Rates of Molecular Evolution and Phylogenomic Inference

Wai Lok Sibon Li

A thesis submitted in fulfilment of the requirements

for the degree of Doctorate of Philosophy in Computer Science,

University of Auckland

November 2010

Abstract

Genome studies have become an integral aspect of modern biology. As a result, there has been a need for methods to analyse genomic data. One aspect of genomic research is the analysis of variation in rate of evolution, both across a genome and between the genomes of species. In this study we explore the relationship between different types of rate heterogeneity. We develop several statistical methodologies to address issues associated with the phylogenomic analysis of genomic data.

Developments were made to the area of lineage-specific variation in evolutionary rates, with improvements in more efficient computational implementations of relaxed molecular clock models and the proposal of new models of rate changes across branches.

The practical application of relaxed molecular clock models was further examined with the proposal of methods of model averaging and model selection for relaxed molecular clock models using Bayesian stochastic search variable selection. Results show that our method identifies the most appropriate model for the underlying distribution of rates across branches in both simulated and real data. Our method of model averaging is particularly useful for preventing poor inference when the correct model is not known.

We examined the correlation in rates of substitution between functionally related genes that are caused by co-evolution of genes. Previously, this correlation was thought to only exist between genes with physically interacting gene products. We demonstrate that these correlations are not limited to genes with protein interactions but often extend to functionally related genes. Such patterns of co-evolution are of concern for the multi-gene analysis of genomic data and how species distances are estimated.

Finally, an attempt was made to develop a high-throughput method for detecting lineage-specific selection through identifying changes in rate of substitution. Results on simulated data indicate that our method had some success in characterising the variation in rate which occurs as a consequence of selection. Our methods were shown to provide significant speed benefits towards phylogenomic analyses.

The outcome of this research has been a progression in methodologies for phylogenomic analysis. Computer software has been developed to allow these methods to be used for understanding rate variations on a genomic scale.

Acknowledgements

First and foremost, I would like to thank my supervisors Alexei Drummond and Allen Rodrigo for all the help and guidance that they have provided over the last four years. They have both been role models for me and I thank them for the patience they have shown.

For ideas contributing to my research, I would like to thank Michael Defoin-Platel, Marc Suchard, David Bryant, Stephane Guindon, Steven Wu, Simon Greenhill, Peter Waddell and Howard Ross. For helping with reading this thesis and providing feedback, I would like to thank Evelyn Kiing, Steven Wu, Danushka Galappaththige, Alana Alexander and Jessie Wu.

I would like to thank my colleagues, the three-year undisputed champions, Steven Wu, Danushka Galappaththige, Peter Tsai, Alethea Rea, Manasa Ramakrishna, Jessica Hayward, Alana Alexander, Louis Ranjard, Melanie Hingston, Kevin Chang, Helen Shearman, Vicky Fan, Shan Wong, William Hu, Denise Kühnert, Jessie Wu, Walter Xie and Steffen Klaere. You guys have been great nakama and have made my academic career thus far as enjoyable as it has been. Additional thanks goes to everyone else at the Bioinformatics Institute, Computational Evolution Group, School of Biological Sciences, Department of Computer Science and Biomatters.

For helping me out with clocking the 50 CPU years or some ridiculous length of computational time that I have used for this project, I thank Yvette Wharton, Peter Tsai and the staff at BeSTGRID. For funding this project, I would like to express my gratitude to Biomatters Ltd. and the Foundation for Research, Science and Technology New Zealand.

I would like to sincerely thank all my friends and family for their support and encouragement over the years, especially to those that have passed away during this

journey. Special thanks go to Akshay Girdhar and Ben Allen for 10 plus years of friendship, not cutting me any slack for being at university for over 8 years and providing beer money for a poor student. Credit also goes to Evelyn Kiing, Kelly Nguyen and Nicholas Lee for amazing tramping, beach and ski trips.

I would like to thank everyone at Aikido Shinryukan and University Judo, especially my senseis Nobuo Takase and Rick Littlewood for teaching me important virtues in life.

For inspirations in life and entertainment over the last few years, thanks goes to Wasalu Muhammad Jaco, Jay Chou, Nasir Jones, Tupac Shakur, Vincenzo Luvineri, Eiichiro Oda, Masashi Kishimoto, Morihei Ueshiba and Shinya Aoki.

Much appreciation and credit goes to my parents Louisa and Jeff Li for teaching me all that I know. I thank you guys for being my role models in life and molding me into who I am today. You guys sacrificed a lot for me and I would not be here writing this otherwise. Mom- thanks for looking after me and keeping me in line over the years. Dad- thanks for leading by example and having a strong set of morals look up to.

A special thank you goes to my wife Evelyn Kiing for sticking around and supporting me, teaching me more about myself than I even knew, sharing ideas about life and keeping a smile on my face.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

The discovery of the structure of DNA and molecular biology (WATSON and CRICK 1953) has lead to the now apparent link between genetics and physiology. Understanding the genetics of an organism is therefore essential for comprehending its biology. Consequently, there has been considerable progress in methods to sequence genetic data (MAXAM and GILBERT 1977; SANGER *et al.* 1977). Over the last decade in particular, significant developments have been made in terms of large-scale retrieval of molecular sequence information (MARDIS 2008; MARGULIES *et al.* 2005b). This advancement in technology has made it now possible to obtain the complete genomes of organisms. In view of this, the current phase has been appropriately coined the "–omic era" in biology (EVANS 2000).

These developments have motivated a number of studies to produce effective analyses for genomic data. The increases in computational time and error rates associated with methods of analyses have become relevant issues due to the growth of sequence data. Given the size of modern datasets, traditional methods of analysis do not perform within reasonable time frames and the number of false positives generated by these methods is in practice unmanageable. Hence there is a demand for more efficient and less error-prone methods to analyse the data.

One area of interest is the link between genomes and the process of evolution. The current understanding of molecular evolution has enabled researchers to relate differences in molecular sequence to the original theories of evolution (DARWIN 1859). Unlike the knowledge available during Darwin's time, biologists now know that variation in phenotypes arises from mutations in the genetic composition of an organism. Selective pressures act on phenotypes that cause a fitness advantage or

disadvantage and subsequently affect the persistence of the mutations at the molecular level. The use of molecular sequence data has helped further understand evolution and selection. In turn, this knowledge has been used to answer other questions in biology.

In this thesis, I include a compilation of studies which describe methods and ideas for improving genome-scale analyses. Using phylogenetics as a foundation, I attempt to improve current methodologies and develop novel techniques for genomic studies, as well as develop a better understanding of the underlying forces of genome evolution. This chapter serves as an introduction to the respective disciplines of genomics and phylogenomics, and how the understanding in rates of molecular evolution has facilitated their progress. I discuss the past, present and future of genomic analyses, and the relevance of my research to the field. Background information is provided to better understand the motivations and concepts behind the work carried out in the following chapters.

1.1 Genomics

In 2001, the sequencing of the human genome marked a significant milestone for the field of biology. At the time, this project in its entirety took over 10 years to complete (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM 2001). Over the decade that followed, major improvements in DNA sequencing technology were made which greatly shortened the time and cost required for sequencing genomes (VENTER *et al.* 2001). In recent years, "next-generation" sequencing technologies have revolutionised sequencing by dramatically improving the associated speed and cost (MARDIS 2008). These new technologies such as Roche 454, Illumina Solexa and Applied Biosystems SOLiD (MARDIS 2008) enable millions of sequence reads to be processed at a time, compared to 96 in traditional Sanger sequencing (SANGER *et al.* 1977). Even more recently, there have been discussions of 3rd generation sequencing using nanopore technology, which promises to further exceed the output of current techniques (BRANTON *et al.* 2008; CLARKE *et al.* 2009; RUSK 2009; TSUTSUI *et al.* 2010; WANUNU *et al.* 2010). As a result of these advances in sequencing technology, the amount of molecular data in repositories is expected to continue its trajectory of exponential growth (BENSON *et al.* 2009).

Genomic analysis has applications in almost all disciplines of biology and has now become an integral component of systems biology (GE *et al.* 2003), functional biology (MOROZOVA and MARRA 2008), evolutionary biology (CHARLESWORTH *et al.* 2001) and health sciences (BALMAIN *et al.* 2003; COLLINS *et al.* 2003; COLLINS and MCKUSICK 2001; REIS-FILHO 2009). In order for genomics to encompass a broader range of applications it is essential to develop effective methods to analyse these overwhelmingly large datasets. Biologists have turned to analyses using computational approaches which are capable of managing such data. Challenges presented in the analysis of genomes include uncovering the underlying species relationships, locating genes, identifying regions of functional significance, and characterising patterns in genome evolution. The developments that have been made to address these issues are further discussed in Sections 1.1.1 and 1.3.

1.1.1 Comparative genomics

The availability of such comprehensive data has enabled comparisons between genomic datasets. Comparative genomics is the comparison of genomes between species to identify functional elements and factors that have shaped the evolution of genomes. An overview of the historical developments of comparative genomics is given in this section.

1.1.1.1 Identifying functional regions

The first major comparative genomic study [1] was carried out shortly after the sequencing of the first bacterial genomes of *Haemophilus influenza* and *Mycoplasma genitalium* (FRASER *et al.* 1995). In this study, the authors identified the similarities between the two genomes in an effort to find the minimal set of genes required for survival. Following this study, more comprehensive comparative genomic studies were performed on bacteria (MCCLELLAND *et al.* 2000) and yeast (CLIFTEN *et al.*

[1] Although comparative genomic studies between the earliest sequenced genomes of viruses were performed, these will not be discussed as these genomes are comparably small. My interest in this thesis lies in the comparison of large genomic datasets.

2001) with the objectives of characterising the evolution across genomes and identifying the functional elements in the non-coding regions of the genomes, respectively.

In the comparative study of the yeast genomes (CLIFTEN *et al.* 2001), the authors identified functional elements through sequence conservation. Evidence of correlation between functional regions and sequence similarity was first formally identified by Tagle *et al.* (1988) in the embryonic genes of prosimians, though this relationship had long been hypothesised (KIMURA 1967; KIMURA 1983). Regions of the genomes that are responsible for essential "housekeeping" functions will likely be conserved across species, as these regions are subject to functional constraint and so mutational changes that occur may disrupt the basic survival of an organism (JUKES and KIMURA 1984; MIYATA *et al.* 1980). The central idea behind this approach is that in the midst of this vast genomic data, differences and similarities between species indicate biological relevance in terms of areas undergoing selection and locations of functional elements. The theory behind these approaches is discussed in detail in Section 1.4.3.2.

It was not until the sequencing of the mouse genome that comparative genomic analyses were performed on large mammalian genomes (INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM 2002). As part of the initial effort to study the mouse genome, a comparative study was done between the mouse and human genomes. A major finding from the human-mouse comparison was that while only 1.5% of the human genome is estimated to be protein coding, roughly 5% of the genome appears to be undergoing purifying selection (INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM 2002). This discovery provided the first definitive indication of the importance of non-coding elements to species function. Consequently, this has alerted researchers to the existence of non-coding regulatory elements in the genome and the extent to which they affect gene expression (DUBCHAK *et al.* 2000; HARDISON 2000; INADA *et al.* 2003; INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM 2002; JOHNSON *et al.* 2004; KEIGHTLEY *et al.* 2005; LOOTS *et al.* 2000; XUE *et al.* 2004). Comparative genomic approaches have since successfully been used to predict functional elements across genomes, such as regulatory elements, genes and chromosomal organisation elements (INTERNATIONAL

MOUSE GENOME SEQUENCING CONSORTIUM 2002; LOOTS *et al.* 2000). Since this finding, a large number of studies have used this approach to identify highly conserved non-coding elements across the genomes of various species groups (BEJERANO *et al.* 2005; GLAZOV *et al.* 2005; KELLIS *et al.* 2003; LEVY *et al.* 2001; PENNACCHIO and RUBIN 2001; SIEPEL *et al.* 2005; STEIN *et al.* 2003; THOMAS *et al.* 2003).

1.1.1.2 Characterising genome evolution

Comparative genomics studies have helped characterise events and patterns that occur in genome evolution. A notable study done recently on genomes of Drosophila species (STARK *et al.* 2007) found signatures of evolutionary change for various functional elements such as protein-coding gene regions, RNA genes, and regulatory motifs. The findings in this study have been helpful towards understanding the evolution of different elements in the genome. A number of other studies have made progress in identifying a range of attributes of genome evolution. Efforts have been taken, for example, to identify the extent to which selection occurs across the genome (INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM 2002), to denote differences in function through gene preservation across species (KELLIS *et al.* 2003; STEIN *et al.* 2003), to determine the difference in rate of substitution between coding and non-coding regions (SUBRAMANIAN and KUMAR 2003) and to reconstruct species relationships (DELSUC *et al.* 2005; SJOLANDER 2004).

1.1.1.3 Computational improvements for comparative genomics

In comparative genomics where large genome-scale datasets are under consideration, the use of computationally-intensive, traditional methods of analysis are impractical. Efficient approaches to address standard problems are a necessity to cope with the increase in data.

A notable example is the development of sequence alignment methods for genomes. Alignments are used to identifying the mutation events that occurred throughout evolutionary descent and are crucial for comparative analyses. Several algorithms have been proposed which have proved to be of practical speeds for genome-scale

alignments. A study by Kent *et al.* (2002) altered the physical computation of basic alignment methods to improve efficiency in large datasets by storing frequently used data in the memory. In contrast, Blanchette *et al.* (2004) used a heuristic approach which computes alignments as sequence blocks, effectively sacrificing alignment accuracy for increased speeds.

Apart from software, the problem has also been overcome by developments in hardware usage. Traditional alignment algorithms such as Smith-Waterman (SMITH and WATERMAN 1981) and CLUSTAL (THOMPSON *et al.* 1994) have recently been ported to platforms that can utilise graphics processing units (GPUs) for computation (LIU *et al.* 2006; MANAVSKI and VALLE 2008). These platforms allow the parallelisation of the alignment computation onto a number of smaller processors on the GPU, effectively reducing alignment times.

These studies have thus demonstrated the need for the development of highly efficient methods for comparative genomics. This objective of developing high-throughput methods is adopted in the work performed in Chapter 5.

1.2 Phylogenetics

Phylogenetics is the study of evolutionary relationships within a set of organisms. The use of phylogenetics dates back to the earliest evolutionary studies by Darwin (1859). Phylogenetics relies on the concept that characteristics of lineages change over descent and hence, the similarities shared between lineages are most likely a reflection of their evolutionary history (SOKAL and MICHENER 1958). Phylogenetics has applications in many areas in biology; including systematics (QUEIROZ and DONOGHUE 1988), population genetics (NEE *et al.* 1995), conservation biology (BARKER 2002; MOULTON *et al.* 2007), biogeography (KNOWLES 2009), palaeontology (BUNCE *et al.* 2009; ORLANDO *et al.* 2009; WRAY *et al.* 1996) and systems biology (MEDINA 2005). It has also been proven as a useful tool outside of biology, with applications in linguistics and cultural evolution (GRAY and ATKINSON 2003; GRAY and JORDAN 2000). Several of these applications of phylogenetics are discussed in more detail in Sections 1.3 and 1.4.

Darwin noted in his book *On the Origin of Species* (DARWIN 1859) that the relationships between species can be represented as a tree diagram. Under Darwin's scheme, the union of two branches on a tree represents the common ancestor between the two lineages; hence the overall tree topology (shape of the tree) represents the evolutionary relationships amongst all the taxa. While not specified in Darwin's time, the lengths of the branches have since come to represent a measure of distance or divergence from an ancestor to its descendant (SOKAL and MICHENER 1958).

1.2.1 Molecular phylogenetics

Initially, when molecular sequence data were not understood, phylogenetics was performed using data on the presence or absence of morphological traits between species. Since the genesis of the structure of DNA (WATSON and CRICK 1953) and the development of sequencing technologies (SANGER *et al.* 1977), molecular sequences have become a common source of data for phylogenetic comparisons. In molecular phylogenetics, the history of evolutionary descent can be reconstructed using molecular sequence data from DNA, RNA, codons and amino-acids. The most common pathway is to first obtain homologous sequences for a group of taxa then generate a sequence alignment, which indicates the evolutionary changes that the sequences have accumulated over time. In a sense, a sequence alignment is a reconstruction of the mutation events (substitutions, insertions and deletions) that have been fixed over the descent of the taxa concerned. The phylogeny among the taxa can then be estimated by tracing back the events of mutation and reconstructing the most likely evolutionary history, given the data observed (FELSENSTEIN 2004).

1.2.2 Modern phylogenetics

Given the number of taxa on a tree n, the number of possible labelled rooted tree topologies is equal to:

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{1.1}$$

This means that with a linear increase in the number of taxa, the complexity of tree space for tree topology reconstruction increases exponentially. Besides topology, the length of each branch on the tree has to be estimated. Complex approaches are thus required to ensure that the tree obtained is a sensible estimate of the true tree shared between the data.

The idea of parsimony methods was first proposed by Edwards and Cavalli-Sforza (1963) which reconstructed the phylogeny based on the minimum number of evolutionary changes required to explain the variation. However, Felsenstein (1981) was the first to establish the proper framework for tree likelihoods, which are commonly used in modern phylogenetics. An advantage of using likelihood-based approaches is the ability to parameterise models of evolution. Models of evolution attempt to capture the features of the underlying processes of evolution which occurred in order for a given dataset to be observed. Such models include models of the substitution process in nucleotide (for example, FELSENSTEIN 1981; HASEGAWA *et al.* 1985; KIMURA 1980; TAVARÉ 1986), amino-acid (ADACHI and HASEGAWA 1996; DAYHOFF *et al.* 1978; JONES *et al.* 1992; WHELAN and GOLDMAN 2001) and codon (GOLDMAN and YANG 1994; MUSE and GAUT 1994) sequences, models of rate heterogeneity across sites (REEVES 1992; YANG 1994), and rate heterogeneity across branches (for example, RAMBAUT and BROMHAM 1998; SANDERSON 1997; THORNE *et al.* 1998). Tree likelihoods are computed by calculating $P(D|\tau,\theta)$, the probability of observing some data (D), given the tree (τ) and other associated parameters of the models (θ) (FELSENSTEIN 1981).

1.2.2.1 Maximum likelihood

Apart from methods of tree likelihood computation, Felsenstein proposed mechanisms to search through tree space for the tree that has the highest probability of generating the observed data (FELSENSTEIN 1973; FELSENSTEIN 1981). Maximum likelihood (FELSENSTEIN 1981) is the most commonly used method for phylogenetic estimation of molecular data. Maximum likelihood approaches work by searching through tree space and finding the single most probable tree, precisely, the tree that maximises the likelihood. Variants to Felsenstein's algorithm have since been developed to

heuristically reduce the computational time associated with tree searching (for example, DEMPSTER *et al.* 1977; FARRIS 1972; GASCUEL 1997; HENDY and PENNY 1982; HENDY and PENNY 1989).

1.2.2.2 Bayesian phylogenetics

More recently, methods have been developed for using Bayesian inference to estimate phylogeny (MAU and NEWTON 1997; MAU *et al.* 1999; RANNALA and YANG 1996; YANG and RANNALA 1997). Bayesian phylogenetics differs from maximum likelihood in that a set of credible trees is used to infer the true tree instead of solely using the most probable tree. Bayesian inference of phylogeny is implemented using the statistical technique of Markov-chain Monte Carlo (MCMC) which searches through tree space and samples trees and values of parameters.

As a result of Bayes' theorem the posterior probability (the probability given a set of conditions) of a tree and its parameters can be expressed as:

$$P(\tau,\theta \mid D) = \frac{P(D \mid \tau,\theta) \cdot P(\tau,\theta)}{P(D)} \tag{1.2}$$

$P(D \mid \tau,\theta)$ again represents the likelihood function. $P(\tau,\theta)$ is the prior probability distribution, which represents *a priori* knowledge of the model parameters and the evolution processes. $P(D)$ is the probability of the data, which in general is unknown but is not compulsory for the phylogenetic estimation, in which case the equation can be expressed as:

$$P(\tau,\theta \mid D) \propto P(D \mid \tau,\theta) \cdot P(\tau,\theta) \tag{1.3}$$

Trees and parameters are sampled by the MCMC in the posterior probability distribution. Whether a proposed tree or parameter in the MCMC is accepted is determined by the Metropolis-Hastings algorithm (HASTINGS 1970; METROPOLIS *et al.* 1953). For each tree or parameter proposed in the MCMC, the likelihood of the new parameter is calculated. The probability of accepting a parameter change α is

proportional to the improvement in likelihood over the current state. The probability of accepting a new parameter value θ^* is calculated as:

$$\alpha = \min\left(1, \frac{P(D \mid \tau, \theta^*)P(\tau, \theta^*)}{P(D \mid \tau, \theta)P(\tau, \theta)} \cdot \frac{P(\theta \mid \theta^*)}{P(\theta^* \mid \theta)}\right) \tag{1.4}$$

The first component of the probability calculation is derived from the Bayes' theorem equation in Equation 1.2 and corresponds to the change in likelihood from the previous parameter value. The second component is the proposal ratio which calculates the probability of the move itself.

What is not shown in this equation are the proposal kernels associated with each parameter. Proposal kernels are mechanisms that specify how changes in parameters are proposed during the MCMC run. Proposal kernels do not directly affect the calculation of likelihoods themselves. Effective proposal kernels enable more effective traversal of parameter space.

As the Metropolis-Hastings algorithm tends to sample more probable states, it would be expected that a high percentage of samples will have parameter values close to the true distribution, so if the overall distributions of the tree and model parameters in the posterior samples are examined, reasonable estimates of the target distribution can be obtained (HUELSENBECK and RONQUIST 2001; MAU and NEWTON 1997).

Bayesian MCMC-based approaches provide certain advantages over maximum likelihood for estimation of phylogeny. As Bayesian MCMC produces a set of trees sampled from the posterior distribution, it provides an explicit estimate of credibility for the tree and associated parameter values. Another benefit of Bayesian inference is the ability to specify prior probability distributions which represent prior knowledge and assumptions about the data at hand. These priors can improve the estimates made by the MCMC by providing guidelines for likely values of the parameters. Priors can also be used to limit parameter values to sensible estimates and to improve the convergence of the MCMC-chain. However, priors have to be specified with caution as strong priors can lead to bias in the posterior, producing misleading estimates. The

added benefits provided by Bayesian estimation of phylogeny has been an influential factor in its increase in popularity which now rivals that of maximum likelihood approaches (BEAUMONT and RANNALA 2004).

1.3 Phylogenomics

Phylogenomics is the intersection between phylogenetics and genomics. In essence, phylogenomics is the application of phylogenetic inference to genomic data. By incorporating information on the underlying relationships of the species examined, more sensible inferences can be made when comparing their genomes. In contrast to direct comparative genomic analyses, phylogenomic approaches enable comparison among a subset of related species without having to consider all of the species at hand. As a result, the power to detect evolutionary changes is better under these approaches. The major developments in phylogenomic methods that are relevant to this thesis are discussed in this section.

1.3.1 Characterising genome evolution

Early work on phylogenomic analyses was done by Eisen *et al.* (1997) which focused on the application of phylogenetics to detect occurrences of gene duplication and gene loss across the genomes of gut bacterias *Escherichia coli* and *Helicobacter pylori*. This study was the first of many utilities of phylogenomic approaches to identify events that occur in genome evolution. Methods so far have been developed to detect gene duplication (EISEN 1998; ZMASEK and EDDY 2001), horizontal gene transfer (PAZOS and VALENCIA 2001; WHITAKER *et al.* 2009), gene relocations (BHUTKAR *et al.* 2007), sequence homology (DROSOPHILA 12 GENOMES CONSORTIUM 2007) and recombination (GÜRTLER 1999). In particular, Bhutkar *et al.*'s study (2007) on identifying gene relocations found that using a phylogenomic approach permitted the detection of lineage-specific gene relocations which would otherwise be overlooked by pure comparative genomic approaches. Identification of these genome evolution events is often difficult without the use of phylogenomic methods as the direct comparisons of sequences does not allow for the lineage-specificity of evolution patterns.

1.3.2 Covariation of genome evolution across species

Another study by Eisen (1998) outlined a procedure for detecting changes in a gene's function by examining the genetic distances on a phylogeny. The basis for this approach is that if a gene has the same function as its homolog in another species then their sequences should be similar and thus the genetic distance would be low. This relationship between sequence homology and gene function was further explored by Pellegrini *et al.* (1999), who developed a method to predict the biological pathway involvement of proteins. This method, now called phylogenetic profiling, examines the homology of a protein across a set of taxa to predict its pathway involvements based on similarity to the profiles of each biological pathway.

Apart from correlations in gene homology among functionally related proteins, Fryxell (1996) and Pazos *et al.* (1997) also noted the correlations in the gene tree branch lengths of physically interacting proteins. These observations lead to the development of phylogenomic techniques, analogous to phylogenetic profiling, which use this covariation in gene tree branch lengths to predict protein-protein interactions (JUAN *et al.* 2008a; PAZOS and VALENCIA 2001) and even entire interactomes (JUAN *et al.* 2008b; PAZOS *et al.* 2005). This relationship between rate of substitution and gene function is of focus in Chapter 4 and is further elaborated in Section 1.4.3.1.

1.3.3 Identifying selection and functional constraints

A common challenge in comparative genomics has been the identification of functionally important elements of the genome (See Section 1.1.1). Several extensions have been made to comparative genomic approaches to address this problem using phylogenomic techniques. Siepel *et al.* (2005) proposed an algorithm to identify highly conserved elements in a set of genomes using a hidden Markov model which considers the underlying species phylogeny. Siepel and colleagues further pursued this direction by allowing for lineage-specific changes in sequence conservation (SIEPEL *et al.* 2006). Lineage-specific changes in conservation are of interest in biology as they denote differences in the underlying functions within the organism. Where sequence conservation is present in one lineage but not another, functional

constraint has been relaxed in the latter group. Relaxed functional constraints suggest a lack of purifying selection in the biological function that the region serves. The detection of lineage-specific rate changes is of interest in Chapter 5.

Phylogenomic approaches were also predominantly used in the initial studies of the first 12 Drosophila genomes (DROSOPHILA 12 GENOMES CONSORTIUM 2007; STARK *et al.* 2007). In these studies, the analyses were performed with the assumption of a known underlying phylogeny shared between the species across all genomic regions. The gene homology, selection in genes (DROSOPHILA 12 GENOMES CONSORTIUM 2007), and genome conservation (STARK *et al.* 2007) were examined across the species. By assuming a species topology, the authors were able to find occurrences of these evolutionary events that were only specific to certain lineages. These studies of Drosophila found that assuming the species phylogeny produced more precise assessments of the evolution and selection in these genomes.

1.3.4 The genome tree problem

Phylogenomic methods have also been applied to improve the resolution of species phylogenies. Species phylogenies are of central interest in biology as they provide a better understanding of the evolution of species and the relationships among them. In early studies of species relationships, attempts at species tree estimation were computed using small gene sets (for example, COLLINS *et al.* 1994; GUADET *et al.* 1989). However, it has been shown that incongruence often exists in the species relationships and the ability to resolve the correct species topology depends on the set of genes used (GIRIBET *et al.* 2001; HWANG *et al.* 2001; KOPP and TRUE 2002; LÖYTYNOJA and MILINKOVITCH 2001; MASON-GAMER and KELLOGG 1996; ROKAS *et al.* 2003a).

Brown *et al.* (2001) first proposed the idea of using concatenation of alignments from homologous genes to estimate the underlying species phylogeny. This approach effectively reduces the effects of stochastic bias on the tree estimation. Although their study highlighted a promising methodology, the authors only examined 23 genes across 45 species and thus the results were fairly inconclusive. Later, Rokas *et al.*

(2003b) showed, by using a set of 106 genes across seven yeast species, that the phylogeny can be better resolved using genome-scale gene concatenations.

This phylogenomic approach taken by Rokas et al. was later applied to a number of datasets to better resolve the species relationships in animals (PHILIPPE *et al.* 2005a; ROKAS *et al.* 2005), fungi (JAMES *et al.* 2006), bacteria (BROCHIER *et al.* 2002) and plants (QIU *et al.* 2006). More recently, similar approaches have been applied to resolving the important issue of the genome tree of life (CICCARELLI *et al.* 2006; DELSUC *et al.* 2005). Apart from the incongruence in species phylogeny, the genetic distance between species can also be better approximated using genome-scale analyses.

Although Rokas *et al.* (2003b) and others have demonstrated the capabilities of this approach, studies have showed that the method still suffers from systematic biases and is sensitive to the choice of models (HOLLAND *et al.* 2006; PHILLIPS *et al.* 2004; RODRIGUEZ-EZPELETA *et al.* 2007). As a result, alternatives to solving the problem of tree reconstruction using multi-gene analysis have been proposed and research in this area is ongoing (HELED and DRUMMOND 2010; HOLLAND *et al.* 2006; LEIGH *et al.* 2008; LIU and PEARL 2007; PHILLIPS *et al.* 2004).

Apart from distance-based estimates of phylogeny, genome trees have been reconstructed using other measures of genome similarity. So far, genome trees have been estimated on the basis of the consensus trees of individual gene trees (SICHERITZ-PONTEN and ANDERSSON 2001), evolutionary distances between orthologous sequences (GRISHIN *et al.* 2000; WOLF *et al.* 2001), gene order (WOLF *et al.* 2001) and gene loss across species (LIN and GERSTEIN 2000; NATALE *et al.* 2000).

In summary, phylogenomic approaches can be used to enhance comparative genomic studies and overcome the limitations posed. The worth of such approaches has been demonstrated here by the wide range of applications to which they have been applied to. The potential benefits such methods provide make phylogenomics a valuable tool for comparative genomic studies. In this thesis, the phylogenomic problems that I will focus on are the estimation of species phylogenies and distances, characterisation of

rates of substitution across genomes and the identification of changes in rate of substitution across species.

1.4 Rates of molecular evolution

In the early stages of evolutionary genetics, it was believed that mutations that occurred were mostly fixed in the population of a species. It was not until Kimura proposed the neutral theory of evolution (KIMURA 1968; KIMURA 1983) that biologists began to understand that not all mutations are fixed in the population as substitutions. The difference between molecular mutation rate and substitution rate is dependent on many factors, including the rates of mutation (BAER *et al.* 2007), metabolic rates, body sizes, generation times (BRITTEN 1986; MARTIN and PALUMBI 1993), population sizes and structures (OHTA 1972), selectional constraints (JUKES and KIMURA 1984; MIYATA *et al.* 1979), and genetic drift (KIMURA 1968). When deviation from the neutral rate of substitution occurs, it can generally be expected that this deviation is the result of one or more of these factors.

1.4.1 Gene-specific versus lineage-specific variation in rate

The rate of molecular evolution varies greatly across different regions of a genome (SUBRAMANIAN and KUMAR 2003) as well as across different species (GAUT *et al.* 1992). One way to therefore consider the variation in rate of evolutionary change is to partition the variation into two categories: (1) gene-specific variation, which refers to the variation in rate of substitution in a particular gene (or homologous region of the genome) that is shared across a set of species; and (2) lineage-specific variation, which refers to a change in overall substitution rate of a species or set of monophyletic species.

Gene-specific variation in rate can occur as a consequence of variation in sequence conservation along the genome. Such conservation is generally linked to the functional constraints of the sequence, especially where the sequence codes for an important function that is crucial for the well-being of the species (JUKES and KIMURA 1984; MIYATA *et al.* 1980). This effect has been repeatedly demonstrated, whereby

genes that code for important functions are strictly conserved and highly invariable (HEISS *et al.* 1998; TAGLE *et al.* 1988; WOOLFE *et al.* 2005). In contrast, regions of the genome where the conservation is relaxed will evolve at the neutral rate of substitution. Adaptive evolution is also known to occur, where changes in a particular function are promoted and the substitution rate is faster than neutral (for example, MONDRAGON-PALOMINO *et al.* 2002). A notable example of gene-specific effects is the development of drug resistance in the human immunodeficiency virus (HIV) within a host, where mutations allowing HIV to evade immune responses are more rapidly fixed in the host population than when under neutral selective pressures (KELLEHER *et al.* 2001; OGG *et al.* 1998; PHILLIPS *et al.* 1991; RAMBAUT *et al.* 2004; ROSS and RODRIGO 2002; WEI *et al.* 2003; YUSIM *et al.* 2002).

Lineage-specific variation in rate of substitution is known to arise from differences in the natural characteristics of a lineage. In the early days of molecular biology, the idea was proposed that the rates of substitution were uniformly distributed across lineages (SIBLEY and AHLQUIST 1984; ZUCKERKANDL *et al.* 1965). This idea of a "molecular clock" was, however, the subject of much criticism (for example, KIMURA 1983; OHTA 1972; OHTA 1987). The molecular clock hypothesis was further denounced by evidence from Bonner *et al.* (1980) and Britten (1986) which respectively indicated variations in substitution rate across the lineages of the primate species and across different taxonomic groups in mammals.

Britten's (1986) study on mammalian rate variation highlighted multiple factors that affected the rate of substitution across lineages. A species that has low DNA repair efficiency (thus high mutation rate) is likely to have higher substitution rate if strong selective pressures are absent. On the other hand, differences in population size and structure influence the fixation rate of a given mutation as a substitution. For instance, where a species has a small population size, fixation of mutations is likely to be much faster. Martin and Palumbi (1993) further provided mechanisms for rate heterogeneity caused by differences in the physiological properties of a species. Martin and Palumbi noted that differences in metabolic rates, body sizes and generation time lead to discrepancies in the rate at which mutations occur over a time interval. Since these developments, a large number of studies have also demonstrated the presence of

lineage-specific effects. Rate heterogeneity across lineages has been well characterised across a range of species such as mammals (WU and LI 1985), plants (GAUT *et al.* 1992) and RNA viruses (JENKINS *et al.* 2002; KORBER *et al.* 1998).

Besides the individual contributions of these two effects, heterogeneity can stem from lineage-gene-specific variation in rates. Lineage-gene-specific effects refer to variation in the rate of substitution that occurs in a particular sequence region on a particular lineage. Blundell and Wood (1975) were one of the first to note the presence of lineage-gene-specific rate changes. The authors found rapid rates of substitution in insulin genes of hystricomorph rodents and guinea pigs. This acceleration in rate is attributed to a change in function and how insulin operates in these rodents (BLUNDELL and WOOD 1975; STEINER *et al.* 2003). In general, lineage-gene-specific rate changes arise from selection that is specific to a particular lineage. Such selection events correspond to biological functions that are important only to that lineage.

1.4.2 Why is rate of evolution interesting?

In the previous section, known factors contributing to variation in the rate of substitution were summarised. For any given sequence, the rate at which it undergoes substitution is a product of how different factors have acted together on the sequence. By partitioning the rate and isolating the effects that have led to rate heterogeneity in the data, it is possible to determine the underlying factors that caused the rates.

Through identifying lineage-specific changes in rates, improvements can be made to the estimation of divergence times in evolutionary history. The earliest estimates of divergence time by Zuckerkandl *et al.* (1965) provided time estimates under an assumption of constant rate across branches. However, benchmark studies by Drummond *et al.* (2006) and Ho *et al.* (HO *et al.* 2007a; HO *et al.* 2005; HO *et al.* 2007b) have demonstrated that improvements in divergence time estimations can be made by taking into account the variation in rate of substitution across lineages. This point is emphasised as divergence time estimation is of considerable focus in ancestral studies and palaeontology (for example, BUNCE *et al.* 2009; ORLANDO *et al.* 2009;

WRAY *et al.* 1996). Also, as divergence time and population parameters are confounded, better estimates of divergence time can effectively improve the estimates of ancestral population sizes and species mutation rates (KIMURA 1983; NACHMAN and CROWELL 2000; RANNALA and YANG 2003; TAKAHATA 1986; TAKAHATA *et al.* 1995; YANG 1997). Debruyne *et al.* (2008) demonstrated this point by co-estimating the rates and the population parameters of divergence times, population sizes over time and migration patterns in the DNA of ancient woolly mammoths. The estimation of lineage-specific rates is explored in Chapters 2 and 3.

Gene-specific effects provide signatures of selection that occur in unison across species. On the other hand, lineage-gene-specific rate changes indicate selection that occurs on a particular species or set of related species (NIELSEN 2005). Identifying selection is useful as regions of a genome that undergo selection are of functional importance to the species (HARDISON 2003). This relationship between selection and function is explained below.

1.4.3 The relationship between function, selection and rate

The relationship between the rate of substitution and gene regions that are functionally important was proposed by Kimura (1983) in his work on the neutral theory of evolution. The physical attributes of an organism are determined by the sequence composition of the functional elements in its genome. When mutations occur in a functional element, they can alter the expressed phenotype of certain attributes. If the new phenotype causes a disadvantage towards survival or reproductive success, the individual is considered less fit and the mutation is less likely to be passed down to the next generation. Accordingly, these deleterious mutations have a lower chance of being fixed in the population. On the other hand, if the mutation provides beneficial change then it is more likely to replace the existing alleles in the population (KIMURA 1983).

Ultimately, what is observed is a correlation between the genomic regions of a species that are undergoing selection and regions which contain functional elements. Areas

undergoing strong selection often indicate regions of high functional importance for the species.

1.4.3.1 Co-evolution of substitution rates in genes

For a given biological process, gene products (such as proteins and regulatory elements) work cooperatively to perform the necessary tasks required for the process to be undertaken. Interactions occur between these gene products to ultimately form a pathway of interactions. Interestingly, a large number of studies have observed covariation in rates of substitution between genes that have physically interacting gene products (ATWELL *et al.* 1997; JUCOVIC and HARTLEY 1996; PAGÈS *et al.* 1997; PAZOS *et al.* 1997; POUMBOURIOS *et al.* 2003). Although this pattern has been observed in many studies, the mechanism behind this co-evolution is currently still unclear. So far two competing hypotheses have been proposed. The first hypothesis is that as interactions take place between the gene products of these genes, their binding sites must be preserved in order for a function to be maintained. When a mutation occurs in one gene in an interacting pair, changes in the physical structure of the gene product may arise which can affect the physical binding surface required for interaction (FRYXELL 1996). Compensatory mutations must therefore take place in the pairing gene in order for the physical interaction to be maintained. Under this hypothesis, the presence of compensatory mutations is the cause of correlated rates of substitution between these physically interacting genes.

The second and more simple hypothesis states that as these genes that have interacting gene products all contribute towards a particular biological function, selection pressures acting on a particular function will impact the rate of substitution across a number of genes involved in that function (HAKES *et al.* 2007). Therefore, selective pressures that act on functionally related genes cause these observed similarities in rate of substitution. So far, there is evidence to support both these hypotheses and it has been suggested that the observed co-evolution is a product of both effects (JUAN *et al.* 2008a; KANN *et al.* 2009).

Regardless of the underlying cause, these patterns of co-evolution are known to exist across species. Coevolving genes will tend to covary in their rates at different lineages due to their functional coupling, so that one gene cannot easily undergo many genetic changes without the other also. Therefore, physically interacting genes often covary across lineages. As a result of this covariation in rate, Fryxell (1996) was able to observe similarities in the gene tree branch lengths of genes that have physically interacting protein products. This pattern of gene tree correlation was used by Pazos *et al.* (1997), who outlined a method that predicted protein interactions based on pairwise similarities in gene trees. This approach has shown success in predicting protein interactions and modifications to this strategy have since been proposed (GERTZ *et al.* 2003; GOH *et al.* 2000; GOH and COHEN 2002; JUAN *et al.* 2008a; JUAN *et al.* 2008b; KIM *et al.* 2004; PAZOS *et al.* 2005; PAZOS and VALENCIA 2001; RAMANI and MARCOTTE 2003; SATO *et al.* 2003; TAN *et al.* 2004). The relationship between gene tree branch lengths and functional involvement of proteins is further explored in Chapter 4.

1.4.3.2 Detecting selection

Understanding the role of functional elements in an organism's genome provides insights into the biology and function of the organism. It has therefore been of considerable interest in computational biology to identify functionally important elements in the genome. As regions of the genome that have undergone selection often correspond to functional elements, a large number of studies have focused on detecting incidence of selection.

The most commonly applied technique to detecting selection in protein coding genes has been the use of the ratio between the rate of non-synonymous (dn) and synonymous (ds) substitutions. A remarkable study by Messier and Stewart (1997) first pointed out that gene variants of the primate lysosome gene that have an unequal number of *dn* to *ds* substitutions are subject to selection. This is because for amino-acid coding genes, only *dn* mutations affect the physical composition of the protein that it codes for, and so selection only affects these mutations. As *ds* substitutions

represent the rate of substitution under neutral selective pressures, any deviations in the proportion of *dn* to *ds* substitutions are indicators of selective pressures.

Using this observation made by Messier and Stewart, Yang (1998) developed a likelihood framework for detecting selection in nucleotide sequences of amino-acid coding regions. This procedure that Yang proposed was highly influential and its application spawned a large number of selection studies (for example, BARGELLONI *et al.* 1998; FORD 2001; HUTTLEY *et al.* 2000). However, as Yang's method examines the codon positions in genes, a limitation to his approach is its inapplicability to non-coding regions. As the importance of non-coding regions of the genome to biological function has become increasingly apparent, alternatives for identifying selection have been pursued.

More recently, studies have attempted to detect selection events through directly looking at gene-specific changes in substitution rate (for example, HARDISON 2000; INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM 2002; WONG and NIELSEN 2004). The principle of these studies is that under the neutral theory of evolution, if no selection is occurring, mutations are fixed into substitutions at a neutral rate (KIMURA 1968; KIMURA 1983). When selection takes place, it alters the rate of substitution by either (1) rejecting mutations as substitution in order to preserve function (purifying selection), or (2) rapidly accepting changes in order to promote change in function (positive selection). Regions of the genome that are functional are likely to have undergone selection and will have non-neutral rates of mutation (NIELSEN 2001). However, it should be noted that gene-specific variations in rates of substitution are also influenced by genetic drift and population effects. Variations in rate do not necessarily reflect selection, as signals of selection are confounded by these effects.

High-throughput methods for detecting selection through rate heterogeneity (for example, MARGULIES *et al.* 2005a; RIVAS and EDDY 2001; SIEPEL *et al.* 2006) have become increasingly popular due to the availability of genomes and the well established correlation between functional elements and variation in rate (GLAZOV *et al.* 2005; HARDISON 2000; INADA *et al.* 2003; JOHNSON *et al.* 2004; LEVY *et al.* 2001;

LOOTS *et al.* 2000; LUNTER *et al.* 2006). These methods are favoured over other methods of detecting selection as they can be applied globally to both coding and non-coding regions of the genome. Although high-throughput procedures generally sacrifice specificity for efficiency on large datasets, they are preferred over lab testing of the sequences due to the costly and time consuming aspects of this technique. In Chapter 5, extensions are made to these current methods of detecting selection through changes in substitution rate.

1.4.4 Estimating the rates of substitution across lineages

The process of molecular substitution is a progression over a period of time, in which changes occur gradually. However, what is actually observed through molecular sequences is the end products of the process. If we examine the phylogeny of a set of homologous sequences, the branch lengths correspond to the genetic distances between a node and its ancestor.

What biologists are often actually interested in though are the two variables of divergence time and rate of substitution. Separating the genetic distance into rate and divergence time is, however, a rather difficult task, as the two variables are confounded and manifest as the total genetic distance (THORNE and KISHINO 2002). Because of this difficulty, little attention has been given towards the modelling of lineage-specific rate heterogeneity across a tree. Until recently, most phylogenetics has been performed with the assumption that either (i) the product of rate and time (the branch length) is estimated independently for every branch or (ii) the rates across branches are constant, and therefore conform to a strict "molecular clock" (ZUCKERKANDL *et al.* 1965). However, when comparing species that have been separated for long periods of time the molecular clock hypothesis often does not hold (BRITTEN 1986; GAUT *et al.* 1992; JENKINS *et al.* 2002; KORBER *et al.* 1998; WU and LI 1985). Only in recent years, has there been development of relaxed molecular clock models, which allow separate modelling of divergence times and lineage-specific rate heterogeneity (HUELSENBECK *et al.* 2000; RAMBAUT and BROMHAM 1998; SANDERSON 1997; THORNE *et al.* 1998).

1.4.4.1 Relaxed molecular clock models

Relaxed molecular clock models relax the assumptions of the molecular clock hypothesis and allow rates to vary across branches. By modelling the process of rate of substitution and divergence time individually, these two effects that are confounded as the distance can be separated. The use of relaxed molecular clock models over strict molecular clock models has been proven in many cases to improve the accuracy of phylogenetic estimation and the estimation of divergence times (DRUMMOND and RAMBAUT 2007; HO *et al.* 2007a; HO *et al.* 2005; HO *et al.* 2007b). Under a Bayesian phylogenetic setting, rates can be sampled for each branch from a distribution which is assumed to be underlying distribution of rates. The probability of a branch having a particular rate is then determined by calculating its likelihood given the branch-rates distribution and other parameters (DRUMMOND *et al.* 2006).

Relaxed molecular clock models can be classified as uncorrelated or autocorrelated models. In uncorrelated models, rates are drawn from a distribution with a mean value that is a shared globally across the branches (RAMBAUT and BROMHAM 1998; SANDERSON 1997). Autocorrelated models on the other hand, come from the empirical observation that in some datasets the rate appears to be "inherited", as the rate of a branch is similar to its ancestral rate. Accordingly, in autocorrelated models, the mean of the distribution from which a rate is drawn from is dependent on the rate of its ancestral node on the tree (ARIS-BROSOU and YANG 2002; THORNE *et al.* 1998). These relaxed molecular clock methods are biologically relevant as they mimic the behaviour of changes in rate along a lineage.

In Chapters 2 and 3, the use of relaxed molecular clock models is investigated and several improvements to the area of relaxed phylogenetics are proposed.

1.5 Conclusion

In recent years, the importance of methods for genomic analysis has been emphasised. Phylogenetics is a general tool which can be applied to comparisons across multiple genomic datasets. Through using phylogenetic approaches, the differences in rates of substitution within and between genomes can be estimated. Development of such

phylogenomic methods is crucial towards grasping the evolutionary processes that occur in genome evolution. By further understanding rate variation within and across species, it is possible to improve existing techniques in phylogenomic inference.

1.5.1 Organisation of this thesis

The aim of this thesis is to improve the applications of phylogenomic methods to genome studies. Specifically, I attempt to accomplish this by further understanding differences in rates of molecular evolution. There are three essential focuses in this work: (1) to improve current methodologies for estimating the rates of substitution, (2) to broaden our knowledge of the underlying nature of rate heterogeneity, and (3) to develop phylogenomic techniques that draw inferences from these rate variations.

In Chapter 2, I propose methods for improving estimation of lineage-specific rates with relaxed molecular clock models. I introduce a novel computational approach for sampling rates across branches which improves efficiency over implementations currently used in Bayesian phylogenetic frameworks. I also propose a new distribution model, namely the inverse Gaussian distribution, for modelling the rates across branches in relaxed molecular clock models. My methods were benchmarked against other approaches for relaxed molecular clocks using a standardised mammalian dataset. Parts of this chapter have been accepted for publication in the international peer-reviewed journal, *Molecular Biology and Evolution*.

In Chapter 3, I extend the work carried out in Chapter 2 to describe a method for model averaging of relaxed molecular clock models using Bayesian stochastic search variable selection. This model averaging technique subsequently permitted the accurate calculation of Bayes factors for model selection. Parts of this chapter have been accepted for publication in the international peer-reviewed journal, *Molecular Biology and Evolution*.

In Chapter 4, I identify a pattern of across-species covariation in rates among genes that are functionally related. This pattern has implications towards phylogenetic estimation and analysis of phylogenomic data which I discuss in the chapter. Parts of

this chapter have been published in the international peer-reviewed journal, *PLoS ONE*.

In Chapter 5, I describe a high-throughput phylogenomic method for detecting selection by identifying gene-specific and lineage-gene-specific heterogeneity in rate of substitution. Its use is demonstrated on a simulated phylogenomic dataset.

In Chapter 6, I discuss the implications of my research to the present and future of phylogenomic analyses. I outline potential extensions to my study and possible directions for future work.

Chapter 2. Relaxed phylogenetics

Parts of this chapter have been published by the author as Li, W. L. S. and A. J. Drummond (in press). "Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics." *Molecular Biology and Evolution*.

2.1 Introduction

In phylogenetic reconstruction, there has been limited research in the modelling of lineage-specific heterogeneity in rate of substitution. To this day, much of phylogenetics is still carried out under the unconstrained phylogenetic model in which rates are not estimated independently of times and thus lineage-specific rates are permitted to have arbitrarily large variance among branches (GUINDON and GASCUEL 2003; HUELSENBECK and RONQUIST 2001; SWOFFORD 2003). Part of the reason for this is that rates and their respective divergence times manifest themselves only as a product (the branch lengths between the species) and these genetic distances are difficult to decode back into their component rates and times without auxiliary information.

Estimating rates of substitution across branches holds great biological relevance as it is crucial for estimating the divergence time when tracing back the evolutionary history of taxa (BUNCE *et al.* 2009; DRUMMOND *et al.* 2006; ORLANDO *et al.* 2009; WRAY *et al.* 1996). Also, rates themselves are indicators of differences between species in terms of the processes influencing their molecular evolution. Variations in rate of substitution across species are known to be caused by differences in population size and structure, generation time, body size, mutation rates, metabolic rates, and

selection (BAER *et al.* 2007; BRITTEN 1986; GILLESPIE 1991; GRAUR and LI 2000; MARTIN and PALUMBI 1993). Partitioning the genetic distances into divergence times and rates allows one to reconstruct the temporal aspect of evolutionary history and dissect the processes involved (ADACHI and HASEGAWA 1995; BUNCE *et al.* 2009; GLAZKO and NEI 2003; GU 1998; MARTIN and PALUMBI 1993). It should be noted that these approaches assume that rate variation among branches is independent to rate variation across sites, an assumption questioned by advocates of the importance of heterotachy (PHILIPPE *et al.* 2005b; ZHOU *et al.* 2007). Nevertheless, our focus in this research in on modelling rate variation among branches, and for our purposes this process will be assumed to be independent of rate variation across sites.

For many years, the phylogenetic community took the notion of branch lengths on a tree as a representation of distance between the species in units of substitutions per site (FELSENSTEIN 1981). However, a more biologically relevant way to consider the branch lengths is to treat the distances as the product of divergence time from the common ancestor, and the rate at which the substitutions occurred. The classical approach for rate/divergence time estimation is to force the rates to conform to a "molecular clock", which assumes that rates are equal across all branches on a tree (ZUCKERKANDL *et al.* 1965), though in reality these rates of substitution differ across species. Strict molecular clocks are therefore generally confined to analyses within a species or among a few closely related species.

As a result, the idea of relaxed molecular clocks has been developed, where the rates of substitution are permitted to vary across branches of the tree (RAMBAUT and BROMHAM 1998; SANDERSON 1997; THORNE *et al.* 1998). Relaxed molecular clock models have in recent years been accepted into the broader field of phylogenetics. This is mainly due to the biological relevance of these models, as it is well established that rates of substitution naturally vary across species and lineages (BRITTEN 1986; GAUT *et al.* 1992; WU and LI 1985). It has also been demonstrated that the use of relaxed molecular clock models can, in some circumstances, improve the accuracy of phylogenetic estimation (DRUMMOND *et al.* 2006; HO *et al.* 2007a; HO *et al.* 2005; HO *et al.* 2007b). Hence, relaxed phylogenetic methods are expected to improve

estimation of divergence times, and perhaps even the accuracy of estimated tree topologies.

Various relaxed molecular clock methodologies have been devised to date to account for this rate heterogeneity in phylogenetic reconstruction (ARIS-BROSOU and YANG 2002; RAMBAUT and BROMHAM 1998; SANDERSON 1997; THORNE *et al.* 1998; YODER and YANG 2000). Several attempts have been made to assess the performance of these existing relaxed clock models in dating speciation events (DRUMMOND *et al.* 2006; HO *et al.* 2007a; HO *et al.* 2005; HO *et al.* 2007b; LEPAGE *et al.* 2007) but few papers have addressed in detail the performance of relaxed clocks in the context of the co-estimation of the tree topology and the divergence times.

Although relaxed molecular clock models enable us to take into account some of the variability of substitution rates across species, there is no single model that is consistently the most appropriate across all datasets. In sequences where the substitutions conform to almost molecular clock-like behaviour, assuming a strict molecular clock can sometimes be more practical and accurate for the analysis (DRUMMOND *et al.* 2006; REAL *et al.* 2005). As the underlying processes of substitution that occurred vary from gene to gene, different relaxed molecular clock models may also be better suited to some datasets than others (DRUMMOND *et al.* 2006; HO *et al.* 2005). Therefore a logical question to ask is how important is having the correct choice of model when running phylogenetic analyses? And also which models are the most appropriate for a given dataset?

In this chapter we outline two improvements to the estimation of rates with relaxed clock models: (1) a novel computational implementation of relaxed molecular clock models which samples rates as quantiles on a distribution rather than directly sampling the rate, and (2) the application of the inverse Gaussian distribution as a branch-rates distribution model. We assess the ability of eight different relaxed clock models at estimating the true tree topology on an empirical dataset of 12 mammalian species sampled at 1056 gene loci (RANWEZ *et al.* 2007). Comparison of different models is generally difficult, as models are often implemented in different frameworks. To allow for comparison between different relaxed molecular clock

models, the models compared in this chapter were all implemented in the BEAST software package (DRUMMOND and RAMBAUT 2007).

2.2 Materials and Methods

2.2.1 The general relaxed phylogenetics framework

In relaxed phylogenetics, the rates and divergence times are co-estimated, in addition to the distribution parameters for a distribution from which the rates are drawn. As a derivation from Equation 1.2, using the properties of the joint conditional distribution, a relaxed molecular clock model can be expressed as:

$$P(\tau,\mathbf{r},\mathbf{t},\boldsymbol{\Omega},\theta \mid D) = \frac{P(D \mid \tau,\mathbf{r},\mathbf{t},\theta)\cdot P(\mathbf{r} \mid \tau,\mathbf{t},\boldsymbol{\Omega},\theta)\cdot P(\mathbf{t} \mid \tau,\theta)\cdot P(\tau)\cdot P(\theta)\cdot P(\boldsymbol{\Omega})}{P(D)} \qquad (2.1)$$

where $\mathbf{r}$ is a vector of rates of length $2n-2$, and n is the number of taxa such that $\mathbf{r}=\{r_1,r_2,...,r_{2n-2}\}$. The vector $\mathbf{t}=\{t_1,t_2,...,t_{2n-2}\}$ represents the divergence times. $\boldsymbol{\Omega}$ is a set containing the model parameters associated with the branch-rates distribution of the clock model and θ represents all other parameters used in the likelihood calculation, such as parameters for models of the sequence substitution process.

Values of $\mathbf{r}$, $\mathbf{t}$, $\boldsymbol{\Omega}$, θ and τ can then be sampled using the Markov-chain Monte Carlo (MCMC) algorithm. The acceptance probability for a proposed change in the states of the parameters is dependent on the parameter being changed. If τ^*, θ^*, $\boldsymbol{\Omega}^*$, $\mathbf{t}^*$ and $\mathbf{r}^*$ represent proposed new values for the parameters, the acceptance probabilities of a proposed move for the different parameters can be expressed as:

$$\alpha_1 = \min\left(1, \frac{P(D \mid \tau^*,\mathbf{r},\mathbf{t},\theta)\cdot P(\mathbf{r} \mid \tau^*,\mathbf{t},\boldsymbol{\Omega},\theta)\cdot P(\mathbf{t} \mid \tau^*,\theta)\cdot P(\tau^*)\cdot P(\theta)\cdot P(\boldsymbol{\Omega})}{P(D \mid \tau,\mathbf{r},\mathbf{t},\theta)\cdot P(\mathbf{r} \mid \tau,\mathbf{t},\boldsymbol{\Omega},\theta)\cdot P(\mathbf{t} \mid \tau,\theta)\cdot P(\tau)\cdot P(\theta)\cdot P(\boldsymbol{\Omega})} \times \frac{P(\tau,\theta,\boldsymbol{\Omega},\mathbf{t},\mathbf{r} \mid \tau^*,\theta,\boldsymbol{\Omega},\mathbf{t},\mathbf{r})}{P(\tau^*,\theta,\boldsymbol{\Omega},\mathbf{t},\mathbf{r} \mid \tau,\theta,\boldsymbol{\Omega},\mathbf{t},\mathbf{r})}\right) \qquad (2.2)$$

$$\alpha_2 = \min\left(1, \frac{P(D \mid \tau, \mathbf{r}, \mathbf{t}, \theta^*) \cdot P(\mathbf{r} \mid \tau, \mathbf{t}, \boldsymbol{\Omega}, \theta^*) \cdot P(\mathbf{t} \mid \tau, \theta^*) \cdot P(\tau) \cdot P(\theta^*) \cdot P(\boldsymbol{\Omega})}{P(D \mid \tau, \mathbf{r}, \mathbf{t}, \theta) \cdot P(\mathbf{r} \mid \tau, \mathbf{t}, \boldsymbol{\Omega}, \theta) \cdot P(\mathbf{t} \mid \tau, \theta) \cdot P(\tau) \cdot P(\theta) \cdot P(\boldsymbol{\Omega})} \times \frac{P(\tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r} \mid \tau, \theta^*, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r})}{P(\tau, \theta^*, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r} \mid \tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r})}\right) \tag{2.3}$$

$$\alpha_3 = \min\left(1, \frac{P(D \mid \tau, \mathbf{r}, \mathbf{t}, \theta) \cdot P(\mathbf{r} \mid \tau, \mathbf{t}, \boldsymbol{\Omega}^*, \theta) \cdot P(\mathbf{t} \mid \tau, \theta) \cdot P(\tau) \cdot P(\theta) \cdot P(\boldsymbol{\Omega}^*)}{P(D \mid \tau, \mathbf{r}, \mathbf{t}, \theta) \cdot P(\mathbf{r} \mid \tau, \mathbf{t}, \boldsymbol{\Omega}, \theta) \cdot P(\mathbf{t} \mid \tau, \theta) \cdot P(\tau) \cdot P(\theta) \cdot P(\boldsymbol{\Omega})} \times \frac{P(\tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r} \mid \tau, \theta, \boldsymbol{\Omega}^*, \mathbf{t}, \mathbf{r})}{P(\tau, \theta, \boldsymbol{\Omega}^*, \mathbf{t}, \mathbf{r} \mid \tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r})}\right) \tag{2.4}$$

$$\alpha_4 = \min\left(1, \frac{P(D \mid \tau, \mathbf{r}, \mathbf{t}^*, \theta) \cdot P(\mathbf{r} \mid \tau, \mathbf{t}^*, \boldsymbol{\Omega}, \theta) \cdot P(\mathbf{t}^* \mid \tau, \theta) \cdot P(\tau) \cdot P(\theta) \cdot P(\boldsymbol{\Omega})}{P(D \mid \tau, \mathbf{r}, \mathbf{t}, \theta) \cdot P(\mathbf{r} \mid \tau, \mathbf{t}, \boldsymbol{\Omega}, \theta) \cdot P(\mathbf{t} \mid \tau, \theta) \cdot P(\tau) \cdot P(\theta) \cdot P(\boldsymbol{\Omega})} \times \frac{P(\tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r} \mid \tau, \theta, \boldsymbol{\Omega}, \mathbf{t}^*, \mathbf{r})}{P(\tau, \theta, \boldsymbol{\Omega}, \mathbf{t}^*, \mathbf{r} \mid \tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r})}\right) \tag{2.5}$$

$$\alpha_5 = \min\left(1, \frac{P(D \mid \tau, \mathbf{r}^*, \mathbf{t}, \theta) \cdot P(\mathbf{r}^* \mid \tau, \mathbf{t}, \boldsymbol{\Omega}, \theta) \cdot P(\mathbf{t} \mid \tau, \theta) \cdot P(\tau) \cdot P(\theta) \cdot P(\boldsymbol{\Omega})}{P(D \mid \tau, \mathbf{r}, \mathbf{t}, \theta) \cdot P(\mathbf{r} \mid \tau, \mathbf{t}, \boldsymbol{\Omega}, \theta) \cdot P(\mathbf{t} \mid \tau, \theta) \cdot P(\tau) \cdot P(\theta) \cdot P(\boldsymbol{\Omega})} \times \frac{P(\tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r} \mid \tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r}^*)}{P(\tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r}^* \mid \tau, \theta, \boldsymbol{\Omega}, \mathbf{t}, \mathbf{r})}\right) \tag{2.6}$$

where α_1 to α_5 are the acceptance ratios for changes in τ, θ, $\boldsymbol{\Omega}$, $\mathbf{t}$ and $\mathbf{r}$, respectively. Accordingly the global α (the Hastings ratio), as defined by the Metropolis-Hastings algorithm (HASTINGS 1970; METROPOLIS *et al.* 1953), can be given by:

$$\alpha = \begin{cases} \alpha_1 & \text{if } 0 \le U < \varphi_1 \\ \alpha_2 & \text{if } \varphi_1 < U < \varphi_2 \\ \alpha_3 & \text{if } \varphi_2 < U < \varphi_3 \\ \alpha_4 & \text{if } \varphi_3 < U < \varphi_4 \\ \alpha_5 & \text{if } \varphi_4 < U \le 1 \end{cases} \tag{2.7}$$

where U is a randomly generated number used to determine the type of move that occurs. The values of φ_1 to φ_4 are proposal probabilities, which are specified values for deciding the probability of each type of move, such that $0 \le \varphi_1 < \varphi_2 < \varphi_3 < \varphi_4 \le 1$.

Under this Metropolis-Hastings approach to sampling, any volume of parameter space will be sampled proportional to its proposal probability. Because this is a sample-based estimation method, its accuracy is dependent on the number of effectively independent samples drawn.

2.2.2 Methods of implementing relaxed molecular clock models

The most common approach to relaxed clock models is to model the heterogeneity in rates of substitution across branches with a parametric distribution. The assumption of this approach is that the underlying distribution of rates across branches is distributed according to the specified distribution. The exact shape of the distribution is determined by the distribution parameters in $\boldsymbol{\Omega}$ which can be estimated by sampling them using MCMC. In this section, existing procedures used to sample rates and distribution parameters in Bayesian approaches to relaxed molecular clock models are outlined. We then propose a novel procedure which samples rates using the inverse cumulative distribution function.

2.2.2.1 The conventional implementation of relaxed phylogenetics

The basic strategy to implement a relaxed molecular clock model is to sample values in $\mathbf{r}$, $\mathbf{t}$ and $\boldsymbol{\Omega}$ independently of each other (RANNALA and YANG 2007). Under this approach, values for rates of substitution are sampled by drawing rates from a continuous distribution. The probability of the rate is then calculated given the distribution parameters in $\boldsymbol{\Omega}$ (for example, mean and standard deviation parameters of the lognormal distribution). However, when the distribution parameters are also treated as random variables in addition to $\mathbf{r}$, then the standard parameterisation can be difficult to produce efficient proposal kernels for because of the strong correlation between the rate values on individual branches and the parent distribution parameters. Alternatives to this implementation have since been proposed which improve the convergence and speed of relaxed molecular clock methods (DRUMMOND *et al.* 2006).

2.2.2.2 Discretizing relaxed clocks

In BEAST, the current implementation of relaxed molecular clock models is to discretize the branch-rates distribution as a set of regular intervals across the distribution (DRUMMOND *et al.* 2006). Rather than sampling continuous values for rates, this method assigns a categorical value to each branch representing a rate category. Each rate category corresponds to a quantile q, on the probability density function (PDF) of the branch-rates distribution, with each category being a regular distance in probability apart from the previous category on the PDF. The total number of categories is equal to the number of branches on the tree, $2n-2$. The rate corresponding to a particular positive integer category, c, is determined by:

$$r_c = F^{-1}\left(\frac{c-1/2}{2n-2} \mid \mathbf{\Omega}\right) \tag{2.8}$$

where F^{-1} is the inverse cumulative distribution function (quantile function) of the branch-rates distribution. The values of $\mathbf{r}$ are therefore sampled as discrete variables and take on the values of rate that correspond to each rate category. As the particular values of $\mathbf{r}$ are determined independently of the distribution parameters, the rate of convergence is improved over the conventional implementation in 2.2.2.1.

However, there are issues associated with this implementation of relaxed molecular clock models. Under the discretized branch rates implementation, given a particular distribution and its distribution parameters, there are only a fixed number of possible values that the rate of a branch can take on. Thus this implementation lacks an ability to accurately determine values of rate. This is particularly problematic when reconstructing the phylogeny for a small number of taxa, as the number of rate categories modelling the whole distribution is small and each category takes on a very distinct value compared to other categories.

2.2.2.3 Sampling rates as quantiles

We propose a new approach for implementing relaxed molecular clock models under a Bayesian phylogenetic framework. In this procedure, rates of substitution are

sampled using MCMC by representing the rate on each branch by its corresponding probability in the cumulative probability distribution of the branch-rate distribution model. In comparison to previously outlined implementations, a more computationally convenient strategy is to sample values of $q \in (0,1)$ rather than the actual rates, r on individual branches. q can then be interpreted as a rate using the inverse values of the cumulative distribution function (CDF) of the branch-rates distribution. The rate of the *i*th branch, r_i, can be defined in terms of its corresponding probability in the CDF (q_i):

$$r_i = F^{-1}(q_i \mid \mathbf{\Omega}) \tag{2.9}$$

Each value of q_1 to q_{2n-2} can then be estimated via MCMC.

In contrast to the conventional approach developed by Rannala and Yang (2007) where values of r are directly sampled, sampling the q values allows the MCMC to independently sample the parent parameters (Ω) and the rate parameters (q) separately while still getting excellent convergence of the Markov chain. This method therefore improves the convergence of the MCMC analysis over the Rannala-Yang approach and effectively reduces the computational time required for an equivalent analysis. Additionally, since the quantile function parameterisation automatically draws rates from the prior defined by the parent distribution, no per-branch terms associated with the rate model need to be added to the calculation of the prior density, so long as the prior on the $\mathbf{q}$ values is unit uniform.

Also, as values of rate are sampled on a continuous scale, this implementation does not pose the same problem that discretizing the branch-rates distribution does with respect to not being able to provide precise estimates of rate.

2.2.3 Branch-rates distribution models

In relaxed phylogenetics, the underlying distribution of branches can be modelled by non-negative, parametric distributions. Individual branch rates are sampled from the

distribution and the likelihood of the rate given the distribution can be calculated. The rates can be sampled from the branch-rates distribution in either an uncorrelated or autocorrelated manner. In the uncorrelated rates models, the rate for each branch is drawn independently from a distribution that is shared globally across all branches. The autocorrelated rate model is an alternative in which each branch rate is dependent on the rate of its parent branch and the divergence time from the parent.

Two commonly used branch-rates distribution models are the lognormal and exponential distributions (WELCH and BROMHAM 2005). We propose the use of the inverse Gaussian distribution as a new distribution to model the rate variation across branches. The inverse Gaussian distribution has certain properties in its PDF that differ from the distributions currently used as branch-rates distributions, and may provide a better fit for certain data. We describe how rates are drawn from each of these branch-rates distribution models.

2.2.3.1 The uncorrelated lognormal distribution (LN) model

The lognormal distribution is commonly used as the distribution of branch-rates (DRUMMOND *et al.* 2006; RANNALA and YANG 2007; WELCH and BROMHAM 2005; YANG 2006). The distribution of values drawn from the lognormal distribution is normally distributed after a log transform. The lognormal distribution can be used to describe variables that are the product of multiple positive independent random variables. The shape of the lognormal distribution is controlled by two parameters: the mean μ, and shape parameter S. S is equivalent to the standard deviation σ, of the distribution after a log transform. When implementing a lognormal branch-rates distribution model, the probability of a rate r_i on a branch given the distribution parameters has the form:

$$P(r_i \mid \mu, S) = \frac{1}{r_i\sqrt{2\pi S^2}} \exp\left[\frac{-\left(\ln(r_i/\mu) - \frac{S^2}{2}\right)^2}{2S^2}\right] \qquad (2.10)$$

2.2.3.2 The uncorrelated exponential distribution (E) model

Another common branch-rates distribution is the exponential distribution. Unlike the lognormal distribution, it is a one parameter distribution, controlled by the rate parameter λ. In relaxed clock models, the exponential is parameterised with $\mu = 1/\lambda$. The probability distribution has the form:

$$P(r_i \mid \mu) = \frac{\exp\left[\frac{-r_i}{\mu}\right]}{\mu} \tag{2.11}$$

2.2.3.3 The uncorrelated Inverse Gaussian distribution (IG) model

Besides the standard branch-rates distributions, the use of the inverse Gaussian distribution as a model for the distribution of rates across branches is investigated. The suggestion of using alternative distributions stems from Kitazoe *et al.* (2007), who found that alternative models of rate distribution were more competent at modelling the rate heterogeneity across branches in mammalian mitochondrial proteins.

The shape of the inverse Gaussian distribution is determined by two parameters: the mean, μ, and the shape parameter, λ. The probability of a rate for the inverse Gaussian distribution model can be expressed as:

$$P(r_i \mid \mu, \lambda) = \left(\frac{\lambda}{2\pi r_i^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda(r_i - \mu)^2}{2\mu^2 r_i}\right) \tag{2.12}$$

Alternatively, the probability of a rate can be parameterised by the standard deviation σ rather than λ. Since

$$\sigma = \sqrt{\left(\frac{\mu^3}{\lambda}\right)} \tag{2.13}$$

Equation 2.12 can be expressed as:

$$P(r_i \mid \mu, \sigma) = \left(\frac{\mu^3}{2\pi r_i^3 \sigma^2} \right)^{\frac{1}{2}} \exp\left(\frac{-\mu^3 (r_i - \mu)^2}{2\mu^2 r_i \sigma^2} \right) \tag{2.14}$$

Overall, the probability distribution function of the IG is similar to the LN when the coefficient of variation is low (less than 1). Takagi *et al.* suggested that as the coefficient of variation increases, the IG distribution tends to have a sharper peak, while the lognormal has a flatter tail (1997). The shapes of the LN and IG distributions were explicitly compared. Figure 2.1 shows the differences in overall skewness and kurtosis of the two distributions across different values of mean and standard deviation. From this figure, it can be seen that the skewness and kurtosis of the two distributions are similar when the ratio of μ to σ is large (i.e. the coefficient of variation is small). When σ increases, the skewness of the LN distribution increases much more rapidly than IG. This difference in skewness is more prominent when a small μ is considered (Figure 2.1a). The differences in kurtosis between the two distributions show a similar pattern to the skewness but at a much larger scale of increase (Figure 2.1b). Overall, the two distributions are very different in shape when the parameters correspond to large coefficient of variations. Where the coefficient of variation is high, the lognormal has a much sharper and less symmetric distribution.

This finding is contrary to Takagi *et al.*'s (1997) study. The reason for the difference in result is that in Takagi *et al.*, the authors only visually compared the peaks of the distributions without taking into account the long tails that follow. By evaluating the entire function, we found that although the peak of the IG distribution was sharper, its tail had a much gentler slope. Given the flatter tail of the PDF of the IG distribution, when employed as a branch-rates distribution, lighter penalties are given to abrupt increases in rate of substitution. The IG distribution is therefore more liberal in allowing for relatively faster rates of evolution within the tree. This property may be suitable for datasets where there are rogue taxa with exceptionally fast rates. An example of such datasets is mammalian data where the rodent lineages have

accelerated rates (BRITTEN 1986; LI *et al.* 1996; MARTIN and PALUMBI 1993; WU and LI 1985).

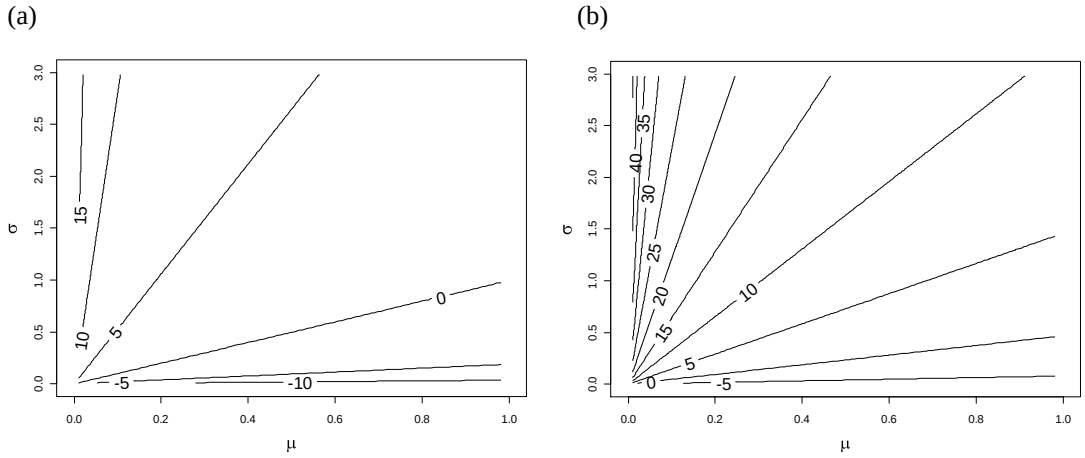


Figure 2.1 Contour plots of (a) skewness of LN minus skewness of IG and (b) kurtosis of LN minus kurtosis of IG on a log scale across different values of μ and σ.

Parameter values of μ and σ shown reflect those often used in phylogenetic analyses of real biological datasets.

2.2.3.4 The autocorrelated lognormal distribution model

Autocorrelated relaxed molecular clock models (ARIS-BROSOU and YANG 2002; THORNE *et al.* 1998; YANG and NIELSEN 2002) are based on the fact that rates of substitution are often similar among monophyletic groups of species, such as within taxonomic groups (BRITTEN 1986). Several studies by Ho *et al.* (HO *et al.* 2007a; HO *et al.* 2005; HO *et al.* 2007b) have demonstrated through simulation studies and the analysis of metazoan and primate data that these types of models better fit the behaviour of rates across species. Under Thorne and Kishino's implementation of the autocorrelated model (KISHINO *et al.* 2001; THORNE *et al.* 1998), the rate of a branch is drawn from a distribution with the mean equal to the rate of the parent branch r_p and the variance proportional to the time t since divergence from the parent node. By substituting these two parameters in the equation for the uncorrelated lognormal model shown in Equation 2.10, an autocorrelated lognormal model can be expressed as:

$$P(r_i \mid r_{p,i}, t_i, S) = \frac{1}{r_i\sqrt{2\pi t_i S^2}} \exp\left[\frac{-\left(\ln(r_i / r_{p,i}) - \frac{t_i S^2}{2}\right)^2}{2t_i S^2}\right] \quad (2.15)$$

Since a factor of t is also taken into account here, the S^2 parameter in the autocorrelated model is the increase in variance of the child rate per unit time of divergence from the parent. It differs from the S^2 in the uncorrelated model and cannot be compared directly. Considering a path of rates from the root to a tip of the tree, this model approximates a geometric Brownian motion, whereby the logarithm of the rate varies over time according to Brownian motion, and so the variance of the distribution increases proportionally to the divergence time from its parent.

2.2.4 Dataset

We used the OrthoMaM dataset v4.0 (RANWEZ *et al.* 2007), which contains alignments of orthologous genomic markers shared between 25 placental mammals (*Ornithorhynchus anatinus*, *Monodelphis domestica*, *Echinops telfairi*, *Loxodonta africana*, *Dasypus novemcinctus*, *Ochotona princeps*, *Oryctolagus cuniculus*, *Cavia porcellus*, *Spermophilus tridecemlineatus*, *Mus musculus*, *Rattus norvegicus*, *Tupaia belangeri*, *Microcebus murinus*, *Otolemur garnettii*, *Macaca mulatta*, *Pongo pygmaeus abelii*, *Pan troglodytes*, *Homo sapiens*, *Erinaceus europaeus*, *Sorex araneus*, *Myotis lucifugus*, *Equus caballus*, *Bos taurus*, *Canis familiaris* and *Felis catus*). All CDS markers that shared orthology across the 25 species in OrthoMaM were obtained, for a total of 1056 alignments. For this analysis, we wanted to obtain a set of alignments where the true species topology between the sequences is well established and uncontroversial. We retained the 1056 alignments shared by 12 mammalian species: *C. familiaris*, *F. catus*, *H. sapiens*, *P. troglodytes*, *P. pygmaeus abelii*, *M. mulatta*, *M. murinus*, *O. garnettii*, *M. musculus*, *R. norvegicus*, *O. princeps*, and *O. cuniculus*. The phylogeny between these species is well supported by literature (BASHIR *et al.* 2005; NOVACEK 2001; PRASAD *et al.* 2008; REYES *et al.* 2004; STEIPER and YOUNG 2006; YODER 1997) and is shown in Figure 2.2.

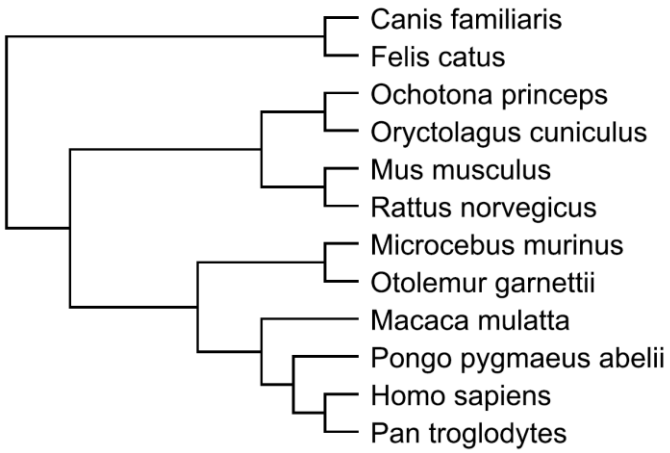


Figure 2.2 The tree topology used as the true species topology for our mammalian dataset.

2.2.5 Algorithm Implementation

The models and relaxed clock implementations were written in Java 1.5, and form part of the BEAST (DRUMMOND and RAMBAUT 2007) software package.

2.2.6 Relaxed clock model priors

We compare three distributions that can be used to model the variation in rate of substitution across branches: the LN distribution, the E distribution, and the IG distribution. E and LN were already implemented in BEAST (DRUMMOND and RAMBAUT 2007) and are commonly used to model the rates across branches (for example, DRUMMOND and RAMBAUT 2007; HO *et al.* 2007a; HO *et al.* 2005; HO *et al.* 2007b; YANG 2007).

As no time-calibration data was available for this dataset, absolute rate and divergence time cannot be separated. Instead the relative rates were estimated by setting the mean rate μ to 1.0 for all distributions. The priors used for each of the branch-rates distribution models are shown in Table 2.1.

Table 2.1 A summary of rate distribution priors used in the relaxed clock models of the mammalian dataset.

Distribution	Parameter	Prior boundaries of parameter
Exponential (E)	Rate, λ	1.0 (σ = 1.0)
Lognormal (LN)	S	0.0 - 10.0
Inverse Gaussian (IG)	Standard deviation, σ	0.0 - 10.0
Autocorrelated LN	S^2	0.0 - ∞

It should be noted that in our implementation, IG was parameterised with the standard deviation parameter σ rather than the parameter λ. The motivation lies within the relationship between σ and λ in the IG as shown in Equation 2.13. As σ and λ are inversely related in the IG, the hyperprior (the prior distribution of a parameter of a prior distribution) for the MCMC that is naturally imposed on the distribution parameter, if it was parameterised with λ, will be an inverse of the natural hyperprior on S in LN. Sampling σ in IG was therefore done to improve consistency of priors across the models compared. However, there were still discrepancies as the LN model is parameterised in BEAST with the S parameter. The relationship between S and σ for the LN is:

$$\sigma = \sqrt{\mu^2 (e^{S^2} - 1)} \tag{2.16}$$

An implicit hyperprior is hence placed on S, where the LN implementation samples σ in exponential space compared to in uniform space for the IG. In effect, comparisons between branch-rates models also take into account the comparisons of both the branch-rates model itself and the specified priors associated.

It should also be noted that for IG, the parameterisation of $\sigma \in (0,10)$ is not equivalent to the parameterisation of LN where $S \in (0,10)$ in terms of the range of the values sampled for the standard deviation parameter. The actual σ of LN can be expressed as:

$$\sigma = \mu \sqrt{e^{S^2} - 1} \tag{2.17}$$

Because of this relationship, an upper limit of 10 for S actually corresponds to an upper bound of 5.18×10^{21} for σ. This differs from the IG where the upper bound of σ is 10. From the analysis of the results of the uncorrelated LN model, we found that the maximum posterior estimate of σ was 2.3 across all alignments. The distribution of σ values indicates that for a majority of the alignments the heterogeneity of rates was well within 3 standard deviations and no σ values actually exceeded the IG's upper limit of σ = 10 (shown in Figure 2.3). Therefore, the difference in the upper limit of the prior on σ does not significantly affect the accuracy of the results in either model.

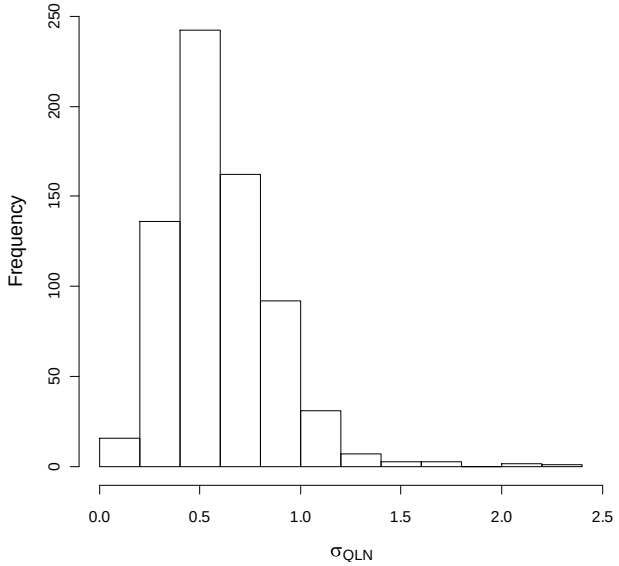


Figure 2.3 Histogram showing the distribution of the standard deviations σ of the analysis of 1056 mammal genes under an uncorrelated LN model.

The standard deviations were estimated using a lognormal model where rates were sampled as quantiles.

2.2.6.1 Choosing priors for autocorrelated models

Additional priors are required for autocorrelated models as their parameterisation differs from uncorrelated models. For autocorrelated models, the mean rate cannot actually be fixed since the mean rate for each branch is the parent rate r_p. Instead a strong prior was placed on the rate of the root node with a Normal(1, 0.01)

distribution. As the root rate is approximately 1, the branch-rates distribution of its descendant nodes will have a mean rate that tightly fluctuates around 1.

A prior was also set on the S^2 parameter for autocorrelated models. We investigated a suitable prior for the distribution of S^2 values across the 12 mammalian species. A gamma distributed prior with parameters k = 2 and θ = 0.5 and an exponentially distributed prior with parameter λ = 1 were compared. Analyses were run with each of these two prior settings on a random sample of 200 genes from the dataset. The posterior estimates of S^2 were retained for each of the genes across the two settings. Figure 2.4 shows a frequency histogram of the posterior mode point estimates of S^2 across these 200 genes compared to the prior distribution. The gamma distributed prior (Figure 2.4a) appears to slightly better match the resulting histogram of posterior point estimates than the exponential distribution (Figure 2.4b). A Kolmogorov-Smirnov (K-S) test was performed on the two sets of results to verify how well each empirical distribution of point estimates matched its prior distribution. Both K-S tests were statistically significant, indicating that both data distributions differed from their parent distributions (both p-values < 0.0001).

This was further investigated by fitting the corresponding parametric distribution to each histogram of point estimates using maximum likelihood fitting of univariate distributions (VENABLES and RIPLEY 2002). The K-S tests were then recomputed using the optimal parameters from the maximum likelihood fit, which were λ = 1.19 for the exponential distribution and k = 3.824, θ = 4.678 for the gamma distribution. Of the two distributions, only the fitted gamma distribution adequately described the empirical distribution of S^2 point estimates (p-value = 0.762). Although the maximum likelihood parameter estimates provided a better fit to the data, a gamma distribution with k = 3.824 and θ = 4.678 would have a large mean for the S^2 estimates. To maintain a distribution mean of 1 and a reasonable fit of data, the original gamma distribution with k = 2 and θ = 0.5 was used as the prior distribution of S^2 for the autocorrelated model.

(a) (b)

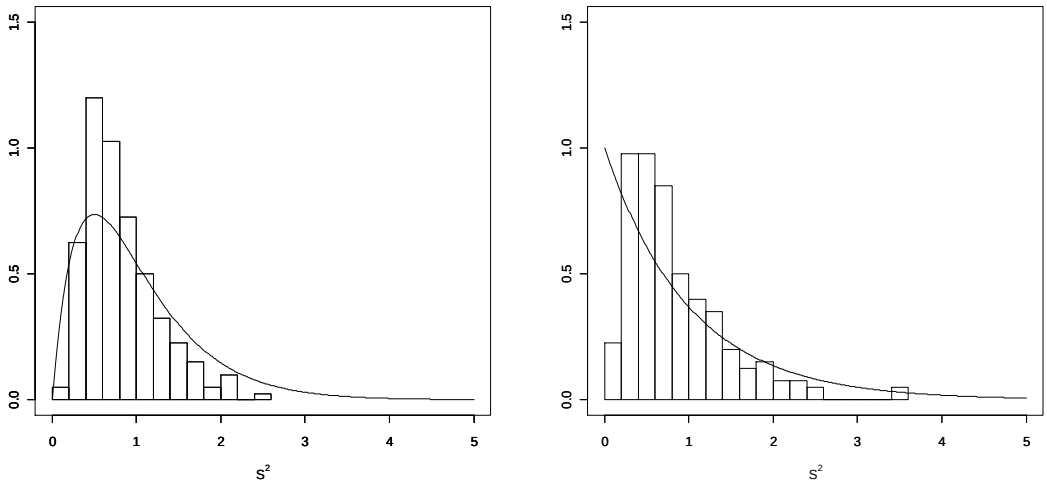


Figure 2.4 Histograms of the parameter estimates of S^2 across 200 random genes for the M_{ACLN} model using two different prior distributions (a) Gamma(k = 2, θ = 0.5) and (b) Exponential(λ = 1).

The line on the plot shows the probability distribution function of the parent distribution.

2.2.7 Other MCMC priors

A Yule pure birth process was used as a prior on the speciation process (EDWARDS 1970; YULE 1924). In general, the Yule process best models the speciation across large divergence times such as those used in comparisons on the species level (DRUMMOND *et al.* 2006). The general time-reversible model (GTR) (TAVARÉ 1986) with gamma distributed heterogeneity across sites (YANG 1994) (GTR + Γ) was used as the nucleotide substitution model. A $1/x$ prior was placed on the relative rate parameters of the GTR.

2.2.8 Proposal kernels for MCMC

The following proposal kernels in BEAST were used for the MCMC the analyses performed:

- Scale operator on the S and σ values of the LN and IG distribution models, respectively
- Scale operator on the birth rate of the Yule prior

- Scale operator on root height of trees
- Uniform operator on internal node heights
- Subtree slide, narrow exchange and Wilson-Balding operators on tree topology (for analyses where the topology was not constrained)
- Scale operators on the substitution frequencies of the GTR model
- Delta exchange on equilibrium base frequencies of the alignment
- Scale operator on α value of the Γ model
- Uniform Operator on each value of q
- Up-down operator on the rates on the tree against the heights of the internal nodes (for autocorrelated models only)
- Scale operator on the node rates (for autocorrelated models only)

2.2.9 Normalising the mean rate on branches

Part of the analysis was focused on the relative rates of branches to each other. For the implementations of the relaxed clock, as branches rates are only drawn from a distribution with a mean rate of 1, the actual mean rate of the branches on a tree will only loosely conform to 1. Also the computations of mean rate account for the values of **t** and so mean rate varies depending on the estimates of divergence times. As a result, the mean rate of the branches in the posterior distribution differs from the mean rate parameterised in the prior.

We normalised the rates of the branches in the posterior samples *a posteriori*. For each tree in the posterior sample, the normalised rate r_{ij}' on each branch was recomputed as:

$$r_{ij}' = \frac{r_{ij}\hat{r}}{\sum r_{ij}t_{ij} / \sum t_{ij}} \tag{2.18}$$

where $\hat{r}$ is the mean rate we want to normalise to, which in our case is 1. Consequently, the divergence times must be adjusted to obtain the correct product of time and rate. The normalised divergence time t_{ij}' of each branch can be expressed as:

$$t_{ij}' = \frac{t_{ij} \sum r_{ij} t_{ij} / \sum t_{ij}}{\bar{r}} \tag{2.19}$$

Under this normalisation, the rates on a particular tree in the posterior samples are bound by the relationship:

$$\frac{\sum r_{ij}' t_{ij}'}{\sum t_{ij}'} = \hat{r} \tag{2.20}$$

2.3 Results

2.3.1 Comparing rate clock implementations and models

We compared the reconstruction of gene trees in the mammalian dataset across eight molecular clock models: (1) an uncorrelated LN distribution model with a discretized branch rates implementation (M_{DLN}), (2) an uncorrelated E distribution model with a discretized branch rates implementation (M_{DE}), (3) an uncorrelated IG distribution model with a discretized branch rates implementation (M_{DIG}), (4) an uncorrelated LN distribution model with a sampling rates as quantiles implementation (M_{QLN}), (5) an uncorrelated E distribution model with a sampling rates as quantiles implementation (M_{QE}), (6) an uncorrelated IG distribution model with a sampling rates as quantiles implementation (M_{QIG}), (7) a LN distribution model with autocorrelated branch rates (M_{ACLN}), and (8) a strict molecular clock model (M_{SC}). A summary of these models is shown in Table 2.2.

For this analysis, we realised that it would be of worth to additionally compare each of these implementations to the Yang-Rannala implementation of relaxed molecular clocks (RANNALA and YANG 2007). The precise Yang-Rannala method, however, is currently not implemented in BEAST. As both implementations sample values of rate from the same distribution, our expectation is that the results of the Yang-Rannala model would be effectively identical to our implementation in which the rates are sampled as quantiles.

Table 2.2 A summary of the different molecular clock models compared in this chapter.

Model	Branch rate implementation	Branch-rates distribution	Autocorrelated/ Uncorrelated
$\mathbf{M_{DLN}}$	Discretized	LN	Uncorrelated
$\mathbf{M_{DE}}$	Discretized	E	Uncorrelated
$\mathbf{M_{DIG}}$	Discretized	IG	Uncorrelated
$\mathbf{M_{QLN}}$	Samples quantiles	LN	Uncorrelated
$\mathbf{M_{QE}}$	Samples quantiles	E	Uncorrelated
$\mathbf{M_{QIG}}$	Samples quantiles	IG	Uncorrelated
$\mathbf{M_{ACLN}}$	Continuous	LN	Autocorrelated
$\mathbf{M_{SC}}$	-	Strict molecular clock	Uncorrelated

For each of the 1056 sequences, an MCMC analysis was run for a chain length of 25,000,000. If all parameters of interest had not converged after the run, the chain was rerun for longer until the effective sample sizes (ESS) were above 100. The ESS is a measure of mixing and convergence in MCMC. The ESS of an MCMC chain is the number of independent samples of the posterior that the MCMC sample is equivalent to. Its calculation takes into account the length of the chain and effects of autocorrelation that occur between subsequent samples. This measure was used to provide an assessment of whether the chain had been run for a sufficient number of samples or whether different proposal kernels were needed. In a number of the genes, we were unable to obtain convergence in the MCMC chain even after 200 million steps. With the given data, models and proposal kernels, the MCMC runs could not be made to converged within practical running time. In cases where this occurred in one or more of the models being compared, the genes were excluded from the analysis. Although this is a potential source of bias, the effect should be small as this only occurred in approximately 8% of the MCMC chains for each model compared. In total, 695 genes were used for the comparison between models. In all runs, the first 10% of the MCMC samples were treated as burn-in and discarded.

The results of the MCMC runs from the eight models were summarised in terms of accuracy in topology reconstruction (Table 2.3). Whether the true tree is the point estimate for the topology (TTPE) and the posterior probability of the true tree (PPTT)

are indicators of the strength of support for the true tree given the model. Out of these two measures, TTPE has greater importance in terms of assessing the accuracy of the phylogenetic reconstruction. If the true tree is not captured within the 95% credible set (TT95), the phylogenetic estimation is usually poor and on the whole provides a bad estimate of the tree. The number of unique trees in the 95% credible set (NU95) is an indicator of the uncertainty associated with the point estimates. A smaller number of trees in the 95% credible set indicates that the Bayesian MCMC analysis is more conclusive.

We first compared the effects of rate clock implementation and branch-rates distribution on the accuracy of our uncorrelated models, as measured by PPTT. As the values of PPTT were not normally distributed, the comparison was performed by computing a Scheirer-Ray-Hare extension of the Kruskal-Wallis test (SCHEIRER *et al.* 1976). This test is a non-parametric alternative to a two-way analysis of variance (ANOVA) and tests for the effects of each factor (in this case the factors are rate clock implementation and the branch-rates distribution) on the values. The test indicated that there was no significant interaction between the factors of relaxed molecular clock implementation and branch-rates model (p-value = 0.642). Each of these factors was then examined individually. For the factor of rate clock implementation, the associated p-value was 0.920, suggesting that using discretized branch rates did not affect the accuracy of the tree estimation in terms of PPTT. The p-value of differences among branch-rates distributions was < 0.0001, suggesting strong evidence that using different branch-rates distributions impacted the accuracy of tree topology estimation.

Table 2.3 Statistics related to the accuracy in estimation of topology for the mammalian dataset.

695 genes were used for this analysis.

Model	Proportion of times true tree is point estimate (TTPE)	Mean posterior probability of true tree (PPTT)	Mean number of unique trees in 95% credible set (NU95)	Proportion of times true tree appears in 95% credible set (TT95)
$\mathbf{M_{DLN}}$	0.508	0.424	54.391	0.863
$\mathbf{M_{DE}}$	0.455	0.366	95.968	0.855
$\mathbf{M_{DIG}}$	0.509	0.421	61.104	0.863
$\mathbf{M_{QLN}}$	0.531	0.428	67.210	0.876
$\mathbf{M_{QE}}$	0.450	0.343	125.636	0.866
$\mathbf{M_{QIG}}$	0.525	0.420	75.704	0.873
$\mathbf{M_{ACLN}}$	0.505	0.433	32.131	0.850
$\mathbf{M_{SC}}$	0.560	0.516	7.037	0.777

More detailed pairwise comparisons between the models were performed. As none of the measures of accuracy were normally distributed [*p*-value < 0.05 in all variables using Shapiro-Wilk tests (SHAPIRO and WILK 1965)], non-parametric significance tests were used. For the continuous variables of PPTT and NU95, significance was tested using non-parametric *t*-tests with sampling of bootstrap replicates (EFRON 1981). For the binomial variables of TT95 and TTPE, McNemar's tests were used (MCNEMAR 1947).

The difference in accuracy between the discretized branch rates implementation (M_D) and the new approach of sampling rates as quantiles (M_Q) was further compared. By using the M_Q models, the accuracy of topology estimation was significantly improved for LN in terms of TT95 and TTPE (*p*-value = 0.027 and 0.034, respectively). These values were also generally higher in M_Q for IG, though not statistically significant. The confidence of the tree estimation decreased when using M_Q implementations, with NU95 increasing significantly in all three distributions (*p*-value < 0.0001 in all cases); however, it should be noted that accuracy should always be considered a priority above confidence.

The use of each branch-rates distribution model for modelling mammalian data was then compared. In looking at the comparisons between the three distributions across both M_D and M_Q, we observe that the LN and IG models performed comparably across all measures of accuracy. LN was actually significantly better than IG in terms of PPTT (*p*-value < 0.0001 and = 0.0003 for M_D and M_Q, respectively), but was not significantly different in terms of TTPE or TT95. Out of these two models, the LN had a much lower NU95 (*p*-value = 0.0004 for both M_D and M_Q), showing an advantage in terms of how conclusive its predictions are. The E distribution performed worse than the other two distributions across all measures of model performance, with significantly worse TTPE, PPTT and NU95 (all *p*-values < 0.0001 for M_D and M_Q).

The pairwise comparison between LN and IG was examined in more detail. The pairwise values of PPTT for each alignment are shown in Figure 2.5. Under the criterion of PPTT, a majority of the analyses performed slightly better in LN than IG.

In a small number of alignments, the choice of one branch-rates distribution over another dramatically improved PPTT, though the number of outliers outside of the 95% confidence interval was less than 5%.

In conclusion, out of these three branch-rates distributions, LN in general fits this mammalian dataset the best, although no major differences were found between LN and IG. The E distribution appeared to perform worse than either of the two-parameter distributions for modelling rates in these mammalian genes. However, the data analysed here contained only a small number of taxa so further empirical studies are needed to confirm these results.

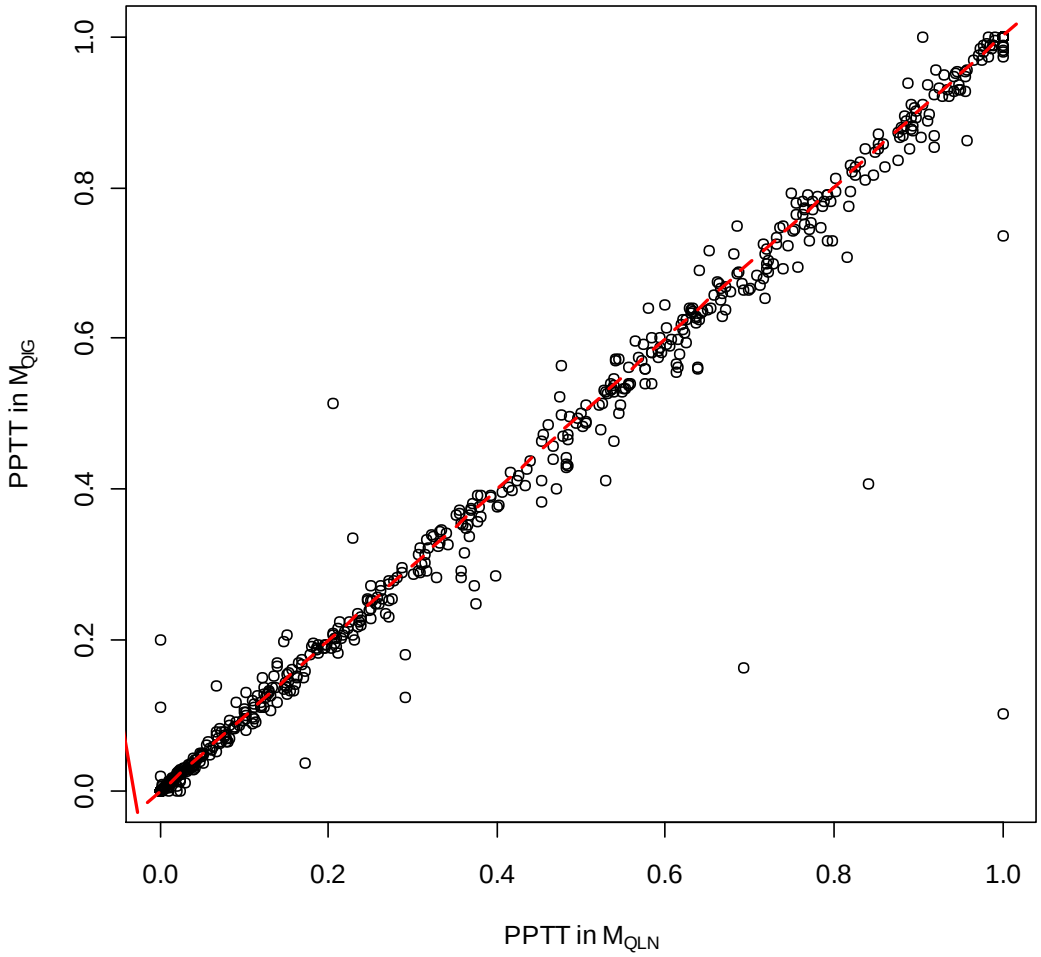


Figure 2.5 Scatterplot of PPTT for comparisons between M_{QLN} and M_{QIG} using the mammalian alignments.

The figure indicates that the two models estimate tree topology to similar accuracies. Overall, the posterior probability of the true true is slightly higher in M_{QLN}, suggesting a slightly better fit of model in this mammalian dataset.

We then compared the relaxed molecular clock models to a strict clock approach to identify the relevance of relaxed molecular clock methods to the mammalian data. M_{SC} was shown to provide significantly better accuracy in terms of PPTT (*p*-value < 0.0001 in a cases), and also had a much lower NU95 (*p*-value < 0.0001). However, the TT95 value for M_{SC} was much lower than relaxed clock methods. Specifically, in 22.3% of the alignments in the dataset, the phylogenetic estimation excluded the true tree from the 95% credible set with a strict clock model, compared to a mean of 13.6% across relaxed clock models. The TT95 for M_{SC} was significantly lower than all relaxed clock approaches (*p*-values < 0.0001 in all comparisons), indicating that is more prone to providing bad estimates of topology. In datasets where M_{SC} contains the true tree in the 95% credible set, the corresponding mean estimate of the *S* parameter in M_{QLN} is 0.503, whereas its mean is 0.660 in the remainder. These two means are significantly different ($p < 0.0001$; using a non-parametric *t*-test). This shows that the M_{SC} model is not robust to data that is not clock-like, but probably performs well on the large fraction of relatively clock-like alignments in this data set. When the data is not clock-like, the strict clock method will more often fail to accurately estimate the phylogeny. The use of relaxed molecular clock methods is therefore justified. The relaxed molecular clock models do have larger credible sets than the strict molecular clock, and thus trade precision for accuracy. However in the context of 13,749,310,575 possible tree topologies (for 12 taxa) an increase in the credible set from 7 to 67 (see Table 2.3) is quite modest.

M_{ACLN} was compared to M_{QLN} to determine whether autocorrelated or uncorrelated models were better for phylogenetic reconstruction of mammal phylogenies using these data. As the branch-rates distribution model implementation of M_{ACLN} in BEAST is implemented in a continuous manner, a more fair comparison is to compare it to M_{QLN} rather than M_{DLN}. The results showed that M_{ACLN} was less accurate than M_{QLN}, with significantly lower TTPE and TT95 values (*p*-value = 0.0336 and 0.0003, respectively). The NU95 for M_{ACLN} was about half that of M_{QLN}, making it significantly more conclusive even though its accuracy was lower (*p*-value < 0.0001). Overall, since lower accuracy was observed in phylogenetic reconstruction, the use of autocorrelated models did not appear suitable for these mammalian data.

Obviously a comparison of these phylogenetic reconstructions with an unrooted Felsenstein model, such as that used in MrBayes (HUELSENBECK and RONQUIST 2001) would also be of interest. However in order to achieve such a comparison additional assumptions are required to root the unrooted trees. This is a worthwhile subject of future research.

2.3.2 Quantifying lineage-specific rates

One comparison worth noting is the estimated relative rates of substitution across the different branches. As the rates of evolution across mammalian species are well characterised (for example, LI *et al.* 1996; MARTIN and PALUMBI 1993; STEIPER and YOUNG 2006; WU and LI 1985), it was worthwhile comparing the conclusions that these methods drew to what previous studies have found.

Only the results of M_{QLN} were examined in this analysis as it provided the best accuracy in the previous section. Using the program Tree Annotator in BEAST (DRUMMOND and RAMBAUT 2007), the maximum clade credibility tree was computed for each MCMC run, which provided a single estimate of the gene tree from the samples in the posterior. If the topology of the maximum clade credibility tree did not match that of the true tree topology (Figure 2.2) then that gene was omitted from the analysis. The resulting dataset contained maximum clade credibility trees from 510 genes.

The average relative rate across genes was calculated for each of the branches on the tree. For branch i, gene j, the deviation Δ_i, in rate over what is expected on average across all branches is calculated as:

$$\Delta_i = \frac{\sum_{j=1}^{n} \ln\left(\frac{r_{ij}'}{r_{\cdot j}'}\right)}{n} \tag{2.21}$$

where $r_{\cdot j}$ is the average rate of that gene. The empirical distribution of r across all genes for a particular branch tends to be significantly skewed. The relative rates were

log transformed so that the distribution of rate values was less skewed. Based on this calculation, small values of Δ_i indicate that the rates on a branch are relatively slow and large values indicate rates that are relatively fast.

In some of the maximum clade credibility trees, there are branch rates that are equal to zero. This occurs when two sister sequences are identical and thus the genetic distance between them is zero. To allow these relative deviations to be calculated, all zero branch rates were replaced with half of the minimum non-zero value across all branch lengths.

The values of Δ_i for this dataset are annotated on the species tree in Figure 2.6. The Δ_i values for humans and chimpanzees (*H. sapiens* and *P. troglodytes*) suggest that their rates of substitution are slowest across the 12 mammalian species examined, with Δ_i of -1.186 and -1.189, respectively. The small Δ_i also appears to somewhat extend to the monophyletic group of great apes. The relative rates observed in these ape taxa are well characterised in previous studies of mammalian evolution. The slow rates in humans, chimps and apes in general have been attributed to their increased generation times, low metabolic rates and large body sizes (BRITTEN 1986; LI *et al.* 1996; MARTIN and PALUMBI 1993; WU and LI 1985). On the other hand, the values of Δ_i suggest that there was a change in rate of substitution in the ancestor of the rodent species. The rate on the rodent ancestral branch appeared fastest among the species examined, with Δ_i of 0.609. Compared to the great apes, and even other mammalian species, rodents have much shorter generation times, faster metabolic rates and smaller body sizes, thus contributing to much faster rates of substitution (BRITTEN 1986; LI *et al.* 1996; MARTIN and PALUMBI 1993; WU and LI 1985). Both our findings here are consistent with the literature, which verifies that the relaxed molecular clock methods used here are, to some degree, capable of uncovering the underlying rates of branches (BRITTEN 1986; LI *et al.* 1996; MARTIN and PALUMBI 1993; STEIPER and YOUNG 2006; WU and LI 1985).

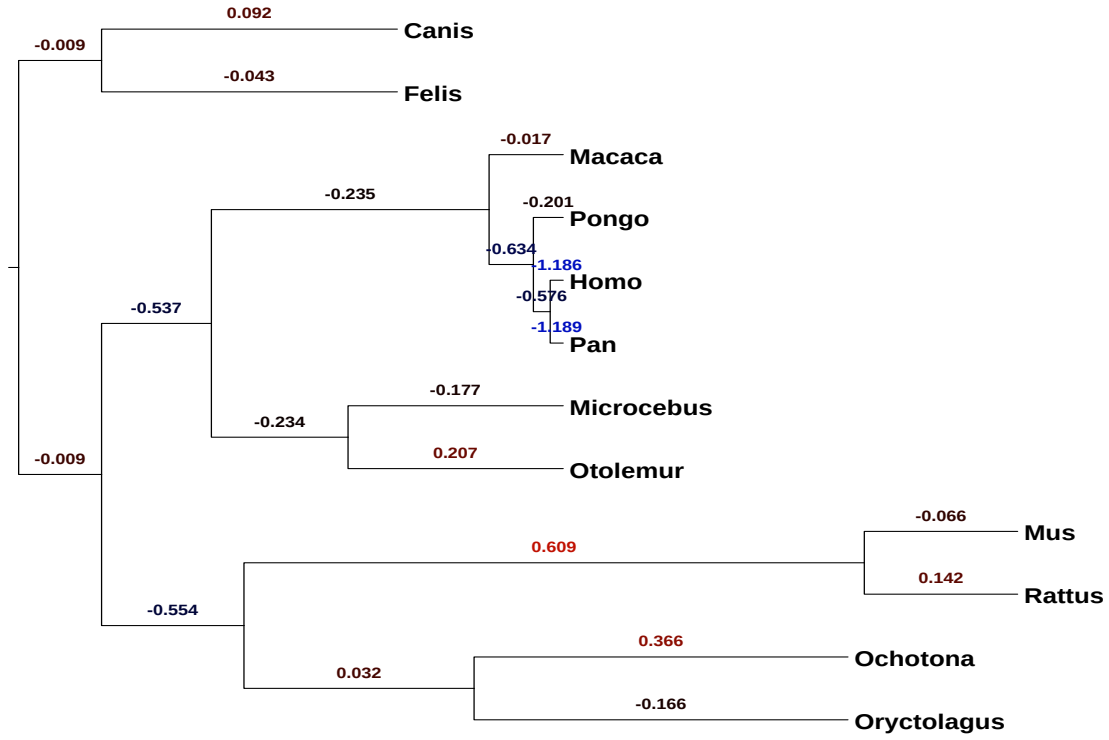


Figure 2.6 The species tree between the 12 mammalian species, annotated with values of Δ_i on the branches.

The branches are annotated with their respective Δ_i values, denoting the relative rate of a branch, averaged across all 510 genes compared. Δ_i values are coloured by their relative values across the tree, with red being exceptionally fast and blue being exceptionally slow. The branch lengths on the tree represent estimates of relative divergence times between the species.

2.3.3 Which genes estimated the tree topology incorrectly?

As an empirical analysis, the characteristics of datasets where the estimation of topology is rather error prone were also of interest. For M_{QLN}, the results were separated into two groups based on whether or not the phylogenetic reconstruction was able to correctly capture the true species topology within the 95% credible set. Of the 948 genes used in this analysis, the model was able to capture the true tree in 812 alignments (86%) but not in the other 136 (14%).

We compared summary statistics of the posterior point estimates of the MCMC runs between these two groups (Table 2.4). The statistics relating to the variation in the rates across branches (the covariance, coefficient of variation and the S parameter of the lognormal distribution S_{LN}) were all significantly higher in trees where the model

was unable to capture the true tree (all p-values < 0.001, non-parametric). These results imply that for data that is less molecular clock-like, the estimation of the tree topology is more difficult. To further verify this finding, tests of correlation were performed between the values of S_{LN} and PPTT on the 948 genes to identify any relationships between rate heterogeneity across branches and accuracy of tree estimation. The Pearson's product-moment correlation (R) between these two variables was -0.333, which indicated a strong negative correlation between S_{LN} and PPTT (p-value < 0.0001). A linear regression of S_{LN} by PPTT fitted a line with slope of -0.209. This slope value was significantly different from zero (p-value < 0.0001), indicating a significant negative relationship between the rate heterogeneity in the data and the accuracy of tree estimation.

The mean rate across the branches also appeared to be significantly lower when the topology was not captured in the 95% credible set (p-value < 0.001). As a result, the estimates of root heights on average became significantly higher (p-value < 0.001), corresponding to larger divergence times. The posterior estimates of the speciation rate (the birth-rate parameter of the Yule model; p-value < 0.01) and the log likelihood of the Yule model (p-value < 0.001) became significantly lower, as these statistics also correspond to trees where the divergence times are overall longer. This change in divergence time was to compensate for the low rate of substitution, in order to maintain the same genetic distance. As divergence times are essentially constant across genes, the differences in rate and divergence time estimations are attributable to underestimation of the mean rate.

Table 2.4 The mean values of statistics related to the estimation of topology and rate for datasets where the 95% credible set contained the true tree versus those that did not.

The *p*-values shown are calculated using a non-parametric two-sided *t*-test. 948 genes were used for this analysis and the results for M_{QLN} were used.

	Mean Posterior estimates		
	95% credible set contains true tree	**95% credible set does not contain true tree**	***p*-value**
Mean rate	0.925	0.864	0.000
Root height	0.139	0.160	0.001
Birth rate of Yule process	18.149	16.329	0.005
Yule process log likelihood	18.589	17.179	0.000
S_{LN}	0.562	0.740	0.000
Covariance	0.008	0.029	0.000
Coefficient of variation	0.574	0.735	0.000

2.4 Discussion

In this study, we have found that the choice of molecular clock models has great influence on both the accuracy and precision of phylogenetic inference. It is therefore important in phylogenetics to take into account the presence of homogeneity in rate of substitution across branches. The study has shown that lineage-specific rate changes not only affect the rate estimation itself but extend to the accuracy of tree topology estimation. The application of relaxed molecular clock methods can in most cases improve the phylogenetic estimation over the use of a strict molecular clock.

From empirical observation of the phylogenetic reconstructions, it was found that the accuracy of phylogenetic estimation is negatively correlated with increasing rate heterogeneity across branches. This finding is consistent with literature, as Kuhner and Felsenstein (1994) found that phylogenetic estimation with parsimony methods is more difficult when the tree is less molecular clock-like. Where the rates of substitution were underestimated, the tree reconstruction also found difficulty recovering the true tree topology. What this again confirms is that correct estimation of rates is crucial not only for divergence time estimates, but also estimation of topology. This finding provides a greater incentive for the use of relaxed phylogenetic methods.

We presented an alternative algorithm for computing relaxed clock estimates by sampling the rates as cumulative probabilities. This implementation is equivalent to a full implementation of a relaxed clock, but can improve convergence of the MCMC. This proposal is an improvement over the current implementation of discretized branch rates used in BEAST (DRUMMOND *et al.* 2006), which has been criticized for its shortcomings (RANNALA and YANG 2007); these are specifically its lack of ability to allow for identical rates (though this was later corrected in a subsequent release), treatment of similar rates as identical rates, and its inability to accurately estimate branch rates when a small number of rate categories are used. Our results found that by discretizing the branch-rates distribution, the accuracy of phylogenetic reconstruction can decrease. Based on these results and the reasoning given in 2.2.2.2, we do not recommend discretizing branch rates when the number of taxa being compared is small (<25). The use of discretized branch rate implementations can be justified when the number of taxa is large (>30), as these methods improve convergence of the MCMC. From experience with the mammalian dataset, we estimate that the chain length required to converge M_D models is roughly half the length required to converge M_Q models.

We also proposed the idea of using the inverse Gaussian distribution as an alternative to model rates of substitution across branches. In this instance, the two-parameter models of lognormal and inverse Gaussian were more suitable overall than the one-parameter exponential model. The shapes of these two distributions appear to be more appropriate for modelling the rate heterogeneity in a larger number of mammalian genomic sequences. Another possible explanation for the superiority of two-parameter models is the added flexibility of having two parameters to control the shape of the distribution function. The flexibility allows the distribution to fit a larger range of different shapes depending on the data.

Although the comparisons performed here demonstrated differences between some of the models of rate variation, more comprehensive empirical testing is required to determine the relevance of different branch-rates models to various datasets. Also, further comparisons are required to distinguish between the more similar models of

LN and IG. As the differences in these two branch-rates models are more subtle, datasets with a larger number of taxa are required to differentiate between them.

A valid method of comparison between the different models is the calculation of Bayes factors for model selection (JEFFREYS 1961; KASS and RAFTERY 1995). The Bayes factors were not compared in this chapter as BEAST currently implements an approximation of the marginal likelihood which is known to provide extremely poor estimates (BEERLI and PALCZEWSKI 2010). The calculation of Bayes factors are further explored in Chapter 3.

It should be noted that a drawback of the inverse Gaussian distribution is that there is no closed-form for its quantile function. Though quantile values can be accurately approximated using Newton-Raphson (TJALLING 1995), it is computationally slower than those with a closed-form. In practice, using this mammalian dataset, we have found that the inverse Gaussian quantile calculation slows the entire MCMC by roughly two fold. Difficulties also exist in obtaining – from quantiles – extreme values at the tails ends of the distribution. Convergence of the Newton-Raphson in some cases could not be reached when the quantile being approximated was less than 0.01 or more than 0.99. This problem was overcome by only sampling values for quantiles within the assessable range (i.e. $\mathbf{q} \in [0.01,0.99]$). We believe that this issue will not cause any problems when modelling realistic homologous sequence data unless extremely divergent species are being compared that have massive heterogeneity in rates, for example, between species from different kingdoms.

Autocorrelated models were also found to be less suitable than uncorrelated models for this mammalian dataset. In mammals, the genetic distances between species are small as divergence times are relatively recent and rates are slow compared to viral or prokaryotic datasets (BRITTEN 1986). Autocorrelated models are known to be more appropriate for data that spans longer divergence times or where rates of substitution are faster (HO *et al.* 2007a; HO *et al.* 2005; HO *et al.* 2007b). More extensive testing is required to determine the relevance of such models across different datasets.

In this study, the value of large phylogenomic datasets like OrthoMaM (RANWEZ *et al.* 2007) for empirical studies was demonstrated. Datasets such as OrthoMaM provide a good empirical testing ground for models of rate heterogeneity across branches. Besides being used for benchmarking, this datasets can also be used to characterise the lineage-specific and gene-specific variations in rate both across and between mammalian genomes.

In this chapter, we have identified two improvements to current relaxed molecular clock methods. Both the aspects outlined in this chapter were crucial in mediating the developments in the following chapter. In Chapter 3, these methods are extended towards more practical applications in Bayesian phylogenetics.

Chapter 3. Model averaging and model selection in relaxed phylogenetics

Parts of this chapter have been published by the author as Li, W. L. S. and A. J. Drummond (in press). "Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics." *Molecular Biology and Evolution*.

3.1 Introduction

Recent studies by Drummond et al. (2006), Ho et al. (2009; 2007a; 2005; 2007b) and Lepage et al. (2007) have demonstrated the relevance of molecular clock models to a range of datasets. The authors of these studies have unanimously found that the most appropriate model choice is inconsistent across different datasets and is dependent on the data at hand. Although in the benchmark study in Chapter 2 it was found that certain models were on average better suited for the mammalian alignments examined, these models will unlikely be optimal for all other datasets. Considering the above, given a particular dataset, it is difficult to predict the suitable choice of molecular clock model. Nonetheless, model choice is important, as specification of inappropriate models can increase the error and bias in phylogenetic estimation.

Like any other statistical modelling technique, relaxed molecular clock methods suffer from problems of model misspecification and uncertainty. Model misspecification is a deep-rooted problem that plagues a range of applications of mathematics across the sciences and can cause errors and bias in the resulting analysis. In Bayesian phylogenetics, as with any statistical inference task, a sensible balance between

practicality and parameter-richness is required. A good model is not necessarily the most parameter-rich, but instead is a model that captures the essential features of the hypothesis being tested without introducing unnecessary error, bias and over-fitting. In a complex process such as evolution, the model will always be misspecified in the sense that all evolutionary models are severe simplifications of reality. Our aim therefore is to choose a model, or a set of models that are (1) able to test the hypothesis and (2) are most plausible given the data at hand. There are two general approaches to evaluating data in light of alternative models: model averaging and model selection. Model averaging allows the data to be evaluated by a weighted average over a set of models. The benefit of model averaging is that uncertainty in models can be incorporated into the analysis. Also, in some cases where multiple models appear to fit the data well, inferences from these models can be combined. In the case of a nested family of models, model averaging can also be used to investigate the importance of different parameters in explaining the data. Results inferred by model averaging are not based on or biased towards a single model, but rather the data itself determines which model or set of models are most probable.

In Bayesian statistics, a common approach to model averaging is stochastic sampling of the model space with Markov chain Monte Carlo (MCMC). Most methods utilise reversible jump MCMC (rjMCMC) (CLYDE 1999; HOETING *et al.* 1999), which allows the MCMC to jump between spaces of varying dimensions (GREEN 1995). However, reversible jump can be difficult to implement in some contexts. As a result, it is often not implemented in software packages for Bayesian MCMC. Another technique is Bayesian stochastic search variable selection (BSSVS). Although it is less computationally efficient than rjMCMC, it is easier to implement and has already found several applications in Bayesian phylogenetics (for example, GRAY *et al.* 2009; LEMEY *et al.* 2009).

Arguably, the most appropriate technique for selecting between two models in a Bayesian setting is the calculation of the Bayes factor (JEFFREYS 1961; KASS and RAFTERY 1995). Bayes factors (BF) can be used to evaluate evidence for one model over another. Though certain heuristics have been proposed (NEWTON and RAFTERY

1994), accurate calculation of BFs is most easily achieved by the same computational techniques as used for model averaging in an MCMC framework.

In Bayesian phylogenetics, model selection has been previously implemented for substitution models (GOWRI-SHANKAR and RATTRAY 2007; HUELSENBECK *et al.* 2004), the rate of nucleotide change (SUCHARD *et al.* 2001) and site heterotachy (PAGEL and MEADE 2008). The application of model averaging and model selection to relaxed molecular clock models has not yet been examined. In recent years, the standard approach has been to approximate the BF with estimated marginal likelihoods obtained from two independent MCMC analyses of the same data using different modelling assumptions. Software packages such as BEAST (DRUMMOND and RAMBAUT 2007) provide a posterior sample of the likelihood which can be used to estimate the marginal likelihood through computing the harmonic mean of the posterior sample. This approach can be interpreted as an approximation of the marginal likelihood using importance sampling, where the posterior distribution is the importance distribution and the prior distribution is the target distribution (NEWTON and RAFTERY 1994). However, this approximation is known to often provide extremely poor estimates of the marginal likelihood (BEERLI and PALCZEWSKI 2010), as the posterior distribution is often not a good importance distribution for the prior distribution. This is especially the case when there is a lot of data, because then the posterior typically has much smaller variance than the prior.

In this chapter, we outline a strategy for model averaging of relaxed molecular clock models under phylogenetic inference with Bayesian MCMC. Consequently, such model averaging allows for accurate calculation of Bayes factors for model selection. Instead of rjMCMC, our method employs BSSVS (GEORGE and MCCULLOCH 1993). We show that our method can improve phylogenetic estimation compared to using a single model when the underlying distribution of rates is unknown. We also demonstrate that by using our method we can accurately estimate the BFs needed to perform model selection.

3.2 Methods

3.2.1 Model averaging

In Chapter 2, we provided a means of sampling the rate of substitution on each branch as cumulative probability values. Effectively, a value of q describes the rate of a branch relative to the other branches. As a result, it is possible to change the underlying distribution of the rates without altering the ordering of the rates, and without changing the probability of the rates given the rate distribution. Given a set of values for q, rate values can be obtained for any parametric distribution for which the inverse-CDF can be easily computed.

Using MCMC, we can sample the underlying distribution itself. The sampling mechanism is based on stochastic search variable selection (SSVS) (GEORGE and MCCULLOCH 1993). We define an indicator variable, $i \in \{1,2,...,N\}$, where N is the total number of branch-rates distributions to be considered. Each value of i refers to a different branch-rates distribution, F_i with a set of associated distribution parameters, $\boldsymbol{\omega}_i$ which models the underlying distribution of rates across all branches. Accordingly, i, along with all of the distribution parameters $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_N\}$ are sampled in the MCMC.

The probability of the sequence data can be computed given the tree and the values of $\mathbf{q}$ and F_i. Consider a Markov chain at time k, with a state of $\phi_k = \{i, \mathbf{q}, \boldsymbol{\Omega}, \tau, \mathbf{t}, \theta\}$. If a new branch-rates distribution is proposed, F_j (i.e., $\phi_k' = \{j, \mathbf{q}, \boldsymbol{\Omega}, \tau, \mathbf{t}, \theta\}$), it will be accepted, with probability

$$\alpha = \min\left(1, \frac{P(D \mid \tau, \mathbf{t}, \mathbf{q}, j, \boldsymbol{\Omega}_j) P(j)}{P(D \mid \tau, \mathbf{t}, \mathbf{q}, i, \boldsymbol{\Omega}_i) P(i)}\right) \tag{3.1}$$

where $P(i)$ and $P(j)$ are the prior probabilities of models i and j, respectively. Consequently, the proportion of samples that have a particular value of the indicator variable i will estimate the posterior probability of the corresponding branch-rates

model. Also, the resulting posterior distribution of trees will be a model-averaged posterior distribution, weighted by the probabilities of the branch-rates models considered.

In our implementation of the method, we use a uniform prior on i (i.e. $P(i) = 1/N$), where we assume that there is no prior knowledge as to which model is preferred; thus the prior probability of each model is equal, and the ratio of the posterior probabilities of two models is equivalent to the Bayes factor. This Bayes factor calculation is further explained in the section that follows.

3.2.2 Bayes factor calculation

The model averaging above provides a means for BF calculation. For a particular branch-rates model F_x, its posterior probability can be expressed as $\Pr(F_x \mid D)$. For two different branch-rates model F_i and F_j, the general calculation of BF is:

$$\mathrm{BF}_{ij} = \frac{\int \Pr(D \mid \theta_i, F_i)\, f_i(\theta_i)\, d\theta_i}{\int \Pr(D \mid \theta_j, F_j)\, f_j(\theta_j)\, d\theta_j} = \frac{\Pr(D \mid F_i)}{\Pr(D \mid F_j)} \tag{3.2}$$

where $f_x(\theta_x)$ is the prior distribution on the parameters of model F_x. Using Bayes' theorem (Equation 1.2), we can rearrange the calculation as:

$$\mathrm{BF}_{ij} = \frac{\Pr(D \mid F_i)}{\Pr(D \mid F_j)} = \frac{\dfrac{\Pr(F_i \mid D)\Pr(D)}{\Pr(F_i)}}{\dfrac{\Pr(F_j \mid D)\Pr(D)}{\Pr(F_j)}} = \frac{\Pr(F_i \mid D)\Pr(F_j)}{\Pr(F_j \mid D)\Pr(F_i)} \tag{3.3}$$

As the prior probabilities, $\Pr(F_i)$ and $\Pr(F_j)$ were defined *a priori*, the calculation of BF is possible with a single model-averaged run. In our case where a uniform prior is used and the prior probabilities are equal, the calculation can be further simplified as:

$$\mathrm{BF}_{ij} = \frac{\Pr(F_i \mid D)}{\Pr(F_j \mid D)} \tag{3.4}$$

The calculation above is for an example of BF calculation when averaging between two models. Newton and Raftery (1994) demonstrated that this calculation of BF for any two models is identical for any given N, provided uniform priors are used.

In Kass and Raftery (1995), the authors provide guidelines for interpreting the Bayes factors in terms of hypothesis testing. The interpretations of a BF for evidence against branch-rates model F_i are shown in Table 3.1.

Table 3.1 Guidelines for interpretations of Bayes factor values as defined by Kass and Raftery (1995).

Value of BF	Evidence against F_i
< 1	Negative support
1 to 3	Barely worth mentioning
3 to 20	Positive
20 to 150	Strong
>150	Very strong

3.2.3 Algorithm implementation

Our models and relaxed clock implementations were written in Java 1.5, and is part of the BEAST (DRUMMOND and RAMBAUT 2007) software package. In this chapter, all relaxed molecular clock models compared were implemented with our method of sampling values of rates as quantiles (M_Q models) outlined in Section 2.2.2.3.

3.2.4 MCMC priors

The priors for the relaxed molecular clock models and branch-rates distributions were consistent with those used in Chapter 2 and are outlined in Section 2.2.6. In addition to the priors for relaxed molecular clock models, a Yule pure birth process was used as a prior on the speciation process (EDWARDS 1970; YULE 1924). For analyses that used the Hasegawa-Kishino-Yano (HKY) nucleotide substitution model (HASEGAWA *et al.* 1985), a $1/x$ prior was placed on the transition-transversion parameter, κ. In

analyses where the general time-reversible (GTR) model (TAVARÉ 1986) was used, a $1/x$ prior was placed on the relative rate parameters.

3.2.5 Proposal kernels for MCMC

The priors for the relaxed molecular clock models and branch-rates distributions were consistent with those used in Chapter 2 and are outlined in Section 2.2.8. The following additional proposal kernels were used for the parameters of the model-averaged models:

- Uniform Integer Operator on the indicator variable i
- Scale operators on the substitution frequencies of the GTR model and κ of the HKY model

3.3 Results

3.3.1 Simulated data

We decided to benchmark our model-averaged MCMC in terms of how well it could recover the true underlying distribution of the rates. We used a balanced tree of 32 taxa plus an outgroup to simulate sequence alignments. The divergence times on each branch were all set to 5 time units, except the outgroup branch which had a length of 30 to make the tree ultrametric. For each of the branches on a tree, we assigned a rate of substitution drawn from either an exponential (E) distribution with a mean of 0.005 (D_E) or a lognormal (LN) distribution with mean of 0.005 and variance = 0.004 (S^2 parameter of 0.5) (D_{LN}). The rates assigned to the branches on the simulated trees are uncorrelated rates rather than autocorrelated. 100 realisations of rates were simulated under each of the two models D_E and D_{LN}. Alignments of 1000 nucleotides in length were generated from each of the 200 trees using Seq-gen (RAMBAUT and GRASSLY 1997) under a Hasegawa-Kishino-Yano (HASEGAWA *et al.* 1985) nucleotide substitution model with gamma-distributed rate heterogeneity across sites (YANG 1994) (HKY + Γ model) with a transition-transversion ratio of 3.0 and shape

parameter of 0.5. The model parameters used here are identical to those used in the simulation studies of Ross *et al.* (2008).

Each alignment was analysed with BEAST using a model-average of LN and E ($M_{LN,E}$). The mean for both distributions was fixed at 0.005 but the standard deviation parameter of LN was estimated by MCMC. The tree topologies were constrained to the true tree topology but divergence times were estimated. A HKY + Γ nucleotide site model was used in the analysis, with model parameters estimated. The analyses were each run for 50,000,000 steps with the initial 5,000,000 steps discarded as the burn-in. The convergence of the chains was verified by checking that all effective sample sizes (ESS) were greater than 100.

The results of this analysis demonstrate that our model averaging method generally yields a higher posterior probability for the true underlying model that the rates were drawn from (Figure 3.1). For the 100 alignments with rates drawn from D_E, 83 had a higher posterior probability for E than for LN and the mean posterior probability of the E model was 0.77. For the D_{LN} alignments, all 100 of the runs had a higher posterior probability for LN and the mean posterior for the LN model was 0.999. Out of the 100 runs where the rates of substitution were drawn from D_E, 97 of the analyses contained the true distribution in the 95% credible set of models. For alignments with rates of substitution drawn from D_{LN}, all 100 of the runs contained the true distribution in the 95% credible set. In this specific set up, it appears the model-averaging technique is better able to predict the underlying rate distribution when the true distribution is LN than it is when the true distribution is E.

The BF for each of the runs was calculated and used to interpret the support, under the criteria outlined in Kass and Raftery (1995), for the true underlying branch-rates model of the data (Figure 3.2a). In most cases, there is support for the true underlying branch-rates distribution as simulated in the data. For D_E, 67% showed positive support for the data being E distributed, while only 17% showed some degree of negative support (thus support for the incorrect model). For D_{LN}, all the analyses had strong or more support for the LN model, 71% of which showed very strong evidence for the data being LN distributed (Figure 3.2b).

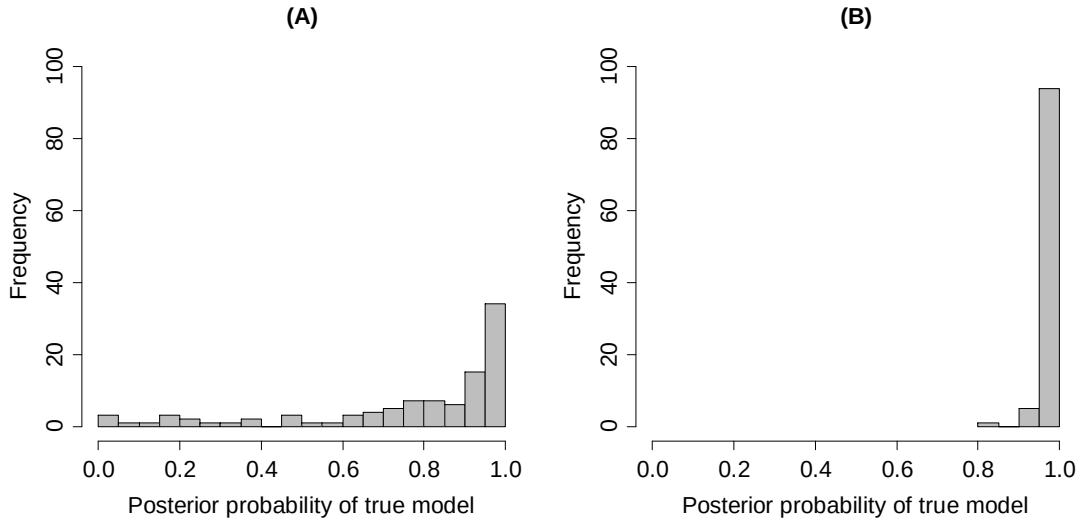


Figure 3.1 Histograms showing the distribution of posterior probabilities from using our model-averaged model on the simulated dataset.

(a) The posterior probabilities of the E distribution when the rates were simulated under D_E. (b) The posterior probabilities of the LN distribution when the rates were simulated under D_{LN}.

These results indicate that our model averaging method is capable of retrieving the true underlying distribution of the rates of substitution. In particular, the statistical power of this method is demonstrated by the fact that both sample distributions D_E and D_{LN} have comparable variances (0.005 and 0.004, respectively), as well as relatively similar density and distribution functions (shown in Figure 3.3). Hence, even when we are drawing rates from two fairly similar distributions, our method is able to differentiate between them.

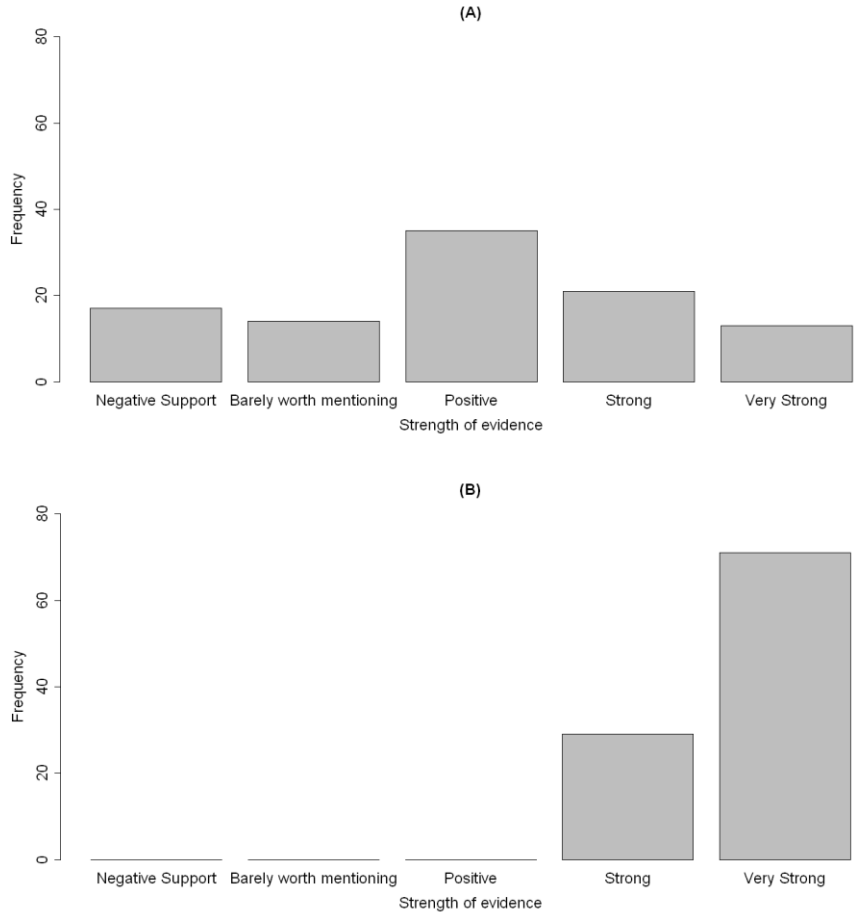


Figure 3.2 Interpretations of the Bayes factors for the (a) D_E and (b) D_{LN} data for support of the distribution the rates were actually drawn from.

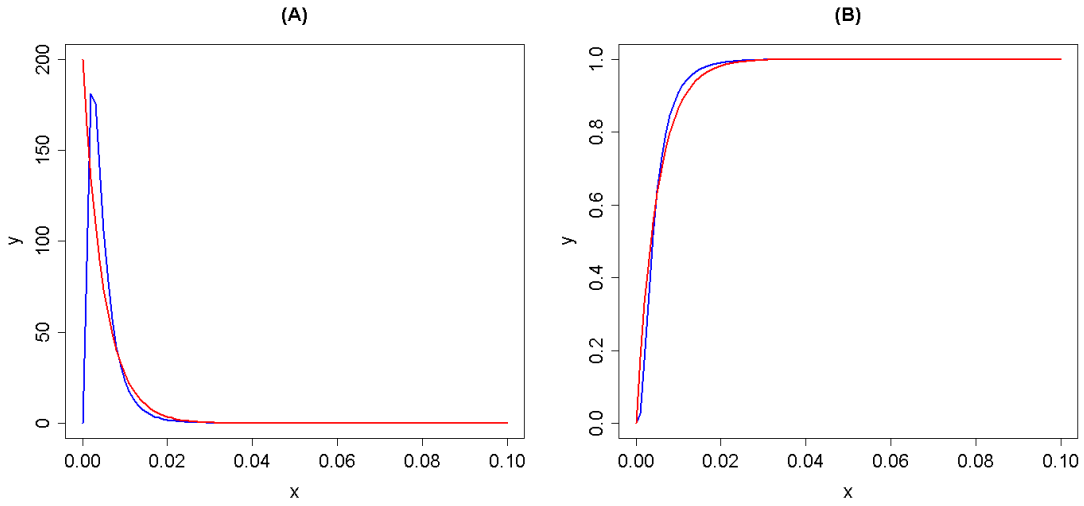


Figure 3.3 The (a) probability density function and (b) cumulative distribution function of the distributions used to generate D_E (red) and D_{LN} (blue).

D_E is characteristic of an exponential distribution with λ = 200 and D_{LN} of a lognormal distribution with M = -5.548 and S^2 = 0.5.

3.3.2 Mammalian data

We again tested our methods on the OrthoMaM 4.0 dataset specified in Section 2.2.4, which contains 1056 genes shared across 12 mammalian species. For each of the analyses using these 1056 sequences, we ran BEAST for 50 million steps, with the first 10% of the MCMC samples treated as burn-in and discarded. If all parameters of interest had not converged after the run (with ESSs above 100), we reran the MCMC for 200 million steps. In some of the genes, convergence in the chains could not be obtained even after 200 million steps. Given the data, models and proposal kernels, the MCMC runs could not be made to converge within practical running time. Consistent with the analysis in Chapter 2, in alignments where this occurred in one or more of the models being compared, the genes were excluded from the analysis. This issue occurred in approximately 10% of the analyses for each model. Again for this analysis, relative rates were estimated rather than absolute rates as no time-calibration data was available. The mean rate of the branch-rates distributions was set to 1 and the accuracy in terms of correctness in topology estimation was compared.

The dataset was first analysed using MCMC runs with a model-average of the LN and E distributions ($M_{LN,E}$). We assumed a GTR model of nucleotide substitution (TAVARÉ 1986) on the data with gamma distributed heterogeneity across sites (YANG 1994) (GTR + Γ). For this analysis the tree topology was not constrained.

One question we wanted to answer was whether the use of model averaging improved the quality of phylogenetic estimation. As there is no known accurate time calibration data available for this dataset, it would be difficult to benchmark our method in terms of rate estimation. Again we benchmark the models in terms of their ability to recover the true tree topology of the taxa.

We compared our analysis with $M_{LN,E}$ on the mammalian dataset to the same analysis with three other settings: using a relaxed clock model with only a lognormal distribution (M_{LN}), only an exponential distribution (M_E), and with a strict molecular clock model (M_{SC}). Table 3.2 shows a summary of the statistics of the analysis using

each of the models. Results indicated that $M_{LN,E}$ and M_{LN} estimated the true tree topology as the point estimate (TTPE) significantly more often than M_E (*p*-value < 0.0001 in both comparisons; non-parametric *t*-tests). The fact that the results of any of the three relaxed molecular clock models only captured the true tree on average ≈85% of the time in the 95% credible set suggests that there is some degree of model misspecification; though the model misspecification is not limited only to the molecular clock but is also contributed by misspecification of other aspects of the evolutionary model. However, realistically we would not expect the evolutionary models used to capture all characteristics of the evolutionary process. Model averaging does not improve the estimation of tree topology in this dataset, when compared to the better of the two models M_{LN}, but does significantly improve performance of the point estimate compared with the M_E model. This suggests that model averaging can protect against poor inference when the correct model is not known. Although M_{SC} often chooses the correct true tree as the point estimate, it significantly more often fails to contain the true tree within the 95% credible set (TT95) (*p*-value < 0.0001).

Table 3.2 Statistics related to the estimation of topology and rate for the mammalian dataset (n=870).

Model	Proportion of times true tree is point estimate (TTPE)	Average posterior probability of true tree (PPTT)	Average number of unique trees in 95% credible set (NU95)	Proportion of times true tree appears in 95% credible set (TT95)
$\mathbf{M_{LN,E}}$	0.513	0.415	98.198	0.857
$\mathbf{M_{LN}}$	0.533	0.437	76.997	0.860
$\mathbf{M_{E}}$	0.462	0.359	124.316	0.853
$\mathbf{M_{SC}}$	0.530	0.493	7.151	0.731

We further examined the biological relevance of different parametric distributions as models of rate heterogeneity in real data. We observed the relative posterior probabilities of each of the two distributions, LN and E, across the analyses of the mammalian genes. From Figure 3.4a, we can see that in a majority of the genes LN is preferred over E. The mean posterior probability of this LN model was 0.563 (hence mean posterior of E = 0.437). In 171 of the 954 genes we compared, the $M_{LN,E}$ model had a posterior probability of over 0.95 for one of the models; 146 (85.4%) of these

genes showed strong preference for the LN model, and 25 (14.6%) showed strong preference for the E model. Our results suggest that on average the LN distribution better models the rates of substitution across branches in this mammalian dataset.

A comparison was then done to identify the differences in the rates found by LN and E. Table 3.a shows the mean branch rates found in trees that had a posterior probability of over 0.95 for LN and for E. It can be seen that the distribution of the relative rates estimated differs significantly between the two models (6 of 21 tests were significant with adjusted *p*-value<0.05, using non-parametric *t*-tests), in which case the choice of different branch-rates model is often necessary to better model the rates.

Table 3.3 The mean branch rates found in each of the two models LN and E for trees that had strong support for a single model.

The number of samples used were 146 and 25 for LN and E, respectively. *p*-values of the non-parametric *t*-tests were adjusted with the false discovery rate correction of Benjamini and Hochberg (BENJAMINI and HOCHBERG 1995)

	Canis	Felis	Homo	Pan	Pongo	Macaca	Microcebus
LN	1.128	1.054	0.940	0.938	0.988	1.003	0.977
E	1.383	1.478	1.245	1.168	1.118	0.855	1.217
Adjusted *p*	0.033	0.012	0.173	0.301	0.658	0.466	0.173

	Otolemur	Mus	Rattus	Ochotona	Oryctolagus	Internal Node 1	Internal Node 2
LN	1.210	1.094	1.087	1.283	0.984	1.068	0.975
E	1.220	1.479	1.316	1.506	1.627	1.345	1.418
Adjusted *p*	0.966	0.071	0.173	0.311	0.004	0.173	0.194

	Internal Node 3	Internal Node 4	Internal Node 5	Internal Node 6	Internal Node 7	Internal Node 8	Internal Node 9
LN	0.925	0.883	1.032	0.849	1.680	1.072	0.869
E	1.107	1.323	1.309	1.492	1.123	1.066	1.672
Adjusted *p*	0.386	0.091	0.173	0.033	0.006	0.966	0.004

We then expanded on this analysis to a comparison of more than two models. We ran the model-averaging analyses on the mammalian data again, except this time including the inverse Gaussian (IG) distribution, along with the LN and E ($M_{LN,IG,E}$). The posterior probabilities of each of these three distributions for the analysis are shown in Figure 3.4b. The mean posterior probabilities of the LN, IG and E models were 0.314, 0.382 and 0.304, respectively, indicating that there was a slight preference of the IG model over the other two models. The exponential model had a posterior probability greater than 0.95 in 10 of the 967 genes. On the other hand in

238 of the genes, E was not contained in the 95% credible set. However, there were no genes that contained only LN or only IG in the 95% credible set, but rather when the support for the E distribution was low, the 95% credible set contained relatively similar posterior probabilities for LN and IG. This suggests that the characteristics of LN and IG are more similar to each other than they are to E, hence the model averaging did not distinguish between them.

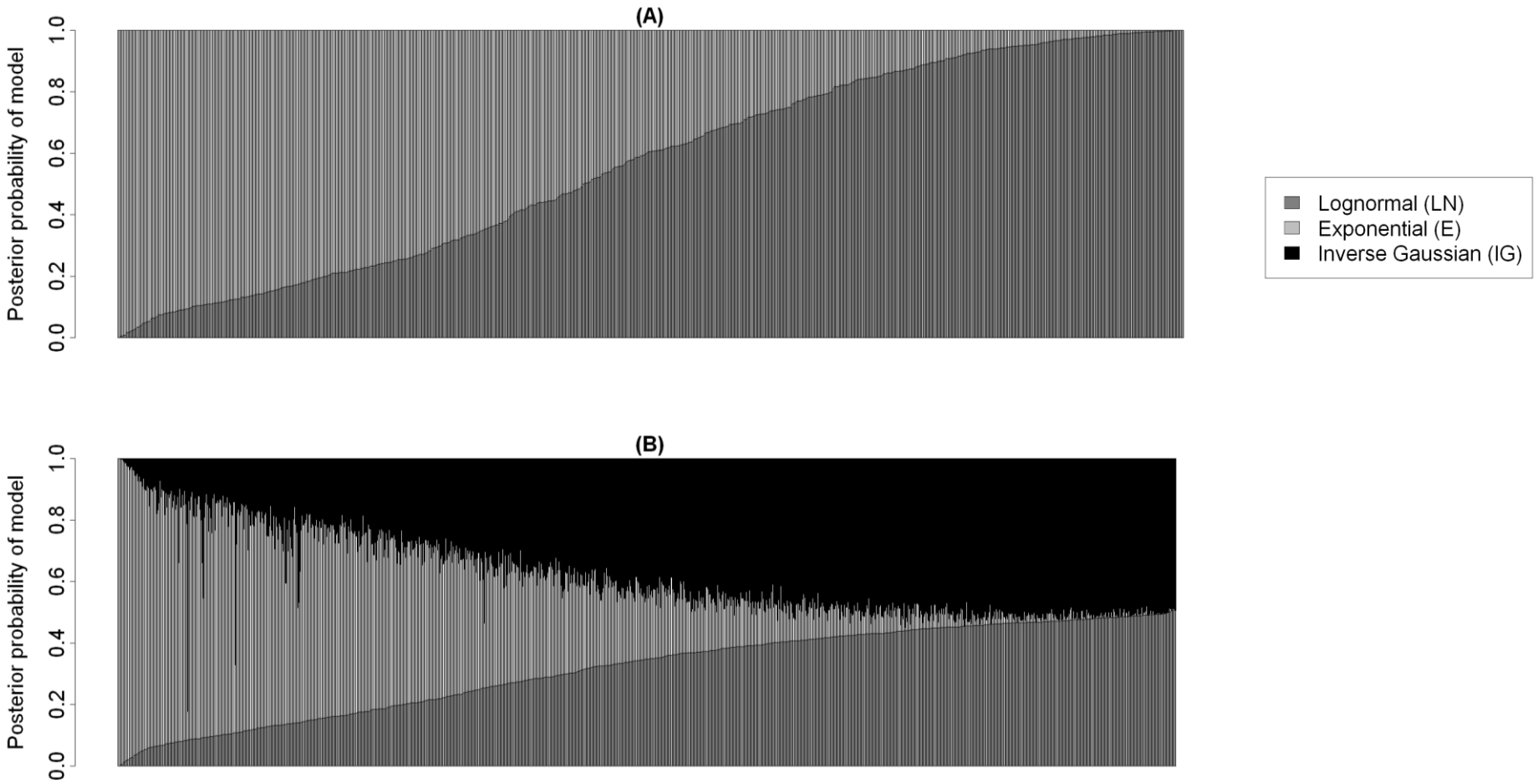


Figure 3.4 Bar plots showing the distribution of posterior probabilities of each distribution when applying our model-averaged models on the mammalian dataset. (a) $M_{LN,E}$ (n = 954) and (b) $M_{LN,IG,E}$ (n = 967). Data is sorted by the posterior probability of LN.

3.3.2.1 Indirect assessment of Bayes factor computations

A simple consistency check of Bayes factor computations involves computing the Bayes factor between two models, with or without a third model in the mixture. If the implementation is correct and the MCMC is run sufficiently long, then the BF computed for a pair of models should be the same whether a third model is present in the set of models to average. We used this fact to test the consistency of our BF estimates. For the $M_{LN,E}$ and $M_{LN,IG,E}$ models that we ran, the log ratio of the posterior probabilities between LN and E: $P(F_{LN} \mid D) / P(F_{E} \mid D)$ was calculated. Because our priors were uniform, the ratio of the posteriors is equivalent to the BF (KASS and RAFTERY 1995). This comparison was carried out as it is known that IG has similarities to LN (TAKAGI *et al.* 1997), and thus the covariance between them is likely to be lower than each of their covariances to E. Comparisons between the BF for each gene are shown in Figure 3.5. Even with the introduction of a third distribution in the $M_{LN,IG,E}$ model, the BF between the LN and E distributions were mostly very similar. The coefficient of determination (R^2) between the log BFs of the two models was 0.925, indicating a very strong correlation between the log BFs (*p*-value < 0.001). It is clear from manual inspection of the plot that variability in the estimate of the BF increases with the log BF. This is to be expected as log BF < -3 or log BF > 3 represents strong support for one model over the other, meaning that the less probable model is seldom sampled, and the relative error of the estimate will be greater.

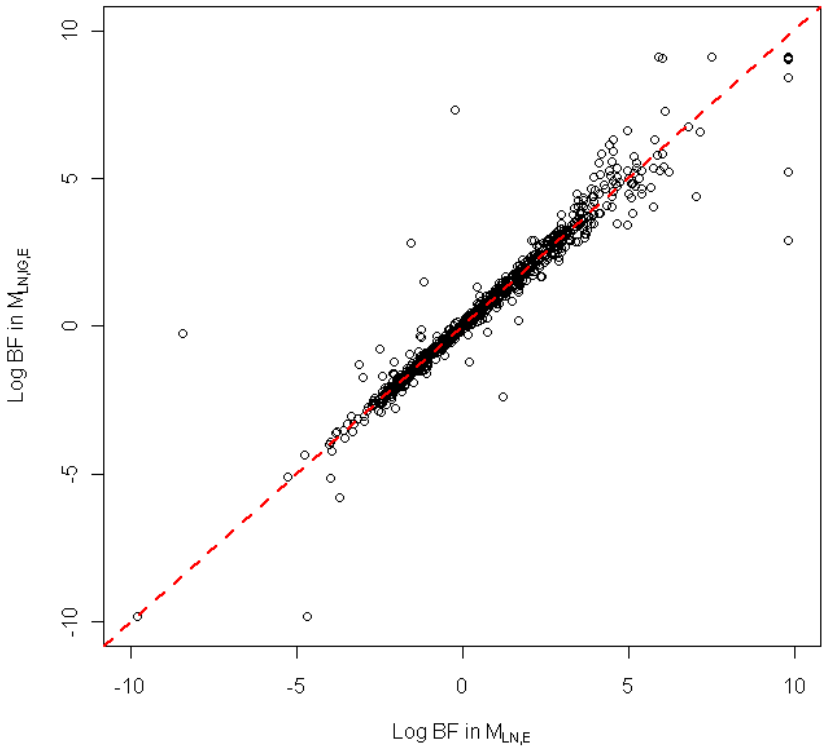


Figure 3.5 Scatter plot of the log Bayes factors of F_{LN} against F_E in the $M_{LN,E}$ model against the same value in the $M_{LN,IG,E}$ model.

Where the posterior probability of a model was zero, we assigned a probability of 0.5/9001 (≈ 0.00005; 9001 is the total number of sampled trees), which is used as a minimum value. 895 genes were used in this comparison.

3.3.2.2 Bayes factor calculation

We then compared the values of the BFs as computed with our model averaging against the approximated value of the Bayes factor as defined by Newton and Raftery (NEWTON and RAFTERY 1994). For simplicity, the two methods of BF calculation were compared with the $M_{LN,E}$ model. As Newton and Raftery's method is an approximation, we can assume any significant differences to our values to be due to sampling error in the importance sampling method. As can be seen in Figure 3.6, the values calculated via approximation appeared to be relatively conservative, while the BFs calculated by model averaging tend to have much larger variation (standard deviations of 1.21 and 1.96, respectively). We removed the presence of outliers by computing leverage coefficients and removing values that were considered significant (HOAGLIN and WELSCH 1978; SOKAL and ROHLF 1995). The R^2 calculated from the outlier-omitted data was 0.043, indicating a lack of correlation between the logged

values of the two BF computations. This suggests that the calculation of BFs with importance sampling provides inaccurate estimates that do not reflect the actual BF, as has been previously shown in other Bayesian phylogenetics contexts (BEERLI and PALCZEWSKI 2010).

The supports of the BFs in terms of hypothesis testing for the lognormal model across the genes are shown Figure 3.7. 504 of the 878 genes compared indicated support for LN, 326 of which showed positive or more support, and 38 of which had very strong evidence against the data being exponentially distributed. For support of the E model, 374 genes showed some support, 213 of which showed positive or more support for the model, while only one gene had very strong evidence against the data being lognormally distributed. What these interpretations essentially show is a quantification of the posterior probabilities that were calculated, which can be used as an assessment for model selection. This demonstrates that calculation of BFs can be used for model choice based on established statistical frameworks for hypothesis testing.

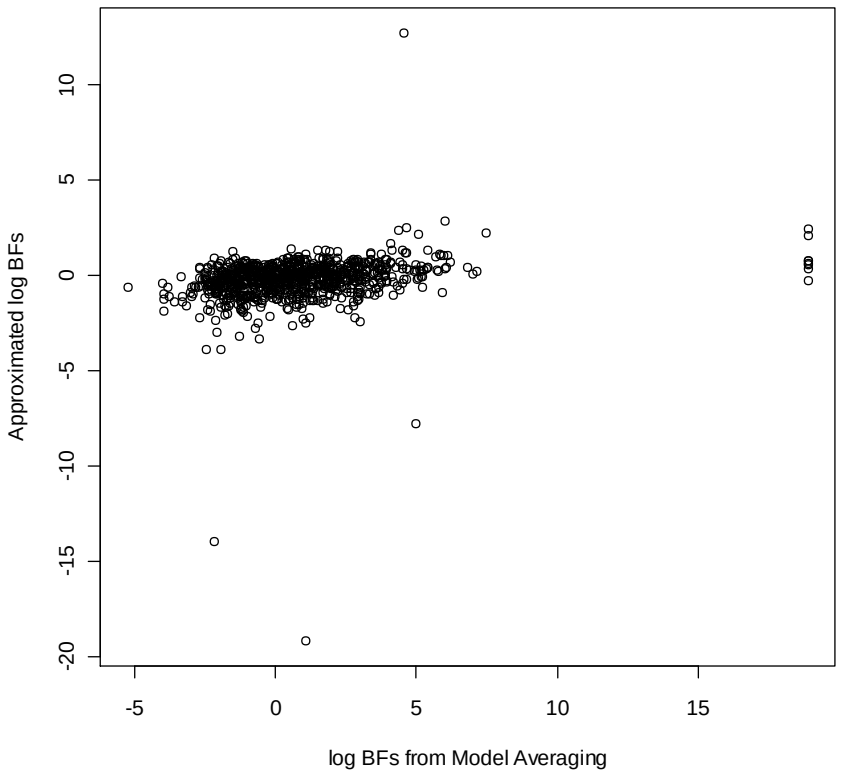


Figure 3.6 Scatter plot of Bayes factor values for the mammalian data calculated by using model averaging versus using an approximation with importance sampling (n = 878).

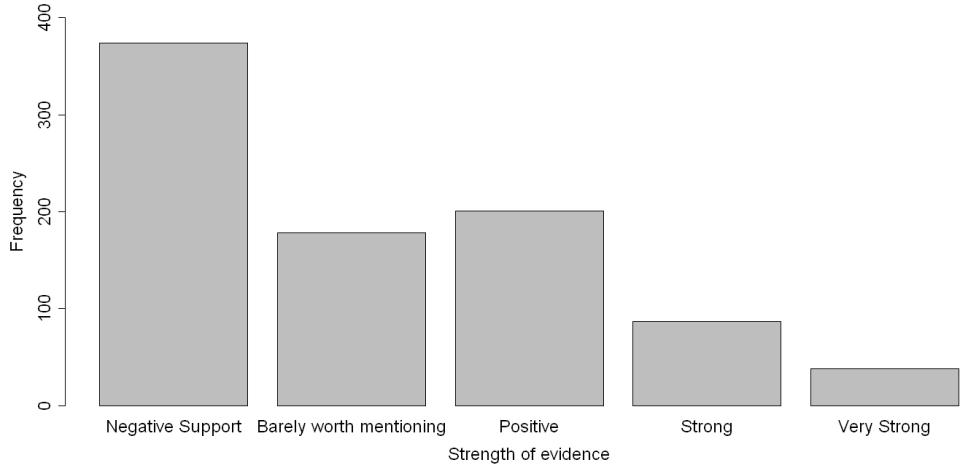


Figure 3.7 Interpretation of the BF for support of the F_{LN} model in the mammalian dataset in $M_{LN,E}$ (n = 878).

3.3.2.3 The effect of prior choice on Bayes factor estimates

We investigated the use of different prior probability specifications to model averaging and BF calculations with relaxed molecular clock models. In the analyses in Chapters 2 and Chapter 3 so far, different priors were used for the variance parameter of the LN and IG distributions. The F_{LN} model had a Uniform(0, 10) prior on the S parameter while the F_{IG} had a Uniform(0, 10) prior on the variance, σ. The upper limit for σ in F_{LN} was therefore actually 5.18×10^{21}, much higher than that of the F_{IG} (this is discussed in detail in Section 2.2.6). In practice, for our analyses in Chapter 2, the maximum σ was found to be 2.3, and so the values close to the upper bound of F_{LN} are unreasonable.

Due to the current proposal mechanisms of model averaging, values for the distribution parameters of each model are sampled even when i does not specify that distribution. For a model F_j, when sampling unlikely values for $\boldsymbol{\omega}_j$ while $i \neq j$, these unlikely values can be accepted with higher probability than when $i = j$ as the change in likelihood is zero (likelihood ratio = 1). Specifically for F_{LN}, it is more likely to accept unreasonably large values of σ when $i \neq$ LN. When these large σ values are in use, a proposed move in i from another distribution to F_{LN} will have a

low acceptance rate. The posterior probability for the LN distribution $P(F_{LN} \mid D)$ may therefore be lower than when a more reasonable range for the prior is used.

The parameterisation of the LN distribution in BEAST was modified to allow equal prior distributions for the σ of LN and IG distributions, with both distributions now sampling σ from a Uniform(0, 10) distribution. The analyses of $M_{LN,E}$ and $M_{LN,IG,E}$ was rerun with 110 randomly chosen alignments. Out of these 110 alignments, 102 converged in each of $M_{LN,E}$ and $M_{LN,IG,E}$ with ESSs above 100 in all variables. Alignments were retained where convergence was reached under both the results where a uniform prior on S in F_{LN} and a uniform prior on σ in F_{LN} were used. 95 genes and 101 genes were retained for $M_{LN,E}$ and $M_{LN,IG,E}$, respectively.

Figure 3.8 shows a comparison of the BFs for support of the F_{LN} against F_{IG} under both equal and unequal priors. The results suggest that by changing the prior on the LN distribution to have a more reasonable range, the posterior probability and BF support for the LN model were improved. Where values of log BF < 3, $P(F_{LN} \mid D)$ was higher with a uniform prior on σ than a uniform prior on S. However, the posterior probability decreased when high support was seen for the LN model (i.e. log BF > 3). The mean posterior probabilities for the LN, IG and E distributions in the $M_{LN,IG,E}$ model were 0.417, 0.297, 0.286, respectively in the 101 genes examined. Compared to 0.314, 0.382 and 0.304 when using a uniform prior on S for LN, the support for the LN distribution model was much higher and the LN was more preferable than both IG and E.

What has been demonstrated here is that the choice of priors for model parameters can have strong influence towards the posterior probabilities of each model in model averaging of relaxed molecular clock models. However, it should be noted that marginal likelihoods are by definition conditional on both the model and its associated priors, and so Bayes factors are comparisons between both the differences in models and priors of two models (KASS and RAFTERY 1995). Priors should in general be considered for each branch-rates model individually and chosen as the most appropriate for that model.

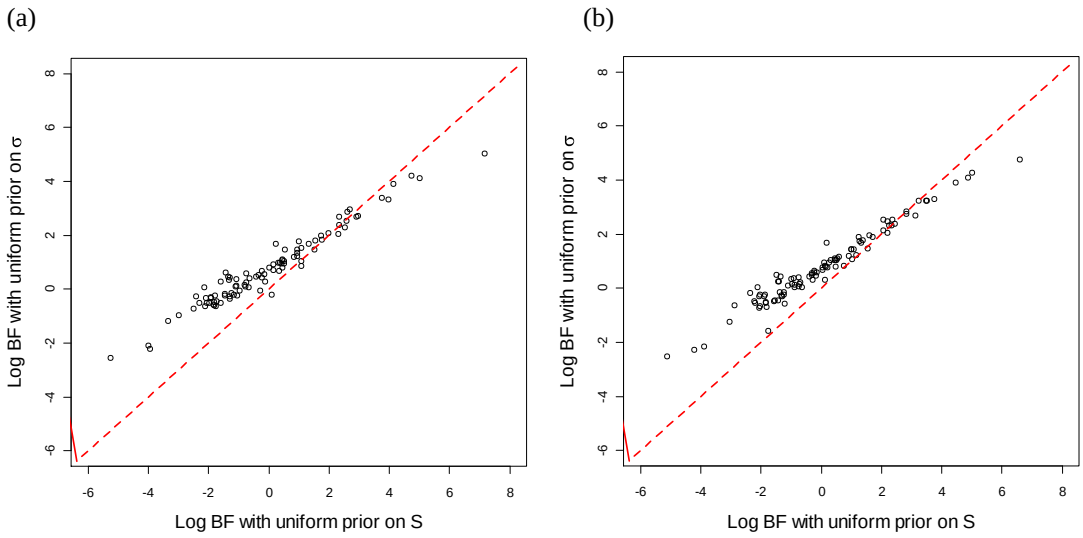


Figure 3.8 Scatterplot of the Bayes factors between F_{LN} and F_{IG} for two different prior distributions for the F_{LN} model for (a) $M_{LN,E}$ and (b) $M_{LN,IG,E}$.

The F_{LN} model was either parameterised with a uniform prior on the *S* parameter or a uniform prior on the σ. Where a Uniform(0, 10) prior on *S* was used, σ had a range of 0 - 5.18×10^{21} rather than 0 - 10 when sampling from a Uniform(0, 10) prior on σ.

3.4 Discussion

In this chapter, we introduce the model averaging of branch-rates models for relaxed molecular clock models, which permitted calculation of Bayes factors for model selection in a Bayesian phylogenetic framework. We believe that our model averaging technique is a positive step towards a phylogenetic analysis scheme that removes the burden of model selection from the user, especially for aspects of the model that are nuisance, and for which strong prior knowledge is not available. In phylogenetic estimation, a large source of bias and error comes from model misspecification. Our method can help eliminate such errors. By allowing the data to select an appropriate branch-rates model or a set of models to represent their characteristics, there is less room for error from manual parameterisation of models.

The method in its current state poses a few issues in terms of application to standard Bayesian phylogenetic analysis. One issue involves convergence assessment for the distribution parameters of the model. Under our current implementation, values of each distribution parameter are sampled regardless of whether the distributions they correspond to are involved in the likelihood calculation at that step in the MCMC.

However, when a particular model is not in use at particular steps, the convergence of its distribution parameters during those steps is irrelevant to the actual convergence of the parameter. The calculated effective sample size (ESS) therefore does not reflect the actual ESS and conventional assessment of convergence cannot be applied here. It should be noted that this problem not only appears with our method but, as pointed out by Green (1995), is also a problem in rjMCMC. Methods to solve similar convergence assessment issues in rjMCMC have been published (BROOKS and GIUDICI 1999; CASTELLOE and ZIMMERMAN 2002) which can be modified for use in our scheme. Also, it is a feasible extension to common convergence assessment programs such as Tracer (RAMBAUT and DRUMMOND 2007) to calculate ESS values conditional on the value of the indicator variable. These modifications are currently being implemented and tested for our method.

In addition, improvements to the mixing of the MCMC can be achieved by making adjustments to the proposal kernels. The way in which new parameter values are proposed, as well as proposals for sampling rates as quantile values can be modified to produce more efficient convergence of the algorithm.

An obvious extension to our model averaging technique would be to allow co-parameterisation of both uncorrelated relaxed clocks and autocorrelated relaxed clocks (ARIS-BROSOU and YANG 2002; THORNE *et al.* 1998). In phylogenetic estimation, it is often of interest as to whether rate of substitution is correlated between parent and child on a tree (HO 2009; LEPAGE *et al.* 2007). Such an extension would allow model selection between uncorrelated and autocorrelated models, which can act as a test of autocorrelation in rates of substitution for a given set of data.

In our analysis of mammalian nuclear genes, we found little evidence to separate the three parametric models compared. The analysis performed in Section 3.3.2.3 indicated that the choice of priors would alter the relative posterior probabilities of each branch-rates model. If direct assessment of the branch-rates models is of interest, then equal priors should be specified in the comparison. However, a larger number of the datasets did seem to favour the two parameter models of lognormal and inverse Gaussian over exponential. This result is in agreement to what was found in Chapter 2.

The method described here opens the way to larger empirical investigations of the relative merits of different relaxed molecular clock models. The underlying process of rate of substitution is complex, so it is unlikely that there is a single model that is optimal across all datasets. Thus, it is important to take a more liberal approach when choosing the appropriate model, and even more crucial to have a wide selection of distributions to choose from. In the BEAST (DRUMMOND and RAMBAUT 2007) software framework we have begun to implement an array of positive continuous parametric distributions to be used as models of rate distribution for model averaging of relaxed clocks, such as Gamma and inverse Gamma.

There are many aspects of modelling the rates of substitution among branches that have yet to be explored and further developments will improve the shortcomings of the methods presented. Further progress will involve the explicit incorporation of the factors that cause rate variation among lineages and we anticipate that model averaging techniques, such as those demonstrated here, will play a role in discovering the factors that are most important in this context.

Chapter 4. Covariation of rates of evolution

Parts of this chapter have been published by the author as Li, W. L. S. and A. G. Rodrigo (2009). "Covariation of Branch Lengths in Phylogenies of Functionally Related Genes." *PLoS ONE* 4(12): e8487.

4.1 Introduction

Estimating lineage-specific substitution rates and divergence dates has become an increasingly important aspect of the reconstruction of evolutionary history (DRUMMOND *et al.* 2006; KISHINO *et al.* 2001; SANDERSON 2002; THORNE *et al.* 1998). Differences in substitution rates from lineage to lineage have been attributed to variation in neutral rates of substitution, population size, generation times, and selective forces. These together are responsible for the non-ultrametric distances on a tree (BROMHAM and PENNY 2003; GILLESPIE 1991) and give rise to lineage-specific variation in molecular evolutionary rates.

More recently there has been focus on the possibility of lineage-gene-specific differences in substitution rate (SIEPEL *et al.* 2006; YANG and NIELSEN 2002). The number of substitutions acquired by a protein-coding gene may increase during periods of rapid adaptive change or decrease because of strong structural or functional constraints on the coded protein. The molecular evidence for such specific selection-mediated substitutions has been the subject of much research since the pioneering paper of Messier and Stewart (1997) (FAY and WU 2003; ROSS and RODRIGO 2002; SAWYER *et al.* 2005; SUMIYAMA *et al.* 2002; WOOLFE *et al.* 2005). These selection-

mediated substitutions are by definition non-neutral and thus would not be expected to be consistent across genes or across lineages.

Proteins do not function individually but in pathways, though this is usually ignored in models of genic evolution (THORNE and KISHINO 2002). In fact, it is reasonable to suggest that natural selection acts on a group of genes that collectively perform a biological function. Under the presence of selection, both functional and structural constraints will be expected to cause the divergence rates of functionally related genes to vary in a coordinated fashion.

Genes whose products physically interact are known to co-evolve, in the sense that there are correlated rates of substitution between genes of interacting proteins (ATWELL *et al.* 1997; JUCOVIC and HARTLEY 1996; PAGÈS *et al.* 1997; PAZOS *et al.* 1997; POUMBOURIOS *et al.* 2003). The way proteins function as physical structures can restrict the mutations that persist in the population through the action of natural selection. This is particularly evident in protein domains involved in direct physical interactions with other proteins, where protein interaction may fail if mutations that change the protein structure occur at the site of interaction. Correlated substitutions that occur within a species lineage can result in similarities in substitution rates across genes. In addition to this, different lineages undergo different extents of selection pressure for any given biological function. Due to this effect of co-evolution, the selection pressures applied to a function are reflected on many or all the genes involved in that function. These two effects in combination have been shown to cause the co-evolution of genes (JUAN *et al.* 2008a; KANN *et al.* 2009).

Accordingly, there is a resemblance in branch lengths in the gene trees of interacting protein coding genes (FRYXELL 1996). Pazos and Valencia (2001) were the first to use this observed pattern of co-evolution across species to predict the interaction between genes. In their study, they were able to predict pairwise interaction of gene products with 79% accuracy in a bacterial dataset (PAZOS *et al.* 2005). Other approaches to predicting gene interactions using co-evolution have also been devised that utilise methods similar to Pazos and Valencia (GERTZ *et al.* 2003; GOH *et al.* 2000; GOH and

COHEN 2002; JUAN *et al.* 2008b; KIM *et al.* 2004; RAMANI and MARCOTTE 2003; SATO *et al.* 2003; TAN *et al.* 2004).

We propose here that co-evolution and similarities in substitution rates across species are not limited purely to interacting gene pairs. Our hypothesis differs from that of Fryxell's (1996) in that we suggest a more general evolutionary relationship: co-evolution occurs not only specifically amongst genes that interact with each other but also amongst genes that are known to be involved in the same biological function or pathway. Co-evolution is partially driven by similarity in selective pressures acting on functionally related genes (HAKES *et al.* 2007). Also, as all genes that interact ultimately form a network in metabolic pathways, it is expected that some "contagious" correlation will extend to functionally related genes. Our argument is supported by recent studies, which show that there is correlation in patterns of evolution amongst genes involved in related biological processes (FRASER *et al.* 2002; HAKES *et al.* 2007; JORDAN *et al.* 2004; JUAN *et al.* 2008a; MARIÑO-RAMÍREZ *et al.* 2006; SHAPIRO and ALM 2008; WALL *et al.* 2005; WOLF *et al.* 2006). In particular, the recent annotation of the 12 drosophila species genome set (DROSOPHILA 12 GENOMES CONSORTIUM 2007) has shown that genes encoding functionally similar proteins have a tendency to evolve at similar rates. Recent studies by Juan *et al.* (2008b) have also found patterns of co-evolution across genes from the interactomes of the NADH-quinone oxidoreductase complex and the flagellar assembly machinery in prokaryotes, though the study did not explicitly state whether or not direct physical interactions occurred between these genes.

Though our hypothesis is supported by literature in theory and results, it has been found that genes operating within the same pathway can vary in selective pressures (LU and RAUSHER 2003; RAUSHER *et al.* 1999). A study by Rausher *et al.* (1999) and its follow up study by Lu *et al.* (2003) have demonstrated that differing selection pressures occur between upstream and downstream genes of the anthocyanin pathway in the *Ipomoea* genus. Hence it should be noted that correlation in evolutionary rates does not necessarily occur amongst proteins that are involved in the same pathway.

The aim of this study was to find how the correlation in branch length varies across the different biological functions. This matter is particularly important for phylogenetic inference and studies of comparative genomics. In particular, we aimed to determine whether the similarities in gene tree branch lengths that are seen in genes that have physically interacting gene products also exist between genes that are functionally related. As a comparison to Rausher *et al.*'s (1999) results, we attempt to determine whether the mode of selection is common within the different pathways in the species examined. In our study, we found that there is a correlation in branch lengths of genes trees from functionally related genes that do not necessarily have physical interactions. Results show that the degree of correlation varies greatly across different biological functions and across different species sets. We discuss the relevance of the findings to gene selection for the purpose of estimating species divergence times.

4.2 Materials and Methods

The aim of this study is to predict the relationship between genes that are functionally related. We hypothesise that correlation between genes can be used to infer the function when the function of some genes in a correlated set is known. The branch lengths of a tree are used here as a basis to detect changes in substitution rate across lineages. We test the effect of the correlation in four individual datasets: two mammalian datasets, a yeast dataset, and a bacterial dataset.

4.2.1 Visualising substitution patterns amongst genes and lineages as a matrix

First we consider a new scheme of visualising variation in substitution rates amongst genes and lineages that uses a matrix of gene tree branch lengths. Consider a collection of orthologous genes from a set of species. If the true species topology is known and assumed to be the same for all genes, the branch lengths of each gene tree can be estimated within a constrained tree topology. This results in a set of gene trees with the same branching patterns but optimised to have gene-specific branch lengths. We can consider a matrix, B, of dimensions $M \times N$, where M is the number of

genes, and N is the number of branches on the tree (N is equal to $2n-3$ in an unrooted tree, where n is the number of taxa). Each entry B_{ij} of the matrix represents the length of branch j in gene tree i. Figure 4.1 demonstrates an example application of this matrix schema, shown as an expression heatmap. Using the matrix representation, it is easy to visualise the relative length of branches in a particular gene, as well as genes or branches that have higher/lower expected numbers of substitution. It should be noted that the order of branches and genes in the matrix is arbitrary, but constant across all genes.

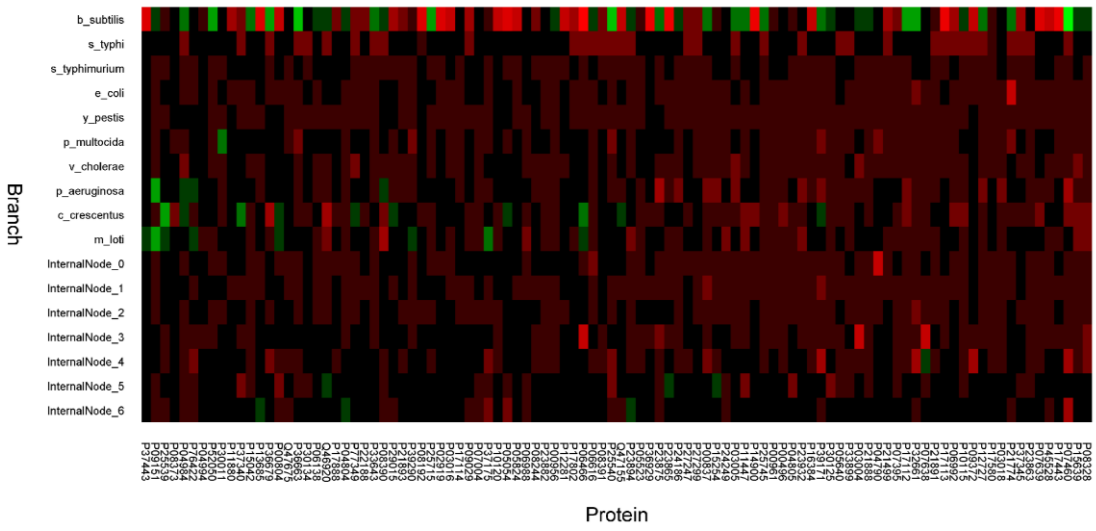


Figure 4.1 An example application of the matrix representation with a sample set of homologous proteins in 10 prokaryotic species.

The matrix is shown as a log scale expression heatmap. Each entry in the matrix is the branch length of a particular branch in a particular gene tree. Matrix entries are colour coded by their lengths relative to the average branch length (the average genetic distances in the phylogeny across all genes). The brighter green the branches, the shorter they are relative to overall species distance. The brighter red, the longer they are relative to their species distance. Using the heatmap, it is easy to identify which branches differ from average.

4.2.2 Matrix transformation

The first step of the analysis procedure was to transform the branch lengths to account for global lineage-specific effects on rate change [for example, the faster rate of evolution on the lineages of mice and rats compared to larger longer-lived mammals such as humans (WU and LI 1985)]. We introduce the branch length transformation in

matrix representation and the procedures described here are analogous to standard procedures used in data transformation in microarray analysis (RITCHIE *et al.* 2007).

First, all zero branch lengths are replaced with the minimum non-zero value in the matrix. All values of the matrix were then log transformed. The empirical distribution of branch lengths across all genes for a particular branch tends to be significantly skewed. An example of this is shown in Figure 4.2, where this distribution can be seen clearly. Matrix entries are therefore log transformed to obtain values that are less skewed.

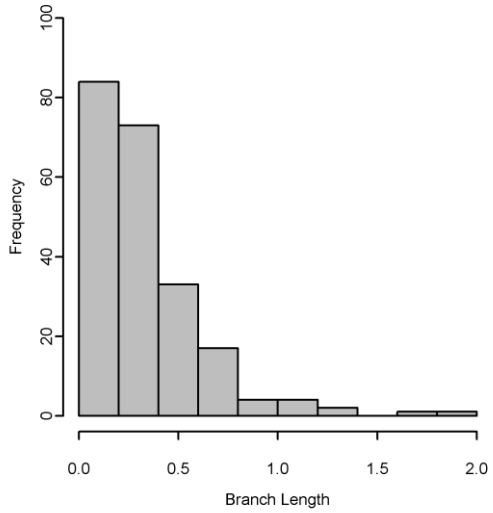


Figure 4.2 Histogram of the gene tree branch lengths on the *P. multocida* branch in the bacterial dataset.

The length of branches is approximately distributed exponentially. The lengths of other branches on the tree also follow similar distributions.

4.2.3 Statistical methods

Statistical methods were applied to the datasets examined to predict the function of a gene by its similarities in gene tree branch lengths to other genes. In doing so, the correlation between the genes can be quantified through how much of the variation in the rate of substitution can be explained by the rates of genes that are functionally related.

4.2.3.1 Generalized linear models

We use generalized linear models (GLM) as a method to predict the function of a gene by its evolution pattern. A GLM is a least squares regression method that uses a link function to model the relationship between sets of independent random variables and the response variable. Binary functions can be modelled by comparing the value predicted by the GLM to cut-offs which determine whether or not the observation is predicted to be involved in the process. A range of cut-off values can be iterated through to control for different false positive and false negative error rates.

In our case, the independent random variables are from rows of the normalised matrix B', where each variable corresponds to the log-transformed length of a branch for a given gene tree. The response variable was a binary variable representing whether the gene was involved in a particular biological function. Specifically, whether each gene is involved in the respective function is tested. By using individual binary GLMs to model each biological function, each gene can be classified as being involved in multiple functions. *Probit* was used as the link function, which transforms the predicted value from the GLM to values on a standard normal distribution before being subject to cut-off points. *Probit* was used as it was observed that the distribution of a particular branch across the set of genes approximately followed a lognormal distribution and so after undergoing log-transformation, the distribution of branch lengths would be roughly normal.

An advantage of using GLMs as our method of identifying correlation is that the method automatically takes into account variation within the same variable. Thus, the method will take into account any variation within a given branch across all the genes, such as effects from the natural species distances.

4.2.3.2 Principal component analysis

Correlation in the gene trees of each dataset was also tested using Principal Component Analysis (PCA). PCA is a method of multivariate analysis which attempts to describe the variation in a set of variables with a number of linear principle components (MANLY 2004). If a large percentage of the variation among the variables

can be explained by a small number of principle components, then there are patterns shared between observations in the dataset. Thus this would indicate that there is correlation among some of the genes in the dataset. In our study, the lengths of each branch on a tree are used as variables for the PCA and so the number of variables for each analysis is equivalent to N.

It should be noted that other methods of multivariate analysis such as linear discriminant analysis and multidimensional scaling were also considered, however the application of various statistical methods to our data was not of primary interest in this study.

4.2.4 Algorithm implementation

The software used to perform this research was written in Java 1.5 and utilises some of the functions and classes from the Phylogenetic Analysis Library (PAL) package version 1.5 (DRUMMOND and STRIMMER 2001).

4.2.5 Phylogenetic analysis

Each of the gene trees were constructed by maximum likelihood with PHYML 3.0 (GUINDON and GASCUEL 2003). Gene tree topologies were constrained to the species tree topologies that were obtained for each dataset. A general time-reversible model (GTR) (TAVARÉ 1986) with gamma distributed heterogeneity across sites (YANG 1994) and a proportion of invariable sites (REEVES 1992) was used as the nucleotide substitution model (GTR + Γ + I) for datasets where nucleotide sequences were examined. A Dayhoff + Γ + I model was used as the amino-acid substitution model (DAYHOFF *et al.* 1978; REEVES 1992; YANG 1994) for the bacterial protein dataset. Equilibrium base frequencies, substitution frequencies (for the GTR model), proportion of invariable sites and distribution shape were estimated from sequence data for each gene.

4.3 Analysis of UCSC mammalian dataset

We examined a dataset containing the genome alignments of 28 vertebrate species (*Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Otolemur garnettii*, *Tupaia belangeri*, *Rattus norvegicus*, *Mus musculus*, *Cavia porcellus*, *Cavia porcellus*, *Sorex araneus*, *Erinaceus europaeus*, *Canis familiaris*, *Felis catus*, *Equus caballus*, *Bos taurus*, *Dasypus novemcinctus*, *Loxodonta africana*, *Echinops telfairi*, *Monodelphis domestica*, *Ornithorhynchus anatinus*, *Gallus gallus*, *Anolis carolinensis*, *Xenopus tropicalis*, *Tetraodon nigroviridis*, *Takifugu rubripes*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Danio rerio*) taken from the University of California, Santa Cruz (UCSC) genome browser database (KAROLCHIK *et al.* 2003). Using the genome browser's database of known genes (HSU *et al.* 2006), we extracted blocks of the genome alignment that corresponded to the coding exons of genes (26645 genes in total). These coding exons were then concatenated to obtain the actual coding sequence of each gene. The genes were then checked for correct codon usage; more specifically genes had to begin with a start codon and end with a stop codon. The genes were also manually inspected to ensure homology between the species. We retained a set of 1043 genes shared across 7 mammalian species: *H. sapiens* (human), *C. familiaris* (dog), *M. musculus* (mouse), *O. cuniculus* (rabbit), *B. taurus* (cow), *R. norvegicus* (rat) and *M. mulatta* (macaque).

We estimated the true species topology of these 7 taxa by concatenating a random sample of the alignments and forming a concatenated alignment of 33263 nucleotide characters. The maximum likelihood topology of the alignment was then estimated using PAUP* (SWOFFORD 2003) under a GTR + Γ + I model (REEVES 1992; TAVARÉ 1986; YANG 1994). The species topology that was used is shown in Figure 4.3. We acknowledge that concatenating data is not the ideal method to estimate the species topology and can often be error prone, however, in this case the topology is generally supported by the literature (BASHIR *et al.* 2005; KAROLCHIK *et al.* 2003; NOVACEK 2001; PRASAD *et al.* 2008; REYES *et al.* 2004; STEIPER and YOUNG 2006; YODER 1997).

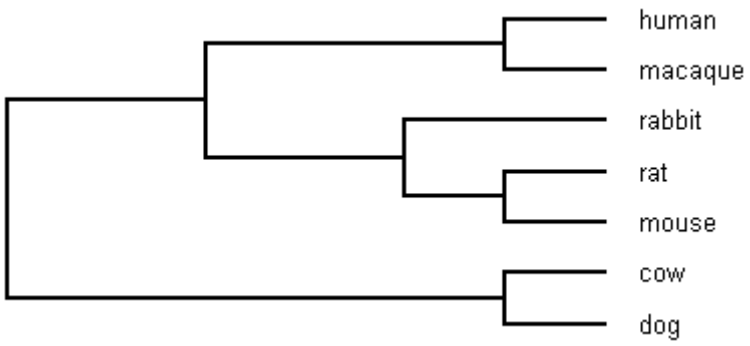


Figure 4.3 The tree topology used as the true species topology for the UCSC mammalian genome dataset.

For each of these genes, we obtained the Gene Ontology (GO) (ASHBURNER *et al.* 2000) terms for the biological processes that they were involved in. To reduce noise, genes that did not have known biological process information were omitted. The resulting dataset contained 607 genes. We then selected GO process terms that were common across 15 or more genes in the dataset. It should be noted that an assumption made here is that the biological function of each gene is identical across the species in the alignment.

Using the GO process terms, we wanted to classify the genes by their underlying biological pathways. There were two ways in which we classified whether or not a gene was involved in a particular pathway: specific terms, which are terms that are exact matches to GO process terms; and truncated terms, which are terms that contain a substring of a term relating to a certain biological pathway or process. Truncated terms allowed grouping of terms which contained similar keywords. For example, the truncated term "apopto" allowed genes labelled with terms such as 'apoptosis' and 'apoptotic program' to be grouped together. Thus some truncated terms are more general while others are more specific.

4.3.1 Results

We first computed a PCA on the data, which showed that 60.5% of the variation in the 11 variables (each variable is the log transformed lengths of a particular branch in the tree) could be explained with first two principle components. This result is

reasonably high and suggests that there is correlation among some of the genes. GLMs were then constructed to analyse the correlation between gene process involvement and relative rates of evolution across branches. For each process, the accuracy of prediction was tested using two-thirds of the data as a training set and one-third as a testing set. Specifically, the GLMs were trained with 405 randomly selected genes and tested for how accurately they could predict gene process involvement on the remaining 202 genes. The Akaike Information Criterion (AIC) was used for model selection in the GLMs (AKAIKE 1974). AIC is a likelihood-based procedure to minimise the number of variables required to make the prediction, and in doing so it attempts to prevent overfitting of data. The predictions from the GLMs were converted to estimates of whether the genes are involved in a process using a range of cut-off values, with each cut-off giving a different set of predictions. This was carried out for each GO process term to obtain the overall accuracy of the predictions.

A randomisation test was then carried out to test the significance of the predictions. For each of the processes, a subset of genes, the same size as the number of genes in the actual process, was randomly selected without replacement. GLMs were then built upon these "random processes". This was repeated 500 times to obtain a null distribution of the false positive rates and true positive rates given the data and number of genes. The area under the Receiver Operator Characteristics curve (AUC) for each of the processes was calculated with the ROCR package in R (SING *et al.* 2005) and is shown in Table 4.1. From the randomisations, we calculated the non-parametric *p*-values of obtaining the actual AUC for each GO process. *p*-values were adjusted with the false discovery rate correction of Benjamini and Hochberg (BENJAMINI and HOCHBERG 1995) to correct for multiple comparisons. A sample of the GO process terms tested is shown in Figure 4.4. For values of AUC, an area of 0.5 indicates that the classifier performs randomly. In contrast, an area of 1.0 would be achieved by a perfect classifier. It can be seen from the receiver operating characteristic (ROC) curves that functionally related genes in these processes have higher correlation than what is expected at random. Out of the 24 processes tested 5 were statistically significant. These results indicate that the similarities in gene tree

phylogeny of genes involved in related biological processes is often greater than what is expected at random.

A PCA plot of the data is shown in Figure 4.5 with the genes annotated with "regulation of transcription" (the most correlated process, according to the GLMs) coloured differently. In this case, the PCA did not seem to differentiate these genes from any of the other genes in the dataset. It appears that there are two different clusters of genes which are separated by the first principle component, however, our analysis found no functional annotations that differentiated the two groups. This correlation may either be a result of correlations between genes which common function is not defined by the GO terms used here or the correlation may be due to other factors besides function.

Table 4.1 Prediction accuracy of the GLMs for the UCSC mammalian genome data, measured by the AUC.

Truncated terms were used that allowed grouping of terms which contained similar keywords. These terms are shown in speech marks. Any variants in spelling of the processes are also taken into account (for example, protein signalling/protein signaling).

GO biological process(es)	Number of genes	AUC	*p*-value	Adjusted *p*-value
"regulation of transcription"	93	0.69	0	0.000
"transcription"	97	0.67	0	0.000
small GTPase mediated signal transduction	24	0.79	0.002	0.012
regulation of transcription, DNA-dependent	73	0.67	0.002	0.012
development/multicellular organismal development	33	0.74	0.006	0.029
"nervous system"	30	0.73	0.028	0.110
protein folding	19	0.78	0.032	0.110
electron transport/electron transfer	20	0.77	0.044	0.132
nervous system development/neurite biosynthesis/neurite formation/neurite growth	19	0.75	0.11	0.293
signal transduction/signaling/signalling	73	0.64	0.134	0.322
G-protein-coupled receptor protein signalling pathway	30	0.70	0.156	0.340
cell-cell signalling	22	0.71	0.218	0.400
"protein biosynthesis/translation/protein anabolism/protein formation/protein synthesis/protein translation"	35	0.67	0.22	0.400
protein transport	27	0.69	0.246	0.400
"apopto"	27	0.69	0.25	0.400

Table 4.1 (continued).

GO biological process(es)	Number of genes	AUC	*p*-value	Adjusted *p*-value
ubiquitin cycle	15	0.75	0.288	0.409
“brain/neuro”	17	0.73	0.29	0.409
potassium ion transport/K+ conductance/potassium transport	16	0.73	0.354	0.472
transcription	48	0.63	0.408	0.515
cell cycle/cell-division cycle	18	0.70	0.47	0.553
cell proliferation	17	0.71	0.484	0.553
apoptosis/apoptotic programmed cell death/programmed cell death by apoptosis	15	0.71	0.508	0.554
ion transport	23	0.66	0.686	0.716
transport	23	0.59	0.982	0.982

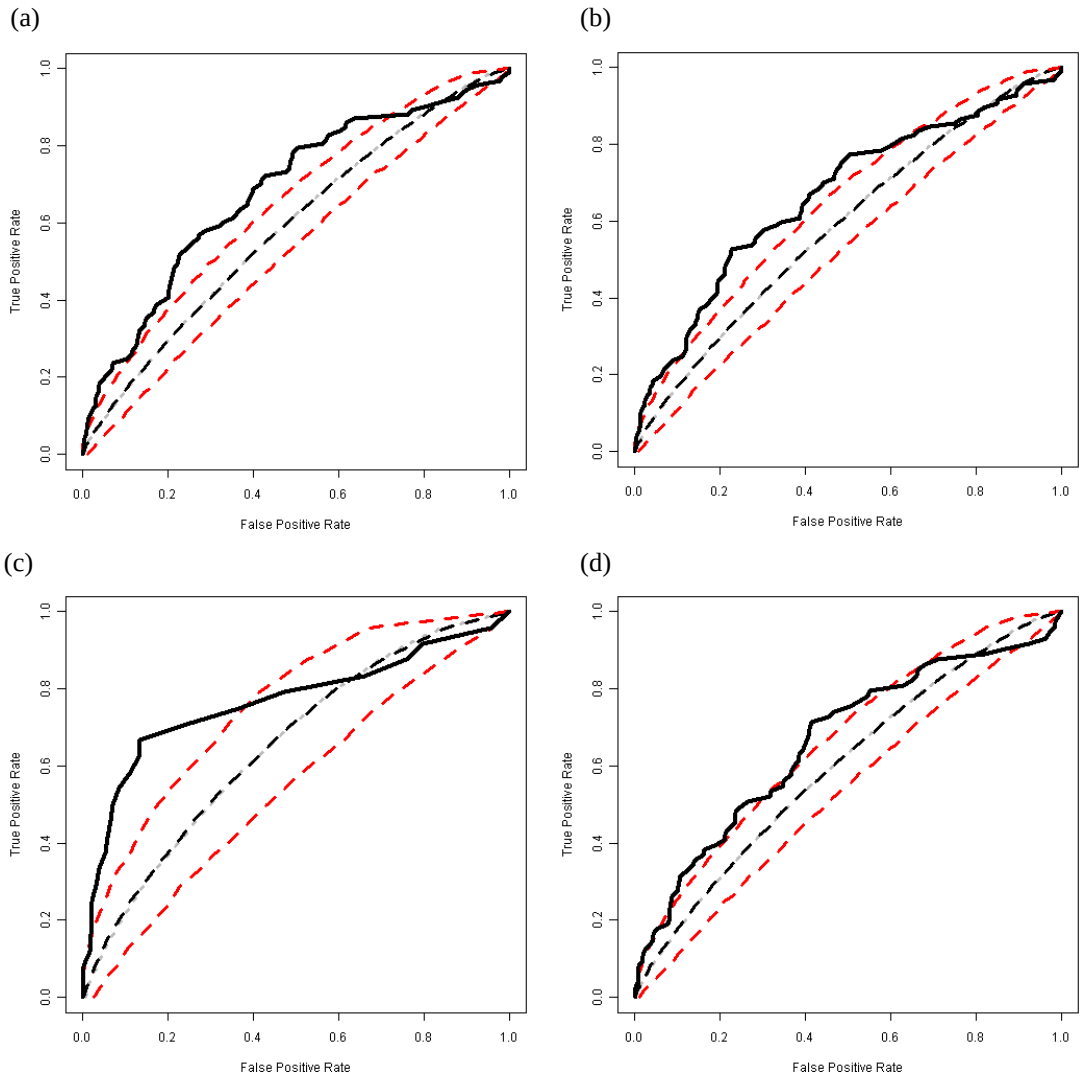


Figure 4.4 ROC curves for the best GO process terms in the UCSC mammalian dataset. (a) "regulation of transcription", (b) "transcription", (c) small GTPase mediated signal transduction, and (d) regulation of transcription, DNA-dependent.

The predictions from the GLMs of each function were estimated using different values of the cut-off point, and error rates were calculated from these predictions. The dotted black line and red lines show the median and confidence intervals of the null distribution.

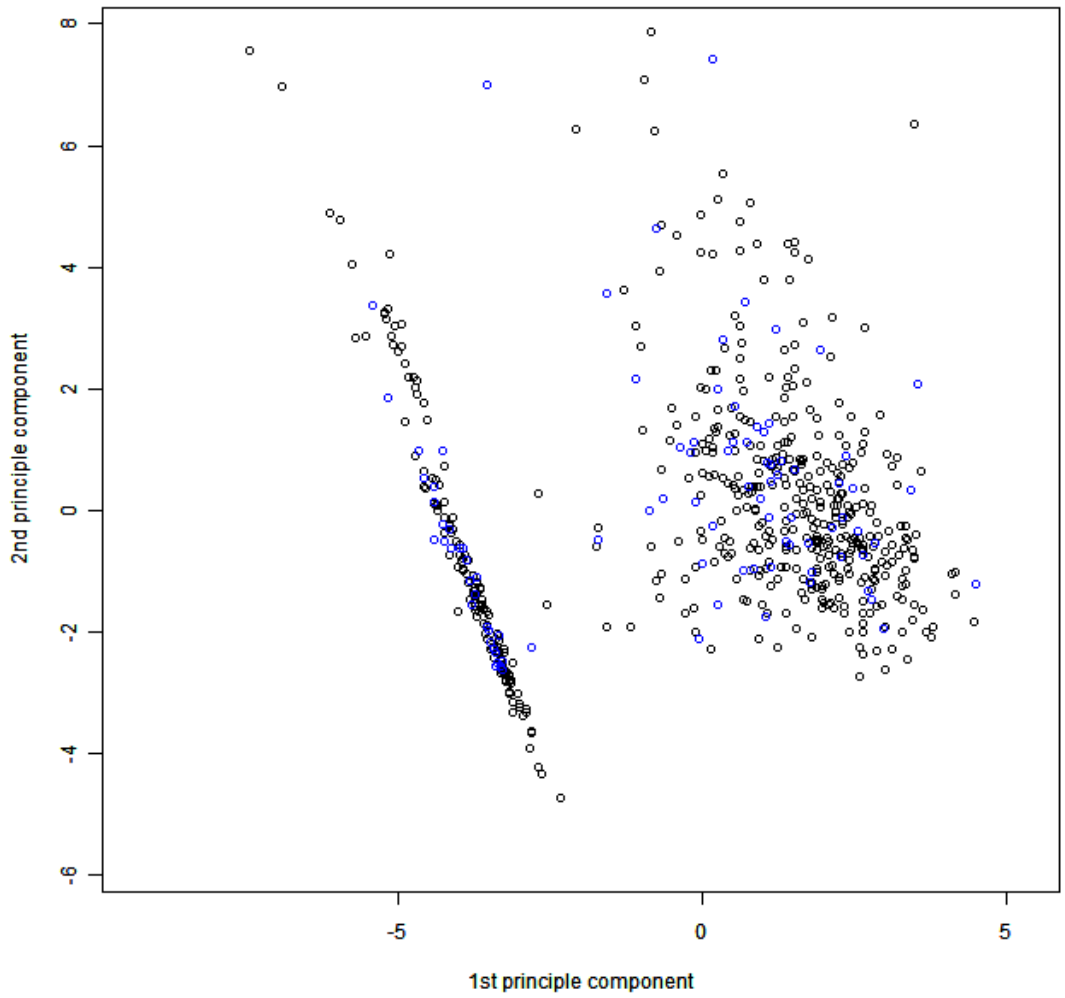


Figure 4.5 A PCA plot of the first two principle components for the UCSC mammalian dataset. Genes annotated with GO process term "regulation of transcription" are coloured blue.

In the other 19 processes, the prediction with GLMs was not significantly better than random. A plausible explanation that we propose is that for these processes, no major selection has occurred in the phylogeny studied. When this is the case, there is a lack of informative differences between the rates of genes involved in a particular process and the average rate across all genes. As there is no differentiation between the genes, the GLMs do not detect any predictable patterns and will lack statistical power.

We tried to prevent this problem from recurring in subsequent analyses by (1) increasing the number of taxa examined, as this increases the chances that at least one branch is informative or (2) choosing species where the divergence times between the species is greater, as the probability of selection having took place on a particular

branch is higher. It should be noted that for (2), the species chosen should not be too genetically divergent as changes in gene function may then occur in some of the homologous sequences.

4.4 Analysis of yeast dataset

We tested the correlation in phylogeny of functional-related genes in the yeast dataset published in Rokas *et al.* (2003b). The dataset consists of 106 homologous genes shared across eight species: seven *Saccharomyces* species (*S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus, S. castellii and S. kluyveri*), and *Candida albicans*. The species topology of these taxa is known and provided by the authors of Rokas *et al.* (2003b). The genes were annotated by their GO process terms as defined in Genbank (BENSON *et al.* 2000). Ninety-four of these genes contained GO terms and were used for this analysis. Only truncated GO terms were used in this analysis as the dataset was small and any exact terms had too few genes associated with them. Only process terms that were shared across 5 or more of the genes were used for the GLM construction.

4.4.1 Results

First a PCA was applied to the data, revealing that two principle components were able to explain 64.8% of the variation in the data (which had 13 variables). This again suggests that there are patterns in the data and that correlation is present among the samples. Rather than using two-thirds of the data as a training set and the rest as a testing set, a leave-one-out test was employed for training GLMs for this dataset. A leave-one-out test is advantageous in small datasets like this where the training set would otherwise be too small to adequately train the model. GLMs were constructed, each time training the models with all but one of the genes. The trained models were applied to get a numerical prediction of the excluded gene. This was repeated until each of the genes in the dataset had been excluded once.

The GO terms used and the average AUC for these terms is shown in Table 4.2. The results show that the predictions for a majority of these genes were insignificant, with

AUC mostly around 0.5. However, out of the 8 processes examined, the genes that were annotated with the GO term of "ER to Golgi vesicle-mediated transport" had a high AUC of 0.93. This indicated a strong covariation in branch lengths across the 4 genes (non-parametric *p*-value < 0.01). A PCA plot of the data was computed but showed no interesting correlations (Figure 4.6).

As this dataset is small it is possible that the overall lack of correlation is due to insufficient data in terms of the size of the taxa set (8 species), the number of total genes (106 genes), and the number of genes that share a common biological process. Although the overall correlation is weak, covariation is observable in gene trees of genes in yeast species that are involved in the process of ER to Golgi vesicle-mediated transport.

Table 4.2 Prediction accuracy of the GLMs for the leave-one out tests of the yeast data, measured by the AUC.

All terms used for this dataset were truncated terms as the number of genes was small.

GO biological process(es) term	Number of genes	AUC
"ribosom"	10	0.55
"fold"	5	0.56
"translat"	5	0.53
"telomere maintenance"	5	0.32
"mitochon"	4	0.45
"ER to Golgi vesicle-mediated transport"	4	0.93
"ribosome biogenesis"	8	0.51
"transcript"	13	0.45

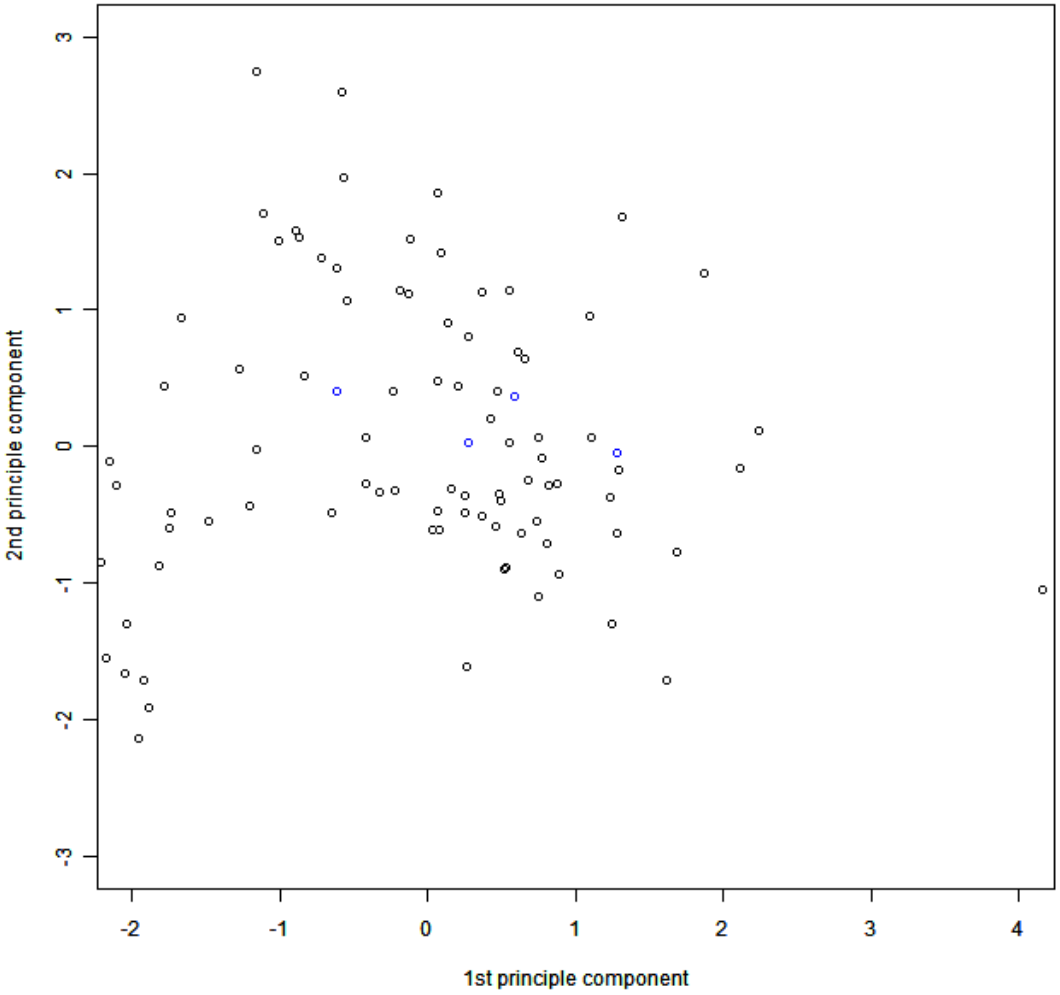


Figure 4.6 A PCA plot of the first two principle components for the yeast dataset.
Genes annotated with GO process term "ER to Golgi vesicle-mediated transport" are coloured blue.

4.5 Analysis of the bacterial dataset

The dataset used in Pazos *et al.* (2005) consists of amino-acid alignments of *Escherichia coli* genes against orthologs in 43 other prokaryotic species. Pazos et al. obtained these alignments by BLASTing (ALTSCHUL *et al.* 1997) the *E. coli* protein sequences against the genomes of other prokaryotic species. Pazos et al. included in the dataset the alignments that have an E-value above a chosen cut-off point.

As the number of species included increases, the number of genes that are homologous between all the species decreases. A subset of the species was selected that was sufficiently large to have a reasonable number of branches in the

corresponding gene trees, but which had a sufficient number of orthologous genes. The subset of species was chosen by finding the ten species that were most frequently present in alignments in Pazos et al.'s dataset (*Bacillus subtilis*, *Mesorhizobium loti*, *Caulobacter crescentus*, *E. coli*, *Salmonella typhi*, *Salmonella typhimurium*, *Yersinia pestis*, *Pasteurella multocida*, *Vibrio cholerae* and *Pseudomonas aeruginosa*) and the gene alignments that contained all ten species were used for further analysis.

4.5.1 Recovering species and gene tree topologies

As the dataset consists of prokaryotes, gene tree topologies can differ from the species topology as a result of horizontal gene transfers (HGT). To filter out genes where the gene relationship may not reflect the underlying species relationships, MCMC analyses were performed using MrBayes (RONQUIST and HUELSENBECK 2003). For each of the genes two independent MCMC runs were performed, each with one cold chain and three heated chains, using a mixed amino-acid model with four gamma (γ) rate categories and allowing invariable sites (i). Prior distributions of tree branch lengths and the gamma shape parameter were set to exponential distributions with λ=10 and the starting tree was chosen randomly. The chains were run for 1100000 steps and sampled every 200 steps, with the first 10% of the MCMC run (550 sample trees) treated as burn-in and discarded.

The posterior distributions were taken and used to estimate the correct relationships amongst the species. Posterior probabilities of each tree topology from the 95% credible set of trees were taken for each gene. The probabilities of each topology for each gene were multiplied to get the joint posterior probability of each topology over all genes, assuming independence of genes. The tree with the highest joint posterior probability was chosen as the best estimate of phylogeny. The procedure here is justified by the fact that if the tree priors for each gene are assumed to be equal, the genes are unlinked, and the prior on tree topologies is uniform, then this calculation is monotonic with the joint posterior probability, as follows. The posterior probability of a given tree τ, over all genes, D_i, is:

$$
\begin{aligned}
& P(\tau \mid D_1, D_2, ..., D_N) \\
& \propto P(D_1, D_2, ..., D_N \mid \tau) P(\tau) \\
& = \prod_{i=1}^{N} P(D_i \mid \tau) P(\tau)
\end{aligned}
\tag{4.1}
$$

If the posterior probabilities are obtained separately for each gene then:

$$
\begin{aligned}
& P(\tau \mid D_1) \times P(\tau \mid D_2) \times ... \times P(\tau \mid D_N) \\
& \propto P(D_1 \mid \tau) P(\tau) \times P(D_2 \mid \tau) P(\tau) \times ... \times P(D_N \mid \tau) P(\tau) \\
& = \prod_{i=1}^{N} P(D_i \mid \tau) P(\tau)^N
\end{aligned}
\tag{4.2}
$$

As can be seen, Equation 4.2 is monotonically and non-linearly proportional to Equation 4.1.

When a particular topology is not found in a gene, a minimum probability is assigned, equivalent to one divided by the number of samples taken in the MCMC analysis. According to this criterion, the most probable tree topology yielded a log joint posterior probability of -2289.62. In contrast, the second most probable tree had a log probability of -2814.34. The most probable species topology found from our MCMC analysis concurs with the one used in Pazos et al.'s study, which is derived from neighbour-joining trees of distances in the 16S rRNA gene (topology in shown in example genes trees in Figure 4.10).

As the issue of HGT needed to be addressed, any genes that had a low probability of their gene tree matching the species tree topology were filtered from the dataset. Genes were excluded if the MrBayes analysis did not contain the species topology within its 95% credible set of trees. As a result, 222 genes out of 471 were excluded from the dataset.

4.5.2 Annotating the dataset with functional information

Gene Ontology (GO) (ASHBURNER *et al.* 2000) annotations on biological processes and molecular functions for *E. coli* genes were obtained from the UniProt (LEINONEN

et al. 2004) and iProClass (WU *et al.* 2004) databases. iProClass contains functional annotations that were electronically determined. These iProClass annotations are determined by high sequence similarity to genes of known function in other species. These annotations were used to increase the amount of annotation for our gene set, as there are insufficient annotations in *E. coli* that have been experimentally identified. Genes containing no GO annotation for known process or function were removed from the dataset. All GO terms used took into account exact synonyms for the same term. The resulting dataset contained alignments of 219 homologous genes from the 10 prokaryotic species.

The number of genes associated with every pair of GO biological process and molecular function was determined. Pairings of GO process and function were used to represent distinct biological functions. The justification for this is that using only one of biological process or molecular function will group together genes that are not necessarily functionally related. This was an improvement over just using information on the GO process term, as we did in our previous analyses. Each gene was therefore labelled with the process-function pairs that it is associated with. This information was later used in training GLMs of each function. The process-function pairs that had less than 7 genes involved were filtered out because training models with a low proportion of positive cases can lead to biased and badly fitting models (FOLEY 1972).

4.5.3 Results

A leave-one-out test was used to assess the accuracy of the predictions with the GLMs. False positive error rates, true positive error rates and the AUC were calculated with the ROCR package in R (SING *et al.* 2005).

Table 4.3 shows a list of process-function pairs used and the accuracy of the trained models as assessed by AUC. It can be seen from Table 4.3 that the AUC for classification appears to vary greatly across the terms. There appears to be strong correlation of phylogenetic signatures among genes that are identified as being involved in both the GO process of "translation" and GO function of "structural constituent of ribosome", with an AUC of 0.92 when trying to predict the function of

these genes. From Figure 4.7a, it can be seen that the false positive rate of predicting gene involvement in this particular function was in general very low across different cut-off points.

The accurate prediction also extends to genes that are identified as being in other ribosomal related functions within "translation", with AUC of 0.80, 0.88 and 0.82 in "tRNA binding", "rRNA binding" and "RNA binding" ("RNA binding" is a generalisation of both types of RNA binding), respectively (Figure 4.7b-d). Upon closer inspection, these four "translation" related RNA functions contain genes that overlap, such that the genes involved in one of the functions were often annotated as being involved in some of the others. This AUC indicates high correlation between branch lengths for the trees of genes annotated as being involved in this process. In contrast, for a majority of the process-function pairs, correlation in gene tree branch lengths was not seen between genes identified as having the same GO terms, with the GLMs performing approximately at random.

Randomisation tests were carried out to determine whether the high correlations in the processes-function pairs were statistically significant. For each pair, a null distribution of 1000 sample replicates was constructed. Each replicate was generated by randomly selecting genes in the dataset to be involved in a null biological function, which represents the expected degree of correlation between the genes under the null hypothesis. The number of genes selected to be involved in the null function in each replicate is equivalent to the number of genes involved in the process-function term. A leave-one-out test was carried out on each of the replicates and the AUC calculated. From these randomisations the *p*-values of obtaining the actual AUC for each GO term combination were calculated. *p*-values were again adjusted with false discovery rate correction (BENJAMINI and HOCHBERG 1995) to correct for multiple comparisons (shown in Table 4.3). It can be seen that the correlation observed in the ribosomal functions of translation was significant to a 5% error rate. This indicates that the high AUC produced by this gene grouping was unlikely caused by sampling effects. Apart from the translation related functions, there were no other functions that were significant.

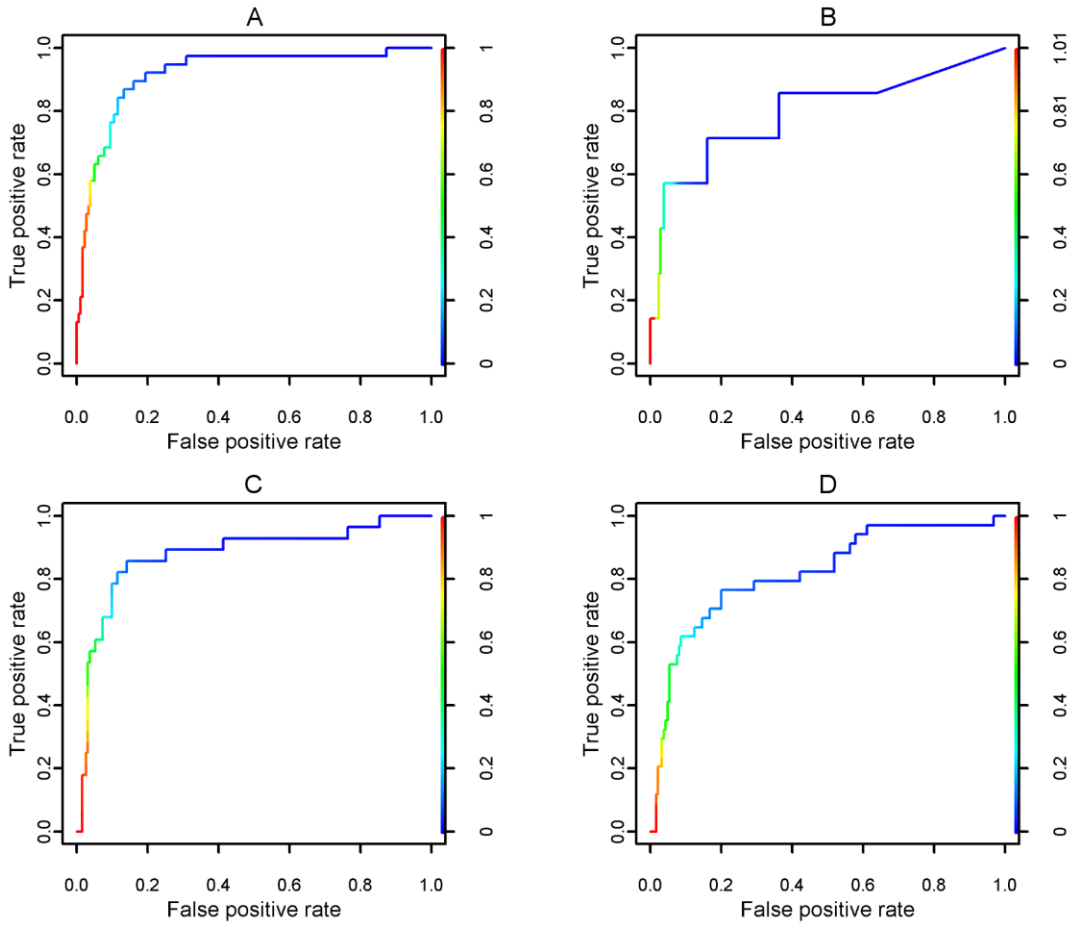


Figure 4.7 Plots of true positive rate against false positive rate for a few example GO process-function pairs in the bacterial dataset.

The predictions from the GLMs of each function were estimated using different values of the cut-off point (shown by the coloured scale on the right), and error rates calculated from these predictions. (a)-(d) shows the accuracy of four related ribosomal functions within the GO process of "translation". The four GO functions are (a) "structural constituent of ribosome", (b) "tRNA binding", (c) "rRNA binding" and (d) "RNA binding", respectively.

Table 4.3 Prediction accuracy of the GLMs for the leave-one out tests on the bacterial data, measured by the AUC.

Different GO process terms and function terms often shared the exact same set of genes. For example the functions of "aminoacyl-tRNA ligase activity", "ATP binding" and "ligase activity" within the "translation" process have the same genes associated with them. These are grouped as a single category in the Table.

GO biological process(es)	GO molecular function(s)	Number of genes	AUC	Adjusted *p*-value
translation	structural constituent of ribosome	38	0.92	0.00
translation	rRNA binding	28	0.88	0.00
translation	RNA binding	34	0.82	0.00
translation	tRNA binding	7	0.80	0.01
translation	protein binding	22	0.69	0.03
translation	aminoacyl-tRNA ligase activity; ATP binding; ligase activity	12	0.71	0.05
regulation of transcription, DNA-dependent	protein binding	8	0.70	0.10
transport	protein binding	7	0.69	0.10
protein folding	protein binding	7	0.66	0.17
DNA replication	protein binding	8	0.67	0.19
tRNA aminoacylation for protein translation	aminoacyl-tRNA ligase activity; ATP binding; ligase activity; nucleotide binding	7	0.62	0.23
DNA repair	hydrolase activity	8	0.61	0.23
translation	nucleotide binding	15	0.59	0.23
response to DNA damage stimulus	hydrolase activity	7	0.55	0.37
transport	ATP binding	7	0.54	0.41
DNA replication	DNA binding	7	0.52	0.41

Table 4.3 (continued).

GO biological process(es)	GO molecular function(s)	Number of genes	AUC	Adjusted *p*-value
SOS response	DNA binding	7	0.49	0.52
metabolic process	transferase activity	10	0.48	0.58
regulation of transcription, DNA-dependent	RNA binding	7	0.43	0.67
metabolic process	protein binding	7	0.39	0.74
metabolic process	catalytic activity	13	0.42	0.74
DNA repair; response to DNA damage stimulus	DNA binding	11	0.41	0.74
cell cycle; cell division	nucleotide binding	8	0.38	0.74
DNA repair; response to DNA damage stimulus	ATP binding; nucleotide binding	7	0.33	0.80
transcription	DNA binding	7	0.29	0.86
transcription	protein binding	8	0.25	0.91

As a control, the accuracy of the prediction was tested for correlation with the number of genes that were used to train the models. It is a known issue in statistics that under-trained models with too few cases of each class produce biased and inaccurate predictions (RAUDYS and JAIN 1991). We computed a linear fit of the number of genes involved in each process against the accuracy of each process in AUC. The coefficient of determination (R^2) was calculated from the linear fit to be 0.39 (p = 0.0007). This indicates bias towards GLMs predicting for functions that have a higher number of genes involved in the function. As seen from the results in Table 4.3, pathways that contained fewer genes in general indicated no correlation in branch length between the genes. Better results are hence expected in some of these pathways as some of these functions become more thoroughly annotated.

For our most significantly correlated process-function of “translation” and “structural constituent of ribosome”, tests were expanded to further investigate the correlation. The size of the null distribution was increased to 10000 replicates. It is noted here that even when the number of replicates was increased, the bootstrap p-value remained at 0.0, indicating that there is <0.0001 chance that the correlation seen was obtained at random. Hence we have strong evidence to reject the null hypothesis that the correlation in gene tree phylogeny between genes labelled with GO terms “translation” and "structural constituent of ribosome" was due to random effects.

The PCA plot of this dataset is shown in Figure 4.8 with genes annotated with “structural constituent of ribosome” coloured separately. The PCA plot indicated that these ribosomal genes were differentiated from other genes with the first principle component, where they generally had much higher values. Therefore, the PCA in this case supported the correlation among the gene trees of the ribosomal proteins.

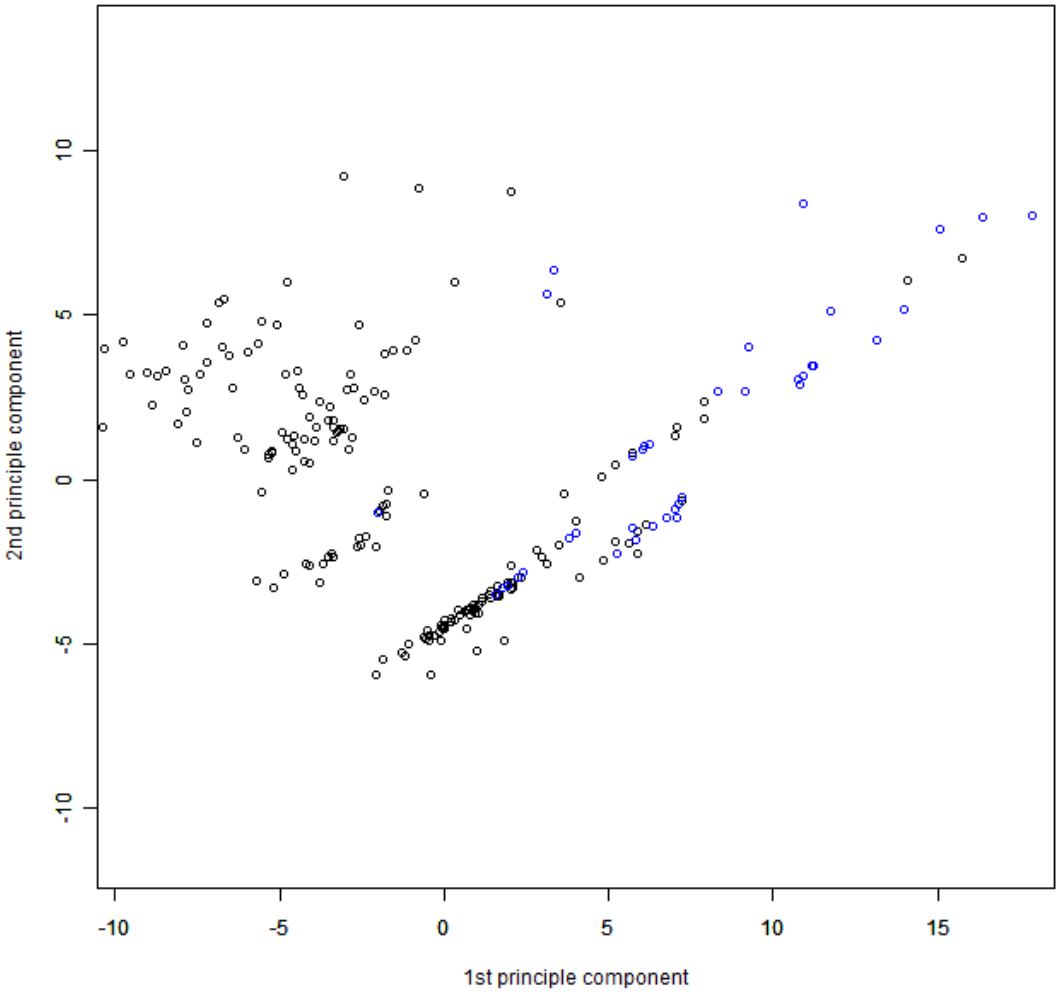


Figure 4.8 A PCA plot of the first two principle components for the bacterial dataset.
Proteins annotated with GO process term "translation" and GO function term "structural constituent of ribosome" are coloured blue.

The covariation in gene tree branch lengths of ribosomal genes was further investigated. We tested whether the correlation in phylogeny occurs only amongst physically interacting genes, or whether correlation extends to non-interacting genes of related function. To test this, the most significantly correlated process-function pair of "translation" and "structural constituent of ribosome" was again used. The known interactions between the genes involved in this biological function were obtained from the Database of Interacting Proteins (XENARIOS *et al.* 2002). Figure 4.9 shows the interaction network of the proteins in our dataset labelled with these particular GO terms. Although a large number of interactions within this pathway occur between the genes, not all the genes contain an interaction with another. In fact some of the

proteins contain few interactions to any of the other proteins. Yet, the correlation in gene tree branch length amongst the proteins shown here was clearly shown in the results of the leave-one-out test. Hence, it can be seen that the correlation in phylogeny between genes is not purely limited to physically interacting genes, but the correlations also exist between functionally related genes operating within the same pathway.

Figure 4.10 shows an example of gene trees from proteins within this pathway. From the example it can be seen that there are similarities in branch lengths between proteins functioning within the same pathway, which is not limited to only proteins that directly interact. These similarities also show distinction from other proteins, as is seen by the dissimilarities of the gene trees to the consensus tree of proteins not involved in the pathway, for example, the relatively shorter length of the *B. subtilis* branch and overall shorter tree lengths.

Figure 4.11 shows the coefficients of the GLMs from modelling the correlation from these proteins. As a comparison, the average values of each coefficient from the 10000 randomisations generated is shown. It should be noted that the coefficients here model the variation in log-transformed branch lengths; therefore a large proportion of the predictor values will be negative, as branch lengths are generally small. From Figure 4.11, it can be seen that the coefficients from the actual process-function term itself differ greatly from that of the randomisations. This indicates that there is a distinction in the gene tree branch lengths of proteins involved in this pathway from those in other pathways. As the intercept value and end predicted value differ between the two models, comparisons cannot be made.

Previous studies have found that for phylogenetic profiling (PELLEGRINI *et al.* 1999) the number and choice of species affects how informative the profiles are (JOTHI *et al.* 2007; SINGH and WALL 2008). As the underlying concept of our analysis is similar to phylogenetic profiling, this may also cause a bias in our results. To test whether the high correlations seen here are biased by species choice, we repeated the leave-one-out analysis. Each time it was repeated, single taxon removal was simulated by excluding the taxon and taking into account the changes in branch length by summing

its adjacent branches. With the removal of taxa, the AUC that was produced by the GLMs of "translation" and "structural constituent of ribosome" did not alter greatly from our original result. From the 10 individual species removals, the AUC ranged from 0.89 to 0.94, with a mean of 0.91. Therefore, the significant correlation is unlikely an effect of bias due to choice of species used in our analysis.

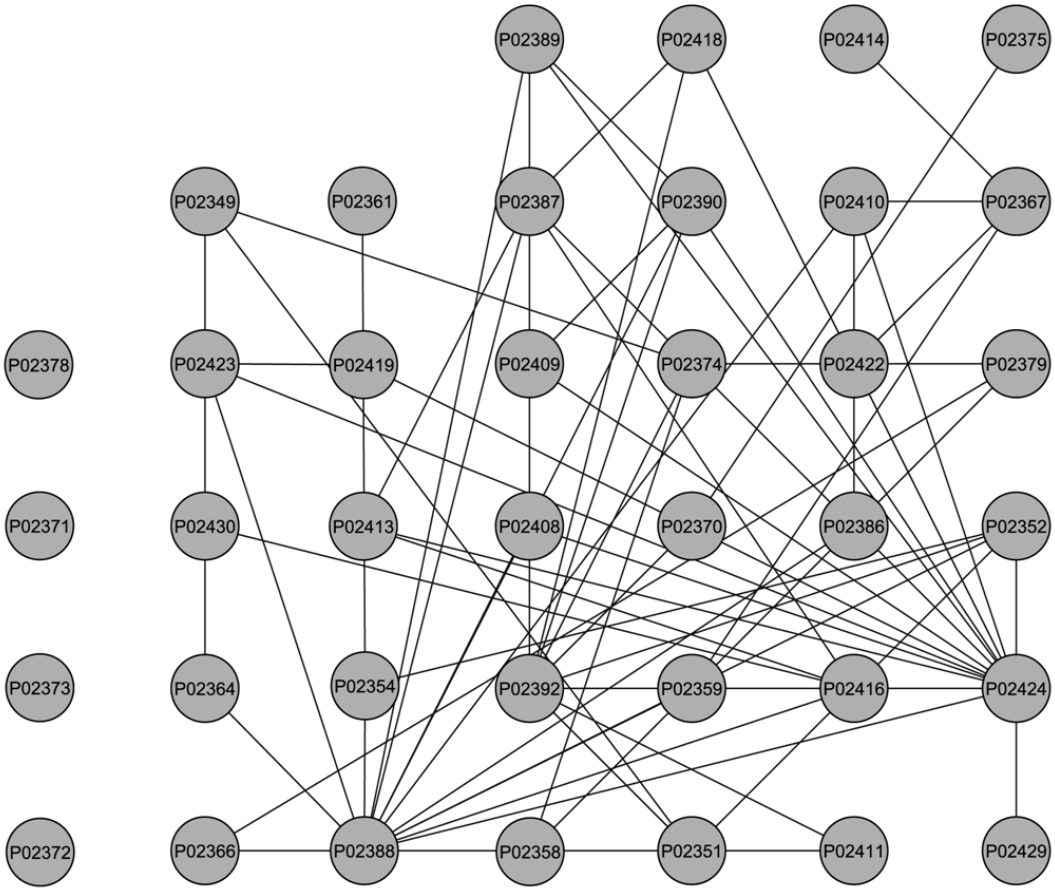


Figure 4.9 The pathway interaction network of proteins in our bacterial dataset annotated as being involved in GO process "translation" and GO function "structural constituent of ribosome".

Proteins P02378, P02371, P02373 and P02372 (in the first column) contain no known physical interactions to any other proteins in our list. The interaction network was constructed using Cytoscape (SHANNON *et al.* 2003).

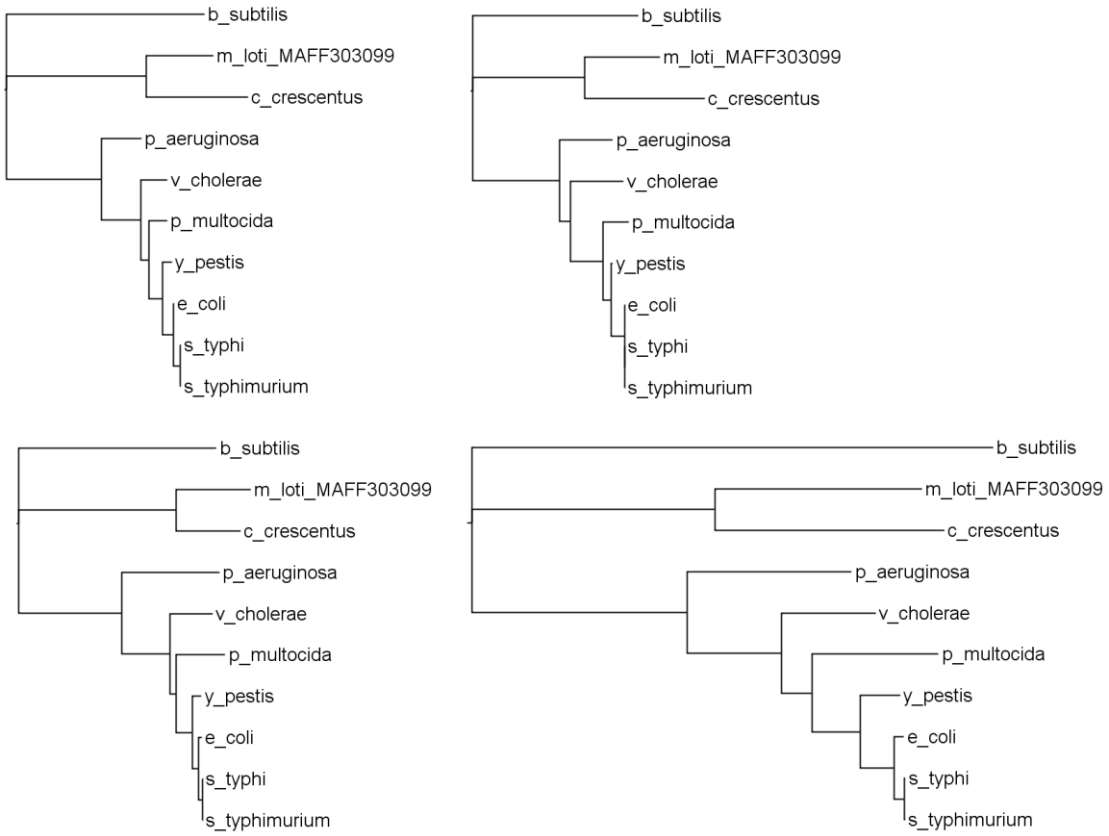


Figure 4.10 Example gene trees of proteins from our bacterial dataset.

These trees are shown here for a detailed analysis of the proteins annotated as being involved in GO process "translation" and GO function "structural constituent of ribosome". From top left to bottom right, the trees are from gene P02386, P02410 (a protein known to physically interact with P02386), P02351 (a protein that does not interact with either of the previous genes but contributes to the pathway) and the consensus of all gene trees in our dataset not labelled with these two GO terms.

(i) "translation" and "structural constituent of ribosome"

Coefficient variable	*B. subtilis*	*S. typhi*	*S. typhimurium*	*E. coli*	*Y. pestis*	*P. multocida*
Mean	0.30705	-0.17759	-0.01228	-0.14300	0.42118	-1.53665
Standard Error	0.00205	0.00042	0.00256	0.00032	0.00175	0.00356

(ii) Randomisations

Coefficient variable	*B. subtilis*	*S. typhi*	*S. typhimurium*	*E. coli*	*Y. pestis*	*P. multocida*
Mean	-0.08033	-0.00268	-0.00706	-0.00029	0.05142	0.00193
Standard Error	0.00016	0.00002	0.00002	0.00003	0.00009	0.00014

(i) "translation" and "structural constituent of ribosome"

V. cholerae	*P. aeruginosa*	*C. crescentus*	*M. loti*	Internal Node 0	Internal Node 1	Internal Node 2	Internal Node 3
-0.07829	0.28773	0.32179	-0.57008	-0.09312	-0.15779	-0.41797	0.00814
0.00232	0.00206	0.00225	0.00163	0.00033	0.00072	0.00149	0.00078

(ii) Randomisations

V. cholerae	*P. aeruginosa*	*C. crescentus*	*M. loti*	Internal Node 0	Internal Node 1	Internal Node 2	Internal Node 3
-0.01319	0.02744	0.02443	-0.01573	0.00752	0.01734	0.02286	0.01531
0.00012	0.00010	0.00014	0.00012	0.00003	0.00007	0.00010	0.00006

(i) "translation" and "structural constituent of ribosome"

Internal Node 4	Internal Node 5	Internal Node 6	Intercept
-0.52385	0.55831	-0.01271	-8.64613
0.00122	0.00166	0.00076	0.03355

Predicted value
0.17515
0.02155

(ii) Randomisations

Internal Node 4	Internal Node 5	Internal Node 6	Intercept
-0.00002	0.01912	0.02145	-0.77317
0.00010	0.00006	0.00007	0.00038

Predicted value
0.18029
0.00009

Figure 4.11 The details of the models built by the GLMs for (i) the proteins labelled with GO process "translation" and GO function "structural constituent of ribosome" and (ii) for the 10000 randomisations of its null distribution.

The end predicted value is obtained by adding the products of each coefficient and its corresponding predictor value, and the intercept value.

4.6 Analysis of OrthoMaM dataset

We tested our methods on the OrthoMaM dataset (version 4.0) (RANWEZ *et al.* 2007). This dataset contains 1056 genes that are shared across 25 mammalian species. After filtering in the same way as the bacterial dataset we obtained a substantial dataset containing 730 genes that were orthologous among 24 mammalian species (*B. taurus*, *C. familiaris*, *Cavia porcellus*, *Monodelphis domestica*, *Ornithorhynchus anatinus*, *Echinops telfairi*, *Loxodonta africana*, *Ochotona princeps*, *Oryctolagus cuniculus*, *Spermophilus tridecemlineatus*, *M. musculus*, *R. norvegicus*, *Tupaia belangeri*, *Microcebus murinus*, *Otolemur garnettii*, *M. mulatta*, *Pongo pygmaeus*, *Pan troglodytes*, *H. sapiens*, *Erinaceus europaeus*, *Sorex araneus*, *Myotis lucifugus*, *Equus caballus* and *Felis catus*). The *Dasypus novemcinctus* species was omitted as we found that by removing the species from the dataset we were able to significantly increase the number of genes that could be used in the analysis. The species topology

was estimated by building a consensus tree from the individual gene trees provided by the OrthoMaM database. We acknowledge the potential for error through using consensus tree methods; however the estimation of the topology with maximum likelihood and Bayesian inference was impractical given the limited computational resources.

4.6.1 Results

We applied the GLM to the OrthoMaM data using a leave-one-out test. Results of this analysis showed no significant correlation between any genes involved in a particular function and the gene tree branch lengths for the genes. Though the data itself are abundant, the terms that were common among the genes were uninformative. For example the most abundant processes-function pairs were: " regulation of transcription, DNA-dependent" with "DNA binding", "signal transduction" with "protein binding", and "regulation of transcription, DNA-dependent" with "transcription factor activity". These terms contain limited information on the underlying pathways themselves and it is likely that not all the genes sharing these terms function within the same pathway. The lack of correlation may also potentially be attributable to the distance between species (the overall tree length of this dataset was roughly 7.5 times shorter than that of the bacterial dataset) and the positive test case to negative test case ratio (the most abundant process-function pair only contained 5.9% of the 730 genes), which is known to cause under-fitting in model fitting.

4.7 Discussion

We have shown that there are correlations between a protein's function and its gene tree branch lengths. This correlation in phylogeny is most likely attributable to the co-evolution of genes that have functionally related gene products. Previous studies of inference from co-evolution have focused primarily on the relationship between genes that have physically interacting gene products. We have shown that correlation in branch lengths extends to genes that are involved in the same functional pathway.

This effect of correlation is demonstrated across a broad range of species, namely mammals, fungi and bacteria.

Hakes *et al.* (2007) proposed the hypothesis of common selection pressures occurring on these genes to account for correlated evolutionary rates in functionally related genes. We may also imagine that the correlations can be caused by the "contagious" propagation of mutations across the genes in the biological pathways responsible for the function. Specifically, mutations in one gene in a pathway may lead to direct compensatory mutations in a set of related genes which in turn can cause compensatory changes in other related genes, causing a cascade of mutations throughout the pathway. Alternatively, it may be that a change in the selective environment leads to changes in the selective pressure to maintain the structures of proteins involved in a given function, so that changes in substitution rates (and branch-lengths) are observed along different lineages.

In our bacterial dataset, we found that the correlation in branch length was particularly high in proteins involved in translation and ribosomal activities. This was most significant in proteins labelled with GO terms "translation" and "structural constituent of ribosome". We found that the overall tree lengths of these proteins are shorter than that of other proteins (average of 2.96 in ribosomal genes and 6.30 in others). This indicates that there is an overall effect across species of purifying selection acting towards the genes coding for these proteins. This is in agreement with literature stating that strong selective pressure occurs across ribosomal and translational genes (BLATTNER *et al.* 1997; LECOMPTE *et al.* 2002). An explanation for the purifying selection across these genes is that functions such as translation are crucial for an organism's basic function and therefore any changes to the protein sequence may cause disruption towards this essential pathway. It can also be seen that the degree to which purifying selection occurs differs across each species lineage. This is indicated by the coefficient values shown in Figure 4.11, as each coefficient varies a different amount to what is expected on average, though stochastic effects should also be considered.

In our analysis, uncovering correlation is limited to identifying genes that experience similar selective regimes. The assumption is that genes with functionally related proteins would undergo similar rates of evolution; yet it is possible for functionally unrelated genes to have undergone rate similarities. A subset of this effect is when trying to find correlations amongst genes from a common biological function where the genes are evolving neutrally or near neutrally. Gene trees of any other neutrally evolving genes will have similarities in branch length to gene trees of this function. This can confuse general classification and correlation methods into believing that these genes should be grouped within this function. This is noted as a limitation to our method but it will also confound any method that is based on identifying equivalent lineage-specific rates of evolution.

How successfully correlation could be detected appeared to vary across the datasets. The likely cause of the differences across datasets is lack of information in some of the datasets. Despite the high significance seen in the correlation of some of the functions, a majority of the functions induced correlations only marginally stronger than random. This suggests that the correlation in branch lengths is weak amongst genes annotated as being involved in those processes. The low correlation may be explained by a range of factors. Firstly, such biases in different processes are possibly due to issues within our dataset. A low number of genes involved in a function to train the model can lead to biased models. As mentioned previously, it is commonly known in statistics that a reasonable number of each case type relative to number of features is required to train accurate models (FOLEY 1972). The test in correlation showed that there was a significant correlation between the number of positive test cases in the processes and the AUC. It is likely that this effect caused some bias in our study where functions with a larger number of genes involved are favoured.

Secondly, errors in the prediction can be caused by incorrect and incomplete functional annotations. Gene annotations in databases are often incomplete and contain errors. In particular for some biological processes such as gene translation, the specific functions of each gene involved in these processes are better known. Relevant processes will therefore have more complete and less erroneous annotations.

A third factor contributing to the discrepancy in correlation is natural variation in amount of selection pressure and gene constraint. Observed co-evolution is an effect of similar selection pressures acting on functionally related proteins (HAKES *et al.* 2007). Where the selection pressures are weak, lesser correlations in substitution rates are expected. In cases where the compensatory mutations are crucial towards the co-evolution, weaker structural constraints between genes with interacting products will result in less co-evolution. Often mutations in amino-acid sequence cause no or small changes to the outcome of protein structure (MINTSERIS and WENG 2005; SHAKHNOVICH *et al.* 2005). While some protein interactions necessarily require co-evolution, others are known to naturally have structural flexibility and can allow for changes in interaction partners without having to make changes to itself (GILLMOR *et al.* 2000; MINTSERIS and WENG 2005). Less constraint between genes would mean that correlated mutations are often inessential thereby resulting in less similarity in substitution rates. This effect is more likely in genes where the sequence of the binding surfaces is proportionally short. In these cases mutations may not have great structural modifications to the gene and compensatory mutations may not occur. As a result, similarities in branches will be less evident. In addition to this, it is possible for functionally related genes to not share common patterns of evolution. As shown by Rausher *et al.* (1999) and Lu *et al.* (2003), genes that produce functionally related proteins can undergo different degrees of selection, as a result of relaxed constraint on some of the genes. It is possible that in our datasets certain genes have become relaxed in one or more species. Consequently, there can be a lack of correlation between such genes and other genes in the pathway it is involved in.

Finally, another explanation for weaker correlation between genes is the definitions of function provided by GO terms. GO provides a set of text vocabularies used to categorise sequences by the general attributes of their biological function. These vocabularies cannot distinguish between different pathway organisations within the function. Hence, it is often the case that functionally unrelated genes may be annotated similarly in GO. In addition, GO terms provide no indication to the specificity of each term. Some function terms are very specific (for example, “protein secretion by the type II secretion system”, “small GTPase mediated signal

transduction") whilst others are very general (for example, "metabolic process", "cell cycle").

This study suggests that when estimating divergence times, care should be taken because gene tree branch lengths may be biased by the function of the gene. Correlated changes in genes are more prominent in genes with gene products of related function; these will affect rate estimation if these genes are treated as multiple "independent" loci. An implication of our finding towards estimating species divergence times in comparative biology is that it is erroneous to estimate species distances using a small number of functionally related genes. As biases in branch length estimation also affect the accuracy of resolving phylogenies, this problem contributes to the systematic bias present in phylogenomic estimations of species topologies. Since the pioneering papers by Brown *et al.* (2001) and Rokas *et al.* (2003b), a number of studies have reconstructed species phylogenies using large genomic datasets (for example, BROCHIER *et al.* 2002; JAMES *et al.* 2006; PHILIPPE *et al.* 2005a; QIU *et al.* 2006; ROKAS *et al.* 2005). Although the methods used by Rokas et al. and Brown et al. reduce the stochastic bias that stems from using a small gene set, studies have shown that using genome-scale phylogenetic estimations make the methods more susceptible and sensitive to systematic biases (HOLLAND *et al.* 2006; LEIGH *et al.* 2008; PHILLIPS *et al.* 2004; RODRIGUEZ-EZPELETA *et al.* 2007). Biases such as those demonstrated here therefore have considerable influence towards tree estimations.

Though these effects have been to some degree recognised, they are often not considered when carrying out comparative analysis between species. A suggestion from our results is that estimating species distances should be performed using multiple loci from genes of a wide range of functions. Our findings support the suggestion made by Thorne and Kishino (2002) of taking into account the correlation of genes when using multiple loci. Thorne and Kishino suggested that when estimating distances using concatenation of genes, to add parameters, models and priors which consider the correlation of substitution rates amongst the genes. Our result provides support to the use of Thorne and Kishino's techniques and as a result raises questions towards the assumption of independence in substitution rate of gene

trees which is commonly made when performing phylogenetic analysis with concatenated data.

Chapter 5. Detecting lineage-specific selection

5.1 Introduction

The link between selection and function has become evident in light of the causes underlying molecular evolution (GILLESPIE 1991; KIMURA 1983). One major finding has been the effect of selection on the rate of substitution across a genome (KIMURA 1967; KIMURA 1983). When positive selection occurs in a given function, mutations occurring in the sequence regions coding for that function are favoured. As a result, mutations in these sequence regions are more easily fixed into the population of the species as substitutions. The opposite is true for regions undergoing purifying selection, where mutations are unlikely to be accepted as substitutions as they may disrupt the fundamental function underlying the sequence. Selection pressures therefore cause changes in substitution rates in genomic regions of functional significance. Consequently, regions of the genome with exceptional rates of substitution often correspond to regions of high functional significance. However, it should be noted that functional regions that have undergone conflicting selection pressures in different directions will not have significantly different rates.

In recent years, there has been emerging interest in using this relationship between function and heterogeneity in substitution rate to determine regions of the genome that are of functional significance (BEJERANO *et al.* 2005; GLAZOV *et al.* 2005; HARDISON 2000; INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM 2002; KIM and PRITCHARD 2007; LOOTS *et al.* 2000; LUNTER *et al.* 2006; SIEPEL *et al.* 2005; WONG and NIELSEN 2004). By using computational procedures for detecting selection, the

search space for finding functional regions can be greatly minimised and so the time and cost required to test each region experimentally is reduced. The continual growth in the availability of genomic data has also contributed to the increase in popularity of high-throughput *in silico* methods for detecting selection.

Out of the current computational methods for detecting selection, methods that detect selection through changes in rate of substitution are advantageous, as they can be applied to non-coding regions of the genome. Comparative studies of the mouse and human genome (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM 2001; INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM 2002) have found that 5% of the human genome is undergoing purifying selection, even though only 1.5% of the genome is estimated to be protein coding. This finding emphasises the functional significance of elements contained within non-coding regions of the genome. Through using these comparative genomic methods for detecting selection with rate changes, studies have found a number of functional elements in non-coding regions, such as regulatory elements and chromosomal organisation elements (DUBCHAK *et al.* 2000; HARDISON 2000; INADA *et al.* 2003; INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM 2002; JOHNSON *et al.* 2004; KEIGHTLEY *et al.* 2005; LOOTS *et al.* 2000; XUE *et al.* 2004).

Although these methods for detecting selection have so far provided researchers with interesting results, one factor that is often neglected is that selection does not necessarily take place in all lineages across the taxa. Due to changes in biological function, selection pressures can be gained or lost over the divergence and are not consistent across all species. Hence, one extension of detecting selection across branches is to find selection that only occurs in a particular species or lineage. Detecting lineage-specific selection can highlight regions that are functionally important for a specific set of species. With this knowledge, it is then possible to identify what separates species on a genetic level and what differentiates lineages from each other.

A number of studies have proposed strategies for identifying lineage-specific selection (ANISIMOVA and YANG 2007; FORSBERG and CHRISTIANSEN 2003; GUINDON

et al. 2004; KOSAKOVSKY POND and FROST 2005; SIEPEL *et al.* 2006; YANG and NIELSEN 2002; YU *et al.* 2006). A majority of these studies have based their methods on the fact that when selection is present, there is a change in the ratio of non-synonymous to synonymous substitutions in protein coding regions (YANG 1998), with the exception of a study by Siepel *et al.* (2006). Siepel *et al.*'s study highlights a method that is based on the changes in rate of substitution under selection. Their algorithm employs a Hidden Markov Model which allows regions to be classified as either conserved or not conserved, where conserved sequences correspond to occurrences of purifying selection. A drawback to their approach is the inability to allow for greater than expected rates of evolution, which transpire under positive selection.

As these current techniques for detecting selection all possess some sort of limitation, there is a need to develop a more general procedure for detecting selection. The ideal approach should not be limited to applications in protein coding regions and should allow selection in only a subset of the species to be identified. The method should also be computationally efficient so that it is feasible for genomic-scale analyses.

In this chapter, we outline a phylogenomic approach for detecting lineage-specific selection. Our method is based on a likelihood framework and identifies selection through deviations in the rate of substitution from what is expected under neutral selective pressures. The algorithms were tested on a dataset containing simulated selection events. The analysis of this method demonstrated promising results for detecting selection in specific simulated datasets but leaves some room for improvement when the data more accurately resembled reality. We discuss the potential benefits of having such procedures for identifying selection, potential improvements for our current method and alternatives for solving the question at hand.

5.2 Algorithm

The algorithms we propose are based on conventional tree likelihood calculations, such as those used in maximum likelihood and Bayesian phylogenetics approaches (FELSENSTEIN 1981). The topology between the species $\bar{\tau}$ and the divergence times $\bar{\mathbf{t}}$

are assumed to be known and are consistent across the genes compared. The assumption is also made that the underlying species substitution rates across a tree when selective pressures are neutral is known, which is denote as a vector $\bar{\mathbf{r}} = \{\bar{r}_1, \bar{r}_2, ... \bar{r}_{2n-2}\}$, where n is the number of taxa. $\bar{\theta}$ on the other hand, denotes the model parameters when selective pressures are neutral. Given some sequence alignment data D, the likelihood of the data given $\bar{\tau}$, $\bar{\theta}$, $\bar{\mathbf{t}}$ and $\bar{\mathbf{r}}$, $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}})$ can be calculated.

Suppose we attempt to maximise the likelihood $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}})$ by optimising the rate of a single branch rate in $\bar{\mathbf{r}}$. If the improvement in likelihood is significant, this would mean that the newly proposed value for the branch rate is preferred over its neutral rate (the rate of substitution when no selective pressures are present) for these particular data. Whether the improvement in likelihood is statistically significant can be determined using a standard statistical test called the likelihood ratio test (LRT) (SOKAL and ROHLF 1995). LRTs test the null hypothesis that the difference in two likelihoods is not significant. Given the two likelihoods $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}})$ and $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}}')$, where $\bar{\mathbf{r}}'$ is the newly proposed value for the rates, the likelihood ratio statistic G can be calculated as:

$$G = 2 \ln \left(\frac{P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}}')}{P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}})} \right) \qquad (5.1)$$

The significance of the likelihood change can then be determined by approximating the distribution of G by the χ^2 distribution. In essence, these methods find any major departures in rate of substitution from the neutral species rate across the genome.

Effectively, the rate of a branch can be interpreted as the sum of gene-specific effects, lineage-specific effects and lineage-gene-specific effects. The rate r_{ij} on a branch i in gene j can be summarised as:

$$r_{ij} = g_j + l_i + h_{ij} + \varepsilon \quad (5.2)$$

where g are gene-specific rate changes, l are lineage-specific rate changes, h are lineage-gene-specific rate changes and ε is an error term. By comparing r_{ij} on a particular gene tree branch with its corresponding neutral value, $\bar{r}_{ij}$, the effects of l can effectively be partitioned out. As values in $\bar{\mathbf{r}}$ represent the underlying neutral rates on the lineages, g can be partitioned by effectively scaling all the rates by a factor, k, such that $P(D \mid \bar{\tau}', \bar{\mathbf{t}}, \bar{\theta}, k\bar{\mathbf{r}})$ is maximised. The effects of g are in itself interesting, as these describe rate changes, and possibly selection pressures that are shared across all taxa. The remaining rate variations after l and g are partitioned out are h which are the result of lineage-specific selection. It should be noted that in this chapter, we interchangeably refer to lineage-gene-specific rate changes as lineage-specific selection as the former is the result of the latter.

Under this framework, two alternatives to the algorithm were proposed. The methods were implemented in Java 1.5 and utilise some of the functions and classes from the Phylogenetic Analysis Library (PAL) package version 1.5 (DRUMMOND and STRIMMER 2001) and BEAST (DRUMMOND and RAMBAUT 2007). The program uses the phylogenetic estimation software PAUP* (SWOFFORD 2003) to determine model parameters for the nucleotide site models that were used.

5.2.1 Algorithm 1 (A1)

The first algorithm proposed A1, involves a more direct application of LRTs. Under A1, any branch that significantly differs in rate from its rate in $\bar{\mathbf{r}}$ is considered an event of selection. The procedure for A1 for a given gene is summarised below:

1) Estimate the true neutral rates of substitution $\bar{\mathbf{r}}$ and the neutral substitution model parameters $\bar{\theta}$ (using PAUP*).

2) Calculate the likelihood of the data given $\bar{\tau}$, $\bar{\mathbf{t}}$, $\bar{\theta}$ and $\bar{\mathbf{r}}$, $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}})$.

3) Find the value of a coefficient k, which maximises the likelihood $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, k\bar{\mathbf{r}})$.

4) Perform a LRT for the change in likelihood from $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},\bar{\mathbf{r}})$ to $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},k\bar{\mathbf{r}})$ with one degree of freedom (*df*).

5) If $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},k\bar{\mathbf{r}})$ significantly improves the likelihood, set $\bar{\mathbf{r}}'=k\bar{\mathbf{r}}$ and note a change in gene-specific rate changes. Otherwise $\bar{\mathbf{r}}'=\bar{\mathbf{r}}$.

6) For each branch i, the rate of the branch is scaled to r_i such that the change in branch i maximises the likelihood $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},r_i,\bar{r_1}',\bar{r_2}',...,\overline{r_{2n-2}}')$.

7) Compare the likelihood of each of the optimisations carried out in (6). The event of change in branch rate, $i_{\max}$, that causes the biggest improvement from $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},\overline{r_{i_{\max}}}',\bar{r_1}',\bar{r_2}',...,\overline{r_{2n-2}}')$ to $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},r_{i_{\max}}\bar{r_1}',\bar{r_2}',...,\overline{r_{2n-2}}')$ is retained, where $\overline{r_{i_{\max}}}'$ is the original rate of the branch $r_{i_{\max}}$ in $\bar{\mathbf{r}}'$.

8) A LRT is performed with *df* = 1 on the likelihood improvement from $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},\overline{r_{i_{\max}}}',\bar{r_1}',\bar{r_2}',...,\overline{r_{2n-2}}')$ to $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},r_{i_{\max}}\bar{r_1}',\bar{r_2}',...,\overline{r_{2n-2}}')$. If the change in likelihood is significant, the rate of $\overline{r_{i_{\max}}}'$ is replaced in $\bar{\mathbf{r}}'$ with its preferred rate $r_{i_{\max}}$ and branch $i_{\max}$ is classified as being under lineage-specific selection.

9) Repeat steps (6) to (8) until the change in likelihood from altering $i_{\max}$ is not statistically significant.

This procedure is shown diagrammatically in Figure 5.1.

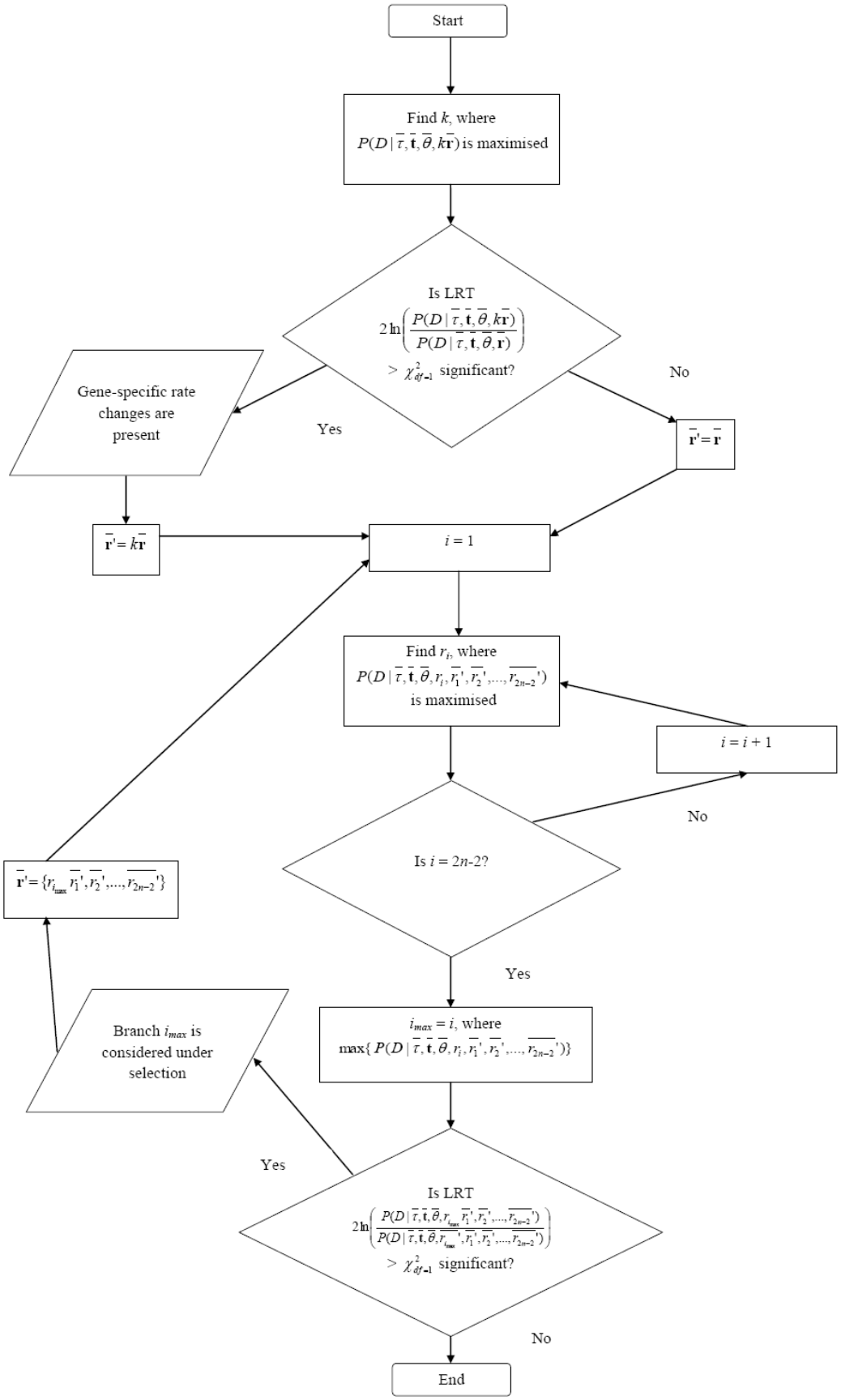


Figure 5.1 Flow diagram describing the logic in algorithm A1.

5.2.2 Algorithm 2 (A2)

For the second algorithm A2, a more conservative approach is taken to detect lineage-specific selection. The basis for this algorithm is that if a single branch is crucial towards the change in the likelihood, then changes in any other branch are comparatively minor. If optimisations of all other branches do not collectively cause a significant change in likelihood, then the unchanged branch has a significant contribution to the difference in tree likelihood. That branch is therefore likely to have a rate of substitution that differs from its neutral rate.

This approach was proposed due to the concern that the method used in A1 imposes a lot of strain on the likelihood calculation. As each time $i_{\max}$ is found in A1, the algorithm attempts to account for as much improvement as possible for the difference in likelihood from the neutral tree to the actual tree for the data, all with one degree of freedom. The change in likelihood from altering the rate of one branch in effect tries to accommodate the likelihood change from all the changes in rate across the branches. As a result, it is possible that changes in a particular branch may overestimate its contribution to the differences in likelihood and produce false positives. The procedure for A2 for a given gene is summarised below:

1) - 5) Repeat steps (1) to (5) in A1 to first partition the data for gene-specific effects on rate.

6) For each branch i, constrain the branch to its rate $\bar{r}_i'$ in $\bar{\mathbf{r}}'$ and simultaneously optimise all other branches to maximise $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{r}_i', r_1, r_2, ..., r_{2n-2})$.

7) For each branch i that we constrained, perform a LRT on the improvement in likelihood from $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}}')$ to $P(D \mid \bar{\tau}, \bar{\theta}, \bar{r}_i', r_1, r_2, ..., r_{2n-2})$, with $df = 2n-3$ (for a rooted tree). If the improvement is not significant, i is classified as being under selection.

This procedure is shown diagrammatically in Figure 5.2. It should be noted that A2 is computationally slower than A1, as multiple branches are optimised simultaneously. Unfortunately, as will become clear, a serious limitation of A2 is that when no selection is present in any of the branches, the algorithm will incorrectly interpret this

as presence of selection in all branches. This is discussed in more detail in the benchmarking of A2 in Section 5.4.2.

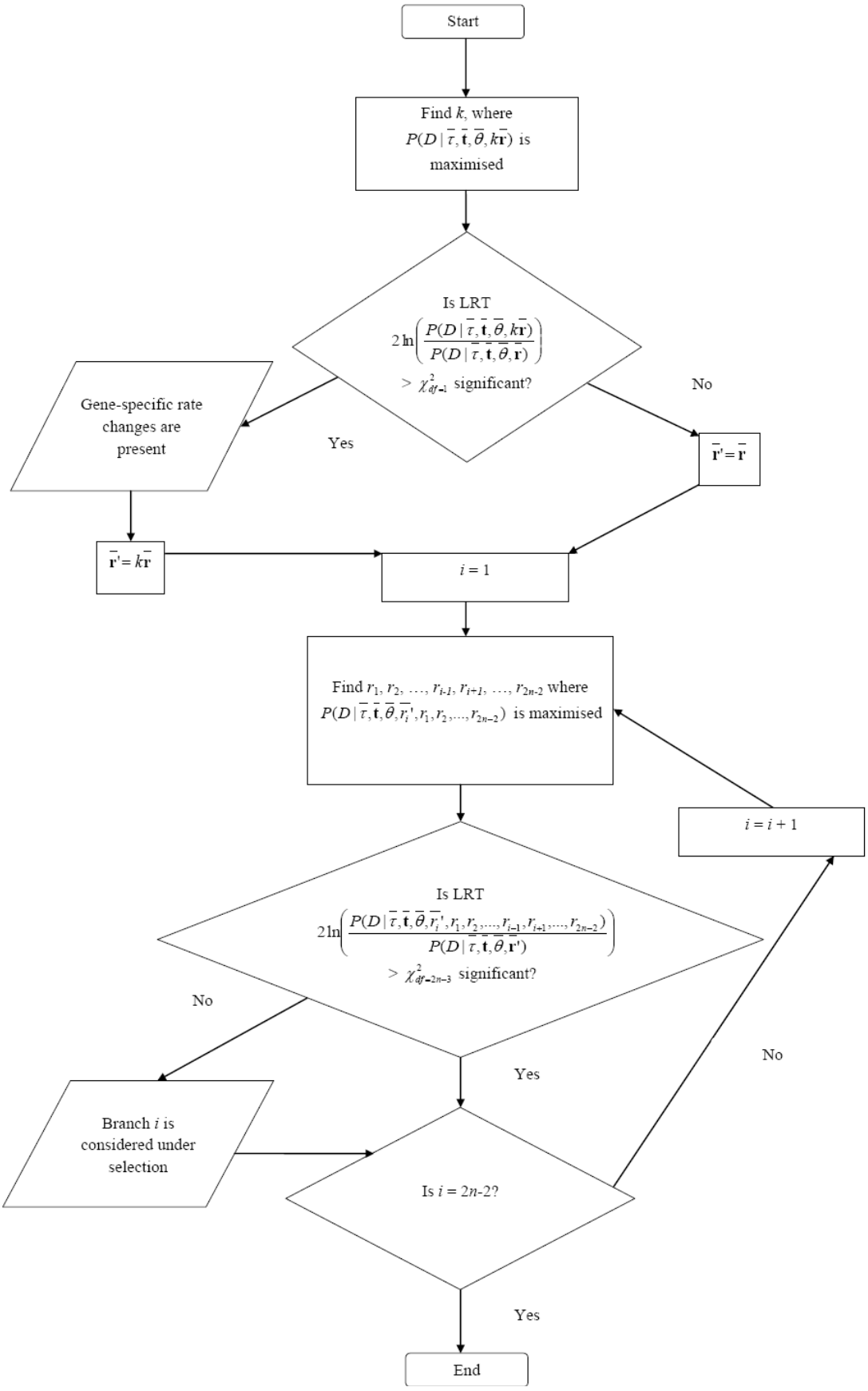


Figure 5.2 Flow diagram describing the logic in algorithm A2.

5.2.3 Limitations and perspectives

The procedures proposed here are in essence greedy algorithms. Greedy algorithms solve problems heuristically by making choices towards reaching local optima given its current state. The idea is that making successive decisions for local optima will yield a solution close to the global optima (JUNGNICKEL 2008). It is possible and often the case that greedy algorithms will choose a sub-optimal solution. Such methods have the benefit of being computationally much more efficient as fewer variables have to be co-optimised. This last point coincides with our aim of developing a high-throughput procedure for detecting selection.

Nonetheless, we acknowledge that there are several issues with this algorithm. First of all, our method is heavily dependent on the accuracy of alignment and tree reconstructions. Any errors that occur in the alignment, tree reconstruction or model specification will be directly reflected on the algorithm. In particular, it is crucial for the species topology that is used to be correct; otherwise rates of substitution that are found will be highly biased. The application of these methods towards each dataset should thus be carefully considered, as gene tree topologies can differ from their species tree topology due to recombination, horizontal gene transfer (in bacteria) and incomplete lineage sorting.

Secondly, the algorithms proposed rely on and are very sensitive to the true neutral distances in $\bar{\mathbf{r}}$. Inaccurate branch rate estimates in $\bar{\mathbf{r}}$ will bias the calculation of $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}}')$ and subsequently the test statistic *G* for each LRT performed. Under current methods in phylogenetics, it is virtually impossible to get a precise estimate of the branch lengths of a neutral species tree. Also the rate of neutral substitution is known to vary across genomes. An instance of this that has been demonstrated, is the higher neutral substitution rate in coding regions compared to non-coding regions found in humans and chimpanzees (SUBRAMANIAN and KUMAR 2003). Thus the underlying true tree varies depending on the region of the genome. We recognise that the use of $\bar{\mathbf{r}}$ is often unreasonable, and that using a global representation of neutral distances for all sequences across a genome is over generalised.

We also acknowledge that these algorithms are incorrect in assuming that divergence time is consistent across all the genes. Due to coalescence, the divergence times for each individual gene can differ as populations segregate over time and some alleles are isolated at different rates from others (NEI 1987). We should therefore take into account that the true species distances underlying each gene tree actually vary slightly in each gene. The variation this effect causes is, however, relatively small when divergence times between the taxa span over long periods, for example, between different species groups (MADDISON and KNOWLES 2006; NEI 1987). These effects will be ignored for the purpose of this study. Coalescent effects of incomplete lineage sorting can also result in gene tree topologies being different from their species tree topology (DEGNAN and SALTER 2005; ROSENBERG 2002); however, it would be the responsibility of the users of these methods to test for topological concordance of gene and species trees.

Another problem that these algorithms pose is the issue of multiple testing. Under these schemes, a large number of hypothesis tests are performed which can lead to a significant number of false positive errors. A potential solution for this would be to compare the likelihood ratio statistic to a different null distribution (explicitly, the distribution of the maximum value of $2n-2$ chi-squared random values). However, using this distribution poses difficulties when the alternative hypothesis involves more than one branch under positive selection. I have not attempted such adjustments, so the tests described here will tend to be over-sensitive.

5.3 Methods

Benchmarking these methods with real data is difficult. First of all, estimates of the true tree shared between a set of species can be imprecise in terms of the branch rates, divergence times and in many cases the topology. Secondly, the exact processes of evolution that underlie the changes in rate of substitution are unknown. Thus, the actual events of selection that occurred in real datasets are not known and can only be speculated.

The most appropriate means to benchmark these methods is to use simulated data. In this section, we outline a simulation pipeline that was developed which simulates phylogenies, rates, selection events and alignments based on well established models of evolution. Data was simulated using this pipeline to benchmark the algorithms proposed.

5.3.1 Simulation of data

Firstly, trees in time units were simulated to model the species divergence process. Using Phylogen (RAMBAUT 2003), species divergences were simulated under a birth-death model with birth rate of 0.025 and death rate of 0.0125. The birth-death process was run until the trees had 15 taxa. This produced a time tree with simulated taxa and divergence times.

Rates of substitution were then simulated on each of the time tree branches. Rate values were drawn from an uncorrelated lognormal model of rate heterogeneity across branches (ARIS-BROSOU and YANG 2002; RAMBAUT and BROMHAM 1998; THORNE *et al.* 1998) (see Section 2.2.3.1 for details). Different parameter values for mean rate of the distribution were then used to benchmark these algorithms across different levels of divergence. Different divergence levels were simulated as a number of studies have demonstrated that the power associated with comparative methods is dependent on the phylogenetic distance across the species (EDDY 2005; MARGULIES *et al.* 2005a; MIGNONE *et al.* 2003; THOMAS *et al.* 2003). Rates were drawn from lognormal distributions with means (μ_{LN}) of 0.0005, 0.001 and 0.002 and a shape parameter S^2 of 0.05 (the standard deviations of the lognormal distributions were 0.000113, 0.000226 and 0.000453, respectively). The branch lengths on the divergence time trees were multiplied by their rates to produce a tree in substitutions per site. The general lengths of the trees and their branches were made to closely resemble distances between species in real mammalian nuclear datasets. The distributions of tree lengths and mean branch lengths in each of these datasets are shown in Figure 5.3 and Figure 5.4, respectively.

Events of selection were then simulated on the trees. Our aim was to simulate the effects of selection on the rate of substitution. However, no studies thus far have explicitly quantified the change in rates of substitution which occur as a result of significant selection pressures. On the other hand, the effect of selection on the ratio of non-synonymous (dn) to synonymous (ds) substitutions is well characterised. Yang (1998) noted that in primate lysosome genes, the ratio of dn/ds (ω) in lineages where positive selection occurred was between 3.3 and 7.3. Yang determined these values with codon-based likelihood models and found where the likelihood was significantly different, observed values of ω in their data was between these values. Where lineage-specific selection was not present, the background ω (ω_0) was roughly 0.5 in the primate lysosome genes. These values of ω_0 correspond to presence of gene-specific effects.

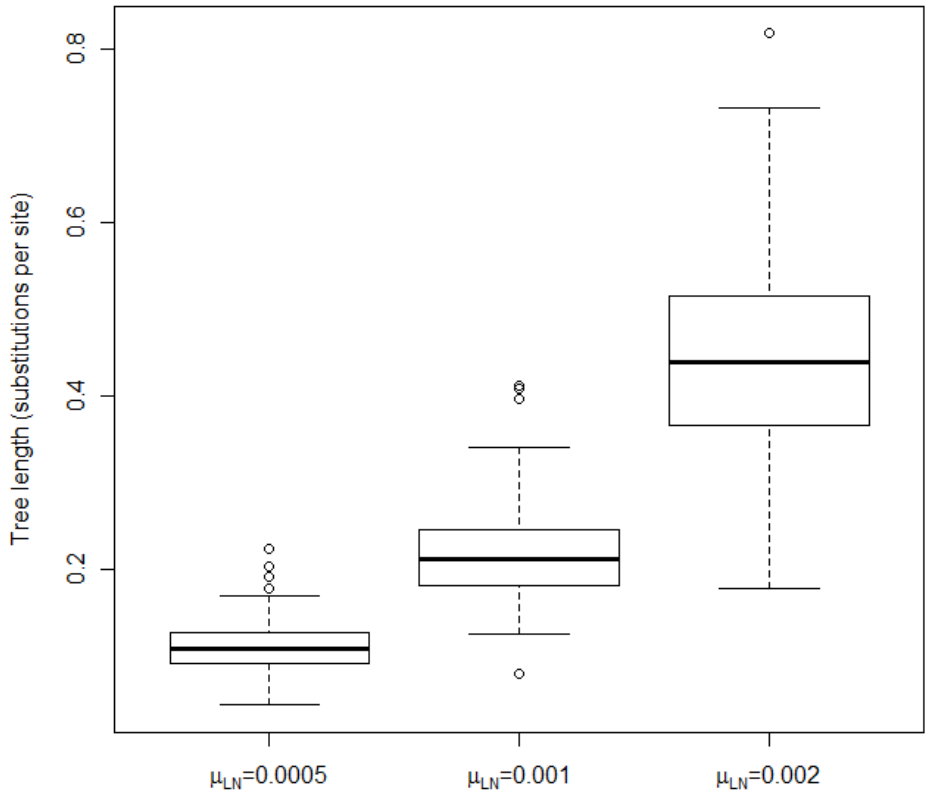


Figure 5.3 The distribution of tree lengths for the datasets under three different values of mean rate, μ_{LN} = 0.0005, 0.001, 0.002.

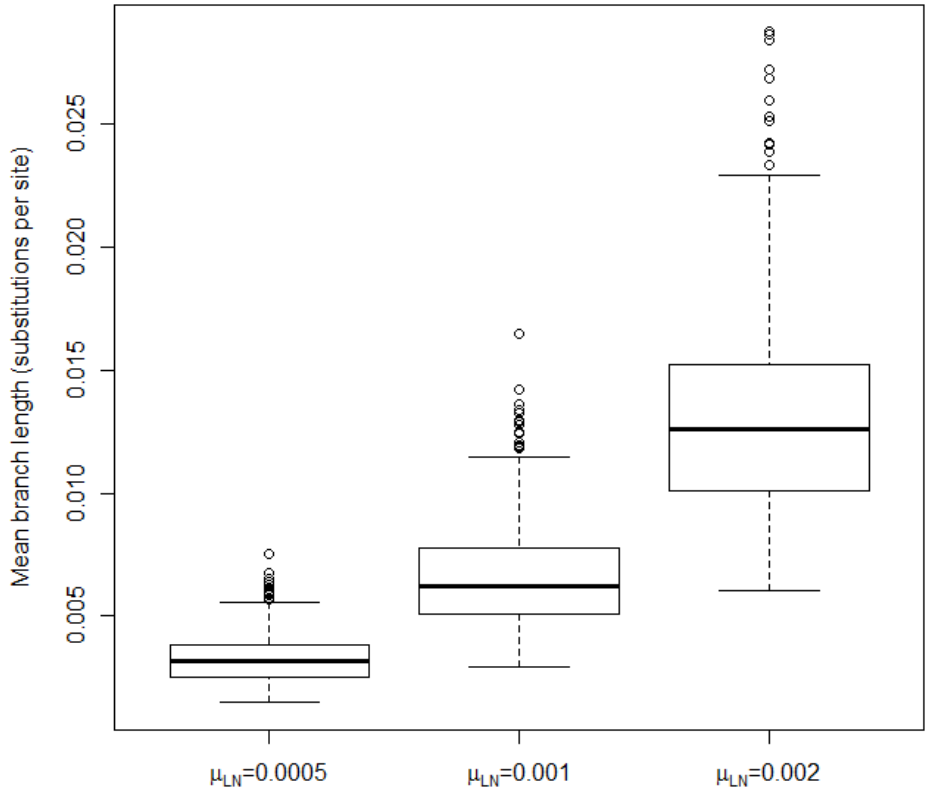


Figure 5.4 The distribution of mean branch lengths for the datasets under three different values of mean rate, μ_{LN} = 0.0005, 0.001, 0.002.

These threshold values were used as guidelines for simulating changes in rate of substitution. Consider the following: for a given gene sequence, *ds* substitutions are unaffected by selection and evolve constantly at a rate equal to the neutral rate of substitution. As *ds* does not change under selection, increases to dn/ds must be a result of an increase in rate of *dn* substitutions. Therefore when $dn/ds > 1$ (corresponding to positive selection), the number of substitutions on the branch increases by the number of additional *dn* substitutions that occurred.

For the standard genetic code, excluding stop codons, there are 41 possible *dn* changes and 20 *ds* changes. Given a value of neutral rate of susbstitution r, if selection occurs and dn/ds changes, then the rate of substitution should change according to:

$$r^* = r\left(\frac{41(dn/ds)+20}{41+20}\right) = r\left(\frac{41\omega+20}{61}\right) \quad (5.3)$$

where r^* is the new value of rate. Essentially, Equation 5.3 denotes the change in rate of substitution on a particular branch as a result of selection.

In the analysis performed, the values of ω and ω_0 for branches on the tree were altered. Gene-specific rate changes (g) corresponding to selection occurring across all lineages on the region, was simulated with parameter ω_0 equal to 0.5 and 1.0. ω_0 = 1.0 denoted regions where no gene-specific effects were present. Where ω_0 = 0.5, the gene-specific effects simulated were similar to those Yang (1998) found in primate lysosome genes.

Consistent with Yang's study, a ω value of 5.0 was simulated on random branches to model occurrences of lineage-specific selection (h). In some datasets, we simulated observations that had one branch under lineage-specific selection (N = 1). In other datasets, a control experiment was done where no lineage-specific selection events were simulated (N = 0). Purifying lineage-specific selection (ω < 1) was not simulated in this study, as there are no definitive indications of appropriate threshold values for these selection events.

In total, there are 12 combinations of parameters for the simulated data – three different values for mean rate of substitution across branches (μ_{LN} = 0.0005, 0.001, 0.002), two values for occurrences of selection across all taxa (ω_0 = 0.5, 1.0) and two scenarios for occurrences of lineage-specific selection where ω = 5.0 (N = 0, 1). 100 trees were simulated for each of these 12 datasets. For every one of these trees, alignments were generated using the program Evolver from the PAML software package (YANG 2007) under a K80 model (KIMURA 1980) with a transition/transversion rate, κ of 10.

5.4 Results

We applied the algorithms A1 and A2 to detecting selection in the simulated datasets. For both algorithms, a general time-reversible model (GTR) (TAVARÉ 1986) was used as the nucleotide substitution model. PAUP* (SWOFFORD 2003) was used to estimate the parameters of the substitution model for each sequence individually. In which case, $\bar{\theta}$ was approximated with estimates of the substitution model parameters θ. The species topology $\bar{\tau}$ and branch lengths (equal to $\bar{\mathbf{t}} \times \bar{\mathbf{r}}$) were constrained to the values on the substitutions per site trees generated; this tree does not contain any changes in branch length from altering ω and thus represents the genetic distances between the species under neutral selective pressures.

A strict molecular clock (ZUCKERKANDL *et al.* 1965) was assumed for the heterogeneity in rate across branches, as it was of interest to maintain high computational speeds for the algorithms. As a strict molecular clock is assumed, the algorithm does not estimate divergence times and rates separately. Rather, values of k and $\mathbf{r}$ that are estimated by these algorithms are directly inferred from the deviation in branch length $\mathbf{t} \times \mathbf{r}$ from neutral ($\bar{\mathbf{t}} \times \bar{\mathbf{r}}$). This direct inference from branch lengths does not affect the likelihood calculations or LRTs computed by the algorithm.

5.4.1 Detecting across-lineage selection

We first examined how effective these methods were at estimating the gene-specific rate changes. Table 5.1a shows the results of whether the method identified the change in k as significant across different parameters. As the procedure for detecting gene-specific changes was the same for A1 and A2, their predictions are identical. The results showed that when no lineage-specific selection was simulated, our method could in most cases accurately detect whether there was a change in gene-specific effects. In five of the six experiments where $N = 0$, the null hypothesis that there was no change in likelihood was correctly retained or rejected, significant to a 5% error rate. The mean estimates of the k parameter are shown in Table 5.2a. The parameters $\omega_0 = 1$ and $\omega_0 = 0.5$ correspond to k of 1 and 0.664, respectively. The estimated

values of k were in most cases very close to the true value of k when no lineage-specific selection events were simulated.

The success varied once lineage-gene-specific effects (N = 1) were factored into the simulated data. Whether our method identified the change in k as significant for these datasets is shown in Table 5.1b. Under these settings, the performance of the algorithms were suboptimal, with high false positive rates of between 30% and 52% when no gene-specific rate changes were simulated (ω_0 = 1). When gene-specific rate changes were simulated (ω_0 = 0.5), the false negative rates were between 1% and 17%, greatly increased from when N = 0. A decrease in accuracy can be observed across the datasets as the mean rate of the branches decreases. This is likely because scaling the length of longer branches has a greater effect on the likelihood than scaling shorter branches. Under these conditions, the estimates of the value of k were on average overestimated (Table 5.2b). The variance in the estimates of k also became bigger, as shown by the 95% confidence intervals. The cause of this overestimation and uncertainty is that the algorithm mistook an increase in rate in a single branch as an increase in rate across the entire gene. As detecting selection across the whole gene always occurs before detecting selection on branches, these algorithms will attempt to account for as much of the change in likelihood as possible by scaling k before accounting for the change by scaling the branches.

Table 5.1 The proportion of samples that identified the change in *k* to be significant across the different parameter values.

(a) When no selection events were simulated on any of the branches (N = 0) (b) When one selection event was simulated on one of the branches (N = 1).

(a)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.07	1
0.001	0.04	1
0.002	0.01	1

(b)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.3	0.83
0.001	0.46	0.89
0.002	0.52	0.99

Table 5.2 The means and 95% confidence intervals for the estimates of *k* across the different parameter values.

The true values for *k* should be 1.0 and 0.664 for ω_0 = 1.0 and ω_0 = 0.5, respectively. (a) When no selection events were simulated on any of the branches (N = 0) (b) When a selection event was simulated on one of the branches (N = 1).

(a)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.998 [0.889,1.119]	0.665 [0.578,0.761]
0.001	1.008 [0.937,1.081]	0.657 [0.607,0.709]
0.002	0.998 [0.945,1.049]	0.659 [0.609,0.703]

(b)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	1.077 [0.926,1.403]	0.774 [0.599,1.071]
0.001	1.098 [0.968,1.334]	0.787 [0.634,1.084]
0.002	1.081 [0.988,1.334]	0.742 [0.635,1.013]

5.4.2 Detecting lineage-specific selection

The individual accuracies of A1 and A2 at detecting lineage-specific selection were then investigated. Table 5.3 shows the false positive and false negative error rates for A1 under each of the 12 datasets. When no lineage-gene-specific effects were present (N = 0), A1 correctly retained the null hypothesis of no selection to a 5% error rate in all datasets. In particular, when ω_0 = 0.5 no errors were produced across all values of μ_{LN} (Table 5.3a).

We hypothesised that the false positives found when ω_0 = 1.0 were a result of changes in k being falsely accepted. When k is incorrectly identified as significantly different, the lengths of the branches being compared at the detecting lineage-specific

selection stage of the algorithm will not be the correct lengths. This can cause biases in all the branch lengths and lead to errors in the predictions. When the results of the analyses across different values of μ_{LN} were examined, it was found that out of the 12 observations that incorrectly identified k as significantly different, only one (8.3%) contained false positives of lineage-specific events. The mean false positive rate among these 12 observations was 0.6%. In contrast, where the null hypothesis of k was correctly retained, 213 of the 288 observations (74.0%) contained false positives, with a mean false positive rate of 4.8%. Contrary to our hypothesis, where A1 made an incorrect prediction on change in k, the false positive rates in predicting lineage-specific selection events actually decreased. Therefore, the incorrect prediction of gene-specific rate changes in this algorithm did not appear to have any negative effects on the accuracy of detecting lineage-gene-specific rate changes.

When selection was simulated on a branch (N = 1), the false positive rate increased in datasets where ω_0 = 0.5 (Table 5.3b). The false positive rate increased as the distances between the taxa became larger, with on average 10% false positives in data where μ_{LN} = 0.0005 and 20% in μ_{LN} = 0.002. In contrast, when ω_0 = 1.0 the false positive rates were only ≈ 5% across all values of μ_{LN}. The false negative rates were also much worse when ω_0 = 0.5 (Table 5.3c), with error rates of between 50% and 58% compared to 8% and 18% when ω_0 = 1.0.

Table 5.3 The false positive and false negative error rates of A1 for detecting lineage-specific selection across the parameter space of the simulations.

The tables are as follows: (a) false positive rate, $N = 0$ (b) false positive rate, $N = 1$ (c) false negative rate, $N = 1$. Where $N = 0$, a false negative rate cannot be calculated.

(a)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.051	0.000
0.001	0.046	0.000
0.002	0.043	0.000

(b)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.052	0.095
0.001	0.046	0.174
0.002	0.046	0.204

(c)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.180	0.580
0.001	0.130	0.500
0.002	0.080	0.580

What was generally observed for A1 is that when gene-specific and lineage-gene-specific events occur individually, the method is able to predict the events of selection fairly accurately. But when the rate of substitution on a branch is influenced by multiple effects, the algorithm fails to separate the confounded effects in the rates. This is problematic as even in the most liberal of real datasets, rates on a branch are almost always affected by multiple effects of rate change. In summary, although A1 can accurately detect events of selection in specific conditions, it is incapable of delineating confounded effects on rate heterogeneity in the data and is therefore unlikely to be robust enough for use with real datasets.

Table 5.4 shows the results of the same analysis performed on A2. When $N = 0$, the algorithm performed extremely poorly, with false positive rates of over 95% across all parameters in the simulated data (Table 5.4a). Upon closer inspection, it was found that in a majority of the cases, A2 assumed that every branch on the tree had significant changes in rate. The reason for this high error rate is that when there are no selection events across the branches of the tree, the likelihood of the tree is not

actually different to $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},k\bar{\mathbf{r}})$. Hence, optimising any set of branches would not significantly alter the likelihood from $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},k\bar{\mathbf{r}})$ itself. Since the lack of significant change in likelihood under A2 implies selection, false positives are generated easily under these data conditions. This problem is a fault that we failed to account for when devising this algorithm.

For datasets where N = 1, the false positive rates were significantly lower than when selection was absent (Table 5.4b). It was observed that when the mean genetic distance increased, the false positive rate decreased while the false negative rates increased (Table 5.4c). Where ω_0 = 0.5, the false negative rate was poor with mean rates of between 44% and 77%. On the other hand, the false negative rate appeared interestingly low for datasets where ω_0 = 1.0, with on average 93% of the selection events being captured by the algorithm when μ_{LN} was 0.0005.

Table 5.4 The false positive and false negative error rates of A2 for detecting lineage-specific selection across the parameter space of the simulations.

The tables are as follows: (a) false positive rate, N = 0 (b) false positive rate, N = 1 (c) false negative rate, N = 1. Where N = 0, a false negative rate cannot be calculated.

(a)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.960	0.965
0.001	0.964	0.978
0.002	0.962	0.980

(b)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.371	0.299
0.001	0.215	0.149
0.002	0.142	0.160

(c)

	Gene specific effects	
Mean branch rate (μ_{LN})	**Not present (ω_0=1)**	**Present (ω_0=0.5)**
0.0005	0.070	0.440
0.001	0.090	0.710
0.002	0.140	0.770

When the results were closely examined, it was found that in quite a few of the simulated alignments where N = 1 and ω_0 = 1, A2 was able to achieve perfect classification of the lineage-specific selection events. The relationship between the prediction accuracies and the characteristics of the data simulated was further investigated. Figure 5.5 shows the error rates for each observation plotted against the relative length of the branch under selection over the total tree length. Overall, as the ratio between selection event branch length and tree length increased, the false positive error rate associated with the estimation decreased. This indicates that A2 is generally more accurate at detecting lineage-specific selection events that occur on relatively longer branches of the tree.

We explicitly compared observations that could classify the selection events without any errors to those that contained at least some false positives or false negatives (Figure 5.6). Again it can be seen that where the prediction was good, the relative length of the selection branch was generally larger. This effect appeared more prominent when μ_{LN} was smaller (Figure 5.6a-c).

The reason for the correlation is this: where the selection branch is proportionally longer, changing its length has a greater impact on the total likelihood of the tree. When this is the case, optimising the rate of these branches from its rate in $\bar{\mathbf{r}}$ will cause a large improvement in likelihood which will be identified as significant in a LRT. Under A2, when selection is being tested on a branch that is not under selection, the length of the long selection branch will be optimised, leading to a significant LRT and thus correctly rejecting the branch as not under selection. When the long branch itself is fixed, all changes in other branches will have a comparatively lesser effect on the likelihood. In this scenario, the LRTs will not be significant and accordingly A2 will correctly classify the branch as being under selection. Hence, where the selection branch is proportionally longer it takes up a larger proportion of the tree likelihood and the error rate will generally be low.

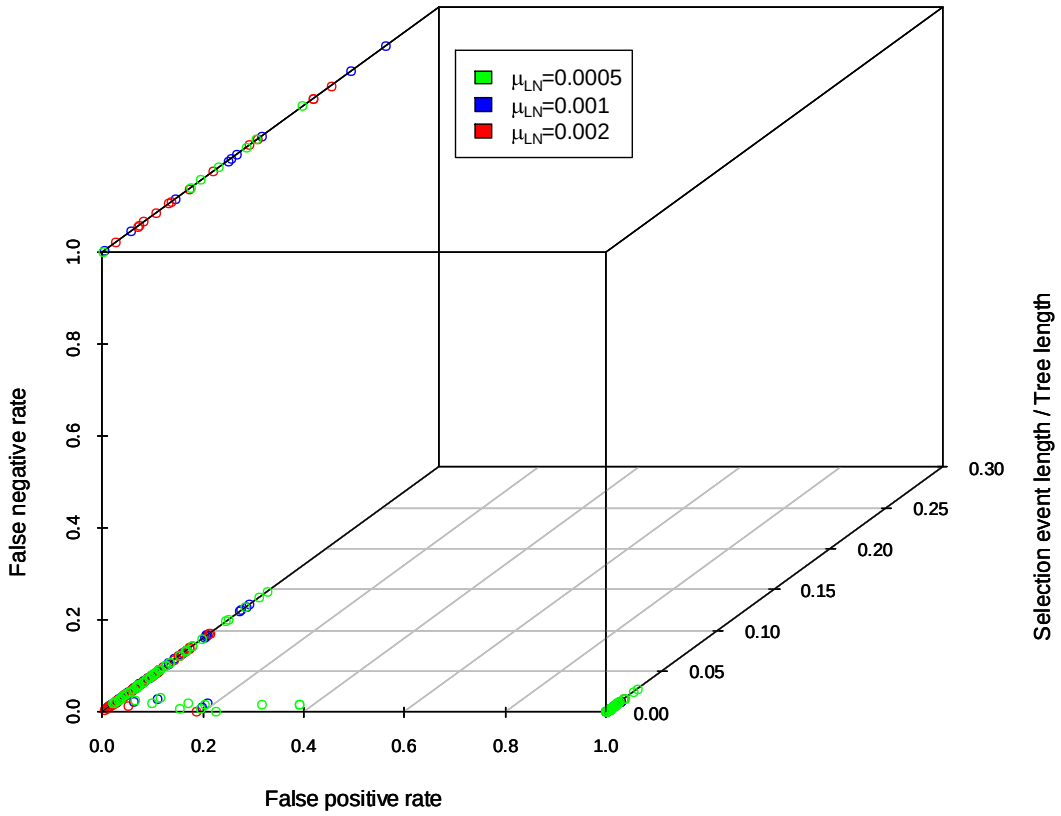


Figure 5.5 Scatter plot of error rates generated using A2 against the relative length of the branch undergoing selection to the total length of the tree.

For this comparison, only conditions where there was a simulated selection event on the branches (N =1) and no gene-specific effects (ω_0 = 1.0) were examined.

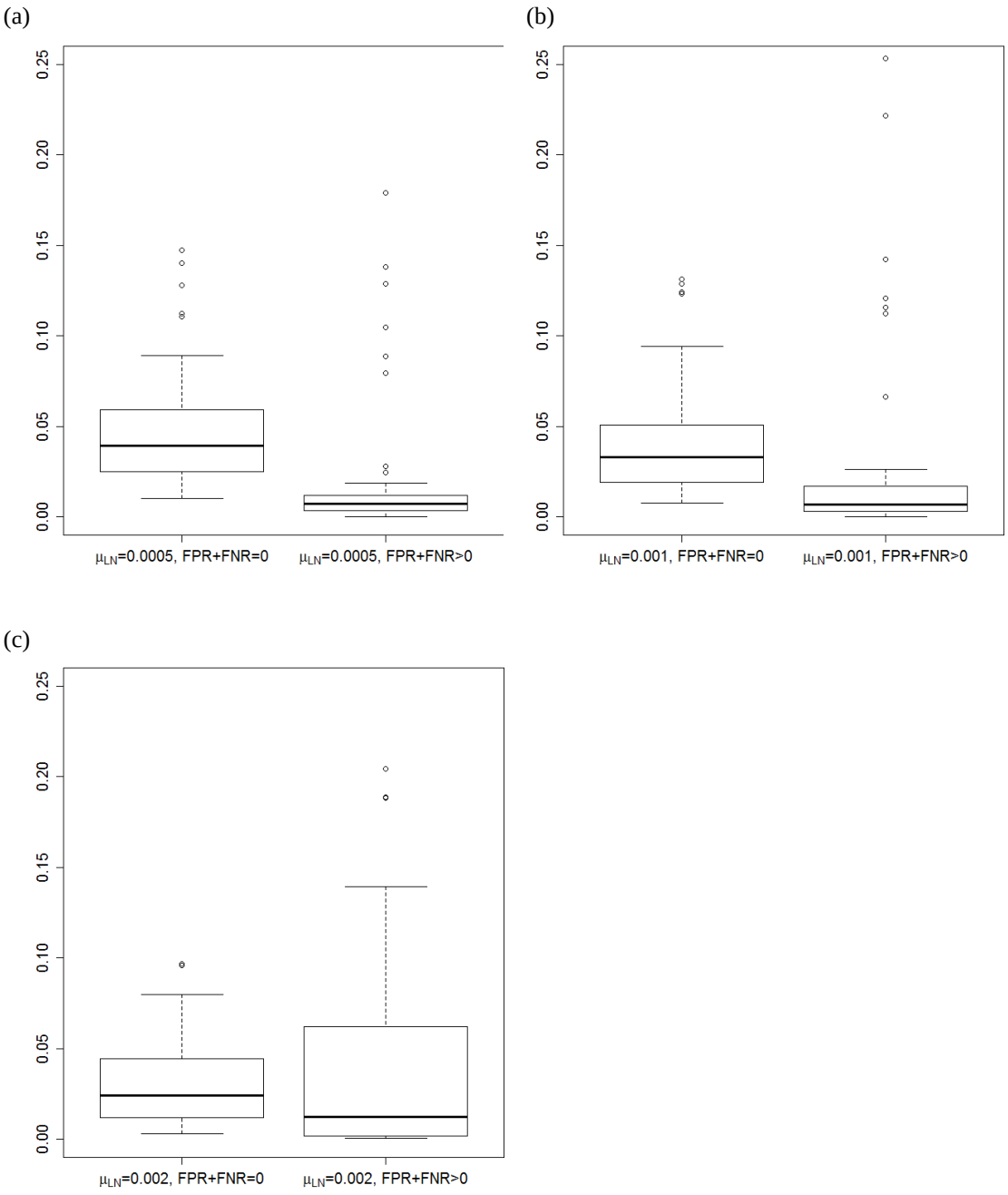


Figure 5.6 Boxplots of the relative length of the selection branch for (a) μ_{LN} = 0.0005 (b) μ_{LN} = 0.001 (c) μ_{LN} = 0.002 compared between datasets with zero error rate and those with some error.
The two boxplots represent a zero error rate (FPR+FNR = 0) and a non-zero error rate (FPR+FNR > 0), respectively.

In retrospect, there are obvious problems with the procedure defined in A2 which prevent it from retaining the null hypotheses that are tested even when the hypothesis is true. When the null hypothesis was not true and selection events were present, false positive rates were still fairly high overall. A2 favoured datasets where the branch

under selection was relatively long compared to the rest of the tree and performed well under precise conditions. A2, like A1 shared the same problems of not being able to distinguish between gene-species and lineage-gene-specific rate changes when both effects were factored into the simulated data.

5.5 Discussion

The potential benefit of methods to detect lineage-specific selection such as those proposed here is their capability to provide a prediction of functionally important regions of the genome. The identification of lineage-specific selection events highlights the importance of certain functions towards a particular species or set of species. Finding functional regions is important in biology, as it helps us understand the underlying biological processes that take place, as well as similarities and differences between species on a molecular level.

Here we have outlined a procedure for detecting selection that occurs across all lineages and within specific lineages given a set of taxa. The method attempts to attribute changes in the rate of substitution from what is expected under neutrality to gene-specific and lineage-gene-specific effects. A useful property of this method is that it is based on established statistical frameworks, namely tree likelihoods and likelihood ratio tests. The main advantage of our techniques is the short computational time required to perform such analyses. For the algorithm A1, the computational time used to analyse all 1200 of the simulated datasets was under five hours on a single 2.4GHz central processing unit. With the exponential growth in amount of genomic data that is becoming publicly available, computational speed benefits such as those offered by these methods are crucial for genome-scale analyses.

The algorithm A1 which performed better out of the two methods, showed some degree of success in identifying rate changes under certain conditions, but when data became increasingly realistic the accuracy of the algorithm began to deteriorate. Much work is required in order for this method to be effective for application toward real data. The known problems and limitations of the algorithms are listed in 5.2.3. The major problems associated with these techniques are its greedy nature and dependence

on an accurate estimate of distances under neutrality. In this study it was found that these algorithms experienced trouble separating confounded effects of rate change on the tree. This suggests that such greedy algorithm-based approaches may not be adequate for identifying changes in the rate of substitution as the underlying process of molecular substitution is extremely intricate. We also acknowledge that the assumption of a neutral tree is an unreasonable assumption to make for real datasets, as neutral rates of substitution are known to vary across different regions of the genome (SUBRAMANIAN and KUMAR 2003). The current understanding of how rates vary across different regions of a genome is yet insufficient for us to establish what should be considered neutral for a particular genomic region. Also the definition of neutral tree is subjective and dependent on the methods of analysis. However, a number of promising phylogenomic studies have been published, suggesting that reasonable estimates of neutral distances across a tree may be on the horizon (GU *et al.* 2005; GU and ZHANG 2004).

One modification to our algorithm that will likely improve the results is to eliminate the greedy nature of these approaches. Under this proposal, $P(D|\bar{\tau},\bar{\mathbf{t}},\bar{\theta},\bar{\mathbf{r}})$ is directly compared to the likelihood of the maximum likelihood tree for the data $P(D|\bar{\tau},\mathbf{t},\bar{\theta},\bar{\mathbf{r}})$. Using likelihood decomposition (see FELSENSTEIN 1981), each of the branches from these two trees can be compared in terms of their contribution towards the total tree likelihood. A LRT can then be used to determine the significance of the difference in likelihood in each branch. This approach, however, is computationally inefficient and would go against our original goal of creating a high-throughput method.

An alternative approach to our algorithm would be to scale the value of $\mathbf{r}'$ to $k\mathbf{r}$ in the initial step of the algorithm, regardless of whether or not the change in k is considered significant. The benefit of this scheme is that it permits gene-specific rates to differ from the neutral rate even when there is insufficient signal to reject the null hypothesis. However, rescaling the gene-specific rate without sufficient evidence at this point could induce false positives in the latter LRTs performed in the lineage-gene-specific stage of the algorithm.

Also for this analysis, significance levels were not adjusted with multiple comparison procedures (MCP) despite the extensive number of hypothesis tests that are performed. The application of corrections for multiple comparisons can be used to adjust the threshold for rejecting the null hypothesis to lower the number of false positive errors. However, the use of MCPs generally causes an increase in false negative error rates which are often considered more important in some studies. Having low false negative rates is problematic, as analyses can lack significant findings and interesting events of selection will often not be identified. The use of MCPs and the choice of technique for correction are dependent on the importance of false positives versus false negatives for the given study. As it is arbitrary to use MCPs, we maintained simplicity in this study by not using them.

An alternative strategy to tackling the problem of detecting lineage-specific rate changes would be to use a Markov-chain Monte Carlo (MCMC) approach with Bayesian stochastic search variable selection (BSSVS) (GEORGE and MCCULLOCH 1993). We propose the following algorithm: begin by constraining the tree topology to $\bar{\tau}$ and divergence times to $\bar{\mathbf{t}}$, then calculating the likelihood $P(D \mid \bar{\tau}, \bar{\mathbf{t}}, \bar{\theta}, \bar{\mathbf{r}})$. For each branch i, an indicator variable s_i = [0, 1] is assigned which denotes whether i is considered to be under selection. If s_i = 1, then i is allowed to vary from its branch rate in $\bar{\mathbf{r}}$ by a factor of r_i. The value of k is also allowed to vary to account for gene-specific rate changes. Each of the parameters of $\mathbf{s}$, $\mathbf{r}$ and k can be sampled in the MCMC using the Metropolis-Hastings algorithm (HASTINGS 1970; METROPOLIS *et al.* 1953) and the probability of accepting a proposed change can be determined by the improvement in likelihood over its previous state. The resulting posterior distributions of k and s_i should provide us with predictions of gene-specific and lineage-specific rate changes on the tree, respectively. Also, conditional on the value of s_i being equal to 1, changes in rate can be determined by whether $\bar{r}_i$ is not contained within the 95% credible set of r_i.

Unlike the greedy algorithm-based approach that was used here, MCMC-based approaches are less likely to return sub-optimal solutions. Also, MCMC provides

added flexibility by allowing us to set informative prior probability distributions on parameters. Useful priors for this application include priors on the likely distribution of k, the probability of a branch being under selection, $P(s=1)$ and the prior distribution of r. A potential drawback of using an MCMC-based approach is that the added accuracy may be a trade-off to the speed.

Another possible extension to improve such methods would be to allow for scaling of subtrees rather than just scaling individual branches. This may more suitably model selection events across branches as selection is often localised among groups of monophyletic species (BRITTEN 1986; GILLESPIE 1991; JOHNSON *et al.* 2001). Accordingly, the appropriate modification to the procedure would be to optimise the likelihood by scaling the rate of a set of monophyletic branches rather than just scaling individual branches.

A side product that was generated out of this study is the simulation pipeline that was developed. Currently, a broad range of tools exist to simulate various aspects of molecular evolution (RAMBAUT 2003; RAMBAUT and GRASSLY 1997; YANG 2007) but these tools are all unlinked and independent to each other. We look to develop our simulation pipeline into an inclusive software package for simulating biological data and molecular evolution. Our goal is for the pipeline to encompass a variety of models for simulating rates of substitution, divergence time processes, selection events, genome evolution patterns and patterns of sequence evolution under nucleotide, protein and codon substitution models. Such a simulation tool could be valuable in studies where benchmarking of algorithms is needed or where researchers may want to model the evolutionary processes. We hope to continue developing this pipeline and keep up with the growing knowledge of models for the underlying processes of evolution.

Overall, we believe there is great potential in high-throughput methods for detecting lineage-specific selection, as they can provide information on the importance of functions across lineages, which would otherwise be difficult to obtain. With the continual growth of genomic data, the added efficiency that these methods can

provide is crucial. In future, we look to uplift the limitations that these methods impose and to reduce the error rates associated with their use.

Chapter 6. Conclusion

In this thesis, I have developed methods for improving how phylogenomic analyses are performed, based on knowledge of variations in the substitution process within and between genomes. As differences in the rate of substitution are characteristic of changes in the underlying evolutionary process, identifying rate heterogeneity can help us better understand genome evolution. This work has explored the importance of the different types of rate variation, and their relevance to phylogenomics.

In this chapter, I discuss the various areas of research carried out in this thesis with respect to their significance for the future of phylogenomics. I outline possible extensions to the work and open questions that have stemmed from the research.

6.1 Relaxed phylogenetics

In Chapter 2, I outlined a novel computational implementation of relaxed molecular clock models in a Bayesian phylogenetics setting. This approach of sampling rates as quantiles has considerable benefits over existing implementations in terms of convergence and efficiency. The use of the inverse Gaussian distribution as a model of rate heterogeneity across branches was also proposed. Although the inverse Gaussian distribution model did not prove to be better than conventional models, the suggestion for its use entertained the idea of alternatives for modelling the rate heterogeneity, which then lead to the development of model averaging techniques in the subsequent chapter. Our findings indicated that by relaxing the assumption of constant rates across branches, better estimates of phylogeny can often be achieved.

As relaxed phylogenetics continues to develop, an ongoing question will be whether there are better ways to model the heterogeneity in rates of substitutions across branches. Specifically, researchers want to know if there are any models that more appropriately represent the behaviour of rate changes across the tree. It is then important to understand which models are most appropriate for particular datasets. Extensive testing is still required to determine the models that are most appropriate for specific data sets. In spite of the work done in Chapter 3, testing the relevance of models is still essential for providing better prior knowledge for model specifications. In future, the types of benchmarks performed here can be more thoroughly performed given the prospect of next-generation sequencing technologies to providing empirical datasets.

Even though they are crucial to understanding the evolutionary histories, the modelling of lineage-specific rates of substitution has often been ignored and consequently relaxed phylogenetic methods are currently still in their infancy. By showing that the quality of tree estimation can be improved using relaxed phylogenetic methods, I hope that my work will increase awareness and use of relaxed molecular clock methods. Ultimately, my goal is to make rate and divergence time estimations part of the norm in phylogenetic analysis.

6.2 Model averaging for relaxed phylogenetics

In my initial study on relaxed phylogenetics, I found the most relevant models for tree estimation on a dataset of mammalian genes. However, there was not one single model that was shown to be consistently optimal across all genes. Also, the relevance of each model when considering a novel dataset remains unclear. Due to these reasons, the most practical solution is to let the data itself determine which are the most appropriate model(s).

The method of model averaging and model selection that I described in Chapter 3 was shown to accurately identify the correct underlying distribution of the substitution rates across branches. Such methods are particularly useful when, for a given dataset, little information is known about the rate variation across branches. Out of the

research done in this thesis, this method will most likely have the greatest impact on the field of phylogenetics. This work presents opportunities for possible future extensions, including the application of model averaging to substitution models (see HUELSENBECK *et al.* 2004) and making the model averaging of relaxed molecular clock models more comprehensive. Model misspecification is still a significant problem in phylogenetics and I firmly believe that phylogenetic analyses should ideally be performed using model averaging techniques.

6.3 Covariation of functionally related genes

In this study, I demonstrated that genes with functionally related gene products co-evolve across species, causing them to have similar rates of substitution. The result of this co-evolution is correlation between the gene tree branch lengths of functionally related genes. Based on this observation, the rate of substitution across different sites of the genome should not be assumed independent. The phenomenon described in this work is important for the application of phylogenomic methods. Our findings suggest that when performing multi-gene analyses, choosing a set of functionally related genes will bias the estimation of species distances.

A question that this work raises is how this correlation could be estimated in an analytical sense, given a set of gene alignments. The motivation behind this question is that the ability to measure this correlation will allow for it to be taken into account when performing joint estimation of rates under multi-gene analyses. So far, this appears to be a rather difficult question to answer. As genes that are functionally related do not all have correlated branch lengths necessarily, the degree of correlation is hard to quantify. Also, by incorporating this information into an analysis, the number of parameters that have to be estimated would increase significantly. Nonetheless, this effect is important to consider if accurate estimates of species divergence times or substitution rates are of interest.

Through my study, progress was made in understanding this pattern of lineage-gene co-evolution. However more in depth analyses are still required, as there remain some elusive aspects of this covariation. Firstly, an interesting question is what is the true

underlying cause of the co-evolution? So far, there have been two hypotheses used to explain why correlated rates of substitution are observed: (1) constraints and compensatory mutations occur to preserve the physical interactions between the protein products of genes, and hence the rates of substitution co-evolve (FRYXELL 1996; PAZOS *et al.* 1997), and (2) common selective pressures act on all genes associated with a particular biological function (HAKES *et al.* 2007). While it is true that these two effects are non-conflicting and that the co-evolution observed may be a mixture of both effects (JUAN *et al.* 2008a; KANN *et al.* 2009), it is relevant to determine which of these effects is the major cause. My understanding is this: for (1) to be true, it must be demonstrated that either the mutations in one gene provoked mutations in other functionally related genes, or the sequence conservation of one gene caused sequence conservation in other functionally related genes. Identifying this correlation may be more easily achieved through population genomic studies or through *in vitro* experiments rather than comparative genomics. For (2) to be true, it must be shown that either "contagious" effects of substitution do not occur, specifically mutations in one gene do not lead to mutations in genes that it is functionally related to; or that functionally related genes evolve simultaneously rather than successively.

A second question that will enable us to better understand these covariation patterns is whether the co-evolution gradually disperses over the number of "connections" that it has. The notion of connectivity can be quantified in at least two ways: one is the number of physical protein interactions a gene has to the rest of the genes in its biological pathway, and another is the path distance (in the pathway interaction network) from the initial mutation event that provoked the compensatory mutations. If hypothesis (1) is true, then we may expect the effect of co-evolution to dampen in genes that code for proteins that are less connected. A simple way to test this theory would be to look at the pairwise similarities between the genes trees of functionally related genes and compare these similarities to their connectivity with one another. If co-evolution does diminish over connectivity, then a positive correlation between these factors should be observed. The answer to this question will inform whether the correlation in gene tree branch lengths can be quantified in an analytical sense.

6.4 Detecting lineage-specific selection

In Chapter 5, I outline a procedure to detect gene-specific and lineage-gene-specific rate changes across sequences. Although the procedure outlined did not achieve a desirable result, it was an endeavour to solve the important problem of detecting selection across branches. Such methods are a great prospect in identifying signatures of selection and provide a means for novel discovery in genomics. Efficient approaches to detecting selection, such as those proposed here, are necessary given the increase in sequence data and recent developments of sequencing technologies.

As various assumptions are made by my method about changes in rate of substitution, these problems need to be addressed in order to improve the performance of the proposed method. Future work in this area will need to first deal with the issues presented in the chapter, in order to work towards addressing the problem of detecting selection-driven rate changes in phylogenomic data. The issues involved include defining a “neutral” or “average” rate of substitution across genomes and relaxing the assumptions of equal divergence times across genes.

Based on experience I have gained through this thesis, I proposed an alternative approach to solving the question of detecting selection. This MCMC-based approach appears rather promising in its design and could potentially overcome a few of the issues experienced by our previous approach. In future, I intend to implement and assess this approach for detecting selection.

In the chapter, a pipeline was also developed for simulating molecular evolution. Simulation of evolution has its merits for benchmarking newly developed methods and testing how realistic models are. I hope to expand on this simulation program to encompass a larger set of models of evolution and to simulate known patterns of evolution.

6.5 General notes

A recurring theme throughout this thesis seems to be the complexity of the variation in rates of substitution. This intricacy in rate heterogeneity can cause bias and error in

phylogenetic analyses. A strong focus should therefore be placed on consolidating the knowledge of all the factors that impact the rates of substitution, so that these effects can be accounted for when performing phylogenetic analyses on a genomic scale. Furthermore, the types of rate variations that occur are not independent of one another. The lineage-specific rates are known to have autocorrelation between branches (BRITTEN 1986; HO *et al.* 2005), gene-specific rates are known to correlate from gene to gene (LI and RODRIGO 2009; PAZOS *et al.* 1997), and lineage-gene-specific effects are affected by both. Hence, genomic sequence data should not be analysed as independent entities.

One way to think of phylogenomic data is as a matrix of genes by lineages, bound by a tree on the horizontal axis and another on the vertical axis. The trees on each axis cluster together rows or columns that have dependent rates. Under this scheme, it is possible to more appropriately annotate rate changes across genomes.

Models have been developed that account for the correlation in rates among lineages (for example, ARIS-BROSOU and YANG 2002; THORNE *et al.* 1998) but models to account for correlation in rates amongst genes have seen less investigation (but see THORNE and KISHINO 2002). Models of covariation between genes are necessary if we are to fully understand the myriad contributions to heterogeneity in rates of evolution. From there, we will have the tools required to explicitly partition the variation in rates for phylogenomic datasets and be able to differentiate between gene-specific, lineage-specific and lineage-gene-specific effects on the rate of evolution.

6.6 Final remarks

The field of biology is ever-changing and computational biologists are continually developing new methods in an effort to keep up with the growth of data that we are being challenged by. Once the sequencing of data has slowed down and becomes more or less saturated, the next wave in genomics will need to focus on developing techniques to adequately mine the data we are already in the process of collecting. Consequently, there should be a focus on developing methodologies that can be used to facilitate discoveries in genomic data.

As such, researchers are beginning to learn that the biology of genomes is far more complex than they once thought. As the understanding of the processes that underlie genome complexity progresses, it is important to relax the assumptions that are made in biological inferences. Throughout this thesis, I have come to realise that older models and methodologies no longer suffice given our current knowledge on the processes that shape genomes. With respect to modelling the evolutionary history among organisms, a fair amount of progress has yet to be made before all factors that influence the evolution are incorporated. The accuracy of phylogenetic reconstructions will be further improved as more of these factors are accounted for by the models. In future, my aim is to contribute towards unravelling the pieces of the puzzle that constitute the evolution of genomes.

References

ADACHI, J., and M. HASEGAWA, 1995 Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: Heterogeneity among amino acid sites. J Mol Evol **40:** 622-628.

ADACHI, J., and M. HASEGAWA, 1996 Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol **42:** 459-468.

AKAIKE, H., 1974 A new look at the statistical model identification. Automatic Control, IEEE Transactions on **19:** 716-723.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res **25:** 3389-3402.

ANISIMOVA, M., and Z. YANG, 2007 Multiple Hypothesis Testing to Detect Lineages under Positive Selection that Affects Only a Few Sites. Mol Biol Evol **24:** 1219-1228.

ARIS-BROSOU, S., and Z. YANG, 2002 Effects of Models of Rate Evolution on Estimation of Divergence Dates with Special Reference to the Metazoan 18S Ribosomal RNA Phylogeny. Syst Biol **51:** 703-714.

ASHBURNER, M., C. A. BALL, J. A. BLAKE, D. BOTSTEIN, H. BUTLER *et al.*, 2000 Gene Ontology: tool for the unification of biology. Nat Genet **25:** 25-29.

ATWELL, S., M. ULTSCH, A. M. DE VOS and J. A. WELLS, 1997 Structural Plasticity in a Remodeled Protein-Protein Interface. Science **278:** 1125-1128.

BAER, C. F., M. M. MIYAMOTO and D. R. DENVER, 2007 Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet **8:** 619-631.

BALMAIN, A., J. GRAY and B. PONDER, 2003 The genetics and genomics of cancer. Nat Genet **33:** 238 - 244.

BARGELLONI, L., S. MARCATO and T. PATARNELLO, 1998 Antarctic fish hemoglobins: Evidence for adaptive evolution at subzero temperature. Proc Natl Acad Sci U S A **95:** 8670-8675.

BARKER, G. M., 2002 Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. Biological Journal of the Linnean Society **76:** 165-194.

BASHIR, A., C. YE, A. L. PRICE and V. BAFNA, 2005 Orthologous repeats and mammalian phylogenetic inference. Genome Res **15:** 998-1006.

BEAUMONT, M. A., and B. RANNALA, 2004 The Bayesian revolution in genetics. Nat Rev Genet **5:** 251-261.

BEERLI, P., and M. PALCZEWSKI, 2010 Unified Framework to Evaluate Panmixia and Migration Direction Among Multiple Sampling Locations. Genetics**:** genetics.109.112532.

BEJERANO, G., A. C. SIEPEL, W. J. KENT and D. HAUSSLER, 2005 Computational screening of conserved genomic DNA in search of functional noncoding elements. Nat Methods **2:** 535-545.

BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57:** 289-300.

BENSON, D. A., I. KARSCH-MIZRACHI, D. J. LIPMAN, J. OSTELL, B. A. RAPP *et al.*, 2000 GenBank. Nucl Acids Res **28:** 15-18.

BENSON, D. A., I. KARSCH-MIZRACHI, D. J. LIPMAN, J. OSTELL and E. W. SAYERS, 2009 GenBank. Nucl Acids Res **37:** D26-31.

BHUTKAR, A., S. M. RUSSO, T. F. SMITH and W. M. GELBART, 2007 Genome-scale analysis of positionally relocated genes. Genome Res **17:** 1880-1887.

BLANCHETTE, M., W. J. KENT, C. RIEMER, L. ELNITSKI, A. F. A. SMIT *et al.*, 2004 Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. Genome Res **14:** 708-715.

BLATTNER, F. R., G. PLUNKETT, III, C. A. BLOCH, N. T. PERNA, V. BURLAND *et al.*, 1997 The Complete Genome Sequence of Escherichia coli K-12. Science **277:** 1453-1462.

BLUNDELL, T. L., and S. P. WOOD, 1975 Is the evolution of insulin Darwinian or due to selectively neutral mutation? Nature **257:** 197-203.

BONNER, T. I., R. HEINEMANN and G. J. TODARO, 1980 Evolution of DNA sequences has been retarded in Malagasy primates. Nature **286:** 420-423.

BRANTON, D., D. W. DEAMER, A. MARZIALI, H. BAYLEY, S. A. BENNER *et al.*, 2008 The potential and challenges of nanopore sequencing. Nat Biotech **26:** 1146-1153.

BRITTEN, R. J., 1986 Rates of DNA sequence evolution differ between taxonomic groups. Science **231:** 1393-1398.

BROCHIER, C., E. BAPTESTE, D. MOREIRA and H. PHILIPPE, 2002 Eubacterial phylogeny based on translational apparatus proteins. Trends in Genetics **18:** 1-5.

BROMHAM, L., and D. PENNY, 2003 The modern molecular clock. Nat Rev Genet **4:** 216-224.

BROOKS, S. P., and P. GIUDICI, 1999 Convergence Assessment for Reversible Jump MCMC Simulations, pp. 733-742 in *Bayesian statistics 6*, edited by J. M. BERNARDO, J. O. BERGER, A. P. DAWID and A. F. M. SMITH. Oxford University Press, Oxford.

BROWN, J. R., C. J. DOUADY, M. J. ITALIA, W. E. MARSHALL and M. J. STANHOPE, 2001 Universal trees based on large combined protein sequence data sets. Nat Genet **28:** 281-285.

BUNCE, M., T. H. WORTHY, M. J. PHILLIPS, R. N. HOLDAWAY, E. WILLERSLEV *et al.*, 2009 The evolutionary history of the extinct ratite moa and New Zealand Neogene paleogeography. Proc Natl Acad Sci U S A **106:** 20646-20651.

CASTELLOE, J. M., and D. L. ZIMMERMAN, 2002 Convergence Assessment for Reversible Jump MCMC Samplers, pp. in *Technical Report*. University of Iowa, Iowa.

CHARLESWORTH, D., B. CHARLESWORTH and G. A. T. MCVEAN, 2001 Genome sequences and evolutionary biology, a two-way interaction. Trends in Ecology & Evolution **16:** 235-242.

CICCARELLI, F. D., T. DOERKS, C. VON MERING, C. J. CREEVEY, B. SNEL *et al.*, 2006 Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science **311:** 1283-1287.

CLARKE, J., H.-C. WU, L. JAYASINGHE, A. PATEL, S. REID *et al.*, 2009 Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nano **4:** 265-270.

CLIFTEN, P. F., L. W. HILLIER, L. FULTON, T. GRAVES, T. MINER *et al.*, 2001 Surveying Saccharomyces Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis. Genome Res **11:** 1175-1186.

CLYDE, M. A., 1999 Bayesian Model Averaging and Model Search Strategies, pp. 157-185 in *Bayesian Statistics 6*, edited by J. M. BERNARDO, J. O. BERGER, A. P. DAWID and A. F. M. SMITH. Oxford University Press, Oxford.

COLLINS, F. S., E. D. GREEN, A. E. GUTTMACHER and M. S. GUYER, 2003 A vision for the future of genomics research. Nature **422:** 835-847.

COLLINS, F. S., and V. A. MCKUSICK, 2001 Implications of the Human Genome Project for Medical Science. JAMA **285:** 540-544.

COLLINS, M. D., P. A. LAWSON, A. WILLEMS, J. J. CORDOBA, J. FERNANDEZ-GARAYZABAL *et al.*, 1994 The Phylogeny of the Genus Clostridium: Proposal of Five New Genera and Eleven New Species Combinations. Int J Syst Bacteriol **44:** 812-826.

DARWIN, C., 1859 *On the Origin of Species*, London.

DAYHOFF, M. O., R. M. SCHWARTZ and B. C. ORCUTT, 1978 A model of evolutionary change in proteins. In Atlas of Protein Sequences and Structure **5:** 345-352.

DEBRUYNE, R., G. CHU, C. E. KING, K. BOS, M. KUCH *et al.*, 2008 Out of America: Ancient DNA Evidence for a New World Origin of Late Quaternary Woolly Mammoths. Current Biology **18:** 1320-1326.

DEGNAN, J. H., and L. A. SALTER, 2005 Gene Tree Distributions under the Coalescent Process. Evolution **59:** 24-37.

DELSUC, F., H. BRINKMANN and H. PHILIPPE, 2005 Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet **6:** 361-375.

DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN, 1977 Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39:** 1-38.

DROSOPHILA 12 GENOMES CONSORTIUM, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature **450:** 203-218.

DRUMMOND, A., and A. RAMBAUT, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology **7:** 214.

DRUMMOND, A., and K. STRIMMER, 2001 PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics **17:** 662-663.

DRUMMOND, A. J., S. Y. W. HO, M. J. PHILLIPS and A. RAMBAUT, 2006 Relaxed Phylogenetics and Dating with Confidence. PLoS Biol **4:** e88.

DUBCHAK, I., M. BRUDNO, G. G. LOOTS, L. PACHTER, C. MAYOR *et al.*, 2000 Active Conservation of Noncoding Sequences Revealed by Three-Way Species Comparisons. Genome Res **10:** 1304-1306.

EDDY, S. R., 2005 A Model of the Statistical Power of Comparative Genome Sequence Analysis. PLoS Biol **3:** e10.

EDWARDS, A. W. F., 1970 Estimation of the Branch Points of a Branching Diffusion Process. Journal of the Royal Statistical Society. Series B (Methodological) **32:** 155-174.

EDWARDS, A. W. F., and L. L. CAVALLI-SFORZA, 1963 The reconstruction of evolution. Annals of Human Genetics **27:** 105-106.

EFRON, B., 1981 Nonparametric estimates of standard error. Biometrika **68:** 589-599.

EISEN, J. A., 1998 Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Res **8:** 163-167.

EISEN, J. A., D. KAISER and R. M. MYERS, 1997 Gastrogenomic delights: A movable feast. Nat Med **3:** 1076-1078.

EVANS, G. A., 2000 Designer science and the "omic" revolution. Nat Biotech **18:** 127-127.

FARRIS, J. S., 1972 Estimating Phylogenetic Trees from Distance Matrices. The American Naturalist **106:** 645-668.

FAY, J. C., and C.-I. WU, 2003 Sequence Divergence, Functional Constraint, and Selection in Protein Evolution. Annual Review of Genomics and Human Genetics **4:** 213-235.

FELSENSTEIN, J., 1973 Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Systematic Zoology **22:** 240-249.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol **17:** 368-376.

FELSENSTEIN, J., 2004 *Inferring Phylogenies*. Sinauer Associates, Inc, Sunderland, Massachusetts.

FOLEY, D. H., 1972 Considerations of Sample and Feature Size. IEEE Transactions on Information Theory **18:** 618-626.

FORD, M. J., 2001 Molecular Evolution of Transferrin: Evidence for Positive Selection in Salmonids. Mol Biol Evol **18:** 639-647.

FORSBERG, R., and F. B. CHRISTIANSEN, 2003 A Codon-Based Model of Host-Specific Selection in Parasites, with an Application to the Influenza A Virus. Mol Biol Evol **20:** 1252-1259.

FRASER, C. M., J. D. GOCAYNE, O. WHITE, M. D. ADAMS, R. A. CLAYTON *et al.*, 1995 The Minimal Gene Complement of Mycoplasma genitalium. Science **270:** 397-404.

FRASER, H. B., A. E. HIRSH, L. M. STEINMETZ, C. SCHARFE and M. W. FELDMAN, 2002 Evolutionary Rate in the Protein Interaction Network. Science **296:** 750-752.

FRYXELL, K. J., 1996 The coevolution of gene family trees. Trends in Genetics **12:** 364-369.

GASCUEL, O., 1997 BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol **14:** 685-695.

GAUT, B. S., S. V. MUSE, W. D. CLARK and M. T. CLEGG, 1992 Relative rates of nucleotide substitution at the rbcl locus of monocotyledonous plants. J Mol Evol **35:** 292-303.

GE, H., A. J. M. WALHOUT and M. VIDAL, 2003 Integrating 'omic' information: a bridge between genomics and systems biology. Trends in Genetics **19:** 551-560.

GEORGE, E. I., and R. E. MCCULLOCH, 1993 Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association **88:** 881-889.

GERTZ, J., G. ELFOND, A. SHUSTROVA, M. WEISINGER, M. PELLEGRINI *et al.*, 2003 Inferring protein interactions from phylogenetic distance matrices. Bioinformatics **19:** 2039-2045.

GILLESPIE, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, New York.

GILLMOR, S. A., T. TAKEUCHI, S. Q. YANG, C. S. CRAIK and R. J. FLETTERICK, 2000 Compromise and accommodation in ecotin, a dimeric macromolecular inhibitor of serine proteases. J Mol Biol **299:** 993-1003.

GIRIBET, G., G. D. EDGECOMBE and W. C. WHEELER, 2001 Arthropod phylogeny based on eight molecular loci and morphology. Nature **413:** 157-161.

GLAZKO, G. V., and M. NEI, 2003 Estimation of Divergence Times for Major Lineages of Primate Species. Mol Biol Evol **20:** 424-434.

GLAZOV, E. A., M. PHEASANT, E. A. MCGRAW, G. BEJERANO and J. S. MATTICK, 2005 Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Res **15:** 800-808.

GOH, C.-S., A. A. BOGAN, M. JOACHIMIAK, D. WALTHER and F. E. COHEN, 2000 Co-evolution of proteins with their interaction partners. J Mol Biol **299:** 283-293.

GOH, C.-S., and F. E. COHEN, 2002 Co-evolutionary Analysis Reveals Insights into Protein-Protein Interactions. J Mol Biol **324:** 177-192.

GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11:** 725-736.

GOWRI-SHANKAR, V., and M. RATTRAY, 2007 A Reversible Jump Method for Bayesian Phylogenetic Inference with a Nonhomogeneous Substitution Model. Mol Biol Evol **24:** 1286-1299.

GRAUR, D., and W.-S. LI, 2000 *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc. , Sunderland, Massachusetts.

GRAY, R. D., and Q. D. ATKINSON, 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature **426:** 435-439.

GRAY, R. D., A. J. DRUMMOND and S. J. GREENHILL, 2009 Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. Science **323:** 479-483.

GRAY, R. D., and F. M. JORDAN, 2000 Language trees support the express-train sequence of Austronesian expansion. Nature **405:** 1052-1055.

GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711-732.

GRISHIN, N. V., Y. I. WOLF and E. V. KOONIN, 2000 From Complete Genomes to Measures of Substitution Rate Variability Within and Between Proteins. Genome Res **10:** 991-1000.

GU, X., 1998 Early Metazoan Divergence Was About 830 Million Years Ago. J Mol Evol **47:** 369-371.

GU, X., W. HUANG, D. XU and H. ZHANG, 2005 GeneContent: software for whole-genome phylogenetic analysis. Bioinformatics **21:** 1713-1714.

GU, X., and H. ZHANG, 2004 Genome Phylogenetic Analysis Based on Extended Gene Contents. Mol Biol Evol **21:** 1401-1408.

GUADET, J., J. JULIEN, J. LAFAY and Y. BRYGOO, 1989 Phylogeny of some Fusarium species, as determined by large-subunit rRNA sequence comparison. Mol Biol Evol **6:** 227-242.

GUINDON, S., and O. GASCUEL, 2003 A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol **52:** 696 - 704.

GUINDON, S., A. G. RODRIGO, K. A. DYER and J. P. HUELSENBECK, 2004 Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci U S A **101:** 12957-12962.

GÜRTLER, V., 1999 The role of recombination and mutation in 16S-23S rDNA spacer rearrangements. Gene **238:** 241-252.

HAKES, L., S. C. LOVELL, S. G. OLIVER and D. L. ROBERTSON, 2007 Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc Natl Acad Sci U S A **104:** 7999-8004.

HARDISON, R. C., 2000 Conserved noncoding sequences are reliable guides to regulatory elements. Trends in Genetics **16:** 369-372.

HARDISON, R. C., 2003 Comparative genomics. PLoS Biol **1:** E58.

HASEGAWA, M., H. KISHINO and T.-A. YANO, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol **22:** 160-174.

HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97-109.

HEISS, N. S., S. W. KNIGHT, T. J. VULLIAMY, S. M. KLAUCK, S. WIEMANN *et al.*, 1998 X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. Nat Genet **19:** 32-38.

HELED, J., and A. J. DRUMMOND, 2010 Bayesian Inference of Species Trees from Multilocus Data. Mol Biol Evol **27:** 570-580.

HENDY, M. D., and D. PENNY, 1982 Branch and bound algorithms to determine minimal evolutionary trees. Mathematical Biosciences **59:** 277-290.

HENDY, M. D., and D. PENNY, 1989 A Framework for the Quantitative Study of Evolutionary Trees. Syst Biol **38:** 297-309.

HO, S. Y. W., 2009 An examination of phylogenetic models of substitution rate variation among lineages. Biology Letters **5:** 421-424.

HO, S. Y. W., S.-O. KOLOKOTRONIS and R. G. ALLABY, 2007a Elevated substitution rates estimated from ancient DNA sequences. Biology Letters **3:** 702-705.

HO, S. Y. W., M. J. PHILLIPS, A. J. DRUMMOND and A. COOPER, 2005 Accuracy of Rate Estimation Using Relaxed-Clock Models with a Critical Focus on the Early Metazoan Radiation. Mol Biol Evol **22:** 1355-1363.

HO, S. Y. W., B. SHAPIRO, M. J. PHILLIPS, A. COOPER and A. J. DRUMMOND, 2007b Evidence for Time Dependency of Molecular Rate Estimates. Syst Biol **56:** 515-522.

HOAGLIN, D. C., and R. E. WELSCH, 1978 The Hat Matrix in Regression and ANOVA. The American Statistician **32:** 17-22.

HOETING, J. A., D. MADIGAN, A. E. RAFTERY and C. T. VOLINSKY, 1999 Bayesian Model Averaging: A Tutorial. Statistical Science **14:** 382-401.

HOLLAND, B. R., L. S. JERMIIN and V. MOULTON, 2006 Improved Consensus Network Techniques for Genome-Scale Phylogeny. Mol Biol Evol **23:** 848-855.

HSU, F., W. J. KENT, H. CLAWSON, R. M. KUHN, M. DIEKHANS *et al.*, 2006 The UCSC Known Genes. Bioinformatics **22:** 1036-1046.

HUELSENBECK, J. P., B. LARGET and M. E. ALFARO, 2004 Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo. Mol Biol Evol **21:** 1123-1133.

HUELSENBECK, J. P., B. LARGET and D. SWOFFORD, 2000 A Compound Poisson Process for Relaxing the Molecular Clock. Genetics **154:** 1879-1892.

HUELSENBECK, J. P., and F. RONQUIST, 2001 MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17:** 754-755.

HUTTLEY, G. A., S. EASTEAL, M. C. SOUTHEY, A. TESORIERO, G. G. GILES *et al.*, 2000 Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Nat Genet **25:** 410-413.

HWANG, U. W., M. FRIEDRICH, D. TAUTZ, C. J. PARK and W. KIM, 2001 Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature **413:** 154-157.

INADA, D. C., A. BASHIR, C. LEE, B. C. THOMAS, C. KO *et al.*, 2003 Conserved Noncoding Sequences in the Grasses. Genome Res **13:** 2030-2041.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

INTERNATIONAL MOUSE GENOME SEQUENCING CONSORTIUM, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

JAMES, T. Y., F. KAUFF, C. L. SCHOCH, P. B. MATHENY, V. HOFSTETTER *et al.*, 2006 Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature **443:** 818-822.

JEFFREYS, H., 1961 *Theory of Probability*. Oxford University Press, London.

JENKINS, G. M., A. RAMBAUT, O. G. PYBUS and E. C. HOLMES, 2002 Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. J Mol Evol **54:** 156-165.

JOHNSON, D. S., B. DAVIDSON, C. D. BROWN, W. C. SMITH and A. SIDOW, 2004 Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance. Genome Res **14:** 2448-2456.

JOHNSON, M. E., L. VIGGIANO, J. A. BAILEY, M. ABDUL-RAUF, G. GOODWIN *et al.*, 2001 Positive selection of a gene family during the emergence of humans and African apes. Nature **413:** 514-519.

JONES, D. T., W. R. TAYLOR and J. M. THORNTON, 1992 The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8:** 275-282.

JORDAN, I. K., L. MARINO-RAMIREZ, Y. I. WOLF and E. V. KOONIN, 2004 Conservation and Coevolution in the Scale-Free Human Gene Coexpression Network. Mol Biol Evol **21:** 2058-2070.

JOTHI, R., T. PRZYTYCKA and L. ARAVIND, 2007 Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinformatics **8:** 173.

JUAN, D., F. PAZOS and A. VALENCIA, 2008a Co-evolution and co-adaptation in protein networks. FEBS Letters **582:** 1225-1230.

JUAN, D., F. PAZOS and A. VALENCIA, 2008b High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci U S A **105:** 934-939.

JUCOVIC, M., and R. W. HARTLEY, 1996 Protein--Protein Interaction: A Genetic Selection for Compensating Mutations at the Barnase--Barstar Interface. Proc Natl Acad Sci U S A **93:** 2343-2347.

JUKES, T. H., and M. KIMURA, 1984 Evolutionary constraints and the neutral theory. J Mol Evol **21:** 90-92.

JUNGNICKEL, D., 2008 The Greedy Algorithm pp. 127-151 in *Graphs, Networks and Algorithms* edited by D. JUNGNICKEL. Springer, Berlin ; New York.

KANN, M. G., B. A. SHOEMAKER, A. R. PANCHENKO and T. M. PRZYTYCKA, 2009 Correlated Evolution of Interacting Proteins: Looking Behind the Mirrortree. J Mol Biol **385:** 91-98.

KAROLCHIK, D., R. BAERTSCH, M. DIEKHANS, T. S. FUREY, A. HINRICHS *et al.*, 2003 The UCSC Genome Browser Database. Nucl Acids Res **31:** 51-54.

KASS, R. E., and A. E. RAFTERY, 1995 Bayes Factors. Journal of the American Statistical Association **90:** 773-795.

KEIGHTLEY, P. D., G. V. KRYUKOV, S. SUNYAEV, D. L. HALLIGAN and D. J. GAFFNEY, 2005 Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res **15:** 1373-1378.

KELLEHER, A. D., C. LONG, E. C. HOLMES, R. L. ALLEN, J. WILSON *et al.*, 2001 Clustered Mutations in HIV-1 Gag Are Consistently Required for Escape from Hla-B27–Restricted Cytotoxic T Lymphocyte Responses. The Journal of Experimental Medicine **193:** 375-386.

KELLIS, M., N. PATTERSON, M. ENDRIZZI, B. BIRREN and E. S. LANDER, 2003 Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:** 241-254.

KENT, W. J., 2002 BLAT—The BLAST-Like Alignment Tool. Genome Res **12:** 656-664.

KIM, S. Y., and J. K. PRITCHARD, 2007 Adaptive Evolution of Conserved Noncoding Elements in Mammals. PLoS Genetics **3:** e147.

KIM, W. K., D. M. BOLSER and J. H. PARK, 2004 Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics **20:** 1138-1150.

KIMURA, M., 1967 On the evolutionary adjustment of spontaneous mutation rates. Genetical Research **9:** 23-34.

KIMURA, M., 1968 Evolutionary Rate at the Molecular Level. Nature **217:** 624-626.

KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16:** 111-120.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KISHINO, H., J. L. THORNE and W. J. BRUNO, 2001 Performance of a Divergence Time Estimation Method under a Probabilistic Model of Rate Evolution. Mol Biol Evol **18:** 352-361.

KITAZOE, Y., H. KISHINO, P. J. WADDELL, N. NAKAJIMA, T. OKABAYASHI *et al.*, 2007 Robust Time Estimation Reconciles Views of the Antiquity of Placental Mammals. PLoS ONE **2:** e384.

KNOWLES, L. L., 2009 Statistical Phylogeography. Annual Review of Ecology, Evolution, and Systematics **40:** 593-612.

KOPP, A., and J. R. TRUE, 2002 Phylogeny of the Oriental Drosophila melanogaster species group: a multilocus reconstruction. Syst Biol **51:** 786-805.

KORBER, B., J. THEILER and S. WOLINSKY, 1998 Limitations of a Molecular Clock Applied to Considerations of the Origin of HIV-1. Science **280:** 1868-1871.

KOSAKOVSKY POND, S. L., and S. D. W. FROST, 2005 A Genetic Algorithm Approach to Detecting Lineage-Specific Variation in Selection Pressure. Mol Biol Evol **22:** 478-485.

KUHNER, M., and J. FELSENSTEIN, 1994 A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates [published erratum appears in Mol Biol Evol 1995 May;12(3):525]. Mol Biol Evol **11:** 459-468.

LECOMPTE, O., R. RIPP, J.-C. THIERRY, D. MORAS and O. POCH, 2002 Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucl Acids Res **30:** 5382-5390.

LEIGH, J. W., E. SUSKO, M. BAUMGARTNER and A. J. ROGER, 2008 Testing Congruence in Phylogenomic Analysis. Syst Biol **57:** 104 - 115.

LEINONEN, R., F. G. DIEZ, D. BINNS, W. FLEISCHMANN, R. LOPEZ *et al.*, 2004 UniProt archive. Bioinformatics **20:** 3236-3237.

LEMEY, P., A. RAMBAUT, A. J. DRUMMOND and M. A. SUCHARD, 2009 Bayesian Phylogeography Finds Its Roots. PLoS Comput Biol **5:** e1000520.

LEPAGE, T., D. BRYANT, H. PHILIPPE and N. LARTILLOT, 2007 A General Comparison of Relaxed Molecular Clock Models. Mol Biol Evol **24:** 2669-2680.

LEVY, S., S. HANNENHALLI and C. WORKMAN, 2001 Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics **17:** 871-877.

LI, W.-H., D. L. ELLSWORTH, J. KRUSHKAL, B. H. J. CHANG and D. HEWETT-EMMETT, 1996 Rates of Nucleotide Substitution in Primates and Rodents and the Generation-Time Effect Hypothesis. Molecular Phylogenetics and Evolution **5:** 182-187.

LI, W. L. S., and A. G. RODRIGO, 2009 Covariation of Branch Lengths in Phylogenies of Functionally Related Genes. PLoS ONE **4:** e8487.

LIN, J., and M. GERSTEIN, 2000 Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res **10:** 808-818.

LIU, L., and D. K. PEARL, 2007 Species Trees from Gene Trees: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions. Syst Biol **56:** 504-514.

LIU, W., B. SCHMIDT, G. VOSS and W. MÜLLER-WITTIG, 2006 GPU-ClustalW: Using Graphics Hardware to Accelerate Multiple Sequence Alignment, pp. 363-374.

LOOTS, G. G., R. M. LOCKSLEY, C. M. BLANKESPOOR, Z. E. WANG, W. MILLER *et al.*, 2000 Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons Science **288:** 136-140.

LÖYTYNOJA, A., and M. C. MILINKOVITCH, 2001 Molecular phylogenetic analyses of the mitochondrial ADP-ATP carriers: The Plantae/Fungi/Metazoa trichotomy revisited. Proc Natl Acad Sci U S A **98:** 10202-10207.

LU, Y., and M. D. RAUSHER, 2003 Evolutionary Rate Variation in Anthocyanin Pathway Genes. Mol Biol Evol **20:** 1844-1853.

LUNTER, G., C. P. PONTING and J. HEIN, 2006 Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model. PLoS Comput Biol **2:** e5.

MADDISON, W. P., and L. L. KNOWLES, 2006 Inferring Phylogeny Despite Incomplete Lineage Sorting. Syst Biol **55:** 21-30.

MANAVSKI, S., and G. VALLE, 2008 CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. BMC Bioinformatics **9:** S10.

MANLY, B. F. J., 2004 *Multivariate Statistical Methods: A Primer*. Chapman and Hall/CRC, Boca Raton.

MARDIS, E. R., 2008 The impact of next-generation sequencing technology on genetics. Trends in Genetics **24:** 133-141.

MARGULIES, E. H., J. P. VINSON, NISC COMPARATIVE SEQUENCING PROGRAM, W. MILLER, D. B. JAFFE *et al.*, 2005a An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci U S A **102:** 4795-4800.

MARGULIES, M., M. EGHOLM, W. E. ALTMAN, S. ATTIYA, J. S. BADER *et al.*, 2005b Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:** 376-380.

MARIÑO-RAMÍREZ, L., O. BODENREIDER, N. KANTZ and I. K. JORDAN, 2006 Co-evolutionary Rates of Functionally Related Yeast Genes. Evolutionary Bioinformatics **2006:** 295-300.

MARTIN, A. P., and S. R. PALUMBI, 1993 Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci U S A **90:** 4087-4091.

MASON-GAMER, R. J., and E. A. KELLOGG, 1996 Testing for Phylogenetic Conflict Among Molecular Data Sets in the Tribe Triticeae (Gramineae). Syst Biol **45:** 524-545.

MAU, B., and M. A. NEWTON, 1997 Phylogenetic Inference for Binary Data on Dendograms Using Markov Chain Monte Carlo. Journal of Computational and Graphical Statistics **6:** 122-131.

MAU, B., M. A. NEWTON and B. LARGET, 1999 Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55:** 1-12.

MAXAM, A. M., and W. GILBERT, 1977 A new method for sequencing DNA. Proc Natl Acad Sci U S A **74:** 560-564.

MCCLELLAND, M., L. FLOREA, K. SANDERSON, S. W. CLIFTON, J. PARKHILL *et al.*, 2000 Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three Salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. Nucl Acids Res **28:** 4974-4986.

MCNEMAR, Q., 1947 Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12:** 153-157.

MEDINA, M., 2005 Genomes, phylogeny, and evolutionary systems biology. Proc Natl Acad Sci U S A **102:** 6630-6635.

MESSIER, W., and C.-B. STEWART, 1997 Episodic adaptive evolution of primate lysozymes. Nature **385:** 151.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER and E. TELLER, 1953 Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics **21:** 1087-1092.

MIGNONE, F., G. GRILLO, S. LIUNI and G. PESOLE, 2003 Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. Nucl Acids Res **31:** 4639-4645.

MINTSERIS, J., and Z. WENG, 2005 Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A **102:** 10930-10935.

MIYATA, T., S. MIYAZAWA and T. YASUNAGA, 1979 Two types of amino acid substitutions in protein evolution. J Mol Evol **12:** 219-236.

MIYATA, T., T. YASUNAGA and T. NISHIDA, 1980 Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci U S A **77:** 7328-7332.

MONDRAGON-PALOMINO, M., B. C. MEYERS, R. W. MICHELMORE and B. S. GAUT, 2002 Patterns of Positive Selection in the Complete NBS-LRR Gene Family of Arabidopsis thaliana. Genome Res **12:** 1305-1315.

MOROZOVA, O., and M. A. MARRA, 2008 Applications of next-generation sequencing technologies in functional genomics. Genomics **92:** 255-264.

MOULTON, V., C. SEMPLE and M. STEEL, 2007 Optimizing phylogenetic diversity under constraints. Journal of Theoretical Biology **246:** 186-194.

MUSE, S., and B. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11:** 715-724.

NACHMAN, M. W., and S. L. CROWELL, 2000 Estimate of the Mutation Rate per Nucleotide in Humans. Genetics **156:** 297-304.

NATALE, D., U. SHANKAVARAM, M. GALPERIN, Y. WOLF, L. ARAVIND *et al.*, 2000 Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biology **1:** research0009.0001 - research0009.0019.

NEE, S., E. C. HOLMES, A. RAMBAUT and P. H. HARVEY, 1995 Inferring population history from molecular phylogenies. Philos Trans R Soc Lond B Biol Sci **349:** 25-31.

NEI, M., 1987 *Molecular evolutionary genetics*. Columbia University Press, New York.

NEWTON, M. A., and A. E. RAFTERY, 1994 Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society. Series B (Methodological) **56:** 3-48.

NIELSEN, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity **86:** 641-647.

NIELSEN, R., 2005 Molecular Signatures of Natural Selection. Annual Review of Genetics **39:** 197-218.

NOVACEK, M. J., 2001 Mammalian phylogeny: Genes and supertrees. Current Biology **11:** R573-R575.

OGG, G. S., X. JIN, S. BONHOEFFER, P. R. DUNBAR, M. A. NOWAK *et al.*, 1998 Quantitation of HIV-1-Specific Cytotoxic T Lymphocytes and Plasma Load of Viral RNA. Science **279:** 2103-2106.

OHTA, T., 1972 Population size and rate of evolution. J Mol Evol **1:** 305-314.

OHTA, T., 1987 Very slightly deleterious mutations and the molecular clock. J Mol Evol **26:** 1-6.

ORLANDO, L., J. L. METCALF, M. T. ALBERDI, M. TELLES-ANTUNES, D. BONJEAN *et al.*, 2009 Revising the recent evolutionary history of equids using ancient DNA. Proc Natl Acad Sci U S A **106:** 21754-21759.

PAGEL, M., and A. MEADE, 2008 Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. Philosophical Transactions of the Royal Society B: Biological Sciences **363:** 3955-3964.

PAGÈS, S., A. BÉLAÏCH, J.-P. BÉLAÏCH, E. MORAG, R. LAMED *et al.*, 1997 Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: Prediction of specificity determinants of the dockerin domain. Proteins: Structure, Function, and Genetics **29:** 517-527.

PAZOS, F., M. HELMER-CITTERICH, G. AUSIELLO and A. VALENCIA, 1997 Correlated mutations contain information about protein-protein interaction. J Mol Biol **271:** 511-523.

PAZOS, F., J. A. G. RANEA, D. JUAN and M. J. E. STERNBERG, 2005 Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome. J Mol Biol **352:** 1002-1015.

PAZOS, F., and A. VALENCIA, 2001 Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. **14:** 609-614.

PELLEGRINI, M., E. M. MARCOTTE, M. J. THOMPSON, D. EISENBERG and T. O. YEATES, 1999 Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A **96:** 4285-4288.

PENNACCHIO, L. A., and E. M. RUBIN, 2001 Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet **2:** 100-109.

PHILIPPE, H., N. LARTILLOT and H. BRINKMANN, 2005a Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol **22:** 1246-1253.

PHILIPPE, H., Y. ZHOU, H. BRINKMANN, N. RODRIGUE and F. DELSUC, 2005b Heterotachy and long-branch attraction in phylogenetics. BMC Evolutionary Biology **5:** 50.

PHILLIPS, M. J., F. DELSUC and D. PENNY, 2004 Genome-Scale Phylogeny and the Detection of Systematic Biases. Mol Biol Evol **21:** 1455-1458.

PHILLIPS, R. E., S. ROWLAND-JONES, D. F. NIXON, F. M. GOTCH, J. P. EDWARDS *et al.*, 1991 Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature **354:** 453-459.

POUMBOURIOS, P., A. L. MAERZ and H. E. DRUMMER, 2003 Functional Evolution of the HIV-1 Envelope Glycoprotein 120 Association Site of Glycoprotein 41. J Biol Chem **278:** 42149-42160.

PRASAD, A. B., M. W. ALLARD, N. C. S. PROGRAM and E. D. GREEN, 2008 Confirming the Phylogeny of Mammals by Use of Large Comparative Sequence Data Sets. Mol Biol Evol **25:** 1795-1808.

QIU, Y.-L., L. LI, B. WANG, Z. CHEN, V. KNOOP *et al.*, 2006 The deepest divergences in land plants inferred from phylogenomic evidence. Proc Natl Acad Sci U S A **103:** 15511-15516.

QUEIROZ, K., and M. J. DONOGHUE, 1988 Phylogenetic Systematics and the Species Problem. Cladistics **4:** 317-338.

RAMANI, A. K., and E. M. MARCOTTE, 2003 Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity. J Mol Biol **327:** 273-284.

RAMBAUT, A., 2003 Phylogen v1.1 Available at: http://evolve.zoo.ox.ac.uk/, pp.

RAMBAUT, A., and L. BROMHAM, 1998 Estimating divergence dates from molecular sequences. Mol Biol Evol **15:** 442-448.

RAMBAUT, A., and A. J. DRUMMOND, 2007 Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer, pp.

RAMBAUT, A., and N. C. GRASSLY, 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13:** 235-238.

RAMBAUT, A., D. POSADA, K. A. CRANDALL and E. C. HOLMES, 2004 The causes and consequences of HIV evolution. Nat Rev Genet **5:** 52-61.

RANNALA, B., and Z. YANG, 1996 Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J Mol Evol **43:** 304-311.

RANNALA, B., and Z. YANG, 2003 Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. Genetics **164:** 1645-1656.

RANNALA, B., and Z. YANG, 2007 Inferring Speciation Times under an Episodic Molecular Clock. Syst Biol **56:** 453-466.

RANWEZ, V., F. DELSUC, S. RANWEZ, K. BELKHIR, M.-K. TILAK *et al.*, 2007 OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. BMC Evolutionary Biology **7:** 241.

RAUDYS, S. J., and A. K. JAIN, 1991 Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Trans. Pattern Anal. Mach. Intell. **13:** 252-264.

RAUSHER, M. D., R. E. MILLER and P. TIFFIN, 1999 Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol **16:** 266-274.

REAL, L. A., J. C. HENDERSON, R. BIEK, J. SNAMAN, T. L. JACK *et al.*, 2005 Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. Proc Natl Acad Sci U S A **102:** 12107-12111.

REEVES, J. H., 1992 Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol **35:** 17-31.

REIS-FILHO, J., 2009 Next-generation sequencing. Breast Cancer Research **11:** S12.

REYES, A., C. GISSI, F. CATZEFLIS, E. NEVO, G. PESOLE *et al.*, 2004 Congruent Mammalian Trees from Mitochondrial and Nuclear Genes Using Bayesian Methods. Mol Biol Evol **21:** 397-403.

RITCHIE, M. E., J. SILVER, A. OSHLACK, M. HOLMES, D. DIYAGAMA *et al.*, 2007 A comparison of background correction methods for two-colour microarrays. Bioinformatics **23:** 2700-2707.

RIVAS, E., and S. EDDY, 2001 Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics **2:** 8.

RODRIGUEZ-EZPELETA, N., H. BRINKMANN, B. ROURE, N. LARTILLOT, B. F. LANG *et al.*, 2007 Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies. Syst Biol **56:** 389-399.

ROKAS, A., N. KING, J. FINNERTY and S. B. CARROLL, 2003a Conflicting phylogenetic signals at the base of the metazoan tree. Evolution & Development **5:** 346-359.

ROKAS, A., D. KRUGER and S. B. CARROLL, 2005 Animal Evolution and the Molecular Signature of Radiations Compressed in Time. Science **310:** 1933-1938.

ROKAS, A., B. L. WILLIAMS, N. KING and S. B. CARROLL, 2003b Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425:** 798-804.

RONQUIST, F., and J. P. HUELSENBECK, 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572-1574.

ROSENBERG, N. A., 2002 The Probability of Topological Concordance of Gene Trees and Species Trees. Theoretical Population Biology **61:** 225-247.

ROSS, H. A., S. MURUGAN and W. L. SIBON LI, 2008 Testing the Reliability of Genetic Methods of Species Identification via Simulation. Syst Biol **57:** 216-230.

ROSS, H. A., and A. G. RODRIGO, 2002 Immune-Mediated Positive Selection Drives Human Immunodeficiency Virus Type 1 Molecular Variation and Predicts Disease Duration. J. Virol. **76:** 11715-11720.

RUSK, N., 2009 Cheap third-generation sequencing. Nat Meth **6:** 244-244.

SANDERSON, M. J., 1997 A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. Mol Biol Evol **14:** 1218-1231.

SANDERSON, M. J., 2002 Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. Mol Biol Evol **19:** 101-109.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A **74:** 5463-5467.

SATO, T., Y. YAMANISHI, K. HORIMOTO, H. TOH and M. KANEHISA, 2003 Prediction of Protein-Protein Interactions from Phylogenetic Trees Using Partial Correlation Coefficient. Genome Informatics **14:** 496-497.

SAWYER, S. L., L. I. WU, M. EMERMAN and H. S. MALIK, 2005 Positive selection of primate TRIM5 Î± identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A **102:** 2832-2837.

SCHEIRER, C., W. RAY and N. HARE, 1976 The analysis of ranked data derived from completely randomized factorial designs. Biometrics **32:** 429-434.

SHAKHNOVICH, B. E., E. DEEDS, C. DELISI and E. SHAKHNOVICH, 2005 Protein structure and evolutionary history determine sequence space topology. Genome Res **15:** 385-392.

SHANNON, P., A. MARKIEL, O. OZIER, N. S. BALIGA, J. T. WANG *et al.*, 2003 Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res **13:** 2498-2504.

SHAPIRO, B. J., and E. J. ALM, 2008 Comparing Patterns of Natural Selection across Species Using Selective Signatures. PLoS Genetics **4:** e23.

SHAPIRO, S. S., and M. B. WILK, 1965 An analysis of variance test for normality (complete samples). Biometrika **52:** 591-611.

SIBLEY, C., and J. AHLQUIST, 1984 The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J Mol Evol **20:** 2-15.

SICHERITZ-PONTEN, T., and S. G. ANDERSSON, 2001 A phylogenomic approach to microbial evolution. Nucleic Acids Res **29:** 545-552.

SIEPEL, A., G. BEJERANO, J. S. PEDERSEN, A. S. HINRICHS, M. HOU *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15:** 1034-1050.

SIEPEL, A., K. POLLARD and D. HAUSSLER, 2006 New methods for detecting lineage-specific selection, pp. in *Proceedings of the 10th International Conference on Research in Computational Molecular Biology*.

SING, T., O. SANDER, N. BEERENWINKEL and T. LENGAUER, 2005 ROCR: visualizing classifier performance in R. Bioinformatics **21:** 3940-3941.

SINGH, S., and D. P. WALL, 2008 Testing the Accuracy of Eukaryotic Phylogenetic Profiles for Prediction of Biological Function. Evolutionary Bioinformatics **4:** 217–223.

SJOLANDER, K., 2004 Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics **20:** 170-179.

SMITH, T. F., and M. S. WATERMAN, 1981 Identification of common molecular subsequences. J Mol Biol **147:** 195-197.

SOKAL, R. R., and C. D. MICHENER, 1958 A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull **38:** 1409-1438.

SOKAL, R. R., and F. J. ROHLF, 1995 *Biometry : the principles and practice of statistics in biological research*. New York : Freeman, c1995.

STARK, A., M. F. LIN, P. KHERADPOUR, J. S. PEDERSEN, L. PARTS *et al.*, 2007 Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature **450:** 219-232.

STEIN, L. D., Z. BAO, D. BLASIAR, T. BLUMENTHAL, M. R. BRENT *et al.*, 2003 The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics. PLoS Biol **1:** e45.

STEINER, D. F., S. J. CHAN, J. M. WELSH and S. C. M. KWOK, 2003 Structure and Evolution of the Insulin Gene. Annual Review of Genetics **19:** 463-484.

STEIPER, M. E., and N. M. YOUNG, 2006 Primate molecular divergence dates. Molecular Phylogenetics and Evolution **41:** 384-394.

SUBRAMANIAN, S., and S. KUMAR, 2003 Neutral Substitutions Occur at a Faster Rate in Exons Than in Noncoding DNA in Primate Genomes. Genome Res **13:** 838-844.

SUCHARD, M. A., R. E. WEISS and J. S. SINSHEIMER, 2001 Bayesian Selection of Continuous-Time Markov Chain Evolutionary Models. Mol Biol Evol **18:** 1001-1013.

SUMIYAMA, K., N. SAITOU and S. UEDA, 2002 Adaptive Evolution of the IgA Hinge Region in Primates. Mol Biol Evol **19:** 1093-1099.

SWOFFORD, D. L., 2003 PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). , pp. Sinauer Associates, Sunderland, Massachusetts.

TAGLE, D. A., B. F. KOOP, M. GOODMAN, J. L. SLIGHTOM, D. L. HESS *et al.*, 1988 Embryonic var epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus) : Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol **203:** 439-455.

TAKAGI, K., S. KUMAGAI, I. MATSUNAGA and Y. KUSAKA, 1997 Application of inverse Gaussian distribution to occupational exposure data. Ann Occup Hyg **41:** 505-a-514.

TAKAHATA, N., 1986 An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genetics Research **48:** 187-190.

TAKAHATA, N., Y. SATTA and J. KLEIN, 1995 Divergence Time and Population Size in the Lineage Leading to Modern Humans. Theoretical Population Biology **48:** 198-221.

TAN, S.-H., Z. ZHANG and S.-K. NG, 2004 ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. Nucl. Acids Res. **32:** W69-72.

TAVARÉ, S., 1986 Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences, pp. 57-86 in *American Mathematical Society: Lectures on Mathematics in the Life Sciences*. Amer Mathematical Society.

THOMAS, J. W., J. W. TOUCHMAN, R. W. BLAKESLEY, G. G. BOUFFARD, S. M. BECKSTROM-STERNBERG *et al.*, 2003 Comparative analyses of multi-species sequences from targeted genomic regions. Nature **424:** 788-793.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

THORNE, J. L., and H. KISHINO, 2002 Divergence Time and Evolutionary Rate Estimation with Multilocus Data. Syst Biol **51:** 689 - 702.

THORNE, J. L., H. KISHINO and I. S. PAINTER, 1998 Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol **15:** 1647-1657.

TJALLING, J. Y., 1995 Historical development of the Newton-Raphson method. SIAM Rev. **37:** 531-551.

TSUTSUI, M., M. TANIGUCHI, K. YOKOTA and T. KAWAI, 2010 Identifying single nucleotides by tunnelling current. Nat Nano **5:** 286-290.

VENABLES, W. N., and B. D. RIPLEY, 2002 *Modern Applied Statistics with S*. Springer, New York.

VENTER, J. C., M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL *et al.*, 2001 The sequence of the human genome. Science **291:** 1304-1351.

WALL, D. P., A. E. HIRSH, H. B. FRASER, J. KUMM, G. GIAEVER *et al.*, 2005 Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A **102:** 5483-5488.

WANUNU, M., W. MORRISON, Y. RABIN, A. Y. GROSBERG and A. MELLER, 2010 Electrostatic focusing of unlabelled DNA into nanoscale pores using a salt gradient. Nat Nano **5:** 160-165.

WATSON, J. D., and F. H. C. CRICK, 1953 A Structure for Deoxyribose Nucleic Acid. Nature **171:** 737-738.

WEI, X., J. M. DECKER, S. WANG, H. HUI, J. C. KAPPES *et al.*, 2003 Antibody neutralization and escape by HIV-1. Nature **422:** 307-312.

WELCH, J. J., and L. BROMHAM, 2005 Molecular dating when rates vary. Trends in Ecology & Evolution **20:** 320-327.

WHELAN, S., and N. GOLDMAN, 2001 A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Mol Biol Evol **18:** 691-699.

WHITAKER, J. W., G. A. MCCONKEY and D. R. WESTHEAD, 2009 The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. Genome Biol **10:** R36.

WOLF, Y., L. CARMEL and E. KOONIN, 2006 Unifying measures of gene function and evolution. Proc R Soc Lond B Biol Sci **273:** 1507-1515.

WOLF, Y., I. ROGOZIN, N. GRISHIN, R. TATUSOV and E. KOONIN, 2001 Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evolutionary Biology **1:** 8.

WONG, W. S. W., and R. NIELSEN, 2004 Detecting Selection in Noncoding Regions of Nucleotide Sequences

10.1534/genetics.102.010959. Genetics **167:** 949-958.

WOOLFE, A., M. GOODSON, D. K. GOODE, P. SNELL, G. K. MCEWEN *et al.*, 2005 Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol **3:** e7.

WRAY, G. A., J. S. LEVINTON and L. H. SHAPIRO, 1996 Molecular Evidence for Deep Precambrian Divergences Among Metazoan Phyla. Science **274:** 568-573.

WU, C. H., H. HUANG, A. NIKOLSKAYA, Z. HU and W. C. BARKER, 2004 The iProClass integrated database for protein functional analysis. Computational Biology and Chemistry **28:** 87-96.

WU, C. I., and W. H. LI, 1985 Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci U S A **82:** 1741-1745.

XENARIOS, I., L. SALWINSKI, X. J. DUAN, P. HIGNEY, S.-M. KIM *et al.*, 2002 DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucl. Acids Res. **30:** 303-305.

XUE, W., J. WANG, Z. SHEN and H. ZHU, 2004 Enrichment of transcriptional regulatory sites in non-coding genomic region. Bioinformatics **20:** 569-575.

YANG, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol **39:** 306-314.

YANG, Z., 1997 On the estimation of ancestral population sizes of modern humans. Genetics Research **69:** 111-116.

YANG, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol **15:** 568-573.

YANG, Z., 2006 *Computational Molecular Evolution*. Oxford University Press, New York.

YANG, Z., 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol **24:** 1586-1591.

YANG, Z., and R. NIELSEN, 2002 Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Mol Biol Evol **19:** 908-917.

YANG, Z., and B. RANNALA, 1997 Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol **14:** 717-724.

YODER, A. D., 1997 Back to the future: A synthesis of strepsirrhine systematics. Evolutionary Anthropology: Issues, News, and Reviews **6:** 11-22.

YODER, A. D., and Z. YANG, 2000 Estimation of Primate Speciation Dates Using Local Molecular Clocks. Mol Biol Evol **17:** 1081-1090.

YU, X.-J., H.-K. ZHENG, J. WANG, W. WANG and B. SU, 2006 Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics **88:** 745-751.

YULE, U., 1924 A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character **213:** 21-87.

YUSIM, K., C. KESMIR, B. GASCHEN, M. M. ADDO, M. ALTFELD *et al.*, 2002 Clustering Patterns of Cytotoxic T-Lymphocyte Epitopes in Human Immunodeficiency Virus Type 1 (HIV-1) Proteins Reveal Imprints of Immune Evasion on HIV-1 Global Variation. J. Virol. **76:** 8757-8768.

ZHOU, Y., N. RODRIGUE, N. LARTILLOT and H. PHILIPPE, 2007 Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evolutionary Biology **7:** 206.

ZMASEK, C. M., and S. R. EDDY, 2001 A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics **17:** 821-828.

ZUCKERKANDL, E., L. PAULING, V. BRYSON and H. J. VOGEL, 1965 Evolutionary divergence and convergence in proteins, pp. 97-166 in *Evolving Genes and Proteins*. Academic Press, New York.

Appendix

Appendix A. List of genes used

A.1 OrthoMam dataset in Chapters 2 and 3

Ensembl references of the genes used:

ENSG00000000460, ENSG00000001084, ENSG00000002746, ENSG00000003393, ENSG00000004487, ENSG00000004534, ENSG00000005108, ENSG00000005156, ENSG00000005187, ENSG00000005483, ENSG00000005812, ENSG00000005884, ENSG00000006114, ENSG00000007202, ENSG00000008086, ENSG00000008294, ENSG00000009335, ENSG00000009830, ENSG00000010256, ENSG00000010803, ENSG00000011021, ENSG00000011198, ENSG00000011376, ENSG00000011465, ENSG00000012504, ENSG00000012963, ENSG00000014257, ENSG00000016864, ENSG00000018280, ENSG00000021776, ENSG00000022840, ENSG00000023318, ENSG00000023909, ENSG00000025434, ENSG00000028116, ENSG00000029725, ENSG00000032389, ENSG00000033030, ENSG00000033178, ENSG00000035403, ENSG00000035687, ENSG00000036257, ENSG00000036473, ENSG00000036565, ENSG00000036828, ENSG00000037474, ENSG00000038002, ENSG00000039139, ENSG00000039537, ENSG00000040199, ENSG00000042088, ENSG00000044446, ENSG00000047188, ENSG00000047249, ENSG00000047315, ENSG00000049167, ENSG00000049883, ENSG00000051341, ENSG00000052723, ENSG00000053108, ENSG00000053900, ENSG00000054282, ENSG00000054965, ENSG00000057663, ENSG00000057704, ENSG00000060688, ENSG00000061918, ENSG00000061936, ENSG00000062725, ENSG00000063761, ENSG00000064692, ENSG00000064933, ENSG00000065060, ENSG00000065183, ENSG00000065308, ENSG00000065485,

ENSG00000065491, ENSG00000065609, ENSG00000066135, ENSG00000066422, ENSG00000067208, ENSG00000067248, ENSG00000067704, ENSG00000068793, ENSG00000069667, ENSG00000069702, ENSG00000070061, ENSG00000070718, ENSG00000070785, ENSG00000071553, ENSG00000071909, ENSG00000072121, ENSG00000072134, ENSG00000072315, ENSG00000072422, ENSG00000072682, ENSG00000073282, ENSG00000073614, ENSG00000073737, ENSG00000073792, ENSG00000074054, ENSG00000074706, ENSG00000074755, ENSG00000074771, ENSG00000075213, ENSG00000075420, ENSG00000075539, ENSG00000075568, ENSG00000075643, ENSG00000075856, ENSG00000076003, ENSG00000077063, ENSG00000077232, ENSG00000077380, ENSG00000077420, ENSG00000077514, ENSG00000077782, ENSG00000077943, ENSG00000078070, ENSG00000078114, ENSG00000078401, ENSG00000078674, ENSG00000078687, ENSG00000078725, ENSG00000079156, ENSG00000079335, ENSG00000079785, ENSG00000079931, ENSG00000079950, ENSG00000080166, ENSG00000080493, ENSG00000080561, ENSG00000080618, ENSG00000080644, ENSG00000080815, ENSG00000081019, ENSG00000081087, ENSG00000081177, ENSG00000081181, ENSG00000081479, ENSG00000081803, ENSG00000081923, ENSG00000082213, ENSG00000082482, ENSG00000082805, ENSG00000083067, ENSG00000083307, ENSG00000083782, ENSG00000084070, ENSG00000084110, ENSG00000084693, ENSG00000084754, ENSG00000085365, ENSG00000085491, ENSG00000085760, ENSG00000085840, ENSG00000086200, ENSG00000086232, ENSG00000086827, ENSG00000086848, ENSG00000087053, ENSG00000087111, ENSG00000087206, ENSG00000087253, ENSG00000087263, ENSG00000088298, ENSG00000088325, ENSG00000088538, ENSG00000088756, ENSG00000088930, ENSG00000089022, ENSG00000089091, ENSG00000089123, ENSG00000089775, ENSG00000090316, ENSG00000090376, ENSG00000090487, ENSG00000090861, ENSG00000090863, ENSG00000091138, ENSG00000091157, ENSG00000091428, ENSG00000091436, ENSG00000091482, ENSG00000091490, ENSG00000092068, ENSG00000092140, ENSG00000092208, ENSG00000092421, ENSG00000092931, ENSG00000093100, ENSG00000093144, ENSG00000093167, ENSG00000094804, ENSG00000094963, ENSG00000095319, ENSG00000095794, ENSG00000096093, ENSG00000099139, ENSG00000099219, ENSG00000099250, ENSG00000099284, ENSG00000099956, ENSG00000100201, ENSG00000100220, ENSG00000100280, ENSG00000100281, ENSG00000100296,

ENSG00000100354, ENSG00000100372, ENSG00000100473, ENSG00000100504, ENSG00000100505, ENSG00000100526, ENSG00000100578, ENSG00000100592, ENSG00000100644, ENSG00000100731, ENSG00000100814, ENSG00000100983, ENSG00000100997, ENSG00000101323, ENSG00000101333, ENSG00000101349, ENSG00000101464, ENSG00000101542, ENSG00000101782, ENSG00000101901, ENSG00000102362, ENSG00000102383, ENSG00000102384, ENSG00000102452, ENSG00000102471, ENSG00000102595, ENSG00000102763, ENSG00000102893, ENSG00000102900, ENSG00000102908, ENSG00000103044, ENSG00000103051, ENSG00000103494, ENSG00000103549, ENSG00000103569, ENSG00000103599, ENSG00000103671, ENSG00000103707, ENSG00000104067, ENSG00000104133, ENSG00000104154, ENSG00000104299, ENSG00000104313, ENSG00000104537, ENSG00000104549, ENSG00000104611, ENSG00000104723, ENSG00000105173, ENSG00000105810, ENSG00000105851, ENSG00000105856, ENSG00000105877, ENSG00000105929, ENSG00000105953, ENSG00000105976, ENSG00000106066, ENSG00000106069, ENSG00000106070, ENSG00000106100, ENSG00000106105, ENSG00000106344, ENSG00000106459, ENSG00000106524, ENSG00000106546, ENSG00000106772, ENSG00000106799, ENSG00000106829, ENSG00000106993, ENSG00000107443, ENSG00000107447, ENSG00000107518, ENSG00000107625, ENSG00000107651, ENSG00000107672, ENSG00000107679, ENSG00000107862, ENSG00000107863, ENSG00000108018, ENSG00000108039, ENSG00000108270, ENSG00000108423, ENSG00000108510, ENSG00000108576, ENSG00000108578, ENSG00000108587, ENSG00000108588, ENSG00000108666, ENSG00000108753, ENSG00000108854, ENSG00000109111, ENSG00000109184, ENSG00000109381, ENSG00000109466, ENSG00000109572, ENSG00000109670, ENSG00000109771, ENSG00000109819, ENSG00000109920, ENSG00000110395, ENSG00000110400, ENSG00000110436, ENSG00000110497, ENSG00000110514, ENSG00000110675, ENSG00000110693, ENSG00000110713, ENSG00000110871, ENSG00000111058, ENSG00000111218, ENSG00000111725, ENSG00000111799, ENSG00000111817, ENSG00000111880, ENSG00000111886, ENSG00000111913, ENSG00000112159, ENSG00000112208, ENSG00000112210, ENSG00000112242, ENSG00000112246, ENSG00000112280, ENSG00000112282, ENSG00000112319, ENSG00000112379, ENSG00000112419, ENSG00000112624, ENSG00000112664, ENSG00000112679, ENSG00000112773, ENSG00000112893, ENSG00000112902, ENSG00000112992,

ENSG00000113048, ENSG00000113194, ENSG00000113263, ENSG00000113272, ENSG00000113273, ENSG00000113282, ENSG00000113361, ENSG00000113492, ENSG00000113494, ENSG00000113578, ENSG00000113580, ENSG00000113638, ENSG00000113648, ENSG00000113657, ENSG00000113716, ENSG00000113742, ENSG00000113946, ENSG00000114054, ENSG00000114331, ENSG00000114388, ENSG00000114423, ENSG00000114439, ENSG00000114757, ENSG00000114770, ENSG00000115137, ENSG00000115159, ENSG00000115183, ENSG00000115211, ENSG00000115232, ENSG00000115239, ENSG00000115252, ENSG00000115267, ENSG00000115290, ENSG00000115295, ENSG00000115464, ENSG00000115474, ENSG00000115750, ENSG00000115761, ENSG00000115806, ENSG00000115896, ENSG00000115902, ENSG00000116044, ENSG00000116117, ENSG00000116353, ENSG00000116667, ENSG00000116668, ENSG00000116688, ENSG00000116704, ENSG00000116711, ENSG00000116745, ENSG00000116748, ENSG00000116874, ENSG00000116957, ENSG00000117000, ENSG00000117115, ENSG00000117360, ENSG00000117481, ENSG00000117528, ENSG00000117543, ENSG00000117758, ENSG00000117906, ENSG00000118058, ENSG00000118200, ENSG00000118246, ENSG00000118257, ENSG00000118407, ENSG00000118513, ENSG00000118514, ENSG00000118729, ENSG00000119042, ENSG00000119125, ENSG00000119185, ENSG00000119487, ENSG00000119537, ENSG00000119682, ENSG00000119688, ENSG00000119689, ENSG00000119844, ENSG00000119888, ENSG00000119913, ENSG00000119927, ENSG00000120262, ENSG00000120526, ENSG00000120594, ENSG00000120697, ENSG00000120708, ENSG00000120798, ENSG00000120800, ENSG00000120868, ENSG00000121031, ENSG00000121053, ENSG00000121486, ENSG00000121644, ENSG00000121940, ENSG00000122025, ENSG00000122121, ENSG00000122335, ENSG00000122507, ENSG00000122591, ENSG00000122707, ENSG00000122779, ENSG00000122870, ENSG00000122882, ENSG00000122912, ENSG00000123066, ENSG00000123191, ENSG00000123213, ENSG00000123219, ENSG00000123240, ENSG00000123473, ENSG00000123600, ENSG00000124120, ENSG00000124198, ENSG00000124201, ENSG00000124207, ENSG00000124228, ENSG00000124786, ENSG00000124818, ENSG00000125037, ENSG00000125124, ENSG00000125149, ENSG00000125255, ENSG00000125304, ENSG00000125409, ENSG00000125450, ENSG00000125630, ENSG00000125675, ENSG00000125851, ENSG00000125863, ENSG00000125885, ENSG00000126010, ENSG00000126016,

ENSG00000126822, ENSG00000126858, ENSG00000127463, ENSG00000127481, ENSG00000127688, ENSG00000127863, ENSG00000128573, ENSG00000128585, ENSG00000128708, ENSG00000129083, ENSG00000129295, ENSG00000129460, ENSG00000129493, ENSG00000129566, ENSG00000129595, ENSG00000129636, ENSG00000129675, ENSG00000129691, ENSG00000130413, ENSG00000131374, ENSG00000131459, ENSG00000131725, ENSG00000131773, ENSG00000131778, ENSG00000131979, ENSG00000132300, ENSG00000132361, ENSG00000132434, ENSG00000132437, ENSG00000132600, ENSG00000132669, ENSG00000132837, ENSG00000132842, ENSG00000132906, ENSG00000132915, ENSG00000133103, ENSG00000133104, ENSG00000133121, ENSG00000133302, ENSG00000133657, ENSG00000133800, ENSG00000133958, ENSG00000134028, ENSG00000134247, ENSG00000134255, ENSG00000134265, ENSG00000134278, ENSG00000134317, ENSG00000134363, ENSG00000134398, ENSG00000134453, ENSG00000134504, ENSG00000134508, ENSG00000134569, ENSG00000134574, ENSG00000134644, ENSG00000134769, ENSG00000134775, ENSG00000134900, ENSG00000134910, ENSG00000134982, ENSG00000134987, ENSG00000135048, ENSG00000135063, ENSG00000135318, ENSG00000135336, ENSG00000135338, ENSG00000135541, ENSG00000135720, ENSG00000135750, ENSG00000135775, ENSG00000135870, ENSG00000135905, ENSG00000135972, ENSG00000136040, ENSG00000136110, ENSG00000136141, ENSG00000136153, ENSG00000136161, ENSG00000136167, ENSG00000136169, ENSG00000136237, ENSG00000136243, ENSG00000136381, ENSG00000136404, ENSG00000136531, ENSG00000136631, ENSG00000136731, ENSG00000136811, ENSG00000136813, ENSG00000136824, ENSG00000136854, ENSG00000136936, ENSG00000136960, ENSG00000136986, ENSG00000137055, ENSG00000137073, ENSG00000137145, ENSG00000137177, ENSG00000137200, ENSG00000137275, ENSG00000137393, ENSG00000137478, ENSG00000137497, ENSG00000137563, ENSG00000137702, ENSG00000137710, ENSG00000137764, ENSG00000137872, ENSG00000137936, ENSG00000137942, ENSG00000138119, ENSG00000138193, ENSG00000138303, ENSG00000138448, ENSG00000138669, ENSG00000138741, ENSG00000138760, ENSG00000138942, ENSG00000139344, ENSG00000139436, ENSG00000139517, ENSG00000139618, ENSG00000139668, ENSG00000139719, ENSG00000139737, ENSG00000139767, ENSG00000139780, ENSG00000139921, ENSG00000140009, ENSG00000140199, ENSG00000140265,

ENSG00000140382, ENSG00000140455, ENSG00000140688, ENSG00000140694, ENSG00000140740, ENSG00000141027, ENSG00000141298, ENSG00000141325, ENSG00000141349, ENSG00000141404, ENSG00000141485, ENSG00000141627, ENSG00000141642, ENSG00000141665, ENSG00000143036, ENSG00000143093, ENSG00000143153, ENSG00000143183, ENSG00000143353, ENSG00000143493, ENSG00000143498, ENSG00000143552, ENSG00000143669, ENSG00000143751, ENSG00000143799, ENSG00000143889, ENSG00000143940, ENSG00000143970, ENSG00000143995, ENSG00000144224, ENSG00000144283, ENSG00000144290, ENSG00000144426, ENSG00000144468, ENSG00000144580, ENSG00000144635, ENSG00000144644, ENSG00000144645, ENSG00000144815, ENSG00000144843, ENSG00000145103, ENSG00000145348, ENSG00000145675, ENSG00000145730, ENSG00000145819, ENSG00000145826, ENSG00000145868, ENSG00000145996, ENSG00000146122, ENSG00000146233, ENSG00000146281, ENSG00000146409, ENSG00000146416, ENSG00000146576, ENSG00000146918, ENSG00000146963, ENSG00000147065, ENSG00000147475, ENSG00000147647, ENSG00000147649, ENSG00000147862, ENSG00000147894, ENSG00000148225, ENSG00000148498, ENSG00000149295, ENSG00000149305, ENSG00000149311, ENSG00000149499, ENSG00000149573, ENSG00000149582, ENSG00000150086, ENSG00000150394, ENSG00000150593, ENSG00000150722, ENSG00000150764, ENSG00000150961, ENSG00000151067, ENSG00000151332, ENSG00000151360, ENSG00000151413, ENSG00000151503, ENSG00000151553, ENSG00000151572, ENSG00000151617, ENSG00000151657, ENSG00000151665, ENSG00000151692, ENSG00000151694, ENSG00000151812, ENSG00000151835, ENSG00000152133, ENSG00000152217, ENSG00000152377, ENSG00000152457, ENSG00000152487, ENSG00000152503, ENSG00000152578, ENSG00000152683, ENSG00000152904, ENSG00000152942, ENSG00000153207, ENSG00000153234, ENSG00000153246, ENSG00000153294, ENSG00000153347, ENSG00000153406, ENSG00000153820, ENSG00000153982, ENSG00000153989, ENSG00000153993, ENSG00000154080, ENSG00000154162, ENSG00000154174, ENSG00000154217, ENSG00000154309, ENSG00000154310, ENSG00000154447, ENSG00000154710, ENSG00000154803, ENSG00000154889, ENSG00000155324, ENSG00000155363, ENSG00000155465, ENSG00000155827, ENSG00000155897, ENSG00000155903, ENSG00000156103, ENSG00000156113, ENSG00000156273, ENSG00000156304, ENSG00000156395, ENSG00000156463,

ENSG00000156469, ENSG00000156502, ENSG00000156531, ENSG00000156642, ENSG00000156687, ENSG00000156958, ENSG00000157107, ENSG00000157168, ENSG00000157350, ENSG00000157426, ENSG00000157470, ENSG00000157542, ENSG00000157680, ENSG00000157796, ENSG00000157851, ENSG00000158019, ENSG00000158079, ENSG00000158161, ENSG00000158258, ENSG00000158486, ENSG00000158525, ENSG00000158560, ENSG00000158636, ENSG00000158941, ENSG00000158966, ENSG00000159086, ENSG00000159167, ENSG00000159322, ENSG00000159461, ENSG00000159708, ENSG00000159921, ENSG00000160392, ENSG00000160551, ENSG00000161526, ENSG00000162129, ENSG00000162191, ENSG00000162374, ENSG00000162402, ENSG00000162623, ENSG00000162692, ENSG00000162695, ENSG00000162819, ENSG00000162869, ENSG00000162877, ENSG00000162885, ENSG00000162927, ENSG00000162929, ENSG00000163072, ENSG00000163093, ENSG00000163312, ENSG00000163328, ENSG00000163507, ENSG00000163512, ENSG00000163541, ENSG00000163576, ENSG00000163611, ENSG00000163617, ENSG00000163625, ENSG00000163669, ENSG00000163686, ENSG00000163689, ENSG00000163697, ENSG00000163728, ENSG00000163755, ENSG00000163781, ENSG00000163832, ENSG00000163932, ENSG00000163933, ENSG00000163946, ENSG00000164023, ENSG00000164099, ENSG00000164111, ENSG00000164124, ENSG00000164169, ENSG00000164176, ENSG00000164188, ENSG00000164190, ENSG00000164199, ENSG00000164209, ENSG00000164270, ENSG00000164292, ENSG00000164300, ENSG00000164303, ENSG00000164306, ENSG00000164338, ENSG00000164347, ENSG00000164398, ENSG00000164418, ENSG00000164440, ENSG00000164463, ENSG00000164483, ENSG00000164494, ENSG00000164532, ENSG00000164542, ENSG00000164597, ENSG00000164619, ENSG00000164715, ENSG00000164733, ENSG00000164761, ENSG00000164823, ENSG00000164879, ENSG00000164930, ENSG00000164938, ENSG00000164941, ENSG00000164953, ENSG00000164961, ENSG00000165046, ENSG00000165072, ENSG00000165097, ENSG00000165185, ENSG00000165194, ENSG00000165195, ENSG00000165280, ENSG00000165475, ENSG00000165490, ENSG00000165533, ENSG00000165626, ENSG00000165672, ENSG00000165832, ENSG00000165868, ENSG00000165891, ENSG00000165990, ENSG00000165997, ENSG00000166068, ENSG00000166073, ENSG00000166111, ENSG00000166123, ENSG00000166128, ENSG00000166147, ENSG00000166167, ENSG00000166224, ENSG00000166263,

ENSG00000166266, ENSG00000166352, ENSG00000166394, ENSG00000166415, ENSG00000166451, ENSG00000166507, ENSG00000166510, ENSG00000166548, ENSG00000166596, ENSG00000166689, ENSG00000166734, ENSG00000166847, ENSG00000166902, ENSG00000166997, ENSG00000167004, ENSG00000167081, ENSG00000167207, ENSG00000167210, ENSG00000167447, ENSG00000167910, ENSG00000168137, ENSG00000168143, ENSG00000168228, ENSG00000168246, ENSG00000168385, ENSG00000168389, ENSG00000168685, ENSG00000168710, ENSG00000168813, ENSG00000168883, ENSG00000168904, ENSG00000169019, ENSG00000169085, ENSG00000169118, ENSG00000169180, ENSG00000169359, ENSG00000169375, ENSG00000169398, ENSG00000169410, ENSG00000169435, ENSG00000169439, ENSG00000169499, ENSG00000169504, ENSG00000169570, ENSG00000169679, ENSG00000169744, ENSG00000169826, ENSG00000169857, ENSG00000170017, ENSG00000170390, ENSG00000170836, ENSG00000170927, ENSG00000170959, ENSG00000170962, ENSG00000171444, ENSG00000171467, ENSG00000171557, ENSG00000171723, ENSG00000171724, ENSG00000171759, ENSG00000171824, ENSG00000171914, ENSG00000172292, ENSG00000172296, ENSG00000172817, ENSG00000172869, ENSG00000172878, ENSG00000172954, ENSG00000172986, ENSG00000172995, ENSG00000173210, ENSG00000173218, ENSG00000173226, ENSG00000173406, ENSG00000173473, ENSG00000173542, ENSG00000173692, ENSG00000173744, ENSG00000173905, ENSG00000174231, ENSG00000174238, ENSG00000174953, ENSG00000175166, ENSG00000175224, ENSG00000175426, ENSG00000175662, ENSG00000175832, ENSG00000176040, ENSG00000177479, ENSG00000177791, ENSG00000178035, ENSG00000178075, ENSG00000178105, ENSG00000178467, ENSG00000179761, ENSG00000180357, ENSG00000180776, ENSG00000181090, ENSG00000181982, ENSG00000182010, ENSG00000182149, ENSG00000182179, ENSG00000182271, ENSG00000182621, ENSG00000182667, ENSG00000182732, ENSG00000182901, ENSG00000182973, ENSG00000183166, ENSG00000183662, ENSG00000183690, ENSG00000183763, ENSG00000183765, ENSG00000183833, ENSG00000184014, ENSG00000184060, ENSG00000184445, ENSG00000184454, ENSG00000184459, ENSG00000184611, ENSG00000184983, ENSG00000185238, ENSG00000185344, ENSG00000185774, ENSG00000185920, ENSG00000186031, ENSG00000186073, ENSG00000186094, ENSG00000186409, ENSG00000186532, ENSG00000187147, ENSG00000187164,

ENSG00000187240, ENSG00000187555, ENSG00000187672, ENSG00000187772, ENSG00000188001, ENSG00000188487, ENSG00000188596, ENSG00000188921, ENSG00000189091, ENSG00000196074, ENSG00000196083, ENSG00000196116, ENSG00000196305, ENSG00000196367, ENSG00000196482, ENSG00000196547, ENSG00000196549, ENSG00000196660, ENSG00000196663, ENSG00000196781, ENSG00000196792, ENSG00000197157, ENSG00000197275, ENSG00000197323, ENSG00000197324, ENSG00000197555, ENSG00000197635, ENSG00000197822, ENSG00000197893, ENSG00000197930, ENSG00000198130, ENSG00000198231, ENSG00000198265, ENSG00000198382, ENSG00000198408, ENSG00000198431, ENSG00000198513, ENSG00000198586, ENSG00000198643, ENSG00000198648, ENSG00000198650, ENSG00000198663, ENSG00000198689, ENSG00000198722, ENSG00000198752, ENSG00000198836, ENSG00000198846, ENSG00000204120, ENSG00000204217, ENSG00000204711, ENSG00000204842, ENSG00000205060, ENSG00000205268, ENSG00000206561, ENSG00000213619, ENSG00000215305

A.2 UCSC Mammalian dataset in Chapter 4

Genes that were used had to contain homologous blocks shared among all species being compared, as well as containing annotations for their Gene Ontology (GO) terms. The Genbank accession numbers of the genes used:

NM_003636, NM_004753, NM_032341, NM_018090, NM_024544, NM_000864, NM_017761, NM_203401, NM_005517, NM_004102, NM_001525, NM_013411, NM_014408, NM_012333, NM_033553, NM_002574, NM_003629, NM_147192, NM_016491, NM_152268, NM_002633, NM_001005353, NM_001037341, NM_015139, NM_015640, NM_001938, NM_020703, NM_005272, NM_000849, NM_000560, NM_001010935, NM_004980, NM_006594, NM_000862, NM_005399, NM_003528, NM_005850, NM_012113, NM_018997, NM_000396, NM_005987, NM_002964, NM_005698, NM_006912, NM_145729, NM_002241, NM_207005, NM_004106, NM_003001, NM_177398, NM_004528, NM_022716, NM_001002294, NM_033343, NM_002597, NM_130782, NM_002922, NM_002871, NM_015999, NM_016243, NM_000707, NM_014873, NM_002107,

NM_145214, NM_033445, NM_001100, NM_000740, NM_002236, NM_004040, NM_199346, NM_006449, NM_133329, NM_144949, NM_022055, NM_002954, NM_006577, NM_004161, NM_014482, NM_003096, NM_032601, NM_019885, NM_003124, NM_016058, NM_001747, NM_016079, NM_005667, NM_012275, NM_004288, NM_199204, NM_014621, NM_006357, NM_004226, NM_002491, NM_015049, NM_003352, NM_006891, NM_014617, NM_079420, NM_003284, NM_025216, NM_022915, NM_006216, NM_002601, NM_024409, NM_014521, NM_080678, NM_000861, NM_206831, NM_004162, NM_005201, NM_182935, NM_005875, NM_014240, NM_005283, NM_001295, NM_153273, NM_144499, NM_006545, NM_006407, NM_013259, NM_024638, NM_212543, NM_004547, NM_019069, NM_032638, NM_004637, NM_000539, NM_006153, NM_014245, NM_004617, NM_014445, NM_174878, NM_001038628, NM_006232, NM_001004312, NM_021101, NM_006241, NM_013261, NM_003102, NM_001024921, NM_005339, NM_198353, NM_130902, NM_020368, NM_006835, NM_004464, NM_152545, NM_004407, NM_005172, NM_014485, NM_000670, NM_020395, NM_152621, NM_016269, NM_153426, NM_002494, NM_172174, NM_014885, NM_007080, NM_001957, NM_004744, NM_006745, NM_007281, NM_004477, NM_001737, NM_004291, NM_003633, NM_001025, NM_001884, NM_173665, NM_005668, NM_003135, NM_001284, NM_004384, NM_005340, NM_020240, NM_000758, NM_014402, NM_052971, NM_198431, NM_002715, NM_001964, NM_005642, NM_020768, NM_152407, NM_024028, NM_014443, NM_001025071, NM_004045, NM_000171, NM_004270, NM_003314, NM_001445, NM_004219, NM_001034838, NM_033644, NM_004417, NM_003945, NM_015980, NM_022754, NM_001031677, NM_003052, NM_000505, NM_017838, NM_020666, NM_002752, NM_002406, NM_001012418, NM_003913, NM_006567, NM_012241, NM_006877, NM_005325, NM_003529, NM_021052, NM_080596, NM_003533, NM_022110, NM_002122, NM_014260, NM_006703, NM_001014, NM_003427, NM_004117, NM_003137, NM_003017, NM_078467, NM_006653, NM_000409, NM_018141, NM_003192, NM_018960, NM_015388, NM_014936, NM_138733, NM_001402, NM_001010844, NM_183050, NM_000735, NM_006416, NM_006813, NM_080743, NM_006581, NM_030784, NM_153453, NM_018479, NM_198392, NM_031287, NM_000838, NM_003381, NM_014161, NM_002947, NM_005738,

NM_019059, NM_007342, NM_018947, NM_024014, NM_005523, NM_017946, NM_006658, NM_012322, NM_005402, NM_031903, NM_012412, NM_182827, NM_001306, NM_002069, NM_021151, NM_012449, NM_012129, NM_004126, NM_005221, NM_024637, NM_005837, NM_002649, NM_012328, NM_006136, NM_012281, NM_005302, NM_016352, NM_020299, NM_145808, NM_020632, NM_013252, NM_145230, NM_002052, NM_001037804, NM_004430, NM_015066, NM_001394, NM_006571, NM_014682, NM_001023, NM_000756, NM_015169, NM_144650, NM_025054, NM_005648, NM_001442, NM_152565, NM_033285, NM_000989, NM_005034, NM_032041, NM_015713, NM_003506, NM_004215, NM_080651, NM_000497, NM_000498, NM_133497, NM_134428, NM_014143, NM_006570, NM_002173, NM_004432, NM_016410, NM_006914, NM_004297, NM_001827, NM_017594, NM_005384, NM_139286, NM_004697, NM_031219, NM_016322, NM_004789, NM_007209, NM_002721, NM_000976, NM_017422, NM_001002295, NM_183005, NM_001004470, NM_178815, NM_000242, NM_000399, NM_018649, NM_018055, NM_005041, NM_032562, NM_015190, NM_207012, NM_033022, NM_006926, NM_005411, NM_006926, NM_005411, NM_004329, NM_014391, NM_002729, NM_005063, NM_017902, NM_005521, NM_006562, NM_005029, NM_000936, NM_001665, NM_000990, NM_000315, NM_000728, NM_001017, NM_006512, NM_003476, NM_012153, NM_145803, NM_020929, NM_003654, NM_012456, NM_000139, NM_014502, NM_004111, NM_002696, NM_001997, NM_033036, NM_031492, NM_021173, NM_007173, NM_004771, NM_003063, NM_006006, NM_001001522, NM_000073, NM_002105, NM_006176, NM_024551, NM_002014, NM_018463, NM_020375, NM_002234, NM_014449, NM_001975, NM_006931, NM_004054, NM_015509, NM_031412, NM_018048, NM_006143, NM_006205, NM_016072, NM_004982, NM_004985, NM_001659, NM_003217, NM_000486, NM_000423, NM_014212, NM_001798, NM_006601, NM_173353, NM_000899, NM_001946, NM_004950, NM_007035, NM_000277, NM_080911, NM_057180, NM_022491, NM_014365, NM_025126, NM_004004, NM_005870, NM_001260, NM_000982, NM_007106, NM_001629, NM_203487, NM_012158, NM_080818, NM_138280, NM_175929, NM_032859, NM_001641, NM_194430, NM_006109, NM_172314, NM_138460, NM_006156, NM_001002000, NM_030631, NM_002013, NM_001663, NM_000953, NM_002806, NM_199421, NM_007374, NM_005982,

NM_145165, NM_002382, NM_004450, NM_006029, NM_130469, NM_000153, NM_021161, NM_002487, NM_018648, NM_003134, NM_133639, NM_020857, NM_003645, NM_005254, NM_004580, NM_004330, NM_016630, NM_018285, NM_018200, NM_018602, NM_004378, NM_004049, NM_014300, NM_003847, NM_001033088, NM_198243, NM_020677, NM_001424, NM_006539, NM_001012981, NM_013258, NM_005205, NM_015927, NM_004352, NM_005954, NM_012106, NM_004165, NM_004691, NM_001138, NM_006742, NM_012320, NM_016101, NM_007014, NM_017853, NM_001906, NM_018975, NM_002386, NM_006086, NM_014604, NM_003963, NM_007278, NM_000546, NM_021210, NM_144997, NM_016084, NM_003593, NM_005165, NM_001007025, NM_003885, NM_057178, NM_002985, NM_002984, NM_000978, NM_006160, NM_032339, NM_000526, NM_201434, NM_025233, NM_003734, NM_001661, NM_004527, NM_025237, NM_004160, NM_013351, NM_007225, NM_012329, NM_016077, NM_022559, NM_000626, NM_016627, NM_000891, NM_001050, NM_002522, NM_003409, NM_007163, NM_181654, NM_014177, NM_148923, NM_145173, NM_024552, NM_001031, NM_024292, NM_001033930, NM_001806, NM_001020, NM_003827, NM_006179, NM_031896, NM_014501, NM_003969, NM_080831, NM_024411, NM_000490, NM_020157, NM_080820, NM_021067, NM_153289, NM_021809, NM_021081, NM_032883, NM_024331, NM_182970, NM_006282, NM_172110, NM_005985, NM_173485, NM_003489, NM_000219, NM_001757, NM_001236, NM_006272, NM_003277, NM_007310, NM_006477, NM_013387, NM_001013440, NM_006941, NM_012323, NM_006855, NM_006953, NM_003916, NM_175859, NM_002893, NM_000284, NM_003410, NM_021242, NM_019886, NM_205856, NM_006639, NM_005296, NM_032553, NM_004085, NM_021029, NM_080390, NM_018476, NM_206915, NM_001006640, NM_198057, NM_004458, NM_178152, NM_001000, NM_004541, NM_002351, NM_006375, NM_004484, NM_001449, NM_033642, NM_004344, NM_004992, NM_014221

A.3 Yeast dataset in Chapter 4

Genes that were used had to contain homologous blocks shared among all species being compared, as well as containing annotations for their Gene Ontology (GO) terms. The Genbank locus tags of the genes used:

YAL053W, YAR007C, YBL015W, YBL091C, YBR039W, YBR056W, YBR070C, YBR110W, YBR126C, YBR162C, YBR179C, YBR198C, YCL054W, YCR017C, YDL006W, YDL031W, YDL116W, YDL126C, YDL148C, YDL166C, YDL195W, YDL215C, YDL238C, YDR021W, YDR054C, YDR072C, YDR101C, YDR176W, YDR361C, YDR443C, YDR465C, YDR484W, YDR531W, YEL037C, YER005W, YER087W, YER090W, YFR044C, YGL001C, YGL192W, YGL205W, YGL225W, YGL253W, YGR005C, YGR094W, YGR194C, YGR285C, YHL014C, YHR019C, YHR137W, YIL109C, YIR008C, YJL100W, YJR117W, YKL104C, YKR089C, YLL029W, YLR253W, YLR389C, YML021C, YML110C, YMR015C, YMR186W, YNL062C, YNL104C, YNL155W, YNL201C, YNL248C, YNL287W, YNR038W, YOL049W, YOL097C, YOL145C, YOR025W, YOR158W, YOR197W, YOR361C, YPL028W, YPL104W, YPL106C, YPL169C, YPL195W, YPL210C, YPR074C, YPR140W, YPR181C, YIL088C, YIL090W, YIL125W, YJL085W, YJL087C, YJR068W, YJR072C, YKL034W, YKL120W, YKR071C, YKR099W, YML096W, YMR041C, YMR203W, YMR277W, YNL082W, YNL123W, YNL220W, YNL313C, YNR008W

A.4 Bacterial dataset in Chapter 4

Genes that were used had to contain homologous blocks shared among all species being compared, as well as containing annotations for their Gene Ontology (GO) terms. SwissProt accession numbers of the proteins used:

P00496, P00574, P00575, P00577, P00579, P00583, P00803, P00822, P00831, P00832, P00894, P00955, P00956, P00957, P00960, P00961, P02339, P02349, P02351, P02352, P02354, P02358, P02359, P02361, P02364, P02366, P02367, P02370, P02371, P02372, P02373, P02374, P02375, P02378, P02379, P02386, P02387, P02388, P02389, P02390, P02392, P02408, P02409, P02410, P02411,

P02413, P02414, P02416, P02418, P02419, P02422, P02423, P02424, P02429, P02430, P02919, P02990, P02995, P02998, P03002, P03003, P03016, P03018, P03033, P03810, P03815, P03844, P04036, P04079, P04286, P04381, P04475, P04790, P04994, P05082, P05640, P05797, P06138, P06139, P06612, P06616, P06710, P06711, P06978, P06981, P06982, P06988, P06992, P07011, P07015, P07016, P07019, P07020, P07028, P07395, P07649, P07671, P07682, P07813, P08178, P08179, P08192, P08193, P08244, P08312, P08324, P08330, P08373, P08398, P08400, P08402, P08576, P08577, P08885, P09029, P09030, P09097, P09151, P09156, P09160, P09453, P09625, P10408, P10443, P11096, P11537, P11665, P11880, P12008, P12281, P12283, P13685, P14900, P15254, P15639, P16659, P17112, P17114, P17579, P17802, P17888, P17952, P19641, P19675, P21499, P21774, P21888, P21889, P21891, P21893, P22188, P22565, P22938, P23863, P23875, P23893, P23932, P24167, P24233, P24253, P24554, P25521, P25522, P25532, P25715, P25717, P25845, P26281, P27247, P27299, P27511, P27851, P28306, P28637, P28691, P28909, P29464, P29680, P30134, P30747, P30749, P30867, P30958, P31059, P32052, P32168, P32661, P32662, P33138, P33398, P33582, P33643, P33899, P34086, P36663, P36679, P36879, P36929, P37149, P37186, P37340, P37443, P37764, P37765, P37768, P39290, P43672, P45528, P45802, P45803, P52062, P52097, P76256, P77241, P77488, P77645, Q46920, Q47675, Q59384

A.5 OrthoMam dataset in Chapter 4

Genes that were used had to contain homologous blocks shared among all species being compared, as well as containing annotations for their Gene Ontology (GO) terms. Ensembl references of the genes used:

ENSG00000001084, ENSG00000002746, ENSG00000003393, ENSG00000004487, ENSG00000004534, ENSG00000005156, ENSG00000005187, ENSG00000005483, ENSG00000005812, ENSG00000007202, ENSG00000008086, ENSG00000008294, ENSG00000009335, ENSG00000009830, ENSG00000010256, ENSG00000011021, ENSG00000011198, ENSG00000011376, ENSG00000011465, ENSG00000012504, ENSG00000012963, ENSG00000016864, ENSG00000018280, ENSG00000021776,

ENSG00000023318, ENSG00000023909, ENSG00000025434, ENSG00000028116, ENSG00000029725, ENSG00000033030, ENSG00000033178, ENSG00000035403, ENSG00000035687, ENSG00000036257, ENSG00000036473, ENSG00000036565, ENSG00000036828, ENSG00000037474, ENSG00000038002, ENSG00000039139, ENSG00000039537, ENSG00000042088, ENSG00000044446, ENSG00000047249, ENSG00000047315, ENSG00000051341, ENSG00000053900, ENSG00000057663, ENSG00000060688, ENSG00000061918, ENSG00000061936, ENSG00000062725, ENSG00000063761, ENSG00000064933, ENSG00000065485, ENSG00000065491, ENSG00000065609, ENSG00000066135, ENSG00000066422, ENSG00000067208, ENSG00000067704, ENSG00000068793, ENSG00000069667, ENSG00000069702, ENSG00000070061, ENSG00000070718, ENSG00000070785, ENSG00000071553, ENSG00000071909, ENSG00000072134, ENSG00000072315, ENSG00000072422, ENSG00000072682, ENSG00000073282, ENSG00000073737, ENSG00000074054, ENSG00000074706, ENSG00000074755, ENSG00000074771, ENSG00000075213, ENSG00000075568, ENSG00000075643, ENSG00000075856, ENSG00000076003, ENSG00000077063, ENSG00000077232, ENSG00000077380, ENSG00000077420, ENSG00000077514, ENSG00000077782, ENSG00000077943, ENSG00000078070, ENSG00000078401, ENSG00000078725, ENSG00000079335, ENSG00000079785, ENSG00000079931, ENSG00000079950, ENSG00000080166, ENSG00000080493, ENSG00000080618, ENSG00000080644, ENSG00000081087, ENSG00000081177, ENSG00000081181, ENSG00000081479, ENSG00000081803, ENSG00000081923, ENSG00000082482, ENSG00000082805, ENSG00000083067, ENSG00000083307, ENSG00000083782, ENSG00000084070, ENSG00000084110, ENSG00000084693, ENSG00000084754, ENSG00000085491, ENSG00000085760, ENSG00000085840, ENSG00000086200, ENSG00000086232, ENSG00000086827, ENSG00000087053, ENSG00000087253, ENSG00000087263, ENSG00000088930, ENSG00000089022, ENSG00000089091, ENSG00000089775, ENSG00000090316, ENSG00000090376, ENSG00000090487, ENSG00000090861, ENSG00000091138, ENSG00000091157, ENSG00000091428, ENSG00000091436, ENSG00000092068, ENSG00000092140, ENSG00000092208, ENSG00000093100, ENSG00000093144, ENSG00000093167, ENSG00000094804, ENSG00000094963, ENSG00000095319, ENSG00000095794, ENSG00000099139, ENSG00000099250, ENSG00000099284, ENSG00000100280, ENSG00000100281, ENSG00000100372, ENSG00000100504, ENSG00000100526,

ENSG00000100592, ENSG00000100644, ENSG00000100814, ENSG00000100983, ENSG00000100997, ENSG00000101323, ENSG00000101333, ENSG00000101349, ENSG00000101464, ENSG00000101542, ENSG00000101901, ENSG00000102362, ENSG00000102383, ENSG00000102452, ENSG00000102471, ENSG00000102595, ENSG00000102893, ENSG00000102900, ENSG00000102908, ENSG00000103044, ENSG00000103051, ENSG00000103549, ENSG00000103569, ENSG00000103671, ENSG00000103707, ENSG00000104067, ENSG00000104133, ENSG00000104154, ENSG00000104299, ENSG00000104313, ENSG00000104549, ENSG00000104723, ENSG00000105173, ENSG00000105810, ENSG00000105851, ENSG00000105856, ENSG00000105877, ENSG00000105929, ENSG00000105953, ENSG00000105976, ENSG00000106066, ENSG00000106069, ENSG00000106070, ENSG00000106100, ENSG00000106105, ENSG00000106344, ENSG00000106459, ENSG00000106546, ENSG00000106799, ENSG00000106829, ENSG00000107447, ENSG00000107518, ENSG00000107651, ENSG00000107862, ENSG00000107863, ENSG00000108018, ENSG00000108039, ENSG00000108270, ENSG00000108423, ENSG00000108576, ENSG00000108578, ENSG00000108587, ENSG00000108753, ENSG00000108854, ENSG00000109111, ENSG00000109381, ENSG00000109466, ENSG00000109572, ENSG00000109670, ENSG00000109819, ENSG00000109920, ENSG00000110395, ENSG00000110400, ENSG00000110436, ENSG00000110514, ENSG00000110693, ENSG00000110713, ENSG00000110871, ENSG00000111058, ENSG00000111725, ENSG00000111799, ENSG00000111817, ENSG00000111880, ENSG00000111886, ENSG00000112159, ENSG00000112208, ENSG00000112210, ENSG00000112242, ENSG00000112246, ENSG00000112280, ENSG00000112282, ENSG00000112319, ENSG00000112379, ENSG00000112419, ENSG00000112664, ENSG00000112679, ENSG00000112893, ENSG00000112902, ENSG00000112992, ENSG00000113263, ENSG00000113272, ENSG00000113273, ENSG00000113282, ENSG00000113361, ENSG00000113494, ENSG00000113578, ENSG00000113580, ENSG00000113648, ENSG00000113657, ENSG00000113716, ENSG00000113946, ENSG00000114054, ENSG00000114388, ENSG00000114423, ENSG00000114439, ENSG00000114770, ENSG00000115137, ENSG00000115159, ENSG00000115211, ENSG00000115252, ENSG00000115267, ENSG00000115290, ENSG00000115464, ENSG00000115474, ENSG00000115750, ENSG00000115806, ENSG00000115896, ENSG00000115902, ENSG00000116353, ENSG00000116688, ENSG00000116704, ENSG00000116711,

ENSG00000116745, ENSG00000116748, ENSG00000116874, ENSG00000116957, ENSG00000117000, ENSG00000117115, ENSG00000117360, ENSG00000117528, ENSG00000117543, ENSG00000117758, ENSG00000118058, ENSG00000118246, ENSG00000118257, ENSG00000118513, ENSG00000118514, ENSG00000118729, ENSG00000119042, ENSG00000119125, ENSG00000119185, ENSG00000119537, ENSG00000119682, ENSG00000119688, ENSG00000119689, ENSG00000119844, ENSG00000119927, ENSG00000120262, ENSG00000120594, ENSG00000120697, ENSG00000120708, ENSG00000120798, ENSG00000120800, ENSG00000120868, ENSG00000121031, ENSG00000121053, ENSG00000121486, ENSG00000122025, ENSG00000122121, ENSG00000122707, ENSG00000122779, ENSG00000122870, ENSG00000122882, ENSG00000122912, ENSG00000123191, ENSG00000123213, ENSG00000123240, ENSG00000124120, ENSG00000124198, ENSG00000124201, ENSG00000124207, ENSG00000124818, ENSG00000125124, ENSG00000125255, ENSG00000125409, ENSG00000125630, ENSG00000125675, ENSG00000125851, ENSG00000125863, ENSG00000126010, ENSG00000126016, ENSG00000126822, ENSG00000126858, ENSG00000127688, ENSG00000128573, ENSG00000128585, ENSG00000128708, ENSG00000129083, ENSG00000129493, ENSG00000129566, ENSG00000129675, ENSG00000129691, ENSG00000130413, ENSG00000131374, ENSG00000131459, ENSG00000131725, ENSG00000131773, ENSG00000131979, ENSG00000132361, ENSG00000132434, ENSG00000132437, ENSG00000132600, ENSG00000132669, ENSG00000132837, ENSG00000132842, ENSG00000132906, ENSG00000132915, ENSG00000133121, ENSG00000133657, ENSG00000133800, ENSG00000134028, ENSG00000134255, ENSG00000134265, ENSG00000134278, ENSG00000134398, ENSG00000134453, ENSG00000134504, ENSG00000134508, ENSG00000134569, ENSG00000134574, ENSG00000134644, ENSG00000134769, ENSG00000134775, ENSG00000134900, ENSG00000134982, ENSG00000134987, ENSG00000135318, ENSG00000135336, ENSG00000135720, ENSG00000135750, ENSG00000135775, ENSG00000135972, ENSG00000136040, ENSG00000136141, ENSG00000136167, ENSG00000136169, ENSG00000136237, ENSG00000136243, ENSG00000136381, ENSG00000136531, ENSG00000136731, ENSG00000136824, ENSG00000136936, ENSG00000136960, ENSG00000136986, ENSG00000137177, ENSG00000137275, ENSG00000137393, ENSG00000137497, ENSG00000137563, ENSG00000137702, ENSG00000137710, ENSG00000137764, ENSG00000137872,

ENSG00000137936, ENSG00000138193, ENSG00000138303, ENSG00000138448, ENSG00000138669, ENSG00000138741, ENSG00000138760, ENSG00000139344, ENSG00000139436, ENSG00000139517, ENSG00000139618, ENSG00000139767, ENSG00000139921, ENSG00000140009, ENSG00000140199, ENSG00000140265, ENSG00000140382, ENSG00000140455, ENSG00000140694, ENSG00000140740, ENSG00000141027, ENSG00000141298, ENSG00000141349, ENSG00000141404, ENSG00000141485, ENSG00000141642, ENSG00000141665, ENSG00000143153, ENSG00000143493, ENSG00000143498, ENSG00000143669, ENSG00000143799, ENSG00000143889, ENSG00000143970, ENSG00000143995, ENSG00000144224, ENSG00000144283, ENSG00000144290, ENSG00000144635, ENSG00000144644, ENSG00000144843, ENSG00000145348, ENSG00000145675, ENSG00000145730, ENSG00000145996, ENSG00000146122, ENSG00000146233, ENSG00000146281, ENSG00000146409, ENSG00000146918, ENSG00000147065, ENSG00000147647, ENSG00000147862, ENSG00000149295, ENSG00000149305, ENSG00000149311, ENSG00000149573, ENSG00000150086, ENSG00000150394, ENSG00000150593, ENSG00000150722, ENSG00000150764, ENSG00000150961, ENSG00000151067, ENSG00000151332, ENSG00000151413, ENSG00000151503, ENSG00000151617, ENSG00000151657, ENSG00000151692, ENSG00000151694, ENSG00000151835, ENSG00000152217, ENSG00000152377, ENSG00000152503, ENSG00000152578, ENSG00000152683, ENSG00000152904, ENSG00000152942, ENSG00000153234, ENSG00000153294, ENSG00000153406, ENSG00000153982, ENSG00000153989, ENSG00000153993, ENSG00000154080, ENSG00000154162, ENSG00000154309, ENSG00000154310, ENSG00000154447, ENSG00000154710, ENSG00000154803, ENSG00000155465, ENSG00000155827, ENSG00000155897, ENSG00000155903, ENSG00000156103, ENSG00000156113, ENSG00000156273, ENSG00000156395, ENSG00000156463, ENSG00000156502, ENSG00000156531, ENSG00000156642, ENSG00000156687, ENSG00000156958, ENSG00000157350, ENSG00000157426, ENSG00000157542, ENSG00000157680, ENSG00000157851, ENSG00000158079, ENSG00000158161, ENSG00000158258, ENSG00000158486, ENSG00000158525, ENSG00000158560, ENSG00000158941, ENSG00000158966, ENSG00000159086, ENSG00000159167, ENSG00000159322, ENSG00000159461, ENSG00000159921, ENSG00000160551, ENSG00000161526, ENSG00000162402, ENSG00000162623, ENSG00000162692, ENSG00000162695, ENSG00000162877, ENSG00000162885,

ENSG00000162927, ENSG00000163072, ENSG00000163093, ENSG00000163541, ENSG00000163669, ENSG00000163781, ENSG00000163932, ENSG00000163933, ENSG00000164023, ENSG00000164099, ENSG00000164111, ENSG00000164176, ENSG00000164190, ENSG00000164199, ENSG00000164209, ENSG00000164270, ENSG00000164292, ENSG00000164300, ENSG00000164303, ENSG00000164306, ENSG00000164347, ENSG00000164398, ENSG00000164418, ENSG00000164463, ENSG00000164494, ENSG00000164532, ENSG00000164619, ENSG00000164715, ENSG00000164733, ENSG00000164761, ENSG00000164879, ENSG00000164930, ENSG00000164941, ENSG00000164953, ENSG00000164961, ENSG00000165097, ENSG00000165194, ENSG00000165195, ENSG00000165280, ENSG00000165475, ENSG00000165672, ENSG00000165832, ENSG00000165891, ENSG00000165997, ENSG00000166068, ENSG00000166073, ENSG00000166111, ENSG00000166123, ENSG00000166128, ENSG00000166147, ENSG00000166167, ENSG00000166224, ENSG00000166266, ENSG00000166394, ENSG00000166507, ENSG00000166548, ENSG00000166902, ENSG00000167004, ENSG00000167081, ENSG00000167207, ENSG00000167910, ENSG00000168137, ENSG00000168228, ENSG00000168385, ENSG00000168685, ENSG00000168710, ENSG00000168813, ENSG00000168883, ENSG00000169118, ENSG00000169180, ENSG00000169359, ENSG00000169375, ENSG00000169398, ENSG00000169410, ENSG00000169435, ENSG00000169504, ENSG00000169679, ENSG00000169744, ENSG00000169826, ENSG00000169857, ENSG00000170017, ENSG00000170390, ENSG00000170836, ENSG00000170927, ENSG00000170962, ENSG00000171444, ENSG00000171467, ENSG00000171557, ENSG00000171724, ENSG00000171759, ENSG00000171824, ENSG00000171914, ENSG00000172292, ENSG00000172296, ENSG00000172817, ENSG00000172869, ENSG00000172878, ENSG00000172954, ENSG00000172986, ENSG00000173210, ENSG00000173218, ENSG00000173226, ENSG00000173473, ENSG00000173542, ENSG00000173692, ENSG00000173744, ENSG00000174231, ENSG00000174238, ENSG00000175166, ENSG00000175426, ENSG00000175832, ENSG00000176040, ENSG00000177479, ENSG00000177791, ENSG00000178035, ENSG00000178467, ENSG00000179761, ENSG00000181090, ENSG00000182179, ENSG00000182621, ENSG00000182667, ENSG00000182732, ENSG00000182901, ENSG00000183166, ENSG00000183763, ENSG00000183765, ENSG00000183833, ENSG00000184060, ENSG00000184445, ENSG00000184611, ENSG00000184983, ENSG00000185238,

ENSG00000185344, ENSG00000185920, ENSG00000186031, ENSG00000186094, ENSG00000187164, ENSG00000187240, ENSG00000187555, ENSG00000187772, ENSG00000188487, ENSG00000189091, ENSG00000196074, ENSG00000196083, ENSG00000196116, ENSG00000196305, ENSG00000196367, ENSG00000196482, ENSG00000196547, ENSG00000196549, ENSG00000196660, ENSG00000196792, ENSG00000197157, ENSG00000197275, ENSG00000197323, ENSG00000197324, ENSG00000197555, ENSG00000197635, ENSG00000197822, ENSG00000197930, ENSG00000198130, ENSG00000198408, ENSG00000198431, ENSG00000198586, ENSG00000198643, ENSG00000198648, ENSG00000198650, ENSG00000198689, ENSG00000198722, ENSG00000198752, ENSG00000198836, ENSG00000198846, ENSG00000204217, ENSG00000204842, ENSG00000205060, ENSG00000205268, ENSG00000213619, ENSG00000215305

Appendix B. Publications

OPEN ACCESS Freely available online

PLoS one

Covariation of Branch Lengths in Phylogenies of Functionally Related Genes

Wai Lok Sibon Li[1,2], **Allen G. Rodrigo**[1,3]*

1 Bioinformatics Institute, University of Auckland, Auckland, New Zealand, **2** Department of Computer Science, University of Auckland, Auckland, New Zealand, **3** School of Biological Sciences, University of Auckland, Auckland, New Zealand

Abstract

Recent studies have shown evidence for the coevolution of functionally-related genes. This coevolution is a result of constraints to maintain functional relationships between interacting proteins. The studies have focused on the correlation in gene tree branch lengths of proteins that are directly interacting with each other. We here hypothesize that the correlation in branch lengths is not limited only to proteins that directly interact, but also to proteins that operate within the same pathway. Using generalized linear models as a basis of identifying correlation, we attempted to predict the gene ontology (GO) terms of a gene based on its gene tree branch lengths. We applied our method to a dataset consisting of proteins from ten prokaryotic species. We found that the degree of accuracy to which we could predict the function of the proteins from their gene tree varied substantially with different GO terms. In particular, our model could accurately predict genes involved in translation and certain ribosomal activities with the area of the receiver-operator curve of up to 92%. Further analysis showed that the similarity between the trees of genes labeled with similar GO terms was not limited to genes that physically interacted, but also extended to genes functioning within the same pathway. We discuss the relevance of our findings as it relates to the use of phylogenetic methods in comparative genomics.

Citation: Li WLS, Rodrigo AG (2009) Covariation of Branch Lengths in Phylogenies of Functionally Related Genes. PLoS ONE 4(12): e8487. doi:10.1371/journal.pone.0008487

Editor: Wayne Delport, University of California San Diego, United States of America

Received June 23, 2009; **Accepted** November 25, 2009; **Published** December 29, 2009

Funding: Wai Lok Sibon Li was partially funded by Biomatters during the period of research outlined in this paper. The commercial funder was not involved in any of the following aspects of the research: study design; collection, analysis, and interpretation of data; writing of the paper; or decision to submit for publication. An early revision of the manuscript was first reviewed by the funder, but no changes were made as a result of the review. The algorithm described here may be implemented in the funding company's software packages. This will pose no restrictions toward anyone interested in reproducing or building on the algorithm, as the described algorithm uses generic statistical methods that are not exclusive to the authors' study, and that have previously been published. No patents have been placed on the methodology described. In addition, all data used in this paper is publicly available. The involvement of the funder does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

Competing Interests: Wai Lok Sibon Li was partially funded by Biomatters during the period of research outlined in this paper. The commercial funder was not involved in any of the following aspects of the research: study design; collection, analysis, and interpretation of data; writing of the paper; or decision to submit for publication. An early revision of the manuscript was first reviewed by the funder, but no changes were made as a result of the review. The algorithm described here may be implemented in the funding company's software packages. This will pose no restrictions toward anyone interested in reproducing or building on the algorithm, as the described algorithm uses generic statistical methods that are not exclusive to our study and that have previously been published. No patents have been placed on the methodology described. In addition, all data used in this paper is publicly available. The involvement of the funder does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: a.rodrigo@auckland.ac.nz

Introduction

Estimating lineage-specific substitution rates and divergence dates has become an increasingly important aspect of the reconstruction of evolutionary history [1–4]. Differences in substitution rates from lineage to lineage have been attributed to variation in neutral rates of substitution, population size, generation times, and selective forces. These together are responsible for the non-ultrametric distances on a tree [5,6] and gives rise to lineage-specific variation in molecular evolutionary rates.

More recently there has been focus on the possibility of lineage-gene-specific differences in substitution rate [7,8]. The number of substitutions acquired by a protein-coding gene may increase during periods of rapid adaptive change or decrease because of strong structural or functional constraints on the coded protein. The molecular evidence for such specific selection-mediated substitutions has been the subject of much research since the pioneering paper of Messier and Stewart [9,10–14]. These selection-mediated substitutions are by definition non-neutral and therefore would not be expected to be consistent across genes or across lineages.

The proteins that genes encode do not function individually but rather within entire pathways, though this is usually ignored in models of genic evolution [15]. In fact, it is reasonable to suggest that natural selection acts on a group of genes that collectively perform a biological function. Under the presence of selection, both functional and structural constraints will be expected to cause the divergence rates of functionally-related genes to covary.

Physically interacting genes are known to co-evolve, in the sense that there are correlated rates of substitution between genes of interacting proteins [16–20]. The way proteins function as physical structures can constrain the mutations that are allowed to persist. This is particularly evident in protein domains involved in direct physical interactions with other proteins, where protein interaction may fail if mutations that change the protein structure occur at the site of interaction. Correlated substitutions that occur within a species lineage can result in similarities in substitution rates across species. In addition to this, different lineages undergo different extents of selection pressure for any given biological

function. Due to this effect of coevolution, the selection pressures applied to a function are reflected on many or all the genes involved in that function. These two effects in combination have been shown to cause the coevolution of genes [21,22].

Accordingly, there is resemblance in branch lengths in the gene trees of interacting protein coding genes [23]. Pazos and Valencia [24] were the first to use this observed pattern of coevolution across species to predict the interaction between genes. In their study, they were able to predict pairwise interaction of gene products with 79% accuracy in the dataset used [25]. Other approaches to predicting gene interactions using coevolution have also been devised that utilize methods similar to Pazos and Valencia [21,26–32].

We argue here that coevolution and similarities in substitution rates across species are not limited purely to interacting gene pairs. Our hypothesis differs from that of Fryxell's [23] in that we suggest a more general evolutionary relationship: coevolution occurs not only specifically amongst genes that interact with each other but also amongst genes that are known to be involved in the same biological function. Coevolution is partially driven by similarity in selective pressures acting on functionally related genes [33]. Also, as all genes that interact ultimately form a network in metabolic pathways, it is expected that some "contagious" correlation will extend to functionally related genes. Our argument is supported by recent studies, which show that there is correlation in patterns of evolution amongst genes involved in related biological processes [21,33–39]. In particular, recent studies by Juan et al. [21] have found patterns of coevolution across genes from the interactomes of the NADH-quinone oxidoreductase complex and the flagellar assembly machinery, though the study did not explicitly state whether or not direct physical interactions occurred between these genes.

Though our hypothesis is supported by literature in theory and results, it has been found that genes operating within the same pathway can vary in selective pressures. A study by Rausher et al. [40] and its follow up study by Lu et al. [41] have demonstrated that differing selection pressures occur between upstream and downstream genes of the anthocyanin pathway in the *Ipomoea* genus. Hence it should be noted that correlation in evolutionary rates does not necessarily occur amongst genes in all pathways.

The aim of our study was to find how the correlation in branch length varies across the different biological functions. This matter is particularly important for phylogenetic inference and studies of comparative genomics. In particular, we aimed to determine whether the similarities in gene tree branch lengths that are seen in genes that have physically interacting gene products also exist between genes that are functionally related. As a comparison to Rausher et al.'s results, we attempt to determine whether the mode of selection is common within the different pathways in our set of species. In our study, we found that there is a correlation in branches lengths of genes trees from functionally related genes that do not necessarily have physical interactions. Results show that the degree of correlation varies greatly across different biological functions. We also discuss the findings of our study towards gene choice when computing species divergences.

Materials and Methods

The aim of our study is to predict the relationship between genes that are functionally related. We hypothesize that correlation between genes can be used to infer the function when the function of some genes in a correlated set is known. The species phylogeny is used here as a basis to detect changes in substitution rate across lineages.

Visualizing Substitution Patterns amongst Genes and Lineages As a Matrix

First we consider a new scheme of visualizing variation in substitution rates amongst genes and lineages which uses a matrix of gene tree branch lengths. Consider a collection of orthologous genes from a set of species. If the true species topology is known and assumed to be the same for all genes, all the gene trees can be built with the topology constrained. This results in a set of genes trees with the same branches but optimized to have gene-specific branch lengths. We can consider a matrix, B, of dimensions $M \times N$, where M is the number of genes, and N is the number of branches on the tree, N (N is equal to $2n$-3 in an unrooted tree, where n is the number of taxa). Each entry B_{ij} of the matrix represents the length of branch j in gene tree i. It should be noted that the order of branches and genes in the matrix is arbitrary, but constant across all genes.

Matrix Transformation

The first step of our analysis procedure is to transform the branch lengths to allow for our models to take into account global species-specific effects (e.g. the faster rate of evolution on the lineages of mice and rats compared to larger longer-lived mammals such as humans). We introduce a procedure to transform within the matrix notation. The procedures described here are analogous to standard procedures used in data transformation in microarray analysis [42].

In this procedure, all zero branch lengths are replaced with the minimum non-zero value in the matrix. In the analysis of our dataset, the lower bound of zero was never reached. All values of the matrix were then log transformed. The empirical distribution of branch lengths across all genes for a particular branch tends to be significantly skewed. An example of this is shown in Figure S1, where this distribution can be seen clearly. Matrix entries are therefore log transformed to obtain values that are less skewed.

Generalized Linear Models

We use Generalized Linear Models (GLM) as a method to predict the function of a gene by its evolution pattern. A GLM is a least squares regression method that uses a link function to model the relationship between sets of independent random variables and the response variable. Binary functions can be modeled by comparing the value predicted by the GLM to cut-offs which determine whether or not the observation is predicted to be involved in the process. A range of cut-off values can be iterated through to control for different false positive and false negative error rates.

In our case, the independent random variables are from rows of the matrix B', where each variable corresponds to the normalized length of a branch for a given gene tree. The response variable was a binary variable representing whether the gene was involved in a particular biological function. Specifically, we are testing whether each gene is involved in the respective function. By using individual binary GLMs to model each biological function, each gene can be classified as being involved in multiple functions. *Probit* was used as the link function.

An advantage of using GLMs as our method of identifying correlation is that the method automatically takes into account variation within the same variable. Thus, the method will take into account any variation within a given branch across all the genes, such as effects from the natural species distances.

Dataset Compilation

We take our dataset from that used in Pazos et al. [25] which consists of amino-acid alignments of *Escherichia coli* genes against

orthologs in other prokaryotic species. Pazos et al. obtained these alignments by BLASTing [43] of the *E. coli* protein sequences against the genomes of other prokaryotic species. Pazos et al. included in the dataset the top hits that have an E-value above a chosen cut-off point.

As the number of species included increases, the number of genes that are homologous between all the species decreases. We wanted to choose a set of species that not only allowed for a reasonable number of branches in the gene trees, but also had a sufficient number of orthologous genes. We chose our species set by finding the ten species that were most frequently present in Pazos et al.'s dataset and took the gene alignments that contained all ten species (*Bacillus subtilis*, *Mesorhizobium loti*, *Caulobacter crescentus*, *Escherichia coli*, *Salmonella typhi*, *Salmonella typhimurium*, *Yersinia pestis*, *Pasteurella multocida*, *Vibrio cholerae* and *Pseudomonas aeruginosa*).

Recovering Species and Gene Tree Topologies

As the dataset consists of prokaryotes, gene tree topologies can differ from the species topology as a result of horizontal gene transfers (HGT). To filter out genes where the gene relationship may not reflect the underlying species relationships, MCMC analyses were performed using MrBayes [44]. For each of the genes, we computed two runs, each with one cold chain and three heated chains, under a mixed amino-acid model with four gamma (γ) rate categories and allowing invariable sites (i). Prior distributions of tree branch lengths and the gamma shape parameter were set to exponential distributions with $\lambda = 10$ and the starting tree was set to random. The chains were run for 1100000 steps and sampled every 200 steps, with the first 500 trees discarded.

The posterior distributions were taken and used to determine the correct relationships amongst the species. Probabilities of each tree topology from the 95% credible set of trees was taken for each gene. The probabilities of each topology for each gene were multiplied to get the joint posterior probability of each topology over all genes, assuming independence of genes. The tree with the highest joint posterior probability was chosen as the best estimate of phylogeny. The procedure here is justified by the fact that if the tree priors for each gene are assumed to be equal, and the genes are unlinked, then this calculation is monotonic with the joint posterior probability, as follows. The posterior probability of a given tree, τ, over all genes, D_i, is:

$$\begin{aligned} & P(\tau|D_1,D_2,\ldots,D_N) \\ & \propto P(D_1,D_2,\ldots,D_N|\tau)P(\tau) \\ & = \prod_{i=1}^{N} P(D_i|\tau)P(\tau) \end{aligned} \quad (1)$$

If the posterior probabilities are obtained separately for each gene then:

$$\begin{aligned} & P(\tau|D_1) \times P(\tau|D_2) \times \ldots \times P(\tau|D_N) \\ & \propto P(D_1|\tau)P(\tau) \times P(D_2|\tau)P(\tau) \times \ldots \times P(D_N|\tau)P(\tau) \\ & = \prod_{i=1}^{N} P(D_i|\tau)P(\tau)^N \end{aligned} \quad (2)$$

As can be seen, Eqn (2) is monotically (but non-linearly) proportional to Eqn (1).

When a particular topology is not found in a gene, a minimum probability is assigned, equivalent to one divided by the number of samples taken in the MCMC analysis. According to this criterion, the most probable tree topology yielded a log probability of -2289.62. In contrast, the second most probable tree had a log probability of -2814.34. The most probable species topology found from our MCMC analysis concurs with the one used in Pazos et al.'s study, which is derived from neighbor-joining trees of distances in the 16S rRNA gene.

As the issue of HGT needed to be addressed, any genes that had significant uncertainty as to whether they had the species topology were filtered from the dataset. Genes were excluded if the MrBayes analysis did not contain the species topology we found to be the most probable within the 95% credible set of trees. As a result, 222 genes out of 471 were excluded from the dataset.

Dataset Annotation

Gene Ontology (GO) [45] annotations on biological processes and molecular functions that the *E. coli* genes are involved in were obtained from the UniProt [46] and iProClass [47] databases. iProClass contains functional annotations that were electronically determined. These annotations are determined by high sequence similarity to genes of known function in other species. These annotations were used to increase the amount of annotation for our gene set, as there are insufficient annotations in *E. coli* that have been experimentally identified. Genes containing no GO annotation for known process or function were removed from our dataset. All GO terms used took into account exact synonyms for the same term. The resulting dataset contained alignments of 219 homologous genes from the 10 prokaryotic species.

For every possible combination of GO biological process and molecular function, we found the number of genes that were involved in both GO terms. We use pairings of GO process and function here as a representation of distinct biological functions. Our justification for this is that using only one of biological process or molecular function will group together genes that are not necessarily functionally related. Each gene was labeled with the process-function pairs that it is involved in. This information is later used in training and benchmarking GLMs of each function. We filtered out process-function pairs that had less than 7 genes involved because training models with a low number of positive cases can lead to biased and badly fit models [48]. An assumption made here is that the biological function of each gene is identical across the species in the alignment.

Algorithm Implementation

Our program was written in Java 1.5 and utilizes some of the functions and classes from the Phylogenetic Analysis Library (PAL) package version 1.5 [49].

Phylogenetic Analysis

Each of the gene trees were constructed by maximum likelihood with PHYML 3.0 [50]. Gene tree topologies were constrained to the species topology that we found previously. A Dayhoff + γ + i model with 8 relative substitution rate categories was used [51] Equilibrium amino-acid frequencies, proportion of invariable sites and distribution shape were estimated from sequence data of each gene.

Results

A leave-one-out test was used to benchmark the accuracy of the GLMs. We constructed GLMs, each time training the models with all but one of the genes. The trained models were applied to get a

numerical prediction of the excluded gene. This was repeated with each of the genes in the dataset. The predictions from the GLMs were converted to estimates of whether the gene is involved in a process for a range of cut-off values. This was carried out for each of the process-function pairs to obtain the overall prediction accuracy of each term pair. As measures of accuracy, false positive error rates, true positive error rates and the Receiver Operator Characteristics (ROC) area under the curve were calculated with the ROCR package in R [52].

Table 1 shows a list of process-function pairs used and the accuracy of the models as assessed by ROC area. For values of ROC area, an area of 0.5 indicates that the classifier performs randomly. In contrast, an area of 1.0 would be achieved by a perfect classifier. It can be seen from Table 1 that the ROC areas for classification appears to vary greatly across the terms. There appears to be good correlation in genes that are identified as being involved in both the GO process of ''translation'' and GO function of ''structural constituent of ribosome'', with a ROC area of 0.92 when trying to predict the function of these genes. From Figure 1a, it can be seen that the false positive rate of predicting gene involvement in this particular function was in general very low across.

The accurate prediction also extends to genes that are identified as being in other ribosomal related functions within ''translation'', with ROC areas of 0.80, 0.88 and 0.82 in ''tRNA binding'', ''rRNA binding'' and ''RNA binding'' (''RNA binding'' is a generalization of both types of RNA), respectively (Figure 1b–d). Upon closer inspection, these four ''translation'' related RNA functions contain genes that overlap, such that the genes involved in one of the functions were often involved in some of the others. This ROC area indicates low correlation between the trees of genes annotated as being involved in this process. In contrast, for a majority of the process-function pairs, correlation in gene tree branch lengths was not seen between genes identified as having the same GO terms, with the GLMs performing approximately at random.

Randomization tests were carried out to determine whether the high correlations in our processes-function pairs are statistically significant. For each pair, a null distribution of 1000 sample replicates was constructed. Each replicate was generated by randomly selecting genes in the dataset to be involved in a null biological function. The number of genes selected to be involved in the null function in each replicate is equivalent to the number of genes involved in the process-function term. A leave-one-out test

Table 1. Prediction accuracy of the GLMs for the leave-one out tests, measured by the ROC area under the curve.

GO biological process(es)	GO molecular function(s)	Number of genes	ROC area	Adjusted *p*-value
translation	structural constituent of ribosome	38	0.92	0.00
translation	rRNA binding	28	0.88	0.00
translation	RNA binding	34	0.82	0.00
translation	tRNA binding	7	0.80	0.01
translation	protein binding	22	0.69	0.03
translation	aminoacyl-tRNA ligase activity; ATP binding; ligase activity	12	0.71	0.05
regulation of transcription, DNA-dependent	protein binding	8	0.70	0.10
transport	protein binding	7	0.69	0.10
protein folding	protein binding	7	0.66	0.17
DNA replication	protein binding	8	0.67	0.19
tRNA aminoacylation for protein translation	aminoacyl-tRNA ligase activity; ATP binding; ligase activity; nucleotide binding	7	0.62	0.23
DNA repair	hydrolase activity	8	0.61	0.23
translation	nucleotide binding	15	0.59	0.23
response to DNA damage stimulus	hydrolase activity	7	0.55	0.37
transport	ATP binding	7	0.54	0.41
DNA replication	DNA binding	7	0.52	0.41
SOS response	DNA binding	7	0.49	0.52
metabolic process	transferase activity	10	0.48	0.58
regulation of transcription, DNA-dependent	RNA binding	7	0.43	0.67
metabolic process	protein binding	7	0.39	0.74
metabolic process	catalytic activity	13	0.42	0.74
DNA repair; response to DNA damage stimulus	DNA binding	11	0.41	0.74
cell cycle; cell division	nucleotide binding	8	0.38	0.74
DNA repair; response to DNA damage stimulus	ATP binding; nucleotide binding	7	0.33	0.80
transcription	DNA binding	7	0.29	0.86
transcription	protein binding	8	0.25	0.91

Different GO process terms and function terms often shared the exact same set of genes. For example the functions of ''aminoacyl-tRNA ligase activity'', ''ATP binding'' and ''ligase activity'' within the ''translation'' process have the same genes involved in them. These are grouped as a single category in the table.
doi:10.1371/journal.pone.0008487.t001

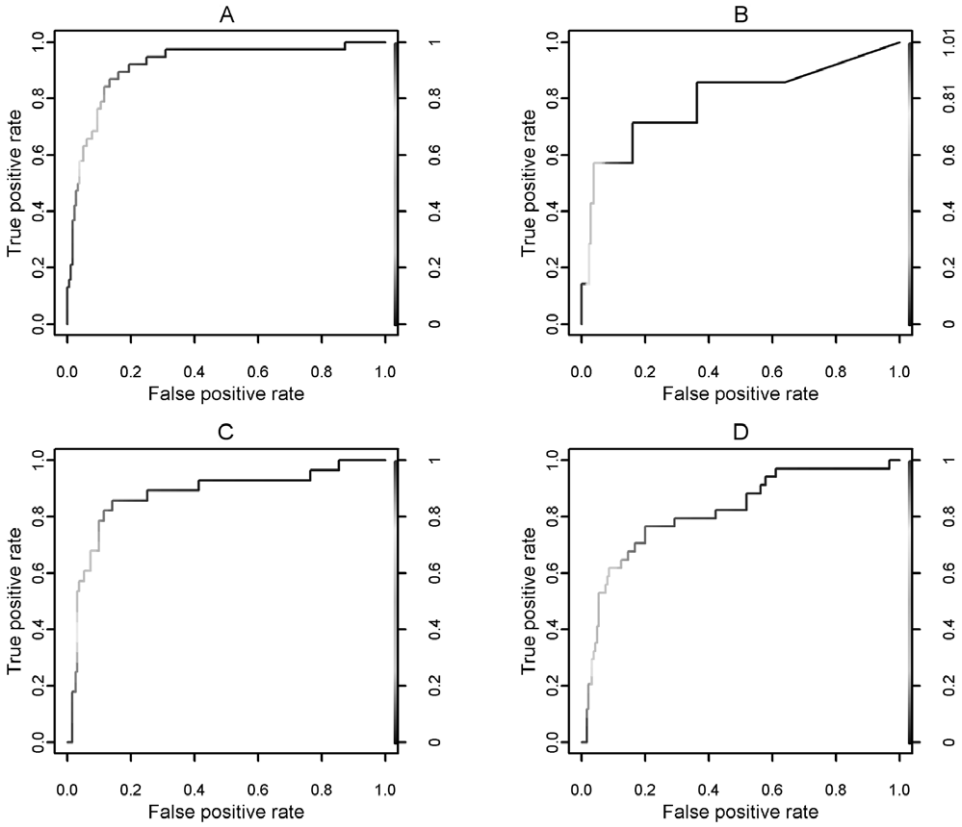


Figure 1. Plots of true positive rate against false positive rate for a few example GO process-function pairs. The predictions from the GLMs of each function were estimated using different values of the cut-off point (shown by the colored scale on the right), and error rates calculated from these predictions. (a)–(d) shows the accuracy of four related ribosomal functions within the GO process of "translation". The four GO functions are "structural constituent of ribosome", "tRNA binding", "rRNA binding" and "RNA binding", respectively.
doi:10.1371/journal.pone.0008487.g001

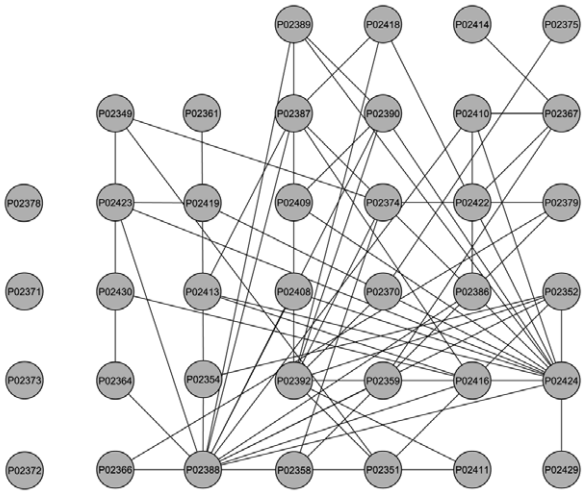
was carried out on each of the replicates and the ROC area calculated. From these randomizations the *p*-values of obtaining the actual ROC areas for each GO term combination were calculated. *p*-values were adjusted with false discovery rate correction [53] to correct for multiple comparisons (shown in Table 1). It can be seen that the correlation observed in the ribosomal functions of translation was highly significant to a 5% error rate. This indicates that the high ROC areas produced by these gene grouping were unlikely caused by sampling effects. Apart from the translation related functions, there were no other functions that were significant.

As a control, we tested whether the accuracy of the prediction was directly correlated to the number of genes that were used to train the models. It is a known issue in statistics that under-trained models with too few cases of each class produce biased and inaccurate predictions. We computed a linear fit of the number of genes involved in each process against the accuracy of each process in ROC area. The coefficient of determination (r^2) was calculated from the linear fit to be 0.39 ($p = 0.0007$). This indicates bias towards GLMs predicting for functions that have a higher number of genes involved in the function. As seen from the results in Table 1, pathways that contained fewer genes in general indicated no correlation in branch length between the genes. We would hence expect better results in some of these pathways as some of these functions become more thoroughly annotated.

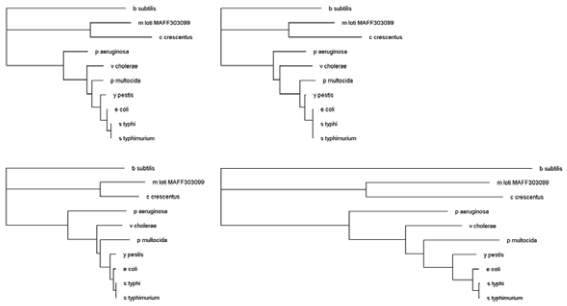
For our most significantly correlated process-function of "translation" and "structural constituent of ribosome", tests were expanded to further investigate the correlation. The size of the null distribution was increased to 10000 replicates. It is noted here that even when the replicates was increased, the *p*-value remained at 0.0, indicating that there is <0.0001 chance that the correlation seen was obtained at random. Therefore we have strong evidence to reject the null hypothesis that the correlation in gene tree phylogeny between genes labeled with GO terms "translation" and "structural constituent of ribosome" was due to random effects.

We computed physical comparisons of the characteristics of proteins involved in this process, relative to other genes and processes. We tested whether the correlation in phylogeny occurs only amongst physically interacting genes, or whether correlation extends to non-interacting genes of related function. To test this, the most significantly correlated process-function pair of "translation" and "structural constituent of ribosome" was again used. The known interactions between the genes involved in this biological function were obtained from the Database of Interacting Proteins [54]. Figure 2a shows the interaction network of the proteins in our dataset labeled with these particular GO terms. Although a large number of interactions within this pathway occur between the genes, not all the genes contain an interaction with another. In fact some of the proteins contain few interactions to any of the other proteins. Yet, the correlation in gene tree branch length amongst the proteins shown here was clearly shown in the results of the leave-one-out test. Hence, it can be seen that the correlation in phylogeny between genes is not purely limited to physically interacting genes, but the correlations also exist between functionally related genes operating within the same pathway.

A



B



C

(i) "translation" and "structural constituent of ribosome"

Coefficient variable	*B. subtilis*	*S. typhi*	*S. typhimurium*	*E. coli*	*Y. pestis*	*P. multocida*
Mean	0.30705	-0.17759	-0.01228	-0.14300	0.42118	-1.53665
Standard Error	0.00205	0.00042	0.00256	0.00032	0.00175	0.00356

(ii) Randomisations

Coefficient variable	*B. subtilis*	*S. typhi*	*S. typhimurium*	*E. coli*	*Y. pestis*	*P. multocida*
Mean	-0.08033	-0.00268	-0.00706	-0.00029	0.05142	0.00193
Standard Error	0.00016	0.00002	0.00002	0.00003	0.00009	0.00014

(i) "translation" and "structural constituent of ribosome"

V. cholerae	*P. aeruginosa*	*C. crescentus*	*M. loti*	Internal Node 0	Internal Node 1	Internal Node 2	Internal Node 3
-0.07829	0.28773	0.32179	-0.57008	-0.09312	-0.15779	-0.41797	0.00814
0.00232	0.00206	0.00225	0.00163	0.00033	0.00072	0.00149	0.00078

(ii) Randomisations

V. cholerae	*P. aeruginosa*	*C. crescentus*	*M. loti*	Internal Node 0	Internal Node 1	Internal Node 2	Internal Node 3
-0.01319	0.02744	0.02443	-0.01573	0.00752	0.01734	0.02286	0.01531
0.00012	0.00010	0.00014	0.00012	0.00003	0.00007	0.00010	0.00006

(i) "translation" and "structural constituent of ribosome"

Internal Node 4	Internal Node 5	Internal Node 6	Intercept
-0.52385	0.55831	-0.01271	-8.64613
0.00122	0.00166	0.00076	0.03355

Predicted value
0.17515
0.02155

(ii) Randomisations

Internal Node 4	Internal Node 5	Internal Node 6	Intercept
-0.00002	0.01912	0.02145	-0.77317
0.00010	0.00006	0.00007	0.00038

Predicted value
0.18029
0.00009

Figure 2. A detailed analysis of the proteins in our dataset annotated as being involved in GO process ''translation'' and GO function ''structural constituent of ribosome''. (a) The pathway interaction network of these proteins, as shown in Cytoscape [64]. Proteins P02378, P02371, P02373 and P02372 (in the first column) contain no known physical interactions to any other proteins in our list. (b) Example gene trees of proteins from our dataset. From top left to bottom right, the trees are from gene P02386, P02410 (a protein known to physically interact with P02386), P02351 (a protein that does not interact with either of the previous genes but contributes to the pathway) and the consensus of all gene trees in our dataset not labeled with these two GO terms. (c) The models built by the GLMs for (i) the proteins labeled with the two GO terms and (ii) for the 10000 randomizations of the null distribution. The end predicted value is obtained by adding the products of each coefficient and its corresponding predictor value, and the intercept value.
doi:10.1371/journal.pone.0008487.g002

Figure 2b shows an example of gene trees from proteins within this pathway. From the example it can be seen that there are similarities in branch lengths between proteins functioning within the same pathway, which is not limited to only proteins that directly interact. These similarities also show distinction from other proteins, as is seen by the dissimilarities of the gene trees to the consensus tree of proteins not involved in the pathway.

Figure 2c shows the coefficients of the GLMs from modeling the correlation from these proteins. As a comparison, the average values of each coefficient from the 10000 randomizations generated is shown. It should be noted that the coefficients here model the variation in log_e transformed branch lengths; therefore a large proportion of the predictor values will be negative, as branch lengths are generally small. From Figure 2c, we see that the coefficients from the actual process-function term itself differ greatly from that of the randomizations. This indicates that there is a distinction in the branch lengths of proteins in this pathway. As the intercept value and end predicted value differ between the two models, comparisons cannot be made.

Previous studies have found that for phylogenetic profiling [55] the number and choice of species affects how informative the profiles are [56,57]. As the underlying concept of our analysis is similar to phylogenetic profiling, this may cause a bias in our results. To test whether the high correlations seen here are biased by species choice, we repeated the leave-one-out analysis. Each time it was repeated, we simulated a single taxa removal by excluding and combining columns corresponding to the branches. With the removal of taxa, the ROC area that was produced by the GLMs of "translation" and "structural constituent of ribosome" did not alter greatly from our original result. From the 10 individual species removals, the ROC areas ranged from 0.89 to 0.94, with a mean of 0.91. Therefore, the significant correlation is unlikely an effect of bias due to choice of species used in our analysis.

Discussion

We have shown here that there are correlations between a protein's function and its gene tree branch lengths. This correlation in phylogeny is most likely attributable to the coevolution of genes that have functionally related gene products. Previous studies of inference from coevolution have focused primarily on the relationship between genes that have physically interacting gene products. We show that correlation in branch lengths extends to genes that are involved in the same functional pathway.

Hakes et al. [33] proposed the hypothesis of common selection pressures occurring on these genes to account for correlated evolutionary rates in functionally-related genes. We may also imagine that the correlations can be caused by the "contagious" propagation of mutations across the genes in the biological pathways responsible for the function. Specifically, mutations in one gene in a pathway may lead to direct compensatory mutations in a set of related genes which in turn can cause compensatory changes in other related genes, causing a cascade of mutations throughout the pathway. Alternatively, it may be that a change in the selective environment leads to changes in the selective pressure to maintain the structures of proteins involved in a given function, so that changes in substitution rates (and branch-lengths) are observed along different lineages.

In our study, we found that the correlation in branch length was particularly high in proteins involved in translation and ribosomal activities. This was most significant in proteins labeled with GO terms "translation" and "structural constituent of ribosome". We found that the overall tree lengths of these proteins are shorter than that of other proteins (average of 2.96 in ribosomal genes and 6.30 in others). This indicates that there is an overall effect across species of purifying selection acting towards the genes coding for these proteins. This is in agreement with literature stating that strong selective pressure occurs across ribosomal and translational genes [58,59]. An explanation for the purifying selection across these genes is that functions such as translation are crucial for an organism's basic function and therefore any changes to the protein sequence may cause disruption towards this essential pathway. It can also be seen that the degree to which purifying selection occurs differs across each species lineage. This is indicated by the coefficient values shown in Figure 2c, as each coefficient varies a different amount to what is expected on average.

In our analysis, uncovering correlation is limited to identifying genes that experience similar selective regimes. The assumption is that genes with functionally related proteins would undergo similar rates of evolution; yet it is possible for functionally unrelated genes to have undergone rate similarities. A subset of this effect is when trying to find correlations amongst genes from a common biological function where the genes are evolving neutrally or near neutrally. Gene trees of any other neutrally evolving genes will have similarities in branch length to gene trees of this function. This can confuse general classification and correlation methods into believing that these genes should be grouped within this function. This is noted as a limitation to our method but it will also confound any method that is based on identifying equivalent lineage-specific rates of evolution.

Despite the high significance seen in the correlation of some of the functions, a majority of the functions performed only marginally better than random. This suggests that the correlation in branch lengths is weak amongst genes annotated as being involved in those processes. The low correlation may be explained by a range of factors. Firstly, such biases in different processes are possibly due to issues within our dataset. A low number of genes involved in a function to train the model can lead to biased models. As mentioned previously, it is commonly known in statistics that a reasonable number of each case type relative to number of features is required to train accurate models [48]. The test in correlation showed that there was a significant correlation between the number of positive test cases in the processes and the ROC area. It is likely that this effect caused some bias in our study where functions with a larger number of genes involved are favored.

In addition, errors in the prediction can be caused by incorrect and incomplete functional annotations. Gene annotations in databases are often incomplete and contain errors. In particular for some biological processes such as gene translation, the specific functions of each gene involved in these processes are better known. Relevant processes will therefore have more complete and less erroneous annotations.

A second factor contributing to the discrepancy in correlation is natural variation in amount of selection pressure and gene constraint. Observed coevolution is an effect of similar selection pressures acting on functionally related proteins [33]. Where the selection pressures are weak, lesser correlations in substitution rates are expected. In cases where the compensatory mutations are crucial towards the coevolution, weaker structural constraints between genes with interacting products will result in less coevolution. Often mutations in amino-acid sequence cause no or small changes to the outcome of protein structure [60,61]. While some protein interactions necessarily require coevolution, others are known to naturally have structural flexibility and can allow for changes in interaction partners without having to make

changes to itself [60,62]. Less constraint between genes would mean that correlated mutations are often inessential thereby resulting in less similarity in substitution rates. This effect is more likely in genes where the sequence of the binding surfaces is proportionally short. In these cases mutations may not have great structural modifications to the gene and compensatory mutations may not occur. As a result similarities in branches will be less evident. In addition to this, it is possible for functionally related genes to not share common patterns of evolution. As shown by Rausher et al. [40] and Lu et al. [41], genes that produce functionally related proteins can undergo different degrees of selection, as a result of relaxed constraint on some of the genes. It is possible that in our dataset certain genes have become relaxed in one or more species. As a result, there can be a lack of correlation between such genes and other genes in the pathway it is involved in.

Another explanation for weaker correlation between genes is the definitions of function provided by GO terms. GO provides a set of text vocabularies used to categorize sequences by the general attributes of their biological function. These vocabularies cannot distinguish between different pathway organizations within the function. Hence, it is often the case that functionally unrelated genes may be annotated similarly in GO.

In addition, GO terms provide no indication towards the specificity of each term. Some function terms are very specific (for example, ''protein secretion by the type II secretion system'', ''small GTPase mediated signal transduction'') whilst others are very general (for example, ''metabolic process'', ''cell cycle'').

As part of the study, we applied the same tests to the OrthoMam dataset (version 4.0) [63]. After filtering we obtained a substantial dataset containing 730 genes that were orthologous among 24 mammalian species. Results of this analysis showed no significant correlation between any genes involved in a particular function and the gene tree branch lengths for the genes. Though the data itself is abundant, the terms that were common among the genes were uninformative. For example the most abundant processes-function pairs were: ''regulation of transcription, DNA-dependent'' with ''DNA binding'', ''signal transduction'' with ''protein binding'', and ''regulation of transcription, DNA-dependent'' with ''transcription factor activity''. These terms contain limited information on the underlying pathways themselves. It is likely that not all the genes function within the same pathway. The lack of correlation may also potentially be attributable to the distance between species (the overall tree length of this dataset was roughly 7.5 times shorter than that of the bacterial dataset) and the positive test case to negative test case ratio (the most abundant process-function pair only contained 5.9% of the 730 genes), which is known to cause under-fitting in model fitting.

Our study also suggests that when estimating divergence times, care should be taken because gene tree branch lengths may be biased by the function of the gene. Correlated changes in genes are more prominent in genes with gene products of related function; these will affect rate estimation if these genes are treated as multiple ''independent'' loci. An implication of our finding towards estimating species divergence times in comparative biology is that it is erroneous to estimate species distances using a small number of functionally-related genes. Though these effects have been to some degree recognized, they are often not considered when carrying out comparative analysis between species. A suggestion from our results is that estimating species distances should be performed using multiple loci from genes of a wide range of functions. Our findings support the suggestion made by Thorne and Kishino [15] of taking into account the correlation of genes when using multiple loci. Thorne and Kishino suggested that when estimating distances using concatenation of genes, to add parameters, models and priors which consider the correlation of substitution rates amongst the genes. Our result provides support to the use of Thorne and Kishino's techniques and as a result raises questions towards the common assumption of independence in substitution rate of gene trees.

Supporting Information

Figure S1 Histogram of the gene tree branch lengths on the *P. multocida* branch. The length of branches is approximately distributed exponentially. The lengths of other branches on the tree also follow similar distributions.
Found at: doi:10.1371/journal.pone.0008487.s001 (0.08 MB TIF)

Acknowledgments

The authors would like to thank Alexei Drummond, David Bryant, and Marc Suchard for ideas contributing to the study.

Author Contributions

Conceived and designed the experiments: WLSL AGR. Performed the experiments: WLSL. Analyzed the data: WLSL. Contributed reagents/materials/analysis tools: WLSL. Wrote the paper: WLSL AGR.

References

1. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed Phylogenetics and Dating with Confidence. PLoS Biology 4: e88.
2. Kishino H, Thorne JL, Bruno WJ (2001) Performance of a Divergence Time Estimation Method under a Probabilistic Model of Rate Evolution. Mol Biol Evol 18: 352–361.
3. Sanderson MJ (2002) Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. Mol Biol Evol 19: 101–109.
4. Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15: 1647–1657.
5. Gillespie JH (1991) The Causes of Molecular Evolution. New York: Oxford University Press.
6. Bromham L, Penny D (2003) The modern molecular clock. Nat Rev Genet 4: 216–224.
7. Yang Z, Nielsen R (2002) Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Mol Biol Evol 19: 908–917.
8. Siepel A, Pollard K, Haussler D (2006) New methods for detecting lineage-specific selection. ;DOI: 10.1007/11732990.
9. Messier W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. Nature 385: 151.
10. Ross HA, Rodrigo AG (2002) Immune-Mediated Positive Selection Drives Human Immunodeficiency Virus Type 1 Molecular Variation and Predicts Disease Duration. J Virol 76: 11715–11720.
11. Sumiyama K, Saitou N, Ueda S (2002) Adaptive Evolution of the IgA Hinge Region in Primates. Mol Biol Evol 19: 1093–1099.
12. Fay JC, Wu CI (2003) Sequence Divergence, Functional Constraint, and Selection in Protein Evolution. Annual Review of Genomics and Human Genetics 4: 213–235.
13. Sawyer SL, Wu LI, Emerman M, Malik HS (2005) Positive selection of primate TRIM5 Î± identifies a critical species-specific retroviral restriction domain. Proceedings of the National Academy of Sciences of the United States of America 102: 2832–2837.
14. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.
15. Thorne JL, Kishino H (2002) Divergence Time and Evolutionary Rate Estimation with Multilocus Data. Systematic Biology 51: 689–702.
16. Atwell S, Ultsch M, De Vos AM, Wells JA (1997) Structural Plasticity in a Remodeled Protein-Protein Interface. Science 278: 1125–1128.
17. Jucovic M, Hartley RW (1996) Protein–Protein Interaction: A Genetic Selection for Compensating Mutations at the Barnase–Barstar Interface. Proceedings of

the National Academy of Sciences of the United States of America 93: 2343–2347.
18. Pagès S, Bélaïch A, Bélaïch JP, Morag E, Lamed R, et al. (1997) Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: Prediction of specificity determinants of the dockerin domain. Proteins: Structure, Function, and Genetics 29: 517–527.
19. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. Journal of Molecular Biology 271: 511–523.
20. Poumbourios P, Maerz AL, Drummer HE (2003) Functional Evolution of the HIV-1 Envelope Glycoprotein 120 Association Site of Glycoprotein 41. J Biol Chem 278: 42149–42160.
21. Juan D, Pazos F, Valencia A (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proceedings of the National Academy of Sciences 105: 934–939.
22. Kann MG, Shoemaker BA, Panchenko AR, Przytycka TM (2009) Correlated Evolution of Interacting Proteins: Looking Behind the Mirrortree. Journal of Molecular Biology 385: 91–98.
23. Fryxell KJ (1996) The coevolution of gene family trees. Trends in Genetics 12: 364–369.
24. Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng 14: 609–614.
25. Pazos F, Ranea JAG, Juan D, Sternberg MJE (2005) Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome. Journal of Molecular Biology 352: 1002–1015.
26. Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, et al. (2003) Inferring protein interactions from phylogenetic distance matrices. Bioinformatics 19: 2039–2045.
27. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. Journal of Molecular Biology 299: 283–293.
28. Goh CS, Cohen FE (2002) Co-evolutionary Analysis Reveals Insights into Protein-Protein Interactions. Journal of Molecular Biology 324: 177–192.
29. Kim WK, Bolser DM, Park JH (2004) Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics 20: 1138–1150.
30. Ramani AK, Marcotte EM (2003) Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity. Journal of Molecular Biology 327: 273–284.
31. Sato T, Yamanishi Y, Horimoto K, Toh H, Kanehisa M (2003) Prediction of Protein-Protein Interactions from Phylogenetic Trees Using Partial Correlation Coefficient. Genome Informatics 14: 496–497.
32. Tan SH, Zhang Z, Ng SK (2004) ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. Nucl Acids Res 32: W69–72.
33. Hakes L, Lovell SC, Oliver SG, Robertson DL (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proceedings of the National Academy of Sciences 104: 7999–8004.
34. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary Rate in the Protein Interaction Network. Science 296: 750–752.
35. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV (2004) Conservation and Coevolution in the Scale-Free Human Gene Coexpression Network. Mol Biol Evol 21: 2058–2070.
36. Mariño-Ramírez L, Bodenreider O, Kantz N, Jordan IK (2006) Co-evolutionary Rates of Functionally Related Yeast Genes. Evolutionary Bioinformatics 2006: 295–300.
37. Shapiro BJ, Alm EJ (2008) Comparing Patterns of Natural Selection across Species Using Selective Signatures. PLoS Genetics 4: e23.
38. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, et al. (2005) Functional genomic analysis of the rates of protein evolution. Proceedings of the National Academy of Sciences 102: 5483–5488.
39. Wolf Y, Carmel L, Koonin E (2006) Unifying measures of gene function and evolution. Proceedings of the Royal Society B: Biological Sciences 273: 1507–1515.
40. Rausher MD, Miller RE, Tiffin P (1999) Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol 16: 266–274.
41. Lu Y, Rausher MD (2003) Evolutionary Rate Variation in Anthocyanin Pathway Genes. Mol Biol Evol 20: 1844–1853.
42. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, et al. (2007) A comparison of background correction methods for two-colour microarrays. Bioinformatics 23: 2700–2707.
43. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25: 3389–3402.
44. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29.
46. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, et al. (2004) UniProt archive. Bioinformatics 20: 3236–3237.
47. Wu CH, Huang H, Nikolskaya A, Hu Z, Barker WC (2004) The iProClass integrated database for protein functional analysis. Computational Biology and Chemistry 28: 87–96.
48. Foley DH (1972) Considerations of Sample and Feature Size. IEEE Transactions on Information Theory 18: 618–626.
49. Drummond A, Strimmer K (2001) PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics 17: 662–663.
50. Guindon S, Gascuel O (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology 52: 696–704.
51. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. Atlas of Protein Sequences and Structure 5: 345–352.
52. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940–3941.
53. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300.
54. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucl Acids Res 30: 303–305.
55. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. PNAS 96: 4285–4288.
56. Jothi R, Przytycka T, Aravind L (2007) Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinformatics 8: 173.
57. Singh S, Wall DP (2008) Testing the Accuracy of Eukaryotic Phylogenetic Profiles for Prediction of Biological Function. Evolutionary Bioinformatics 4: 217–223.
58. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, et al. (1997) The Complete Genome Sequence of Escherichia coli K-12. Science 277: 1453–1462.
59. Lecompte O, Ripp R, Thierry J-C, Moras D, Poch O (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucl Acids Res 30: 5382–5390.
60. Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. Proceedings of the National Academy of Sciences 102: 10930–10935.
61. Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E (2005) Protein structure and evolutionary history determine sequence space topology. Genome Res 15: 385–392.
62. Gillmor SA, Takeuchi T, Yang SQ, Craik CS, Fletterick RJ (2000) Compromise and accommodation in ecotin, a dimeric macromolecular inhibitor of serine proteases. Journal of Molecular Biology 299: 993–1003.
63. Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, et al. (2007) OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. BMC Evolutionary Biology 7: 241.
64. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13: 2498–2504.