CHAPTER

An Introduction to Phylogenetics and the Tree of Life

2

Tom A. Williams[*,1], **Sarah E. Heaps**[*,†]

[*]*Institute for Cell and Molecular Biosciences, The Medical School, Newcastle University, Newcastle upon Tyne, United Kingdom*

[†]*School of Mathematics and Statistics, Newcastle University, Newcastle upon Tyne, United Kingdom*

[1]*Corresponding author: e-mail address: tom.williams2@ncl.ac.uk*

1 INTRODUCTION

Phylogenetic trees are fundamental to organising, understanding and testing hypotheses about the evolution of biological diversity. Early phylogenies were based on morphology: useful for multicellular eukaryotes, but much less so when inferring relationships among prokaryotes or among the different branches of the tree of life, most of which is microbial. Although comparisons of biochemical properties provided some insight into bacterial relationships, they proved unreliable at deeper taxonomic levels, and by 1960, it seemed that a universal phylogeny was out of reach, with the only unambiguous division in the microbial world separating the eukaryotes from the structurally simpler prokaryotes (Stanier & van Niel, 1962). This situation changed completely with the advent of molecular sequencing, which provided biologists with a rich new source of information about evolutionary history (Zuckerkandl & Pauling, 1965) that was just as relevant for prokaryotes and microbial eukaryotes as for animals, plants and fungi. The greatest early success of the sequencing era came when Carl Woese and colleagues showed that the ribosomal RNA (rRNA) sequences of prokaryotes clustered into two groups that were at least as divergent from each other as they were from the rRNA genes of eukaryotes, demonstrating that the prokaryotes comprised two distantly related lineages, the Bacteria and Archaea (Woese & Fox, 1977; Woese, Kandler, & Wheelis, 1990).

The discovery of the Archaea demonstrated the power of sequence data for investigating relationships among prokaryotes, and in the intervening years, analyses of rRNA and, more recently, whole genome sequences have become standard approaches in molecular evolution and systematics. The advantages of sequences over other types of data—such as morphology, physiology and biochemistry—for inferring phylogenies are clear, particularly in the case of prokaryotes, microbial eukaryotes and viruses. Sequence data are highly informative, and today, millions of characters

Methods in Microbiology, Volume 41, ISSN 0580-9517, http://dx.doi.org/10.1016/bs.mim.2014.05.001

can be analysed simultaneously in single-gene or concatenated multiple sequence alignments. With contemporary (i.e. "next-generation") sequencing technologies, new sequences are cheap and relatively easy to obtain, and the number of sequences in public databases is so large that the data needed to address many unanswered or new evolutionary questions is already available. From the biological point of view, one of the greatest strengths of sequence-based phylogenies is the capability they provide for inferring relationships among organisms for which other meaningful points of comparison do not really exist. For example, all cellular organisms synthesise proteins on a ribosome, so a tree based on rRNA can include the bacterium *Escherichia coli*, the archaeon *Sulfolobus solfataricus* and the eukaryotes *Saccharomyces cerevisiae* and *Homo sapiens*, organisms which would otherwise be difficult or impossible to fit into a single, meaningful classification. Much of the early excitement around sequence-based phylogenies was due to their potential use in constructing a universal tree of life that would include all cellular organisms (Woese et al., 1990). In fact, much progress has been made on this issue in the sequencing era (Embley & Martin, 2006), although the relationships among the major lineages of cellular life remain actively debated (Ciccarelli et al., 2006; Cox, Foster, Hirt, Harris, & Embley, 2008; Foster, Cox, & Embley, 2009; Gribaldo, Poole, Daubin, Forterre, & Brochier-Armanet, 2010; Williams, Foster, Cox, & Embley, 2013; Williams, Foster, Nye, Cox, & Embley, 2012), as will be discussed in more detail below.

Another major advantage of sequence data is that it is unambiguously categorical: there are 4 possible states (A, C, G and T) for each nucleotide position, and 20 for each amino acid. As a result, sequences are considerably more amenable to rigorous statistical analysis than phenotypic characters, whose states must often be encoded in a somewhat arbitrary way (Stevens, 1991). This categorical character of sequence data is important—sequences may represent the richest source of information about prokaryotic evolution currently available, but as with other kinds of data, they can be positively misleading (Felsenstein, 1978) if analysed using inappropriate methods. Thus, while obtaining sequences is easier than ever before, careful phylogenetic analysis using the best available methods remains a time-consuming and potentially challenging task. With the right tools in hand, the process of building phylogenies can be relatively straightforward, but it is not automatic—each step (Figure 1), from collecting and aligning sequences to choosing the most appropriate phylogenetic model and building the trees, involves making decisions that may change the outcome. The aim of this chapter is to provide a practical guide to each of these steps and to introduce some of the best and most frequently used software for phylogenetic analysis. In order to make our discussion more concrete, we will work through an attempt to resolve one of the most interesting and controversial questions in phylogenetics—the relationship between Bacteria, Archaea and Eukarya, the three major lineages of cellular life.

Following Woese's discovery of the Archaea, the question naturally arose as to which of the prokaryotic groups (Bacteria or Archaea), if either, was more closely related to the eukaryotes. This question is complex because of the symbiogenic origins of eukaryotic cells (Sagan, 1967): all eukaryotes have a mitochondrion or

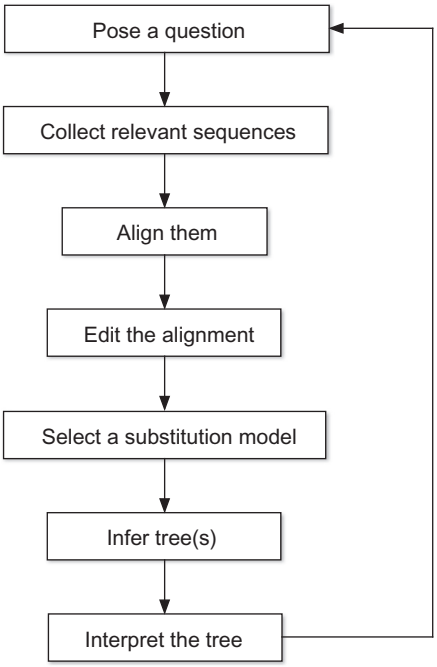


FIGURE 1

A workflow for phylogenetic analysis. The outline of a generic approach that can be used to address many questions in phylogenetics. In this chapter, we decided to investigate the relationship between Archaea and eukaryotes. This decision motivated our selection of SSU ribosomal RNA sequences for analysis, and the properties of that dataset suggested a particular approach to alignment and phylogenetic modelling. The resulting trees were then interpreted in the light of the original question, helping to focus discussion on their most relevant features.

mitochondria-related organelle that descends from a free-living alphaproteobacterium (Andersson et al., 1998; Esser et al., 2004), and many also possess a plastid descended from cyanobacteria (Martin et al., 2002). Thus, different compartments of eukaryotic cells have different phylogenetic origins. However, the genetic and ultrastructural similarities between mitochondria and plastids and their bacterial relatives are sufficiently strong that a broad consensus now exists on the origins of these organelles. Instead, contemporary debate focuses on the phylogenetic affinity of the eukaryotic nucleocytoplasmic lineage, which is often taken to represent the original host cell for these bacterial partners (Embley & Martin, 2006). Early analyses of rRNA led by Woese and coworkers (Woese, 1987; Woese & Fox, 1977) suggested that each of the three "domains" of life—Bacteria, Archaea and Eukarya—were monophyletic; in other words, that all Archaea, for example, are more closely related to each other than any of them are to Bacteria or eukaryotes. Combined with

evidence from analyses of ancient gene duplications which suggested that the root of this "universal tree" lay on the branch leading to Bacteria (Gogarten et al., 1989; Iwabe, Kuma, Hasegawa, Osawa, & Miyata, 1989), these results led to the now-famous rooted three-domains tree (Woese et al., 1990), in which the Eukarya and Archaea form monophyletic sister groups to the exclusion of Bacteria. This tree represents the dominant hypothesis for the deepest branches of the tree of life and as such plays an important role in modern evolutionary biology. In this chapter, our goal will be to investigate whether it remains the most strongly supported hypothesis given currently available sequence data and statistical models.

2 STEP 1: POSING A QUESTION

The first step in any phylogenetic analysis is to frame the question you are attempting to answer, or the hypothesis you wish to test. This provides a rationale for choosing the sequences to analyse and a framework for interpreting the results. In the present case, our aim is to test whether the three-domains tree is supported from contemporary sequence data. Consulting the literature, we can see that a number of alternatives to the three-domains tree have been proposed. Several of these involve the placement of the eukaryotes (or at least, the set of conserved eukaryotic genes encoding the ribosome and related cellular components) within the Archaea, as the sister group to the Crenarchaeota (Lake, Henderson, Oakes, & Clark, 1984), the Thaumarchaeota (Kelly, Wickstead, & Gull, 2011), or the Thermoplasmatales (Pisani, Cotton, & McInerney, 2007). It may be worth keeping some of these alternative hypotheses in mind as we analyse and interpret our results.

3 STEP 2: CHOOSING RELEVANT SEQUENCES

Since our question addresses the relationships between domains, we will need to include sequences from all three domains of life—Bacteria, Archaea and Eukarya. A pervasive problem that affects all attempts to resolve inter-domain trees, as well as many smaller-scale phylogenetic analyses, is that individual genes often do not contain sufficient phylogenetic information (or signal) to produce a well-resolved species tree. As a result, many modern analyses attempt to combine signal from multiple genes using either "supermatrices" or "supertrees". In the supermatrix approach (de Queiroz & Gatesy, 2007), alignments from individual genes are simply concatenated and analysed as if they represented one large gene, although some aspects of the evolutionary model may be allowed to vary among the constituent genes. In the supertree approach (Bininda-Emonds, 2004), individual trees are inferred separately for each gene, and the information in these trees is then combined to produce a consensus estimate of the species tree. These methods have a number of advantages when the goal is to infer a species tree: for example, trees inferred from supermatrices are usually very well resolved, with high support values (see Section 5.3 and 5.4.) for most or all branches. However, these methods also add an additional layer of complexity to phylogenetic analyses, and they introduce a number of

difficulties and caveats. In particular, the supermatrix approach necessarily assumes that all the genes in the matrix are evolving on the same underlying species tree, and violation of this assumption (e.g. due to horizontal gene transfer in some genes) can lead to the recovery of trees that are strongly supported but incorrect (see, e.g., Moreira & Lopez-Garcia, 2005). For more information, see chapter 'Reconciliation Approaches to Determining HGT, Duplications, and Losses in Gene Trees' by Kamneva and Ward, in this volume provides a discussion dealing with these cases. Here, we will sidestep these issues by focusing on the phylogenetic analysis of just one gene—that encoding the RNA component of the small subunit of the ribosome (16S rRNA). This is the most frequently used gene in prokaryotic phylogeny (see, for instance, chapter 'The All-Species Living Tree Project' by Yarza and Munoz, in this volume) and is also well suited for analysis of inter-domain relationships because of its ubiquity and very slow evolutionary rate. The phylogenetic methods we will apply can be easily extended to model the evolution of protein-coding sequences; for those interested in building supermatrices or supertrees, we recommend first consulting the extensive literature on these methods, to which Rannala and Yang (2008) provide an excellent entry point. Finally, it is important to bear in mind that the analysis of any single gene, no matter how broadly distributed or well conserved, provides only one perspective on the evolution of the organisms that encode it. Our aim here is to thoroughly analyse a single gene in order to introduce some of the most important concepts in phylogenetic analysis; state-of-the-art work typically involves a much larger sample of genes to provide a much more robust estimate of species phylogenies—the interested reader should consult chapter 'Reconciliation Approaches to Determining HGT, Duplications, and Losses in Gene Trees' by Kamneva and Ward in this volume.

3.1 OBTAINING 16S rRNA SEQUENCES FOR BACTERIA, ARCHAEA AND EUKARYA

Due to their historical importance as phylogenetic markers in the era before complete genome sequencing, 16S rRNA sequences are available for a very wide range of Bacteria, Archaea and Eukarya. Here, we will make use of a publicly available dataset of 36 sequences that was analysed by one of the present authors in Williams et al. (2012). The sequences can be obtained from the public repository Dryad (see Table 1, which provides links to the data, software and Web resources referenced in this chapter). This dataset is useful for our purposes here because it is relatively small, and so each step of the analysis can be performed quickly. It is also an interesting dataset for illustrating the impact that different decisions made during the analysis can have on the inferred phylogeny, as we will see in section 7 below. From the relevant Dryad page, download and extract the archive "rrna.tar". Our re-analyses will require the "ssu.fa" and "ssu_all.fa" files inside this archive. These files are partially redundant: "ssu.fa" (hereafter the SSU dataset) contains 32 small subunit rRNA sequences from Bacteria, Archaea and Eukarya; "ssu_all.fa" contains the same 32 sequences and 4 new sequences from recently sequenced Archaea (Thaumarchaeota, Aigarchaeota and Korarchaeota; hereafter

Table 1 The Freely Available Resources Referenced in This Chapter

Resource	Description	Link
16S ribosomal RNA sequence dataset	36 rRNA sequences from Bacteria, Archaea and eukaryotes analysed in Williams et al. (2012)	http://datadryad.org/resource/doi:10.5061/dryad.0hd1s
Muscle	Popular alignment tool	http://www.drive5.com/; http://www.ebi.ac.uk/Tools/msa/muscle/
Jalview	Alignment viewer	http://www.jalview.org/
TrimAl	Alignment masking (editing) program	http://trimal.cgenomics.org/; http://phylemon2.bioinfo.cipf.es/
jModelTest2	Model comparison	https://code.google.com/p/jmodeltest2/
RAxML	Maximum likelihood inference of phylogeny	http://www.exelixis-lab.org/; http://www.phylo.org/sub_sections/portal/ (Web server)
PhyloBayes, PhyloBayes-MPI	Bayesian inference of phylogeny; implements the CAT and CAT+GTR models	http://www.phylobayes.org; http://www.phylo.org/sub_sections/portal/
Dendroscope	Tree viewer	http://ab.inf.uni-tuebingen.de/software/dendroscope/
FigTree	Tree viewer	http://tree.bio.ed.ac.uk/software/figtree/
AWTY (Nylander, Wilgenbusch, Warren, & Swofford, 2008)	Graphical exploration of MCMC convergence in Bayesian phylogenetic inference	http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php

the SSU+TAK dataset). We will analyse these two datasets using exactly the same protocol, in order to investigate whether slight changes in taxon sampling can influence the inferred tree.

3.2 A NOTE ON THE AVAILABILITY AND USE OF DATA AND METHODS

Openness and reproducibility are fundamental to scientific progress. In principle, ensuring reproducibility in phylogenetic analyses should be straightforward because sequences and alignments can be easily shared over the internet. Further, the analyses are all computational, which should help to limit the role of human bias or error. Unfortunately, reproducibility in phylogenetics is generally low because researchers often fail to make their datasets publicly available (Drew et al., 2013). One of the many benefits of publishing the raw materials of your phylogenetic analyses in a public repository is that it allows others to build on or refine your work; thus, beyond

the immediate context of this practical, we encourage you to explore and experiment with the datasets we use here. In the same spirit, all of the methods used in this tutorial are freely available and open source; this ensures that the scientific community can investigate and verify the algorithms used, and represents another important component of ensuring reproducibility in phylogenetic research.

Most phylogenetic software is designed to run on Unix-based systems, of which the most popular today are OS X and the various distributions of Linux. If you have access to one of these systems, you will find installing and running the tools used in our tutorial much more straightforward than if you are limited to a computer running Windows. As the purpose of this chapter is to introduce the basic principles of phylogenetics, and not to provide a Unix handbook, we have designed our tutorial so that all the analyses can be run using a Web browser and one or more of the Web servers that various laboratories generously provide for free online (see Table 1). However, we strongly encourage the interested reader to run the analyses locally, as familiarity with Unix and the ability to compile and run academic software is a prerequisite for any serious phylogenetic work. Throughout the rest of this chapter, we generally do not provide the exact commands you will need to run the analyses—for information on the basic operation of the different tools, refer to the available documentation on their Web sites (Table 1).

If you want to run the analyses locally but only have access to a Windows machine, one option is to set up a virtual machine running a variant of Linux, such as Ubuntu (http://www.ubuntu.com/). VirtualBox (https://www.virtualbox.org/) is a free and relatively easy-to-use virtualisation program available for Windows and other operating systems that will help you to set up a virtual machine. If you are planning to do a lot of bioinformatic analysis on your computer, installing a Linux distribution directly to your hard drive (i.e. dual-booting with Windows) may be the best option, as this will allow your analyses to make full use of the underlying hardware.

4 STEP 3: ALIGNING SEQUENCES AND EDITING THE ALIGNMENT

With the SSU rRNA datasets in hand, the next step in our analysis is to align the sequences. Sequence alignment is critical because although we assume that the rRNA genes of contemporary organisms are all descended from an ancestral gene that was present in their common ancestor, these genes have experienced not only point mutations (i.e. one nucleotide being substituted for another) but also insertions and deletions (gains or losses of one or more nucleotides) during their evolutionary history. As a result, the sequences from different organisms may have different lengths, and so (for example) the 100th nucleotide in the *H. sapiens* sequence is not necessarily equivalent to the 100th nucleotide of the *E. coli* sequence, or to the 100th nucleotide of the ancestral gene. Our inference of the phylogenetic tree will be based on comparisons of *homologous* nucleotides—that is, the nucleotides in contemporary genes that are descended from an ancestral nucleotide in their

common ancestor (see Figure 2 for a visual clarification of this important idea). The purpose of sequence alignment is to arrange these homologous nucleotides into columns, or sites, in the final alignment file, ready for subsequent phylogenetic analysis.

Since the evolutionary history of a gene family is almost never known with certainty, the alignment that we infer will represent a hypothesis (or rather, a set of hypotheses) about that history that is likely to be wrong in some respects. Clearly, the quality of the sequence alignment is absolutely critical to the reliability of any subsequent phylogenetic analysis. Consideration of even the very small-scale example in Figure 2 indicates that the number of possible alignments and substitution histories

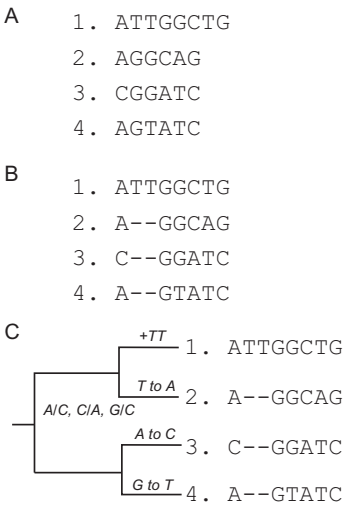


FIGURE 2

Sequence alignments represent inferences of homology, and as such contain information about the evolutionary history of a gene family. (A) Four related gene sequences sampled from contemporary organisms. (B) One possible alignment of the three sequences, identifying eight homologous sites (alignment columns). If this alignment is correct, then it is likely that sequence 1 contains a "TT" insertion after the first position; thus, its fourth nucleotide (G) is equivalent to the second position G in the three other sequences. (C) One possible evolutionary history for these sequences, based on the alignment in (B). The homologous nucleotides in columns 6 and 8 of the alignment are C, G in sequences 1 and 2, but A, C in sequences 3 and 4. This provides some evidence that sequences 1 and 2 are more closely related to each other than either is to 3 or 4 and vice versa. Inferred ancestral sequences and an associated substitutional history are mapped onto the interior branches of the phylogenetic tree. Inferences of homology can be made with reference to the reconstructed ancestral sequences: for example, the nucleotides T, A, T and T in column 7 of the alignment in (B) are homologous because they all descend from the same nucleotide in position 5 of the ancestral sequence.

for any real sequence dataset is likely to be astronomical. Given this context, the aim of sequence alignment is generally not to find the best global alignment, but rather to produce the best alignment possible within a reasonable computational time, and also to identify (and somehow deal with) the regions of this alignment that are likely to be problematic or unreliable. A large number of sequence alignment programs have been developed for this purpose; their relative strengths and weaknesses are reviewed in Thompson, Linard, Lecompte, and Poch (2011). In this tutorial, we will use the Muscle aligner (Edgar, 2004), a popular choice that performs reasonably well in alignment benchmarks. Muscle can either be downloaded or run locally, or via online Web servers such as Phylemon—see Table 1 for details. An excellent alternative to using a single alignment tool such as Muscle is to align your sequences with a number of different methods and then produce a consensus alignment; this is a slightly more elaborate and time-consuming approach, and so we will not use it here. Nonetheless, it does tend to produce better alignments and may be something to consider for your own analyses in the future (see Wallace, O'Sullivan, Higgins, & Notredame, 2006 for a discussion of this approach). An alternative strategy, which is generally more complex than sequential sequence alignment and tree building, is to simultaneously infer the alignment and phylogenetic tree from unaligned sequence data (see, e.g., Redelings & Suchard, 2005 for further details).

For the time being, align both the SSU and SSU+TAK datasets using Muscle with the default parameters, either on your own computer or online. If you open the outfile file in a text editor, you will see that the program has introduced gaps ("-") in the sequences in order to align the homologous positions against each other. This will be most obvious if you use an alignment viewing program such as Jalview (Waterhouse, Procter, Martin, Clamp, & Barton, 2009; Table 1), which also allows you to colour the nucleotides by type or sequence identity in order to visualise the parts of the alignment that are more or less similar to each other (more or less *conserved*). Viewing the alignment in this way, it will soon become obvious that some regions are much more conserved than others. Conserved regions are relatively easy to align because few changes have taken place and homologous positions are usually straightforward to identify. However, you may notice other parts of the alignment that contain many more gaps, and where the case for positional homology looks much weaker. Since SSU rRNA is one of the most conserved genes known, the alignment will look quite reasonable overall, although you may notice more ambiguity near the 3′-end of the molecule (e.g. beyond position 1100 in the alignment); often alignments will contain many more of these disordered regions. This observation raises the question of how these less reliable parts of the alignment should be treated. Opinions differ on this issue: on the one hand, including these regions in your analysis might introduce noise or other non-historical signal, potentially reducing the reliability of the inferred tree (Talavera & Castresana, 2007). On the other hand, over-zealous removal of the less conserved regions of the alignment is likely to remove genuine signal that could help to resolve the tree. In this tutorial, we will remove these regions of the alignment, but we encourage you to investigate other perspectives on this issue (see, e.g., Lee, 2001).

Once the decision to remove poorly aligning regions has been made, the next question is how these regions should be selected. This is sometimes done manually, by visualising the alignment in a program such as Jalview and simply deleting the columns or regions that look unreliable, and to which no obvious manual improvements can be made. The alternative to this manual approach is to use an automated "masking" program that uses some set of rules to decide which parts of the alignment to remove. In principle, expert opinion may be as good as or better than the heuristics employed by these programs. However, use of masking programs has the great benefit of reproducibility: if you include the program and settings used in the Methods section of your paper, then others will be able to see exactly what you did. For that reason, we will use an automated masking program here, although—as with the issue of whether the alignment should be edited at all—opinions differ on this subject. We will use trimAl (Capella-Gutierrez, Silla-Martinez, & Gabaldon, 2009), a masking program which uses the level of conservation of the sequence alignment to determine which of several modes (of varying strictness) should be used to select the regions to be deleted. As with the other steps in this tutorial, it is well worth exploring the alternatives to this program in your own research. Some of the most widely used examples include Gblocks (Castresana, 2000) and BMGE (Criscuolo & Gribaldo, 2010). TrimAl can be run locally or via the Phylemon Web server (Table 1); you should select the "automated1" mode when trimming the SSU and SSU+TAK datasets and then inspect the edited alignments using Jalview.

5 STEP 4: THE THEORY OF FITTING AND SELECTING A PHYLOGENETIC MODEL

Steps 4 and 5 in our analysis involve fitting phylogenetic models in one of two inferential frameworks (frequentist or Bayesian) and then comparing these models in a principled fashion. We first provide an introduction to the statistical background of the models and the basis for statistical inference and model comparison. We then illustrate these ideas through analyses of our SSU and SSU+TAK datasets. The impatient reader may wish to skip forward to the practical guidelines in section 6, dipping back into the statistical theory to gain a better understanding of the underpinning ideas.

5.1 MARKOV NUCLEOTIDE SUBSTITUTION MODELS

Statistical models for molecular evolution provide a structured framework for describing the complex and uncertain relationships between the molecular sequences for a collection of species. They are generally based on a set of simplifying assumptions about the evolutionary process that allows these relationships to be explained using a reasonably small number of parameters. When the values of these parameters, including the phylogeny, are fixed at one possible set of values they might take, the model attaches probabilities to the different alignments that could be observed, and these probabilities will be different for different sets of parameters. When we fit models to data, we then learn which parameter values are more or less consistent with the observed data according to the model we have assumed.

Most phylogenetic models assume that substitutions in molecular sequences can be modelled using continuous time Markov processes (CTMPs). A crucial property of these models, called the Markov property, is that the future state of the process (i.e. the remaining time before the next substitution and the character state resulting from the next substitution) depends only on the current state of the process and not on its past given this current state. Consider a single site of our sequence of nucleotides evolving over time on one branch of the tree. The CTMP describing the substitutions along that branch can be characterised by an instantaneous rate matrix Q. Given the length of the branch in question, Q determines the transition probabilities of the process. A transition probability is the probability of being in state j at the end of the branch given that the process was in state i at the start of the branch. These probabilities, for all possible combinations of states i and j, form a transition matrix. Standard models assume that the CTMP on any particular branch of the tree is time reversible and stationary. The assumption of reversibility means that the transition matrix would be the same if the process ran forwards or backwards in time. The assumption of stationarity means that (i) the probability of the character state being A, C, T or G does not change over time, that is, over the duration of the branch (these probabilities are called the stationary distribution, or probabilities, of the process) and (ii) the probability of transitioning from one state to another in a window of time depends only on the size of the window and not on its position in time. A consequence of making these assumptions is that the likelihood function of a tree is not affected by the position of the root. This makes it impossible to learn about the root position. Nevertheless, the assumptions are made in order to simplify the mathematics underpinning the model. In particular, they allow the instantaneous rate matrix Q to be decomposed in the following form:

$$Q = \begin{pmatrix} - & \rho_{AG}\pi_G & \rho_{AC}\pi_C & \rho_{AT}\pi_T \\ \rho_{AG}\pi_A & - & \rho_{GC}\pi_C & \rho_{GT}\pi_T \\ \rho_{AC}\pi_A & \rho_{GC}\pi_G & - & \rho_{CT}\pi_T \\ \rho_{AT}\pi_A & \rho_{GT}\pi_G & \rho_{CT}\pi_C & - \end{pmatrix}$$

where the dashes on the diagonal signify that these elements are defined to ensure the rows sum to zero. The terms ρ_{AG}, ρ_{AC}, ρ_{AT}, ρ_{GC}, ρ_{GT} and ρ_{CT} are called exchangeability parameters and they can be interpreted as the instantaneous rates of change between the different character states. The terms π_A, π_G, π_C and π_T are the probabilities from the stationary distribution of the process. Note that there is a different instantaneous rate of change between all possible pairs of nucleotides. This is called the general time-reversible (GTR) model (Tavaré, 1986). Other commonly used models for DNA are special cases. For example, the HKY85 model (Hasegawa, Kishino, & Yano, 1985) is a special case with only two distinct exchangeability parameters: the transition rate α and the transversion rate β. In other words, under the HKY85 model, $\rho_{AG} = \rho_{CT} = \alpha$ and $\rho_{AC} = \rho_{AT} = \rho_{GC} = \rho_{GT} = \beta$. The K80 model (Kimura, 1980), like the HKY85 model, assumes different rates for transitions and transversions. However, it imposes the additional constraint that all the stationary probabilities are equal, that is, $\pi_A = \pi_G = \pi_C = \pi_T = 1/4$. This constraint on the stationary probabilities is also imposed by the JC69 model (Jukes & Cantor, 1969) which additionally

assumes that the rates of change between all nucleotides are the same. A more detailed description of these and other Markov models of nucleotide substitution can be found in Yang (2006, Chapter 1).

So far we have only discussed the process on one branch of the tree. Under standard phylogenetic models, a transition matrix governed by the same instantaneous rate matrix Q applies to every branch of the tree. In order to extend the model to allow for all of the sites in our molecular sequence, standard models then assume that the same Markov process applies to every site and that sites are independent of each other. This assumption of independence can be interpreted as follows. Consider a small fixed tree on four taxa and assume that we know the branch lengths and the values of all the substitution model parameters in our model. Suppose we want to use our model to find the probability of having the nucleotides A, A, T, A at site number i (for any i). Independence simply means that this probability would not be affected by what happens at any other site(s). As you may already have guessed, this assumption is made for mathematical convenience, rather than to represent any real biological understanding.

In most molecular sequences, there will be heterogeneity in the extent to which different sites are conserved due to varying selective (or functional) constraints. This can be accommodated in our model by modifying it to allow each site to evolve at its own rate. In order to share information on rates between sites, these rates are assumed to come from a probability distribution, typically a gamma distribution with mean equal to one, which is often denoted as $\Gamma(\alpha, \alpha)$. The shape of this distribution influences the degree of rate heterogeneity between sites and is determined by the single parameter α. For computational convenience, the gamma distribution is often replaced by a discrete approximation with M categories. Although the continuous distribution is only recovered in the limit as M approaches infinity, a small number of categories generally provides a good model for the heterogeneity in the data, for example, $M = 4$ is a popular choice (Yang & Rannala, 2012). Models that allow gamma rate heterogeneity are typically denoted by, for example, HKY85 + Γ or GTR + Γ. Note that these extended models still assume that the processes at different sites are independent of each other.

The standard phylogenetic model described so far makes a number of simplifying assumptions about the evolutionary process. For example, we assume that there is a single composition vector, $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$, which applies to all branches of the tree and at all sites. In other words, we are assuming that sequence composition (the proportion of A, G, C and T bases) tends to remain constant across sites and over evolutionary time. However, these assumptions are violated in many real datasets. For example, the GC content of 16S rRNA varies from 45% to 74% across the diversity of sampled Bacteria, Archaea and Eukarya (Cox et al., 2008). Although simplifying assumptions make statistical models simpler and inference more computationally tractable, they can also impact on inferences about the underlying phylogeny, as we will see in our analysis of the SSU and SSU + TAK datasets. Motivated by such inferential concerns, models have been developed which allow sequence composition to vary across sites (Lartillot & Philippe, 2004) and across

branches of the tree, that is, over time (Foster, 2004). Models with the latter property can have the additional benefit of allowing data to be informative about the root position. In this tutorial, we will use the CAT model (Lartillot & Philippe, 2004) which allows sequence composition to vary across sites. The CAT model assigns each site to one of K groups, where K is less than or equal to the total number of sites in the alignment. Each group of sites has its own composition vector, and there is a probability vector of length K which determines the probability of a site belonging to each of the K groups. The number of groups K is unknown and treated as another parameter of the model. In statistical terms, this is a mixture model with an unknown number of mixture components.

Although we have focused here on the analysis of nucleotide data, protein evolution is also generally modelled using CTMPs and so the same principles apply. However, because there are many more amino acids than nucleotides, the most general reversible, stationary model (the GTR model) contains many more parameters than the equivalent model for nucleotides. As a result, the most commonly used models for amino acid substitutions are empirical models in which fixed values, inferred from large protein datasets, are assumed for the exchangeability parameters or stationary probabilities or both (see, e.g., Le & Gascuel, 2008). However, the very large concatenated protein datasets used in modern phylogenomic analyses are often analysed using the same models discussed here.

5.2 INFERRING PHYLOGENIES UNDER MARKOV SUBSTITUTION MODELS

There are two main schools of thought regarding the appropriate framework for statistical inference of phylogenies using Markov models: frequentist and Bayesian. The two schools differ fundamentally in their interpretations of probability and parameters. Frequentist inference is based on the idea that probability represents a long run relative frequency. For example, the probability of obtaining a head on a coin toss would be interpreted as the limit of the proportion of heads obtained if the coin toss was repeated infinitely many times under identical conditions. In contrast, in Bayesian inference, probability is regarded as "degree of belief" and it is subjective. In the coin toss example, *your* probability of obtaining a head is interpreted as a measurement of how strongly *you* believe a head will occur. Consequently, probability is still a meaningful concept in the case of one-off uncertain events such as whether Species A diverged from Species B. Applying the frequentist definition of probability in the context of non-repeatable events presents conceptual difficulties.

The frequentist interpretation of parameters is that they are fixed but unknown constants and so cannot have probabilities associated with them. The only probabilities considered come from repeated realisations of the model. In contrast, Bayesian inference treats parameters and all other unknowns as random variables to which we assign probabilities. This facilitates straightforward quantification of uncertainty through probability statements. A more detailed comparison between the two schools of thought can be found in our Online Supplement (http://dx.doi.org/10.1016/bs.

mim.2014.05.001). For now, we note that one great practical advantage of the Bayesian approach is that it provides a very natural framework for formulating complex hierarchical (multi-level) models. In a phylogenetic context, this is why the CAT and other highly structured models are generally formulated in the Bayesian framework. Such models often perform very well on large datasets.

5.3 FREQUENTIST INFERENCE

Phylogenetic software for inference based in the frequentist paradigm, such as RAxML (Stamatakis, 2006) and PhyML (Guindon et al., 2010), generally fits models using the method of maximum likelihood. Let $\boldsymbol{y} = (y_{ij})$ denote an alignment of molecular sequences in which y_{ij} is the character state (e.g. nucleotide) at the jth site for species i. Given a particular Markov substitution model, we can write down a likelihood function $p(\boldsymbol{y}|\theta)$ which depends on the data $\boldsymbol{y}$ and unknowns θ. These unknowns include the parameters of the substitution model, the branch lengths and the tree topology. The likelihood function provides a measure of how likely the data $\boldsymbol{y}$ are given the unknowns θ. For the purposes of maximum likelihood estimation, however, it is regarded as a function of the unknowns θ given the data $\boldsymbol{y}$ and is often written as $L(\theta|\boldsymbol{y})$ rather than $p(\boldsymbol{y}|\theta)$. The maximum likelihood estimate $\hat{\theta}$ of θ is the value of θ which maximises the likelihood function. Finding the maximum is an optimisation problem which must be solved numerically if no closed form solution exists. For standard phylogenetic models, the likelihood must be maximised numerically. For a fixed tree, standard numerical optimisation routines are used to find the model parameters and branch lengths which maximise the likelihood. In principle, this procedure should be repeated for all possible trees to find that which leads to the greatest (maximised) likelihood. However, the number of trees on n taxa grows very rapidly with n. For example, there are 15 unrooted trees on 5 taxa, over 2 million unrooted trees on 10 taxa and over 2×10^{20} trees on 20 taxa. This means that an exhaustive search is only feasible for alignments based on a small number of taxa. For larger alignments, the search over tree space is generally heuristic and search algorithms may become stuck in a local maximum. This can occur if the maximised likelihood for a tree is not the global maximum, but it is greater than the maximised likelihoods for the neighbouring trees which are explored during the tree search. As a result, there is no guarantee that the search will find the maximum likelihood tree, that is, the global maximum.

In a frequentist analysis, the quantification of uncertainty about the maximum likelihood tree is usually based on nonparametric bootstrapping (Felsenstein, 1985). Suppose there are N sites in the alignment. Boostrapping involves resampling the sites (alignment columns) with replacement N times in order to generate a bootstrap sample of the same size as the original alignment. The maximum likelihood estimation procedure is then repeated to compute the bootstrap tree as the maximum likelihood tree for the bootstrap alignment. Typically, 100–1000 bootstrap samples are generated and analysed in this way. For every clade in the original maximum likelihood tree, the percentage of bootstrap trees that contain that clade are then

computed. These percentages are called the bootstrap support values, and they are generally marked on phylograms visualising the maximum likelihood tree, or a so-called consensus tree which only includes the most commonly occurring clades in the bootstrap samples. Bootstrap support values are difficult to interpret beyond an intuitive sense that a high bootstrap support value indicates consistent support for a clade across sites. More rigorous interpretations are not easy to defend, although see Yang (2006, Chapter 6).

5.4 BAYESIAN INFERENCE

As remarked above, in the Bayesian approach to inference, probability distributions are assigned to parameters (and, more generally, unknowns). Consider data $\boldsymbol{y}$ which are modelled as having arisen from a statistical model which depends on unknowns θ. Given the chosen model, we can write down a likelihood function $p(\boldsymbol{y}|\theta)$ which provides a measure of how likely the data $\boldsymbol{y}$ are conditional on the unknowns θ. We quantify our initial uncertainty about θ through a probability distribution $\pi(\theta)$ called the prior distribution which indicates how likely we believe each possible value of θ is before seeing the data. The likelihood is then used to update the prior distribution in light of the data $\boldsymbol{y}$. This yields a posterior distribution $\pi(\theta|\boldsymbol{y})$ which summarises all of our uncertainty about θ after seeing the data. The rule for updating the prior distribution is called Bayes' Theorem which can be written as

$$\pi(\theta|\boldsymbol{y}) = \frac{\pi(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})}$$

where

$$p(\boldsymbol{y}) = \int \pi(\theta)p(\boldsymbol{y}|\theta)\mathrm{d}\theta$$

is a normalising constant, called the *marginal likelihood*, which will be discussed further in Section 5.7. In general, this integral cannot be evaluated in closed form. It can also be very difficult to approximate numerically, especially for complex models where θ involves a large number of unknown quantities. In fact, difficulty in computing this integral had been one of the main objections to the Bayesian approach until the early 1990s when computational algorithms called Markov chain Monte Carlo (MCMC) methods effectively solved the computational problem.

MCMC methods work by generating samples $\theta^{[1]}, \theta^{[2]}, \ldots$ from the posterior distribution $\pi(\theta|\boldsymbol{y})$ in such a way that knowledge of the normalising constant $p(\boldsymbol{y})$ is not required. A large number of samples then yields an approximation to the posterior distribution $\pi(\theta|\boldsymbol{y})$. There is, however, an ambiguity in this statement: what constitutes "a large number" of samples? The problem is that the samples drawn from the posterior distribution are not independent of each other. Instead, they form a Markov chain (hence the name Markov chain Monte Carlo) whose stationary distribution is equal to the posterior distribution. It is therefore necessary to run the chain for long enough that it can "forget" the value for θ at which it was initialised and thereby

reach its stationary distribution. Sometimes, movement towards and around the stationary distribution can be very slow. In such cases, we say the chain suffers from slow convergence and poor mixing properties. Unfortunately, these are common problems in phylogenetic analyses because it can be very difficult for the chain to move around tree space. It is therefore good practice to run two or more MCMC chains, initialised at different starting points, and to check that they converge to the same distribution. This will be explained further in the application to the SSU datasets.

Most Bayesian phylogenetic software uses a Metropolis-within-Gibbs sampler in which the unknowns (e.g. model parameters, branch lengths, the tree) in θ are updated one at a time from their conditional posteriors given the current values of all other unknowns. Most, if not all, of these updates cannot be performed as direct simulations from the required distribution. Samples are therefore generated using Metropolis Hastings steps. In a Metropolis Hastings step, a new value for the unknown is generated from a distribution called a proposal distribution, which we choose. This proposed value is then accepted or rejected with a certain probability called the acceptance probability. The acceptance probability is carefully formulated so that the resulting Markov chain has an equilibrium distribution equal to the posterior with density $\pi(\theta|\boldsymbol{y})$. If the proposal is accepted, then the chain moves to the proposed new value. If it is rejected, the chain remains at its current value.

In a Bayesian analysis, the post-data uncertainty about the parameters and other unknowns is represented completely by the posterior distribution. From this, we can compute any posterior quantities of interest, for example, Bayesian credible intervals for parameters or the posterior probability in support of a particular hypothesis. Given an MCMC sample from the posterior, histograms are commonly used to visualise the shapes of marginal posterior distributions for model parameters. The latter information can also be presented more concisely by using the MCMC samples to compute posterior summaries, such as posterior means and Bayesian credible intervals. Similarly, in a phylogenetic context, the posterior distribution for the tree can be computed (approximately) by the proportion of times each tree was sampled during MCMC. The posterior probability of a clade can be approximated likewise by the proportion of sampled trees which contained that clade. As noted above, these posterior probabilities have unambiguous definitions. A posterior probability of 0.93 for a clade is the posterior probability that the "true" tree contains that clade given the chosen model and prior. A word of warning here: the statement "given the chosen model" represents the very strong assumption that the model we have chosen is the actual mechanism by which the data were generated.

5.5 MODEL COMPARISON AND ASSESSMENT

GEP Box famously wrote "All models are wrong but some are useful" (Box, 1979). It is important to realise that we can never hope to find the "true" model by which our data were generated. At best we hope to find a model which is sufficiently flexible to allow us to capture the most important aspects of the evolutionary process.

Implicitly, we hope that models that provide a good fit to the data also provide biologically reasonable inferences about the underlying phylogeny. In the statistical literature, model choice is guided by the principle that we should seek out models which provide a good fit to the data without incorporating redundant or meaningless parameters. In the following sections, we consider model choice in the frequentist and Bayesian inferential frameworks. We then apply some of the principles discussed in this section to a comparison between the GTR model and the CAT model for the SSU and SSU+TAK alignments.

5.6 FREQUENTIST METHODS

In a frequentist framework, a variety of likelihood-based methods are available to help in deciding which model provides the best fit to the data. These include likelihood ratio tests and penalised likelihood criteria.

The likelihood ratio test can be applied in comparisons of nested models. Models M_0 and M_1 are nested if the null model M_0 can be derived from the alternative model M_1 by placing constraints on the values of the parameters of M_1. For example, the JC69 model is nested within the K80 model because setting $\alpha = \beta$ in the K80 model produces the JC69 model. The likelihood ratio test statistic is defined as

$$LR = -2\ln\left\{L\left(\hat{\theta}_0|\boldsymbol{y},M_0\right)/L\left(\hat{\theta}_1|\boldsymbol{y},M_1\right)\right\} = -2\left\{\ln L\left(\hat{\theta}_0|\boldsymbol{y},M_0\right) - \ln L\left(\hat{\theta}_1|\boldsymbol{y},M_1\right)\right\}$$

where $L\left(\hat{\theta}_0|\boldsymbol{y},M_0\right)$ is the likelihood of the null model evaluated at the maximum likelihood estimate $\hat{\theta}_0$ of its parameters and similarly for $L\left(\hat{\theta}_1|\boldsymbol{y},M_1\right)$. Under certain regularity conditions, if the null model is true, the likelihood ratio statistic asymptotically (for large samples) follows a χ^2 distribution on r degrees of freedom where r is the number of constraints imposed on the alternative model M_1 to recover the null model M_0. In the JC69/K80 example, $r = 1$. In phylogenetics, an immediate problem with the likelihood ratio test is that the nesting condition demands that the same tree is used for both models, even though the tree with the highest maximised likelihood may differ between models. In practice, however, Posada and Crandall (2001) found that the model chosen by likelihood ratio tests was generally robust against different choices of the common tree.

An alternative approach that does not require models to be nested is to use penalised likelihood criteria which attempt to provide an indication of model fit (as measured by the maximised log-likelihood), adjusted by a term which penalises model complexity. Examples of such model selection tools are the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) defined as

$$\text{AIC} = -2\ln L\left(\hat{\theta}|\boldsymbol{y},M\right) + 2n \quad \text{and} \quad \text{BIC} = -2\ln L\left(\hat{\theta}|\boldsymbol{y},M\right) + n\ln N$$

Here, $L\left(\hat{\theta}|\boldsymbol{y},M\right)$ is the likelihood of the model evaluated at the maximum likelihood estimate $\hat{\theta}$ of its parameters, n is the number of free parameters in the model

and N is the size of the observed sample, usually taken to be the number of sites in the alignment. For a comparison between the various methods of model selection in phylogenetics, see Posada and Buckley (2004).

For standard models of nucleotide substitution, likelihood ratio tests and comparisons via AIC and BIC can be carried out using jModelTest 2 (Darriba, Taboada, Doallo, & Posada, 2012; see Table 1).

5.7 BAYESIAN MODEL CHOICE

The Bayesian framework offers a principled approach to dealing with model uncertainty. A central role is played by the marginal likelihood $p(\mathbf{y})$, defined earlier as the normalising constant in Bayes' Theorem, which is involved in the computation of Bayes factors and posterior model probabilities. Note that it depends on the prior distribution as well as the model. Bayes factors are often used to compare pairs of models and have an interpretation in terms of the posterior odds of one model in comparison with the other. Posterior model probabilities can be interpreted as the posterior probability that each model is correct if it is assumed that the collection of models being compared contains the "true" model, that is, the true data generating mechanism. For further details on formal Bayesian model choice and the related concept of Bayesian model averaging, see our Online Supplement (http://dx.doi.org/10.1016/bs.mim.2014.05.001).

Despite its attractive framework, formal Bayesian model choice is often hampered by the difficulty of computing the marginal likelihood, which is generally a very difficult numerical integration problem. Accordingly, most Bayesian phylogenetic programs do not provide functions to perform this calculation. An exception is MrBayes (Ronquist et al., 2012), which allows users to approximate marginal likelihoods using either the harmonic mean or the (greatly superior) stepping-stone method (Xie, Lewis, Fan, Kuo, & Chen, 2011). For practical guidance on how to perform the calculation, see Section 4.4 of the MrBayes Manual (http://mrbayes.sourceforge.net/).

In situations where it is not possible to compute the marginal likelihood for all models under consideration, there are a variety of informal methods that allow comparisons to be made. These include cross-validation and posterior predictive model checking. Both can be carried out using the Bayesian program PhyloBayes (Lartillot, Lepage, & Blanquart, 2009) through the cvrep/pb/readcv/sumcv commands and the ppred command, respectively. In PhyloBayes, cross-validation involves dividing the data into a training set and a validation set and then approximating the probability of the validation data given the training data under each possible model. Ideally, this procedure would be repeated for every possible partition of the data into training and validation sets, averaging the probability across partitions. However, this is computationally infeasible and so in practice only a small random sample from the set of possible partitions is used. The model with the highest cross-validation score is judged to be best-fitting. Unfortunately, cross-validation can often be computationally prohibitive because the posterior needs to be recomputed for every partition of

the data. Consequently, we will not use it in this chapter. Posterior predictive model checking provides a less computationally demanding approach to model choice, and a practical example can be found in the next section. It is based on the posterior predictive distribution of a hypothetical replicate $\boldsymbol{y}_{\text{rep}}$ of the data $\boldsymbol{y}$ which could have been observed under the model M

$$p(\boldsymbol{y}_{\text{rep}}|\boldsymbol{y},M) = \int p(\boldsymbol{y}_{\text{rep}}|\theta,M)p(\theta|\boldsymbol{y},M)\mathrm{d}\theta$$

These model checks can be based on test quantities $T(\boldsymbol{y})$ which are simple summaries designed to capture aspects of the data that we want the model to capture adequately. For example, in PhyloBayes, one of the predefined test quantities is the biochemical specificity, measured by the mean number of distinct character states in a column. In general, it will not be possible to compute the posterior predictive distribution of a test quantity $T(\boldsymbol{y})$ analytically; however, given MCMC draws $\theta^{[1]},\theta^{[2]},\ldots,\theta^{[B]}$ from the posterior distribution of a model M, it is easy to build up a numerical approximation by simulating a replicated dataset $\boldsymbol{y}_{\text{rep}}^{[\text{i}]}$ for each draw from the posterior and computing $T(\boldsymbol{y}_{\text{rep}}^{[i]})$. The idea is that if the model fits well then the observed data should look plausible under the posterior predictive distribution. A simple way of checking this graphically is to plot the posterior predictive distribution for the test quantity (for example in a histogram) and then to examine the position of the observed test quantity $T(\boldsymbol{y})$ in that distribution. Poor model fit would be indicated by the observed test quantity lying far into the tails of the posterior predictive distribution. When comparing two or more models, the model for which the observed test quantity lies most centrally provides the best fit to the data in terms of capturing the feature of the data summarised by $T(\boldsymbol{y})$. In PhyloBayes, the position of the observed test quantity relative to the posterior predictive distribution is summarised using a P-value; very small or very large P-values indicate that the observed value lies in one of the tails of the predictive distribution, suggesting poor model fit. We note that even when only one model is being considered, it is good practice to assess its fit to the data using posterior predictive checks like those discussed here.

6 STEP 5: INFERRING TREES—PRACTICAL GUIDELINES FOR FITTING AND COMPARING MARKOV SUBSTITUTION MODELS

6.1 ALIGNMENT FORMATS FOR PHYLOGENY PROGRAMS

While alignments are often generated and viewed in formats such as FASTA (where each sequence is prefaced with a one-line header, starting with a ">") or ClustalW (in which the alignment is formatted for easy visual inspection or printing), most phylogeny packages, including the two we will introduce here, require that the input alignment be in PHYLIP format, perhaps for historical reasons. Converting data between formats is one of the most irritating and mundane aspects of phylogenetics

(and, more generally, bioinformatics), but a number of conversion tools are available to streamline the process. One option is readAl, which is part of the trimAl package that we used in Step 3 to edit the alignment. It can be run locally or on the Phylemon Web server (Table 1). Use readAl to convert your alignments to the standard PHYLIP format using the "-phylip" option and then open the resulting files in a text editor. You will see that all these different alignment formats—FASTA, Clustal, PHYLIP—contain exactly the same information, but presented differently. With PHYLIP-formatted versions of your SSU and SSU+TAK alignments, you are ready to run the phylogeny packages. We will analyse these datasets using two of the models introduced in section 5.1 GTR+Γ and CAT+Γ (i.e. the GTR or CAT models with a gamma distribution for modelling across-site rate variation), and compare the results. In general, when comparing two or more models, they should be fitted in the same inferential framework. Inference for the CAT model is only implemented in the Bayesian paradigm and so we will fit and compare the models in the Bayesian framework. The program we will use does not perform marginal likelihood calculations and so we will compare models using posterior predictive checks. Although unnecessary in this analysis, for the purpose of illustration, we will also fit the GTR model using maximum likelihood.

6.2 INFERRING MAXIMUM LIKELIHOOD PHYLOGENIES USING RAxML

In this tutorial, we will use RAxML (Stamatakis, 2006), one of the most popular maximum likelihood packages. One of the great strengths of RAxML is that it is optimised for modern multi-core processors; if you compile and run it locally, make sure to choose a version appropriate for your system so that you can take advantage of all of the cores in your desktop or laptop processor to run the analyses more quickly. RAxML can be run either locally or online through a Web server such as CIPRES (see Table 1). Whichever option you choose, the most important parameters to set are the model specification and the number of bootstrap replicates. In this case, these will be the GTR+Γ model, and 100 rapid bootstraps. As discussed in section 5.3, maximum likelihood methods rely on heuristic tree search algorithms to look for the maximum likelihood tree and so it is possible that the optimisation algorithm will get stuck in a local maximum. Therefore, it is good practice to rerun RAxML analyses several times using different starting values for the random seed (*-x* and *-p* options).

6.3 BAYESIAN ANALYSES WITH PHYLOBAYES

PhyloBayes (Lartillot et al., 2009) is a Bayesian phylogenetic package. Its speciality is the implementation of CAT and other complex mixture models that often fit alignments of highly divergent sequences (such as the rRNAs from across the tree of life) much better than GTR and other single *Q*-matrix models. PhyloBayes can be installed locally or run on the CIPRES Web server (Table 1); more so than for

the other packages we discuss here, you will have much more flexibility and control over your PhyloBayes analyses if you run them locally. Two versions of PhyloBayes are currently available—a serial (single-core) and parallelised (multi-core) version, the latter called PhyloBayes-MPI (Lartillot, Rodrigue, Stubbs, & Richer, 2013). We will use the single-core version here, but bear in mind that the MPI version can be very useful for speeding up analyses of larger datasets or inference under more complex models. As discussed in section 5.4, the strategy for a Bayesian analysis is to run multiple MCMC chains and then check for convergence. To start a chain running with PhyloBayes, you invoke the "pb" command with the appropriate options for model and prior specification:

```
pb -s -d myAlignment.phy -cat myChainName &
```

Here, -cat indicates that the model to be fitted is the CAT model. Replace this with -gtr for the GTR model. By default, PhyloBayes uses the discrete gamma model with four categories to describe across-site rate heterogeneity. We accept this default here, but it can be changed by the inclusion of appropriate options. The posterior distribution is formed by combining information from the data (the likelihood) with information from the prior and so the choice of prior distribution can and will influence our posterior inferences about the phylogenetic tree. In theory, the prior should be chosen to reflect prior beliefs about the model parameters, branch lengths and tree. However, eliciting prior information is a challenging topic in its own right, and provision of guidelines is well beyond the scope of this tutorial. Here, we simply accept the default prior, but encourage the interested reader to explore the effect on the posterior distribution of changing the prior. Again, this can be achieved by the inclusion of appropriate options with the "pb" command.

Assessing whether MCMC chains have converged is a difficult problem and you are likely to need to generate *at least* tens of thousands of iterations. As discussed previously, we suggest running two chains for each model and comparing the MCMC output. PhyloBayes provides two programs that implement numerical convergence diagnostics for pairs of chains—bpcomp and tracecomp. See the PhyloBayes manual (http://www.phylobayes.org) for some guidelines on how to interpret the results of these diagnostics. The .trace files produced by the "bp" command record a few parameter summaries (e.g. the log-likelihood) for each sample from the posterior distribution, and the numerical diagnostics invoked by the tracecomp program work on this MCMC output. As an alternative, we encourage the reader to examine the convergence of these statistics graphically using the guidelines available in our Online Supplement (http://dx.doi.org/10.1016/bs.mim.2014.05.001), which also provides further comments on assessing mixing in tree space.

The first time you examine the numerical and/or graphical diagnostics, you may judge that the MCMC chains have not converged. If this is the case, you should continue to run the chains until there is no evidence of any lack of convergence. Note that convergence and model checking are completely different things—once your MCMC chains appear to have converged, you can proceed to comparing the fit of the GTR and CAT models.

6.4 POSTERIOR PREDICTIVE CHECKS

The "ppred" program can be used to do posterior predictive checks once your analyses have converged. We will focus on the site-specific biochemical diversity test, which can be performed by invoking ppred with the -sat flag (other tests, such as for compositional homogeneity, are often useful but will not be considered here; see Foster, 2004). Site-specific biochemical diversity refers to the number of different nucleotides (or amino acids) per alignment column—see Figure 3 for more details. Run this test for one of your GTR and one of your CAT chains. Our results for the SSU dataset (no TAK) are plotted in Figure 3: you can see that the range of site-specific diversity predicted under the CAT model overlaps the observed value, whereas GTR predicted levels of diversity that were much too high. This result is one indication that the GTR model does a poor job of capturing the site-specific evolutionary dynamics observed in this alignment. Are the results similar for the SSU +TAK dataset? Before moving on to the next step, we note that PhyloBayes also implements a very popular and flexible model that combines the across-site mixture model of CAT with a GTR matrix for exchangeabilities between nucleotides (or amino acids): CAT+GTR. This model is somewhat more realistic than the CAT model on its own because it also allows the rates of change between different

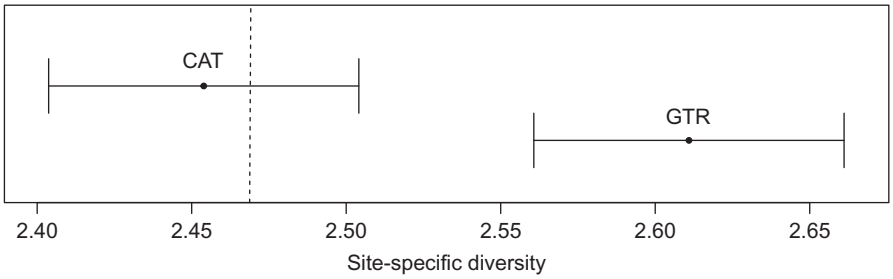


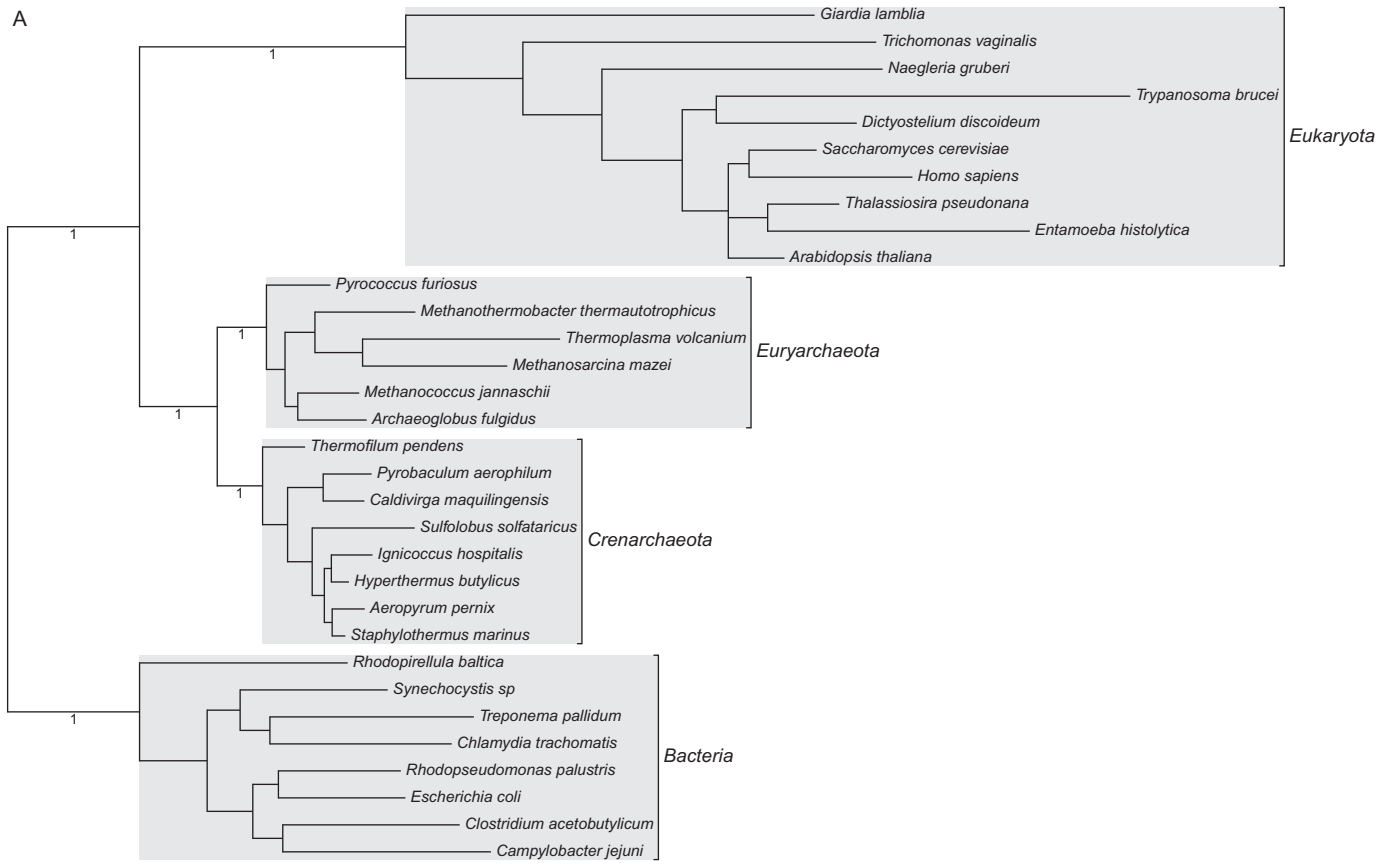
FIGURE 3

Posterior predictive checks for site-specific biochemical diversity on the SSU alignment under the GTR and CAT models. Site-specific diversity is simply the mean number of different nucleotides observed per alignment column. In real datasets, this value tends to be rather low: when you look at an alignment, you will notice that most columns largely consist of only one or two of the four possibilities (the same pattern is usually observed with amino acid data). The observed (i.e. "real") value is plotted as a dashed line; posterior predictive means and an interval two standard deviations either side are also plotted for the GTR and CAT models. (This interval represents a symmetric 95% Bayesian credible interval if it can be assumed that the posterior predictive distribution is approximately normal.) The values predicted under GTR are too high, suggesting that GTR does not adequately model the site-specific nature of the evolutionary process—this may well be because GTR averages nucleotide frequencies and exchangeabilities over the whole alignment. The observed value is, however, consistent with the posterior predictive distribution for the CAT model, indicating that CAT is a more realistic model for this alignment, at least for this aspect of the evolutionary process.

nucleotides to differ. You may like to try fitting the CAT + GTR model to one or both of our alignments and comparing your results with those we discuss below—if you can, use PhyloBayes-MPI for significant speed improvements.

7 STEP 6: INTERPRETING THE PHYLOGENETIC TREE

The phylogeny packages discussed here write out the trees as plain text files in Newick format, which is rather difficult to parse by eye. Instead, we recommend the use of tree viewing software such as Dendroscope or FigTree (Table 1) to visualise the phylogenies. Open and compare the trees inferred for the two datasets under the GTR and CAT models (four trees in total). You will notice some differences between them, both in terms of topology (branching order) and the support values (maximum likelihood bootstraps or Bayesian posterior probabilities) for different groups. Remember that the question we wanted to address with these analyses was the position of the Eukarya relative to the Archaea: do they form monophyletic sister groups, as under the three-domains tree, or do the Eukarya emerge from within the Archaea, as proposed by some alternative hypotheses, such as the eocyte hypothesis of Lake et al. (1984)? To answer this question, we will need to root the trees, in order to establish the direction of ancestor–descendant relationships. Like almost all phylogenetic models in current use, GTR and CAT are stationary and time reversible, meaning that the position of the root has no effect on the probability of the tree, and so cannot be inferred as part of the analysis (see section 5.1 for further details). This means that the position of the root must be established based on independent data—a far from ideal but very commonly encountered situation in phylogenetics. The usual strategy is to root the tree on a known outgroup—a sequence (or group) whose divergence from all others is known to represent the earliest split in the tree; the root can then be placed on the branch that joins the outgroup to the rest of the tree. For example, a phylogenetic analysis of birds might include mammal sequences as an outgroup because the split between birds and mammals is known to predate the radiation of birds. As mentioned in Section 1, analyses of this type place the root of the tree of life on the branch leading to the Bacteria. We will assume that bacterial root here, in order to polarise our trees and investigate the relationship between Eukarya and Archaea. To do this using your tree viewing software, you will first need to confirm that all the Bacteria cluster together (form a clan; Wilkinson, McInerney, Hirt, Foster, & Embley, 2007) in the tree. If this is the case, select the branch joining this cluster to the rest of the tree and reroot; this can be done from the Edit menu in Dendroscope. The tree needs to be "rerooted" rather than simply rooted because phylogeny packages often output trees with an arbitrary root; this can be safely ignored if the analysis was performed using a stationary, time-reversible model, as was the case here. You can now investigate the position of the Eukarya relative to the Archaea in your rooted trees. You may find Figures 4 and 5, which summarise the results that we obtained from our Bayesian analyses, a helpful point of comparison. The analysis of the SSU dataset with the GTR model gave a strongly supported three-domains tree, with maximal posterior support (posterior probability (PP) = 1) for the


A
Giardia lamblia
Trichomonas vaginalis
Naegleria gruberi
Trypanosoma brucei
Dictyostelium discoideum
Saccharomyces cerevisiae
Homo sapiens
Thalassiosira pseudonana
Entamoeba histolytica
Arabidopsis thaliana
Eukaryota
Pyrococcus furiosus
Methanothermobacter thermautotrophicus
Thermoplasma volcanium
Methanosarcina mazei
Methanococcus jannaschii
Archaeoglobus fulgidus
Euryarchaeota
Thermofilum pendens
Pyrobaculum aerophilum
Caldivirga maquilingensis
Sulfolobus solfataricus
Ignicoccus hospitalis
Hyperthermus butylicus
Aeropyrum pernix
Staphylothermus marinus
Crenarchaeota
Rhodopirellula baltica
Synechocystis sp
Treponema pallidum
Chlamydia trachomatis
Rhodopseudomonas palustris
Escherichia coli
Clostridium acetobutylicum
Campylobacter jejuni
Bacteria
1
1
1
1
1

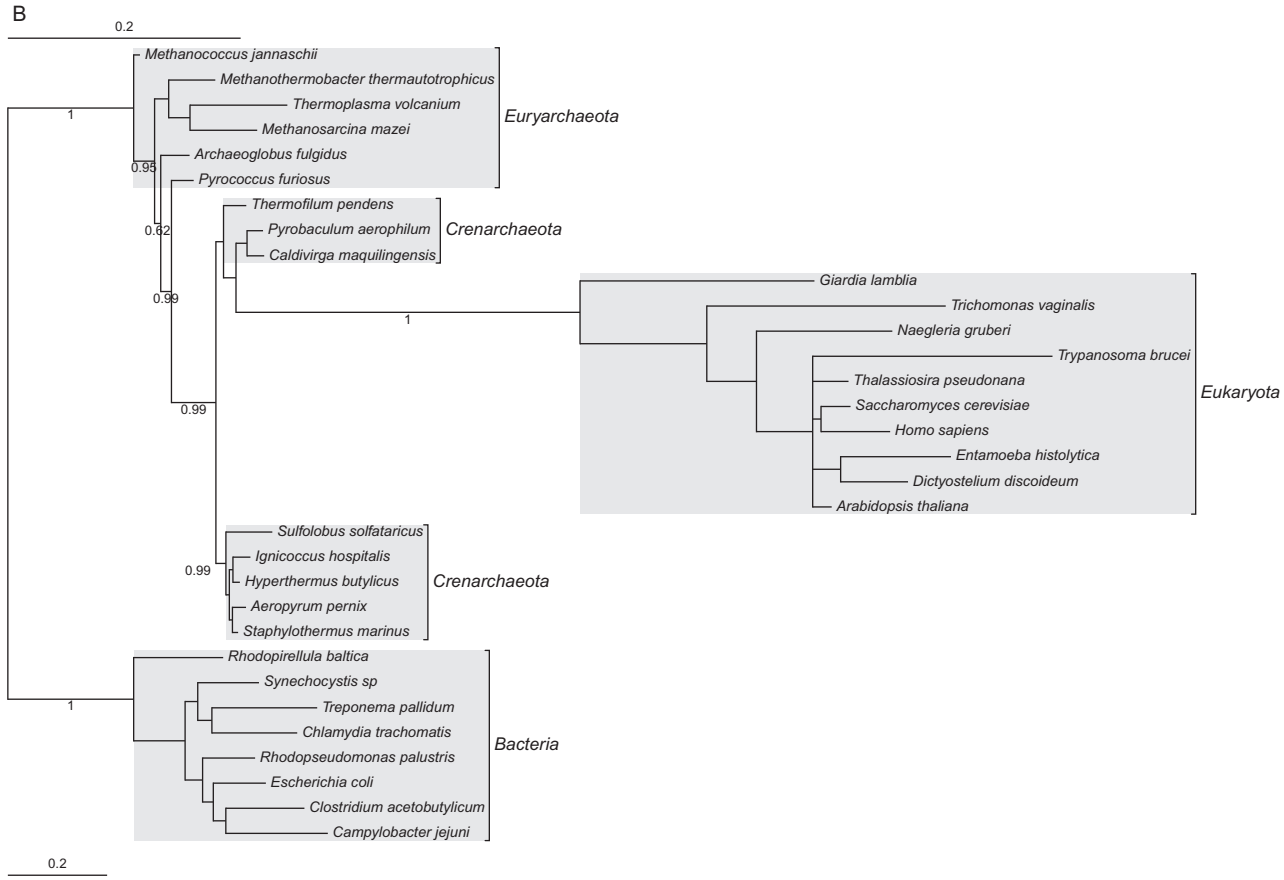
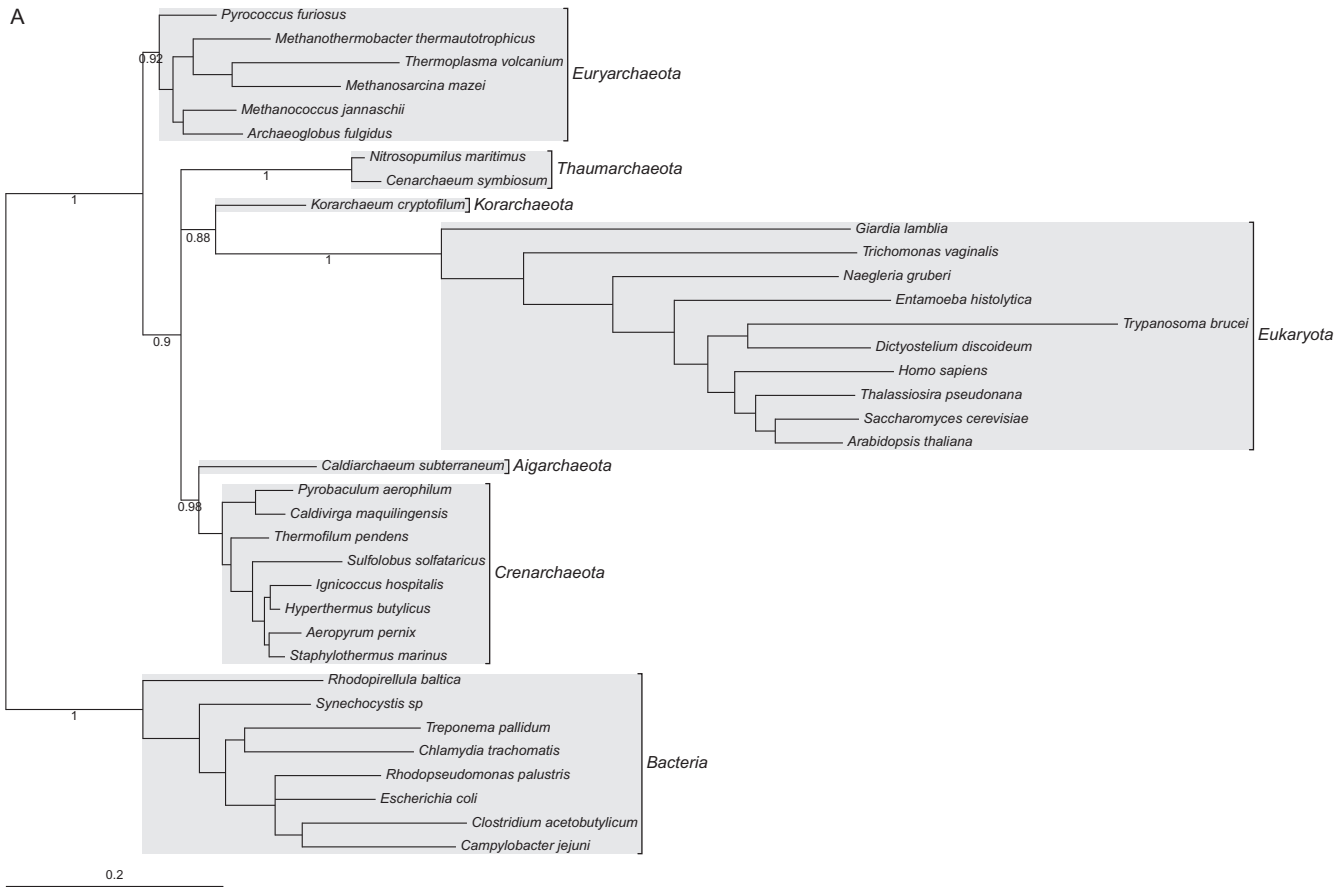


FIGURE 4

Phylogenies inferred from the SSU dataset under (A) the GTR and (B) the CAT models. Branch lengths measure the expected number of substitutions per site. Support values on branches are Bayesian posterior probabilities. In the GTR analysis of this dataset (A), the eukaryotes emerge as the sister group to a monophyletic Archaea with maximal posterior support (PP = 1)—the classic "three-domains" (Woese et al., 1990) tree. When the same alignment was analysed using the CAT model (B), the eukaryotes are placed within the Crenarchaeota, also with very high support (PP = 0.99); this topology corresponds to the eocyte hypothesis (Lake et al., 1984). Both trees were rooted on the branch leading to Bacteria, consistent with analyses of ancient paralogous genes (Gogarten et al., 1989; Iwabe et al., 1989).


A
Pyrococcus furiosus
Methanothermobacter thermautotrophicus
Thermoplasma volcanium
Methanosarcina mazei
Methanococcus jannaschii
Archaeoglobus fulgidus
Euryarchaeota
Nitrosopumilus maritimus
Cenarchaeum symbiosum
Thaumarchaeota
Korarchaeum cryptofilum
Korarchaeota
Giardia lamblia
Trichomonas vaginalis
Naegleria gruberi
Entamoeba histolytica
Trypanosoma brucei
Dictyostelium discoideum
Homo sapiens
Thalassiosira pseudonana
Saccharomyces cerevisiae
Arabidopsis thaliana
Eukaryota
Caldiarchaeum subterraneum
Aigarchaeota
Pyrobaculum aerophilum
Caldivirga maquilingensis
Thermofilum pendens
Sulfolobus solfataricus
Ignicoccus hospitalis
Hyperthermus butylicus
Aeropyrum pernix
Staphylothermus marinus
Crenarchaeota
Rhodopirellula baltica
Synechocystis sp
Treponema pallidum
Chlamydia trachomatis
Rhodopseudomonas palustris
Escherichia coli
Clostridium acetobutylicum
Campylobacter jejuni
Bacteria
0.92
1
0.88
1
0.9
0.98
1
0.2

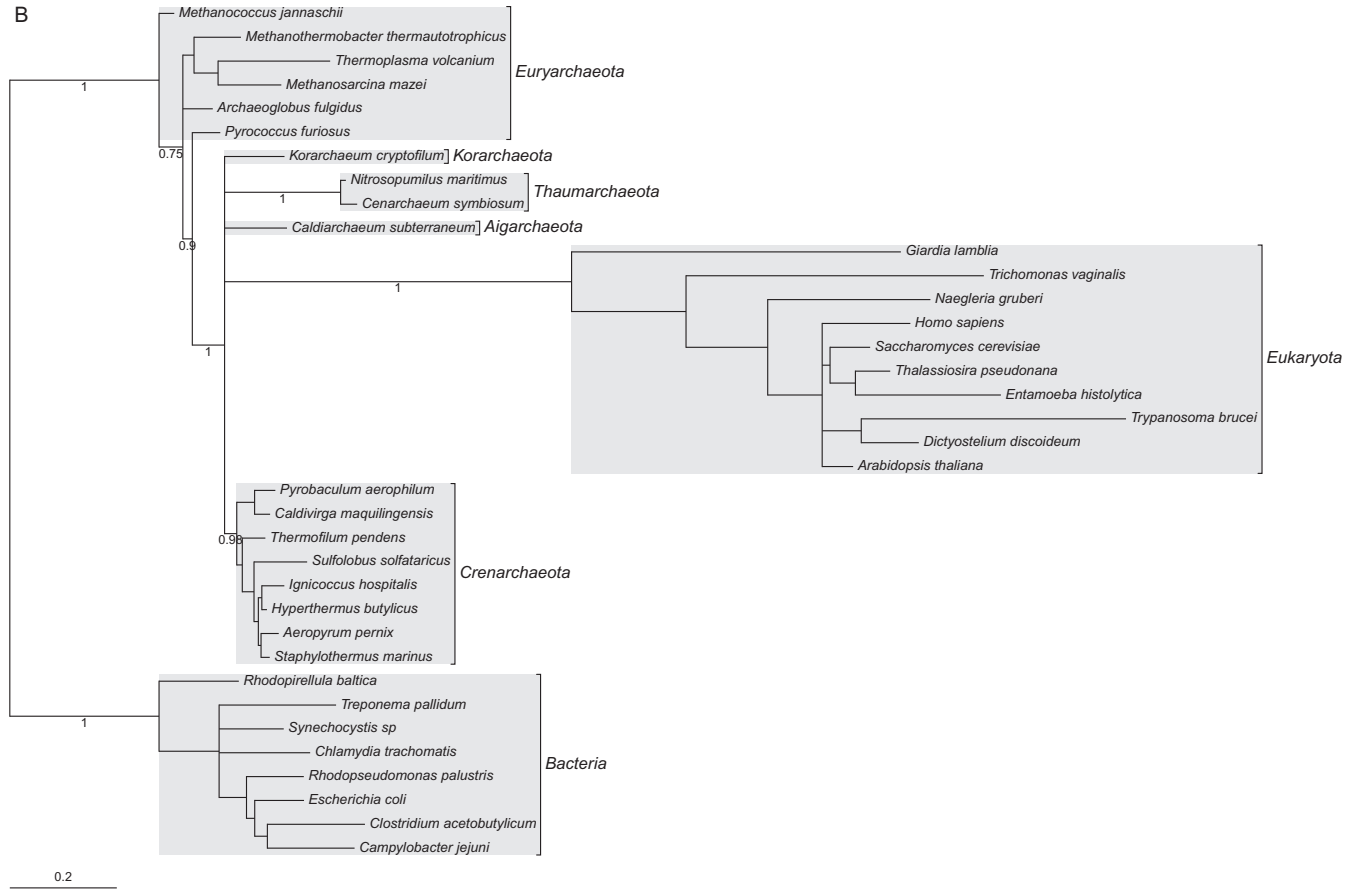


FIGURE 5

Phylogenies inferred from the SSU+TAK dataset under (A) the GTR and (B) the CAT models. Support values on branches are Bayesian posterior probabilities. With the addition of the new sequences from the TAK Archaea (Thaumarchaeota, Aigarchaeota and Korarchaeota), both the GTR (A) and CAT (B) models recover a topology in which the eukaryotes emerge from within the Archaea; the support for a clade of eukaryotes and TACK Archaea is highest (PP = 1) under the CAT model. As in Figure 4, both trees were rooted on the branch leading to Bacteria, consistent with analyses of ancient paralogous genes (Gogarten et al., 1989; Iwabe et al., 1989).

monophyly of the Archaea. *Monophyly* indicates that all Archaea share a common ancestor to the exclusion of eukaryotes and vice versa. If you are unsure which sequences belong to Archaea and which to Eukarya, try a Web search for their names. In our analyses, all other combinations of models and datasets gave an eocyte tree, in which the Eukarya emerge from within the Archaea; that is, some Archaea are more closely related to eukaryotes than they are to other Archaea. The support for this relationship was reasonable in the GTR analysis of the SSU+TAK dataset (PP = 0.9 for a clade of eukaryotes plus TACK), rising to PP = 0.99 and PP = 1 in the analyses of both datasets using the CAT model (PP of 1). Given that our posterior predictive tests suggested that the CAT model fit the data better than the GTR model, and that adding the additional archaeal sequences leads to the recovery of the same topology under both models, these analyses support the eocyte tree over the three-domains tree and suggest that the small subunit rRNA genes of eukaryotes, at least, may descend from within the Archaea—in particular, the "TACK" group containing the Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota.

Having performed these analyses yourself, we hope that you are now more aware of the caveats and limitations that necessarily accompany results of this type. In this tutorial, we saw that both taxon sampling (the choice of species to be included in the analysis) and the phylogenetic model used can influence the final result. More generally, decisions made at any of the steps of the analysis have the potential to influence the inferred tree, not least of which is the choice of gene to analyse in the first place. Here, we focused on SSU rRNA as one of the most frequently used phylogenetic markers, but any comprehensive attempt to address the three domains/eocyte debate would involve analyses of a much broader range of genes—at the very least, the set of 30–40 other genes, mostly encoding protein components of the ribosome, that are conserved on the genomes of most cellular life forms. As mentioned above, simultaneous analysis of multiple genes raises a number of complications that we did not deal with here, including how to handle the all-too-frequent cases in which the genes being analysed disagree as to the underlying species tree. If you are interested in exploring the three domains/eocyte question further, a good starting point may be some of the recent literature on this topic, which attempts to engage with these issues (Guy & Ettema, 2011; Lasek-Nesselquist & Gogarten, 2013; Williams et al., 2013, 2012), as well as chapter 'Reconciliation Approaches to Determining HGT, Duplications, and Losses in Gene Trees' by Kamneva and Ward in this volume.

CONCLUSIONS

In this chapter, we have attempted to introduce some of the most important aspects of phylogenetic analysis, focusing in particular on sequence alignment, taxon sampling and the selection of an appropriate phylogenetic model. Here, we have concentrated on a hands-on approach; for an excellent, up-to-date theoretical treatment, see Yang and Rannala (2012). All of these decisions have the potential to greatly influence the results of your analysis; as a result, phylogenies must be interpreted tentatively and

with their caveats and limitations in mind. Ultimately, phylogenies are hypotheses of evolutionary relationships that help us to organise biological diversity and to understand the evolution of traits. This is not to say that the process of phylogenetics is arbitrary: there are benchmarks and best practices available for most of the steps, and following these as carefully as possible will increase the robustness of the resulting phylogeny. The ideal situation is when a variety of methods agree on the result; when this is not the case, tests of model fit such as those discussed in step 4 may indicate which phylogeny should be tentatively preferred. We hope that these analyses have provided some sense of the fascination of phylogenetics. The idea that it is possible to make inferences, however cautious, about events that took place billions of years ago through the statistical analysis of sequences from contemporary organisms can be intoxicating, and the advent of next-generation sequencing, combined with continuing advances in statistical methodology, makes this an exciting time to enter the field!

ACKNOWLEDGEMENTS

We thank Cymon J. Cox, Malcolm Farrow, Peter G. Foster and T. Martin Embley for helpful feedback on an earlier draft of this chapter. This work was supported by a Marie Curie postdoctoral fellowship to T. A. W. S. E. H. is supported under a European Research Council Advanced Investigator grant awarded to T. Martin Embley.

REFERENCES

Akaike, H. (1974). New look at statistical-model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M., et al. (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, *396*(6707), 133–140.

Bininda-Emonds, O. R. (2004). The evolution of supertrees. *Trends in Ecology & Evolution*, *19*(6), 315–322.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer, & G. N. Wilkinson (Eds.), *Robustness in statistics*. Waltham, Massachusetts: Academic Press.

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, *311*(5765), 1283–1287.

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 of the United States of America*, *105*(51), 20356–20361.

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, *10*, 210.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: More models, new heuristics and parallel computing. *Nature Methods*, *9*(8), 772.

de Queiroz, A., & Gatesy, J. (2007). The supermatrix approach to systematics. *Trends in Ecology & Evolution*, *22*(1), 34–41.

Drew, B. T., Gazis, R., Cabezas, P., Swithers, K. S., Deng, J., Rodriguez, R., et al. (2013). Lost branches on the tree of life. *PLoS Biology*, *11*(9), e1001636.

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113.

Embley, T. M., & Martin, W. (2006). Eukaryotic evolution, changes and challenges. *Nature*, *440*(7084), 623–630.

Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., et al. (2004). A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*, *21*(9), 1643–1660.

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

Felsenstein, J. (1985). Confidence-limits on phylogenies—An approach using the bootstrap. *Evolution*, *39*(4), 783–791.

Foster, P. G. (2004). Modeling compositional heterogeneity. *Systematic Biology*, *53*(3), 485–495.

Foster, P. G., Cox, C. J., & Embley, T. M. (2009). The primary divisions of life: A phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society of London Series B, Biological sciences*, *364*(1527), 2197–2207.

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., et al. (1989). Evolution of the vacuolar H+-ATPase: Implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(17), 6661–6665.

Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P., & Brochier-Armanet, C. (2010). The origin of eukaryotes and their relationship with the Archaea: Are we at a phylogenomic impasse? *Nature Reviews. Microbiology*, *8*(10), 743–752.

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*(3), 307–321.

Guy, L., & Ettema, T. J. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends in Microbiology*, *19*(12), 580–587.

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*(2), 160–174.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., & Miyata, T. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(23), 9355–9359.

Jukes, T. H., & Cantor, C. R. (1969). *Evolution of protein molecules.* New York: Academic Press.

Kelly, S., Wickstead, B., & Gull, K. (2011). Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proceedings. Biological Sciences/The Royal Society*, *278*(1708), 1009–1018.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120.

Lake, J. A., Henderson, E., Oakes, M., & Clark, M. W. (1984). Eocytes: A new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(12), 3786–3790.

Lartillot, N., Lepage, T., & Blanquart, S. (2009). PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, *25*(17), 2286–2288.

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, *21*(6), 1095–1109.

Lartillot, N., Rodrigue, N., Stubbs, D., & Richer, J. (2013). PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, *62*(4), 611–615.

Lasek-Nesselquist, E., & Gogarten, J. P. (2013). The effects of model choice and mitigating bias on the ribosomal tree of life. *Molecular Phylogenetics and Evolution*, *69*(1), 17–38.

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320.

Lee, M. S. Y. (2001). Unalignable sequences and molecular evolution. *Trends in Ecology & Evolution*, *16*(12), 681–685.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., et al. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(19), 12246–12251.

Moreira, D., & Lopez-Garcia, P. (2005). Comment on "The 1.2-megabase genome sequence of Mimivirus" *Science*, *308*(5725), 1114, author reply 1114.

Nylander, J. A., Wilgenbusch, J. C., Warren, D. L., & Swofford, D. L. (2008). AWTY (are we there yet?): A system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, *24*(4), 581–583.

Pisani, D., Cotton, J. A., & McInerney, J. O. (2007). Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular Biology and Evolution*, *24*(8), 1752–1760.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, *53*(5), 793–808.

Posada, D., & Crandall, K. A. (2001). Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, *50*(4), 580–601.

Rannala, B., & Yang, Z. (2008). Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics*, *9*, 217–231.

Redelings, B. D., & Suchard, M. A. (2005). Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology*, *54*(3), 401–418.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., et al. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.

Sagan, L. (1967). On the origin of mitosing cells. *Journal of Theoretical Biology*, *14*(3), 255–274.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. http://dx.doi.org/10.1214/aos/1176344136.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688–2690.

Stanier, R. Y., & van Niel, C. B. (1962). The concept of a bacterium. *Archives of Microbiology*, *42*, 17–35.

Stevens, P. F. (1991). Character states, morphological variation, and phylogenetic analysis—A review. *Systematic Botany*, *16*(3), 553–583.

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577.

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

Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives. *PLoS One*, *6*(3), e18093.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., & Notredame, C. (2006). M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, *34*(6), 1692–1699.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191.

Wilkinson, M., McInerney, J. O., Hirt, R. P., Foster, P. G., & Embley, T. M. (2007). Of clades and clans: Terms for phylogenetic relationships in unrooted trees. *Trends in Ecology & Evolution*, *22*(3), 114–115.

Williams, T. A., Foster, P. G., Cox, C. J., & Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, *504*(7479), 231–236.

Williams, T. A., Foster, P. G., Nye, T. M., Cox, C. J., & Embley, T. M. (2012). A congruent phylogenomic signal places eukaryotes within the Archaea. *Proceedings. Biological Sciences/The Royal Society*, *279*(1749), 4870–4879.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, *51*(2), 221–271.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090.

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 of the United States of America*, *87*(12), 4576–4579.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, M. H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, *60*(2), 150–160.

Yang, Z. (2006). *Computational molecular evolution*. Oxford: Oxford University Press.

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews. Genetics*, *13*(5), 303–314.

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, *8*(2), 357–366.