A new method for identifying site-specific evolutionary rates and its applications.

A thesis submitted to the National University of Ireland for the Degree of

Doctor of Philosophy

Presented by:

Carla A. Cummins

Department of Biology,

NUI Maynooth,

Co. Kildare, Ireland.

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

October 2011

Supervisor: Dr. James O. McInerney B.Sc., Ph.D. (Galway)

Head of Dept.: Professor Kay Ohlendieck, Dip. Biol., M.Sc. (Konstanz), Ph.D.

Table of Contents

Acknowledgements

Firstly, I'd like to extend my gratitude to my supervisor, James. Thank you for all of the opportunities you provided for me at both undergraduate and postgraduate levels of my education. I have learned so much from you and had a lot of fun over my time as part of your research group. This would have been impossible without all of your advice, guidance and expertise.

My research would not have been possible without funding from Science Foundation Ireland and the excellent computing facilities provided by the NUIM HPC facility.

I wish to thank all of the members of the lab, both old and new, for all of your useful comments and answers when I needed them. I have learned a lot and I wish you all success in the future.

To all my friends, both in the lab and out, you have made this process a whole lot easier. I'd like to thank Leanne and Lahcen for all of the useful discussions (although I'm equally thankful for the useless ones) and for helping me to chill out when I needed to. Theresa and Kevin, I think I underestimated the power of awesome housemates. Thank you for never getting annoyed with me for being lazy around the house or for my abundant complaining, it really allowed me to focus on my PhD without worries. Thanks especially for all your help with the English language, particularly with commas. Thanks to Therese for the vast amounts of knowledge and Fr. Ted quotes you bestowed upon me.

Thanks to Ruth for always being there for all of the ups and downs, and for the epic volumes of laughter you have created over the years. None of this would have been the same, or even possible, without you. Sincerest thanks.

Finally, I would like to thank my family for all their help, support and love, not only during this PhD process, but throughout my entire education (and anything else life could come up with). Without all of their encouragement, I would not be where I am today and, for this, I am very grateful. I would particularly like to thank my parents for all that you have done for me, for being proud of me no matter what and, very importantly, for your impeccable genes.

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed:____________________________________

Carla A. Cummins

Abbreviations Used

DNA	DeoxyriboNucleic Acid
RNA	RiboNucleic Acid
BLAST	Basic Local Alignment Search Tool
MCL	Markov CLustering
ASRV	Among Site Rate Variation
LCA	Last Common Ancestor
ME	Minimum Evolution
LBA	Long Branch Attraction
ML	Maximum Likelihood
JC	Jukes and Cantor model
K2P	Kimura 2 Parameter model
HKY	Hasegawa, Kishino and Yano model
GTR	General Time Reversible model
PAM	Point Accepted Mutation matrices
BLOSUM	BLOcks of amino acid SUbstitution Matrix
NNI	Nearest Neighbour Interchange
SPR	Subtree Pruning and Regrafting
TBR	Tree Bisection and Reconnection
MCMC	Markov Chain Monte Carlo
LRT	Likelihood Ratio Test
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion

MRP	Matrix Representation with Parsimony
GTP	Gene Tree Parsimony
HGT	Horizontal Gene Transfer
TIGER	Tree Independent Generation of Evolutionary Rates
SACW	Successive Approximation Character Weighting
CI	Consistency Index
LQP	LeQuesne Probability
GUI	Graphic User Interface
CLI	Command Line Interface
PTP	Permutation Tail Probability
CV	Composition Vector
PPS	Posterior Predictive Simulation

Index of Figures

Chapter 1 - Introduction

Chapter 2 - TIGER: Tree Independent Generation of Evolutionary Rates

Chapter 3 - TIGER Software and Experimental Applications

Chapter 4 - Mitochondrial Origins

Index of Tables

Chapter 3 - TIGER Software and Experimental Applications

Chapter 4 - Mitochondrial Origins

Abstract

In this thesis, I discuss each stage in the development of a new method for identifying site specific evolutionary rates, from conception of the idea, through the implementation to its application to data. TIGER, or tree independent generation of evolutionary rates, is based largely around the works of LeQuesne (1989), Wilkinson (1998) and Pisani (2004) and the premise that sites in a multi-state character matrix could be scored based on the level of agreement it displays with the other sites. In these earlier studies, however, agreement was measured in binary manner: sites were either compatible with each other or they are not. TIGER allows various degrees of agreement to occur between two sites, allowing it to pick up more subtle signals in the data.

After implementing the method into a software program, it could be applied to data. Using a combination of simulated and empirical datasets, TIGER was shown to produce desirable results. In particular, removal of sites identified by TIGER was shown to improve phylogenetic reconstruction of deeply diverging lineages and of taxa displaying compositional attraction. Additionally, TIGER was applied to a gene content matrix in order to identify HGT signals and integrated into the analysis of a current phylogenetic problem, the origin of the mitochondria.

Although it is widely accepted that eukaryotes have a chimeric genome, the specific “parent” of the mitochondria is, as of yet, unclear. Previous studies have failed to reach agreement regarding this issue for a number of reasons. Exploration of the signals using TIGER and heterogeneous modelling reveal that multiple signals and compositional heterogeneity are among the biggest problems with datasets containing both mitochondrial and α-proteobacterial sequences.

Chapter 1 – Introduction

1.1: Traditional View on Evolution – Tree Thinking

Trees have been used to describe patterns of evolution for hundreds of years. Although Darwin is widely credited as the forefather of evolutionary ideas, it was Lamarck who constructed the first evolutionary tree. In *Philosophie Zoologique* (Lamarck, 1809), Lamarck provided a figure depicting his view on the origins of various animals (Figure 1.1). Darwin popularised the idea of evolution and how all life is related through common ancestry in *The Origin of the Species* (Darwin, 1859). It was Ernst Haeckel, however, who coined the term *phylogenetics* (Haeckel, 1866) thus creating a novel area of science that is well studied today. This is the basis for the use of trees (or tree-like structures) in depicting evolutionary relationships today, and the search for a single tree depicting relationships between every organism on the planet is still very much in progress.

The great diversity in the physical appearance of eukaryotes, particularly plants and animals, provides copious morphological characters on which to base inferences. This means that even before the advent of molecular data (DNA, RNA and protein sequences), phylogenies of these organisms could be constructed (Haeckel, 1866, Snodgrass, 1938). It has never been so simple for prokaryotes. Bacteria lack morphological complexity, so resolution in the bacterial phylogeny did not come until molecular phylogenetics was employed. Even then, the molecules showed that, while eukaryotes appeared to inherit genes in a largely vertical, tree-like fashion, many bacterial genes did not (discussed in further detail in section 1.5). It has become increasingly clear that the use of a tree to describe prokaryotic relationships may not be as accurate or provide as much information regarding the evolution of prokaryotes as network analyses (Fitzpatrick et al., 2006, Halary et al., 2010, Alvarez-Ponce and McInerney, 2011).

1.2: From Species to Tree

1.2.1: Data

As with all sciences, inferred hypotheses, in this case trees, must be based on an observable datum. Morphological data were used to infer early phylogenetic relationships; that is, the physical form and structure of an organism was the basis for classification, but a revolutionary paper in 1965 changed the face of phylogenetic inference. Zuckerkandl and Pauling suggested that molecular data may be used to understand evolutionary processes. They offered the opinion that the best evidence for inferring historical events might lie in the analysis of the macromolecules found in contemporary organisms. They then showed that the relative recentness of common ancestry of a group of animals (as judged against the fossil record) was in good agreement with the relative similarities of some proteins found in those animals (Zuckerkandl and Pauling, 1965). The use of molecular data such as DNA, RNA and protein sequences gave systematics a whole new lease of life (Woese and Fox, 1977, Fitch and Margoliash, 1967, Hasegawa et al., 1991). With current technology, complete genome sequencing can be carried out in as little as two hours and today there are many high-profile genome sequencing projects whose purpose, in part at

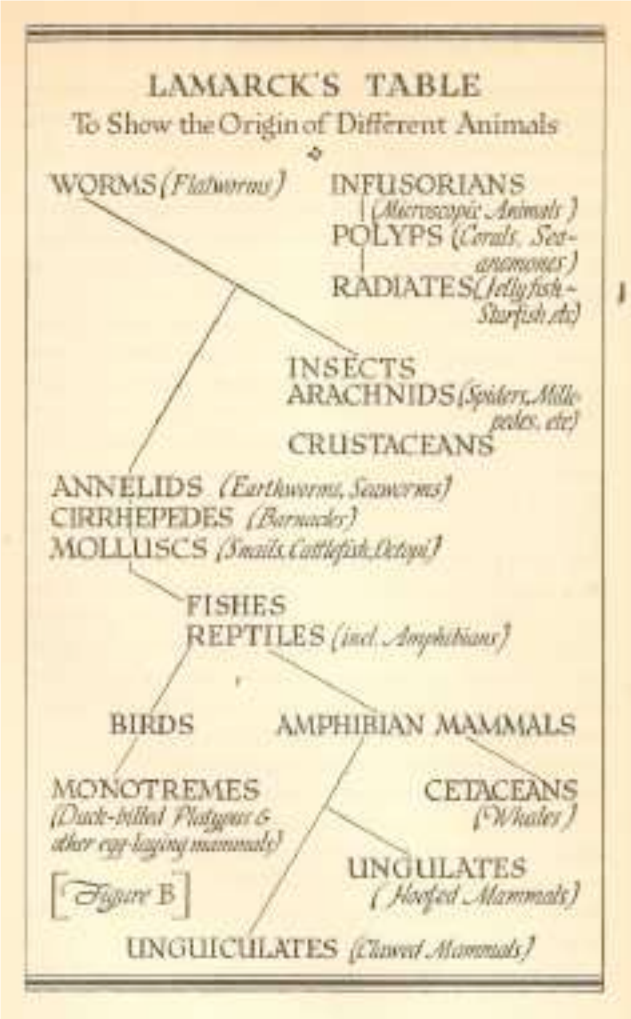
LAMARCK'S TABLE
To Show the Origin of Different Animals
WORMS (Flatworms)
INFUSORIANS (Microscopic Animals)
POLYPS (Corals, Sea-anemones)
RADIATES (Jellyfish - Starfish etc)
INSECTS
ARACHNIDS (Spiders, Millepedes, etc)
CRUSTACEANS
ANNELIDS (Earthworms, Seaworms)
CIRRHEPEDES (Barnacles)
MOLLUSCS (Snails, Cuttlefish, Octopi)
FISHES
REPTILES (incl. Amphibians)
BIRDS
AMPHIBIAN MAMMALS
MONOTREMES (Duck-billed Platypus & other egg-laying mammals)
CETACEANS (Whales)
UNGULATES (Hoofed Mammals)
[Figure B]
UNGUICULATES (Clawed Mammals)

Figure 1.1: Lamarck's view on evolution.

least, is to understand the evolution of the species in which the genomes are found (e.g. Genome 10K http://genome10k.soe.ucsc.edu/).

1.2.2: Homology and Alignment

In 1843, Richard Owen defined homology as "the same organ in different animals under every variety of form and function" (Owen, 1843). This definition holds true, not only for organs, but for proteins too, giving rise to the first step of inferring phylogenetic relationships: detection of homologous gene families. In phylogeny, relationships may only be inferred from homologs, therefore the goal is to identify genes (or proteins) that have diverged from a common ancestor. As all DNA sequences are made up of the same four bases (A, C, G and T), all sequences display some similarity, therefore it is important to be able to identify which genes are similar to each other due to common ancestry rather than just by chance. The most widely used method for detecting homology between sequences is by using BLAST (Basic Local Alignment Search Tool), (Altschul et al., 1997). This uses a sliding window approach to return a value known as the e-value (or expect value). The e-value represents the expectation that the sequences being compared are not related, meaning a lower e-value equals an increased expectation that the sequences are related. Many methods for detecting homologous gene families require BLAST as input. Markov Clustering (MCL) (Enright et al., 2002), for example, creates an undirected graph based on the BLAST hits and produces clusters which may be interpreted as gene families. It is important to note that there are a number of homology subtypes:

- Orthology: homology in different species, usually due to a speciation event.

- Paralogy: homology due to a duplication event. Paralogs may occur within the same species or in different species as with orthology. A subset of paralogy where homology is due to a whole genome duplication, however, is often referred to as ohnology (Wolfe, 2000).
- Xenology: homology due to a horizontal gene transfer.

For the construction of a gene tree (a tree depicting the evolutionary history of a single gene), homologs for that gene in the species of interest must be aligned. Sequence alignment is the act of arranging the sequences so that regions of the sequences that display homology are lined up. This is achieved by inserting gaps (denoted by '-') into the sequences so that homologous characters are aligned in the same column. Many algorithms and softwares exists to perform this step, CLUSTAL, Muscle, PRANK and FSA, to name a few (Thompson et al., 1994, Edgar, 2004, Löytynoja and Goldman, 2008, Bradley et al., 2009). Without this step, we are not comparing residues of the same evolutionary history, thus nullifying all subsequent steps. Accurate alignment is crucial.

1.2.3: Among Site Rate Variation

From the analysis of thousands of gene families, we have observed a number of features of genetic alignments, one of which is amongst site rate variation (ASRV). Different homologous characters can evolve at different rates. Some characters, or sites, may evolve very quickly and be saturated for substitution (i.e. during the evolution of a character, substitutions have been superimposed on top of each other). Others may be constant and unchanging. Since the late 1980's, it has been known that failure to accommodate this variation may cause problems and misestimations in tree inference (Olsen, 1987). Using a model that accounts for ASRV, however, can improve this issue (Yang, 1993). A discrete approximation to the gamma distribution is now widely used to model the ASRV during tree construction (Yang, 1993) and sites are divided into different categories based on their perceived rate. In this case, the rate of evolution is defined as the amount of disagreement a site has with a given tree, the ML tree. Sites that display disagreement with the tree are deemed rapidly-evolving and sites that agree with the tree are slowly-evolving. The sites are split into categories based on their rate and each category is assigned different models of evolution.

Although there are means of accounting for ASRV, oftentimes the range of variation is too great to be adequately described by the model. Often in these cases, sites with extreme rates (i.e. very fast-evolving or very slow-evolving) are removed to narrow the spectrum of variation (Hirt et al., 1999). ASRV will be discussed in further detail later in this thesis (Chapter 2).

1.2.4: Phylogenetic Trees

Over the years, many methods of constructing the evolutionary history of a set of data have been developed. These range from relatively simple neighbour-joining (Saitou and Nei, 1987), for example, to quite complex and parameter rich maximum likelihood (Felsenstein, 1981). However, before we can understand the results of inferring such phylogenetic trees, we must understand the structure and terminology of this field.

Phylogenetic trees are branching mathematical structures that represent the evolutionary relationships between a set of species or sequences, which, in this context, are generally called taxa (plural) or taxon (singular). A phylogenetic tree may be viewed as an acyclic, directed graph. Trees consist of branches and nodes, and the specific combination in which they are arranged is known as a topology (Figure 1.2). There are three types of node: internal, external (usually referred to as a “leaf node”) and root nodes. External, hereafter leaf, nodes are so named because they exist at the extremities of the tree structure. They represent the extant taxa for which the evolutionary relationships are being inferred. Internal nodes represent the last common ancestor (LCA) of all leaf nodes occurring in it’s subtree. A root node is not present in all trees; those without are called unrooted trees. A root gives a tree direction as it represents the LCA for all taxa in the tree, meaning that it arose further back in time than any other node in the tree. A root node is, essentially, a special case internal node. A subtree is, as the name suggests, a subset of the tree as a whole. Given any internal node, a subtree may be pruned from the complete tree producing a smaller tree with the internal node as the root. Branches, much like edges in a graph, connect nodes. They are directed because they, in a way, represent the passage of time. The length of a branch corresponds to the rate at which a sequence is changing; a long branch length means a lot of substitution occurred and a short branch means that little change happened.

1.3 Tree inference

Reconstructing the relationships between taxa is not as simple as we would like. Although there are many methods, ranging in complexity, for inferring a phylogenetic

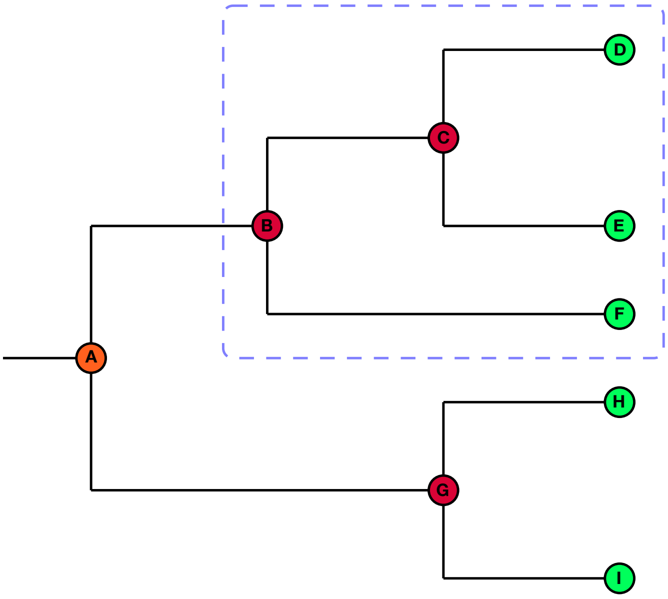

Figure 1.2: Example of a tree. Green, red and orange nodes represent leaf (external), internal and root nodes respectively. All green nodes represent extant species for which the sequence is available. The blue box represents a subtree. Within this subtree node B would be the root node. The branch between nodes A and B is shorter than that between nodes A and G, denoting less evolution/change in sequences from A to B than A to G. Node C represents the last common ancestor of leaf nodes D and E.

tree, none of them are guaranteed to produce the “true” tree. In this section I discuss the main methods used for phylogenetic inference along with their benefits and potential pitfalls.

1.3.1: Distance methods

Using distance matrix approaches for phylogenetic inference was first introduced in 1967 (Cavalli-Sforza and Edwards, 1967, Fitch and Margoliash, 1967). The most basic distance measure is called the p-distance and is a simple count of the number of sites that are not the same for each pair of sequences. The p-distance is the number of nucleotide positions that differ between the two sequences in question.

Pairwise comparison of all sequences in a dataset results in a matrix of distances and the tree that most accurately reflects these distances is the one that is chosen. This, however, provides a very over-simplified view on evolution. It does not take subtle evolutionary events into account; multiple substitutions per site, transition/ transversion ratios and compositional bias are ignored, to name a few. This is where parametric models of evolution come into play. Using models allows an extra layer of complexity to be added to tree inference and they are the basis of all probabilistic inference. These will be discussed in more detail further on in this thesis.

1.3.2: Neighbour Joining

Neighbour joining is a method based on distances, created by Saitou and Nei (Saitou and Nei, 1987). It is a method for iteratively clustering taxa based on the distance between them. The algorithm aims to minimise branch lengths and, for this reason, is similar in nature to the minimum evolution (ME) method (Cavalli-Sforza and Edwards, 1967). It is widely accepted that neighbour joining is an approximation of ME.

As the name suggests, neighbour joining is based on the concept of “neighbours”, i.e. a pair of taxa that are closer to each other than either is to any other taxon in the dataset. The algorithm for the clustering of these neighbours is as follows:

1. Begin with a completely unresolved tree (star topology)
2. Calculate the Q matrix which summarises how close each taxon is to every other one.
3. Cluster the taxa such that of all possible pairs, the pair with the smallest Q value is chosen.
4. These neighbours are then treated as a single unit and step 2 is repeated until the tree is completely resolved.

Although this method is overly-simplistic, it is extremely time efficient and is often used as a “quick and dirty” approximation of the true topology.

1.3.3: Maximum parsimony

The notion of Occam’s razor is the main concept underlying parsimony, which, in simple terms, dictates that the explanation requiring fewest assumptions is the preferred one. This idea was first applied to systematics in the 1960s by various scientists (Edwards and Sforza, 1963, Camin and Sokal, 1965) and maximum parsimony quickly became the method of choice for inferring phylogenies (Stewart, 1993).

In the context of phylogenetic trees, parsimony favours those reconstructions that postulate the fewest number of changes on the tree. Given all possible tree topologies inferable from a given dataset, the tree with the smallest number of changes is deemed to be the correct one. Therefore, due to its nature, parsimony does not account for possible ASRV. Highly heterogeneous substitution rates in neighbouring branches may cause an effect known as long branch attraction (LBA) (Felsenstein, 1978); a phenomenon wherein two (or more) long branches are erroneously inferred to be sister lineages due to the fact that both lineages have long unbroken branches where superimposed substitutions are not detected by the method . Initially, LBA was described in parsimony analyses, but it is now known that, potentially all methods may suffer from this problem. It is now considered a problem of model misspecification. This will be discussed later in the thesis (section 1.5.1).

Because of it's pitfalls, maximum parsimony is now an approach that is less popular than the probabilistic methods of inference (Maximum Likelihood (ML) and Bayesian, described later) for most types of data; however, it remains the most popular method for the analysis of morphological data.

1.3.4: Models of Evolution

All tree inference methods mentioned so far have been non-parametric, meaning they don't rely on parameters of a model of sequence evolution. The more accurate and sophisticated methods of inference (ML and Bayesian) make use of a substitution model. These substitution models represent the probability of one character state changing to another and are essential for calculating the likelihood of a tree given a set of sequences or for estimating the number of substitutions that have occurred since a pair of sequences last shared a common ancestor. The characters can be nucleotides (JC, K2P, HKY, GTR), amino acids (PAM, BLOSUM), codons (Goldman and Yang, 1994) or morphological (Lewis, 2001).

The simplest model was proposed in 1969 by Jukes and Cantor (Jukes and Cantor, 1969). This model assumes equal probabilities for all possible character state changes and an equal base composition. This level of simplicity, however, rarely reflects the observed data and there have been many attempts to elaborate on this model, to make it more realistic. The Kimura 2 parameter (K2P) model added an extra layer of complexity to JC by allowing the transition and transversion substitution rare to vary (Kimura, 1980). In 1985, the HKY model was developed. It not only allowed a varying transition/transversion ratio, but also allowed unequal base frequencies; that is, the observed frequencies of each character are incorporated into the calculation of the distance between sequences (Hasegawa et al., 1985). The most notable improvement, however, came with the development of the general time-reversible (GTR) model that allows a different rate of change from all character states to all other character states. The GTR model consists of a base frequency parameter, as in the HKY model, and a rate matrix Q.

$$Q = \begin{pmatrix} * & R_{AC}\pi_c & R_{AG}\pi_G & R_{AT}\pi_T \\ R_{AC}\pi_A & * & R_{CG}\pi_G & R_{CT}\pi_T \\ R_{AG}\pi_A & R_{CG}\pi_C & * & R_{GT}\pi_T \\ R_{AT}\pi_A & R_{CT}\pi_C & R_{GT}\pi_G & * \end{pmatrix} \qquad [1.2]$$

where R_{ij} is the rate at which base *i* changes to base *j* and π_i is the frequency at which base *i* occurs. The matrix is symmetric, meaning the model is completely time reversible, but as all previous models have simply been special cases of the GTR model, this holds true for many models.

A number of models to describe protein evolution also exist and are similar in structure to those described, the major difference being the larger number of character states and thus larger matrices. Often, one model is not adequate to describe an entire dataset, so multiple models (known as heterogeneous models) may be applied to a single dataset during analysis. These heterogeneous models can fall into three forms:

1. Lineage heterogeneous: many different models may be applied to a single dataset with each subtree effectively "choosing" the model that best fits (Foster, 2004). This allows for diverse lineages to be adequately described by their models during a single analysis.

2. Site heterogeneous: as sites evolve at different rates, they may each be described by a different model. The CAT model (Lartillot and Philippe, 2004) takes this into account. Any number of models from 1 to *n* (where *n* is the total number of sites) may be applied to the data to account for compositional variations between sites.

3. Lineage heterogeneous and site heterogeneous methods can be combined.

1.3.5: Maximum Likelihood

Maximum likelihood is a method that revolves around maximise the likelihood function for a set of parameters given a set of fixed data (usually an alignment of morphological or molecular data). The parameters are a tree and a model of sequence evolution and to maximise the likelihood function one must be fixed. For example, when searching tree space, the model of sequence evolution is fixed and the tree is free to vary while maximising the likelihood. The likelihood function:

$$L(\tau,\theta) = \Pr(d \mid \tau,\theta) \qquad [1.3]$$

states that the likelihood of a tree (τ) and a substitution model (θ) is the probability of observing the data alignment (*d*) given that tree and model. This equation allows us to make a statement about how well the tree describes how the alignment may have arisen given a substitution model, or vice versa. This, however, can only be calculated if all parts of the model are known. For this reason, the model must be optimised to values that maximise the probability of observing the data.

Maximising the likelihood on a single tree is not a difficult task. It involves the use of an optimisation procedure to select between alternative branch lengths under the substitution model in order to maximise the likelihood function. Searching through all possible trees for a given dataset, however, can become computationally intractable. For 5 taxa, there are 15 possible unrooted trees. A small jump to 10 taxa results in an exponential leap in possible trees to 2,027,025. This means that it is virtually impossible to calculate the likelihood on all trees for any reasonable number of taxa.

Using heuristics, the search space may be greatly reduced. From a start tree, the tree is progressively changed to maximise the likelihood function. A large number of tree rearrangement options are available. To name a few:

1. Nearest neighbour interchange (NNI) is a simple exchange of two neighbouring branches.
2. Subtree pruning and regrafting (SPR) involves cutting a subtree from the existing tree and reattaching it at a different location in the tree.
3. Tree-bisection and reconnection (TBR) involves cutting the tree into two pieces at a given branch and reattaching the pieces at the internal branch that maximises the function in question.

Other strategies include hitch-hiking (Charleston, 2001) and simulated annealing (Stamatakis, 2005). Once the tree is changed, the likelihood is calculated. If the likelihood of the altered tree is better than that of the unchanged tree, it is accepted as the current best tree, otherwise we go back to the previous tree. This is repeated until all variations of the tree topology yield a poorer likelihood score (ie. the current tree is better than any perturbed version of itself). Unfortunately, it is very easy to get stuck in a local maximum rather than the global maximum (Figure 1.3).

It is important to note how pivotal the role of the substitution model is to the likelihood calculation. This means that an accurate model of evolution is essential to get the correct tree. For this reason, a number of approaches to model selection are available. This will be discussed in further detail in section 1.3.7. Maximum likelihood remains a popular method as it has been shown to be robust to systematic errors and, given a good model and an adequate amount of data, it will find the true tree (Whelan et al., 2001). Many software programs are available for ML

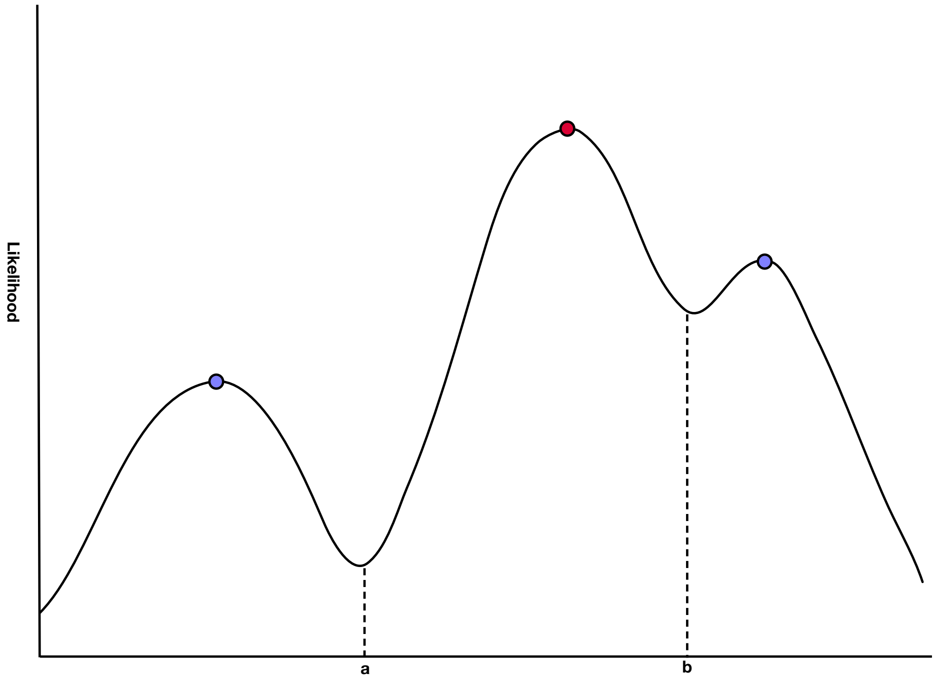

Figure 1.3: Example likelihood distribution. Local maxima are marked with blue dots, while the global maximum is denoted by a red dot. As the likelihood heuristic search does not allow steps down in likelihood, only searches that begin between point a and point b will find the best tree (i.e. the global maximum).

inference (Schmidt et al., 2002, Wilgenbusch and Swofford, 2003, Stamatakis et al., 2005, Guindon and Gascuel, 2003).

1.3.6: Bayesian Inference

Bayesian inference as a mathematical concept was formulated by Thomas Bayes, a British mathematician in the 18th century. His theorem encompasses an idea known as reverse probability. In traditional forward probability, we look at the probability of a certain outcome given a condition. A good example of this is the coin tossing experiment. In phylogenetics, the likelihood is defined as the probability of observing data given a tree and a substitution model. The Bayes theorem seeks the inverse. It seeks to calculate the probability of a model (tree + substitution model) given the data and a prior (see below). The theorem states:

$$\Pr(\theta \mid D) = \frac{\Pr(\theta) \times \Pr(D \mid \theta)}{\Pr(D)} \qquad [1.4]$$

where θ is a model and D is the data. This is the posterior probability and can be considered to be the probability that the tree is “true” (Huelsenbeck et al., 2001). This, however, is impossible to calculate without providing prior knowledge about the possible values of θ. This “knowledge” is known as the prior probability distribution or, simply, prior. The prior can be one of a number of distributions (uniform, exponential, gamma) and provides the method with additional information about the probability of observing the model θ, or Pr(θ). Therefore, the probability of a model given the observed data is a combination of the prior (Pr(θ)) and the likelihood function (Pr(D| θ)) with the probability of the data (Pr(D)) as a normalising factor. As was the case with likelihood, it is impossible to sample all possible trees for most datasets. The denominator of this equation becomes impossible to calculate with any moderately sized dataset, but the introduction of a tree sampling method known as Markov Chain Monte Carlo (MCMC) overcomes this issue. The most commonly used MCMC method is the Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970). The steps are as follows:

1. Choose an arbitrary starting tree.

2. Make a random change to to the current tree.

3. Calculate the ratio (r) of the posterior probabilities between the two trees from the new tree to the old tree. This may have two outcomes:

    (a) $r > 1$: the new tree is more probable than the old one; new tree is accepted.

    (b) $r < 1$: the new tree is less probable than the old one; the new tree is either accepted (with a probability of r) or rejected.

4. Repeat from step 2 with updated current tree.

This method is effective for a number of reasons. Firstly, the use of ratios between the posteriors for both trees allows denominators to be cancelled out, making the equation computationally tractable. As the acceptance rate for trees with a lower posterior are accepted with the probability r, small steps down in probability are accepted at quite a high rate allowing the MCMC chain to get out of local maxima. This means Bayesian methods are less likely to get stuck in local maxima and are more likely to produce a set of good trees.

The MCMC procedure can continue indefinitely. For this reason, a single MCMC run or "chain" provides little information, so multiple chains are run concurrently. The analysis ends when the chains converge on the same answer. Several software programs implement Bayesian inference for phylogenetics including MrBayes (Huelsenbeck and Ronquist, 2001), BEAST (Drummond and Rambaut, 2007) and Phylobayes (Lartillot et al., 2009).

1.3.7: Model Testing

Phylogenetic analyses are conditional on the model. Model choice is, therefore, a very important part of any analysis. Models that do not adequately describe the data can often lead to incorrect inferences (Gaut and Lewis, 1995, Sullivan and Swofford, 1997, Foster, 2004, Cox et al., 2008). Two categories of model testing can be used:

1. Model choice: allows the user to pick the best model relative to all models tested.
2. Model adequacy: tests whether the model describes the data.

A number of statistical criteria are available to facilitate model choice. The likelihood ratio test (LRT) is used to compare two models, a null and alternative hypothesis. As the name suggests, the ratio of the likelihoods of the models in question is used to calculate a p-value, which allows acceptance or rejection of the null hypothesis. Unfortunately, the LRT can only be used to test nested models (models whose parameters are subsets of one another) against each other and is, therefore, limited in its uses (Keane et al., 2006).

The Akaike information criterion (AIC) allows comparison of non-nested models and is one of the most popular means for model choice (Akaike, 1973). As with the LRT, it does not reveal any information regarding the overall adequacy of a model, but rather the relative merits of the models being tested. The “best” model is the one with the lowest AIC score. The AIC is defined as:

$$AIC = 2k - 2\ln(L) \qquad [1.5]$$

where k is the number of parameters used by the model and L is the optimised likelihood. From this formula it is clear that the AIC not only rewards a well fitting model, but also punishes excessive parameter use. An alternative statistic is the Bayesian information criterion (BIC). This is very similar to the AIC, but penalises parameters more severely (Schwarz, 1978).

In an ML framework, the Goldman test is a test utilised to investigate the overall adequacy of a given model to describe the data (Goldman, 1993, Whelan et al., 2001). Given a model of evolution, the test assumes that, for any test statistic S, the observed data would not deviate greatly from that which would be expected to have arisen from the model in question. This is the null hypothesis, or model. The alternative hypothesis states that the data is described by the broadest model possible; one only adhering to the rules of basic probability. If the null hypothesis does not describe the data, the alternative hypothesis can always describe it. Monte Carlo simulations are used to generate the null distribution using the model; free parameters are estimated from the data. The value of S for the observed data should fall within this distribution if the model correctly describes the data. In the case of the Goldman-Cox test, a specific case of the Goldman test, S is the natural log of the difference in likelihoods between the null and alternative hypotheses.

In a Bayesian framework, model assessment is carried out using posterior predictive simulation (Bollback, 2002). Similarly to the Goldman test, posterior predictive simulation expects that the model, if correct, should be able to predict future observations. In general, future observations are unavailable, so predicted observations are simulated using both the model and the posterior distribution of an MCMC run. Given a test statistic *S*, the value of *S* for the observed data should fall within the distribution of *S* values for the projected posterior observations.

1.3.8: Data Amalgamation

So far, inference of relationships for a single gene family has been discussed. There is an inherent problem with the use of a single gene to infer a species phylogeny: not all genes reflect the evolutionary history of the species as a whole, so the phylogeny of a single gene should only be interpreted as the evolutionary history of that gene in particular (Doolittle and Brown, 1994, Timmis et al., 2004). Although this is true to varying degrees for different organisms, care should always be taken when making assumptions about the predictive power of a single gene phylogeny. For the inference of a species phylogeny, it is now advisable to use data amalgamation approaches. This allows many genes to be included in the inference of a tree and, thus, more signals in the data would be detected. Combining data allows us to view the most prominent evolutionary history in the data as a whole, giving us a much more realistic species phylogeny.

Two main means of amalgamating numerous datasets to form one tree exist (Figure 1.4):

1. Supertree methods: A supertree approach is one that creates a single tree by combining the information from a number of input trees. Generally in biology, the input trees are gene trees (i.e: a tree inferred from a single gene family). By combining the relationships seen in the gene trees, a summary of how the taxa are related (often called a species tree) may be generated. A large number of methods to do this are available; matrix representation with parsimony (MRP), quartet joining and gene tree parsimony (GTP), to name a few (Baum, 1992, Ragan, 1992, Slowinski and Page, 1999).

2. Supermatrix methods: A supermatrix approach is one where all alignments are concatenated to form one large composite matrix of sequences. It is important to note the effect of missing data on this approach; some say that low coverage of data (lots of missing data) does not compromise the method (Philippe et al., 2004), but it has recently been show that the specific distribution of the missing data has a greater impact on accuracy (Sanderson et al., 2010).

1.4: Assessing Support

Once a tree is inferred, it is important to know how much confidence to place in this tree. For this reason, a number of statistical methods are available to measure the support of each node on a tree. Numerical values represent the level of support a particular node receives within a defined range, generally from 0% to 100% (or 0.0 to 1.0). Bootstrap support (Efron, 1979)is commonly used to assess support in most frameworks (ML, parsimony etc.), but posterior probabilities are often used to assess support in a Bayesian framework.

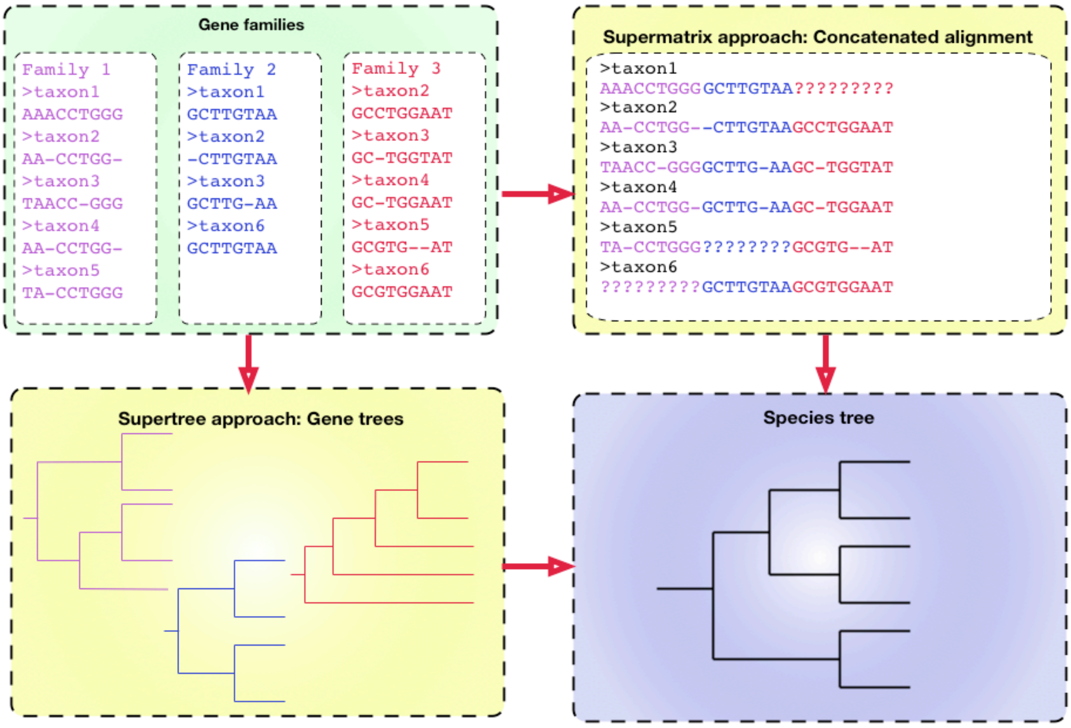

Figure 1.4: Summary of data amalgamation approaches. Starting with a number of gene families, a species tree can be obtained by either a supertree or a supermatrix approach.

1.4.1: Bootstrap

In statistics, bootstrapping is a technique that can estimate properties of a distribution by resampling the observed data. By resampling the data at hand, a dataset with approximately the same distribution as the original may be constructed. By creating a number of these resampled datasets, test statistics may be evaluated on the resampled datasets and used for significance testing.

Bootstrapping was first applied to phylogenetics in 1985 (Felsenstein, 1985). Under this method, datasets of identical dimensions to the observed data are created by resampling columns in the alignment. Resampling with replacement is employed here. This means that the same column may be sampled any number of times, so not every column will occur in all resampled datasets. A defined number of datasets are created in this fashion and each one is used to infer a phylogeny (under the phylogenetic method of choice). The resulting trees are subjected to a majority rule consensus method (Margush and McMorris, 1981). Majority rule consensus methods, as a supertree method of sorts (see section 1.3.8), produce a tree summarising a group of input trees; in this case the trees resulting from the resampled datasets are the input trees. As the name suggests, nodes that occur in the majority of input trees are included in the final tree. In a bootstrapping situation, the percentage of input trees in which a certain node occurs is that nodes bootstrap value. If a certain node occurs in all resampled trees, it is clearly a well supported relationship in the data and it would receive a bootstrap value of 100%.

1.4.2: Posterior Probability

As discussed in section 1.3.6, the Bayesian phylogenetic method employs an MCMC approach to sample trees based on their posterior probabilities. This produces a posterior probability distribution. Using this sample of probable trees, features of the trees may be investigated. For support assessment, the most interesting feature to investigate is the individual posterior probability of each node (Huelsenbeck et al., 2001). This is calculated using a consensus method, where the posterior probability of a given node is the proportion of trees in which it is found. This is clearly very similar to the bootstrapping method, except that rather than resampling the data to infer a set of trees, the sampling process is built into the MCMC chain.

1.5: Sources of Phylogenetic Error

1.5.1: Long Branch Attraction

Long branch attraction (LBA) was first described as a problem in parsimony analyses (Felsenstein, 1978). It occurs when two or more species are erroneously drawn together by their rapid evolution. Rapidly evolving sequences will produce long branches on a tree (as branch length is dictated by the number of differences a sequence displays with the other taxa) and as these data display little in common with any other taxa in the dataset, they are seen to have more in common with each other than they do with the rest of the data. In parsimony, this problem occurs due to the inability of the method to account for superimposed substitutions.

As LBA is such a pervasive problem in parsimony, its effects on other methods have been assessed. ML was believed to be resistant to inconsistencies (Felsenstein, 1973, Yang, 1994), however, situations where model misspecification occurs, LBA can become an issue (Gaut and Lewis, 1995, Sullivan and Swofford, 1997). As Bayesian inference uses similar formulae to ML, it is believed that it is affected by LBA under the same conditions as ML (Kolaczkowski and Thornton, 2004).

Several methods to avoid LBA have been proposed. These include optimal outgroup choice (Wheeler, 1990) and selective sampling (Aguinaldo et al., 1997), but increasing taxon sampling is the most widely used and effective method available. Increasing taxon sampling serves to break up long branches by placing more distance between the problematic taxa (Hendy and Penny, 1989). This approach has been repeatedly proven effective and has resulted in more accurate phylogenies (Hillis, 1996, Pollock et al., 2002, Poe, 2003, Holton and Pisani, 2010).

1.5.2: Compositional Attraction

Compositional attraction is, as the name suggests, when two or more taxa are “attracted” to each other due to similarities in their base or amino acid composition. It has been known for some times that the proportion of A+T content of a genome is rarely equal to the G+C content (Sueoka, 1962). The mechanistic basis for this is unclear, however (Mooers and Holmes, 2000). It was often believed that it was selectively advantageous for organisms that are exposed to high temperatures to have a more GC rich genome (as guanine and cytosine form a stronger bond and, therefore, a more stable mRNA). This, however, is not the case as associations between the GC content of a genome and the organism's optimal living temperature are not seen (Bernardi, 1995, Hughes et al., 1999, Galtier and Lobry, 1997).

Whatever the mechanisms contributing to this bias are, it is clear that the biases have a huge impact on phylogenetic reconstruction (Lanave et al., 1984, Foster and Hickey, 1999). This is due to the difficulties in inferring true genetic distances and substitution rates. Early attempts to correct for this included LogDet transformation (Lockhart et al., 1994) or paralinear distances (Lake, 1994), but these neglected ARSV and were simply inaccurate in a different way. In recent years, methods to account for biases in composition have been developed (see section 1.3.4). These methods, applicable to both an ML (Galtier and Gouy, 1998) and a Bayesian framework (Foster, 2004), allow the composition of each branch of the tree to vary, producing more realistic estimates of the substitution process.

1.5: Horizontal Gene Transfer

When working with prokaryotic phylogeny, it is very important to consider the confounding effects of horizontal gene transfer (HGT). HGT is the process of attaining genetic material from an organism without being the offspring of that organism (vertical inheritance). Several mechanisms of genetic transfer exist:

• Conjugation: transfer of material through a tubular structure called a pilus. The genetic material is in the form of a plasmid, which is a circular DNA structure. These can very greatly in size.

• Transduction: transfer of genes via a bacteriophage.

• Transformation: the uptake of DNA by the recipient cell. A cell that is able to perform this transformation is called "competent". While some bacterial species are always competent, others competency is mediated by physiological factors (Ochman et al., 2000).

HGT was first described in 1951 (Freeman, 1951) and for many years its importance was underestimated. In the late 90's, with complete genome sequences becoming more readily available, the data clearly suggested that not all genes in a given organism were of the same descent, and that HGT was not a rare event (Nelson et al., 1999, Deppenmeier et al., 2002, Galagan et al., 2002). Despite skepticism (Kurland, 2000, Kurland et al., 2003, Eisen, 2000), hypotheses regarding the extent and ubiquity of HGT became popular. This was notably driven by Ford Doolittle, who, in a 1999 paper, stated that "the history of life cannot properly be represented as a tree".

In recent years, many scientists have departed from the use of trees in the investigation of prokaryotic evolution in favour of networks. One type of network used is known as a phylogenetic network and is based on split decomposition (Bandelt and Dress, 1992, Bryant and Moulton, 2004). Similarly to supertrees, this works by summarising all of the signals present in a set of input trees, except that relationships are not forced to adhere to a tree structure. If the input trees disagree about a certain clade, that clade will appear as a network displaying all of the relationships present (for an example, see Fitzpatrick et al., 2006). Gene sharing networks have also been constructed (Halary et al., 2010, Alvarez-Ponce and McInerney, 2011). This involves creating a network of homology, where a node represents a gene and an edge between two nodes represents homology in their sequences. In this way, the extent of sharing between species can be viewed.

1.6: Aims of this Thesis

In this thesis I wish to present a new method for site rate identification (TIGER), its capabilities based on test data, its software implementation and its application to unknown datasets. Chapter 2 will discuss the mathematical basis for the method along with results regarding its performance on a number of datasets. These datasets are either simulated or empirical, but in all cases the characteristics of the data are known prior to analysis. The reason for using this kind of dataset in the testing of a new method is quite clear; there is no way to know if the method is performing as it should unless the features of the data are known.

Chapter 3 sees the implementation of the method in the form of a software program. With the growing size of datasets, given the genomic sequencing driven world of phylogenetics, it is no longer feasible to apply mathematical methods to data manually as LeQuesne did. For efficiency and easy pipeline integration, TIGER was implemented in Python under a command line interface. Also in chapter 3, an alternative application of TIGER is explored.

Chapter 4 shows the incorporation of TIGER into the analysis if a controversial issue: the placement of the mitochondria within the α-proteobacteria. This has been studied using several different datasets and experimental approaches, but I believe that, until recently, phylogenetic methods were not sophisticated enough to account for such complex data. The data clearly displays severe compositional heterogeneity and a broad range of ASRV, but the methods employed in earlier studies to account for this were suboptimal. Current modelling techniques to account for the variations, along with TIGER, were applied to the dataset.

Chapter 2 - TIGER: Tree Independent Generation of Evolutionary Rates

I would like to begin this chapter by making a statement. The work in this chapter has been published in a paper in Systematic Biology that was co-written with James O. McInerney. I would like to make it clear that, while James helped with the writing and some experimental design, I carried out all of the work presented here. This includes development of the method, data collection and analysis (in consultation with James).

2.1: Introduction

Homologous characters evolve at different rates. Within a given data matrix some characters might evolve at an appropriate rate to resolve the branching order of the taxa in question (Townsend, 2007) while others might exhibit high levels of homoplastic noise and others might be too slowly evolving and therefore mute with respect to phylogenetic statements (Delsuc et al., 2005, Kluge and Farris, 1969, Townsend, 2007, Philippe et al., 2005). A character could be considered important if it contains useful information about the phylogeny of the group of interest and if it is relatively free of homoplasy for that group. Therefore, for deep phylogenetic relationships, a slowly evolving character might prove useful, whereas for shallower relationships, a more rapidly evolving character could prove to be more useful. Character-state substitution rate (i.e. the rate at which a characters state is transformed into a different state) is an important factor to consider when ranking the informativeness of characters. Knowing the rate of evolution of a character *a priori* can greatly facilitate the treatment of characters for phylogeny reconstruction.

A number of efforts have been made to evaluate character-specific evolutionary rates. In 1969, Farris introduced successive approximations character weighting (SACW) in order to weight characters according to a perceived importance assigned to them (Farris, 1969). This weighting scheme sought to ensure that characters with a higher degree of correlation with the phylogenetic history were more highly regarded during reconstructions. Farris defined this correlation as the consistency index (CI) for a matrix, or the goodness of the fit of the characters within the matrix to a given tree. The CI for an individual character on a particular tree is derived as the minimum possible character length divided by the observed character length on the considered tree. So, when a character fits on a tree without apparent homoplasy, the CI value is unity. If additional *ad hoc* hypotheses need to be invoked to explain the evolution of the character on the tree, then the CI value will be less than one (Farris, 1969). The CI for a data matrix is obtained by averaging the CI values for all the characters in the matrix. Therefore a tree must be initially inferred. In his description of the method, Farris pre-weighted characters according to a weighting system devised by Le Quesne (Le Quesne, 1969), though he indicated that initial character weights set to unity would also work. As a consequence of the approach, characters that tend to disagree with the initial tree are given a lower weighting in subsequent analyses, in contrast to characters that tend to agree with this initial tree, whose weight remains high.

In the late 1980s Olsen (1987) noted that among-site rate variation (ASRV) could cause problems in phylogenetic inference (Olsen, 1987) and he attempted to accommodate this variation using a model-based approach that employed a normal distribution. Using a model to account for rate variation across sites can increase the probability of finding the correct phylogenetic tree topology, compared with a method that does not account for rate variation (Yang, 1993). By using an evolutionary model that neglects to account for ASRV, sequences will appear to have undergone fewer mutations overall and will appear to be more similar to their relatives, compared with an analysis using a model that accounts for ASRV. Therefore much of the effort to improve phylogeny reconstruction accuracy has focused on methods that deal with accommodating site rate heterogeneous data (Brinkmann and Philippe, 1999, Farris, 1969, Hirt et al., 1999, Schmidt et al., 2002, Yang, 1996).

Yang (1996) modeled ASRV using the gamma distribution. This distribution has some attractive properties, particularly given that its shape can change from being L-shaped to being hill-shaped, depending on the characteristics of the alignment. Again, this approach tries to incorporate rate variation and it assumes that site rate heterogeneity is well approximated by this model. However, assuming that all sites are free to vary will lead to incorrect estimations when there are sites in the data set that do not or cannot change (Yang, 1996). In 1970, Fitch and Markowitz proposed that for a protein there might be two classes of sites – invariable and variable and they suggested a method of analyzing molecular alignments in order to determine how many positions were invariable and how many were variable (Fitch and Markowitz, 1970). These invariable sites can also confound phylogeny reconstruction and accentuate rate variation across sites. To overcome these issues, some studies have experimented with the removal of sites that violate assumptions of the models that are being used. This has the effect of reducing the range of site-to-site rate variation in the data set.

As an example of a study that effectively reduced site-to-site rate variation, Hirt et al. (1999) not only removed invariant sites, but also removed sites they considered to be fast evolving (Hirt et al., 1999). They identified fast-evolving sites by using two different phylogenetic trees and only removing sites that were considered to be fast evolving on both topologies. In this case, removal of both slow and fast-evolving sites vastly improved the support values for internal branches on the phylogenetic trees and resulted in a robust placement of the Microsporidia.

Many different methods exist for the identification of sites with a high substitution rate (Farris, 1969, Kuhner and Felsenstein, 1994, Brinkmann and Philippe, 1999, Hansmann and Martin, 2000, Schmidt et al., 2002, Pisani, 2004). The majority, though not all, of these methods are tree based. Tree based methods identify rapidly evolving sites based on a tree either provided by the user or inferred by the method before site identification. For instance, TREE-PUZZLE (Schmidt et al., 2002) and DNArates (Maidak et al., 1996, Olsen et al., 1998) estimate evolutionary rates for each character based on a given tree and process of character-state substitution. TREE-PUZZLE can employ a discrete gamma distribution to estimate site rates, with sites allocated to a different category based on their likelihood score on the tree. The DNArates program has been used in conjunction with the fastDNAml program (Olsen et al., 1994) in order to partition alignments of homologous characters into rate categories (Fischer and Palmer, 2005). Fischer and Palmer (2005) used a procedure that is not unlike the SACW approach in order to reweight characters for subsequent analyses. For a data set that was aimed at settling the placement of Microsporidia, they found that early unweighted data sets resulted in a variety of placements of the taxon, while successive rounds of character reweighting tended to result in fewer tree topologies and finally the authors settled on a placement of the microsporidia with the fungi that was best supported by the successively reweighted data.

Brinkmann and Philippe (1999) developed a method known as “slow-fast” where an alignment is split into groups (Brinkmann and Philippe, 1999, Kostka et al., 2008). These groups are generally user-defined taxonomic groups. The evolutionary rate at a given site is calculated as the sum of the number of changes at the same position in all the groups individually. Although groups are, technically, user-defined any prior knowledge of the group will be based on previous tree inferences and, therefore, the “slow-fast” method is, by proxy, a tree based method. In addition, due to the nature of this method it is not suitable for small data sets.

The problem with tree-based methods is that the true tree is rarely known with certainty. Therefore use of an incorrect initial tree can result in incorrect assignation of an evolutionary rate to a character. Each character is compared to the given tree topology, whether correct or incorrect. A character is considered rapidly evolving if it conflicts with the initial tree or has a high level of homoplasy when mapped onto the tree. By assuming a topology prior to site rate identification, a slowly evolving site could potentially appear to be rapidly evolving, simply because the tree onto which it is mapped is incorrect. This initial error can become a source for systematic biases. Therefore, it may be preferable to have a method of determining evolutionary rate for a character that is independent of any *a priori* tree estimation procedure.

Tree-independent approaches to differentially weighting characters for phylogeny reconstruction include the Le Quesne test of character compatibility (Le Quesne, 1969), which provided a ‘co-efficient of character-state randomness’, which could be used, if desired, to exclude characters from subsequent analysis. Essentially, this test evaluates two characters and if they can be mapped onto the same tree without homoplasy, then they are compatible, otherwise they are incompatible. Characters that have the highest amounts of incompatibilities with the other characters might be considered candidates for removal prior to subsequent phylogenetic analysis. Le Quesne later introduced the notion of compatibility within data being indicative of the level of phylogenetic information (Le Quesne, 1989). This work was further extended by Meacham, who developed his "Frequency of Compatibility Attainment" statistic (Meacham, 1994). In 1998, Wilkinson highlighted the advantages of creating split patterns for sites when detecting conflict (Wilkinson, 1998). Conflict, as defined by Le Quesne, becomes much easier to identify and rank when using a universal coding system for sites. Pisani (2004) utilised this idea to identify fast-evolving sites. According to Pisani's method each site in the alignment receives an Le Quesne Probability (LQP) score, which is "[…] the probability of a random character having as low or lower incompatibility with the rest of the data than does the original character" (Pisani, 2004). Pisani used this probability measure to explore arthropod relationships using different strategies for removal of characters with differing LQP values.

Hansmann and Martin (2000), in contrast with the compatibility strategies, proposed a simplistic non-tree based method for identifying rapidly-evolving characters. They used the number of different character states in an alignment column as a proxy for evolutionary rate (Hansmann and Martin, 2000). They cite the intuitiveness of the relationship between higher numbers of polymorphisms at a site and speed of evolution at that site. The set of most polymorphic characters would, therefore, be enriched in homoplastic sites (Hansmann and Martin, 2000). However, each site is treated as a separate entity and consequently, this approach does not include information that may be contained in the data set as a whole, apart from ranking the sites from least to most polymorphic. Additionally, polymorphy does not always equate to rapid-evolution and vice-versa. For example, a site may constantly flip-flop between two character states. This site would appear not terribly polymorphic, yet it is still rapidly evolving.

In this chapter, I present our method, TIGER (Tree Independent Generation of Evolutionary Rates), which is based on a similar concept to Le Quesne (Le Quesne, 1989), Wilkinson (Wilkinson, 1998) and Pisani (Pisani, 2004). TIGER analyzes similarity within characters (Wilkinson, 1998). We expect that fast-evolving characters have lost some, most, or all of their phylogenetic signal and therefore should demonstrate reduced similarity with other sites that are more slowly-evolving. Rather than comparing sites and only allowing them to be compatible or incompatible, our method allows sites to be scored according to varying degrees of similarity. This approach should provide a more fine-grained, or nuanced result than one that scores sites as being either compatible or incompatible.

In this chapter, synthetic datasets were analysed in order to explore the behaviour of our approach and then, to demonstrate the utility of the method, two well-known problematic data sets were studied. Additionally, we show that tree-based site-removal approaches have significant problems, particularly when the data set contains a systematic bias (e.g. convergent base compositional bias), while our tree-independent approach can overcome these biases.

2.2: Materials and Methods

2.2.1: Set Partitions

Our method is based on the analysis of set partitions at each position in a matrix. This matrix could be any type of data, including alignments of DNA or protein sequences or a matrix of homologous morphological characters.

A partition of a set *X* is a set of nonempty subsets of *X* such that every element *x* in *X* is in exactly one of these subsets. Each character in the matrix is treated as a set and this set is partitioned based on character states. A set partition is denoted, for example, as $\{\{1\}, \{2, 3\}, \{4\}, \{5\}\}$ or 1/2,3/4/5. The partition 1/2,3/4/5 shows that for this character, taxa 2 and 3 have the same character state which is different from all the others, taxon 1, taxon 4 and taxon 5 each have unique character states – both different from each other and different from taxa 2 and 3. In this way, each character's partition is determined in order to enable pairwise comparisons with the rest of the characters in the data set. For example, in a nucleotide alignment of six taxa, character *J*=AAGGGC and character *K*=TTCCCA (assuming the order of the taxa is the same for both characters in this example). The partition set for both *J* and *K* is 1,2/3,4,5/6, despite having different character states.

Using this kind of data transformation, we can measure the degree of similarity between characters based on the similarity of their set partitions. We find that a character with a set partition that is similar to many other characters in the data matrix can usually, though not always, be a more slowly evolving character than a character with a set partition that is less similar to the rest of the characters in the matrix.

Therefore, we can use the average similarity of a character's set partition to the rest of the matrix as a proxy for evolutionary rate.

The rate for the character at position *i* is defined as:

$$r_i = \frac{\sum_{j \neq i} pa(i,j)}{n-1} \qquad [2.1]$$

where *n* is the total number of characters in the matrix. *pa*(*i,j*) is the partition agreement score. This is defined as:

$$pa(i,j) = \frac{\sum_{x \in P(j)} a(x,P(i))}{|P(j)|} \qquad [2.2]$$

where |*P*(*j*)| is the number of groups in the partition of the *j*th character and *a*(*x,P*(*i*)) equals 1 if $x \subseteq A$ for some $A \in P(i)$. *P*(*i*) may be defined as a partition in character *i*.

It should be noted that, (1) constant sites are not included in the scoring of characters as they do not contribute to the score, and (2) given two sets *A* and *B*, if $A \subseteq B$, it is not necessarily commutative and, often, $B \not\subset A$. In this case, $pa(i,j) \neq pa(j,i)$. Also, because the rate is based on averaging of combinations of 1 or 0 values, it will always have a range between 0 and 1. A constant site, i.e., a site with only one character state, will have $r = 1$ given that the *pa*, will be one for every comparison.

For example, consider 2 sites *A*=CTTAA and *B*=AGGGG with partition sets 1/2,3/4,5 and 1/2,3,4,5, respectively. $pa(A, B) = 0.5$ as, of 2 partitions in *B* ({1} and {2,3,4,5}), only $\{1\} \subseteq P(A)$ as *P*(*A*) may be {1}, {2,3} or {4,5}. As {2,3,4,5} is not a subset of any partition in A, $a(\{1\}, P(A)) = 1$ and $a(\{2,3,4,5\}, P(A)) = 0$ $\therefore$ $pa(A, B) = 1/2$. As mentioned, this calculation is not commutative, so $pa(B, A) \neq 0.5$. $pa(B, A) = 1$ because all partitions in $A \subseteq P(B)$.

This approach is designed to measure how much a particular character tends to agree with the other characters in the data. If a character shares partitions with many other characters then it is likely that they hold similar information. This may be viewed as a signal in the data. Conversely, a character whose set partition greatly differs from the other signals in the data may be thought of as noise. To put it another way, a rapidly evolving character is likely to have sustained multiple substitutions, some or all of whom might be superimposed on earlier substitutions, therefore, this character is more likely to have a set partition that agrees less with more slowly evolving characters.

It is reasonable to suggest that a character that shares few partitions with the majority of other characters could be considered rapidly evolving. On the other hand, a slowly evolving character is more likely to share partitions with, or at least have fewer that conflict with, many other characters. The first assumption might not hold true in a situation where all or most characters in a matrix are rapidly evolving. It is most likely to hold true when evolutionary rates are moderate and when there is a gradient of evolutionary rates from slow to fast. Note that the rate of evolution that is assigned to a particular character is measured in arbitrary units and will vary with the data matrix being used. It is not a measure of substitutions per unit of time and indeed there are no units associated with the rate. This method can be used to analyse DNA, protein, morphological or other arbitrary homologous characters.

2.2.2: Binning

It is often useful or convenient to group sites with similar evolutionary rates together and in our implementation of this method a range of rates can be divided into a user specified number of partitions, or bins. Sites are placed into bins depending on their rate value. The slowest rate and the fastest rate are determined and bins are constructed by splitting the rates into equal partitions. In this chapter, I have used a variety of binning schemes, from eight bins to twenty bins. In theory, any number of bins can be constructed, as long as the number is less than or equal to the number of characters in the matrix.

2.2.3: Data Simulations

In order to test the features of the method, a number of artificial nucleotide data sets were generated using a phylogenetic tree and a pre-specified model of nucleotide substitution. In the first instance, we simply wanted to know if data sets with different patterns of ASRV would return different patterns when analysed using TIGER. Secondly, we wanted to see if removing characters had a beneficial effect on the fit of the data matrix to all possible trees or produced the desirable effect of improving the fit of the data to 'good' trees, while worsening the fit of the data to 'bad' trees. Our third simulation experiment involved the evaluation of whether or not the TIGER approach to character removal would improve the likelihood of resolving deep relationships.

In this study, we have used nucleotide data for reasons of ease of interpretation and also because of the ready availability of excellent computer software (Rambaut and Grass, 1997) to generate the data, however, in principle we could have used protein, morphological or any kind of multistate character matrices.

2.2.3.1: Varying gamma shapes

Using Seq-Gen (Rambaut and Grass, 1997), we simulated two data sets over the same 49-taxon tree (Figure 2.1) and we employed a model that used a discrete approximation to the gamma distribution, with four categories of sites. In order to assess whether or not the TIGER algorithm could detect different patterns of ASRV, two different α values were used in simulations – 0.5 and 20.0 reflecting two different distribution shapes – the first is L-shaped and the second is hill-shaped. Both alignments were 999 bp in length and simulated under the JC model (Jukes and Cantor, 1969). We experimented with other models of sequence evolution and different tree shapes and numbers of taxa and the results are essentially the same as presented here, so we only present the results of the JC simulations on this data set.

2.2.3.2: Changing fit of the data to all trees in treespace

Removal of homoplastic characters in a matrix should have the effect of improving the fit of the data to the true tree while worsening the fit of the matrix to trees that are very different from the true tree. However, given that it is possible to edit any tree to change its topology into any other tree, if we perform any data modification it will most likely influence the goodness-of-fit of the data to all trees in some way. Some trees are very similar to the true tree and some are very dissimilar, consequently, while incrementally removing larger numbers of characters (grouped into bins), we investigated the change in fit of the data to all possible phylogenetic trees for an eight-taxon data set. In our experiments, we measured the change in the Consistency Index (CI) for all trees as bins were sequentially removed, starting with the bin containing the most rapidly evolving characters (a total of 10 bins were used in this experiment). In effect, for the set of all trees, T, we computed the CI for the original data set on tree t ($t \in T$) and compared this value to the CI value for the data set with Bin10 removed. We then plotted this value against the 'nodal' distance (Puigbo et al., 2007) between the true tree and tree t (when t is not the true tree). For the true tree, the nodal distance is always zero. We carried out the same procedure when we removed Bin9+Bin10, Bin8+Bin9+Bin10 and Bin7+Bin8+Bin9+Bin10.

2.2.3.3: TIGER rates vs likelihood scores

Using the correct tree and the correct model, site-specific likelihood scores can give a very good estimate of character evolutionary rate. We wished to test how well the TIGER approach could identify these characters without any knowledge of a tree. We used 100 different seven-taxon trees chosen at random from treespace (containing 945 unrooted trees). A nucleotide alignment of 999 positions was generated under the JC model for each of these 100 trees. We generated site-specific likelihood scores in PAUP* (Wilgenbusch and Swofford, 2003) for all 945 trees for each data set and we measured the ranking of sites on each tree to TIGER rankings. That is to say, the site (s) with the highest likelihood value are ranked as #1 and the site(s) with the lowest value as #999 and likewise for TIGER rates. The Euclidian distance between all likelihood rankings and TIGER rankings was calculated. This is a very simple measure of the average difference in rank for a character in the two lists.

2.2.3.4: Deep branching tree

Rapid evolution can obfuscate deep relationships on a tree, often leading to unwanted polytomies. This situation is particularly problematic when long unbroken branches subtend a series of rapid cladogenetic events. To test whether the TIGER approach could help resolve deep relationships where there is very little phylogenetic signal, we used the JC model of sequence evolution to produce 100 simulated 999 bp nucleotide data sets across the eight taxon tree shown in Figure 2.4a. The short, deep branches combined with long terminal branches presents a difficult problem for phylogenetic analysis, mostly due to the confounding effects of rapidly evolving characters. To ensure that the data generated displayed poor phylogenetic resolution, we built a majority-rule consensus tree from ML trees constructed from each of the data sets prior to any site removal. This was repeated after removal of sites dictated by TIGER and to test the performance of a tree based method in this scenario, we also repeated the analysis after removal of rapidly-evolving sites identified by maximum likelihood. The maximum likelihood tree was estimated using PAUP* and the sites were categorized on this tree using TREE-PUZZLE.

2.2.4: Empirical Data

2.2.4.1: Thermus data set

In order to further understand TIGER's functionality, two empirical data sets were used. A 1,273-column alignment of bacterial 16S ribosomal RNA genes known as the *Thermus* data set is well studied (Embley et al., 1993, Mooers and Holmes, 2000) and we used this data set to examine whether the TIGER approach is useful for accounting for base compositional biases. This data set contains three thermophiles, *Aquifex aeolicus*, *Thermotoga maritima* and *Thermus aquaticus* whose sequences are enriched in G and C nucleotides and two mesophiles, *Bacillus subtilis* and *Deinococcus radiodurans* whose nucleotide composition is more balanced. A combination of compositional bias and distant relationships can mean that when there is only a weak phylogenetic signal, it can be overcome by the similarity in base composition of the most rapidly evolving positions in the alignment. In general, many methods of phylogenetic analysis will group the thermophiles together in this data set, despite the fact that there is strong evidence that *T. aquaticus* and *D. radiodurans* are sister-taxa (Embley et al., 1993). We refer to a tree displaying the mesophiles as a monophyletic group to the exclusion of the thermophiles as the ATTRACT tree and this is the tree recovered by most tree inference methods using the whole sequence alignment. We refer to a phylogenetic tree that places *T. aquaticus* and *D. radiodurans* together as the TRUE tree. Due to this well characterized strong compositional attraction we wished to investigate whether site removal using the TIGER approach could influence recovery of the correct tree. However, to demonstrate the different effects of site removal in a tree-independent fashion compared with the traditional ML approaches, we also compared the topology inferred after removal of rapidly evolving sites identified by TIGER with the topology recovered after removal of rapidly evolving sites according to TREE-PUZZLE (Schmidt et al., 2002) and SACW (Farris, 1969). We did not use TREE-PUZZLE to infer the tree, we simply used the method implemented by TREE-PUZZLE to assign evolutionary rates to sites, based on a tree that we supplied to the software.

2.3: Results

2.3.1: Varying Gamma Shapes

Our first analysis of the behaviour of the TIGER method focused on the analysis of simulated data sets for 49 taxa with different patterns of rate variation across sites. We chose the 49-taxon data set that is distributed with the MACCLADE software (Maddison, 2004) because it contains a reasonable range of branch lengths and has a moderately large number of taxa (Figure 2.1). There are two interesting points to be made about Figure 2.2. First of all, the two graphs are not the same and furthermore Figure 2.2b, which is generated from the data set with an α parameter of 0.5, is more L-shaped than Figure 2.2a, which was generated from the data with an α parameter of 20. This indicates that the TIGER approach is detecting the different ASRV patterns. What is of further interest is that within each graph there is a clear multimodality. There are four clusters of bars on the histograms (indicated by the alternative shading and clear zones on the diagrams).

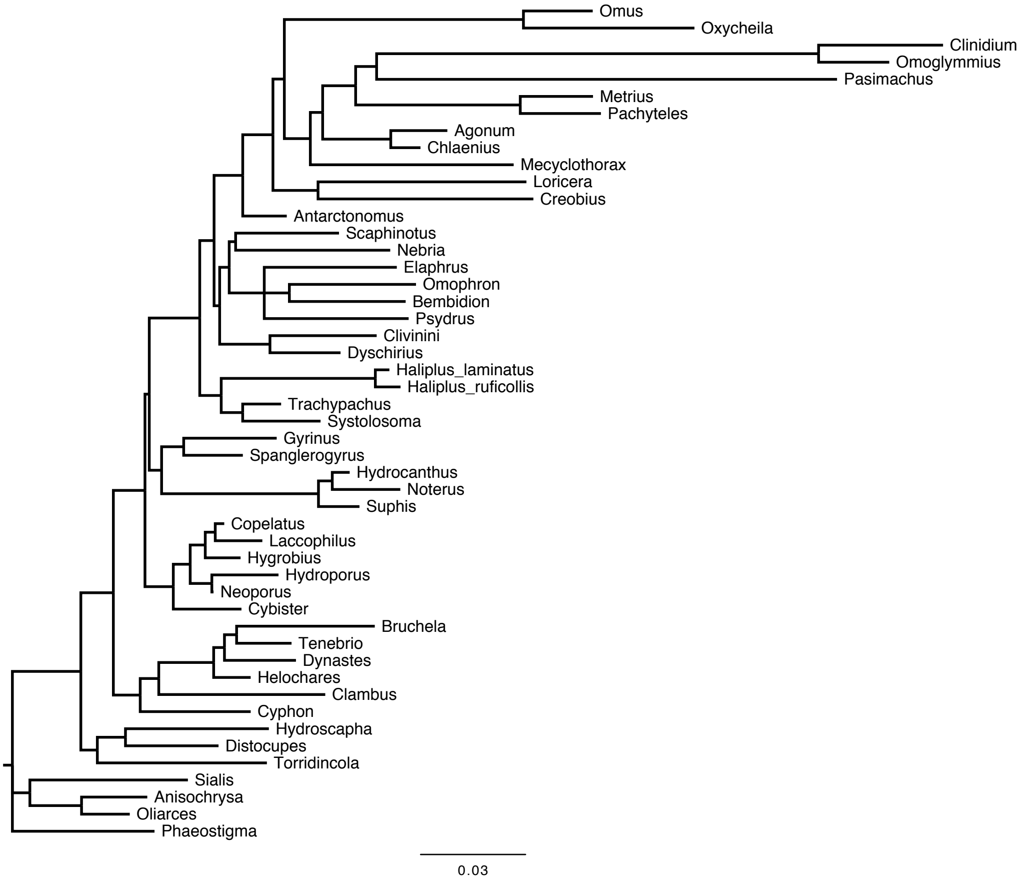

Figure 2.1: 49-taxon tree. This tree was used to simulate data to test TIGER's ability to detect subtle changes in gamma shapes. This data is supplied with the MACCLADE software (Maddison and Maddison, 1992).

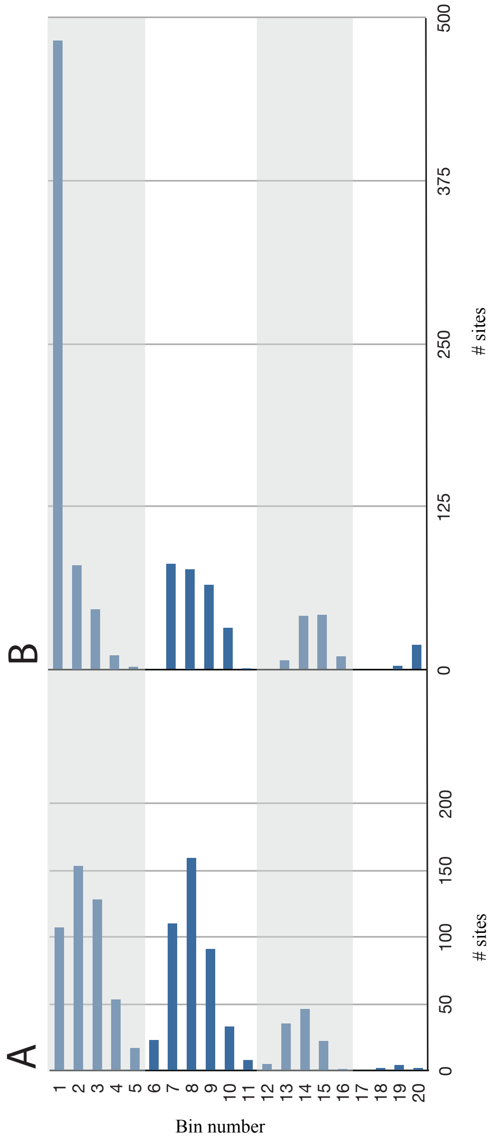

Figure 2.2: Histograms of binning results for two different data sets with different ASRV. (a) A 999 bp, 49-taxon data set generated using the tree in Figure 2.1 and ASRV modeled using a gamma distribution with a shape parameter of 0.5 (b) data set of the same size and topology but with ASRV modeled using a gamma shape parameter of 20.0. The alternating shaded and clear areas indicate the four categories of sites that approximate the gamma distribution.

When the seq-gen software generates data, it uses an approximation to the gamma distribution and in these cases an approximation was employed that used four categories of sites. The TIGER approach has identified these subtle patterns and has placed the different sites into clusters.

2.3.2: True Tree vs Incorrect Trees

If the removal of rapidly evolving characters really is a good idea for improving the chances of recovering the correct phylogenetic tree, then we expect that removal of these characters would improve the goodness-of-fit of the data to the true tree, while worsening the goodness-of-fit of the data to other trees. In order to test this hypothesis, we generated a simulated data set containing eight taxa and using the JC model, according to the protocols previously described. We progressively removed the fastest evolving sites, as judged by the TIGER approach, until we had removed the four fastest categories of sites. We then examined the goodness-of-fit of the data to the correct tree (the tree used to simulate the data) and also the goodness-of-fit of the data to all the other possible trees. We plotted the goodness-of-fit measure (CI) against the nodal distance (as measured by the TOPD/FMTS software (Puigbo et al., 2007)) for the unstripped data set for each possible tree topology and we plotted the

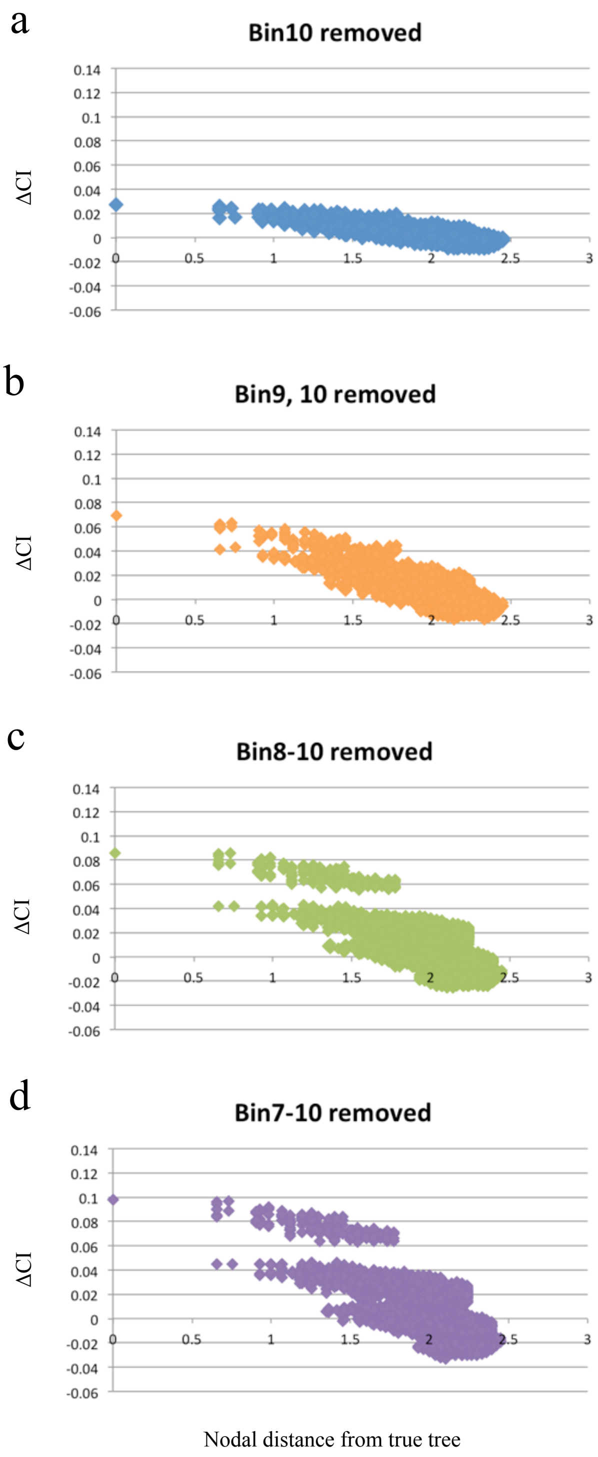

Figure 2.3: Change in CI with increasing site removal. On the abscissa is the nodal distance of a tree from the correct tree and on the ordinate is either the consistency index (CI) or the difference in CI value between the unstripped alignment and the stripped alignment (ΔCI).

change in CI (ΔCI) against nodal distance for each of the data sets where sites were stripped. The results of these experiments are seen in Figure 2.3. In total, there were 10,395 trees examined for each treatment of the data. With all sites included in the alignment, the CI for the correct tree was 0.825. The worst CI value in the data set was 0.612 and the tree with the largest nodal distance from the true tree had a distance of 2.44949 and a CI value of 0.616. In general, there is a negative correlation between CI and nodal distance from the true tree.

When we stripped out the Bin10 category of sites, we saw the CI values increase for some trees and decrease for others. The CI value with the largest increase for any of the 10,395 trees was the CI value for the true tree – an increase to 0.852. In contrast, the tree with the largest nodal distance from the true tree experienced a decrease in CI value and its new value was 0.612. Overall, a total of 5,364 trees (51.6% of the total) saw an increase in CI value, while 5,031 trees experienced a decrease in CI value. Continued site stripping resulted in a progressive increase in CI value for the true tree and a progressive decrease in CI value for the tree with the largest nodal distance from the true tree. When Bin categories 9 and 10 were removed, the values changed to 0.894 and 0.609 respectively, with 5,403 (51.9%) of the trees now experiencing an increase in CI value. When Bin categories 8, 9 and 10 were removed, the values changed to 0.911 for the true tree and 0.601 for the worst tree with 3,811 of the trees having an increased CI value. Finally, when we removed Bin categories 7, 8, 9 and 10, the values changed to 0.923 and 0.597 respectively with 3,257 of the trees experiencing an increase in CI value (31.3%), while 7,138 had a decreased CI value (68.6%).

Therefore, we can see for this data set that site stripping has resulted in a bias in the fit of the data to different trees. In general, those tree topologies that are close to the true tree will begin to fit the data better, while those trees that are least similar in topology to the true will begin to fit the data worse. The tree that is most positively affected by site stripping is the true tree. It must be remembered that the TIGER approach is not tree based and at no time was the TIGER software aware of the topology of the true tree.

2.3.3: TIGER Rates vs Likelihood Scores

To see how well TIGER can approximate site-specific rates we compared it to likelihood scores for each site on every possible 7-taxon unrooted tree. The Euclidian distance from TIGER ranking to the likelihood rankings on all trees were recorded for all data sets, with particular emphasis on where the distance between TIGER rankings and the likelihood rankings on the known true tree fell with respect to the other trees. In 100% of data sets, this distance fell within the top 0.3% of all scores. In 95% of all cases, the distance from TIGER rankings to the likelihood rankings on the true tree was the smallest distance recorded to any tree in the data set.

This shows that the TIGER approach will produce an ordering of the evolutionary rates of the sites that is usually closer to the ranking of sites according to the true tree than to other incorrect trees.

2.3.4: Deep Branching Tree

In order to see whether it is possible for our method to improve the resolution of deep relationships where phylogenetic signal is weak, we simulated 100 different DNA alignments based upon a single phylogenetic tree with long external branches and very short internal branches (see Figure 2.4a). This alignment was designed to represent a difficult problem of phylogenetic inference and was simulated using the JC model of sequence evolution. ML trees for each of the data sets were inferred under the JC model. As expected, prior to removal of rapidly evolving sites, the majority rule consensus analysis using the JC model produced a tree with polytomies and poor resolution (Figure 2.4b), and the only branch that is resolved has a bipartition frequency (BF) of 55% was for a split that separates taxa C and D from the rest. We used the TIGER approach to identify the rapidly evolving characters in the matrices and place all characters into ten bins with increasing evolutionary rate. Removal of the most rapid category of sites, Bin10, which contained between 183 and 502 sites with an

average of 424 between the 100 data sets, entirely resolved all polytomies (Figure 2.4c), with BF ranging from 67% to 99%. We wished to test our method against a tree-based method. We used TREE-PUZZLE (Schmidt et al., 2002) on the same simulated data. Removing the most rapidly evolving category of sites using the TREE-PUZZLE approach (ranging from 269 sites to 481 sites, mean of 334 sites removed) the tree remained equally as unresolved as prior to any site removal, with the BF of the split separating C and D rising to 61 (Figure 2.4b).

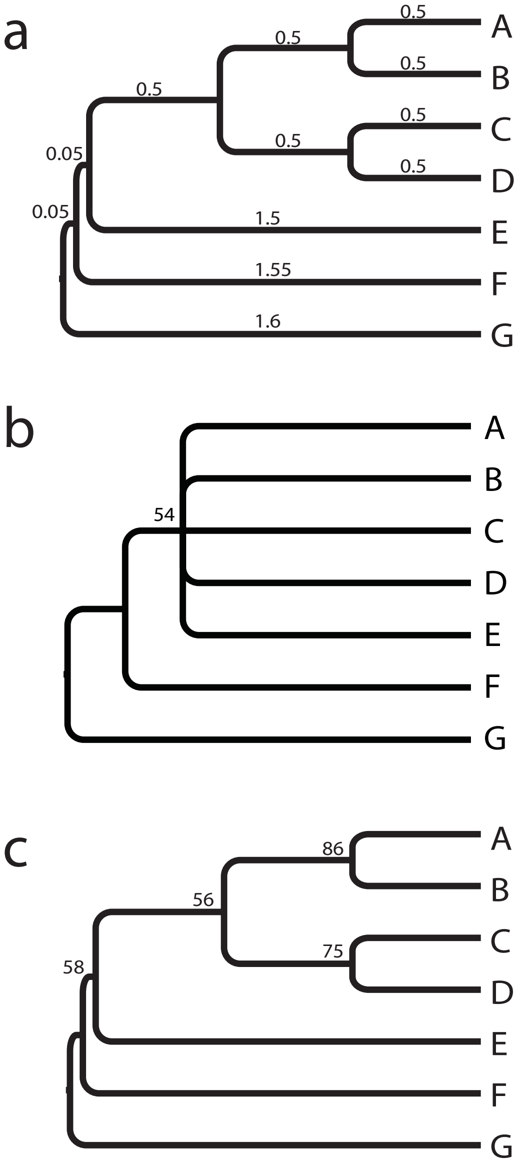

Figure 2.4: Effect of site removal on deep closely-spaced cladogenetic events. (a) The topology of the tree used to generate the simulated data (see text for details of simulation). (b) Majority-rule consensus ML tree after before site removal and also after site removal using maximum likelihood. The bootstrap support value for the unstripped alignments is above the line and the value after site removal using likelihood is below the line. (c) Majority-rule consensus ML tree after removal of Bin10, the fastest evolving sites, according to the TIGER method.

This shows both the pitfall of the tree-based method and the advantage of our tree-independent method. The sites identified as most rapidly evolving by TREE-PUZZLE are those that do not agree with the initial tree inferred by ML. For this reason, removal of these sites does not clarify signals in the data, rather it merely strengthens the signal for the initial groupings. The tree independent method, however, does not need any initial tree, therefore it is not biased towards any single tree and, instead, it picks out genuine signals in the data.

2.3.5: *Thermus* Data Set

The *Thermus* data set consists of 1,273 aligned nucleotide positions from the 16S rRNA gene and is available as supplementary information. Using ML phylogenetic reconstruction implemented in PAUP4.0b10 we examined the differences in tree topology when removing characters judged to be rapidly evolving according to TIGER versus characters judged to be rapidly evolving according to TREE-PUZZLE (with a user-supplied tree, constructed using ML). In addition we used the *reweight* command in PAUP to apply SACW (Farris, 1969) and evaluate the effect that this approach had on the chances of recovering the correct tree. Using the original alignment of 1,273 aligned positions and a GTR+I+G model of sequence evolution, we produced the phylogenetic tree in Figure 2.5a. Using the TREE-PUZZLE software we categorised sites according to the GTR+I+G model using a discrete approximation to the gamma distribution to model ASRV, with a total of eight categories of sites. The category of sites with the fastest rate of evolution was removed from the alignment (a total of 186

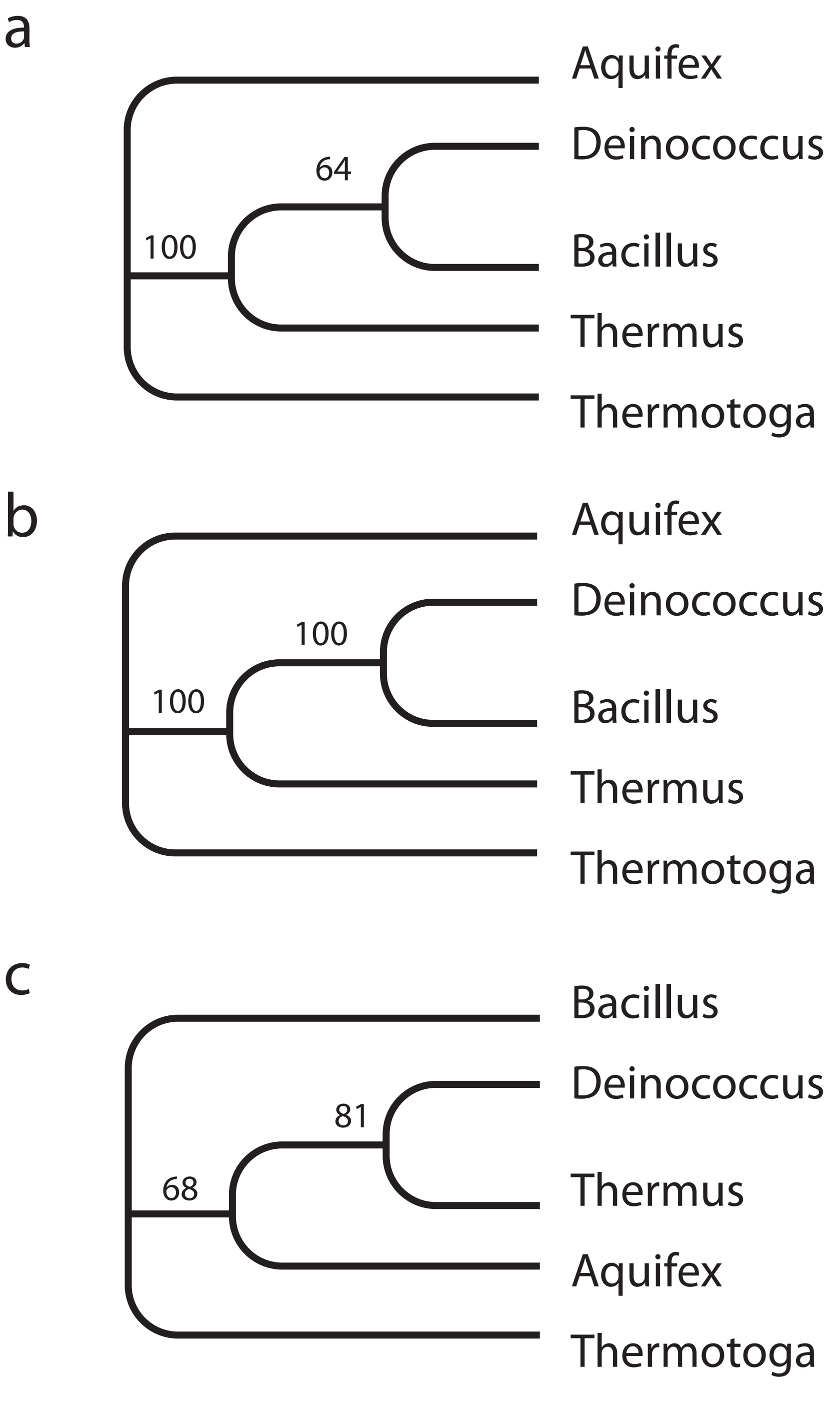

Figure 2.5: Analysis of the *Thermus* data set. (a) Topology and support prior to site removal. (b) The tree recovered after removal of sites identified by PUZZLE and using SACW. (c) The resulting tree after removal of sites identified by TIGER.

sites) and the analysis was re-run using this newer shorter data set (consisting of 1,087 sites). In this case, the same ATTRACT tree was recovered. The most significant difference between the two bootstrap analyses was that the bootstrap support values for the data set with the sites removed were much higher and each of the internal nodes was recovered in 100% of the bootstrap pseudoreplicates (see Figure 2.5b). It must be remembered that the rates of evolution of the sites had been determined using the ATTRACT tree, which is the tree that is obtained in the analysis of the unstripped data set.

In order to investigate the SACW method, we firstly inferred the most parsimonious phylogenetic tree with all sites equally weighted and using an exhaustive search of tree space and the parsimony optimality criterion. Support for this tree was assessed using 1,000 rounds of bootstrap resampling, with the results summarized by a majority-rule consensus procedure. The most parsimonious tree was once again the ATTRACT tree, with bootstrap support values of 92% for the grouping of *D. radiodurans* and *B. subtilis* and 96% for a clan containing *A. aeolicus* and *T. maritima*. Using the *reweight* command in the PAUP software, we weighted the characters according to their CI value on this tree. We then carried out another bootstrap resampling analysis to assess support for groups on the tree. This time the ATTRACT tree was once again recovered, but the support for all internal edges was at 100%.

We used the TIGER approach to identify rapidly evolving sites in the rRNA data set. We placed all sites from the alignment into one of 8 bins according to how rapidly they evolve. The most rapidly evolving category of sites contained 108 sites and these were removed for subsequent ML analysis. Using the GTR+I+G model of sequence evolution on the remaining 1,165 sites, we recovered the TRUE phylogenetic tree. After 1,000 bootstrap replicates, we observed that the grouping of *D. radiodurans* and *T. aquaticus* in 81% of the replicates and the grouping of *T. maritima* and *B. subtilis* was observed in 68% of the replicates. The ATTRACT topology that groups *D. radiodurans* and *B. subtilis* together was seen in 19% of the replicates.

We carried out an additional analysis of the sites that are identified as being rapidly evolving. In all cases, we analysed the most rapidly evolving sites on their own to see if there was any strong phylogenetic signal in those sites. As these sites are saturated for change, we do not expect to see a single phylogenetic signal, rather a number of incongruent signals. In our analyses, only the sites in category 8 of the ML analysis contained any congruent phylogenetic signal. There was 80% bootstrap support for the TRUE tree in these sites. This result demonstrates that not only does such an ML approach result in strong support for the incorrect topology, but also the characters that it discards contain more true phylogenetic signal than the characters that it retains. This needs to be viewed as a systematic error.

2.4: Discussion

In this chapter, I report the development of an algorithm, based on those of Le Quesne, Wilkinson and Pisani that uses similarity in the pattern of character state distributions between characters as a proxy for speed of evolution in a data matrix of homologous characters (Le Quesne, 1989, Wilkinson, 1998, Pisani, 2004). We expect that rapidly evolving characters are likely to lose some, most, or all of their phylogenetic information and will tend to have a character-state distribution that is closer to random than the distribution expected from a more slowly evolving character. A character is assumed to be rapidly evolving if it has a character-state distribution pattern that, on average, is not very similar to the patterns observed in other characters. This assumption is only likely to hold in some (though probably very many) situations. Specifically, in a data matrix where each character is effectively randomised, due to a very rapid rate of evolution or a long evolutionary timespan, we do not expect that this kind of approach will work well. Notwithstanding this caveat (which is a situation that would confound most, if not all, phylogenetic methods), we have observed some very interesting and desirable properties of this approach that make it a useful addition to the phylogenetic arsenal. The TIGER approach identified differing patterns of ASRV, distinguishing alignments that had extreme variation in among-site evolutionary rates from those alignments that had a more even distribution of rates. Additionally, it was able to identify subtleties in the data such as the four clusters of rates in each alignment – a by-product of the simulation process.

The TIGER approach helped improve the fit of the data to the correct tree in our simulations. Removing sites that TIGER identified as being rapidly evolving resulted in a better fit of the data to good trees and worse fit of the data to bad trees, with the true tree being affected most positively. Additionally, using the TIGER approach we could improve the resolution of deep lineages where rapid cladogenesis resulted in very difficult-to-resolve branches. Worryingly, the likelihood approach to removing rapidly evolving sites proved to be problematic – the sites that were removed were those that did not agree with the initial tree, resulting in a situation where, out of 100 simulations, there was little improvement in the recovery of the clades in fast radiations.

Analysis of the ribosomal RNA data set allowed the identification of a number of problems. Firstly, the TIGER approach seems to have some merit as an approach to removing sites that interfere with phylogeny reconstruction. Additionally, two other tree-dependent methods – site identification using a maximum likelihood model of among site rate variation and site identification using the fit of the data to an initially constructed phylogenetic tree – are systematically biased towards favouring the first phylogenetic tree they construct. We, therefore, feel it is important to be cautious when using tree-based methods of assigning evolutionary rates to sites, unless the evolutionary history is known with certainty. We note, however, that a sophisticated compositionally heterogeneous model of sequence evolution is capable of identifying the correct topology for this data set, without the necessity of deleting or reweighting characters (Foster, 2004).

Ultimately, TIGER is an interesting device for identifying characters that do not agree with the majority of the data. We argue here that in many cases this disagreement can be diagnostic of rapid evolution. At the very least, the converse is likely to be true – rapid character evolution is likely to produce a pattern that is not very similar to other characters. Removal of these kinds of characters can greatly improve the accuracy of successive phylogenetic analysis by removing conflicting signals.

There are surely limits to what site removal can accomplish and with certainty site removal is a poor alternative to precise model definition. However, precise model definition comes with a cost. Models that adequately describe the evolution of a set of DNA or protein sequences might, of necessity, be very parameter rich (for example using a combination of Dirichlet processes for both site rate identification (Huelsenbeck and Suchard, 2007) and site-specific profiling (Lartillot and Philippe,

2004) as implemented in the CAT model) and require a large amount of sequence before they become statistically consistent. The most commonly used models of sequence evolution are often inadequate to describe the evolution of the sequences being studied. Model selection approaches often “max-out”, where the most parameter-rich method of analysis is the one that is selected by a Likelihood Ratio Test, Akaike Information Criterion or Bayesian Information Criterion (Keane et al., 2006), indicating that perhaps there are not enough parameters available. Therefore, it might not be an option to use a precisely described model. In the case of the rRNA sequences being analysed in this study, the raw alignment exhibited significant compositional heterogeneity and none of the standard, compositionally-homogeneous, time-reversible models of sequence evolution can adequately account for this heterogeneity. By identifying and removing the most rapidly evolving characters, the models are better able to account for the evolution of the sequences.

We note that bootstrap support values or Bayesian clade probability values are probably meaningless when there is a directed attempt to remove sites that disagree with the rest of the data as this affects the properties of the resampled distribution (see section 1.4.1). It is likely that the support values will tend to increase when incongruent data are removed. When we use bootstrap support values we wish to show that the data have been strongly influenced by the character removal; we do not wish to imply that bootstrapping should follow character removal, as, in most cases, the resulting bootstrap scores are likely to be higher. Alternative approaches to bootstrapping could overcome the bias introduced by site removal. For example, by sampling the bootstrap replicates prior to site removal, running a *tiger* analysis on each replicate and then building trees, the distribution of each replicate will still approximately mirror that of the original alignment, while still providing a statistically meaningful way to measure support.

Given that there are limits to what can be achieved by character removal, we conclude by advising that this method should be used as one part of an overall experimental programme of data exploration. We expect that additional tree-independent methods of analyzing evolutionary rate variation can be developed.

Chapter 3 - TIGER Software and Experimental Applications

3.1: Introduction

In chapter two, I presented the mathematical basis for a new means of calculating evolutionary rates. It was shown that this produced favourable results when used with both simulated and empirical molecular datasets. Manual calculations for few, small datasets are possible, if impractical, but with the advent of whole genome sequencing, the number of genes used in a single analysis has skyrocketed and could easily be in the order of thousands (Holton and Pisani, 2010, Cotton and McInerney, 2010, Hejnol et al., 2009). As these datasets grow, computational methods become essential. For this reason, a software implementation of TIGER became the logical step forward for the method (hereafter *tiger* refers to the software implementation, while TIGER refers to the method itself. The software is available from http://bioinf.nuim.ie/tiger). This software was designed with the goal that it must be efficient, reliable and easily incorporated into users' analyses. As no two users will have the same requirements for *tiger*, it is important to provide a tool that permits the exploration of the data, as well as a simple means to customise the data for the users' needs (Creevey and McInerney, 2005).

Although *tiger* was developed with the evolutionary rates of molecular sequence data in mind, it became clear that it was not restricted to this function. *Tiger*'s scoring system is based upon the level of similarity a site displays with the rest of the sites

(Cummins and McInerney, 2011); the lower the agreement a site displays in relation to the other sites, the lower its score. This implies that the exact meaning of the *tiger* scores will vary for each type of dataset analysed. In the context of molecular sequence data, for example, a low score may mean that the site is rapidly-evolving, as it no longer holds the same information as the majority of the data. For morphological data, a low score may mean that the character in question does not have the same evolutionary history as the other characters, highlighting it as a putative convergence. Given a binary matrix of gene presence or absence characters, where each column in the matrix would represent a gene family, while the rows represent the taxa, a low *tiger* score would suggest that the gene family displays a distribution pattern that is unlike the patterns displayed by the rest of the gene families. Patterns like this may, in part, be explained by horizontal gene transfer (Dagan and Martin, 2007), so we wished to test *tiger*'s capabilities at predicting these events.

HGT is a pervasive phenomenon in prokaryotic biology and is currently thought to be a major influence on prokaryotic evolution (McDaniel et al., 2010, McInerney et al., 2011). For this reason, it is important to be able to computationally predict HGT events. There are a number of approaches available to do this and they fall into three main types:

1. Identifying parts of the genome that differ in its features, such as codon usage or GC content, from the rest of the genome. These alien characteristics may be explained by acquisition of a foreign gene (Moszer et al., 1999, Nakamura et al., 2004).

2. Gene trees that significantly differ in topology from the species tree may be explained by a HGT event (Doolittle, 1999). This, however, is highly dependent on the species tree accurately reflecting the evolutionary history of the organisms in

question (Suchard et al., 2003). Use of an incorrect species tree would lead to the identification of an entirely different set of incongruent gene trees, most probably erroneously.

3. Another approach aims to infer gene gain and gene loss events (Snel et al., 2002, Kunin and Ouzounis, 2003). By mapping gene presence or absence onto a species tree, gene gain and loss events are inferred in a parsimony framework; that is, the solution that postulates fewest gene gain and loss events to explain the gene distribution is preferred. This not only identifies the gene families in which a HGT event occurred, but also when they occurred and how often.

These methods, essentially, aim to identify genes that do not display a similar evolutionary history as the genome as a whole. Using *tiger* to identify gene distribution patterns that differ from the majority, we may be able to infer HGT events without the use of a tree. This can eliminate the possibility of introducing a bias by using the wrong tree and also improves expediency by circumventing the phylogeny inference steps.

3.2: Software Implementation

In order to easily automate *tiger* to run on multiple datasets, and to quickly incorporate it into pipelines, *tiger* is run in a command line interface (CLI). This meant that the language used to implement *tiger* did not require GUI (graphical user interface) capabilities, therefore scripting languages were ideal candidates. These are noted for their excellent text handling capabilities; an important consideration when dealing with sequence data. Ultimately, the two most suitable languages for this implementation are Perl and Python, but with Python's extra ability to incorporate C code (a widely used, extremely fast programming language), it was the clear choice. This allows the possibility of future performance enhancements to be made by optimising the most time intensive parts of the algorithm using C, while still retaining the concise syntax of Python. As *tiger* is implemented in Python, it is platform-independent, so it can be run on any machine with Python capabilities, regardless of the architecture or operating system. Here, I describe the software's various steps and features.

3.2.1: Site Patterns

Firstly, in order to compare sites, they must first be translated to a universal notation. The taxa in the dataset are numbered according to their order in the input. If, in a dataset of five taxa, at a given site, the same character state occurs in taxa 1, 2 and 3 and a different state in taxa 4 and 5 (AAAGG, for example), the "site pattern" will be 1,2,3|4,5. Taxa with an uncertainty (?) in the site in question will be omitted from the pattern, so AA?GG would result in the pattern 1,2|4,5. Users may define a custom list of characters denoting uncertainty. Patterns are created for every site in a given matrix, but in order to reduce computation, only unique patterns are scored, as all sites with the same pattern will obtain the same score. In doing this, each pattern must also be weighted by the number of times it occurs, as some patterns occur more than others. This recoding system also allows any set of characters to be compared to each other, meaning that *tiger* is not constrained to traditional molecular or morphological data.

3.2.2: Scoring and Binning

The scoring system is based on that described in chapter two. This essentially assigns a numerical value (in the range 0.0-1.0) to each site that reflects how much agreement the site pattern displays with all variable sites in the alignment.

When every site is assigned a score they are sorted into bins, or partitions, of the range of scores. This, although not essential, allows the user to easily 1) see the distribution of rates across the range for their data and 2) remove an entire category of data. For example, if the scores range from 0.0 to 1.0, using 10 bins, all sites with a score from 0.0-0.1 will fall into Bin1; 0.4-0.5 will be placed in Bin5 and sites scoring between 0.9 and 1.0 in Bin10. The range is completely data dependent, however, so if the lowest scoring site is 0.03 and the highest 0.899, the bins will be evenly spread across this range. The user may specify how many bins should be used.

3.2.3: File Formats

Tiger is executed using a CLI and accepts data in FastA format. Data must be aligned in advance, in order to ensure that homologous characters are being compared with each other. The software returns a file in NEXUS format (Maddison, et al., 1997). This file format is most suitable as the *Charset* command allows *tiger* to define the sites that fall into each bin. The user may load the sequence into any NEXUS compatible software (PAUP, MrBayes, Mesqite etc) and customise the data for their needs. *Tiger* does not remove any data, but provides the user with a simple means for data removal or reweighting. For this reason, it may also be used for data exploration. For example, by isolating and analysing each bin separately, different signals in the data may be highlighted. In the case of an alignment displaying multiple signals, it is often possible to see clear partitioning of signals into separate bins. By analysing the changes in topology and distances between the bins, sites that may be causing systematic biases can be identified and removed.

To enhance usability, several customisable output features are available in *tiger*. By default, *tiger* outputs the sites in the order they appear in the input alignment with its bin number underneath, however, the sites may also be sorted based on their rate. So while the matrix remains dimensionally and compositionally intact, the sites are rearranged from highest score (slowly evolving sites) to lowest (rapidly evolving sites). This allows the user to see the progression from "good" sites to "bad" ones, the patterns they display and how they interact with the other sites. There are also options to output a file with a list of rates or p-values (see section 3.2.5).

3.2.4: Running Time

Running time is always a consideration when working with large datasets, so I wished to run some benchmarks on *tiger* to assess performance. As previously mentioned, sites with the same pattern will obtain the same score, so only unique patterns of the variable sites are used for the generation of rates. As each pattern has to be compared to every other one, a dataset with n patterns will require n^2-n comparisons. While *tiger* is efficient on small datasets, a variety of factors can affect the number of patterns and, ergo, the running time. To investigate the effect of these factors on running time a number of simulations, each of which were carried out using *seq-gen* (Rambaut and Grass, 1997). Four different tests were set up as follows:

• *Test 1*: DNA datasets with 5, 10, 20 and 50 taxa were simulated under the JC model, using a randomly generated tree topology (Figure A1-A4, Appendix). Branch lengths were randomly generated between 0.0 and 0.5. Sequences were 1,000 bp in length. This tests the effect of taxon number on running time.

• *Test 2*: As more divergent sequences should produce a greater range of patterns than very closely related ones, four datasets were generated using the same procedure as *Test 1*. However, this time, branch lengths were generated in the range 1.0-5.0 (Figure A5-A8).

• *Test 3*: Longer sequences result in a greater number of sites and more chances for a new pattern to arise, so again, four datasets were simulated similar to *Test 1*, but with sequences of 5,000 bp in length.

• *Test 4*: Datatype has the potential to affect pattern number due to a greater number of character states. A five-taxon DNA and amino acid alignment may produce 3,125 and 4,084,101 patterns respectively, producing a much broader distribution of patterns to occur in an amino acid sequence. In this test, four amino acid datasets were simulated across five-, ten-, twenty- and fifty-taxon trees, using the WAG substitution matrix (Rambaut and Grassly, 1997), with branch lengths in the range 0.0-0.5 and a sequence length of 1,000 aa.

The results are summarised in table 3.1. Every test has an increasing number of taxa, ranging from five to fifty. Consistently, taxon number has an effect on the running time; ranging from a 90.2-fold (*test 2*) to a 480.5-fold (*test 3*) increase in minutes taken. *Test 2* tested the hypothesis that greater sequence divergence would produce more random sequences, producing more unique patterns. This is indeed the case and while its effects are minimal, running time is consistently increased. Datatype (*Test 4*) also has minor effect. I postulated that a greater number of character states may give rise to more patterns, but this does not appear to be the case, at least in these simulations. The number of patterns in both DNA data and amino acid data appear to increase at very similar rates. Running time is slightly increased in *Test 4*, however, this is most likely due to the increased complexity of the sites, on account of the greater number of character states present. Increase in sequence length has the greatest effect on both running time and patterns observed. By increasing the length of the sequences 5-fold, the running time can be affected by up to a 25-fold increase. It should be noted that in every case, using fifty taxa results in all sites displaying a different pattern. This is most likely due to the increased probability of a site becoming unique with fifty variables combined with the higher number of potential patterns. For DNA data, fifty taxa can produce orders of billions of unique patterns, making the observation of up to 5,000 unique patterns less surprising.

3.2.5: Permutation Tail Probability (PTP) Test

A major issue with *tiger* is the slightly arbitrary nature of the site removal. Although sequentially removing rapidly evolving sites or entire bins produces favourable results, a means to define when enough sites have been removed remained to be developed. For this reason, the PTP test was implemented. This test is a statistical

Table 3.1. Benchmark running times for *tiger*. *Test 1* is simulated under the JC model; branch lengths ranging from 0.0-0.5; sequence length of 1,000bp. *Test 2* simulated under JC; branch lengths 1.0-5.0; sequence length 1,000. *Test 3* simulated under JC; branch lengths 0.0-0.5; sequence length 5,000. *Test 4* simulated under WAG; branch lengths 0.0-0.5; sequence length 1,000. "Time" is in the format hours:minutes and "Pattern" denotes the number of unique patterns observed in the dataset.

#Taxa		5	10	20	50
Test 1	Time	00:00.3	00:06.2	00:16.2	00:36.9
	Pattern	38	659	1000	1000
Test 2	Time	00:00.5	00:10.7	00:18.0	00:45.1
	Pattern	51	988	1000	1000
Test 3	Time	00:01.9	01:47.0	06:51.2	15:13.0
	Pattern	50	2205	4991	5000
Test 4	Time	00:00.3	00:06.6	00:22.8	00:58.3
	Pattern	36	604	999	1000

randomisation test that is used to test for significance. As with many randomisation tests, the aim is to generate a null distribution of values, under which your observation should fall. For example, in our case, we wish to test whether a site is evolving significantly faster than if it were a completely random site. This is achieved by generating a null distribution of similarity scores between randomly permuted versions of a particular character and the unchanged remainder of the data matrix. For each character, the p-value is calculated as the proportion of times that the permuted version of the character has a lower agreement score with the rest of the data (i.e. that it is more rapidly evolving) than the original character. This allows the user to evaluate whether the average disagreement between the character and the rest of the data is significantly more than might be expected by chance, from a character with the same number of character states and the same composition. A p-value of <0.05 is deemed significant.

This PTP test should identify characters that are highly homoplastic and are the most likely to be misleading during phylogeny reconstruction, highlighting them as candidates for removal. As the rate of disagreement increases from the lowest bin to the highest, we would expect to see an overlap in the sites with high bin numbers and those displaying significant disagreement (i.e. those that evolve quickly are more likely to be found to significantly contribute as little information as a random site). To test this, *tiger* was run with a PTP test on the primate mitochondrial dataset (see section 2.2.4.2; Hayasaka et al. 1988). Of 898 sites in this dataset, only four displayed significance according to the PTP test. This means that these four sites scored worse (or equal to) completely randomised versions of itself in over 95% of cases, showing that it does not contribute to signal in the data whatsoever, merely noise. While these sites should be removed, this test does not account for all types of sites that a user may want to remove. Sites displaying confounding signals, for example, will not be identified by this method because, while it may be misleading, the site still displays enough signal to consistently score better than a random version of itself. Although all sites that were identified as significant fell into Bin10 (Figure 3.1), supporting the hypothesis regarding the correlation of *tiger* scores and p-values, future work developing a “stopping rule” for site removal would greatly improve the process (Goremykin et al., 2010).

3.3: Identifying HGT using *Tiger*

Theoretically, *tiger* can be applied to any type of dataset and the rates can imply a multitude of different things. As earlier mentioned, it has the capability to find gene families with a “strange” distribution in relation to the rest of the families. These genes may be candidates for a HGT event. In identifying these families, the process may emulate a tree based means of identifying HGT events (Doolittle, 1999), but without introducing the potential biases of tree based approaches (Suchard et al., 2003). In addition, *tiger* may be used as a precursor to the the third HGT identification method (section 3.1). This involves positing gene loss and gain events given the distribution of the gene family across the species tree. This would generally involve an exhaustive search of all gene families, but the *tiger* score for each family may be used as a heuristic for families that have undergone HGT.

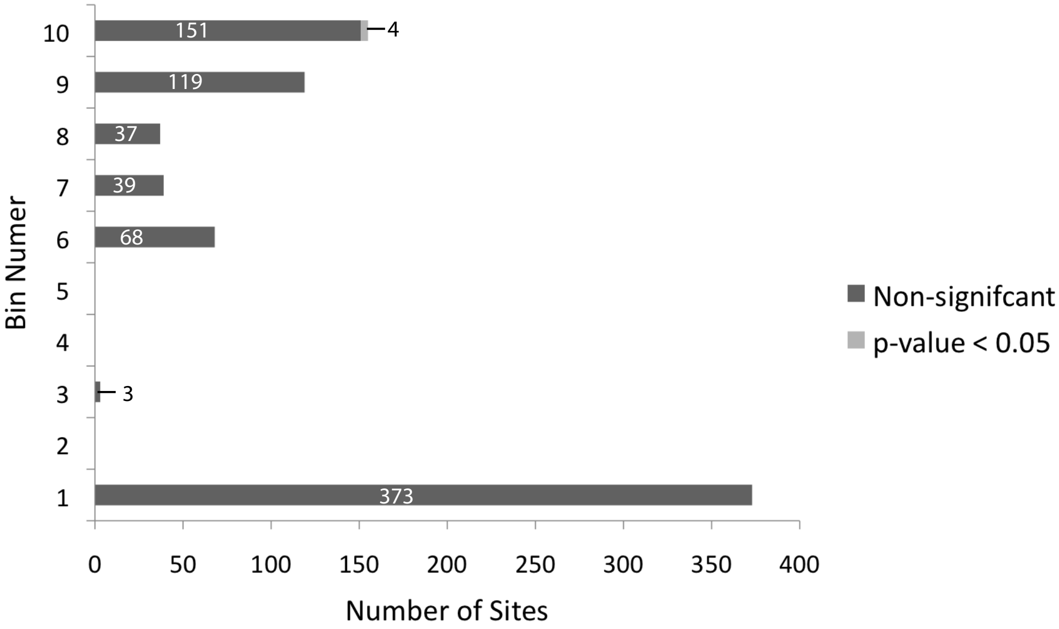

Figure 3.1: Distribution of significant vs non-significant sites in TIGER bins.

3.3.1: Materials and Methods

3.4.1.1: Identification of Gene Families

A total of seven taxa from the α-proteobacteria (*Bartonella quintana*, NC_005955.1; *Bradyrhizobium japonicum*, NC_004463.1; *Brucella abortus*, NC_006932.1; *Caulobacter crescentus*, NC_002696.2; *Rhodobacter sphaeroides*, NC_007493.1; *Rickettsia rickettsii*, NC_010263.2; *Wolbachia* endosymbiont strain TRS, NC_006833.1) were downloaded from NCBI in FastA format. A database with all 20,436 sequences from the seven taxa was formatted for BLAST and searched against itself to discover putative homologs. Markov clustering (MCL), a clustering algorithm, was used to find gene families from the BLAST search results (Enright et al., 2002). This is achieved by implicitly creating an undirected graph of the BLAST hits (where a gene is a node and a hit is represented by an edge) and clustering this based on an algorithm known as flow simulation. The philosophy behind flow simulation clustering is that, given a random walk through the graph, intra-cluster connections are more likely to be visited than inter-cluster connections.

The most important parameter in the MCL algorithm is the inflation parameter (I-value), which affects the granularity of the resulting clusters. As the I-value increases, the resulting clusters are greater in number, but smaller in size; therefore, clustering with higher I-values should produce a sub-clustering of those clustered using a lower I-value. The appropriate inflation value for a given dataset may be calibrated by looking at the differences in clustering using different inflation parameters.

3.3.1.2: Presence/Absence Matrix and Tiger

A Python script was written to create a presence/absence matrix from the gene families; that is, if a species has a gene in a given gene family, it is present, otherwise it is absent. This information may be represented in a matrix of 1's and 0's (1 denoting present and 0 denoting absent) where each site (column) in the matrix is a gene family and each row is a taxon. Only families with 4 or more genes were considered for this analysis, leaving a total of 494 columns in the matrix. As *tiger* can analyse any multi-state character matrix in FastA format, the presence/absence matrix was duly formatted and analysed using *tiger*.

3.3.1.3: Analysis of gene families

Gene families were placed into bins (see section 2.2.2) from one to ten. Those that occur in Bin10 are candidates for a past HGT event. In order to test whether *tiger* is successfully identifying gene families with a HGT event, an ML tree was inferred for each gene family that occurs in Bin10 (i.e. those showing most disagreement with the data). Small subunit ribosomal RNA (16S) was downloaded from http://www.arb-silva.de/ for each of the 7 sequences and an ML tree was inferred. This could act as the species tree. All trees were inferred using PAUP* (Swofford, 2003) and, in every case, parameters were estimated from the datasets (rate matrix, proportion of invariable sites, gamma shape parameter). Additionally, presence and absence characters are mapped onto the species tree to see if a HGT event is required to explain their distribution.

3.3.2: Results

3.3.2.1: MCL I-value calibration

If the distance between two clusterings using different I-values is not great, the lower I-value should be chosen. In this case, the distance is defined as the number of nodes that need to be rearranged in one clustering to form a sub-clustering of another. That is, if 200 nodes in a clustering with an I-value of 2 need to be rearranged to make it a perfect sub-clustering of the clusters obtained using an I-value of 1.2, then the distance between the two clusterings is 200. Higher I-value clusters will produce sub-clusterings of lower I-value clusters, but this is not commutative. As a rule of thumb, if the number of nodes that need to be rearranged is <1% of the total number of nodes, the difference is small and the lower I-value should be chosen. An I-value of 1.2 was chosen for this dataset (Table 3.2).

3.3.2.2: Putative HGT events

Of the 494 families analysed using *tiger*, only seven fell into Bin10 (Figure 3.2). In order to test whether these genes display a different evolutionary history than the species as a whole, we carried out a tree analysis, similar to that of Doolittle. Of the seven gene families under investigation, one contained a paralogous sequence, making it unsuitable for this analysis. Six datasets remained, each of which had only four taxa and of the remaining datasets, only four unique trees were obtained. The species tree was pruned to match the leaf set of each of these four trees and, in all cases, the gene tree was identical to the species tree.

Table 3.2: Matrix showing distance (as described in section 3.3.2.1) between each I-value clustering. I-values are represented in the grey areas. I-values of 2 and 4 produce the greatest distance between clusterings (14). As none of the distances exceed 1% of the total nodes, 6,092, the lowest I-value is sufficient for this dataset.

I-value	*1.4*	*2*	*4*	*6*
1.2	1	4	4	4
1.4		13	12	13
2			14	10
4				13

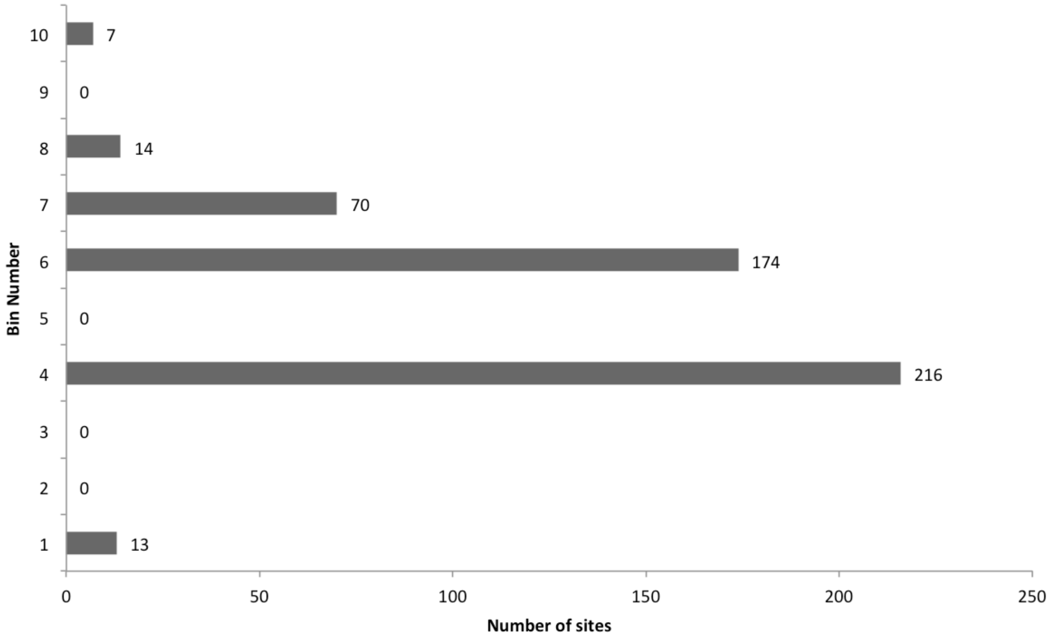

Figure 3.2: Distribution of gene families in TIGER bins.

Gene presence/absence characters were mapped onto the species tree to test whether a HGT event is required to explain the gene distribution (Figure 3.3). Only two unique gene distributions occurred in the seven gene families in Bin10 and neither required a gene gain event to explain their distribution across the species tree. Investigation of the signals in each bin revealed that, in fact, the most prominent signal in the dataset is an artefact of genome reduction. Gene families in which *B.quintana*, *Wolbachia* and *R.rickettsia* are missing the gene are the most frequently observed pattern due to their similarly reduced genome sizes. This is the most homoplastic type of site in the dataset, but its high frequency means that all sites displaying this signal were placed into Bin4. As *tiger* ranks sites based on how much disagreement it displays with the other characters, the families placed in *tiger*'s Bin10 are those that disagree most with this reduced genome attraction. In datasets without this kind of bias, *tiger* would identify the most homoplastic sites.

3.4: Discussion

In this chapter, I described the software implementation of the new method, TIGER, and explored one of its potential uses. Whole genome sequencing is a growing research area and produces huge amounts of data daily. Because of this, datasets are getting bigger, both in taxonomic sampling and the number of genes used. Therefore, in the current scientific environment, software implementations of methods are essential. To compute TIGER rates of evolution, a software implementation, *tiger*, was put in place. An important consideration when dealing with this volume of data is efficiency. Benchmark running times show that *tiger* can process datasets of

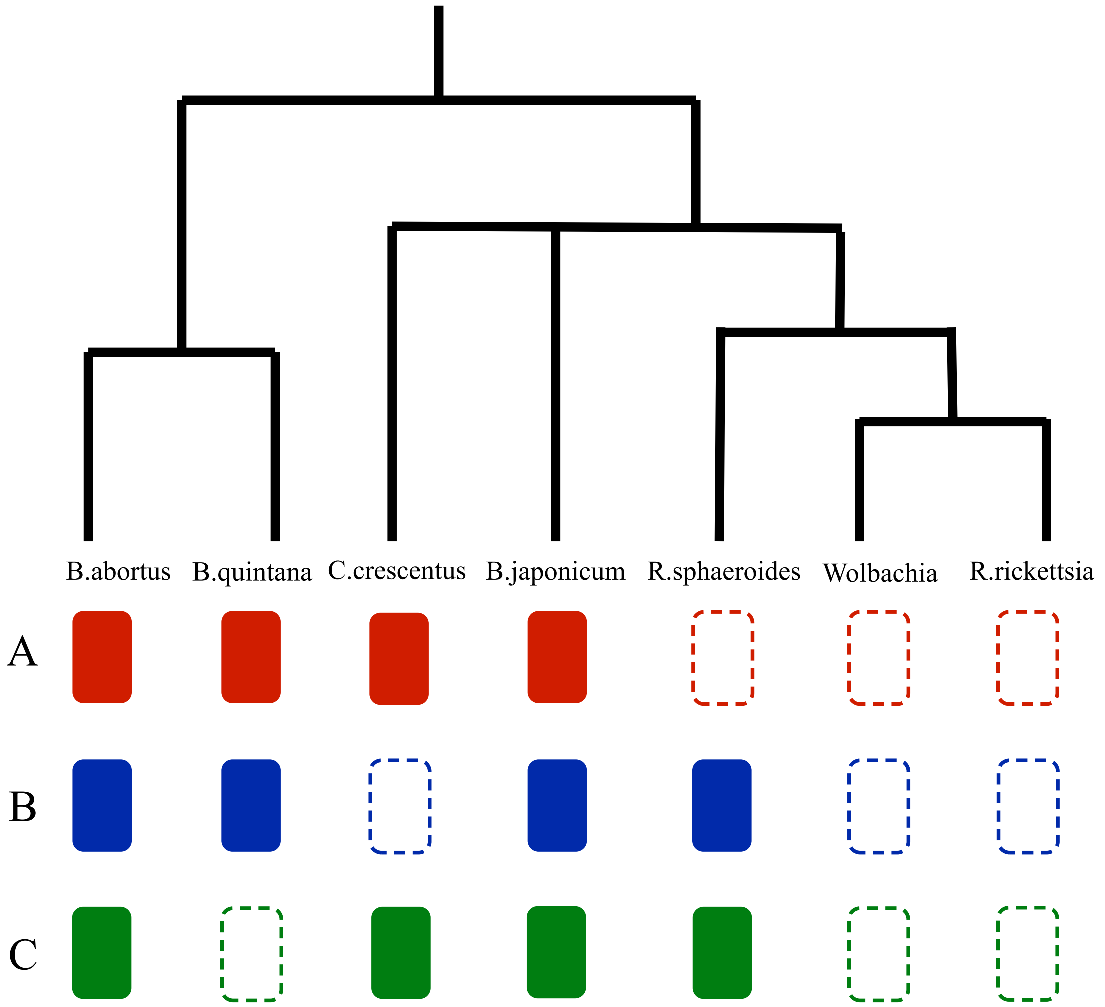

Figure 3.3: Distribution of the genes across the species tree. The 16S rRNA tree is used as the species tree. Filled rectangles represent a gene presence, while unfilled one represent absence. Patterns A and B were found in Bin10 with frequencies of five and two, respectively. Pattern C is the only pattern found in Bin4. It occurs an overwhelming 216 times.

considerable size in acceptable times. Future updates of *tiger* will most likely be based on performance improvements. As Python has the ability to execute C code (a highly efficient programming language), performance would be greatly improved by recoding the most computationally arduous tasks. Also*,* *tiger* would be a very good candidate for a parallelised version due to its nature; each site may be scored independently by a single processor and results collated at the end. This could allow multiple sites to be scored simultaneously.

Tiger's ability to identify HGT events was also explored. Although *tiger* identifies gene families with a distribution pattern unlike the other genes, these genes do not appear to have evolutionary histories that conflict with the species phylogeny or that cannot be explained by differential gene loss. A number of flaws with the experimental design may have caused the result obtained here. Firstly, the size of the dataset would affect the number of individual trees and patterns of loss that can be inferred. This narrows the possibility of a gene tree differing significantly, if at all, from the species tree. Similarly, as small gene families require little HGT to explain their distribution (Dagan and Martin, 2007), a larger dataset may produce larger gene families which are more likely to require HGT events to account for their distributions. Secondly, the method of identifying gene families is questionable. Although MCL is very popular and has been successfully applied to the area of gene family detection (Enright et al., 2002), some curious results were obtained. A number of datasets were tested using MCL and none other than that used for this analysis produced any universally distributed genes. For this reason, the gene families presented here may be unreliable. Despite weaknesses in the experimental procedure, an inherent pitfall of the method as a whole exists; by using only characters for gene presence or absence, this method should only be able to highlight gene transfer events where the host organism did not already possess the gene. The method fails to account for orthologous gene replacement or cases where the host already owned some version of the transferred gene.

Chapter 4 - Mitochondrial Origins

4.1: Introduction

It was shown in chapter two that the TIGER method performed well and produced desirable results on empirical datasets (the *Thermus* dataset and the primate dataset). In both cases the true tree was known with some confidence. We wished to explore both phylogenetic and non-phylogenetic signals in datasets that might address a currently unresolved issue: the origin of the mitochondria. In doing this, we may identify whether there was a role for *tiger* in the analysis of this dataset. Several attempts to elucidate the origins of the mitochondria have been made (Andersson et al., 1998, Esser et al., 2004, Fitzpatrick et al., 2006, Gray et al., 1999), but until recently, phylogenetic modelling methods have not been sophisticated enough to deal with such complex data.

The origins of the eukaryotic cell, particularly the symbiotic origins of the mitochondria, have been of considerable interest for over a century and many hypotheses regarding their emergence exist (Altman, 1890, Margulis, 1981, Martin et al., 2001, Embley and Martin, 2006). Early studies, based largely on the analysis of a single gene (the 16S rRNA was particularly popular for studies of this sort), gave rise to what is known as the "Tree of Life" hypothesis (Woese and Fox, 1977, Woese et al., 1990). This hypothesis postulates that the eukaryotes are a primary lineage of life and that they are the sister group to the archaebacteria. This implies that all eukaryotic cells were derived from an archaebacterium alone. Analyses of this sort highlight the pitfall of single gene phylogenies because, while the 16S rRNA of eukaryotes generally tends to support an archaebacterium sister group relationship, it does not capture the full story of the origins of the eukaryotes. Not all eukaryotic genes reflect this evolutionary history. This became clear as more and more eukaryotic gene sequences became available (Doolittle and Brown, 1994, Brown and Doolittle, 1997, Brown, 2003, Timmis et al., 2004, Gupta, 1998).

Different genes in eukaryotes may display affinities to either archaebacteria or eubacteria. In 1998, a study found that, in yeast, gene function can act as a predictor of its origins (Rivera et al., 1998). Informational genes, such as genes involved in transcription and translation, are more similar to archaebacterial homologs, whereas operational genes involved in metabolism, for example, are more similar to eubacterial homologs. This led to a 2004 study in which Rivera and Lake used a phylogenetic method known as conditioned reconstruction to construct the "Ring of Life" (Rivera and Lake, 2004). The "Ring of Life" hypothesis postulates that eukaryotes were not a primary lineage of life on Earth, rather they are made of a whole genome fusion between two ancient prokaryotes. The conditioned reconstruction method is based on two character states: gene presence or gene absence (similar to that discussed in section 3.4.2.2), and can be used to reconstruct genome fusions (Lake and Rivera, 2004). They found that the eukaryotic genome displayed evidence of a whole genome fusion between two prokaryotes, one from the Crenarchaeota and one from the Proteobacteria, thus providing some evidence to support the "Ring of Life". With poor taxon sampling and an approach based entirely on gene content (i.e. no molecular phylogenies were used. Inferences were based on gene presence and absence characters), this analysis did not incorporate enough information to reach incontestable solutions.

One study used a supertree approach to detect the varying signals present in their data (Pisani et al., 2007). For this analysis, a very large, well sampled dataset consisting of 168 prokaryotic genomes, 18 eukaryotic genomes and over 5,000 genes was used. Supertrees were used to summarise the relationships in multiple input trees (section 1.3.8). The resulting supertree for a given set of input trees displayed the strongest signal present in those trees. By removing all gene trees that are congruent with the initial supertree and, again, subjecting the dataset to supertree analysis, the resulting supertree displayed the strongest sub-signal. In this way, Pisani *et al.* found that the eukaryotes display affinities, in order of strength, to Cyanobacteria, Proteobacteria and Thermoplasmatales. This provides evidence that the eukaryotic genome holds genes not only from its archaebacterial ancestor, but also from what is believed to be the ancestor of the chloroplast (Cyanobacteria) and the mitochondria (Proteobacteria). A 2008 study that takes compositional heterogeneity into account also supports the idea that eukaryotes are not a primary lineage of life on earth (Cox et al., 2008). In this study, however, the phylogenies inferred under a heterogeneous model show that eukaryotes are sister group to the Crenarchaeota. Although these findings conflict with Pisani *et al.*'s findings (as Thermoplasmatales are not a member of the Crenarchaeota), the authors showed, using posterior predictive simulations, that the chosen model fits the extremely heterogeneous data. It was also shown that using a homogeneous model results in the wrong tree. It is possible, therefore, that model misspecification played a part in the results obtained from the Pisani *et al.* supertree analysis.

It is clear that these observations are compatible with the idea that eukaryotic genomes are chimeric and have resulted from a genome fusion event between two primitive prokaryotes. It is also clear that attempts to elucidate the specific archaeal and bacterial lineages that gave rise to the eukaryote has proven difficult (Koonin, 2010). In particular the "parent" of the mitochondria has been an area of considerable interest for decades.

For many years, accepted phylogenetic associations between the mitochondria and bacterial lineages were largely based on 16S rRNA and other single gene analyses (Yang et al., 1985, Gray et al., 1984, Pace et al., 1986, Olsen and Woese, 1993). For data like these, that display extreme base compositional biases and accelerated rate of divergence between the mitochondrial encoded genes and those of their bacterial relatives, analysis of a single gene is both difficult and inaccurate (Gray et al., 1999, Fitzpatrick et al., 2006, Esser et al., 2004). Gene order has also been used to infer a relationship between the mitochondria and bacteria (Sankoff et al., 1992). Again, the accuracy of this approach is questionable due to the prolific rearranging of mitochondrial genes (Gray et al., 1999).

In 1998, the first analysis regarding the parent of the mitochondria that included multiple genes was carried out (Andersson et al., 1998). Two datasets, one containing 12 ribosomal genes and one containing six NADH encoding genes, were used. NJ and parsimony methods were applied to both datasets and all method/dataset combinations place the mitochondria as sister group to the α-proteobacteria. With its minimal taxon sampling and simplistic methods of tree inference, this analysis provides little resolution of the specific "parent" of the mitochondria. In 2004, Esser *et al*. used a dataset of 31 mitochondrial genes that are common to both *Reclimonas americana* and *Marchantia polymorpha* mitochondrial genomes. 14 taxa were used: two mitochondrial genomes, two outgroups and ten α-proteobacteria. The placement of the mitochondria was unresolved using these data, but *Rhodospirillum rubrum* emerged as being more closely related to the mitochondria than the other α-proteobacteria. In 2007, an analysis regarding the effect of HGT on the Esser dataset was carried out (Fitzpatrick et al., 2006). Esser's dataset was also expanded taxonomically to include three mitochondria, two outgroups and 13 α-proteobacteria. A total of 15 genes that showed little evidence of HGT were concatenated and analysed. This resulted in the placement of the mitochondria with the α-proteobacterial group, the Rickettsiales. For completeness, analysis of all 31 genes used by Esser *et al.* was carried out yielding the same result.

Further analysis of this type of data have been carried out (Williams et al., 2007, Sassera et al., 2011, Thrash et al., 2011) and, although both the method of analysis and taxon sampling have progressively improved through time, the problem of compositional heterogeneity seems to be pervasive. This is unsurprising due to the large amount of sequence divergence and the highly biased codon usage patterns of the mitochondria (Nedelcu and Lee, 1998). Heterogeneous models allow multiple composition vectors and rate matrices to be used in a single analysis to account for this sort of bias (see section 1.3.4)(Foster, 2004). Sassera *et al.* (2011) reported that for their dataset of α-proteobacteria alone, no change in topology occurred when using a heterogeneous model over a homogeneous one. However, the introduction of mitochondrial genes would make the compositional bias more pronounced, so we wished to apply these recent methodological modelling advances to allow for differences in composition within our dataset of both α-proteobacteria and mitochondria. These differences can be either lineage or site specific (Foster, 2004, Lartillot and Philippe, 2004) and are both accounted for in this chapter. This potentially allows us to find a model that statistically fits the data while retaining all of the data.

Many analyses dealt with the problem of compositional bias using site removal, with the eventual topology actually being dependent on the method of site removal chosen (Esser et al., 2004, Fitzpatrick et al., 2006). Esser (2004) used TOPAL for site removal (Hansmann and Martin, 2000). This simply uses the number of character states in a site as a proxy for evolutionary rates. Fitzpatrick (2006) used a ML approach to site removal, one that was shown to be problematic in chapter two (see section 2.3.5). We wished to explore the evolutionary history of the mitochondria using (a) site heterogeneous methods, (b) compositionally heterogeneous methods and (c) removal of sites in a tree-independent way. By exploring the signals in the data in this way, we may not produce a definitive answer, but we will be closer to understanding why this problem is so difficult to resolve.

4.2: Materials and Methods

4.2.1: Data

Datasets from an earlier study of mitochondrial origins (Esser et al., 2004, Fitzpatrick et al., 2006) were expanded for this study. Here, we used 30 mitochondrion-encoded genes from 93 taxa made up of six outgroup, seven mitochondrial and 80 α-proteobacteria sequences (Table A1, Appendix). Each gene was aligned separately using Muscle (Edgar, 2004) and concatenated to make a supermatrix with 11,327 characters. Four different smaller datasets were constructed from this data, in order to correct for potential systematic biases (summarised Table 4.1):

- *Tiger* was run on the dataset and 1,504 sites were placed into Bin10. Removal of these sites resulted in a dataset of 9,823 characters.
- Preliminary ML trees (not shown) indicated that within the seven mitochondrial sequences, three were long branched (*L.digitata*, *P. marneffei* and *C. porcellus*) and four were short branched (*Reclinomonas*, *R. salina*, *M. jakobi* and *Marchantia*). Datasets (1) without any short branched mitochondrial taxa and (2) without any long branched taxa were constructed to test whether the placement of the mitochondria was driven by long branch attraction (LBA).
- A final dataset including only the α-proteobacterial sequences was constructed to test whether introduction of mitochondrial sequences disrupted the relationships between the bacteria, possibly indicating a systematic bias.

4.2.2: Tree inference

Bayesian inference was used for the construction of all trees in this chapter. Two sets of analyses were carried out, each to account for a different type of data heterogeneity. As the data is a concatenation of a number of genes which may all evolve at different rates, site rate heterogeneity is likely to be a factor. For this reason, the CAT model was chosen (Lartillot and Philippe, 2004). CAT is a site heterogeneous model that allows each site to be described by a different model. Using Phylobayes (Lartillot et al., 2009), a tree was inferred for each of the five datasets described above under the CAT model.

Table 4.1: Summary of datasets used.

#	Characteristics	Aims
1	All taxa and sites included	To test the raw phylogeny and for use as a control
2	TIGER Bin10 sites removed	To test the effect of site removal on phylogeny reconstruction
3	Long branched mitochondria removed	To test whether exclusion of specific taxa causes a change in topology, indicating an LBA
4	Short branched mitochondria removed	
5	Only α-proteobacteria included	To ensure that the inclusion of the mitochondria and outgroup sequences do not affect the phylogeny, indicating an LBA

To account for lineage specific heterogeneity, a Python platform for phylogenetics called p4 was used (Foster, 2004). p4 allows a number of rate matrices and compositional vectors to be added to the tree before MCMC optimisation, then, during the MCMC process, each lineage may “choose” the best fitting rate matrix and composition vector. For our analysis, the entire dataset was used with constant sites removed. This resulted in 9,122 characters. Firstly, a homogeneous model was chosen using ProtTest (Abascal et al., 2005) to form the basis for the rate matrix used in the compositionally heterogeneous modelling process. Next, two instances of p4 were run: one with two composition vectors (CVs) and one with three CVs. CVs are free to be optimised during the MCMC chain, while the LG model’s rate matrix was fixed in both cases.

4.2.3: Model Testing

Posterior predictive simulations (PPS) were used to test the models in all cases (Bollback, 2002). As mentioned in section 1.3.7, PPS allow the user to test overall model adequacy, rather than relative fit. In order to perform PPS, however, MCMC chains must be run and in order to run a chain, a model must be used. This means that the user must arbitrarily choose a model before the MCMC computation. During the MCMC process, the chain intermittently saves all model parameters (rate matrix, CVs, branch lengths, tree topology etc) so that when the chain finishes, datasets may be simulated using the exact parameters used during the MCMC process. Using an appropriate test statistic, the distribution of the statistic across these simulations can be generated. If the test statistic calculated from the observed data fits into the distribution, the model fits the data.

As mentioned above, two types of heterogeneous models were used: one to account for ASRV, or site heterogeneity, and the other to account for compositional heterogeneity. To account for site heterogeneity, the CAT model was chosen and PPS were carried out using *ppred*, a PPS dedicated software that accompanies the Phylobayes package. In this case, we wished to test whether the model could account for the site heterogeneity in the data. For this purpose, the saturation index was chosen as the test statistic. This allows us to test whether the model sufficiently accounts for sequence saturation. Models that do not account for sequence saturation may lead to systematic errors (Lartillot et al., 2007). This test statistic is implemented in *ppred*. To account for compositional heterogeneity, a homogeneous rate matrix was required and CVs could then be added and optimised. The LG model was chosen based on the BIC score (Schwarz, 1978, Le and Gascuel, 2008). To test whether the compositional heterogeneity in the data was adequately described by the model, the chi-square test for homogeneity was used as the test statistic.

4.2.4: Signal Exploration

By running the concatenated alignment through *tiger*, the characteristics of each bin could be viewed separately. The strongest phylogenetic signal in each bin could be viewed by making a tree. Only bins 7-10 contained enough parsimony informative characters on which to base a phylogeny, so an individual dataset containing the sites placed in each of these bins was constructed and subjected to ML analysis using

PhyML (Guindon et al., 2010). Model testing was carried out using ProtTest (Abascal et al., 2005) and the LG model was used for each tree (Le and Gascuel, 2008).

4.3: Results

4.3.1: Site Heterogeneity

As described above, the CAT model was applied to all five datasets (summarised in Table 4.1) using Phylobayes. Posterior predictive simulations reveal that for each dataset, the model adequately accounts for saturation as measured by the saturation index. As the saturation index is defined as the number of homoplasies per site, a tree topology is needed to reconstruct the full substitution history of a dataset. For this reason, the saturation index is calculated on both the observed data and the simulated data for each posterior predictive sample. This means that both the observed and the simulated statistics are distributions, which are summarised in Table 4.2 using the mean and variance of each distribution.

Knowing that the model accounts for the level of saturation observed in the data, some confidence can be placed in the tree topologies obtained from the analyses. Trees are obtained by taking a consensus of all trees in all of the chains run for each dataset. In these consensus trees, nodes that had a posterior probability of less than 0.6 were collapsed in order to highlight the amount of support different parts of the tree displayed. Trees are shown in Figures 4.1 to 4.5.

Dataset five consisted only of α-proteobacteria. This dataset was important in the analysis to ensure that the inclusion of the mitochondria and the outgroups did not compromise the underlying relationships between the bacteria themselves. All four datasets (1-4) were in agreement with the topology, indicating that LBA, that might be caused by the long branches leading to the mitochondria, is not causing any α-proteobacterial taxa to be "pulled" out of place. Despite this, two different topologies were recovered from the datasets. Datasets 1-3 recover the traditional trees, where mitochondria are sister group to the Rickettsiales. Removing sites dictated by *tiger* results in much better resolution of the tree (Figure 4.2). Dataset four, on the other hand, recovers a tree where the mitochondria are grouped with the α-proteobacteria, excluding the Rickettsiales. This dataset contained only the long-branched mitochondria, so, in the first instance, this was attributed to LBA. Analysis of lineage heterogeneity, however shed some light on the matter.

4.3.2: Lineage Heterogeneity

Using the two models, one with two CVs and one with three CVs, the complete dataset (without invariable sites) was analysed. Posterior predictive simulations reveal that only the model using two CVs adequately describe the compositional heterogeneity of the data (Figure 4.6). Two runs of the model were carried out to test for convergence and vector placement uniformity.

Consensus trees for each of the two runs using two CVs were inferred and the placement of each CV noted. The consensus trees displayed little topological difference, with only the placement of *Azospirillum*, *C.pelagibacter* and *N.hamburgensis* differing between the trees. There was also good agreement regarding the placement of the CVs across the tree. Only six of 193 branches displayed differences in the CV chosen between the two trees (Figure 4.7). Despite the majority of the CAT trees supporting the mitochondria/Rickettsiales grouping, both runs of the heterogeneous model recover the grouping of the α-proteobacteria with the mitochondria, to the exclusion of the Rickettsiales (Figure 4.8). It is clear that the Rickettsiales and the mitochondria largely chose the same composition vector, even though the tree suggests that the mitochondria are more closely related to the other α-proteobacteria. This may suggest that previous grouping of the Rickettsiales with the mitochondria (Fitzpatrick et al., 2006, for example), may have been a compositional attraction.

Inspection of the composition of each group (α-proteobacteria, Rickettsiales, mitochondria and outgroups) reveals that the Rickettsiales and the mitochondria share the highest proportion of gaps compared to the other groups (Figure 4.9). They also, albeit to a lesser degree, share the highest proportions of isoleucine (I), serine (S), asparganine (N), tyrosine (Y) and cysteine (C) (in descending order). It should be noted, however, that a divide among the mitochondria is apparent. When specific mitochondria sequences were removed from the analyses, the topology changed (Figure 4.3 and 4.4). Removal of the long-branched mitochondria resulted in the mitochondria grouping with the Rickettsiales, while removing the short-branched mitochondria produced the alternative topology. Closer inspection of each of these sub-groups of taxa revealed a large difference in the composition (Figure 4.10). Although, at first, the change in topology was assumed to be LBA, it is now very apparently compositional attraction.

Tiger was shown to perform well in counteracting this sort of signal (section 2.3.5), yet the CAT tree recovered after *tiger* analysis still matches the traditional topology (mitochondria with Rickettsiales). The ML trees inferred for each bin could highlight the signal present at each level of evolutionary rate, but each bin returned the same tree topology: the mitochondria/Rickettsia grouping with support values for the clade ranging from 0.596 in Bin9 to 0.859 in Bin10. Closer inspection of these two bins reveals almost identical compositions, each displaying a strong attraction between the short-branched mitochondria and the Rickettsiales (Figures 4.11 and 4.12).

Table 4.2: Summary of posterior predictive simulations to test saturation. Values are "mean +/- variance". It is important in this case to note the p-values. Often in statistics, p-values are used to test significance. In general, a p-value is deemed significant if it occurs in either tail of the distribution (0.05 > p > 0.95). In this case, it is desirable for the statistic to fall within the distribution, rather than the tails. This means that a case where 0.05 < p < 0.95 is acceptable.

Dataset	Observed Saturation	Predicted Saturation	p-value
1	34.1039 +/- 0.443161	34.0475 +/- 0.453095	0.268462
2	32.8123 +/- 0.439826	32.74 +/- 0.453042	0.241923
3	29.9467 +/- 0.373437	29.8925 +/- 0.39001	0.282029
4	31.9003 +/- 0.44786	31.8487 +/- 0.460425	0.297388
5	20.7821 +/- 0.23495	20.6982 +/- 0.247973	0.141148

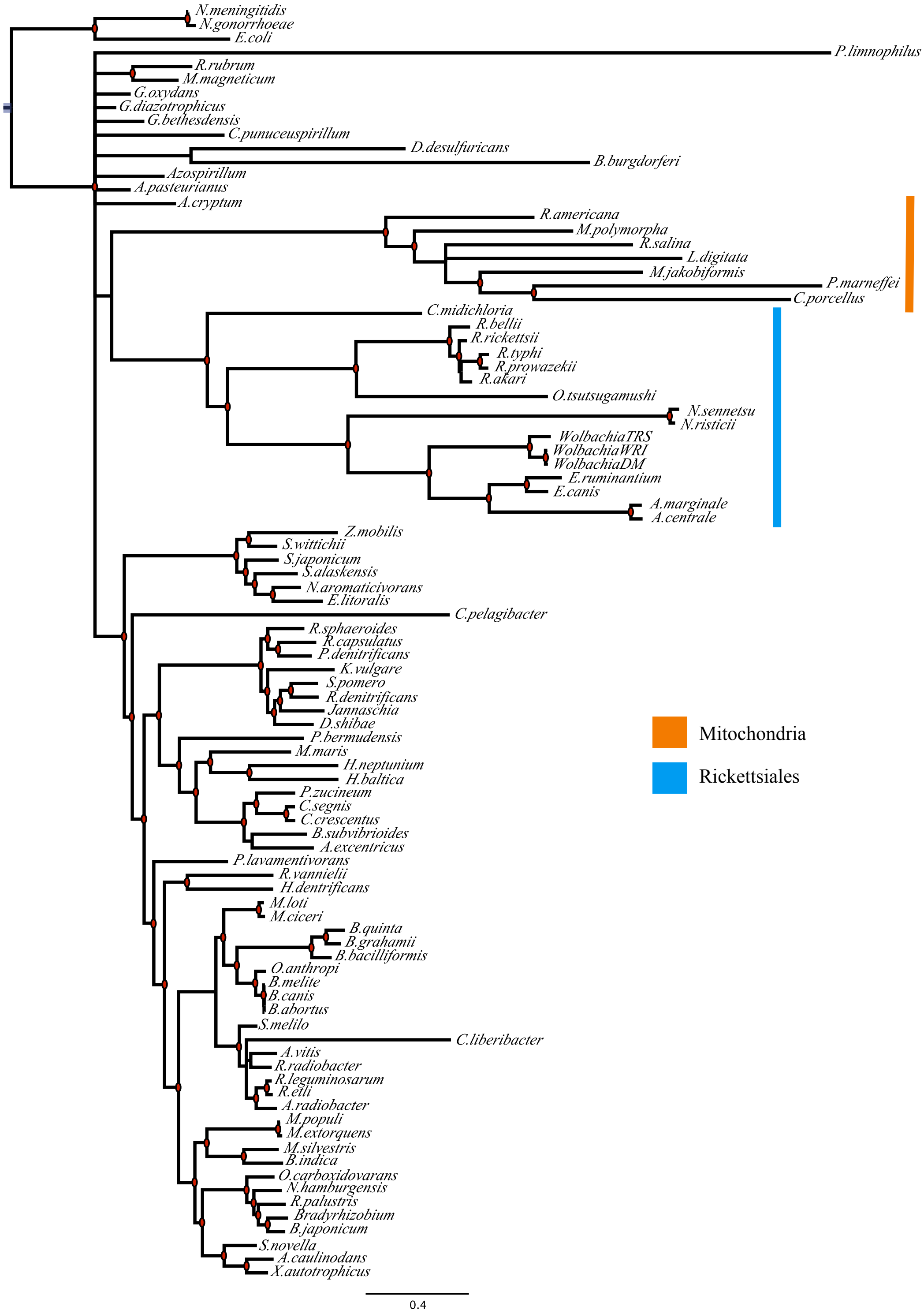

Figure 4.1: CAT tree for dataset 1. Red dots indicate a posterior probability (PP) ≥ 0.95.

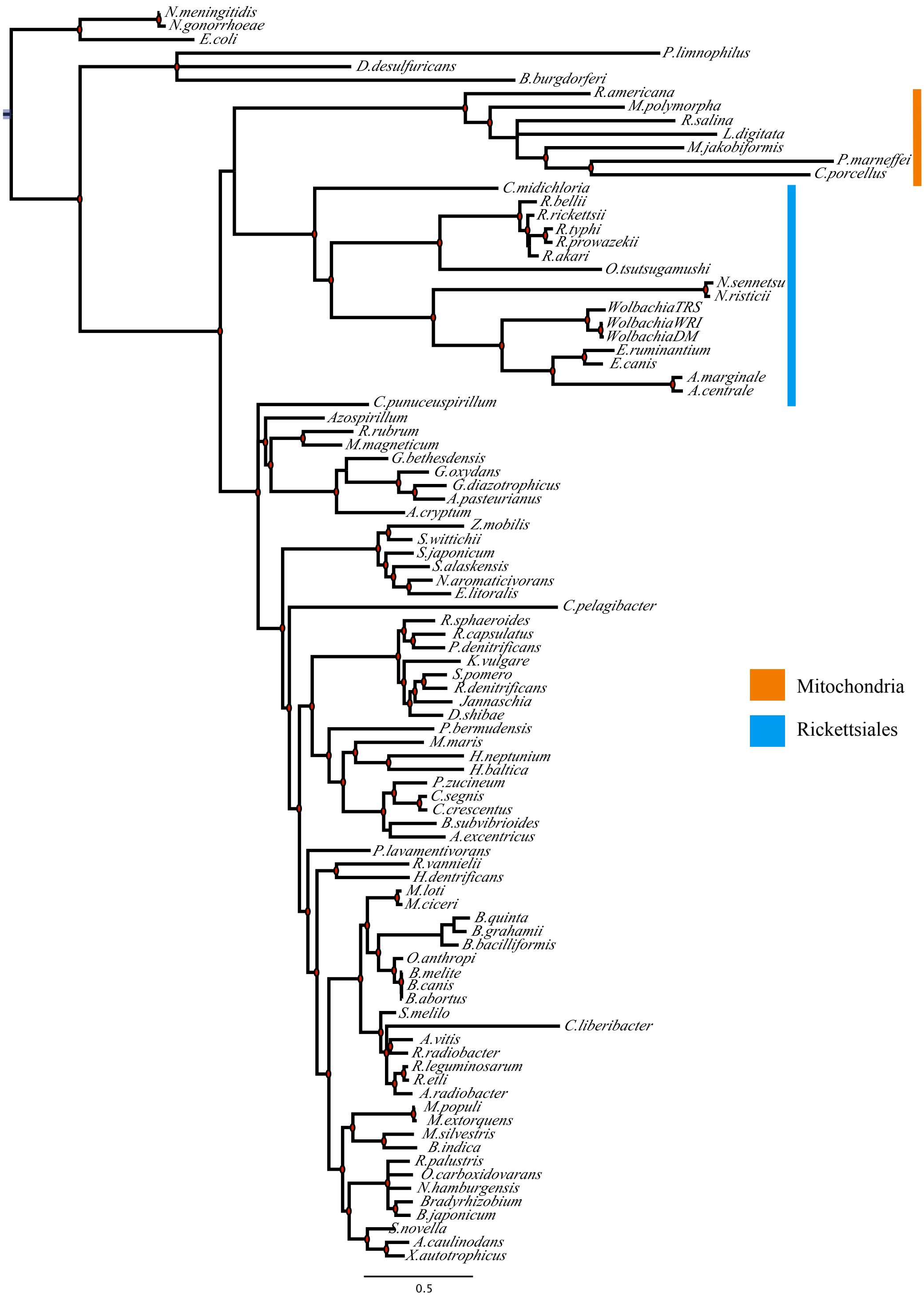

Figure 4.2: CAT tree for dataset 2. Red dots indicate a PP ≥ 0.95.

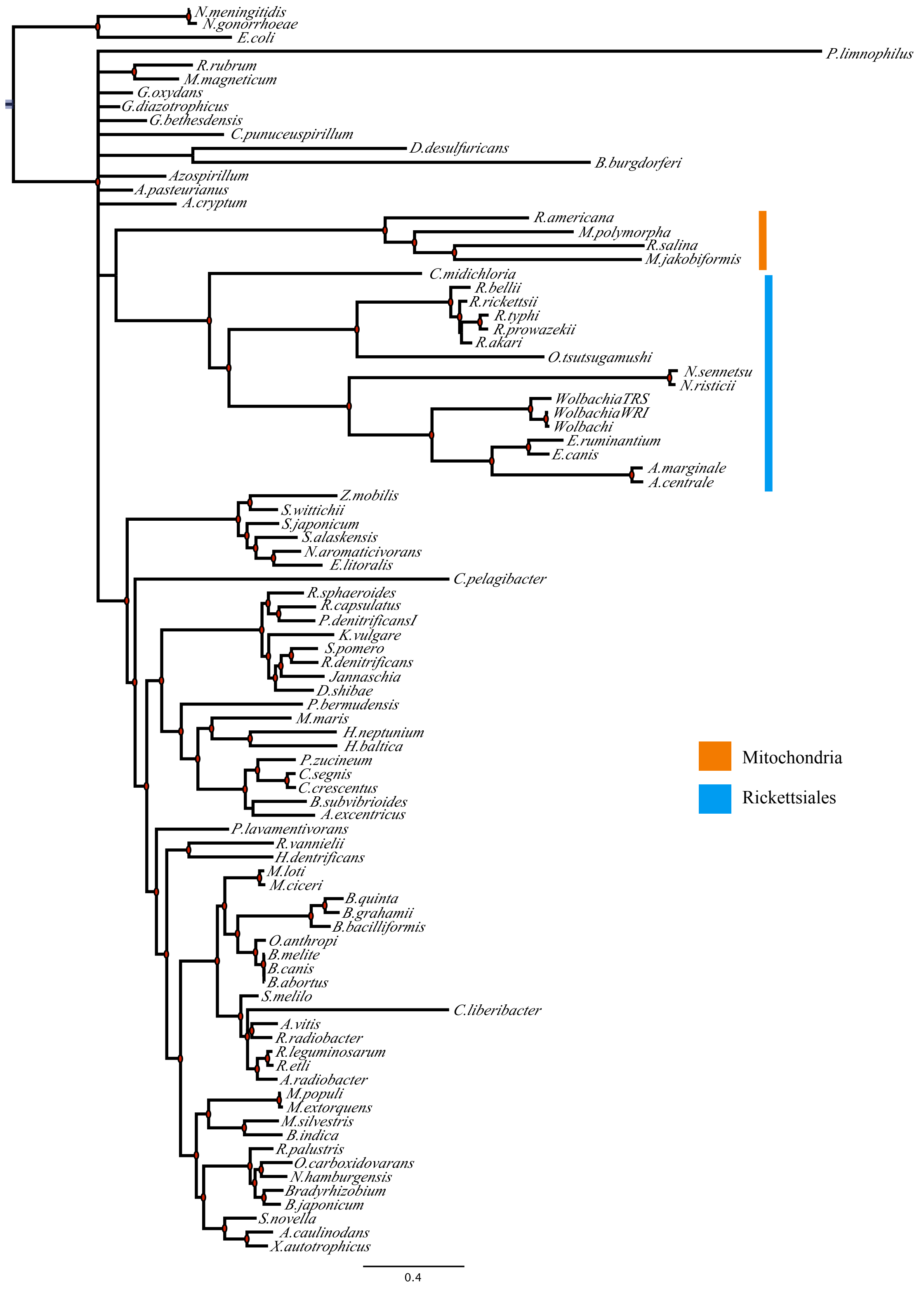

Figure 4.3: CAT tree for dataset 3. Red dots indicate a node where PP ≥ 0.95.

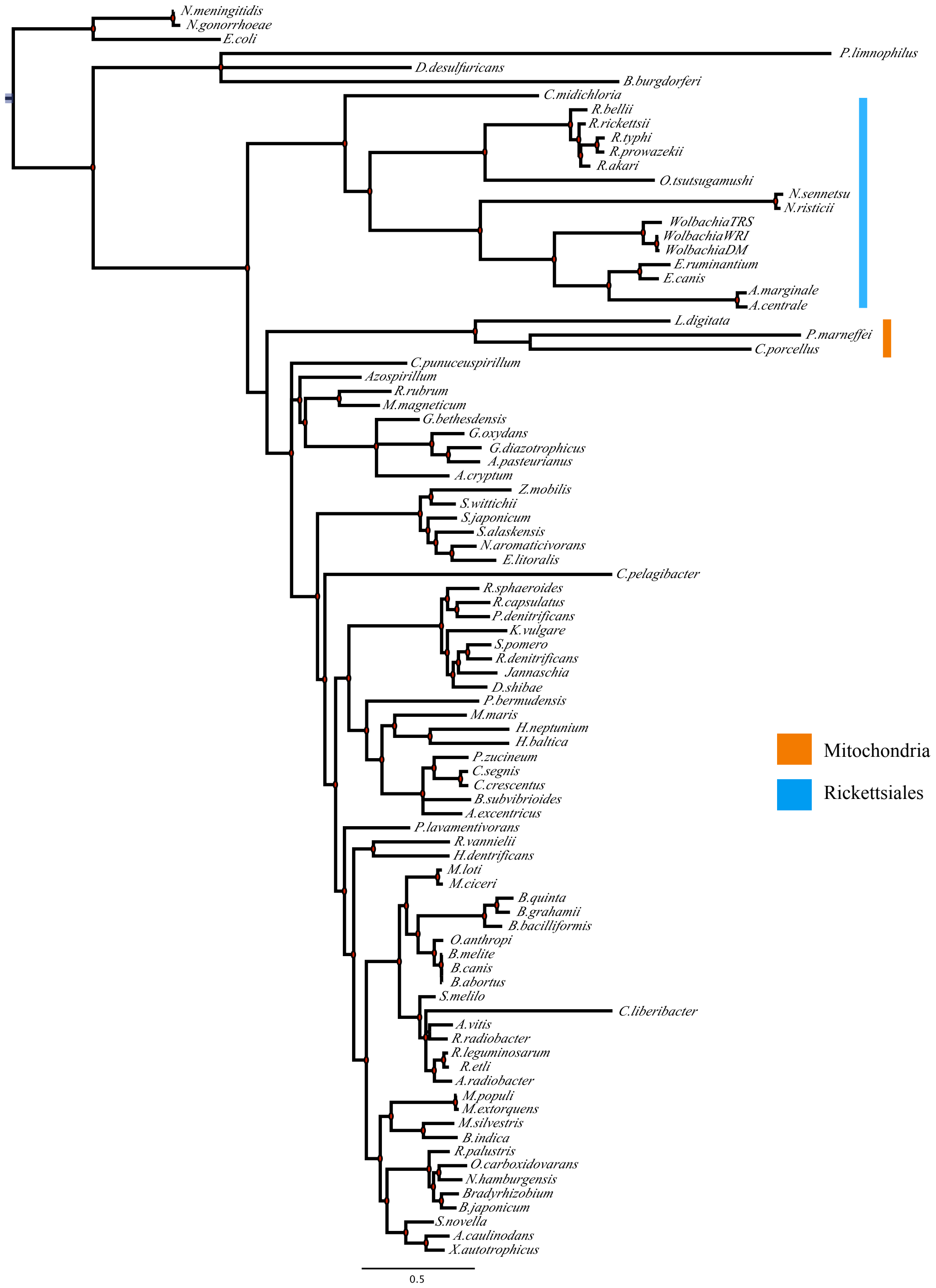

Figure 4.4: CAT tree for dataset 4. Red dots indicate nodes with a PP ≥ 0.95.

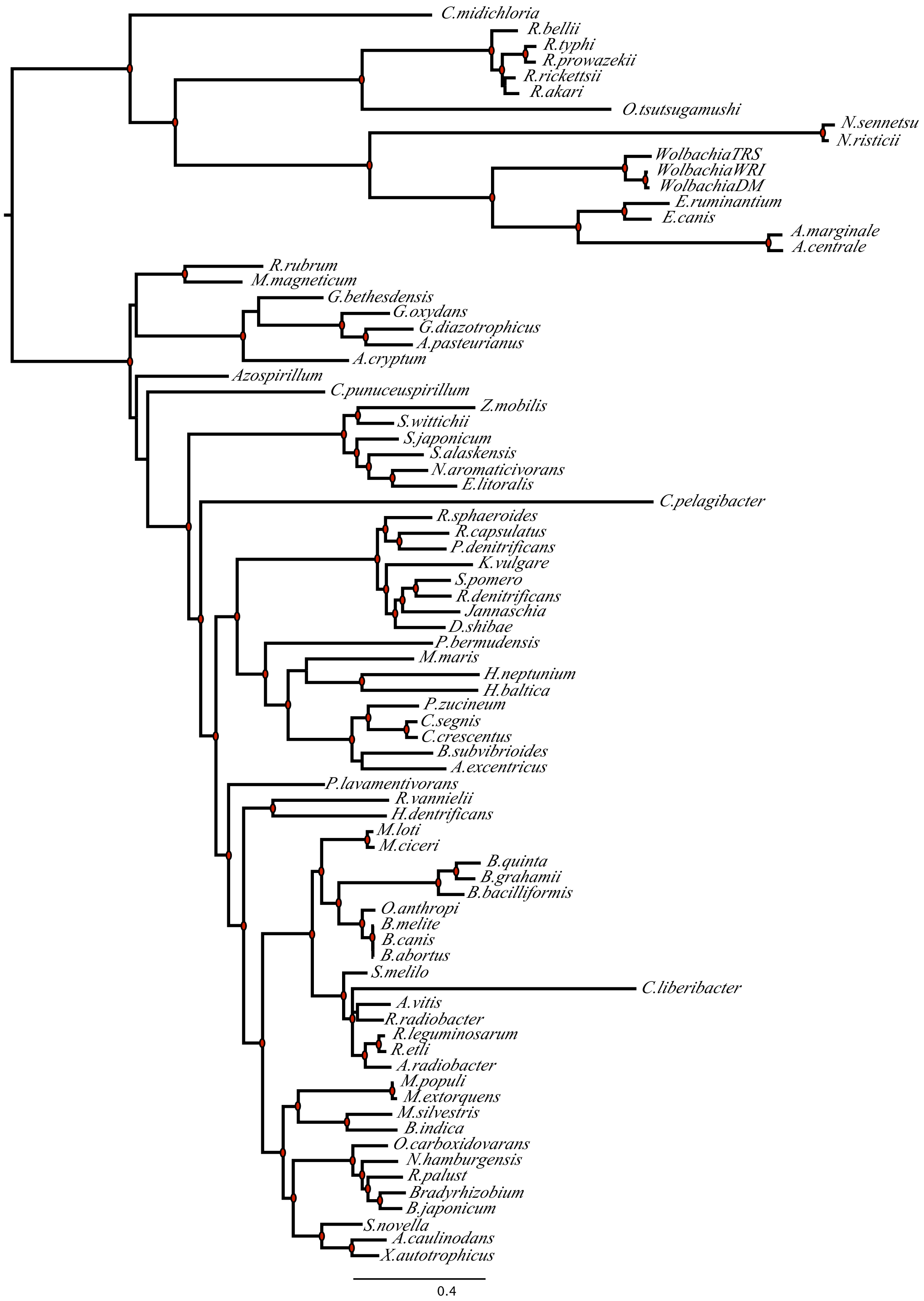

Figure 4.5: CAT tree for dataset 5. Red dots indicate nodes with a PP ≥ 0.95.

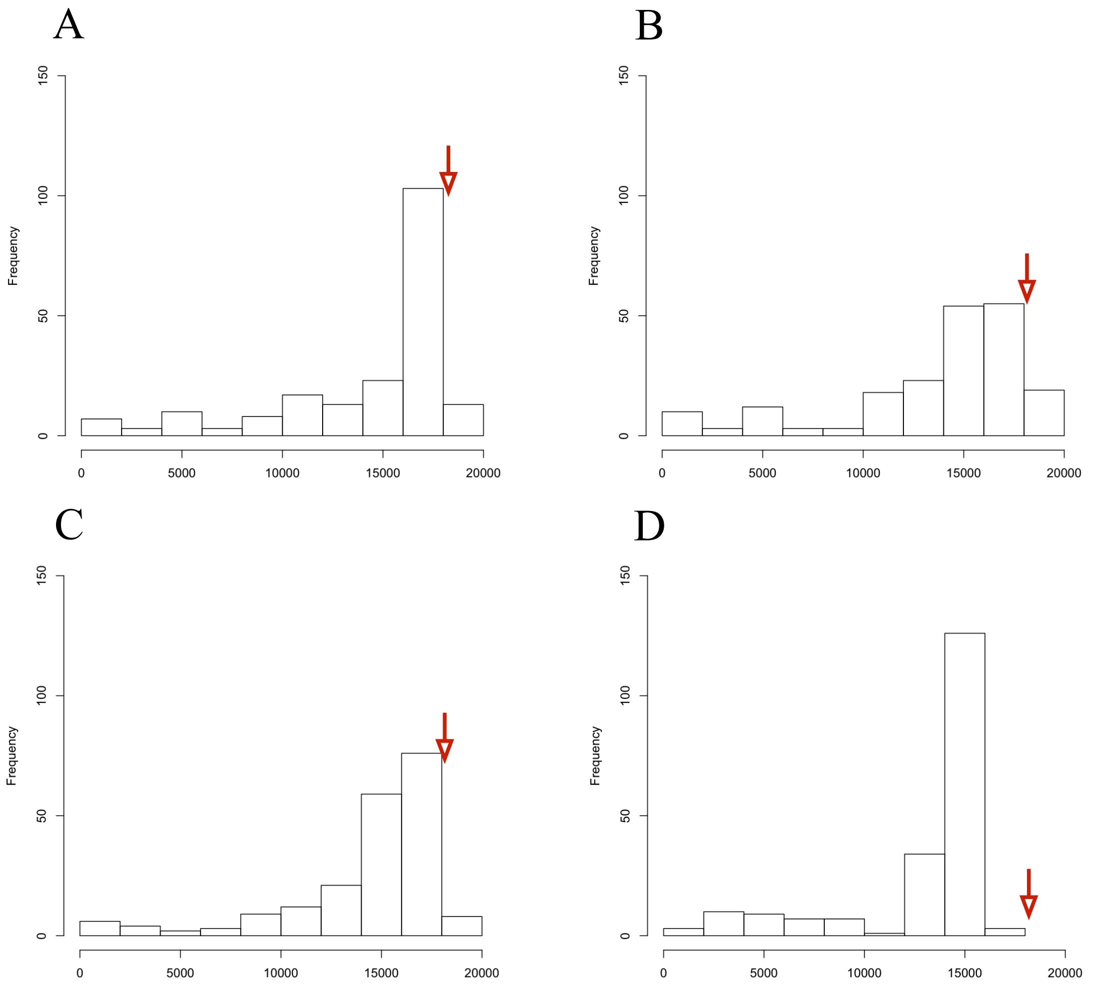

Figure 4.6: Distributions of the χ^2 values for each of the four p4 runs. χ^2 values are on the x-axis in each. Distributions A and B are the posterior predictive simulations for the model using two composition vectors while C and D are for three composition vectors. The χ^2 value for the original dataset (18037.69) is marked in each distribution with a red arrow. P-values are 0.055, 0.085, 0.04 and 0.0 respectively.

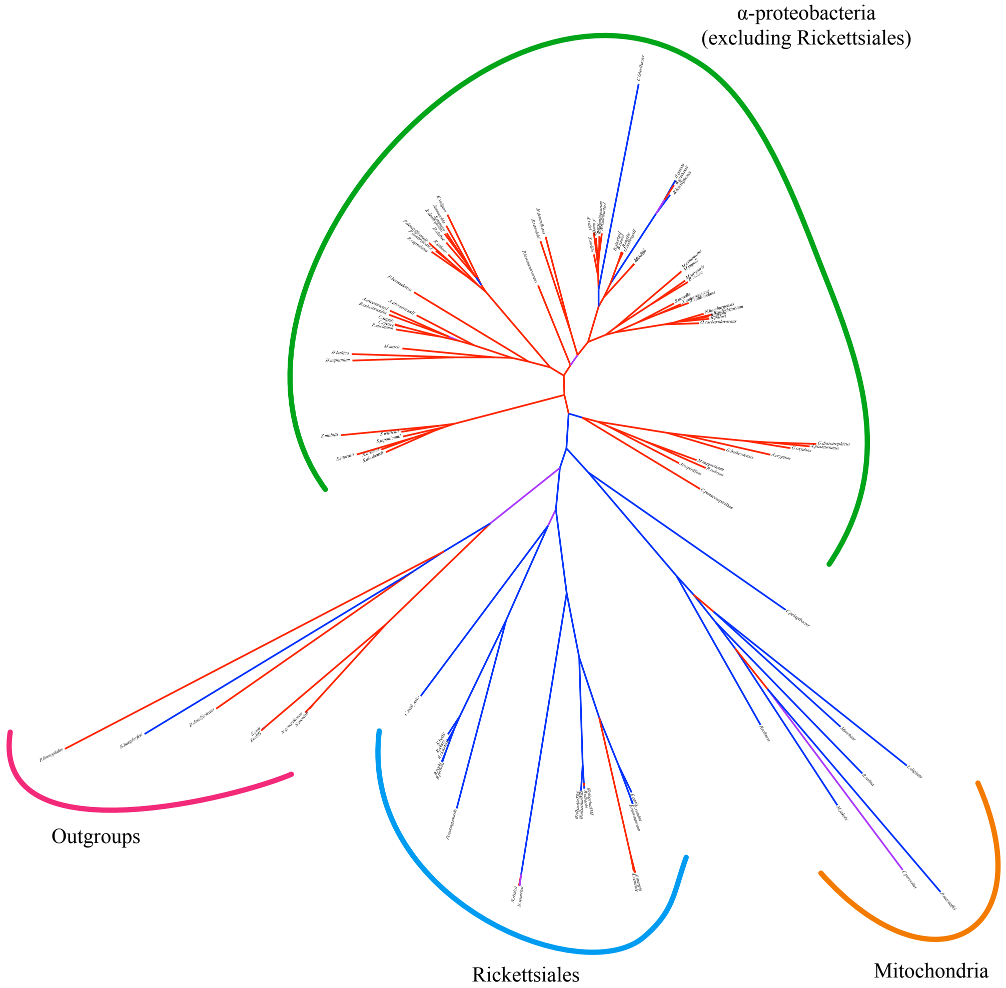

Figure 4.7: Phylogeny of the dataset using two composition vectors. Red denotes the branches with vector A, while blue denotes vector B. Purple represents branches that chose different CVs in each run.

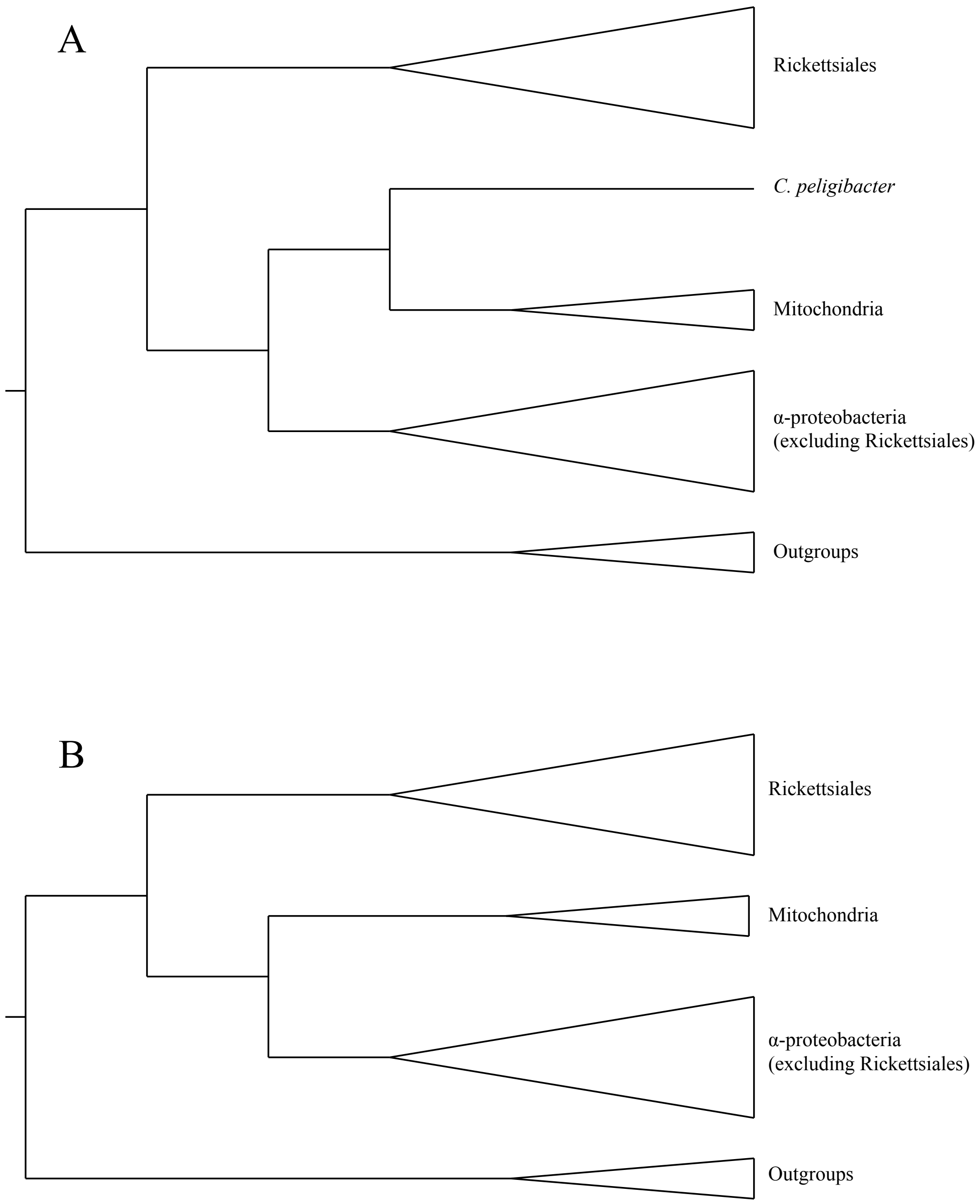

Figure 4.8: Main relationships inferred by a non-homogenous model using two CVs. Trees displayed are collapsed versions of the consensus tree inferred for both runs of this model. In both cases, the mitochondria group with the α-proteobacteria to the exclusion of the Rickettsiales.

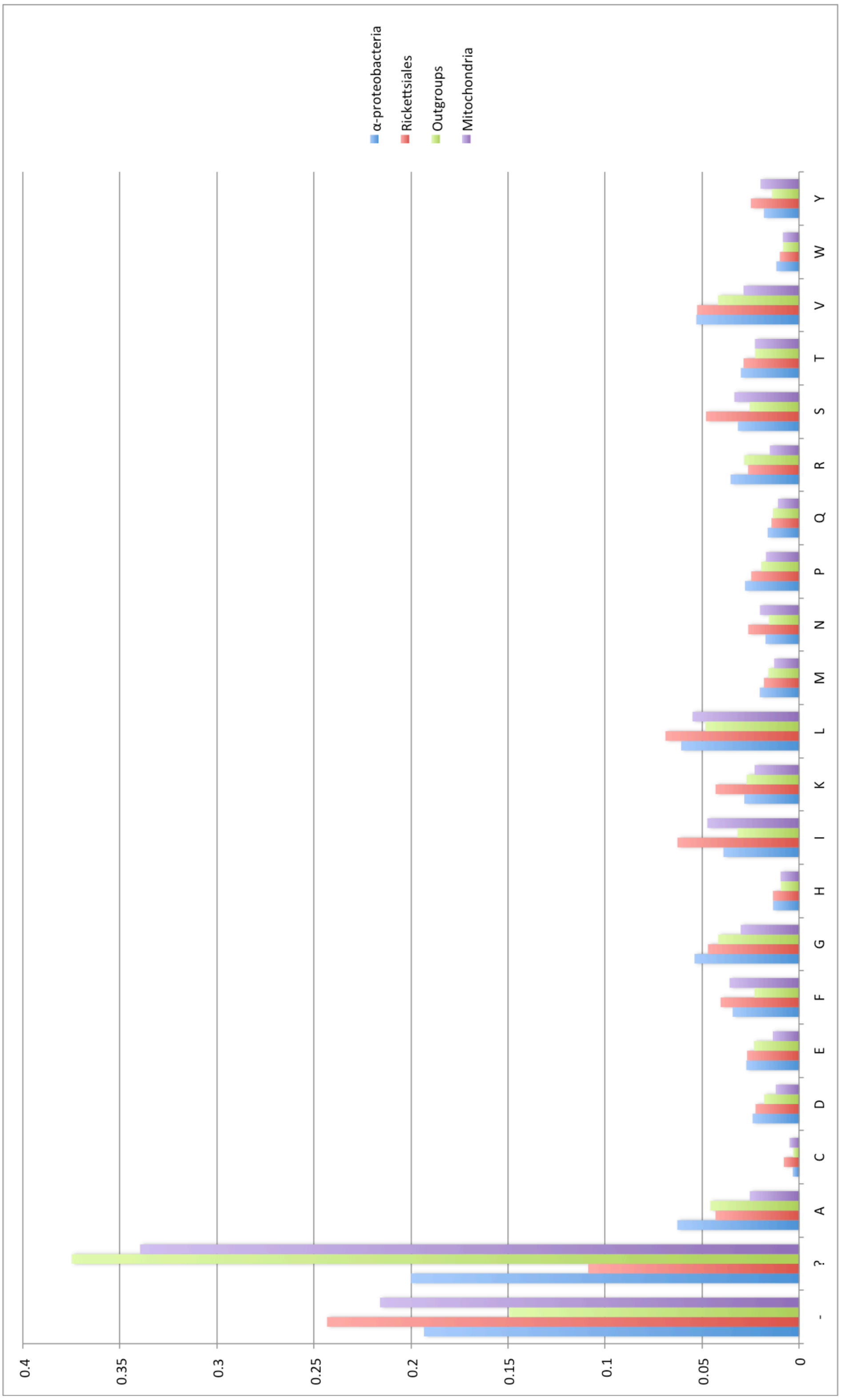

Figure 4.9: Composition of the four main groups in the data. Mean proportion of each character in each of the four groups: the α-proteobacteria, Rickettsiales, outgroups and mitochondria.

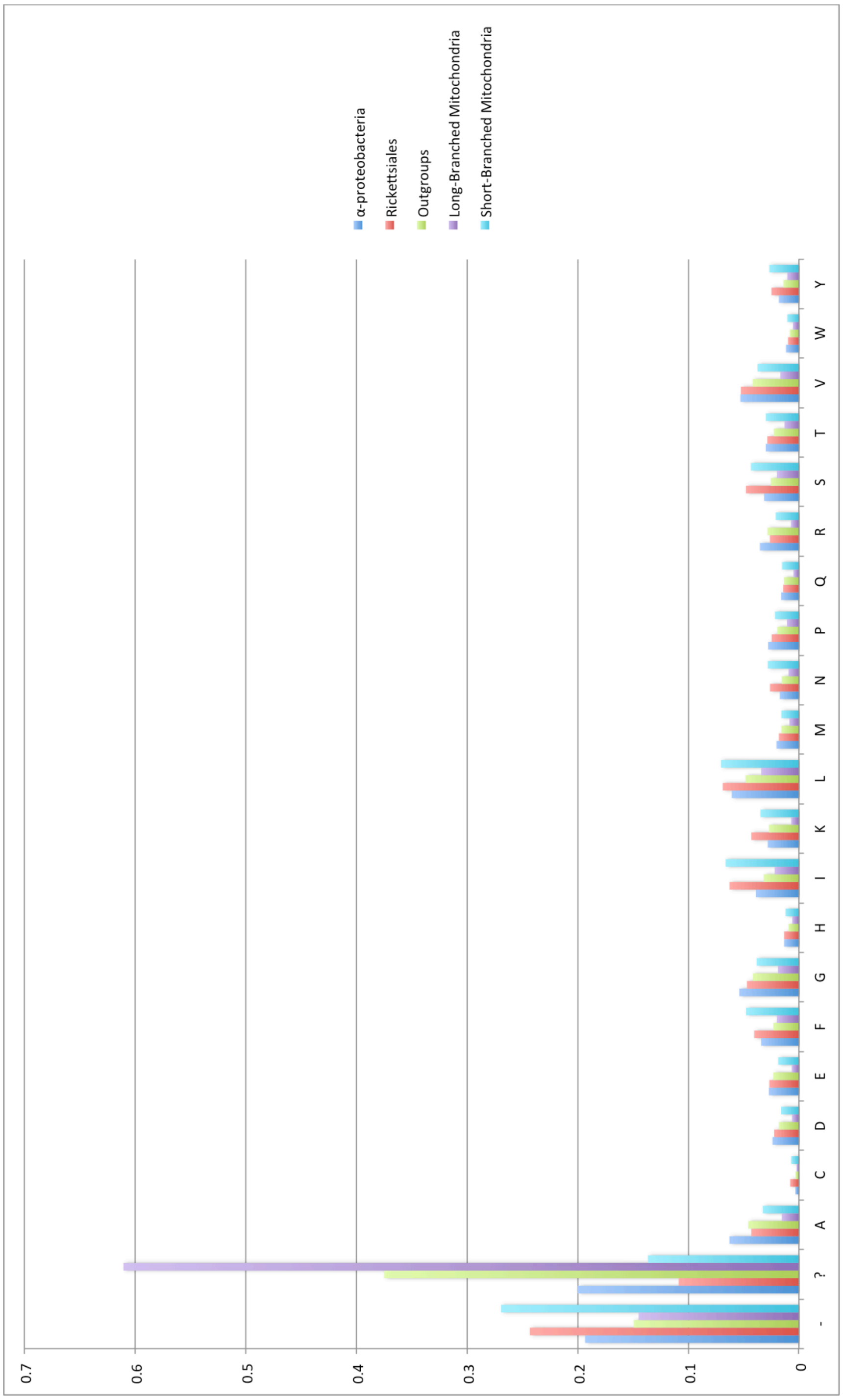

Figure 4.10: Composition within five groups of taxa. Mean proportion of each character in the groups: the α-proteobacteria, Rickettsiales, outgroups, long-branched mitochondria (*L.digitata*, *P.marneffei* and *C.porcellus*) and the short-branched mitochondria (*R.americana*, *M.polymorpha*, *R.salina* and *M.jakobiformis*). A very clear segregation in composition is seen between the two groups of mitochondrial taxa.

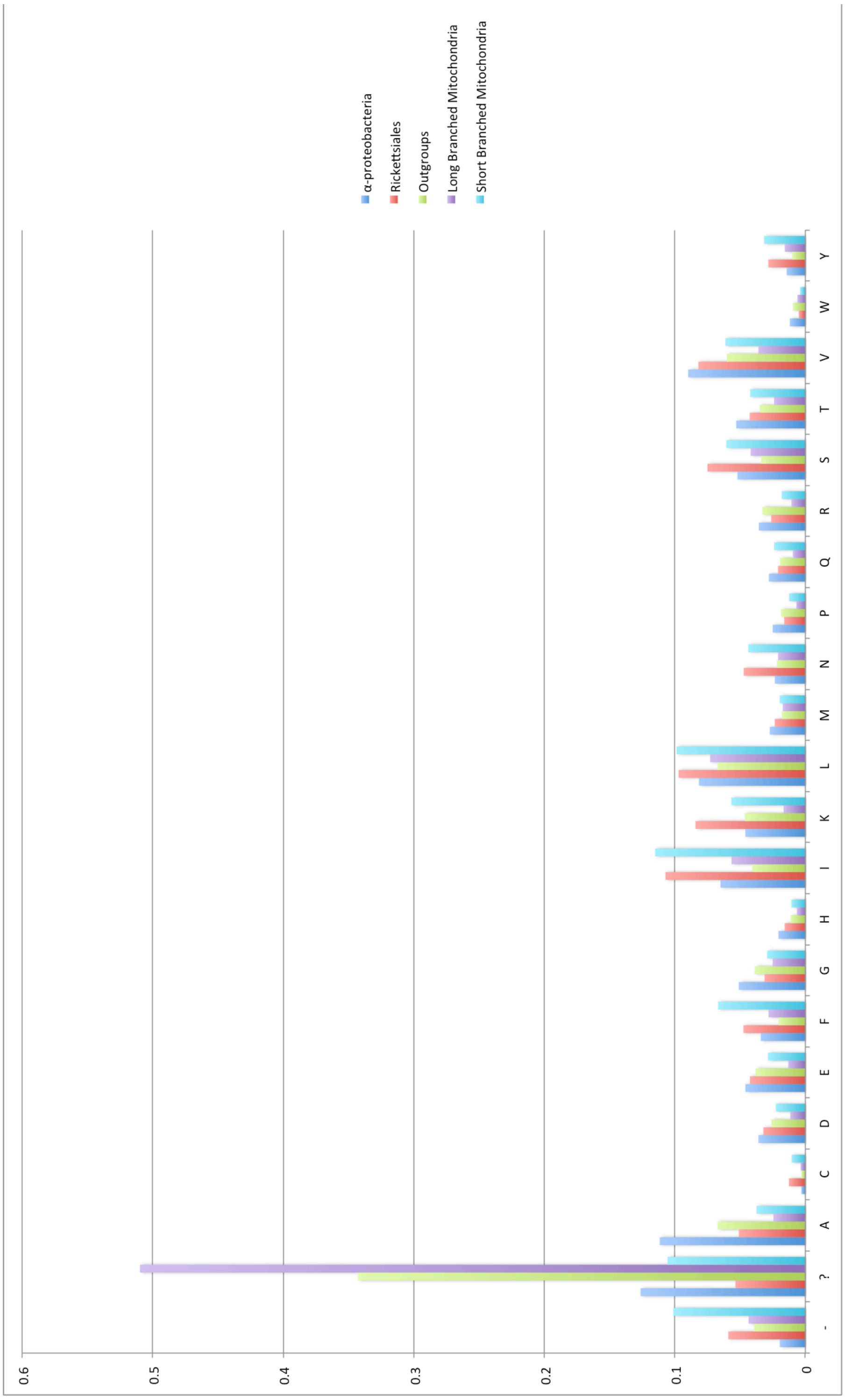

Figure 4.11: Composition of Bin10.

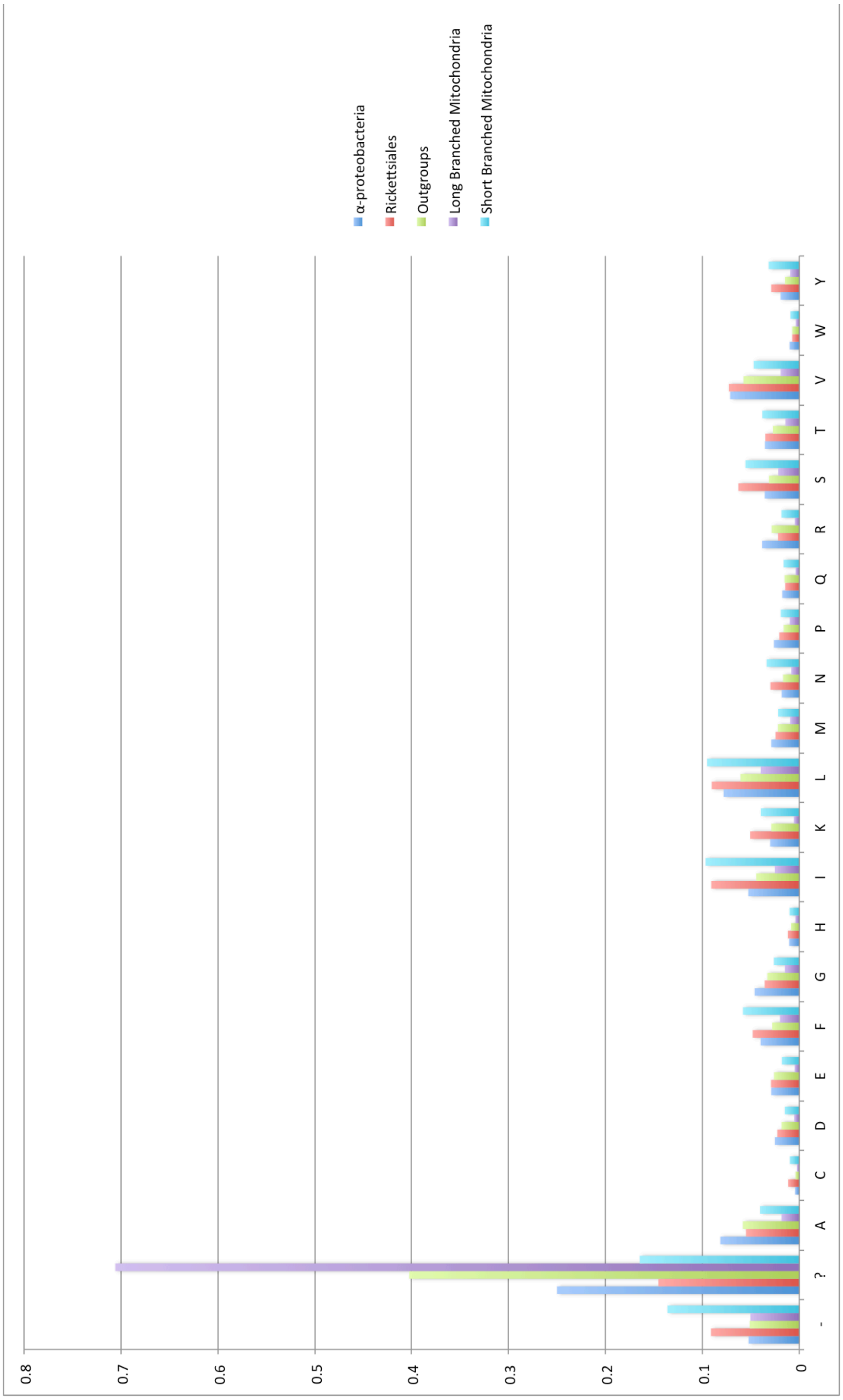

Figure 4.12: Composition of Bin9.

4.4: Discussion

Eukaryotes are not a primary lineage on earth. They are a result of a whole genome fusion between two primitive prokaryotes (Rivera and Lake, 2004). Although this concept is widely accepted, the exact archaebacterial and eubacterial parents of the eukaryotic cell have not been elucidated. While studies suggest that the eukaryotic genes displaying affinities to the archaebacteria may have arisen from either the Thermoplasmatales or the Crenarchaeota (Pisani et al., 2007, Cox et al., 2008), studies have always suggested that the mitochondria originated as an α-proteobacterium, more specifically, a member of the Rickettsiales (Andersson et al., 1998, Esser et al., 2004, Fitzpatrick et al., 2006, Williams et al., 2007). Consistently, however, datasets containing both mitochondrial and α-proteobacterial sequences have failed compositional homogeneity tests and, while this was counteracted with site removal, the method of site identification had a massive impact on the recovered topology (Fitzpatrick et al., 2006). For this reason, I wished to test the effects of the removal of sites identified by *tiger*, but also to apply compositionally heterogeneous models to this problem.

The dataset as a whole was analysed using the CAT model, producing a grouping of the mitochondria with the Rickettsiales, although parts of the tree were unresolved. By removing sites that occurred in *tiger*'s Bin10, the topology remained largely the same, but the resolution greatly improved. Given these datasets under the CAT model, the traditional topology is well supported. Both analyses, however, displayed a clear distinction in branch lengths: the Rickettsiales, mitochondria and the outgroups displayed greater branch lengths than the rest of the α-proteobacteria. To ensure that the grouping of the mitochondria with the Rickettsiales wasn't an LBA, the mitochondrial taxa were split into two groups: the short-branched taxa and the long-branched taxa. If LBA were creating this grouping, we would have expected to see greater support when using only the long branched mitochondrial sequences versus using only the short branched mitochondria. Rather surprisingly, using the long branched mitochondria alone did not improve support for the mitochondria/ Rickettsiales grouping, but changed the topology entirely. A reasonable explanation for this result is compositional heterogeneity.

Heterogeneity in the composition of taxa within a dataset can cause a phenomenon known as compositional attraction (section 1.5.2). This may be accounted for by using models that allow the base compositions of the taxa to vary (Foster, 2004). By applying these modelling methods to data, the attraction of compositionally similar taxa can be weakened. When applied to the data at hand, an important feature of the taxa became apparent: for the most part, the mitochondria share similar patterns of composition to the Rickettsiales. With two composition vectors to choose from, the majority of the mitochondria and the Rickettsiales choose one CV, while most of the outgroup and α-proteobacterial sequences choose the other (Figure 4.7). This suggests that all previous inferences, wherein the mitochondria are placed as a sister group with the Rickettsiales, may have been due to a compositional attraction. Closer analysis of the compositions of the four main groups in the dataset (α-proteobacteria, Rickettsiales, outgroups and mitochondria) show that for many characters, the Rickettsiales and the mitochondria share similarly higher (or lower) proportions of the characters when compared to the other two groups. This leads to the question: “Why did the long branched mitochondria group with the α-proteobacteria despite compositional attraction?”. Again, by inspecting the proportion of each character present in each group of organisms, the answer becomes apparent; this time, however, the long and short branched mitochondria were treated as separate groups. The long branched mitochondria, which show the mitochondria grouping with the α-proteobacteria, display drastically different compositions to their short branched relatives. This suggests that the selection of mitochondrial sequences and their compositions dictate where they will place within the α-proteobacterial tree. Although removal of sites identified by *tiger* was previously shown to negate the effects of compositional bias, that was not the case in this analysis. Analysis of each bin revealed that *tiger* is isolating signals as support for the mitochondria/Rickettsia grouping as different bins display varying support for the clade. Study of the effects of further bin removal on topology are warranted.

This study has advanced our understanding of the features that influence our ideas about the origins of the mitochondria. It is clear that there is an interplay between phylogenetic and non-phylogenetic signals, most prominent of which is compositional attraction. Past work has shown that dense taxon sampling can often help to improve difficult phylogenetic reconstructions, so perhaps further sampling of mitochondrial sequences would elucidate the, clearly, mixed signals present in our seven mitochondria. Furthermore, the number of genes used could be expanded to include mitochondria associated genes as well as mitochondria encoded genes (Cotton and McInerney, 2010). With the potential for dataset expansion and continuous methodological improvements, the origin of the mitochondria may be elucidated soon.

Chapter 5 - General Discussion

This thesis follows the developmental process of a new method, from the conception of an idea, through the implementation of that idea and its applications to both simulated and real world datasets. It is this very process that drives the scientific world. While ideas are the entire foundation for innovation, they cannot progress without forming a means for testing their validity, nor without data upon which to test them. Test data, in particular simulated data, plays a pivotal role in the development of a new method as, in order to test whether the method is producing the correct results, it must first be known what results to expect. This may only be possible when all of the features of a dataset are known with confidence; this is the main desirable characteristic of simulated data. Only when the idea has been validated and tested may it be used to infer hypotheses regarding unknown problems. In this thesis, I present the complete developmental process of a new method for site rate identification, TIGER.

In chapter two, I presented the premise for TIGER. Although many methods exist to identify site specific evolutionary rates, many of them are based on a starting tree, producing results highly dependant on the chosen starting tree (Cummins and McInerney, 2011). The TIGER method avoids this potential bias by remaining completely independent of trees. This method, based largely around that of LeQuesne (1989), Wilkinson (1998) and Pisani (2004), encompasses the idea that it is possible to score sites based on how much they agree with the other sites in the data matrix. Certainly, in the context of molecular data, these scores have been shown to reflect the rate at which a site evolves. This was proven when TIGER, not only detecting differing levels of ASRV as measured by the gamma shape parameter, but also detecting the subtle multi-modality of the data introduced by the use of gamma categories during the simulation process. Additionally, the TIGER scores emulate the site specific likelihood scores on the true tree, while remaining completely oblivious to any tree.

As confidence grew regarding TIGER's ability to assign meaningful scores to sites, the effects of removal of sites identified by TIGER was explored. Firstly, an exhaustive analysis of a simulated dataset was carried out. This is a total evidence approach, as the complete set of characteristics of the data was known and all possible trees for that data were analysed. For this data, it was shown that sequential removal of sets of rapidly-evolving characters resulted in two desirable effects; firstly, the fit of the data to trees closer to the true tree improved, but also, the fit of the data to the trees that are most unlike the true tree got progressively worse. Given this, we began to apply the method to real problems, such as resolving deep divergences or compositional attraction.

Often, when rapid cladogenesis occurs in deep branches of a phylogeny, these branches become difficult to resolve, resulting in polytomies. Through simulation of this problem, it was shown that removing the sites that contribute most to the obfuscation of the signal, as identified by TIGER, can help to resolve these types of deep branches. This result is in contrast to those obtained when a tree-based method of removing noisy sites, as little improvement was seen over 100 simulations of a deeply divergent tree. As TIGER proved its efficacy in all simulations, further testing was carried out on some empirical dataset.

The ribosomal RNA dataset is a well studied one, so many of the features of the data were known previous to analysis (Embley et al., 1993). The most prominent feature of this dataset is the severe compositional bias causing most phylogenetic methods to erroneously group the thermophiles together. This dataset was used to test TIGER's ability to remove these convergent signals. Pleasingly, removal of sites dictated by TIGER reversed the effects of compositional attraction in this dataset. However, additional analysis of the performance of tree-based methods produced worrying results. These methods are inherently biased towards their starting tree, which is, in general, the first tree the method constructs. This means that removing sites using these methods simply improves the support for the initial tree. This is further shown to be the case using a primate mitochondrial dataset (Hayasaka et al., 1988). While parsimony identifies two trees as being equally good, reweighting of characters based on each topology simply results in improved support for the topology used. This exemplifies the cyclic nature of this type of analysis; tree based methods will only improve support for the true tree if the true tree is used to identify the rapidly evolving characters. If, however, the true tree is known with enough confidence to use it in this way, then there should be no need to further improve the support.

Although TIGER has been repeatedly shown to produce desirable results, easily applying it to both large and many datasets was a clear issue. As with many new methods in the area of phylogenetics, a software implementation (*tiger*) became essential. Several features of *tiger* are dedicated to both data exploration and ease of data customisation. Although the method was developed and tested with molecular sequences in mind, during the software development, it became apparent that the method could be applied to any multi-state character matrix and, furthermore, the meaning of the rates are context dependent. For example, in the context of molecular data, the score reflected the evolutionary rate of a given site, and a low score represented little signal in common with the rest of the data, making it a rapidly evolving site. Given a binary matrix of gene presence or absence, *tiger* will rank sites from genes that are universally distributed down to those that display a distribution unlike the rest of the genes. By identifying genes like this, *tiger* may emulate two types of HGT identification methods. The first of these involves identifying genes that have an evolutionary history that is different from the species as a whole (Doolittle, 1999). This involves building both a reliable species tree and numerous gene trees and performing comparisons between the topology of every gene tree to the species tree. With so many steps, this process can become very time consuming and, additionally, the method is very dependent on the species phylogeny. The second method involves inferring gene loss and gain events to explain the patchy distributions of genes (Dagan and Martin, 2007). Using *tiger* to identify genes with the most disparate distribution could speed up this process. Unfortunately, in practice, the dataset analysed was under the influence of a bias caused by three of the seven genomes used displaying significant genome reduction. This caused an attraction between the three taxa, which became the most prominent signal in the dataset. *Tiger*, therefore, only identified sites that disagreed with this signal (the main signal in the data). Perhaps using a different taxon set would produce the results we expected. Despite this, it should be noted that *tiger* did still separate all of the signals into separate bins, from most frequently seen (Bin4) to least frequent (Bin10).

The final stage in the evolution of the TIGER method was its application to real data with an unknown phylogeny. For this, the placement of the mitochondria within the α-proteobacteria was analysed. This problem has been studied a number of times (Andersson et al., 1998, Esser et al., 2004, Fitzpatrick et al., 2006, Williams et al., 2007), but pervasive compositional heterogeneity has always been a problem in datasets containing both mitochondrial and α-proteobacterial sequences. We investigated the effects of site removal using *tiger* on this dataset, along with models to account for heterogeneity within the sequences. Under a model capable of accounting for site rate variation (CAT) and using variations of the dataset (Table 4.1), two topologies were obtained: one grouping the mitochondria with the Rickettsiales, the traditional topology, and one grouping the mitochondria with the α-proteobacteria. Most interestingly, when different subsets of the mitochondria were used, different topologies were inferred, meaning that the topology is highly dependant on the taxonomic sampling of the mitochondria. In my dataset, this phenomenon is down to compositional differences in the mitochondria, perhaps due to the accelerated divergence rate of mitochondrial genes (Gray et al., 1999). So, as composition can differ quite drastically between groups of mitochondria, some are compositionally similar to the Rickettsiales and some are more similar to the other α-proteobacteria. These differential composition similarities dictate where the mitochondria place on the tree, suggesting that, without adequate modelling, what we see are merely composition trees. Using compositionally heterogeneous models, the mitochondria group with the α-proteobacteria. Allowing the taxa to, effectively, chose between two different composition vectors results in the majority of the mitochondria and Rickettsiales choosing one vector, while most of the outgroups and α-proteobacteria chose the other. This further supports the notion that phylogenies portraying the mitochondria as sister group to the Rickettsiales may have been a compositional attraction.

5.1 Future Work

While this body of work forms a cohesive unit, several analyses could expand on the results. Firstly, some improvements could be made to the *tiger* process. Development of a "stopping rule" for site removal would make removal of sites less arbitrary and more accurate (Goremykin et al., 2010), but defining sites that contain useful signal and those containing noise is not an easy process, as the boundaries between the two are not well defined. In a purely computational framework, several avenues could be pursued to improve the performance of *tiger* as a software. Parallelisation or the use of the highly efficient C programming language for the most computationally intensive parts of the calculation could certainly contribute to this. With regards the placement of sites into bins, a more informative approach could be taken. As it stands, *tiger* places sites in bins in a relatively arbitrary nature (by splitting the range of rates into equal partitions). However, by clustering sites based on their scores the contents of each bin could be more meaningful and situations where a single, coherent signal is erroneously split between two bins could be avoided.

Secondly, the role that *tiger* could play outside the area of evolutionary rate identification warrants further investigation. As previously mentioned (section 3.1), the meaning of a TIGER score changes based on context and this was explored when using *tiger* to identify HGT events. Although all applications of the method that are discussed in this thesis are in a biological context, I believe that it may have applications in a wide range of fields. Any area that uses matrices to represent data could benefit from using *tiger* to identify and categorise the signals present. With regards to the identification of HGT using *tiger*, some methodological improvements could be made. Using a larger dataset and an exhaustive comparison of *tiger* rates to the occurrence of HGT events in gene families may produce a more definitive answer on the matter. Similarly, exploring other methods for identifying gene families could produce better results.

Lastly, the effects of additional site removal on the placement of the mitochondria within the α-proteobacteria should be investigated. The TIGER method was shown to work well on reversing the effects of compositional attraction on the *Thermus* dataset (section 2.3.5), but this is not seen when it was applied to the mitochondrial dataset. Removal of additional categories of sites both with and without a compositionally heterogeneous model may produce different topologies and should be studied. Additionally, the effects of broader taxonomic sampling within the mitochondria may shed some light on the compositional complexity within this group and elucidate their true origins.

Chapter 6 - Bibliography

ABASCAL, F., ZARDOYA, R. & POSADA, D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics,* 21**,** 2104.

AGUINALDO, A. M. A., TURBEVILLE, J. M., LINFORD, L. S., RIVERA, M. C., GAREY, J. R., RAFF, R. A. & LAKE, J. A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature,* 387**,** 489-493.

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. 1973. Springer Verlag, 267-281.

ALTMAN, R. 1890. *Die Elementaroganismen und ihre Beziehungen zur den Zellen.,* Leipzig, Germany, Verlag von Veit.

ALTSCHUL, S. F., MADDEN, T. L., SCH‰FFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research,* 25**,** 3389.

ALVAREZ-PONCE, D. & MCINERNEY, J. O. 2011. The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. *Genome Biology and Evolution,* 3**,** 782.

ANDERSSON, S. G. E., ZOMORODIPOUR, A., ANDERSSON, J. O., SICHERITZ-PONTÉN, T., ALSMARK, U. C. M., PODOWSKI, R. M., NÜSLUND, A. K., ERIKSSON, A. S., WINKLER, H. H. & KURLAND, C. G. 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature,* 396**,** 133-140.

BANDELT, H. J. & DRESS, A. W. M. 1992. A canonical decomposition theory for metrics on a finite set. *Advances in mathematics,* 92**,** 47-105.

BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon***,** 3-10.

BERNARDI, G. 1995. The human genome: organization and evolutionary history. *Annual review of genetics,* 29**,** 445-476.

BOLLBACK, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. *Molecular biology and evolution,* 19**,** 1171.

BRADLEY, R. K., ROBERTS, A., SMOOT, M., JUVEKAR, S., DO, J., DEWEY, C., HOLMES, I. & PACHTER, L. 2009. Fast statistical alignment. *PLoS computational biology,* 5**,** e1000392.

BRINKMANN, H. & PHILIPPE, H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Molecular biology and evolution,* 16**,** 817.

BROWN, J. R. 2003. Ancient horizontal gene transfer. *Nature Reviews Genetics,* 4**,** 121-132.

BROWN, J. R. & DOOLITTLE, W. F. 1997. Archaea and the prokaryote-to-eukaryote transition. *Microbiology and Molecular Biology Reviews,* 61**,** 456.

BRYANT, D. & MOULTON, V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution,* 21, 255.

CAMIN, J. H. & SOKAL, R. R. 1965. A method for deducing branching sequences in phylogeny. *Evolution*, 311-326.

CAVALLI-SFORZA, L. L. & EDWARDS, A. W. F. 1967. Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics,* 19, 233.

CHARLESTON, M. A. 2001. Hitch-hiking: A parallel heuristic search strategy, applied to the phylogeny problem. *Journal of Computational Biology,* 8, 79-91.

COTTON, J. A. & MCINERNEY, J. O. 2010. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proceedings of the National Academy of Sciences,* 107, 17252.

COX, C. J., FOSTER, P. G., HIRT, R. P., HARRIS, S. R. & EMBLEY, T. M. 2008. The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences,* 105, 20356.

CREEVEY, C. & MCINERNEY, J. 2005. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics,* 21, 390.

CUMMINS, C. A. & MCINERNEY, J. O. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Systematic Biology,* 60, 833-844.

DAGAN, T. & MARTIN, W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences,* 104, 870.

DARWIN, C. 1859. On the origin of the species by natural selection.

DELSUC, F., BRINKMANN, H. & PHILIPPE, H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature reviews. Genetics,* 6, 361.

DEPPENMEIER, U., JOHANN, A., HARTSCH, T., MERKL, R., SCHMITZ, R. A., MARTINEZ-ARIAS, R., HENNE, A., WIEZER, A., B‰UMER, S. & JACOBI, C. 2002. The genome of Methanosarcina mazei: Evidence for lateral gene transfer between bacteria and archaea. *Journal of molecular microbiology and biotechnology,* 4, 453.

DOOLITTLE, W. F. 1999. Phylogenetic classification and the universal tree. *Science,* 284, 2124.

DOOLITTLE, W. F. & BROWN, J. R. 1994. Tempo, mode, the progenote, and the universal root. *Proceedings of the National Academy of Sciences,* 91, 6721.

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

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* 32, 1792.

EDWARDS, A. W. F. & SFORZA, C. L. L. 1963. The reconstruction of evolution. *Heredity,* 18.

EFRON, B. 1979. Bootstrap methods: another look at the jackknife. *The annals of statistics*, 1-26.

EISEN, J. A. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current opinion in genetics & development,* 10, 606-611.

EMBLEY, T., THOMAS, R. & WILLIAMS, R. 1993. Reduced thermophilic bias in the 16S rDNA sequence from Thermus ruber provides further support for a relationship between Thermus and Deinococcus. *Systematic and applied microbiology,* 16, 25-29.

EMBLEY, T. M. & MARTIN, W. 2006. Eukaryotic evolution, changes and challenges. *Nature,* 440, 623-630.

ENRIGHT, A. J., VAN DONGEN, S. & OUZOUNIS, C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research,* 30, 1575.

ESSER, C., AHMADINEJAD, N., WIEGAND, C., ROTTE, C., SEBASTIANI, F., GELIUS-DIETRICH, G., HENZE, K., KRETSCHMANN, E., RICHLY, E. & LEISTER, D. 2004. A genome phylogeny for mitochondria among - proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular biology and evolution,* 21, 1643.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. *Systematic Biology,* 18, 374.

FELSENSTEIN, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 240-249.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology,* 27, 401.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution,* 17, 368-376.

FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *American Naturalist*, 1-15.

FISCHER, W. M. & PALMER, J. D. 2005. Evidence from small-subunit ribosomal RNA sequences for a fungal origin of Microsporidia. *Molecular phylogenetics and evolution,* 36, 606-622.

FITCH, W. M. & MARGOLIASH, E. 1967. Construction of phylogenetic trees. *Science,* 155, 279-284.

FITCH, W. M. & MARKOWITZ, E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics,* 4, 579-593.

FITZPATRICK, D. A., CREEVEY, C. J. & MCINERNEY, J. O. 2006. Genome Phylogenies Indicate a Meaningful alpha-Proteobacterial Phylogeny and Support a Grouping of the Mitochondria with the Rickettsiales. *Molecular Biology and Evolution,* 23, 74-85.

FOSTER, P. G. 2004. Modeling compositional heterogeneity. *Systematic Biology,* 53, 485.

FOSTER, P. G. & HICKEY, D. A. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of molecular evolution,* 48**,** 284-290.

FREEMAN, V. J. 1951. Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. *Journal of bacteriology,* 61**,** 675.

GALAGAN, J. E., NUSBAUM, C., ROY, A., ENDRIZZI, M. G., MACDONALD, P., FITZHUGH, W., CALVO, S., ENGELS, R., SMIRNOV, S. & ATNOOR, D. 2002. The genome of M. acetivorans reveals extensive metabolic and physiological diversity. *Genome Research,* 12**,** 532.

GALTIER, N. & GOUY, M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular biology and evolution,* 15**,** 871.

GALTIER, N. & LOBRY, J. 1997. Relationships between genomic G+ C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of molecular evolution,* 44**,** 632-636.

GAUT, B. S. & LEWIS, P. O. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular biology and evolution,* 12**,** 152.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. *Journal of molecular evolution,* 36**,** 182-198.

GOLDMAN, N. & YANG, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution,* 11**,** 725.

GOREMYKIN, V., NIKIFOROVA, S. & BININDA-EMONDS, O. 2010. Automated removal of noisy data in phylogenomic analyses. *J. Mol. Evol,* 71**,** 319-331.

GRAY, M. W., BURGER, G. & LANG, B. F. 1999. Mitochondrial evolution. *Science,* 283**,** 1476.

GRAY, M. W., SANKOFF, D. & CEDERGREN, R. J. 1984. On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucleic acids research,* 12**,** 5837.

GUINDON, S., DUFAYARD, J. F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology,* 59**,** 307.

GUINDON, S. & GASCUEL, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology,* 52**,** 696.

GUPTA, R. S. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiology and Molecular Biology Reviews,* 62**,** 1435.

HAECKEL, E. H. P. A. 1866. *Generelle Morphologie der Organismen: allgemeine Grundz¸ge der organischen Formen-Wissenschaft, mechanisch begr¸ndet durch die von Charles Darwin reformirte Descendenz-Theorie*, G. Reimer.

HALARY, S., LEIGH, J. W., CHEAIB, B., LOPEZ, P. & BAPTESTE, E. 2010. Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences,* 107**,** 127.

HANSMANN, S. & MARTIN, W. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *International journal of systematic and evolutionary microbiology,* 50**,** 1655.

HASEGAWA, M., KISHINO, H. & SAITOU, N. 1991. On the maximum likelihood method in molecular phylogenetics. *Journal of molecular evolution,* 32**,** 443-445.

HASEGAWA, M., KISHINO, H. & YANO, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution,* 22**,** 160-174.

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

HAYASAKA, K., GOJOBORI, T. & HORAI, S. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular biology and evolution,* 5**,** 626.

HEJNOL, A., OBST, M., STAMATAKIS, A., OTT, M., ROUSE, G. W., EDGECOMBE, G. D., MARTINEZ, P., BAGUÒ‡, J., BAILLY, X. & JONDELIUS, U. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences,* 276**,** 4261.

HENDY, M. D. & PENNY, D. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Biology,* 38**,** 297.

HILLIS, D. M. 1996. Inferring complex phytogenies. *Nature,* 383**,** 130-131.

HIRT, R. P., LOGSDON, J. M., HEALY, B., DOREY, M. W., DOOLITTLE, W. F. & EMBLEY, T. M. 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences,* 96**,** 580.

HOLTON, T. A. & PISANI, D. 2010. Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single-and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biology and Evolution,* 2**,** 310.

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

HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R. & BOLLBACK, J. P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 294**,** 2310.

HUELSENBECK, J. P. & SUCHARD, M. A. 2007. A nonparametric method for accommodating and testing across-site rate variation. *Systematic Biology,* 56**,** 975.

HUGHES, S., ZELUS, D. & MOUCHIROUD, D. 1999. Warm-blooded isochore structure in Nile crocodile and turtle. *Molecular biology and evolution,* 16**,** 1521.

JUKES, T. & CANTOR, C. 1969. Evolution of protein molecules. *In:* MUNRO, H. (ed.) *Mammalian Protein Metabolism.* Academic Press, New York.

KEANE, T., CREEVEY, C., PENTONY, M., NAUGHTON, T. & MCLNERNEY, J. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology,* 6**,** 29.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution,* 16**,** 111-120.

KLUGE, A. G. & FARRIS, J. S. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Biology,* 18**,** 1.

KOLACZKOWSKI, B. & THORNTON, J. W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature,* 431**,** 980-984.

KOONIN, E. V. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome biology,* 11**,** 209.

KOSTKA, M., UZLIKOVA, M., CEPICKA, I. & FLEGR, J. 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. *BMC bioinformatics,* 9**,** 341.

KUHNER, M. K. & FELSENSTEIN, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution,* 11**,** 459.

KUNIN, V. & OUZOUNIS, C. A. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Research,* 13**,** 1589.

KURLAND, C., CANBACK, B. & BERG, O. G. 2003. Horizontal gene transfer: a critical view. *Proceedings of the National Academy of Sciences,* 100**,** 9658.

KURLAND, C. G. 2000. Something for everyone. *EMBO reports,* 1**,** 92-95.

LAKE, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences,* 91**,** 1455.

LAKE, J. A. & RIVERA, M. C. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Molecular biology and evolution,* 21**,** 681.

LAMARCK, J. B. 1809. *Philosophie zoologique*, C. Martins.

LANAVE, C., PREPARATA, G., SACONE, C. & SERIO, G. 1984. A new method for calculating evolutionary substitution rates. *Journal of molecular evolution,* 20**,** 86-93.

LARTILLOT, N., BRINKMANN, H. & PHILIPPE, H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology,* 7**,** S4.

LARTILLOT, N., LEPAGE, T. & BLANQUART, S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics,* 25**,** 2286.

LARTILLOT, N. & PHILIPPE, H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution,* 21**,** 1095.

LE QUESNE, W. J. 1969. A method of selection of characters in numerical taxonomy. *Systematic Biology,* 18**,** 201.

LE QUESNE, W. J. 1989. The normal deviate test of phylogenetic value of a data matrix. *Systematic Zoology***,** 51-54.

LE, S. Q. & GASCUEL, O. 2008. An improved general amino acid replacement matrix. *Molecular biology and evolution,* 25**,** 1307.

LEWIS, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology,* 50**,** 913.

LOCKHART, P. J., STEEL, M. A., HENDY, M. D. & PENNY, D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular biology and evolution,* 11**,** 605.

LÖYTYNOJA, A. & GOLDMAN, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science,* 320**,** 1632.

MADDISON, W. P. & MADDISON, D. R. 1992. MacClade: analysis of phylogeny and character evolution. *Evolution (PMBD, 185908476).*

MAIDAK, B. L., OLSEN, G. J., LARSEN, N., OVERBEEK, R., MCCAUGHEY, M. J. & WOESE, C. R. 1996. The ribosomal database project (RDP). *Nucleic acids research,* 24**,** 82.

MARGULIS, L. 1981. *Symbiosis in cell evolution: life and its environment on the early Earth,* New York, Freeman.

MARGUSH, T. & MCMORRIS, F. R. 1981. Consensus n-trees. *Bulletin of Mathematical Biology,* 43**,** 239-244.

MARTIN, W., HOFFMEISTER, M., ROTTE, C. & HENZE, K. 2001. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. *Biological chemistry,* 382**,** 1521-1539.

MCDANIEL, L. D., YOUNG, E., DELANEY, J., RUHNAU, F., RITCHIE, K. B. & PAUL, J. H. 2010. High frequency of horizontal gene transfer in the oceans. *Science,* 330**,** 50.

MCINERNEY, J. O., PISANI, D., BAPTESTE, E. & O'CONNELL, M. J. 2011. The public goods hypothesis for the evolution of life on Earth. *Biology Direct,* 6**,** 41.

MEACHAM, C. A. 1994. Phylogenetic relationships at the basal radiation of angiosperms: further study by probability of character compatibility. *Systematic Botany***,** 506-522.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics,* 21**,** 1087.

MOOERS, A. O. & HOLMES, E. C. 2000. The evolution of base composition and phylogenetic inference. *Trends in Ecology & Evolution,* 15**,** 365-369.

MOSZER, I., ROCHA, E. P. C. & DANCHIN, A. 1999. Codon usage and lateral gene transfer in Bacillus subtilis. *Current opinion in microbiology,* 2**,** 524-528.

NAKAMURA, Y., ITOH, T., MATSUDA, H. & GOJOBORI, T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature genetics,* 36**,** 760-766.

NEDELCU, A. & LEE, R. 1998. *In:* ROCHAIX, J. D. (ed.) *The molecular biology of chloroplasts and mitochondria in Chlamydomonas.* Springer Netherlands.

NELSON, K. E., CLAYTON, R. A., GILL, S. R., GWINN, M. L., DODSON, R. J., HAFT, D. H., HICKEY, E. K., PETERSON, J. D., NELSON, W. C. &

KETCHUM, K. A. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature,* 399**,** 323-329.

OCHMAN, H., LAWRENCE, J. G. & GROISMAN, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature,* 405**,** 299-304.

OLSEN, G. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. 1987. Cold Spring Harbor Laboratory Press, 825.

OLSEN, G., PRACHT, S. & OVERBEEK, R. 1998. DNArates. 1.1 ed.

OLSEN, G. J., MATSUDA, H., HAGSTROM, R. & OVERBEEK, R. 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Computer applications in the biosciences: CABIOS,* 10**,** 41.

OLSEN, G. J. & WOESE, C. 1993. Ribosomal RNA: a key to phylogeny. *The FASEB journal,* 7**,** 113.

OWEN, R. 1843. *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals,* London, Longman, Brown, Green and Longman.

PACE, N. R., OLSEN, G. J. & WOESE, C. R. 1986. Ribosomal RNA phylogeny and the primary lines of evolutionary descent. *Cell,* 45**,** 325.

PHILIPPE, H., SNELL, E. A., BAPTESTE, E., LOPEZ, P., HOLLAND, P. W. H. & CASANE, D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Molecular biology and evolution,* 21**,** 1740.

PHILIPPE, H., ZHOU, Y., BRINKMANN, H., RODRIGUE, N. & DELSUC, F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology,* 5**,** 50.

PISANI, D. 2004. Identifying and Removing Fast-Evolving Sites Using Compatibility Analysis: An Example from the Arthropoda. *Systematic Biology,* 53**,** 978-989.

PISANI, D., COTTON, J. A. & MCINERNEY, J. O. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular biology and evolution,* 24**,** 1752.

POE, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. *Systematic Biology,* 52**,** 423-428.

POLLOCK, D. D., ZWICKL, D. J., MCGUIRE, J. A. & HILLIS, D. M. 2002. Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology,* 51**,** 664.

PUIGBO, P., GARCIA-VALLVE, S. & MCINERNEY, J. O. 2007. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics,* 23**,** 1556.

RAGAN, M. A. 1992. Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. *Biosystems,* 28**,** 47-55.

RAMBAUT, A. & GRASS, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences: CABIOS,* 13**,** 235.

RIVERA, M. C., JAIN, R., MOORE, J. E. & LAKE, J. A. 1998. Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences,* 95**,** 6239.

RIVERA, M. C. & LAKE, J. A. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature,* 431**,** 152-155.

SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution,* 4**,** 406.

SANDERSON, M., MCMAHON, M. & STEEL, M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evolutionary Biology,* 10**,** 155.

SANKOFF, D., LEDUC, G., ANTOINE, N., PAQUIN, B., LANG, B. F. & CEDERGREN, R. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences,* 89**,** 6575.

SASSERA, D., LO, N., EPIS, S., D'AURIA, G., MONTAGNA, M., COMANDATORE, F., HORNER, D., PERETÛ, J., LUCIANO, A. M. & FRANCIOSI, F. 2011. Phylogenomic evidence for the presence of a flagellum and cbb3 oxidase in the free-living mitochondrial ancestor. *Molecular biology and evolution*.

SCHMIDT, H. A., STRIMMER, K., VINGRON, M. & VON HAESELER, A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics,* 18**,** 502.

SCHWARZ, G. 1978. Estimating the dimension of a model. *The annals of statistics***,** 461-464.

SLOWINSKI, J. B. & PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology,* 48**,** 814.

SNEL, B., BORK, P. & HUYNEN, M. A. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research,* 12**,** 17.

SNODGRASS, R. E. 1938. *Evolution of the Annelida, Onychophora and Arthropoda*, The Smithsonian institution.

STAMATAKIS, A. An efficient program for phylogenetic inference using simulated annealing. 2005. IEEE, 8 pp.

STAMATAKIS, A., LUDWIG, T. & MEIER, H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics,* 21**,** 456.

STEWART, C. B. 1993. The powers and pitfalls of parsimony.

SUCHARD, M. A., KITCHEN, C. M. R., SINSHEIMER, J. S. & WEISS, R. E. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology,* 52**,** 649.

SUEOKA, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America,* 48**,** 582.

SULLIVAN, J. & SWOFFORD, D. L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution,* 4**,** 77-86.

SWOFFORD, D. L. 2003. PAUP*: phylogenetic analysis using parsimony. 4.0 b10 ed.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research,* 22**,** 4673.

THRASH, J. C., BOYD, A., HUGGETT, M. J., GROTE, J., CARINI, P., YODER, R. J., ROBBERTSE, B., SPATAFORA, J. W., RAPPÈ, M. S. & GIOVANNONI, S. J. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Scientific Reports,* 1.

TIMMIS, J. N., AYLIFFE, M. A., HUANG, C. Y. & MARTIN, W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics,* 5**,** 123-135.

TOWNSEND, J. P. 2007. Profiling phylogenetic informativeness. *Systematic Biology,* 56**,** 222.

WHEELER, W. C. 1990. Nucleic acid sequence phylogeny and random outgroups. *Cladistics,* 6**,** 363-367.

WHELAN, S., LI , P. & GOLDMAN, N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *TRENDS in Genetics,* 17**,** 262-272.

WILGENBUSCH, J. C. & SWOFFORD, D. 2003. Inferring evolutionary trees with PAUP*. *Current protocols in bioinformatics*.

WILKINSON, M. 1998. Split Support and Split Conflict Randomization Tests in Phylogenetic Inference. *Systematic Biology,* 47**,** 673-695.

WILLIAMS, K. P., SOBRAL, B. W. & DICKERMAN, A. W. 2007. A robust species tree for the Alphaproteobacteria. *Journal of bacteriology,* 189**,** 4578.

WOESE, C. R. & FOX, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences,* 74**,** 5088.

WOESE, C. R., KANDLER, O. & WHEELIS, M. L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences,* 87**,** 4576.

WOLFE, K. 2000. Robustness--it's not where you think it is. *Nature genetics,* 25**,** 3.

YANG, D., OYAIZU, Y., OYAIZU, H., OLSEN, G. J. & WOESE, C. R. 1985. Mitochondrial origins. *Proceedings of the National Academy of Sciences,* 82**,** 4443.

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

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of molecular evolution,* 39**,** 306-314.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution,* 11**,** 367-372.

ZUCKERKANDL, E. & PAULING, L. 1965. Molecules as documents of evolutionary history. *Journal of Theoretical Biology,* 8**,** 357-366.

Appendix

Table A1: A list of the 93 taxa used in this study and their taxonomy.

Species	Taxonomy
Acetobacter pasteurianus	α-proteobacteria; Rhodospirillales;
Acidiphilium cryptum	α-proteobacteria; Rhodospirillales;
Agrobacterium radiobacter	α-proteobacteria; Rhizobiales;
Agrobacterium vitis	α-proteobacteria; Rhizobiales;
Anaplasma centrale	α-proteobacteria; Rickettsiales;
Anaplasma marginale	α-proteobacteria; Rickettsiales;
Asticcacaulis excentricus	α-proteobacteria; Caulobacterales;
Azorhizobium caulinodans	α-proteobacteria; Rhizobiales;
Azospirillum	α-proteobacteria; Rhodospirillales;
Bartonella bacilliformis	α-proteobacteria; Rhizobiales;
Bartonella grahamii	α-proteobacteria; Rhizobiales;
Bartonella quintana	α-proteobacteria; Rhizobiales;
Beijerinckia indica	α-proteobacteria; Rhizobiales;
Bradyrhizobium japonicum	α-proteobacteria; Rhizobiales;
Bradyrhizobium	α-proteobacteria; Rhizobiales;
Brevundimonas subvibrioides	α-proteobacteria; Caulobacterales;
Brucella abortus	α-proteobacteria; Rhizobiales;
Brucella canis	α-proteobacteria; Rhizobiales;
Brucella melitensis	α-proteobacteria; Rhizobiales;
Candidatus Liberibacter asiaticus	α-proteobacteria; Rhizobiales;
Candidatus Midichloria mitochondrii	α-proteobacteria; Rickettsiales;

Candidatus Pelagibacter ubique	α-proteobacteria; Rickettsiales;
Candidatus Puniceispirillum marinum	α-proteobacteria; SAR116 cluster;
Caulobacter crescentus	α-proteobacteria; Caulobacterales;
Caulobacter segnis	α-proteobacteria; Caulobacterales;
Dinoroseobacter shibae	α-proteobacteria; Rhodobacterales;
Ehrlichia canis	α-proteobacteria; Rickettsiales;
Ehrlichia ruminantium	α-proteobacteria; Rickettsiales;
Erythrobacter litoralis	α-proteobacteria; Sphingomonadales;
Gluconacetobacter diazotrophicus	α-proteobacteria; Rhodospirillales;
Gluconobacter oxydans	α-proteobacteria; Rhodospirillales;
Granulibacter bethesdensis	α-proteobacteria; Rhodospirillales;
Hirschia baltica	α-proteobacteria; Rhodobacterales;
Hyphomicrobium denitrificans	α-proteobacteria; Rhizobiales;
Hyphomonas neptunium	α-proteobacteria; Rhodobacterales;
Jannaschia	α-proteobacteria; Rhodobacterales;
Ketogulonicigenium vulgare	α-proteobacteria; Rhodobacterales;
Methylobacterium extorquens	α-proteobacteria; Rhizobiales;
Magnetospirillum magneticum	α-proteobacteria; Rhodospirillales;
Maricaulis maris	α-proteobacteria; Rhodobacterales;
Mesorhizobium ciceri	α-proteobacteria; Rhizobiales;
Mesorhizobium loti	α-proteobacteria; Rhizobiales;
Methylobacterium populi	α-proteobacteria; Rhizobiales;
Methylocella silvestris	α-proteobacteria; Rhizobiales;
Neorickettsia risticii	α-proteobacteria; Rickettsiales;
Neorickettsia sennetsu	α-proteobacteria; Rickettsiales;
Nitrobacter hamburgensis	α-proteobacteria; Rhizobiales;

Novosphingobium aromaticivorans	α-proteobacteria; Sphingomonadales;
Ochrobactrum anthropi	α-proteobacteria; Rhizobiales;
Oligotropha carboxidovorans	α-proteobacteria; Rhizobiales;
Orientia tsutsugamushi	α-proteobacteria; Rickettsiales;
Paracoccus denitrificans	α-proteobacteria; Rhodobacterales;
Parvibaculum lavamentivorans	α-proteobacteria; Rhizobiales;
Parvularcula bermudensis	α-proteobacteria; Parvularculales;
Phenylobacterium zucineum	α-proteobacteria; Caulobacterales;
Rhizobium etli	α-proteobacteria; Rhizobiales;
Rhizobium leguminosarum	α-proteobacteria; Rhizobiales;
Rhizobium radiobacter	α-proteobacteria; Rhizobiales;
Rhodobacter capsulatus	α-proteobacteria; Rhodobacterales;
Rhodobacter sphaeroides	α-proteobacteria; Rhodobacterales;
Rhodomicrobium vannielii	α-proteobacteria; Rhizobiales;
Rhodopseudomonas palustris	α-proteobacteria; Rhizobiales;
Rhodospirillum rubrum	α-proteobacteria; Rhodospirillales;
Rickettsia akari	α-proteobacteria; Rickettsiales;
Rickettsia bellii	α-proteobacteria; Rickettsiales;
Rickettsia prowazekii	α-proteobacteria; Rickettsiales;
Rickettsia rickettsii	α-proteobacteria; Rickettsiales;
Rickettsia typhi	α-proteobacteria; Rickettsiales;
Roseobacter denitrificans	α-proteobacteria; Rhodobacterales;
Silicibacter pomeroyi	α-proteobacteria; Rhodobacterales;
Sinorhizobium meliloti	α-proteobacteria; Rhizobiales;
Sphingobium japonicum	α-proteobacteria; Sphingomonadales;
Sphingomonas wittichii	α-proteobacteria; Sphingomonadales;
Sphingopyxis alaskensis	α-proteobacteria; Sphingomonadales;

Starkeya novella	α-proteobacteria; Rhizobiales;
Wolbachia endosymbiont of *Drosophila melanogaster*	α-proteobacteria; Rickettsiales;
Wolbachia endosymbiont strain TRS	α-proteobacteria; Rickettsiales;
Wolbachia sp. wRi	α-proteobacteria; Rickettsiales;
Xanthobacter autotrophicus	α-proteobacteria; Rhizobiales;
Zymomonas mobilis	α-proteobacteria; Sphingomonadales;
Borrelia burgdorferi	Bacteria; Spirochaetes;
Desulfovibrio desulfuricans	δ-proteobacteria; Desulfovibrionales;
Escherichia coli	γ-proteobacteria; Enterobacteriales;
Neisseria gonorrhoeae	β-proteobacteria; Neisseriales;
Neisseria meningitidis	β-proteobacteria; Neisseriales;
Planctomyces limnophilus	Bacteria; Planctomycetes;
Cavia porcellus	Eukaryota; Metazoa;
Laminaria digitata	Eukaryota; Stramenopiles;
Malawimonas jakobiformis	Eukaryota; Excavata;
Marchantia polymorpha	Eukaryota; Plantae;
Penicillium marneffei	Eukaryota; Fungi;
Reclinomonas americana	Eukaryota; Protozoa;
Rhodomonas salina	Eukaryota; Cryptophyta;

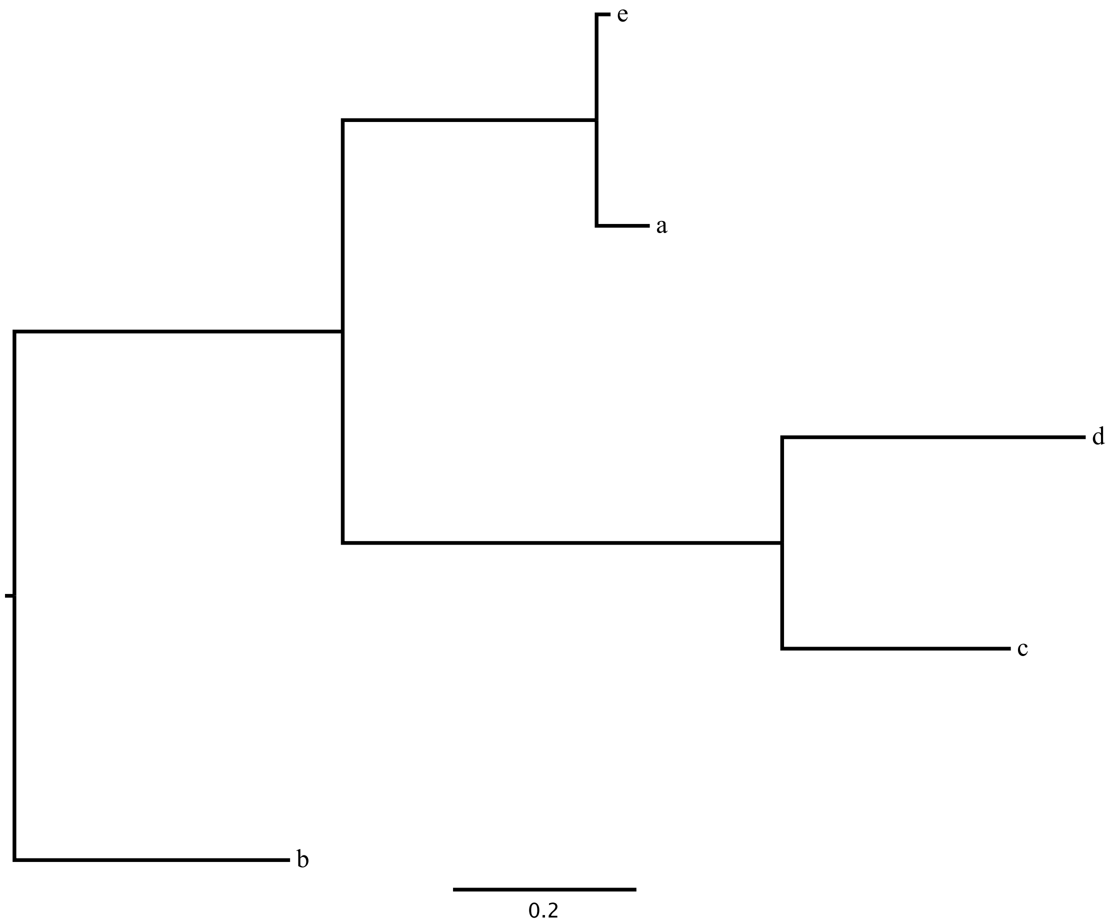

Figure A1: Tree used to simulate *Test 1*, *Test 3* and *Test 4* with five taxa (see section 3.2.4).

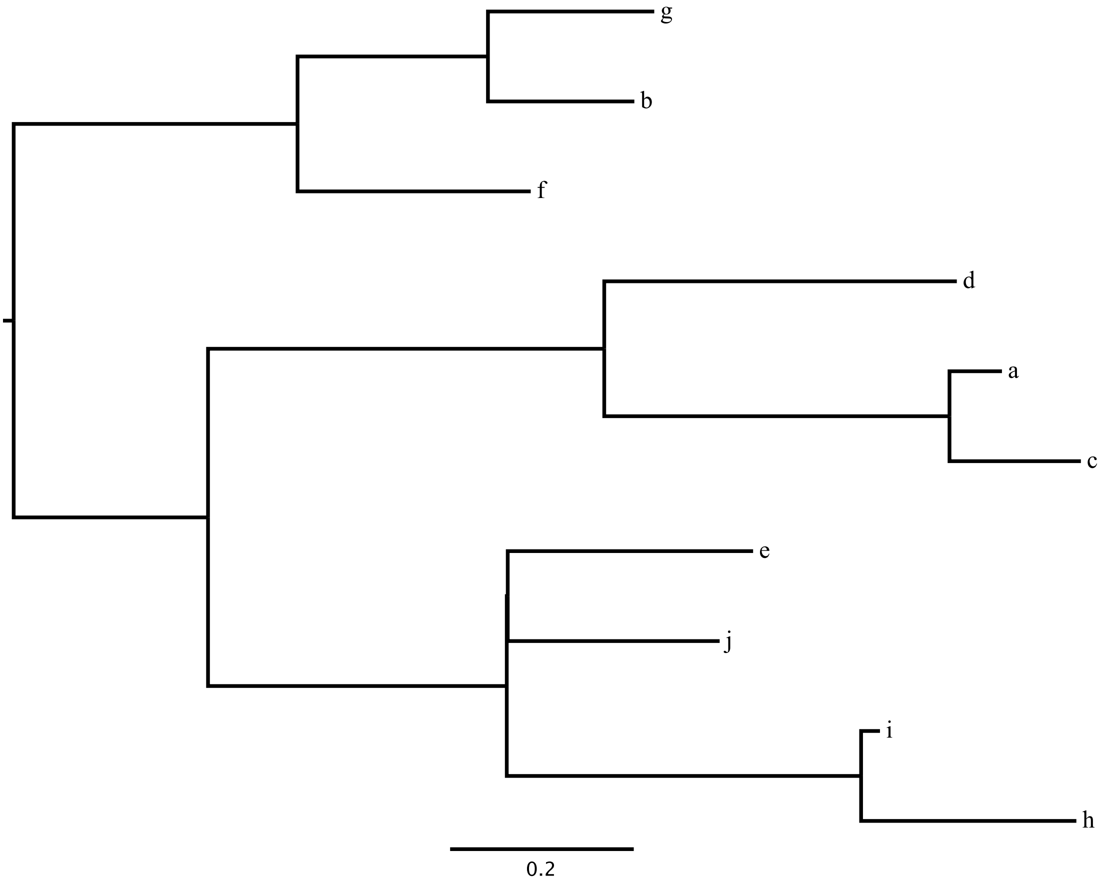

Figure A2: Tree used to simulate *Test 1*, *Test 3* and *Test 4* with ten taxa (see section 3.2.4).

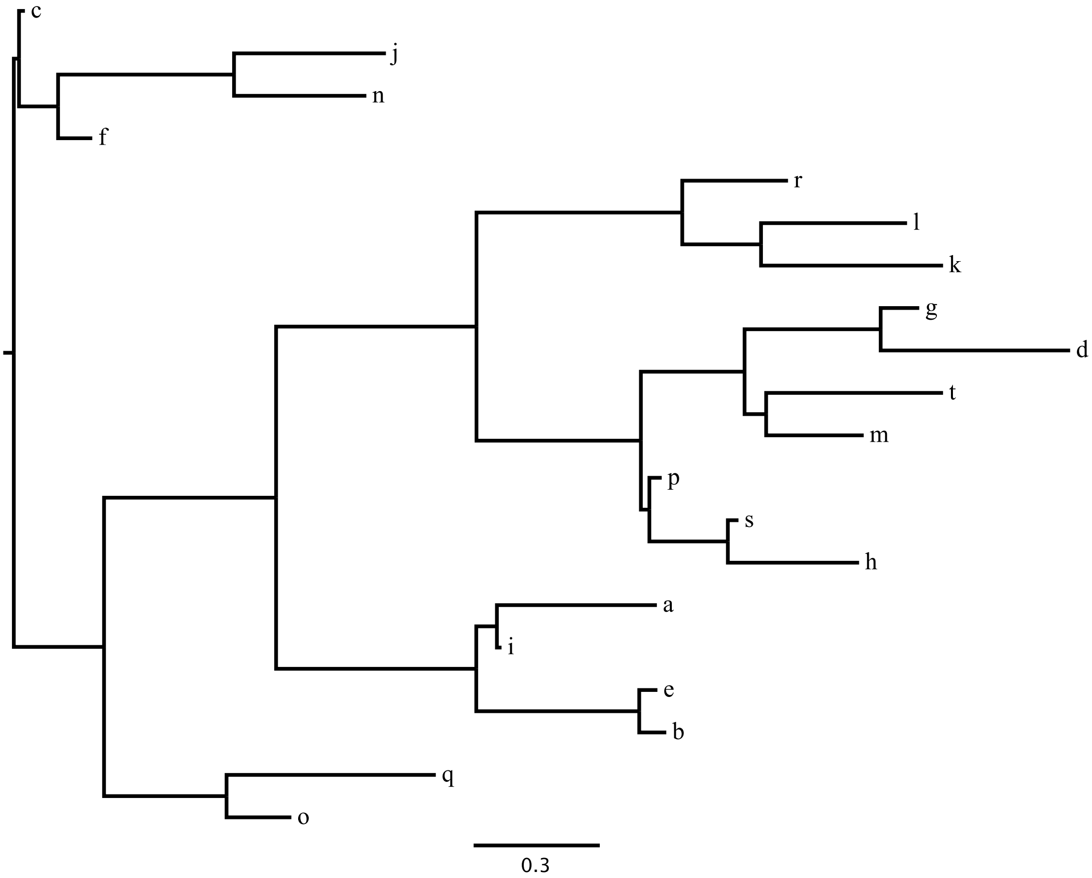

Figure A3: Tree used to simulate *Test 1*, *Test 3* and *Test 4* with 20 taxa (see section 3.2.4).

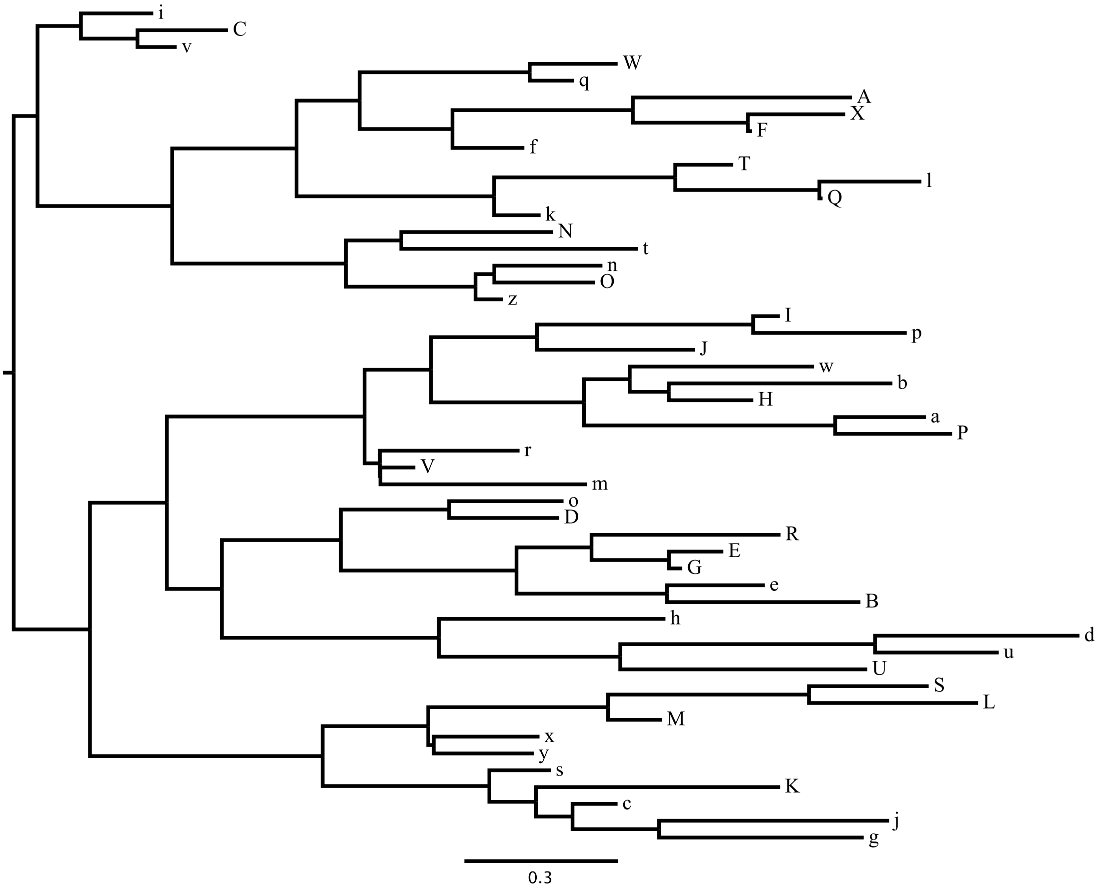

Figure A4: Tree used to simulate *Test 1*, *Test 3* and *Test 4* with 50 taxa (see section 3.2.4).

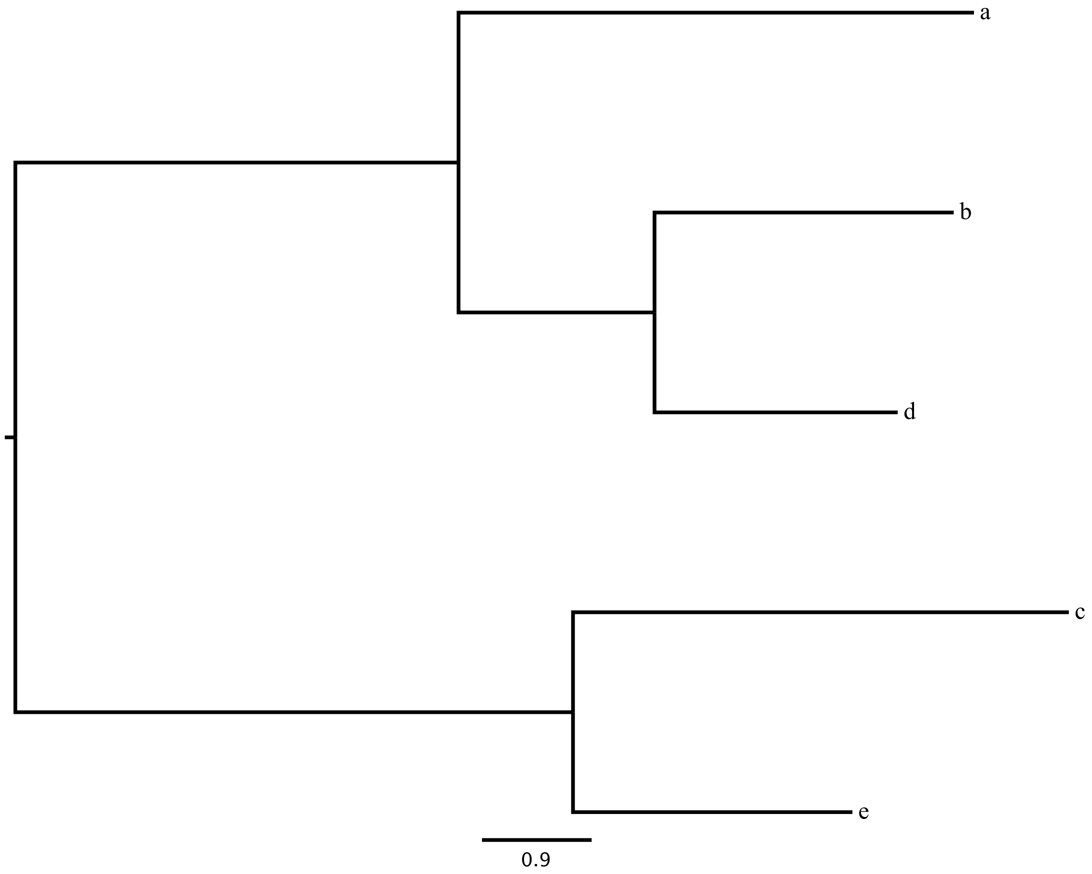

Figure A5: Tree used to simulate *Test 2* with five taxa (see section 3.2.4).

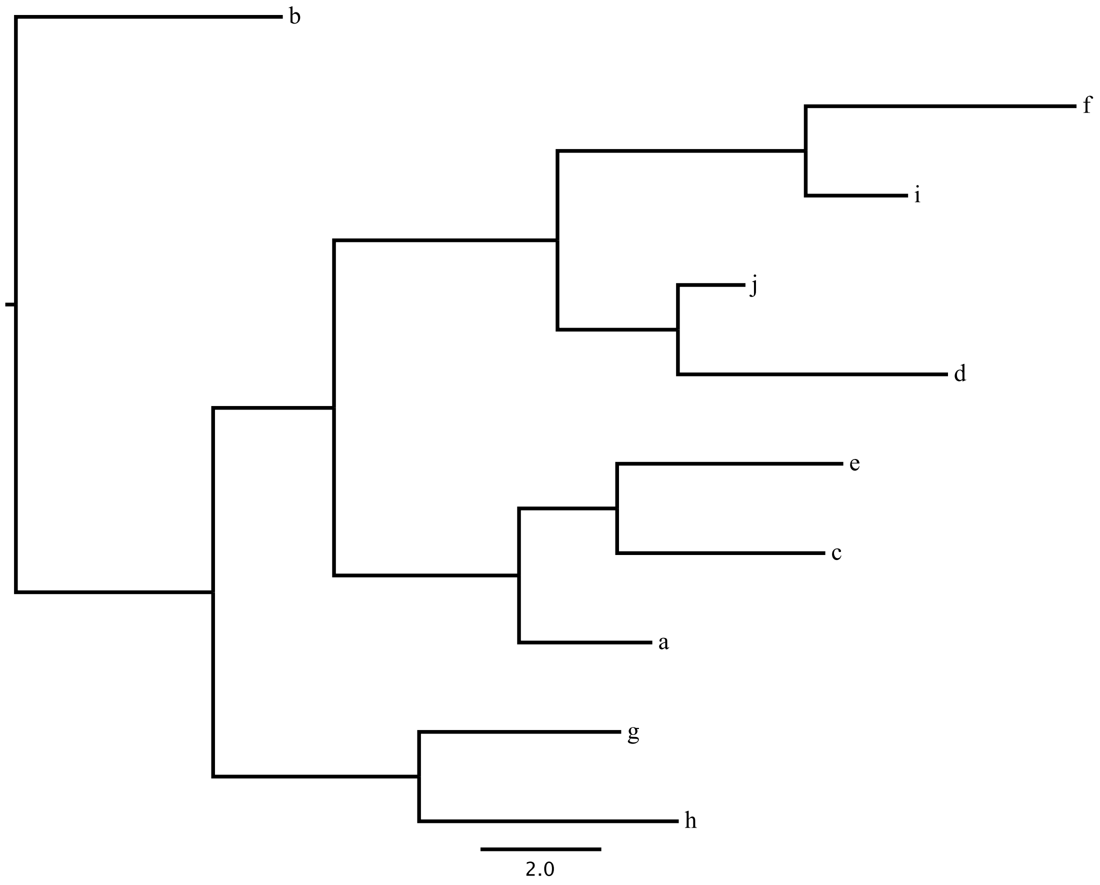

Figure A6: Tree used to simulate *Test 2* with ten taxa (see section 3.2.4).

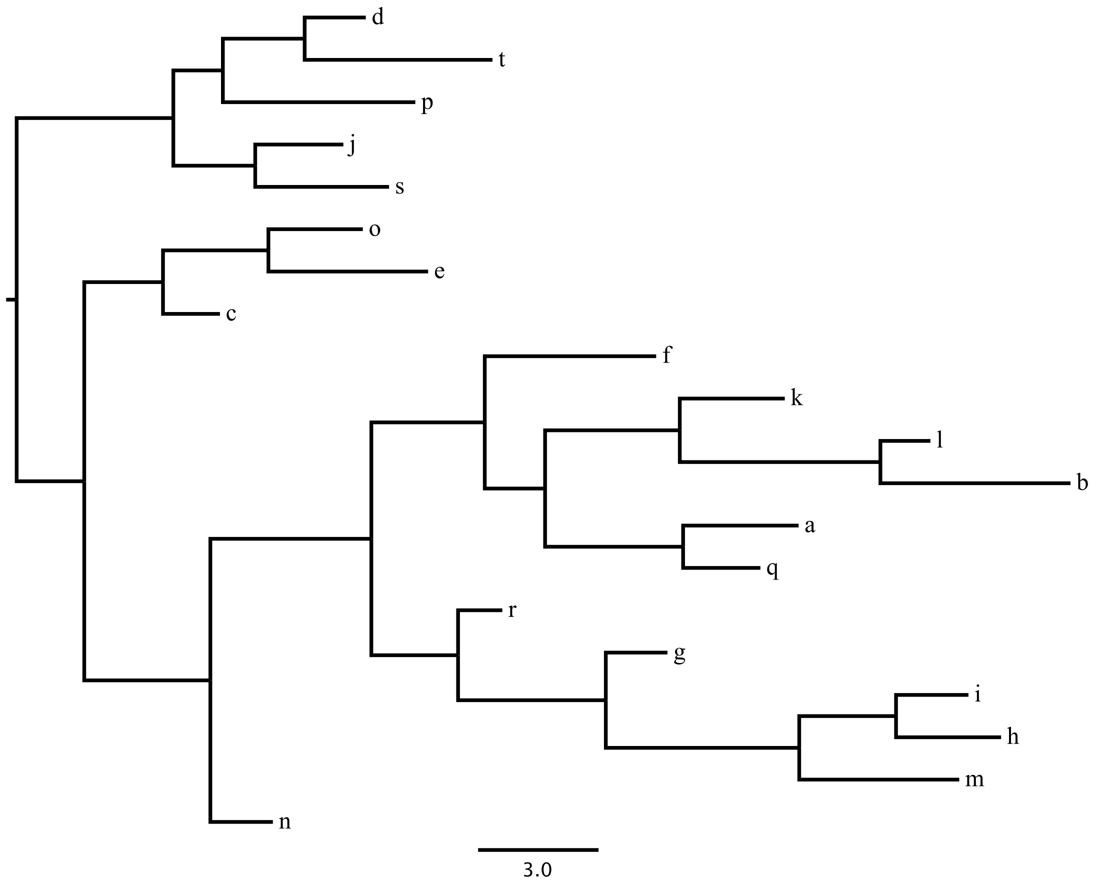

Figure A7: Tree used to simulate *Test 2* with 20 taxa (see section 3.2.4).

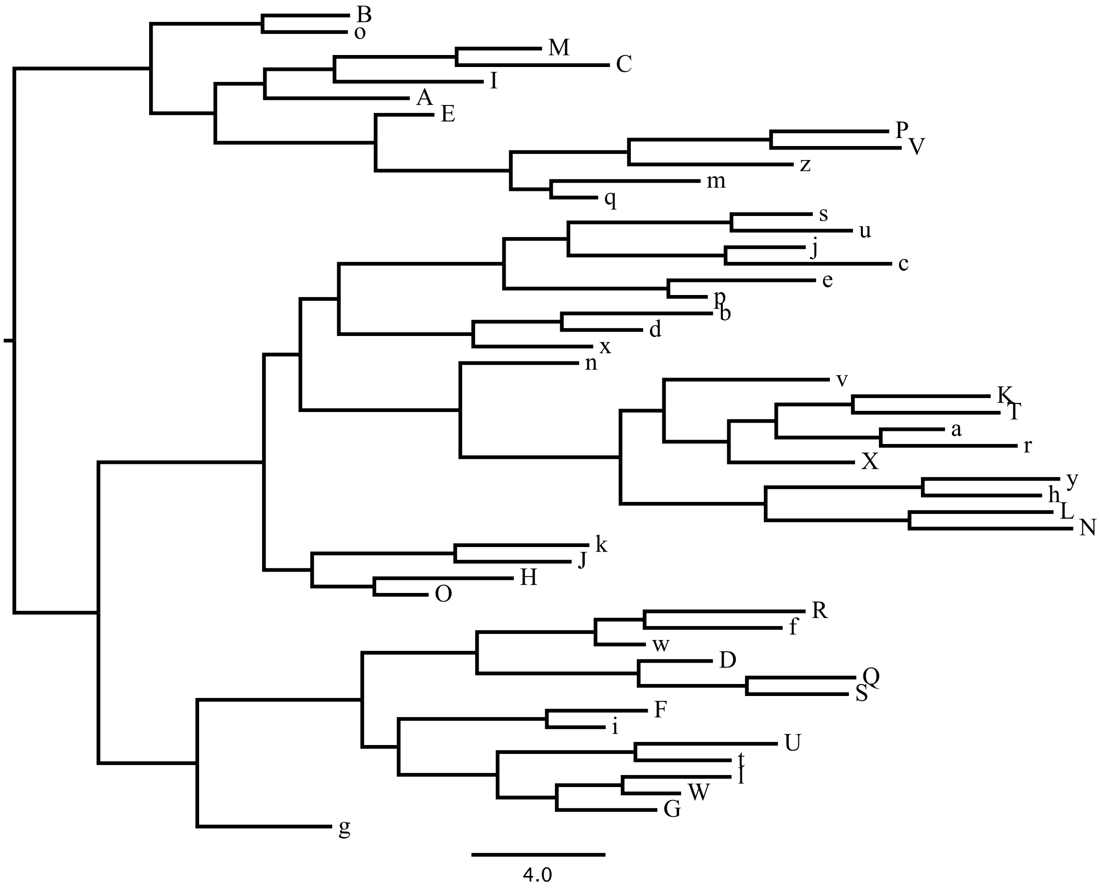

Figure A8: Tree used to simulate *Test 2* with 50 taxa (see section 3.2.4).

Publication

Syst. Biol. 60(6):833–844, 2011

DOI:10.1093/sysbio/syr064
Advance Access publication on July 29, 2011

A Method for Inferring the Rate of Evolution of Homologous Characters that Can Potentially Improve Phylogenetic Inference, Resolve Deep Divergence and Correct Systematic Biases

Carla A. Cummins and James O. McInerney*

Molecular Evolution and Bioinformatics Unit, Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland;
Correspondence to be sent to: Molecular Evolution and Bioinformatics Unit, Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland; E-mail: james.o.mcinerney@nuim.ie.

Received 3 September 2010; reviews returned 14 February 2011; accepted 2 March 2011
Associate Editor: Michael Charleston

Abstract.—Current phylogenetic methods attempt to account for evolutionary rate variation across characters in a matrix. This is generally achieved by the use of sophisticated evolutionary models, combined with dense sampling of large numbers of characters. However, systematic biases and superimposed substitutions make this task very difficult. Model adequacy can sometimes be achieved at the cost of adding large numbers of free parameters, with each parameter being optimized according to some criterion, resulting in increased computation times and large variances in the model estimates. In this study, we develop a simple approach that estimates the relative evolutionary rate of each homologous character. The method that we describe uses the similarity between characters as a proxy for evolutionary rate. In this article, we work on the premise that if the character-state distribution of a homologous character is similar to many other characters, then this character is likely to be relatively slowly evolving. If the character-state distribution of a homologous character is not similar to many or any of the rest of the characters in a data set, then it is likely to be the result of rapid evolution. We show that in some test cases, at least, the premise can hold and the inferences are robust. Importantly, the method does not use a "starting tree" to make the inference and therefore is tree independent. We demonstrate that this approach can work as well as a maximum likelihood (ML) approach, though the ML method needs to have a known phylogeny, or at least a very good estimate of that phylogeny. We then demonstrate some uses for this method of analysis, including the improvement in phylogeny reconstruction for both deep-level and recent relationships and overcoming systematic biases such as base composition bias. Furthermore, we compare this approach to two well-established methods for reweighting or removing characters. These other methods are tree-based and we show that they can be systematically biased. We feel this method can be useful for phylogeny reconstruction, understanding evolutionary rate variation, and for understanding selection variation on different characters. [Compatibility; maximum likelihood; maximum parsimony; molecular phylogeny reconstruction; site rate variation; site removal; simulation; systematic bias.]

Homologous characters evolve at different rates. Within a given data matrix, some characters might evolve at an appropriate rate to resolve the branching order of the taxa in question (Townsend 2007) whereas others might exhibit high levels of homoplastic noise. Some might be too slowly evolving and therefore mute with respect to phylogenetic statements (Kluge and Farris 1969; Delsuc et al. 2005; Philippe et al. 2005; Townsend 2007). A character could be considered important if it contains useful information about the phylogeny of the group of interest and if it is relatively free of homoplasy for that group. Therefore, for deep phylogenetic relationships, a slowly evolving character might prove useful, whereas for shallower relationships, a more rapidly evolving character could prove to be more useful. Character-state substitution rate (i.e., the rate at which a characters state is transformed into a different state) is an important factor to consider when ranking the informativeness of characters. Knowing a priori the rate of evolution of a character can greatly facilitate the treatment of characters for phylogeny reconstruction.

A number of efforts have been made to evaluate character-specific evolutionary rates. Farris (1969) introduced successive approximations character weighting (SACW) in order to weight characters according to a perceived importance assigned to them. This weighting scheme sought to ensure that characters with a higher degree of correlation with the phylogenetic history were more highly regarded during reconstructions. Farris defined this correlation as the consistency index (CI) for a matrix, or the goodness of the fit of the characters within the matrix to a given tree. The CI for an individual character on a particular tree is derived as the minimum possible character length divided by the observed character length on the considered tree. So, when a character fits on a tree without apparent homoplasy, the CI value is unity. If additional ad hoc hypotheses need to be invoked to explain the evolution of the character on the tree, then the CI value will be less than one (Farris 1969). The CI for a data matrix is obtained by averaging the CI values for all the characters in the matrix. Therefore, a tree must be initially inferred. In his description of the method, Farris preweighted characters according to a weighting system devised by Le Quesne (1969), though he indicated that initial character weights set to unity would also work. As a consequence of the approach, characters that tend to disagree with the initial tree are given a lower weighting in subsequent analyses, in contrast to characters that tend to agree with this initial tree, whose weight remains high.

In the late 1980s, Olsen (1987) noted that among-site rate variation (ASRV) could cause problems in phylogenetic inference, and he attempted to accommodate this variation using a model-based approach that employed a normal distribution. Using a model to account for rate variation across sites can increase the probability of finding the correct phylogenetic tree topology compared with a method that does not account for rate variation (Yang 1993). By using an evolutionary model that neglects to account for ASRV, sequences will appear to have undergone fewer mutations overall and will appear to be more similar to their relatives compared with an analysis using a model that accounts for ASRV. Therefore, much of the effort to improve phylogeny reconstruction accuracy has focused on methods that deal with accommodating site rate heterogeneous data (Farris 1969; Yang 1996; Brinkmann and Philippe 1999; Hirt et al. 1999; Schmidt et al. 2002).

Yang (1996) modeled ASRV using the gamma distribution. This distribution has some attractive properties, particularly given that its shape can change from being L shaped to being hill shaped, depending on the characteristics of the alignment. Again, this approach tries to incorporate rate variation and it assumes that site rate heterogeneity is well approximated by this model. However, assuming that all sites are free to vary will lead to incorrect estimations when there are sites in the data set that do not or cannot change (Yang 1996). In 1970, Fitch and Markowitz (1970) proposed that for a protein there might be two classes of sites—invariable and variable and they suggested a method of analyzing molecular alignments in order to determine how many positions were invariable and how many were variable. These invariable sites can also confound phylogeny reconstruction and accentuate rate variation across sites. To overcome these issues, some studies have experimented with the removal of sites that violate assumptions of the models that are being used. This has the effect of reducing the range of site-to-site rate variation in the data set.

As an example of a study that effectively reduced site-to-site rate variation, Hirt et al. (1999) not only removed invariant sites, but also removed sites they considered to be fast evolving (Hirt et al. 1999). They identified fast-evolving sites by using two different phylogenetic trees and only removing sites that were considered to be fast evolving on both topologies. In this case, removal of both slow- and fast-evolving sites vastly improved the support values for internal branches on the phylogenetic trees and resulted in a robust placement of the Microsporidia.

Many different methods exist for the identification of sites with a high substitution rate (Farris 1969; Kuhner and Felsenstein 1994; Brinkmann and Philippe 1999; Hansmann and Martin 2000; Schmidt et al. 2002; Pisani 2004). The majority, though not all, of these methods are tree based. Tree based methods identify rapidly evolving sites based on a tree either provided by the user or inferred by the method before site identification. For instance, TREE-PUZZLE (Schmidt et al. 2002) and DNArates (Maidak et al. 1996; Olsen et al. 1998) estimate evolutionary rates for each character based on a given tree and process of character-state substitution. TREE-PUZZLE can employ a discrete gamma distribution to estimate site rates, with sites allocated to a different category based on their likelihood score on the tree. The DNArates program has been used in conjunction with the fastDNAml program (Olsen et al. 1994) in order to partition alignments of homologous characters into rate categories (Fischer and Palmer 2005). Fischer and Palmer (2005) used a procedure that is not unlike the SACW approach in order to reweight characters for subsequent analyses. For a data set that was aimed at settling the placement of Microsporidia, they found that early unweighted data sets resulted in a variety of placements of the taxon, whereas successive rounds of character reweighting tended to result in fewer tree topologies and finally the authors settled on a placement of the microsporidia with the fungi that was best supported by the successively reweighted data.

Brinkmann and Philippe (1999) developed a method known as “slow-fast” where an alignment is split into groups (Brinkmann and Philippe 1999; Kostka et al. 2008). These groups are generally user-defined taxonomic groups. The evolutionary rate at a given site is calculated as the sum of the number of changes at the same position in all the groups individually. Although groups are, technically, user defined, any prior knowledge of the group will be based on previous tree inferences and, therefore, the slow-fast method is, by proxy, a tree-based method. In addition, due to the nature of this method, it is not suitable for small data sets.

The problem with tree-based methods is that the true tree is rarely known with certainty. Therefore use of an incorrect initial tree can result in incorrect assignation of an evolutionary rate to a character. Each character is compared with the given tree topology, whether correct or incorrect. A character is considered rapidly evolving if it conflicts with the initial tree or has a high level of homoplasy when mapped onto the tree. By assuming a topology prior to site rate identification, a slowly evolving site could potentially appear to be rapidly evolving, simply because the tree onto which it is mapped is incorrect. This initial error can become a source for systematic biases. Therefore, it may be preferable to have a method of determining evolutionary rate for a character that is independent of any a priori tree estimation procedure.

Tree-independent approaches to differentially weighting characters for phylogeny reconstruction include the Le Quesne (1969) test of character compatibility, which provided a “coefficient of character-state randomness” that could be used, if desired, to exclude characters from subsequent analysis. Essentially, this test evaluates two characters and if they can be mapped onto the same tree without homoplasy, then they are compatible, otherwise they are incompatible. Characters that have the highest amounts of incompatibilities with the other characters might be considered candidates for removal prior to subsequent phylogenetic analysis. Le Quesne (1989) later introduced the notion of compatibility within data being indicative of the level of phylogenetic information. This work was further extended by Meacham (1994), who developed his “Frequency of Compatibility

Attainment" statistic. Wilkinson (1998) highlighted the advantages of creating split patterns for sites when detecting conflict. Conflict, as defined by Le Quesne, becomes much easier to identify and rank when using a universal coding system for sites. Pisani (2004) utilized this idea to identify fast-evolving sites. According to the method of Pisani (2004), each site in the alignment receives an Le Quesne Probability (LQP) score, which is "[…] the probability of a random character having as low or lower incompatibility with the rest of the data than does the original character.". Pisani used this probability measure to explore arthropod relationships using different strategies for removal of characters with differing LQP values.

Hansmann and Martin (2000), in contrast with the compatibility strategies, proposed a very simplistic non–tree-based method for identifying rapidly evolving characters. They used the number of different character states in an alignment column as a proxy for evolutionary rate (Hansmann and Martin 2000). They cite the intuitiveness of the relationship between higher numbers of polymorphisms at a site and speed of evolution at that site. The set of most polymorphic characters would, therefore, be enriched in homoplastic sites (Hansmann and Martin 2000). However, each site is treated as a separate entity and consequently, this approach does not include information that may be contained in the data set as a whole, apart from ranking the sites from least to most polymorphic. In this paper, we present our method, TIGER (Tree Independent Generation of Evolutionary Rates), which is based on a similar concept to Le Quesne (1989), Wilkinson (1998) and Pisani (2004). TIGER analyzes similarity within characters (Wilkinson 1998). We expect that fast-evolving characters have lost some, most, or all of their phylogenetic signal and therefore should demonstrate reduced similarity with other sites that are more slowly evolving. Rather than comparing sites and only allowing them to be compatible or incompatible, our method allows sites to be scored according to varying degrees of similarity. This approach should provide a more fine-grained or nuanced result than the one that scores sites as being either compatible or incompatible.

In this report, we analyze synthetic data sets in order to explore the behavior of our approach and then, to demonstrate the utility of the method, we analyze two well-known problematic data sets. Additionally, we show that tree-based site removal approaches have significant problems, particularly when the data set contains a systematic bias (e.g., convergent base compositional bias), whereas our tree-independent approach can overcome these biases.

Methods

Set Partitions

Our method is based on the analysis of set partitions at each position in a matrix. This matrix could be any type of data, including alignments of DNA or protein sequences or a matrix of homologous morphological characters.

A partition of a set X is a set of nonempty subsets of X such that every element x in X is in exactly one of these subsets. We treat each character in the matrix as a set and partition this set based on character states. A set partition is denoted, for example, as {{1}, {2, 3}, {4}, {5}} or 1/2,3/4/5. The partition 1/2,3/4/5 shows that for this character, taxa 2 and 3 have the same character state which is different from all the others, taxon 1, taxon 4 and taxon 5 each have unique character states—both different from each other and different from taxa 2 and 3. In this way, each character's partition is determined in order to enable pairwise comparisons with the rest of the characters in the data set. For example, in a nucleotide alignment of six taxa, character J = AAGGGC and character K = TTCCCA (assuming the order of the taxa is the same for both characters in this example). The partition set for both J and K is 1,2/3,4,5/6, despite having different character states.

Using this kind of data transformation, we can measure the degree of similarity between characters based on the similarity of their set partitions. We find that a character with a set partition that is similar to many other characters in the data matrix can usually, though not always, be a more slowly evolving character than a character with a set partition that is less similar to the rest of the characters in the matrix. Therefore, we can use the average similarity of a character's set partition to the rest of the matrix as a proxy for evolutionary rate.

The rate r_i for the character at position i is defined as:

$$r_i = \frac{\sum_{j \neq i} \mathrm{pa}(i,j)}{n-1} \qquad (1)$$

where n is the total number of characters in the matrix and $\mathrm{pa}(i,j)$ is the partition agreement score. This is defined as

$$\mathrm{pa}(i,j) = \frac{\sum_{x \in P(j)} a(x, P(i))}{|P(j)|}, \qquad (2)$$

where $|P(j)|$ is the number of groups in the partition of the jth character and $a(x, P(i))$ equals 1 if $x \subseteq A$ for some $A \in P(i)$. $P(i)$ may be defined as a partition in character i.

It is important to note that, given two sets A and B, if $A \subseteq B$, it is not necessarily commutative and, often, $B \not\subseteq A$. In this case, $\mathrm{pa}(i,j) \neq \mathrm{pa}(j,i)$. Also, because the rate is based on averaging of combinations of 1 or 0 values, it will always have a range between 0 and 1. A constant site, that is, a site with only one character state, will have $r = 1$ given that the pa, will be one for every comparison.

For example, consider two sites A = CTTAA and B = AGGGG with partition sets 1/2,3/4,5 and 1/2,3,4,5, respectively. $\mathrm{pa}(A, B) = 0.5$ because, out of two partitions in B ({1} and {2,3,4,5}), only $\{1\} \subseteq P(A)$. Given that {2,3,4,5} is not a subset of any partition in A, $a(\{1\}, P(A)) = 1$ and $a(\{2,3,4,5\}, P(A)) = 0 \therefore \mathrm{pa}(A, B) = 0.5$. As mentioned, this calculation is not commutative, so $\mathrm{pa}(B, A) \neq 0.5$. $\mathrm{pa}(B, A) = 1$ because all partitions in $A \subseteq P(B)$.

This approach is designed to measure how much a particular character tends to agree with the other characters in the data. If a character shares partitions with many other characters, then it is likely that they hold similar information. This may be viewed as a signal in the data. Conversely, a character whose set partition greatly differs from the other signals in the data may be thought of as noise. To put it another way, a rapidly evolving character is likely to have sustained multiple substitutions, some or all of whom might be superimposed on earlier substitutions, therefore, this character is more likely to have a set partition that agrees less with more slowly evolving characters.

It is reasonable to suggest that a character that shares few partitions with the majority of other characters could be considered rapidly evolving. On the other hand, a slowly evolving character is more likely to share partitions with, or at least have fewer that conflict with, many other characters. The first assumption might not hold true in a situation where all or most characters in a matrix are rapidly evolving. It is most likely to hold true when evolutionary rates are moderate and when there is a gradient of evolutionary rates from slow to fast. Note that the rate of evolution that is assigned to a particular character is measured in arbitrary units and will vary with the data matrix being used. It is not a measure of substitutions per unit of time and indeed there are no units associated with the rate. This method can be used to analyze DNA, protein, morphological, or other arbitrary homologous characters.

It should be noted that for the current analyses, we did not attempt to deal with missing data. Missing data can be a feature of both molecular and morphological data sets, usually because a particular gene or morphological character has not been sampled or found. Missing data can be accommodated by an appropriate pruning of the characters so that only character states that have been observed are being compared.

Binning

It is often useful or convenient to group sites with similar evolutionary rates together and in our implementation of this method a range of rates can be divided into a user specified number of partitions, or bins. Sites are placed into bins depending on their rate value. The slowest rate and the fastest rate are determined and bins are constructed by splitting the rates into equal partitions. In this paper, we have used a variety of binning schemes, from 8 bins to 20 bins. In theory, any number of bins can be constructed, as long as the number is less than or equal to the number of characters in the matrix.

Data Simulations

In order to test the features of the method, we generated a number of artificial nucleotide data sets, using a phylogenetic tree and a prespecified model of nucleotide substitution. In the first instance, we simply wanted to know if data sets with different patterns of ASRV would return different patterns when analyzed using TIGER. Second, we wanted to see if removing characters had a beneficial effect on the fit of the data matrix to all possible trees or produced the desirable effect of improving the fit of the data to "good" trees while worsening the fit of the data to bad" trees. Our third simulation experiment involved the evaluation of whether or not the TIGER approach to character removal would improve the likelihood of resolving deep relationships.

In this report, we have used nucleotide data for reasons of ease of interpretation and also because of the ready availability of excellent computer software (Rambaut and Grassly 1997) to generate the data; however, in principle we could have used protein, morphological, or any kind of multistate character matrices.

Varying gamma shapes.—Using Seq-Gen (Rambaut and Grassly 1997), we simulated two data sets over the same 49-taxon tree (Maddison 2004) (the tree is available in Supplementary Material, available from http://www.sysbio.oxfordjournals.org/) and we employed a model that used a discrete approximation to the gamma distribution, with four categories of sites. In order to assess whether or not the TIGER algorithm could detect different patterns of ASRV, two different α values were used in simulations—0.5 and 20.0 reflecting two different distribution shapesthe first is L shaped and the second is hill shaped. Both alignments were 999 bp in length and simulated under the JC model (Jukes and Cantor 1969). We experimented with other models of sequence evolution and different tree shapes and numbers of taxa and the results are essentially the same as presented here, so we only present the results of the JC simulations on this data set.

Changing fit of the data to all trees in treespace.—Removal of homoplastic characters in a matrix should have the effect of improving the fit of the data to the true tree whereas worsening the fit of the matrix to trees that are very different from the true tree. However, given that it is possible to edit any tree to change its topology into any other tree, if we perform any data modification it will most likely influence the goodness-of-fit of the data to all trees in some way. Some trees are very similar to the true tree and some are very dissimilar, consequently, whereas incrementally removing larger numbers of characters (grouped into bins), we investigated the change in fit of the data to all possible phylogenetic trees for an eight-taxon data set. In our experiments, we measured the change in the CI for all trees as bins were sequentially removed, starting with the bin containing the most rapidly evolving characters (a total of 10 bins were used in this experiment). In effect, for the set of all trees, T, we computed the CI for the original data set on tree t ($t \in T$) and compared this value with the CI value for the data set with Bin10 removed. We then plotted this value against the "nodal" distance (Puigbo et al. 2007) between the true tree and tree *t* (when *t* is not the true tree). For the true tree, the nodal distance is always zero. We carried out the same procedure when we removed Bin9+Bin10, Bin8+Bin9+Bin10, and Bin7+Bin8+Bin9+Bin10.

TIGER rates versus likelihood scores.—Using the correct tree and the correct model, site-specific likelihood scores can give a very good estimate of character evolutionary rate. We wished to test how well the TIGER approach could identify these characters without any knowledge of a tree. We used 100 different seven-taxon trees chosen at random from treespace (which contains 945 unrooted trees). A nucleotide alignment of 999 positions was generated under the JC model for each of these 100 trees. We generated site-specific likelihood scores in PAUP* (Wilgenbusch and Swofford 2003) for all 945 trees for each data set and we measured the ranking of sites on each tree to TIGER rankings. That is to say, the site(s) with the highest likelihood value are ranked as #1 and the site(s) with the lowest value as #999 and likewise for TIGER rates. The Euclidian distance between all likelihood rankings and TIGER rankings was calculated. This is a very simple measure of the average difference in rank for a character in the two lists.

Deep branching tree.—Rapid evolution can obfuscate deep relationships on a tree, often leading to unwanted polytomies. This situation is particularly problematic when long unbroken branches subtend a series of rapid cladogenetic events. To test whether the TIGER approach could help resolve deep relationships where there is very little phylogenetic signal, we used the JC model of sequence evolution to produce 100 simulated 999 bp nucleotide data sets across the eight taxon tree shown in Figure 2. The short deep branches combined with long terminal branches presents a difficult problem for phylogenetic analysis, mostly due to the confounding effects of rapidly evolving characters. To ensure that the data generated displayed poor phylogenetic resolution, we built a majority-rule consensus tree from maximum likelihood (ML) trees constructed from each of the data sets prior to any site removal. This was repeated after removal of sites dictated by TIGER and to test the performance of a tree based method in this scenario, we also repeated the analysis after removal of rapidly evolving sites identified by ML. The ML tree was estimated using PAUP* and the sites were categorized on this tree using TREE-PUZZLE.

Empirical Testing

Thermus data set.—In order to further understand TIGER's functionality, two empirical data sets were used. A 1273-column alignment of bacterial 16S ribosomal RNA genes known as the *Thermus* data set is well studied (Embley et al. 1993; Mooers and Holmes 2000), and we used this data set to examine whether the TIGER approach is useful for accounting for base compositional biases. This data set contains three thermophiles, *Aquifex aeolicus*, *Thermatoga maritima*, and *Thermus aquaticus* whose sequences are enriched in G and C nucleotides and two mesophiles, *Bacillus subtilis* and *Deinococcus radiodurans* whose nucleotide composition is more balanced. A combination of compositional bias and distant relationships can mean that when there is only a weak phylogenetic signal, it can be overcome by the similarity in base composition of the most rapidly evolving positions in the alignment. In general, many methods of phylogenetic analysis will group the thermophiles together in this data set, despite the fact that there is strong evidence that *T. aquaticus* and *D. radiodurans* are sister taxa (Embley et al. 1993). We refer to a tree displaying the mesophiles as a monophyletic group to the exclusion of the thermophiles as the ATTRACT tree and this is the tree recovered by most tree inference methods using the whole sequence alignment. We refer to a phylogenetic tree that places *T. aquaticus* and *D. radiodurans* together as the TRUE tree. Due to this well-characterized strong compositional attraction, we wished to investigate whether site removal using the TIGER approach could influence recovery of the correct tree. However, to demonstrate the different effects of site removal in a tree-independent fashion compared with the traditional ML approaches, we also compared the topology inferred after removal of rapidly evolving sites identified by TIGER with the topology recovered after removal of rapidly evolving sites according to TREE-PUZZLE (Schmidt et al. 2002) and SACW (Farris 1969). We did not use TREE-PUZZLE to infer the tree, we simply used the method implemented by TREE-PUZZLE to assign evolutionary rates to sites, based on a tree that we supplied to the software.

Primate data set.—It has generally been accepted that humans share a close relationship with orangutans, gorillas, and chimpanzees (Hayasaka et al. 1988; Begun 1992; Adachi and Hasegawa 1995; Shoshani et al. 1996; Ruvolo 1997; Satta et al. 2000; Ebersberger et al. 2007). From this group, it is generally agreed that orangutans are the least closely related to humans and that humans, chimps, and gorillas form a monophyletic group, though there are some conflicting opinions (Schwartz 1984; Grehan and Schwartz 2009).

The relationships of interest, therefore, concern the human, chimpanzee, and gorilla lineages (Satta et al. 2000). The separation of these three lineages is thought to have occurred in quick succession (Hayasaka et al. 1988; Adachi and Hasegawa 1995), and this makes the phylogeny difficult to resolve and the two alternative hypotheses—human, chimp together (HC hypothesis) or chimp, gorilla together (CG hypothesis)—receive almost equal support from this data set. Because of the controversy surrounding this topology, the second empirical data set we use is a well-known primate mitochondrial data set (see Supplementary material) consisting of 12 sequences and 898 aligned nucleotide positions (Hayasaka et al. 1988).

In a parsimony analysis of the data set, with all characters being equally weighted, both the HC and the CG hypotheses are equally good, with 1153 steps required to explain the data. We used the tree-based methods of assigning character evolutionary rates and use alternatively the HC and the CG trees in order to carry out the inferences. We compared and contrasted the results from tree-based analysis with the tree-independent method described here.

Results and Discussion

Varying Gamma Shapes

Our first analysis of the behavior of the TIGER method focused on the analysis of simulated data sets for 49 taxa with different patterns of rate variation across sites. We chose the 49-taxon data set that is distributed with the MACCLADE software (Maddison 2004) because it contains a reasonable range of branch lengths and has a moderately large number of taxa. We simulated two separate data sets that differed by the ASRV model used to generate the data. In the first case, we used a gamma distribution with an α parameter of 20 and in the second the α parameter was set to 0.5, reflecting very different evolutionary scenarios. We then used the TIGER approach to place sites into 20 bins sorted by their rate of evolution (Fig. 1a,b).

There are two interesting points to be made about Figure 1. First of all, the two graphs are not the same and furthermore Figure 1b, which is generated from the data set with an α parameter of 0.5, is more L shaped than Figure 1a, which was generated from the data with an α parameter of 20. This indicates that the TIGER approach is detecting the different ASRV patterns. What is of further interest is that within each graph there is a clear multimodality. There are four clusters of bars on the histograms (indicated by the alternative shading and clear zones on the diagrams). When the seq-gen software generates data, it uses an approximation to the gamma distribution and in these cases an approximation was employed that used four categories of sites. The TIGER approach has identified these subtle patterns and has placed the different sites into clusters.

True Tree versus Incorrect Trees

If the removal of rapidly evolving characters really is a good idea for improving the chances of recovering the correct phylogenetic tree, then we expect that removal of these characters would improve the goodness-of-fit of the data to the true tree while worsening the goodness-of-fit of the data to other trees. In order to test this hypothesis, we generated a simulated data set containing eight taxa and using the JC model, according to the protocols previously described. We progressively removed the fastest evolving sites, as judged by the TIGER approach, until we had removed the four fastest categories of sites. We then examined the goodness-of-fit of the data to the correct tree (the tree used to simulate the data) and also the goodness-of-fit of the data to all the other possible trees. We plotted the goodness-of-fit measure (CI) against the nodal distance (as measured by the TOPD/FMTS software, Puigbo et al. 2007) for the unstripped data set for each possible tree topology and we plotted the change in CI (ΔCI) against nodal distance for each of the data sets where sites were stripped. The results of these experiments are seen in Figure 2. In total, there were 10,395 trees examined for each treatment of the data.

With all sites included in the alignment, the CI for the correct tree was 0.825. The worst CI value in the data set was 0.612 and the tree with the largest nodal distance from the true tree had a distance of 2.44949 and a CI value of 0.616. In general, there is a negative correlation between CI and nodal distance from the true tree.

When we stripped out the Bin10 category of sites, we saw the CI values increased for some trees and decreased for others. The CI value with the largest increase for any of the 10,395 trees was the CI value for the true

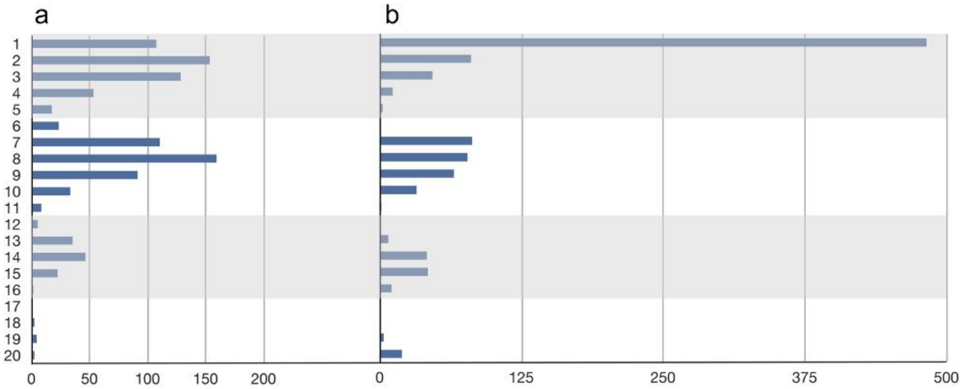

FIGURE 1. Histograms of binning results for two different data sets with different ASRV. a) A 999-bp, 49-taxon data set generated using the tree in S1 and ASRV modeled using a gamma distribution with a shape parameter of 0.5, and (b) data set of the same size and topology but with ASRV modeled using a gamma shape parameter of 20.0. The alternating shaded and clear areas indicate the four categories of sites that approximate the gamma distribution. This figure is available in black and white in print and in color at *Systematic Biology* online.

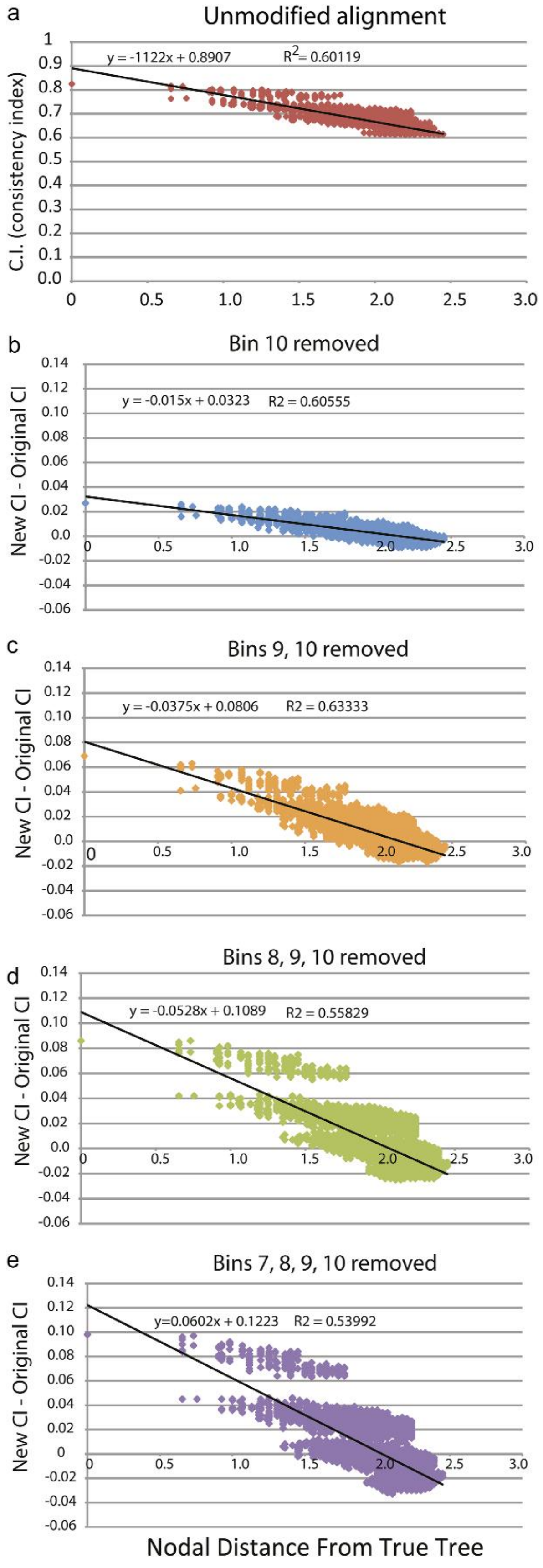

FIGURE 2.

tree—an increase to 0.852. In contrast, the tree with the largest nodal distance from the true tree experienced a decrease in CI value and its new value was 0.612. Overall, a total of 5364 trees (51.6% of the total) saw an increase in CI value, whereas 5031 trees experienced a decrease in CI value.

Continued site stripping resulted in a progressive increase in CI value for the true tree and a progressive decrease in CI value for the tree with the largest nodal distance from the true tree. When Bin categories 9 and 10 were removed, the values changed to 0.894 and 0.609, respectively, with 5403 (51.9%) of the trees now experiencing an increase in CI value. When Bin categories 8, 9 and 10 were removed, the values changed to 0.911 for the true tree and 0.601 for the worst tree with 3811 of the trees having an increased CI value. Finally, when we removed Bin categories 7, 8, 9, and 10, the values changed to 0.923 and 0.597, respectively, with 3257 of the trees experiencing an increase in CI value (31.3%), whereas 7138 had a decreased CI value (68.6%).

Therefore, we can see for this data set that site stripping has resulted in a bias in the fit of the data to different trees. In general, those tree topologies that are close to the true tree will begin to fit the data better, whereas those trees that are least similar in topology to the true will begin to fit the data worse. The tree that is most positively affected by site stripping is the true tree. It must be remembered that the TIGER approach is not tree based and at no time was the TIGER software aware of the topology of the true tree.

TIGER Rates versus Likelihood Scores

To see how well TIGER can approximate site-specific rates we compared it with likelihood scores for each site on every possible seven-taxon unrooted tree. The Euclidian distance from TIGER ranking to the likelihood rankings on all trees were recorded for all data sets, with particular emphasis on where the distance between TIGER rankings and the likelihood rankings on the known true tree fell with respect to the other trees. In 100% of data sets, this distance fell within the top 0.3% of all scores. In 95% of all cases, the distance from TIGER rankings to the likelihood rankings on the true tree was the smallest distance recorded to any tree in the data set.

This shows that the TIGER approach will produce an ordering of the evolutionary rates of the sites that is usually closer to the ranking of sites according to the true tree than to other incorrect trees.

Deep Branching Tree

In order to see whether it is possible for our method to improve the resolution of deep relationships where

←
FIGURE 2. Change in CI with increasing site removal. On the abscissa is the nodal distance of a tree from the correct tree and on the ordinate is either the CI or the difference in CI value between the unstripped alignment and the stripped alignment (ΔCI).

phylogenetic signal is weak, we simulated 100 different DNA alignments based upon a single phylogenetic tree with long external branches and very short internal branches (Fig. 3a). This alignment was designed to represent a difficult problem of phylogenetic inference and was simulated using the JC model of sequence evolution. ML trees for each of the data sets were inferred under the JC model. As expected, prior to removal of rapidly evolving sites, the majority-rule consensus analysis using the JC model produced a tree with polytomies and poor resolution (Fig. 3b), and the only branch that is resolved has a bipartition frequency (BF) of 55% was for a split that separates taxa C and D from the rest. We used the TIGER approach to identify the rapidly evolving characters in the matrices and place all characters into 10 bins with increasing evolutionary rate. Removal of the most rapid category of sites, Bin10, which contained between 183 and 502 sites with an average of 424 between the 100 data sets, entirely resolved all except the basal polytomy (Fig. 3c), with BF ranging from 67% to 99%. We wished to test our method against a tree-based method. We used TREE-PUZZLE (Schmidt et al. 2002) on the same simulated data. Removing the most rapidly evolving category of sites using the TREE-PUZZLE approach (ranging from 269 to 481 sites, mean of 334 sites removed) the tree remained equally unresolved as prior to any site removal, with the BF of the split separating C and D rising to 61 (Fig. 3b).

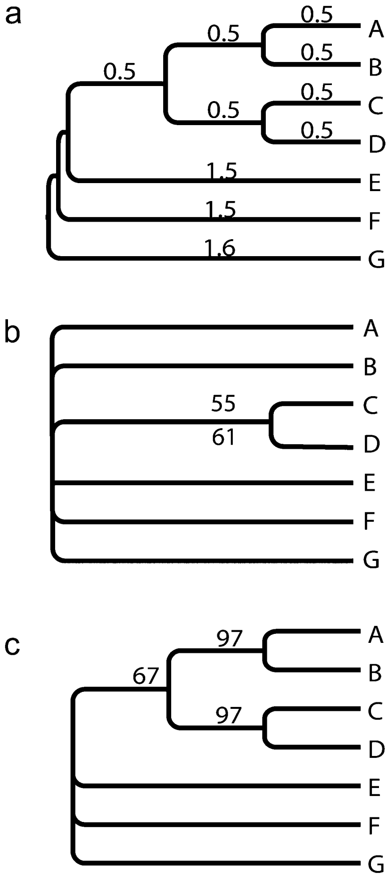

FIGURE 3. Effect of site removal on deep closely spaced cladogenetic events. a) The topology of the tree used to generate the simulated data (see text for details of simulation). b) Majority-rule consensus ML tree after before site removal and also after site removal using ML. The bootstrap support value for the unstripped alignments is above the line and the value after site removal using likelihood is below the line. c) Majority-rule consensus ML tree after removal of Bin10, the fastest evolving sites, according to the TIGER method.

This shows both the pitfall of the tree-based method and the advantage of our tree-independent method. The sites identified as most rapidly evolving by TREE-PUZZLE are those that do not agree with the initial tree inferred by ML. For this reason, removal of these sites does not clarify signals in the data, rather it merely strengthens the signal for the initial groupings. The tree-independent method, however, does not need any initial tree, therefore it is not biased toward any single tree and, instead, it picks out genuine signals in the data.

Thermus Data Set

The *Thermus* data set consists of 1273 aligned nucleotide positions from the 16S rRNA gene and is available as Supplementary Material. Using ML phylogenetic reconstruction implemented in PAUP4.0b10, we examined the differences in tree topology when removing characters judged to be rapidly evolving according to TIGER versus characters judged to be rapidly evolving according to TREE-PUZZLE (with a user-supplied tree, constructed using ML). In addition, we used the *reweight* command in PAUP to apply SACW (Farris 1969) and evaluate the effect that this approach had on the chances of recovering the correct tree. Using the original alignment of 1273 aligned positions (see Supplementary Material) and a GTR+I+G model of sequence evolution, we produced the phylogenetic tree in Figure 4a. Using the TREE-PUZZLE software, we categorized sites according to the GTR+I+G model using a discrete approximation to the gamma distribution to model ASRV, with a total of eight categories of sites. The category of sites with the fastest rate of evolution was removed from the alignment (a total of 186 sites) and the analysis was re-run using this newer shorter data set (consisting of 1087 sites). In this case, the same ATTRACT tree was recovered. The most significant difference between the two bootstrap analyses was that the bootstrap support values for the data set with the sites removed were much higher and each of the internal nodes was recovered in 100% of the bootstrap pseudoreplicates (Fig. 4b). It must be remembered that the rates of evolution of the sites had been determined using the ATTRACT tree, which is the tree that is obtained in the analysis of the unstripped data set.

In order to investigate the SACW method, we first inferred the most parsimonious phylogenetic tree with all sites equally weighted and using an exhaustive search of tree space and the parsimony optimality criterion. Support for this tree was assessed using 1000 rounds of bootstrap resampling, with the results summarized by a majority-rule consensus procedure. The most

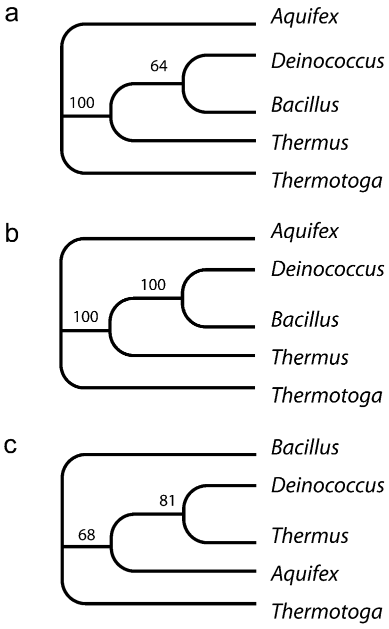

FIGURE 4. Analysis of the *Thermus* data set. a) Topology and support prior to site removal. b) The tree recovered after removal of sites identified by PUZZLE and using SACW. c) The resulting tree after removal of sites identified by TIGER.

parsimonious tree was once again the ATTRACT tree, with bootstrap support values of 92% for the grouping of *D. radiodurans* and *B. subtilis* and 96% for a clan containing *A. aeolicus* and *T. maritima*. Using the *reweight* command in the PAUP software, we weighted the characters according to their CI value on this tree. We then carried out another bootstrap resampling analysis to assess support for groups on the tree. This time the ATTRACT tree was once again recovered, but the support for all internal edges was at 100%.

We used the TIGER approach to identify rapidly evolving sites in the rRNA data set. We placed all sites from the alignment into one of eight bins according to how rapidly they evolve. The most rapidly evolving category of sites contained 108 sites and these were removed for subsequent ML analysis. Using the GTR+I+G model of sequence evolution on the remaining 1165 sites, we recovered the TRUE phylogenetic tree. After 1000 bootstrap replicates, we observed that the grouping of *D. radiodurans* and *T. aquaticus* in 81% of the replicates and the grouping of *T. maritima* and *B. subtilis* was observed in 68% of the replicates. The ATTRACT topology that groups *D. radiodurans* and *B. subtilis* together was seen in 19% of the replicates.

We carried out an additional analysis of the sites that are identified as being rapidly evolving. In all cases, we analyzed the most rapidly evolving sites on their own to see if there was any strong phylogenetic signal in those sites. As these sites are saturated for change, we do not expect to see a single phylogenetic signal, rather a number of incongruent signals. In our analyses, only the sites in Category 8 of the ML analysis contained any congruent phylogenetic signal. There was 80% bootstrap support for the TRUE tree in these sites. This result demonstrates that not only does such an ML approach result in strong support for the incorrect topology but also the characters that it discards contain more true phylogenetic signal than the characters that it retains. This needs to be viewed as a systematic error.

Primate Data Set

Our last analysis involves an 898 bp data set of 12 primate mitochondrial sequences (Hayasaka et al. 1988). Two equally most parsimonious trees, requiring 1153 steps can be obtained by analysis of these sequences. One of these trees places the human and chimpanzee (*Pan troglodytes*) together as sister taxa, whereas the other tree groups the Chimpanzee with the Gorilla. We wanted to investigate two things with this data set. First, in this case, where two phylogenetic hypotheses are strongly competing and where there is no greater support for one topology over the other, whether the TIGER approach would recover the accepted tree (human and chimp together) with confidence. Second, whether the tree-dependent method would be influenced strongly by the tree that is used to determine the evolutionary rate of the characters, or whether it would work well irrespective of the tree that it used initially for character reweighting. More specifically, we wished to see if using a particular tree in order to generate evolutionary rates would tilt the balance in favor of this topology in a bootstrap analysis. In other words, we wanted to explore whether character removal, based on an incorrect tree, could override the (albeit small) amount of extra support for the true tree and subsequently provide strong support for the incorrect tree.

When the tree that places *Homo* and *Pan* together was used in SACW in order to reweight characters according to the CI, then this same tree was recovered in the majority-rule consensus tree following bootstrapping. The bootstrap support value for this relationship was 79%, compared with a 51% value for the equally weighted data set (10,000 bootstrap replicates). We then used the other equally parsimonious tree in order to carry out character weighting for SACW. Using character reweighting according to the CI, we obtained a bootstrap support value of 77% for the grouping of *Pan* and *Gorilla* together. This shows that the initial tree that is used for character weighting can override small phylogenetic signals and because characters that tend not to agree with this initial tree are down weighted, this has a huge affect on which tree is supported in subsequent analyses.

It should be noted that in this particular case, the ML approach to site stripping was not as sensitive as the SACW approach and indeed was quite insensitive to the initial tree that was used for site classification. When the HC hypothesis tree was used, and the

TREE-PUZZLE software was asked to put sites into a total of 10 categories, then a total of 114 sites were put into the fastest category. When the CG hypothesis tree was used, then a total of 121 sites were put into the fastest category. Irrespective of the tree that was used to categorize sites, when category 10 was removed, we always recovered strong support for the HC hypothesis. We should note, however, that when the HC hypothesis was used to categorize sites, the resulting bootstrap support value was 99%, whereas when the CG hypothesis tree was used to categorize sites, then support for the HC hypothesis after site stripping was somewhat lower at 81%.

We used the TIGER approach to categorize characters in a tree-independent manner and to place them into a total of 10 bins according to their average split similarity with the other characters in the matrix. We removed the fastest category of sites, Bin10, which contained a total of 192 characters. We then used maximum parsimony bootstrapping to evaluate support for groups in the phylogeny. We recovered a grouping of *Homo* and *Pan*, with 87% support after 10,000 bootstrap replicates. The alternative hypothesis, grouping *Pan* and *Gorilla* together received 8.8% bootstrap support. Using ML, the HC hypothesis received 90% bootstrap support, whereas the CG hypothesis received 6% support.

Conclusion

In this article, we report the development of an algorithm, based on those of Le Quesne (1989), Wilkinson (1998) and Pisani (2004) that uses similarity in the pattern of character-state distributions between characters as a proxy for speed of evolution in a data matrix of homologous characters. We expect that rapidly evolving characters are likely to lose some, most, or all of their phylogenetic information and will tend to have a character-state distribution that is closer to random than the distribution expected from a more slowly evolving character. A character is assumed to be rapidly evolving if it has a character-state distribution pattern that, on average, is not very similar to the patterns observed in other characters. This assumption is only likely to hold in some (though probably very many) situations. Specifically, in a data matrix where each character is effectively randomized, due to a very rapid rate of evolution or a long evolutionary timespan, we do not expect that this kind of approach will work well. Notwithstanding this caveat (which is a situation that would confound most, if not all, phylogenetic methods), we have observed some very interesting and desirable properties of this approach that make it a useful addition to the phylogenetic arsenal.

The TIGER approach identified differing patterns of ASRV, distinguishing alignments that had extreme variation in among-site evolutionary rates from those alignments that had a more even distribution of rates. Additionally, it was able to identify subtleties in the data such as the four clusters of rates in each alignment—a by-product of the simulation process.

The TIGER approach helped improve the fit of the data to the correct tree in our simulations. Removing sites that TIGER identified as being rapidly evolving resulted in a better fit of the data to good trees and worse fit of the data to bad trees, with the true tree being affected most positively. Additionally, using the TIGER approach, we could improve the resolution of deep lineages where rapid cladogenesis resulted in very difficult-to-resolve branches. Worryingly, the likelihood approach to removing rapidly evolving sites proved to be problematic—the sites that were removed were those that did not agree with the initial tree, resulting in a situation where, out of 100 simulations, there was little improvement in the recovery of the deep diverging rapid cladogenesis tree.

For the ribosomal RNA data set, we observed a number of issues. First, the TIGER approach seems to have some merit as an approach to removing sites that interfere with phylogeny reconstruction. Additionally, two other tree-dependent methods—site identification using a ML model of ASRV and site identification using the fit of the data to an initially constructed phylogenetic tree—are systematically biased toward favoring the first phylogenetic tree they construct. We, therefore, feel it is important to be cautious when using tree-based methods of assigning evolutionary rates to sites, unless the evolutionary history is known with certainty. We note, however, that a sophisticated compositionally heterogeneous model of sequence evolution is capable of identifying the correct topology for this data set, without the necessity of deleting or reweighting characters (Foster 2004).

The point concerning tree-based attribution of evolutionary rate is quite clearly exemplified by the primate mitochondrial data set and maximum parsimony analysis. Here two hypotheses are equally good when using the parsimony criterion. Character reweighting based on one of the two equally most parsimonious trees will skew subsequent analyses toward supporting this particular topology, whereas the same is true for the alternative topology. Ultimately, the TIGER analysis, which does not use a tree, recovers the correct phylogenetic hypothesis (which has been confirmed by numerous other studies) while not using an a priori determined phylogenetic tree in order to do so. We find that support for the grouping of *Pan* and *Gorilla*, to the exclusion of *Homo* is an artifact that is due to the most rapidly evolving sites. This also shows that site stripping can be beneficial for resolution of recent relationships, not just ancient relationships. We should also state here that ML analysis of this data set produces the correct tree, using the Tamura–Nei model, with bootstrap support for (*Homo, Pan*) at 94%.

Ultimately, TIGER is an interesting device for identifying characters that do not agree with the majority of the data. We argue here that in many cases this disagreement can be diagnostic of rapid evolution. At the very least, the converse is likely to be true—rapid character evolution is likely to produce a pattern that is not very similar to other characters. Removal of these kinds of characters can greatly improve the accuracy of successive phylogenetic analysis by removing conflicting signals.

There are surely limits to what site removal can accomplish and with certainty site removal is a poor alternative to precise model definition. However, precise model definition comes with a cost. Models that adequately describe the evolution of a set of DNA or protein sequences might, of necessity, be very parameter rich (e.g., using a combination of Dirichlet processes for both site rate identification, Huelsenbeck and Suchard 2007, and site-specific profiling, Lartillot and Philippe 2004, as implemented in the CAT model) and require a large amount of sequence before they become statistically consistent. The most commonly used models of sequence evolution are often inadequate to describe the evolution of the sequences being studied. Model selection approaches often "max-out," where the most parameter-rich method of analysis is the one that is selected by a likelihood ratio test, Akaike information criterion or Bayesian information criterion (Keane et al. 2006), indicating that perhaps there are not enough parameters available. Therefore, it might not be an option to use a precisely described model. In the case of the rRNA sequences being analysed in this study, the raw alignment exhibited significant compositional heterogeneity and none of the standard, compositionally homogeneous time-reversible models of sequence evolution can adequately account for this heterogeneity. By identifying and removing the most rapidly evolving characters, the models are better able to account for the evolution of the sequences.

We have no good theoretical framework for knowing precisely how many sites to remove from an alignment. It is likely that in many cases there is no need to strip out any sites. At the moment, we only have an ad hoc approach to site stripping and this must be considered a major problem. Ideally, we wish to remove sites that only contribute noise and do not contribute any phylogenetic signal. Our recommendation is the testing of congruence across a progressively larger number of the fastest evolving characters using methods such as the permutation-tail-probability test (Faith and Cranston 1991) or likelihood mapping (Strimmer and von Haeseler 1997). This would result in the removal of sites that show very little consistency with the rest of the data and very little consistency with one another. However, this is also ad hoc and should be used as nothing more than a rule-of-thumb.

We also note that bootstrap support values or Bayesian clade probability values are probably meaningless when there is a directed attempt to remove sites that disagree with the rest of the data. It is likely that the support values will tend to increase when incongruent data are removed. When we use bootstrap support values, we wish to show that the data have been strongly influenced by the character removal; we do not wish to imply that bootstrapping should follow character removal, as, in most cases, the resulting bootstrap scores are likely to be higher.

Given that there are limits to what can be achieved by character removal, we conclude by advising that this method should be used as one part of an overall experimental programme of data exploration. We expect that additional tree-independent methods of analyzing evolutionary rate variation can be developed.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Funding

C.A.C. is funded by a Science Foundation Ireland Research Frontiers Programme award [07/RFP/EEEOBF654] to J.O.Mc.I.

Acknowledgments

The work was supported by the NUIM High-performance computing resource and the Irish Centre for High End Computing.

References

Adachi J., Hasegawa M. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J. Mol. Evol. 40:622–628.

Begun D. 1992. Miocene fossil hominids and the chimp human clade. Science. 257:1929–1933.

Brinkmann H., Philippe H. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Delsuc F., Brinkmann H., Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Ebersberger I., Galgoczy P., Taudien S., Taenzer S., Platzer M., von Haeseler A. 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24:2266–2276.

Embley T., Thomas R., Williams R. 1993. Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between Thermus and Deinococcus. Syst. Appl. Microbiol. 16:25–29.

Faith D., Cranston P. 1991. Could a cladogram this short have arisen by chance alone?: on permutation tests for cladistic structure. Cladistics. 7:1–28.

Farris J. 1969. Successive approximations approach to character weighting. Syst. Zool. 18:374–385.

Fischer W. M., Palmer J. D. 2005. Evidence from small-subunit ribosomal RNA sequences for a fungal origin of Microsporidia. Mol. Phylogenet. Evol. 36:606–622.

Fitch W., Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.

Foster P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Grehan J., Schwartz J. 2009. Evolution of the second orangutan: phylogeny and biogeography of hominid origins. J. Biogeogr. 36: 1823–1844.

Hansmann S., Martin W. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int. J. Syst. Evol. Microbiol. 50:1655–1663.

Hayasaka K., Gojobori T., Horai S. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. Mol. Biol. Evol. 5: 626–644.

Hirt R., Logsdon J., Healy B., Dorey M., Doolittle W., and Embley T. 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. U.S.A. 96:580–585.

Huelsenbeck J., Suchard M. 2007. A nonparametric method for accommodating and testing across-site rate variation. Syst. Biol. 56: 975–987.

Jukes T., Cantor C. 1969. Evolution of protein molecules. In: Munro, editor. Mammalian protein metabolism. New York: Academic Press. p. 240–253.

Keane T., Creevey C., Pentony M., Naughton T., McInerney J. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. 6.

Kluge A. G., Farris J. S. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1–32.

Kostka M., Uzlikova M., Cepicka I., Flegr J. 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinformatics. 9:341.

Kuhner M. K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

Lartillot N., Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Le Quesne W. 1969. A method of selection of characters in numerical taxonomy. Syst. Biol. 18:201–205.

Le Quesne W. 1989. The normal deviate test of phylogenetic value of a data matrix. Syst. Zool. 38:51–54.

Maddison D. R. 2004. Testing monophyly of a group of beetles. Study 1 in Mesquite: A Modular System for Evolutionary Analysis. Version 1.04. Available from: http://mesquiteproject.org.

Maidak B. L., Olsen G. J., Larsen N., Overbeek R., McCaughey M. J., Woese C. R. 1996. The ribosomal database project (RDP). Nucleic Acids Res. 24:82–85.

Meacham C. A. 1994. Phylogenetic relationships at the basal radiation of angiosperms: further study by probability of character compatibility. Syst. Bot. 19:506–522.

Mooers A., Holmes E. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15: 365–369.

Olsen G. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor. Symp. Quant. Biol. 52:825–837.

Olsen G. J., Matsuda H., Hagstrom R., Overbeek R. 1994. fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10:41–48.

Olsen G., Pracht S., Overbeek R. 1998. DNArates. Version 1.1.

Philippe H., Zhou Y., Brinkmann H., Rodrigue N., Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5:50.

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst. Biol. 53:978–989.

Puigbo P., Garcia-Vallve S., McInerney J. O. 2007. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics. 23: 1556–1558.

Rambaut A., Grassly N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Ruvolo M. 1997. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol. Biol. Evol. 14:248–265.

Satta Y., Klein J., Takahata N. 2000. DNA archives and out nearest relative: the trichotomy problem revisited. Mol. Phylogenet. Evol. 14:259–275.

Schmidt H., Strimmer K., Vingron M., von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502–504.

Schwartz J. 1984. The evolutionary relationships of man and orangutans. Nature. 308:501–505.

Shoshani J., Groves C., Simons E., Gunnell G. 1996. Primate phylogeny: morphological vs molecular results. Mol. Phylogenet. Evol. 5:102–154.

Strimmer K., von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 94:6815–6819.

Townsend J. P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Wilgenbusch J. C., Swofford D. 2003. Inferring evolutionary trees with PAUP*. Curr. Protoc. Bioinformatics. Chapter 6: Unit 6.4.

Wilkinson M. 1998. Split support and split conflict randomization tests in phylogenetic inference. Syst. Biol. 47:673.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang Z. 1996. Among-site rate variation and its impact in phylogenetic analyses. Trends Ecol. Evol. 11:367–372.