. However mutations to Arg are still the biggest contributor to genetic disease with ~19.4% of all mutations.

Figure 11 shows a rank order comparison between the frequency of occurrence of the 1 kG and OMIM variants ($r = 0.09$) as well as between 1 kG and Humsavar ($r = 0.31$) and Humsavar and OMIM ($r = 0.51$), normalised for amino acid occurrence. Unlike for the 1 kG data, the disease-associated variants show moderate inverse correlations between their frequency and the frequency of occurrence of the residue type ($r = -0.67$) implying that, at least for OMIM, the mutations to the rarer amino acids (with fewer codons) are more likely to be associated with disease. As with the 1 kG data there is no strong correlation between a residue type being associated with a disease in the OMIM data and the number of codons. For hydrophobicity and size, the disease associated variants show the opposite trend to the 1 kG dataset with a moderate correlation between lower frequency and smaller size ($r = 0.528$, excluding Cys and Trp) but no correlation between frequency and hydrophobicity ($r = 0.289$). It is interesting to note that the least mutable amino acid in the 1 kG data (Trp) turns out to be the residue whose mutation is most likely to result in disease in the OMIM variants and is highly ranked in the Humsavar set. Trp, the largest amino acid, often occurs in specialized roles in proteins as does Cys, the second most frequent variant residue type in OMIM. Amino acids with a lower frequency of occurrence tend to be the more complex amino acids and are frequently found in specialized roles. Mutating them will result in the possible loss or alteration of protein function, hence the over-representation in OMIM and Humsavar. In a number of cases the OMIM and 1 kG variant preferences appear to behave in an opposite way from one another e.g. in Figure 7 Arg most frequently mutates to Gln in the 1000 Genomes and a variantion to Gly is much less common, whilst Arg to Gly is the most common variant in the OMIM dataset and a variation to Gln is rare.

We observe a reasonable correlation between the OMIM and Humsavar mutabilities ($r = 0.51$), but some amino acids appear to behave completely differently in the two datasets. Gly and Ala are much more frequently mutated in the Humsavar set than in OMIM, whilst Gln, Lys and His have mutabilities in the Humsavar set similar to those observed in the 1 kG dataset and much smaller than those in OMIM. This may reflect the larger Humsavar dataset (but this seems unlikely since Gly and Ala are quite common amind acids), so these specific discrepancies may rather reflect the origins of mutations in the two separate datasets.

Structural properties of disease-associated nsSNPs

The disease-associated OMIM variants show a slight preference for buried sites (33%) compared to all residues (27%) in the human proteome (Figure 5A) is even stronger in the Humsavar data (41%). This contrasts with the 'natural' variants of the 1 kG data,

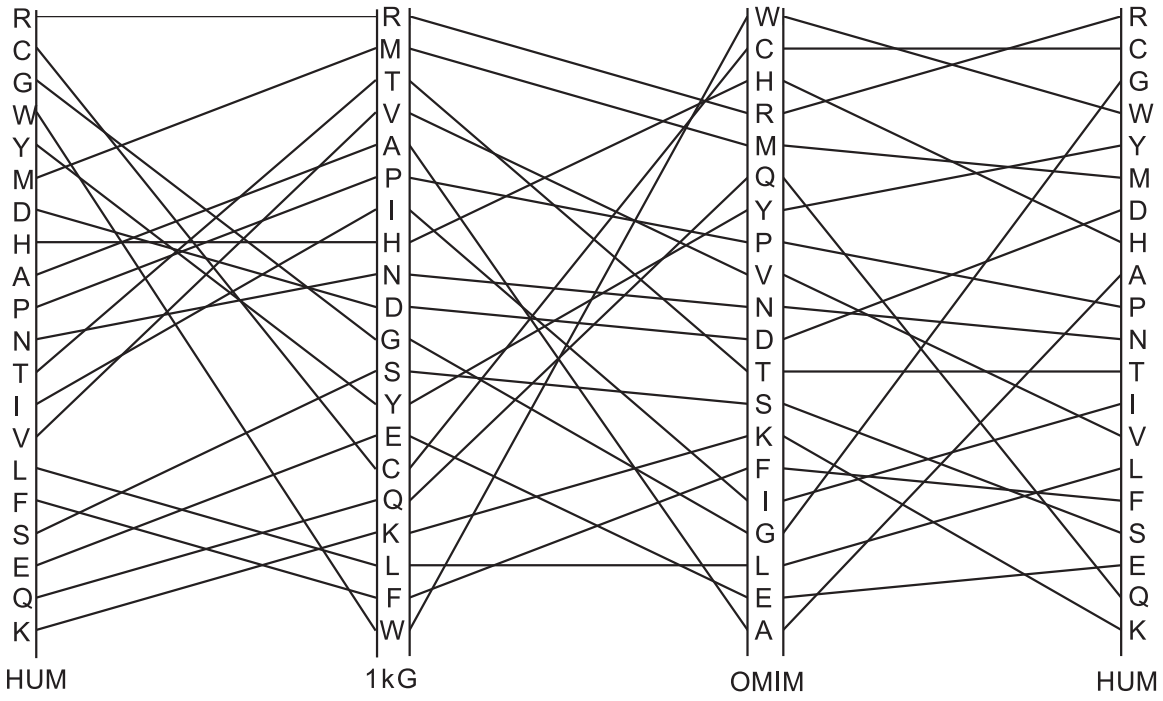


Figure 11. Amino acid mutability rank order plot comparing the mutability scores for 1 kG, OMIM and Humsavar residues. The most mutable amino acids are at the top. Correlation coefficients for 1 kG vs OMIM, 1 kG vs Humsavar and OMIM vs Humsavar are 0.09, 0.17 and 0.51, respectively.
doi:10.1371/journal.pcbi.1003382.g011

which show a decreased preference (18%) for the interior. Our work broadly agrees with a smaller study done by Gong and Blundell [21] that showed 60–65% of disease associated nsSNPs are solvent exposed. We found an almost identical distribution of OMIM and Humsavar variants compared to all residues and the 1 kG variants between the different secondary structures (Figure 5B).

Figure 8A shows the differences in the DOPE scores [42] calculated for each variant during the structural modelling process for the 1 kG, OMIM and Humsavar datasets. The distribution for the disease-associated variants is shifted towards larger positive energies in both datasets, indicating that the variants destabilize the protein slightly more than the non-disease variants. In contrast to the 1 kG data, OMIM mutations are more likely to increase polarity (54%) and more likely to decrease size (51.6%, Figure 8B,C). The two datasets show some detailed differences in size and hydrophobicity changes. The Humsavar variants less frequently reduce size or decrease hydrophobicity compared to OMIM mutations.

Functional annotations

In the OMIM set, 11.2% (209 of 1,864) of the modelled mutations were annotated with a function (Figure 5C and methods). This is less than the distribution for all residues (29.1%) and that seen for the 1 kG variants (15.5%). For the Humsavar data this drops to only 6.5%. This is a surprising finding, which needs further validation. It implies that most disease-associated mutations do not have a direct effect on the proteins' catalytic or binding sites but instead act through other, unannotated residues such as those which affect overall structure and stability or are involved in as yet unidentified protein-protein interfaces.

Conservation

There is a clear difference in the conservation score distribution between natural variants and the OMIM and Humsavar variants (Figure 6). The natural variants occur across the entire range of conservation but the OMIM and Humsavar variants show a peak in the more conserved residues. This is consistent with the idea that mutations in conserved residues often lead to disease.

Discussion

The results presented herein are subject to a few caveats, the most serious being related to the limited and possibly biased disease-associated data in OMIM. There are only ~10,000 variants in our OMIM set and these have variable experimental validation, and may indeed be biased according to scientists' preconceptions that such mutations should correspond to the residues that are most conserved and the amino acid exchanges that generate the largest changes in physicochemical characteristics. The Humsavar set has over 23,000 disease variants, however the requirements for inclusion are based on an annotation of 'involvement in disease'. This annotation is derived from either OMIM annotations or associations found in literature during curation of the SwissProt data. Notwithstanding, the OMIM dataset is one of the best available at the present time, although the coming years will see major expansion and hopefully improvements in such data. The results highlight the complex interplay of features from the level of the DNA up to protein sequence and

structure. The codon CpG dinucleotide content plays a large role in determining which amino acids mutate. This in turn affects the mutability of amino acids and a clear difference was seen between non-disease and disease variants where amino acids that are naturally very mutable, show the opposite trend in the disease-associated data.

The data for the 1000 Genomes provides a new experimental baseline against which amino acid profiles may be compared. Although there might be sequencing biases due to the DNA sequencing techologies used [43], every effort has been made by the 1000 Genomes consortium to correct for this. They estimate that using consensus calling on data produced by multiple platforms results in an error rate of 1–4%, thus having a small but negligible impact on our results. The current results show evidence for some protein selection, mainly in that the variants occur slightly more often on the surface of the protein and are much less likely to be annotated as functional than expected by chance. However, we should note that even the best definition of functional, taken from structural data, is limited. At one level, the definition is rather broad. For example, all residues in contact with a ligand are described as functional, but this is a major underestimate since many cognate ligands are not present in the crystal structures and similarly protein-protein interactions are rarely captured. In addition there are still relatively few complete structures for human proteins, which makes analysis of the effects of variants more difficult.

Even with these caveats, it is clear that the 1 kG variants eschew functional residues as defined here, a trend which is surprisingly even stronger in the OMIM and Humsavar data. The preference for OMIM mutations to be more buried and less functional supports the suggestion that these variants predominantly affect the structure and stability of the protein [4]. This is a similar result to that found by Sunyaev and co-workers [44] on a much smaller set. They found that 35% of disease variants were buried and a more detailed analysis found that ~70% of the variants are located in structurally and functionally important regions. Therefore these disease-associated mutations may well target residues that are remote from the active site, which modulate rather than obliterate the function of the protein. For example, for an enzyme, the primary catalytic residues are rarely targeted, but the 'secondary' residues in the interior (which affect stability) or on the surface, which may affect protein-protein interactions, could modulate function. However, the higher than average conservation scores for OMIM and Humsavar sites suggest that these disease-associated residues, although not defined as 'functional', are still important for the organism. This needs further investigation, with particular attention to how 'functional' residues are defined and whether we can improve on this definition.

Bringing together all the above observations for disease-associated and natural variants in ~1000 humans, we observe that the mutability of amino acids is largely driven by the properties of the DNA and mutational mechanisms, which favour mutations at codons containing a CpG dinucleotide. Therefore mutations to Arg residues are more than twice as common as any other mutation. However there are clearly other factors at play, which determine the frequency of variants, even at the DNA level. Although the disease-associated variants (both OMIM and Humsavar) follow the same pattern as the 1 kG variants (i.e. the same mutations are present in both sets, as dictated by the genetic code), the rank order of amino acids, according to their probability of being disease-associated, is radically different from that expected on the basis of the 1 kG data, with some of the rarer amino acids being shifted to the top of the list.

There is a small but significant impact of the protein structure on amino acid mutability, so that natural variants occur slightly more often in non-conserved regions. 59.4% of variations increase the hydrophobicity of the amino acid and 52.4% increase its size in the natural set, while OMIM variants often result in larger changes in the size and hydrophobicity of the amino acid and are more destabilising on average than 1 kG variants. The Humsavar data supports this idea that disease variants result in more extreme changes. The selection pressures captured in the WAG and PAM matrices 'purify' out the 'natural' variants, removing variants with large changes in size and hydrophobicity. The amino acids all show distinctive exchange profiles, whereby some exchanges are very common and some very rare, which provides an empirical expectation for any specific exchange in humans.

As the cost of sequencing drops rapidly, many more genomes will be sequenced and experimental validation of disease-causing mutations will improve as a result of more data. Much better codon-based models of evolution will be attainable, allowing in turn a better dissection of the impact of selection at the protein level. The data herein will be used to develop an improved method to predict the effects of individual mutations, to explore cancer-related amino acid mutations, to investigate and compare mutational profiles in different organisms as well as improving codon mutation models for human DNA.

Methods

Non-synonymous mutations in humans

UniProt [5] was queried for all reviewed protein sequences belonging to *Homo sapiens*. 19,058 entries were retrieved. The Ensembl transcript ID [45] was obtained for each protein sequence using the mapping provided by UniProt (17,708 UniProt entries were mapped to 40,351 Ensembl transcript IDs). Immunoglobulins and major histocompatibility complex proteins were excluded as they are inherently variable. For every protein, the Ensembl v67 Perl API was used to query the transcript ID in Ensembl for nsSNPs found in the 1 kG data set (as available on 1 August 2012). To reduce the inherent uncertainty involved in determining the ancestral allele, only mutations that occurred in one of the 1000 Genomes described populations were used, with the allele present in all populations considered the ancestral, hence defining the direction of the mutation. This increases the chances that the variant found in the 1 kG data is a mutation away from the ancestral genome. 106,311 mutations were found and this data set, containing the 'natural' variants found in the 1 kG project, will be referred to as the 1 kG set.

Residue conservation scores for each residue in every protein sequence were calculated using the Evolutionary Trace server [35]. Conservation scores for 2,274 sequences could not be calculated due to the methodology used by the Evolutionary Trace server that disregards residues in columns of the multiple alignment containing more than 60% gaps and ranked as being non-conserved, as well as residues judged by the algorithm not to have enough information. This process almost certainly preferentially excludes surface residues (where insertions and deletions are most common) but since we are using the conservation distribution for comparisons, this bias is not significant. The UniProt sequences were used to calculate the relative abundance of amino acids in human proteins. A total of about 10.5 million amino acids were counted. For each protein sequence, the OMIM Mutations search tool (http://www.bioinf.org.uk/omim) was queried with the UniProt entry ID to retrieve variants found in OMIM. Only variants for which the correct amino acid position in the protein has been verified, were used for the OMIM data set and will be

referred to as the OMIM set. 556 of the OMIM mutations were found in the 1 kG set (0.5%). Although these represent a very small fraction we removed them so that they did not bias the results.

The instantaneous rate change matrices were derived using the DCFreq method [36] and the human proteome frequencies.

Mutability of amino acids

A mutability score for every amino acid was calculated by taking the total number of mutations for a specific amino acid in the data and dividing by the frequency of occurrence for the specific amino acid in the human genome. The proportional representation of each amino acid in the human proteome is given in supplemental Table S1.

Statistical validation

We compared the amino acid variant counts in the 1 kG and OMIM data using Fischer's exact test in the R package (R Development Core Team, 2011). Multiple comparison correction was done on the p-values for each amino acid using p.adjust in R with the Benjamini-Hochberg-Yekutieli method [46,47]. P-values lower than 0.01 were considered statistically significant. For correlation values, $r>0.7$ and $r<-0.7$ were considered strong, $0.4<r<0.7$ and $-0.4>r>-0.7$ were considered moderate and $0.3>r>-0.3$ weak or no correlation.

Retrieving human proteins and their structures

The protein structure data set was constructed by first taking all the above mentioned protein sequences and annotating each with their respective Pfam [48] domains. Only proteins for which there were matching entries in the Protein Data Bank (PDB, [49]) were kept. This resulted in a list containing the UniProt identifiers for all known human proteins that have at least one structure in the PDB. For accuracy, the corresponding PDB structures were then filtered to include only X-ray structures. Using the Pfam mapping, only protein structures containing all the protein's Pfam domains were kept. The final list contained 2,139 protein chains and will be referred to as the 3D set.

A set consisting only of human monomeric proteins was also constructed. An algorithm was implemented whereby a protein was classified as being either a multimer or a monomer based on a majority vote. The predictions used were from PISA [50], UniProt, 3DComplex [51], PIQSI [52], PQS-PITA [53–55], relevant PubMed abstracts and REMARK 350 records from the PDB structure file. The oligomeric predictions from each of the servers were collected for every protein in the 3D set. Only when the majority of the servers agreed on the most probable oligomeric state of the protein, was it designated as either a multimer or a monomer. The monomeric protein list contained 325 proteins and will be referred to as the monomer set.

Another homology-based set was constructed using the human models in ModBase [31]. Models with 90–100% sequence identity and coverage were used as templates. This set contained 2,630 models and will be referred to as the model set.

Protein chain annotation

Each protein chain in the 3D, monomer and model sets was annotated with information from various databases and online resources. Information about protein properties such as catalytic residues, metal-binding residues, ligand-binding residues and PROSITE patterns [56] were extracted from PDBsum [34] and additional functional residue annotations were retrieved using SAS (Sequence Annotated by Structure, [33]). The 3D coordinates for each of the proteins in the structure data sets were retrieved from the PDB. To maintain consistency between the PDB and UniProt residue numbering, the SIFTS mapping [57] for each protein chain was used. NACCESS was used to calculate the relative solvent accessibilities for the individual residues in a chain. A cut-off of 5% solvent exposure was used to distinguish between buried and exposed residues.

Mapping nsSNPs to structures

To investigate the effect a nsSNP might have, each individual nsSNP was mapped to its correct amino acid in the protein structure. For every such nsSNP that could be mapped, a homology model of the protein containing the nsSNP was built using Modeller 9v3 [32] with the original protein structure serving as the template. A maximum of 200 steps of conjugate gradient minimization followed by 200 rounds of molecular dynamics at 300 K (using Modeller) was applied to each variant and its structural context analysed. NACCESS was run on all the variant models to identify changes in solvent accessibility. Comparisons of the Modeller DOPE score (Discrete Optimized Protein Energy, [42]) were made between the nsSNP model and the reference structure to estimate the magnitude of change that a variant might cause. The 1 kG models are available in PDBsum (http://www.ebi.ac.uk/pdbsum/) by looking at the specific PDB code of interest.

Supporting Information

Figure S1 Mutabilities of the amino acids for each population. AMR: American admixed, ASN: South East Asian, AFR:African, EUR: European.
(EPS)

Figure S2 The distribution of average protein mutabilites for all human proteins (blue) and disease associated proteins (red).
(EPS)

Figure S3 The amino acid exchanges observed in human protein variants. The 1 kG data set is the top row of each cell and Humsvar(SP) the bottom row of each cell*. Amino acids are arranged by 1 letter code according to increasing hydrophobicity (least hydrophobic is left and most hydrophobic is right) using the Fauchère and Pliska scale. Yellow blocks indicate mutations where there are statistically significant differences between 1 kG and Humsavar. Blue blocks indicate where no mutations were present in the 1 kG data set. White blocks show where there are no statistically significant differences. Green blocks show where there are proportionally more 1 kG mutations compared to Humsavar. Orange blocks show where there are proportionally more Humsavar mutations than 1 kG. The mutability scores (see methods) for the 1 kG and Humsavar sets are shown in the last column. *Note that these matrices are fundamentally different. The 1 kG data set gathers all the observed mutations in the 1 kG project, counting each only once; the Humsavar data set combines information gathered from potentially many individuals but filtered to identify those mutations associated with a disease.
(EPS)

Figure S4 Comparison of the differences in observed mutations in the various sets. Comparison of the differences in the % of observed mutations in the 1 kG (blue) and Humsavar (red) sets for one amino acid mutating to all others e.g. proportionally, more mutations from Lys to Glu are recorded in Humsavar than in the 1 kG set. Each plot shows the results of

mutation from a specific amino acid (e.g. Arg at top left) to every other amino acid.
(EPS)

Figure S5 Comparison of the differences in observed mutations in the various sets. Comparison of the differences in the % of observed mutations in the Humsavar (green) and OMIM (red) sets for one amino acid mutating to all others. Each plot shows the results of mutation from a specific amino acid (e.g. Arg at top left) to every other amino acid.
(EPS)

Table S1 The relative abundances of the various amino acids in the UniProt protein set.
(PDF)

References

1. 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, et al (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
2. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.
3. Iengar P (2012) An analysis of substitution, deletion and insertion mutations in cancer genes. Nucleic Acids Res 40: 6401–6413.
4. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37: D793–D796.
5. UniProt-Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–D148.
6. Stenson PD, Ball E, Howells K, Phillips A, Mort M, et al. (2008) Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet 45: 124–126.
7. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7: 61–80.
8. Steward RE, MacArthur MW, Laskowski RA, Thornton JM (2003) Molecular basis of inherited diseases: a structural perspective. Trends Genet 19: 505–513.
9. Fabre KM, Ramaiah L, Dregalla RC, Desaintes C, Weil MM, et al. (2011) Murine Prkdc polymorphisms impact DNA-PKcs function. Radiat Res 175: 493–500.
10. Minutolo C, Nadra AD, Fernández C, Taboas M, Buzzalino N, et al. (2011) Structure-based analysis of five novel disease-causing mutations in 21-hydroxylase-deficient patients. PLoS One 6: e15899.
11. González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88: 440–449.
12. Bromberg Y, Yachdav G, Rost B (2008) SNAP predicts effect of mutations on protein function. Bioinformatics 24: 2397–2398.
13. Worth CL, Preissner R, Blundell TL (2011) Sdm–a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 39: W215–W222.
14. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.
15. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070.
16. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863–874.
17. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812–3814.
18. Calabrese R, Capriotti E, Fariselli P,Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30: 1237–1244.
19. Nakken S, Alseth I, Rognes T (2007) Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. Neuroscience 145: 1273–1279.
20. Reumers J, Schymkowitz J, Rousseau F (2009) Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations. BMC Bioinformatics 10 Suppl 8: S9.
21. Gong S, Blundell TL (2010) Structural and functional restraints on the occurrence of single amino acid variations in human proteins. PLoS One 5: e9186.
22. Kamaraj B, Purohit R (2013) Computational screening of disease-associated mutations in OCA2 gene. Cell Biochem Biophys: 1–13.
23. Gong S, Worth CL, Bickerton GRJ, Lee S, Tanramluk D, et al. (2009) Structural and functional restraints in the evolution of protein families and superfamilies. Biochem Soc Trans 37: 727–733.
24. Worth CL, Gong S, Blundell TL (2009) Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol 10: 709–720.
25. Subramanian S, Kumar S (2006) Evolutionary anatomies of positions and types of diseaseassociated and neutral amino acid mutations in the human genome. BMC Genomics 7: 306.
26. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, et al. (2005) A universal trend of amino acid gain and loss in protein evolution. Nature 433: 633–638.
27. Hurst LD, Feil EJ, Rocha EPC (2006) Protein evolution: causes of trends in amino-acid gain and loss. Nature 442: E11–2; discussion E12.
28. Walser JC, Furano AV (2010) The mutational spectrum of non-CpG DNA varies with CpG content. Genome Res 20: 875–882.
29. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475.
30. Zuckerkandl E, Derancourt J, Vogel H (1971) Mutational trends and random processes in the evolution of informational macromolecules. J Mol Biol 59: 473–490.
31. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, et al. (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 39: D465–D474.
32. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779–815.
33. Milburn D, Laskowski RA, Thornton JM (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. Protein Eng 11: 855–859.
34. Laskowski RA (2009) PDBsum new things. Nucleic Acids Res 37: D355–D359.
35. Mihalek I, Res I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 336: 1265–1282.
36. Kosiol C, Goldman N (2005) Different versions of the Dayhoff rate matrix. Mol Biol Evol 22: 193–199.
37. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
38. Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5(3): 345–351.
39. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994–997.
40. Akashi H, Osada N, Ohta T (2012) Weak selection and protein evolution. Genetics 192: 15–31.
41. Vitkup D, Sander C, Church GM (2003) The amino-acid mutational spectrum of human genetic disease. Genome Biol 4: R72.
42. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15: 2507–2524.
43. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. (2013) Characterizing and measuring bias in sequence data. Genome Biol 14: R51.
44. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16: 198–200.
45. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, et al. (2010) Ensembl's 10th year. Nucleic Acids Res 38: D557–D562.
46. Hochberg YBY (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Statistical Society 57(1): 289–300.
47. Yekutieli YBD (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29(4): 1165–1188.
48. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–D222.
49. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
50. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372: 774–797.
51. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2: e155.

Acknowledgments

We would like to thank Angela Wilkins for running the large scale conservation analysis, Grecia Lapizco-Encinas for constructing the monomer set, Arjun Ray for doing the analysis of the ModBase models and Ewan Birney for valuable discussions.

Author Contributions

Conceived and designed the experiments: TAPdB RAL SLP BS NG JMT. Performed the experiments: TAPdB RAL. Analyzed the data: TAPdB SLP BS. Wrote the paper: TAPdB RAL JMT. Valuable discussion regarding the method design: NG JMT SLP BS.

52. Levy ED (2007) PiQSi: protein quaternary structure investigation. Structure 15: 1364–1367.
53. Ponstingl H, Henrick K, Thornton JM (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 41: 47–57.
54. Henrick K, Thornton JM (1998) PQS: a protein quaternary structure file server. Trends Biochem Sci 23: 358–361.
55. Ponstingl H, Kabir T, Gorse D, Thornton JM (2005) Morphological aspects of oligomeric protein structures. Prog Biophys Mol Biol 89: 9–35.
56. Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3: 265–274.
57. Velankar S, McNeil P, Mittard-Runte V, Suarez A, Barrell D, et al. (2005) E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res 33: D262–D265.
58. Fauchère JL, Charton M, Kier LB, Verloop A, Pliska V (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 32: 269–278.

Appendix C

C.1 Supplementary Figures

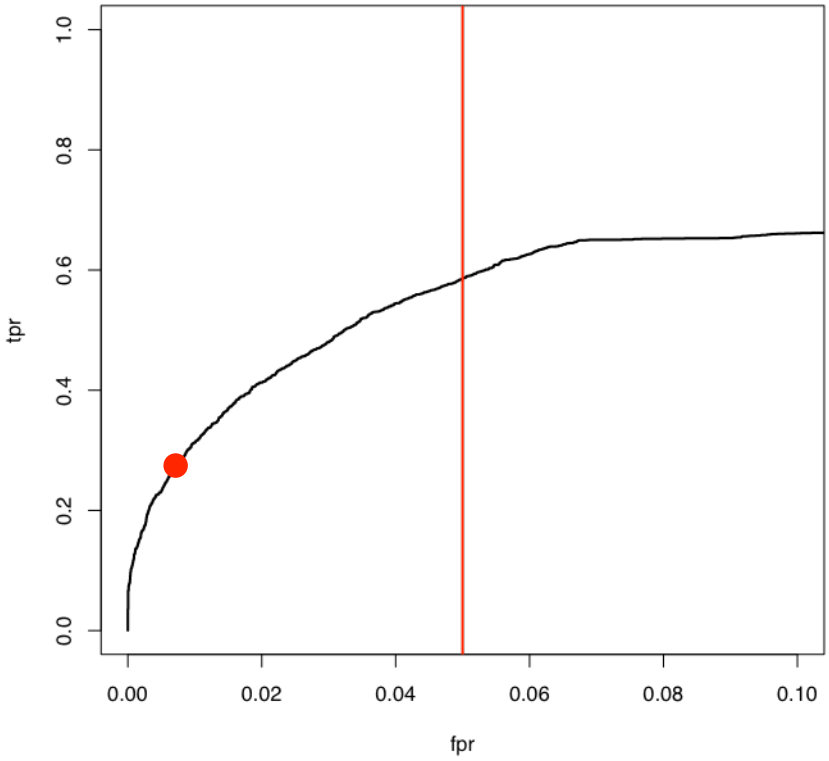


Figure C.1: ROC curve for a simulated dataset with the ω-distribution of the mammal dataset shown in Figure 5.1a, including those which are positively selected, showing the FPR against the TPR. The red line indicates the 5% FPR point, corresponding to a power of approximately 0.6. My aim is to devise a statistical test that approaches this power as closely as possible, without exceeding the chosen FPR. The red dot corresponds to the FPR and TPR found if the $\bar{\chi}^2_1$ threshold of Massingham and Goldman (2005) is used: 0.0075 and 0.27, respectively. Note that the 'cost' of this reduced FPR is that the TPR is reduced by about one half.

References

Ababneh F., Jermiin L. S., Ma C., and Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics* 22: 1225–1231.

Abascal F., Posada D., and Zardoya R. 2007. MtArt: a new model of amino acid replacement for Arthropoda. *Molecular Biology and Evolution* 24: 1–5.

Abby S. S., Tannier E., Gouy M., and Daubin V. 2010. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics* 11: 324.

Adachi J. and Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution* 42: 459–468.

Adachi J., Waddell P. J., Martin W., and Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution* 50: 348–358.

Alexandrov L. B., Nik-Zainal S., Wedge D. C., Campbell P. J., and Stratton M. R. 2013. Deciphering signatures of mutational processes operative in human cancer. *Cell reports* 3: 246–59.

Amberger J., Bocchini C. A., Scott A. F., and Hamosh A. 2009. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Research* 37: D793–D796.

Anderson F. E. and Swofford D. L. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Molecular Phylogenetics and Evolution* 33: 440–451.

Anisimova M. and Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution* 26: 255–271.

Anisimova M., Bielawski J. P., and Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution* 18: 1585–1592.

Anisimova M., Bielawski J. P., and Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution* 19: 950–958.

Anisimova M., Nielsen R., and Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164: 1229–1236.

Arenas M. and Posada D. 2010. The effect of recombination on the reconstruction of ancestral sequences. *Genetics* 184: 1133–1139.

Arndt P. F., Hwa T., and Petrov D. A. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *Journal of Molecular Evolution* 60: 748–763.

Baele G. 2011. Context-dependent evolutionary models for non-coding sequences: an overview of several decades of research and an analysis of Laurasiatheria and Primate evolution. *Evolutionary Biology* 39: 61–82.

Baele G., Van de Peer Y., and Vansteelandt S. 2010. Using non-reversible context-dependent evolutionary models to study substitution patterns in primate non-coding sequences. *Journal of Molecular Evolution* 71: 34–50.

Barros M. C., Sampaio I., and Schneider H. 2008. Novel 12S mtDNA findings in sloths (Pilosa, Folivora) and anteaters (Pilosa, Vermilingua) suggest a true case of long branch attraction. *Genetics and Molecular Biology* 31: 793–799.

Barry D. and Hartigan J. A. 1987. Statistical analysis of hominoid molecular evolution. *Statistical Science* 2: 191–207.

Behjati S., Huch M., Boxtel R. V., Karthaus W., David C, Tarpey P. S., Roerink S., Blokker J., Maddison M., Robinson B., Nik-zainal S., Campbell P., Goldman N., Van M., Wetering D., Cuppen E., Clevers H., and Stratton M. R. 2014. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* 513: 422–425.

Benjamini Y. and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* 57: 289–300.

Benner S. A., Cohen M. A., and Gonnet G. H. 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Engineering* 7: 1323–1332.

Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21: 163–193.

Bielejec F., Lemey P., Baele G., Rambaut A., and Suchard M. A. 2014. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Systematic Biology* 63: 493–504.

Birney E. et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816.

Bishop M. J. and Thompson E. A. 1986. Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology* 190: 159–165.

Blanquart S. and Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution* 25: 842–858.

Bodilis J., Meilo S., Cornelius P., Vos P. D., and Barray S. 2011. A long-branch attraction artifact reveals an adaptive radiation in Pseudomonas. *Molecular Biology and Evolution* 28: 2723–2726.

Boussau B. and Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology* 55: 756–768.

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

Bulmer M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution* 8: 868–883.

Cao Y., Adachi J., Janke A., Pii 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.

Cavalli-Sforza L. L. and Edwards A. W. 1967. Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics* 19: 233–257.

Chang J. T. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences* 137: 51–73.

Chor B. and Snir S. 2004. Molecular clock fork phylogenies: closed form analytic maximum likelihood solutions. *Systematic Biology* 53: 963–967.

Chor B. and Snir S. 2007. Analytic solutions of maximum likelihood on forks of four taxa. *Mathematical Biosciences* 208: 347–358.

Chor B., Hendy M., and Penny D. 2001. Analytic solutions for three taxon MLMC trees with variable rates across sites. *Discrete Applied Mathematics* 155: 750–758.

Chor B., Hendy M., and Snir S. 2006a. Maximum likelihood Jukes-Cantor triplets: analytic solutions. *Molecular Biology and Evolution* 23: 626–632.

Chor B., Khetan A., and Snir S. 2006b. Maximum likelihood molecular clock comb: analytic solutions. *Journal of Computational Biology* 13: 819–837.

Conde L., Vaquerizas J. M., Dopazo H., Arbiza L., Reumers J., Rousseau F., Schymkowitz J., and Dopazo J. 2006. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Research* 34: W621–5.

Connor T. O., Sundberg K., Carroll H., Clement M., and Snell Q. 2010. Analysis of long branch extraction and long branch shortening. *BMC Genomics* 11: S14.

Coulondre C., Miller J. H., Farabaugh P. J., and Gilbert W. 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* 274: 775–780.

Cox C. J. and Foster P. G. 2013. A 20-state empirical amino-acid substitution model for green plant chloroplasts. *Molecular Phylogenetics and Evolution* 68: 218–220.

Cox D. 1961. Tests of separate families of hypotheses. *Proceedings of the 4th Berkeley Symposium* 1: 105:123.

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

Dabert M., Witalinski W., Kazmierski A., Olszanowski Z., and Dabert J. 2010. Molecular phylogeny of acariform mites (Acari, Arachnida): strong conflict between phylogenetic signal and long-branch attraction artifacts. *Molecular Phylogenetics and Evolution* 56: 222–241.

Dacks J. B., Marinets A., Ford Doolittle W, Cavalier-Smith T., and Logsdon J. M. 2002. Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang. *Molecular Biology and Evolution* 19: 830–840.

Dang C. C., Le Q. S., Gascuel O., and Le V. S. 2010. FLU, an amino acid substitution model for influenza proteins. *BMC Evolutionary Biology* 10: 99.

Dang C. C., Lefort V., Le V. S., Le Q. S., and Gascuel O. 2011. ReplacementMatrix: a web server for maximum-likelihood estimation of amino acid replacement rate matrices. *Bioinformatics* 27: 2758–2760.

Darriba D., Taboada G. L., Doallo R., and Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9: 772.

Dayhoff M. O. and Eck R. V. 1968. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*. Ed. by M. O. Dayhoff and R. V. Eck. Washington D.C.: National Biomedical Research Foundation, 33–41.

Dayhoff M. O., Eck R. V., and Eck C. M. 1972. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*. Ed. by M. O. Dayhoff. Vol. 5. Washington D.C.: National Biomedical Research Foundation, 89–99.

Dayhoff M. O., Schwartz R. M., and Orcutt B. C. 1978. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*. Ed. by M.O. Dayhoff. Vol. 5. Washington D.C.: National Biomedical Research Foundation, 345–352.

de Beer T. A. P., Laskowski R. A., Parks S. L., Sipos B., Goldman N., and Thornton J. M. 2013. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Computational Biology* 9: e1003382.

De Maio N., Holmes I., Schlötterer C., and Kosiol C. 2013a. Estimating empirical codon hidden markov models. *Molecular Biology and Evolution* 30: 725–736.

De Maio N., Schlötterer C., and Kosiol C. 2013b. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Molecular Biology and Evolution* 30: 2249–2262.

Dessimoz C., Margadant D., and Gonnet G. H. 2008. DLIGHT Lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. *RECOMB*, 315–330.

dos Reis M., Hay A. J., and Goldstein R. A. 2009. Using non-homogeneous models of nucleotide substitution to identify host shift events: application to the origin

of the 1918 'Spanish' influenza pandemic virus. *Journal of Molecular Evolution* 69: 333–345.

Duncan B. K. and Miller J. H. 1980. Mutagenic deamination of cytosine rsidues in DNA. *Nature* 287: 560–561.

Duret L., Eyre-Walker A., and Galtier N. 2006. A new perspective on isochore evolution. *Gene* 385: 71–74.

Eck R. V. and Dayhoff M. O. 1966. *Atlas of Protein Sequence and Structure*. Silver Spring, Maryland: National Biomedical Research Foundation.

Edwards A. W. F. 1972. *Likelihood*. Cambridge: Cambridge University Press.

Efron B. 2010. *Large-Scale Inference*. Cambridge: Cambridge University Press.

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

Endo T., Ikeo K., and Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution* 13: 685–690.

Eyre-Walker A. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* 152: 675–683.

Fares M. A., Byrne K. P., and Wolfe K. H. 2006. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. *Molecular Biology and Evolution* 23: 245–253.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27: 401–410.

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

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783–791.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* 22: 521–565.

Felsenstein J. 2001. The troubled growth of statistical phylogenetics. *Systematic Biology* 50: 465–467.

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

Fisher R. A. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* I: 3–22.

Fisher R. A. 1925. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* 22: 700–725.

Fitch W. M. and Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155: 279–284.

Fleissner R., Metzler D., and von Haeseler A. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. *Systematic Biology* 54: 548–561.

Flicek P., Amode M. R., Barrell D., Beal K., Brent S., Chen Y., Clapham P., Coates G., Fairley S., Fitzgerald S., Gordon L., Hendrix M., Hourlier T., Johnson N., Kähäri A., Keefe D., Keenan S., Kinsella R., Kokocinski F., Kulesha E., Larsson P., Longden I., McLaren W., Overduin B., Pritchard B., Riat H. S., Rios D., Ritchie G. R. S., Ruffier M., Schuster M., Sobral D., Spudich G.,

Tang Y. A., Trevanion S., Vandrovcova J., Vilella A. J., White S., Wilder S. P., Zadissa A., Zamora J., Aken B. L., Birney E., Cunningham F., Dunham I., Durbin R., Fernández-Suarez X. M., Herrero J., Hubbard T. J. P., Parker A., Proctor G., Vogel J., and Searle S. M. J. 2011. Ensembl 2011. *Nucleic Acids Research* 39: D800–806.

Gaut B. and Lewis P. O. 1995. Success of maximum likelihood phlogeny inference in the four-taxon case. *Molecular Biology and Evolution* 12: 152–162.

Geuten K., Massingham T., Darius P., Smets E., and Goldman N. 2007. Experimental design criteria in phylogenetics: where to add taxa. *Systematic Biology* 56: 609–622.

Goldman N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of DNA substitution and to parsimony analyses. *Systematic Zoology* 39: 345–361.

Goldman N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36: 182–198.

Goldman N. and Whelan S. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Molecular Biology and Evolution* 17: 975–978.

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.

Goldman N., Thorne J. L., and Jones D. T. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149: 445–458.

Golub G. H. and Van Loan C. 1996. *Matrix Computations*. Baltimore: Johns Hopkins University Press.

Goode M., Guindon S., and Rodrigo A. 2008. Modelling the evolution of protein coding sequences sampled from Measurably Evolving Populations. *Genome Informatics*. 21: 150–164.

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

Guindon S., Dufayard J., and Lefort V. 2010. New algorithms and methods to estimate maximum-ikelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59: 307–321.

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

Hatta Y., Tsuchiya N., Matsushita M., Shiota M., Hagiwara K., and Tokunaga K. 1999. Identification of the gene variations in human CD22. *Immunogenetics* 49: 280–286.

Hein J., Wiuf C., Knudsen B., Møller M. B., and Wibling G. 2000. Statistical alignment: computational properties, homology testing and goodness-of-fit. *Journal of Molecular Biology* 302: 265–279.

Hein J., Jensen J. L., and Pedersen C. N. S. 2003. Recursions for statistical multiple alignment. *Proceedings of the National Academy of Sciences of the United States of America* 100: 14960–14965.

Hendy M. D. and Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38: 297–309.

Huelsenbeck J. P. 1995. Performance of phylogenetic methods in simulation. *Systematic Biology* 44: 17–48.

Huelsenbeck J. P. 1997. Is the Felsenstein Zone a fly trap? *Systematic Biology* 46: 69–74.

Huelsenbeck J. P. 1998. Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? *Systematic Biology* 47: 519–537.

Huelsenbeck J. P. and Hillis D. M. 1993. Success of phylogenetic methods in the four-taxon case. *Systematic Biology* 42: 247–264.

Huelsenbeck J. P., Bollback J. P., and Levine A. M. 2002. Inferring the root of a phylogenetic tree. *Systematic Biology* 51: 32–43.

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

Husmeier D. and Wright F. 2001. Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology* 8: 401–427.

Huson D. H. and Scornavacca C. 2010. A survey of combinatorial methods for phylogenetic networks. *Genome Biology and Evolution* 3: 23–35.

Huson D. H., Rupp R, and Scornavacca C. 2011. *Phylogenetic Networks*. Cambridge: Cambridge University Press.

Hwang D. G. and Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America* 101: 13994–14001.

Inagaki Y., Susko E., Fast N. M., and Roger A. J. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. *Molecular Biology and Evolution* 21: 1340–1349.

Jayaswal V., Jermiin L. S., Poladian L., and Robinson J. 2010. Two stationary nonhomogeneous markov models of nucleotide sequence evolution. *Systematic Biology* 60: 74–86.

Jayaswal V., Ababneh F., Jermiin L. S., and Robinson J. 2011. Reducing model complexity of the general Markov model of evolution. *Molecular Biology and Evolution* 28: 3045–3059.

Jensen J. L. and Pedersen A.-M. K. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Advances in Applied Probability* 32: 499–517.

Jones D. T., Taylor W. R., and Thornton J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8: 275–282.

Jordan G. 2011. Analysis of Alignment Error and Sitewise Constraint in Mammalian Comparative Genomics. PhD thesis. University of Cambridge.

Jukes T. H. and Cantor C. R. 1969. Evolution of protein molecules. *Mammalian Protein Metabolism*. Ed. by H. N. Munro. New York: Academic Press, 21–132.

Kelchner S. A. and Thomas M. A. 2007. Model use in phylogenetics: nine key questions. *Trends in Ecology & Evolution* 22: 87–94.

Kelly F. P. 1979. *Reversibility and Stochastic Networks*. Chichester: Wiley.

Kim J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Systematic Biology* 45: 363–374.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16: 111–120.

Kishino H. and Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution* 29: 170–179.

Klosterman P. S., Uzilov A. V., Bendaña Y. R., Bradley R. K., Chao S., Kosiol C., Goldman N., and Holmes I. 2006. XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* 7: 428.

Kosakovsky Pond S. L. and Frost S. D. W. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution* 22: 1208–1222.

Kosakovsky Pond S. L. and Muse S. V. 2005. Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution* 22: 2375–2385.

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

Kosakovsky Pond S. L., Frost S. D. W., Grossman Z., Gravenor M. B., Richman D. D., and Brown A. J. L. 2006a. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Computational Biology* 2: e62.

Kosakovsky Pond S. L., Posada D., Gravenor M. B., Woelk C. H., and Frost S. D. W. 2006b. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22: 3096–3098.

Köser C. U., Holden M. T. G., Ellington M. J., Cartwright E. J. P., Brown N. M., Oglivy-Stuart A. L., Yang Hsu L., Chewapreecha C., Croucher N. J., Harris S. R., Sanders M., Enright M. C., Dougan G., Bentley S. D., Parkhill J., Fraser L. J., Betley J. R., Schulz-trieglaff O. B., Smith G. P., and Peacock S. J. 2013. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *New England Journal of Medicine* 366: 2267–2275.

Kosiol C. and Goldman N. 2005. Different versions of the Dayhoff rate matrix. *Molecular Biology and Evolution* 22: 193–199.

Kosiol C. and Goldman N. 2011. Markovian and non-Markovian protein sequence evolution: aggregated Markov process models. *Journal of Molecular Biology* 411: 910–923.

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

Kosiol C., Vinar T., da Fonseca R. R., Hubisz M. J., Bustamante C. D., Nielsen R., and Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genetics* 4: e1000144.

Kück P., Mayer C., Wägele J.-W., and Misof B. 2012. Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS ONE* 7: e36593.

Kuhner M. K. and Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* 11: 459–468.

Lanave C., Preparata G., Saccone C., and Serio G. 1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* 20: 86–93.

Lartillot N. and Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* 21: 1095–1109.

Le S. Q. and Gascuel O. 2008. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* 25: 1307–1320.

Li T., Hua J., Wright A. M., Cui Y., Xie Q., Bu W., and Hillis D. M. 2014. Long-branch attraction and the phylogeny of true water bugs (Hemiptera: Nepomorpha) as estimated from mitochondrial genomes. *BMC Evolutionary Biology* 14: 99.

Li W.-H., Wu C.-I., and Luo C.-C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* 2: 150–174.

Lindblad-Toh K., Garber M., Zuk O., Lin M. F., Parker B. J., Washietl S., Kheradpour P., Ernst J., Jordan G., Mauceli E., Ward L. D., Lowe C. B., Holloway A. K., Clamp M., Gnerre S., Alföldi J., Beal K., Chang J., Clawson H., Cuff J., Di Palma F., Fitzgerald S., Flicek P., Guttman M., Hubisz M. J., Jaffe D. B., Jungreis I., Kent W. J., Kostka D., Lara M., Martins A. L., Massingham T., Moltke I., Raney B. J., Rasmussen M. D., Robinson J., Stark A., Vilella A. J.,

Wen J., Xie X., Zody M. C., Baldwin J., Bloom T., Chin C. W., Heiman D., Nicol R., Nusbaum C., Young S., Wilkinson J., Worley K. C., Kovar C. L., Muzny D. M., Gibbs R. A., Cree A., Dihn H. H., Fowler G., Jhangiani S., Joshi V., Lee S., Lewis L. R., Nazareth L. V., Okwuonu G., Santibanez J., Warren W. C., Mardis E. R., Weinstock G. M., Wilson R. K., Delehaunty K., Dooling D., Fronik C., Fulton L., Fulton B., Graves T., Minx P., Sodergren E., Birney E., Margulies E. H., Herrero J., Green E. D., Haussler D., Siepel A., Goldman N., Pollard K. S., Pedersen J. S., Lander E. S., and Kellis M. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478: 476–482.

Lindsay J. K. 1974a. Comparison of probability distributions. *Journal of the Royal Statistical Society. Series B* 36: 38–44.

Lindsay J. K. 1974b. Construction and comparison of statistical models. *Journal of the Royal Statistical Society. Series B* 36: 418–425.

Liò P and Goldman N. 1999. Using protein structural information in evolutionary inference: transmembrane proteins. *Molecular Biology and Evolution* 16: 1696–1710.

Lobry J. R. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. *Journal of Molecular Evolution* 40: 326–330.

Lohmueller K. E., Indap A. R., Schmidt S., Boyko A. R., Hernandez R. D., Hubisz M. J., Sninsky J. J., White T. J., Sunyaev S. R., Nielsen R., Clark A. G., and Bustamante C. D. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451: 994–997.

Luenberger D. 1984. *Introduction to Linear and Non-linear Programming*. Reading: Addison-Wesley.

Lunter G. and Hein J. 2004. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* 20 Suppl 1: i216–i223.

Lunter G., Miklós I., Drummond A., Jensen J. L., and Hein J. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* 6: 83.

Martyn I. and Steel M. 2012. The impact and interplay of long and short branches on phylogenetic information content. *Journal of Theoretical Biology* 314: 157–163.

Massingham T. and Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics* 169: 1753–1762.

McGuire G., Denham M. C., and Balding D. J. 2001. Models of sequence evolution for DNA sequences containing gaps. *Molecular Biology and Evolution* 18: 481–490.

Meredith R. W., Janečka J. E., Gatesy J., Ryder O. A., Fisher C. A., Teeling E. C., Goodbla A., Eizirik E., Simão T. L. L., Stadler T., Rabosky D. L., Honeycutt R. L., Flynn J. J., Ingram C. M., Steiner C., Williams T. L., Robinson T. J., Burk-Herrick A., Westerman M., Ayoub N. A., Springer M. S., and Murphy W. J. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* 334: 521–524.

Miklós I., Lunter G. A., and Holmes I. 2004. A "Long Indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution* 21: 529–540.

Mitchison G. and Durbin R. 1995. Tree-based maximal likelihood substitution matrices and hidden markov models. *Journal of Molecular Evolution* 41: 1139–1151.

Moler C. and Loan C. V. 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review* 45: 3–49.

Money D. and Whelan S. 2012. Characterizing the phylogenetic tree-search problem. *Systematic Biology* 61: 228–239.

Morrison D. A. 2007. Increasing the efficiency of searches for the maximum likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences. *Systematic Biology* 56: 988–1010.

Morrison D. A. 2011. *Introduction to Phylogenetic Networks*. Uppsala: RJR Productions.

Murphy W. J., Eizirik E., O'Brien S. J., Madsen O., Scally M., Douady C. J., Teeling E., Ryder O. A., Stanhope M. J., de Jong W. W., and Springer M. S. 2001. Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science* 294: 2348–2351.

Murrell B., Moola S., Mabona A., Weighill T., Sheward D., Kosakovsky Pond S. L., and Scheffler K. 2013. FUBAR: a Fast, Unconstrained Bayesian AppRoximation for inferring selection. *Molecular Biology and Evolution* 30: 1196–1205.

Muse S. V. and Gaut B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* 11: 715–724.

Nasrallah C. A., Mathews D. H., and Huelsenbeck J. P. 2011. Quantifying the impact of dependent evolution among sites in phylogenetic inference. *Systematic Biology* 60: 60–73.

Nei M. and Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3: 418–426.

Nei M. and Jin L. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. *Molecular Biology and Evolution* 6: 290–300.

Neyman J. 1971. Molecular studies of evolution: a source of novel statistical problems. *Statistical Theory and Related Topics*. Ed. by J Gupta and S Yackel. New York: Academic Press, 1–27.

Nielsen R. and Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 929–936.

Nikolaev S. I., Montoya-Burgos J. I., Popadin K., Parand L., and Margulies E. H. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proceedings of the National Academy of Sciences* 104: 20443–20448.

Norris J. R. 1998. *Markov Chains (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge: Cambridge University Press.

Ohta T. and Gillespie J. 1996. Development of neutral and nearly neutral theories. *Theoretical Population Biology* 49: 128–142.

Omilian A. R. and Taylor D. J. 2001. Rate acceleration and long-branch attraction in a conserved gene of cryptic daphniid (Crustacea) species. *Molecular Biology and Evolution* 18: 2201–2212.

Parks S. L. and Goldman N. 2014. Maximum likelihood inference of small trees in the presence of long branches. *Systematic Biology* 63: 798–811.

Philippe H and Laurent J. 1998. How good are deep phylogenetic trees? *Current Opinion in Genetics & Development* 8: 616–623.

Philippe H. and Germot A. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Molecular Biology and Evolution* 17: 830–834.

Plotkin J. B. and Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics* 12: 32–42.

Pol D and Siddall M. 2001. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. *Cladistics* 17: 266–281.

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

Qiu Y. L., Lee J., Whitlock B., Bernasconi-Quadroni F., and Dombrovska O. 2001. Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? Amborella, Nymphaeales, Illiciales, Trimeniaceae, and Austrobaileya. *Molecular Biology and Evolution* 18: 1745–1753.

Redelings B. 2014. Erasing errors due to alignment ambiguity when estimating positive selection. *Molecular Biology and Evolution* 31: 1979–1993.

Redelings B. D. and Suchard M. A. 2005. Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology* 54: 401–418.

Robinson D. M., Jones D. T., Kishino H., Goldman N., and Thorne J. L. 2003. Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution* 20: 1692–1704.

Rodrigue N., Lartillot N., Bryant D., and Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347: 207–217.

Rogers J. S. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Systematic Biology* 46: 354–357.

Rota-Stabelli O., Yang Z., and Telford M. J. 2009. MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. *Molecular Phylogenetics and Evolution* 52: 268–272.

Rubinstein N. D., Doron-Faigenboim A., Mayrose I., and Pupko T. 2011. Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. *Molecular Biology and Evolution* 28: 3297–3308.

Rutschmann F. 2006. Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. *Diversity and Distributions* 12: 35–48.

Rzhetsky A. and Nei M. 1992. Statistical properties of the ordinary least squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution* 35: 367–375.

Rzhetsky A. and Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution* 10: 1073–1095.

Rzhetsky A. and Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. *Molecular Biology and Evolution* 12: 131–151.

Saccone C., Lanave C., Pesole G., and Preparata G. 1990. Influence of base composition on quantitative estimates of gene evolution. *Methods in Enzymology* 183: 570–583.

Saitou N. and Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4: 406–425.

Sanderson M. J., Wojciechowski M. F., Hu J. M., Khan T. S., and Brady S. G. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Molecular Biology and Evolution* 17: 782–797.

Schierup M. H. and Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156: 879–891.

Schneider A., Cannarozzi G. M., and Gonnet G. H. 2005. Empirical codon substitution matrix. *BMC Bioinformatics* 6: 134.

Schulmeister S. 2004. Inconsistency of maximum parsimony revisited. *Systematic Biology* 53: 521–528.

Schwarz R. F., Fletcher W., Förster F., Merget B., Wolf M., Schultz J., and Markowetz F. 2010. Evolutionary distances in the twilight zone–a rational kernel approach. *PloS One* 5: e15788.

Self S. G. and Liang K.-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82: 605–610.

Sharp P. M. 1997. In search of molecular darwinism. *Nature* 387: 111–112.

Sharp P. M. and Li W.-H. 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* 15: 1281–1295.

Shimodaira H. and Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* 16: 1114–1116.

Shriner D., Nickle D. C., Jensen M. A., and Mullins J. I. 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genetical Research* 81: 115–121.

Siepel A. and Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution* 21: 468–488.

Silvey S. D. 1975. *Statistical Inference*. London: Chapman and Hall.

Squartini F. and Arndt P. 2008. Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Molecular Biology and Evolution* 25: 2525–2535.

Stefanović S., Rice D. W., and Palmer J. D. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: amborella or monocots? *BMC Evolutionary Biology* 4: 35.

Stiller J. W. and Hall B. D. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Molecular Biology and Evolution* 16: 1270–1279.

Storey J. D. 2003. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics* 31: 2013–2035.

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

Swanson W. J., Nielsen R., and Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Molecular Biology and Evolution* 20: 18–20.

Tajima F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution* 10: 677–688.

Tamura K. 1992. Estimation of the number of nucleotide substituitions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution* 9: 678–687.

Tamura K. and Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* 10: 512–526.

Tamuri A. U., dos Reis M., Hay A. J., and Goldstein R. A. 2009. Identifying changes in selective constraints: host shifts in influenza. *PLoS Computational Biology* 5: e1000564.

Tamuri A. U., dos Reis M., and Goldstein R. A. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190: 1101–1115.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* 17: 57–86.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.

Thorne J. L. and Kishino H. 1992. Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution* 9: 1148–1162.

Thorne J. L., Kishino H, and Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* 33: 114–124.

von Haeseler A. and Schoniger M. 1998. Evolution of DNA or amino acid sequences with dependent sites. *Journal of Computational Biology* 5: 149–163.

Wald A. 1949. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics* 20: 595–601.

Walser J.-C. and Furano A. V. 2010. The mutational spectrum of non-CpG DNA varies with CpG content. *Genome Research* 20: 875–882.

Whelan S. and Goldman N. 1999. Distributions of statistics used for the comparison of 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. and Money D. 2010. The prevalence of multifurcations in tree-space and their implications for tree-search. *Molecular Biology and Evolution* 27: 2674–2677.

Whelan S., Liò P., and Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics* 17: 262–272.

Whelan S., de Bakker P. I. W., 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.

Wiens J. J. and Hollingsworth B. D. 2000. War of the iguanas: conflicting molecular and morphological phylogenies and long-branch attraction in iguanid lizards. *Systematic Biology* 49: 143–159.

Wilcox T. P., García de León F. J., Hendrickson D. A., and Hillis D. M. 2004. Convergence among cave catfishes: long-branch attraction and a Bayesian relative rates test. *Molecular Phylogenetics and Evolution* 31: 1101–1113.

Wiley E. O. 1981. *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. New York: Wiley.

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

Wong W. S. 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.

Wu C.-I. and Maeda N. 1987. Inequality in mutation rates of the two strands of DNA. *Nature* 327: 169–170.

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. 1994c. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology* 43: 329–342.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* 15: 568–573.

Yang Z. 2000. Complexity of the simplest phylogenetic estimation problem. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 267: 109–116.

Yang Z. 2006. *Computational Molecular Evolution*. Oxford: Oxford University Press.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

Yang Z. 2014. *Molecular Evolution*. Oxford: Oxford University Press.

Yang Z. and Goldman N. 1994. Evaluation and extension of Markov process models for the evolution of DNA (in Chinese, with Abstract in English). *Acta Genetica Sinica* 21: 17–23.

Yang Z. and Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* 46: 409–418.

Yang Z. and Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* 19: 908–917.

Yang Z., Goldman N., and Friday A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution* 11: 316–314.

Yang Z., Nielsen R, and Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* 15: 1600–1611.

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

Yang Z., Wong W. S. W., and Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* 22: 1107–1118.

Yap V. B. and Speed T. 2005. Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evolutionary Biology* 5: 2.

Yu J. and Thorne J. L. 2006. Dependence among sites in RNA evolution. *Molecular Biology and Evolution* 23: 1525–1537.

Zharkikh A and Li W. H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution* 9: 1119–1147.

Zoller S. and Schneider A. 2013. Improving phylogenetic inference with a semiempirical amino acid substitution model. *Molecular Biology and Evolution* 20: 469–479.

Zuckerkandl E. and Pauling L. 1962. Molecular disease, evolution, and genetic heterogeneity. *Horizons in Biochemistry*. Ed. by M. Marsha and B. Pullman. New York: Academic Press, 189–225.

Zuckerkandl E. and Pauling L. 1965. Documents as molecules of evolutionary history. *Journal of Theoretical Biology* 8: 357–366.