PhyloBayes: Bayesian Phylogenetics Using Site-heterogeneous Models

Nicolas Lartillot

Chapter 1.5 PhyloBayes: Bayesian Phylogenetics Using Site-heterogeneous Models

Nicolas Lartillot
Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive,
43 Bld du 11 Novembre 1918, 69622 Villeurbanne cedex, France.
nicolas.lartillot@univ-lyon1.fr
https://orcid.org/0000-0002-9973-7760

Abstract

PhyloBayes is a software program for Bayesian phylogenetic reconstruction. Compared to other programs, its main distinguishing feature is the implementation of the CAT model, which accounts for fine-grained variation across sites in amino acid preferences using a Bayesian non-parametric approach. This chapter provides a detailed step-by-step practical introduction to phylogenetic analyses using PhyloBayes, using as an example a previously published dataset addressing the phylogenetic position of Microsporidia within eukaryotes. Through this historically emblematic case of a long-branch attraction artifact, a complete analysis under site-homogeneous and site-heterogeneous models is conducted and interpreted, thus providing an illustration of why modeling pattern variation is so fundamental for reconstructing deep phylogenies.

How to cite: Nicolas Lartillot (2020). PhyloBayes: Bayesian Phylogenetics Using Site-heterogeneous Models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 1.5, pp. 1.5:1–1.5:16. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

Since the realization, by Zuckerkandl and Pauling (1965), that DNA molecules represent documents of the long-term evolutionary history of species, molecular phylogenetics has gone a long way in progressively deciphering the detailed patterns of diversification across species at all evolutionary scales. Accurately reconstructing the tree of life, however, has turned out to be quite more challenging than anticipated, especially over deep evolutionary times. The long-standing hesitations, over the last 50 years, concerning the position of Microsporidia in the tree of eukaryotes (Brinkmann et al., 2005), or that of nematodes (Aguinaldo et al., 1997; Philippe et al., 2005) and, more recently, ctenophores (Telford et al., 2016), in the metazoan phylogeny, clearly illustrate the difficulty in firmly establishing a definitive picture of the diversification patterns having occurred in the remote evolutionary past.

The first phylogenetic methods, based on distance or on maximum parsimony, have quickly shown important methodological weaknesses and have progressively been replaced by more principled model-based approaches, using either maximum likelihood or Bayesian inference. Even with these probabilistic methods, however, are systematic errors in tree reconstruction still a major plague (Philippe et al. 2011; Chapter 2.1 [Simion et al. 2020]). One main reason is the difficult question of model adequacy: probabilistic approaches are accurate only inasmuch as the underlying model of sequence evolution correctly describes the true evolutionary process. In practice, models are obviously a much idealized description of the true processes, which thus raises the question of which aspects of the evolutionary process are critical and should be correctly captured, in order to mitigate the impact of reconstruction errors.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 1.5; pp. 1.5:1–1.5:16

A book completely handled by researchers.
No publisher has been paid.

In this direction, variation among sites turns out to be a particularly important feature to take into account. The simple models originally used in phylogenetics typically assume that all sites evolve under the same process of nucleotide or amino acid substitutions. In the case of amino acid sequence alignments, such models are typically implemented using so-called empirical amino acid replacement matrices, such WAG (Whelan and Goldman, 2001) or LG (Le and Gascuel, 2008). This, however, amounts to assuming that all sites should visit amino acid states at the same relative frequencies. Yet in practice, there is much variation among sites in substitution patterns (and in particular, in amino acid preferences). As it turns out (and as will be explored in more detail below), explicitly accounting for this heterogeneity across sites is crucial, in order to get more accurate tree reconstructions.

Owing to its complexity, pattern variation across sites is not a trivial aspect of the evolutionary process to model adequately. This problem has motivated (and is still motivating) the development of various approaches (Halpern and Bruno, 1998; Koshi and Goldstein, 1998; Lartillot and Philippe, 2004; Pagel and Meade, 2004; Wang et al., 2008; Le et al., 2008; Quang et al., 2008; Wang et al., 2014, 2018; Susko et al., 2018; Dang and Kishino, 2019). In particular, the CAT model (Lartillot and Philippe, 2004), such as implemented in PhyloBayes (Lartillot et al., 2013), relies on a Bayesian non-parametric approach based on Dirichlet process priors (see Chapter 1.4 [Lartillot 2020] for an introduction on the concepts of Bayesian inference, site-heterogeneity and non-parametric models). In this chapter, a practical application using PhyloBayes is presented, showing how to use both site-homogeneous and site-heterogeneous model, compare their results and evaluate their goodness of fit. The results are then interpreted in the broader context of phylogenomic analysis over broad evolutionary scales.

2 A practical example using PhyloBayes: Microsporidia

As a practical example of how to conduct a Bayesian phylogenetic analysis with PhyloBayes, we consider here a phylogenomic dataset originally assembled by Brinkmann et al. (2005). This dataset is a concatenation of 133 genes (24,000 aligned positions) for 40 taxa (34 eukaryotes and 6 Archaea). It represents an interesting case, for which the inferred position of the fast-evolving Microsporidia in the phylogeny of eukaryotes turns out to be model-dependent – and, more specifically, turns out to depend on whether or not site-specific amino acid preferences are accounted for.

All analyses presented here have been conducted using PhyloBayes MPI, version 1.8. The package can be obtained directly from github (`https://github.com/bayesiancook/pbmpi`). The dataset is also available along with the current version of the program. In what follows, we briefly recall the main points about program usage that are necessary to run through the complete analysis on this particular dataset. For more information, see the manual (provided in the package).

2.1 Running PhyloBayes under the CAT model

PhyloBayes is primarily intended for high-performance computing facilities operating under linux or unix. The package contains a series of programs (`pb_mpi`, `readpb_mpi`, `bpcomp`, `tracecomp`), all of which can be controlled using a command-line interface. Among them, `pb_mpi` implements the Markov chain Monte Carlo sampler targeting the posterior distribution over the parameters of the model chosen by the user. The MCMC sampler cycles over a complex series of Monte Carlo updates (or moves) of the topology, the branch lengths or the substitution model (including the Dirichlet process mixture), and saves the current model

configuration after each cycle. The series of points saved during a run of `pb_mpi` defines a *chain*. Each chain has a name, which is used as the base name for all files produced during the run.

Running a chain using `pb_mpi`

To start our analysis, we run a first chain under the CAT-F81 model on the Microsporidia dataset. This model, which combines uniform exchange rates across amino acid pairs (Felsenstein, 1981, generalized to amino acid states) with site-specific amino acid equilibrium frequency profiles from a Dirichlet process (the CAT model), was introduced by Lartillot and Philippe (2004) and represents the best compromise between computational speed and phylogenetic robustness. Runs under this model are much faster than under alternative models, such as considered below. Therefore, it is generally useful to start with CAT-F81, so as to get a first picture of the problem of interest, before launching computationally more demanding analyses. To run the chain, we type the following command:

```
mpirun -np 32 pb_mpi -d microsporidia.ali -cat -f81 -dgam 4 catf81microspo1 &
```

Here, we have started the analysis in direct mode. On a cluster operated by a job scheduling system, one would instead need to write a script containing, among other information, the command for running `pb_mpi`, and then send it to the queue.

In this command, the `-np 32` option specifies the number of processes running in parallel (this number should be at least 2). The `-d` option is for specifying the dataset. For the model, we combine a Dirichlet process for site-specific equilbirium frequency profiles over amino acids (the `-cat` option) with uniform (or Poisson) exchangeabilities (the `-f81` option). In addition, we allow for rate variation across sites, using a discretized gamma distribution with 4 categories (`-dgam 4`). Finally, we give a name to the chain, here, `catf81microspo1`. Before starting, the chain will output a summary of the model settings.

While the chain is running, a series of files will be produced. The most important are:

- `catf81microspo1.treelist`: list of sampled trees (with branch lengths);
- `catf81microspo1.trace`: the trace file, containing summary statistics (detailed below);
- `catf81microspo1.chain`: contains the parameter configurations visited during the run.

These files will be regularly updated (after each cycle). Note that the trace file contains one line per point saved since the beginning of the run. Thus, the number of lines of the trace file gives a direct indication of the current MCMC sample size. It is always good practice to run at least two chains in parallel and compare the samples obtained under these several independent runs. In the present case, we run four independent chains, which we name `catf81microspo1`, `2`, `3`, and `4`.

The chains will run as long as allowed. PhyloBayes implements a check-pointing system, so that chains can be interrupted at any time (possibly because of a timeout on the cluster) and then restarted. In the present case, on the machine where the example was conducted, the four independent chains save one point about every 30 seconds, or 150 points per hour. We let these chains run for 24 hours ($\sim$ 3500 points) before checking the results.

Checking convergence and mixing (`bpcomp` and `tracecomp`)

Convergence can be first visually assessed by plotting the summary statistics recorded in the trace files as a function of number of iterations. Visual assessment can be conducted while the chain is running. This can be done on the fly, directly from the command line interface, using simple linux utilities such as `gnuplot`. Alternatively, the trace file of `PhyloBayes` is

1.5:4 Bayesian phylogenetics

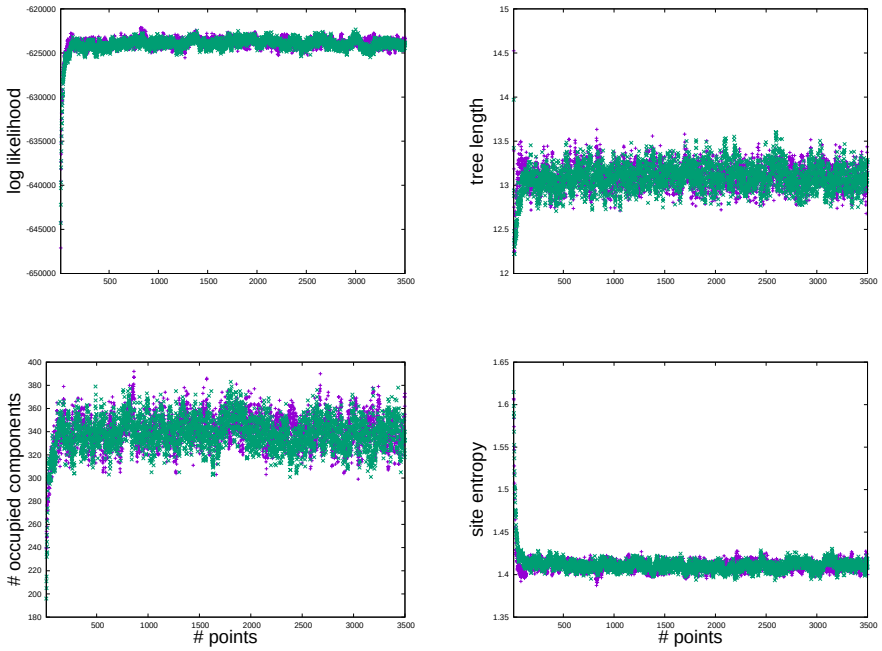

Figure 1 Traceplots for the CAT analyses, showing, as a function the number of points saved in the tracefile, the log likelihood, the tree length, the number of occupied components and the mean site entropy

compatible with the `Tracer` program (Rambaut et al., 2018). Visual assessment is essential, in particular, for gettting a reliable estimate of the burn-in, i.e the number of points before the chain has reached stationarity. In general, it is particularly important to visualize at least the log likelihood (`loglik`, 4th column of the trace file), the total tree length (`length`, column 5), the number of occupied components of the mixture (`Nmode`, column 6) and the mean site entropy (`statent`, column 7), which is a measure of the strength of site-specific amino acid preferences. In the present case, after 24 hours, the four independent chains have saved around 3500 points each. Visualization of the log likelihood and the summary statistics (Figure 1) suggests that the chains have reached convergence after a burnin of 400 to 500 points. We set the burnin conservatively to 500.

After visual inspection, convergence and mixing can be assessed more quantitatively. This can be done using the `tracecomp` program (for checking convergence of the parameters) and the `bpcomp` program (for assessing convergence in tree space). Both use a similar syntax. First, we inspect the trace files using tracecomp:

```
tracecomp -x 500 catf81micro1 catf81micro2 catf81micro3 catf81micro4
```

or, more rapidly

```
tracecomp -x 500 catf81micro?.trace
```

which produces an output summarizing the estimated effective sample size and the discrepancies among the four runs for each column of the trace file:

```
name                effsize     rel_diff
```

```
loglik          101     0.302739
length          614     0.0754705
alpha           611     0.0914714
Nmode           245     0.177897
statent         257     0.126228
statalpha       631     0.317062
kappa           471     0.120433
```

The effective size (second colum, `effsize`) is an estimate of the effective number of independent points produced by each run. As for the discrepancy (third column, `rel_diff`) it measures, for each statistic of the trace file, the deviation among the four chains in the mean value, normalized by the within-chain standard deviation of the statistic and averaged over all pairs of runs. A discrepancy much less than 1 means that the error on the posterior mean estimate for a given quantity is very small compared to the 95% credible interval. Here, all effective sizes are greater than 100 (that is, each chain yields the equivalent of at least 100 independent draws from the posterior distribution), and the discrepancies are less than or slightly above 0.3. The run is thus quite acceptable. Ideally, one would like to achieve effective sample sizes more in the order of 1000 or more, and discrepancies smaller than 0.1. However, this has typically been difficult to achieve in Bayesian phylogenomics, for reasonably large datasets. In the present case, the chains could be run for an additional few days, and the discrepancies would then decrease, eventually landing below 0.1 for all entries of the trace file. Of note, the differences that are implied by these discrepancies are relatively small in practice. For instance, in the case of `alpha` (the α parameter of the gamma distribution of rates across sites), which is the entry with the highest discrepancy (0.32), the 4 posterior mean estimates obtained for the 4 chains are all between 0.64 and 0.67 – thus implying very similar distributions of rates across sites. The reason why the discrepancies look large is that the standard deviation within each chain is small (around 0.03). Thus, the criterion of having all discrepancies below 0.1 is in fact fairly stringent.

Second, we inspect tree lists using `bpcomp`:

```
bpcomp -x 500 catf81micro1 catf81micro2 catf81micro3 catf81micro4
```

or, more rapidly:

```
bpcomp -x 500 catf81micro?.treelist
```

The program writes in the log the largest (`maxdiff`) and mean (`meandiff`) discrepancy observed across all bipartitions:

```
maxdiff     : 0.186092
meandiff    : 0.00323931

bipartition list in : bpcomp.bplist
consensus in        : bpcomp.con.tre
```

In the present case, the maximum difference in bipartition frequencies among the four chains is above 0.1 (equal to 0.186). Detailed inspection of the `bpcomp.bplist` file, however, shows that this is due only to discrepancies between chains about the position of *Glomus* within Fungi. For all other bipartitions, the discrepancy between the chains is below 0.1. Of note, a `maxdiff` of 1, which happens not so rarely in practice, warns us that at least one clade is inferred with a posterior probability of 1 in one chain and 0 in another chain, a clear sign of

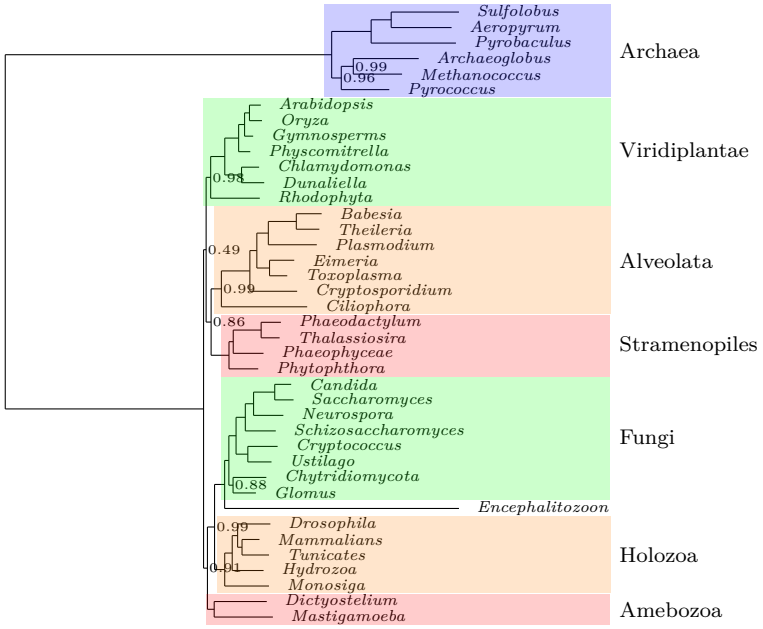

Figure 2 Consensus tree under the CAT-F81 model

MCMC mixing problems. In practice, however, as long as the discrepancies between chains does not directly affect the group of interest, this has usually been considered as acceptable.

Posterior consensus tree

The `bpcomp` program also produces a file containing the consensus obtained by pooling the trees of all of the chains given as arguments (file named `bpcomp.con.tre`). We see from this tree (Figure 2) that CAT-F81 infers that Microsporidia are the sister-group to Fungi, which is in accordance with the currently accepted view (Brinkmann et al., 2005). There is some uncertainty in the early splits at the base of eukaryotes, with posterior probability support values often smaller than 0.95. In fact, most of this lack of support is due to some hesitation about the branching point for the outgroup (Archaea). In the consensus displayed here, it is between Unikonta (Holozoa, Fungi, Amebozoa) and Bikonta (Viridiplantae, Alveolata and Stramenopiles), although this specific rooting point has a posterior probability of only 0.49.

2.2 Running PhyloBayes under other models

We can now run a similar analysis under other models. Here, we consider CAT-GTR, LG (Le and Gascuel, 2008) and GTR. The commands are as follows (in each case, with 4 replicates):

```
mpirun -np 32 pb_mpi -d microsporidia.ali -cat -gtr -dgam 4 catgtrmicrospo1 &
mpirun -np 32 pb_mpi -d microsporidia.ali -ncat 1 -gtr -dgam 4 gtrmicrospo1 &
mpirun -np 32 pb_mpi -d microsporidia.ali -ncat 1 -lg -dgam 4 lgmicrospo1 &
```

Note that the command-line option to select one-matrix models in PhyloBayes is just `-ncat 1`, that is, a mixture with one single category (one single matrix for all sites).

For these three models, the time per cycle is substantially longer than under CAT-F81: 1 point per minute for GTR and LG, and 1 point every 2 minutes for CAT-GTR. For the

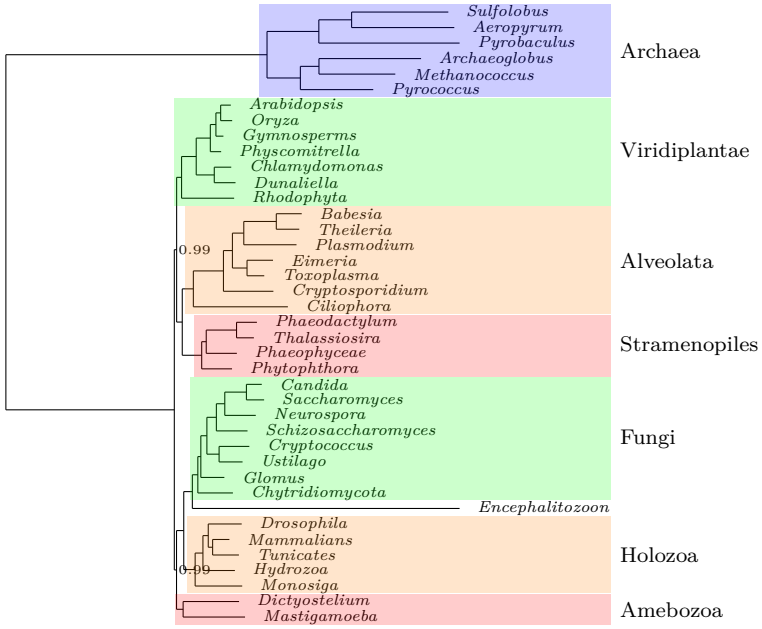

Figure 3 Consensus tree under the CAT-GTR model

CAT-GTR model, we will thus need 5 to 6 days in order for our chains to reach a size of 3500. Checking convergence on CAT-GTR after 5 days gives results similar to those obtained for CAT-F81. With `tracecomp`, effective sizes are all greater than 100. Discrepancies, on the other hand, are a bit larger between independent chains than what was observed with CAT-F81, although still acceptable:

```
name                effsize     rel_diff

loglik              374         0.434818
length              101         0.159781
alpha               980         0.178217
Nmode               466         0.123387
statent             241         0.38135
statalpha           291         0.126557
kappa               704         0.0996022
rrent               1005        0.21717
rrmean              2852        0.0308481
```

With `bpcomp`, reproducibility between chains is high:

```
maxdiff      : 0.0219639
meandiff     : 0.000298828
```

The resulting consensus tree (Figure 3) differs from that obtained under CAT-F81 only for the position of *Glomus*, although this might be due to the lack of convergence of the CAT-F81 analyses concerning the position of this particular taxon. Another remarkable difference, compared to CAT-F81, is the higher posterior probability support values obtained by CAT-GTR for the deep clades of the eukaryotic ingroup. The most probable rooting for eukaryotes

is, again, between Unikonta and Bikonta, although now with a posterior probability greater than 0.95. This pattern is often observed when comparing these 2 models: typically, CAT-F81 tends to be more conservative than CAT-GTR, giving lower clade posterior probabilities. Otherwise, the two models do not differ so much in their point estimates.

With LG and GTR, mixing turns out to be challenging, with different chains stabilizing at different levels. This is clearly detected both by `tracecomp` and `bpcomp`. First, the `tracecomp` output points to a very high discrepancy for the log likelihoods between chains, with a `rel_diff` statistic above 25:

```
name                effsize     rel_diff

loglik              1783        25.8753
length              658         2.24296
alpha               966         1.05931
Nmode               2400        0
statent             943         0.253293
statalpha           2400        0
rrent               761         0.263397
rrmean              1991        0.0536239
```

As for `bpcomp`, it gives us a maximal discrepancy of 1:

```
maxdiff     : 1
meandiff    : 0.025974
```

A closer look at the trace files shows that the log likelihood stabilizes for all chains, after less than 100 cycles. However, some chains stabilize at -727300, while other chains reach only -727600. Obviously, the latter are stuck in a local optimum in tree space. The difference in log likelihood between chains is so large in the present case that we can fairly exclude the chains stabilizing at -727600, as not representative of the true equilibrium, and concentrate on those chains that reach the highest average log likelihood. Restricting the `tracecomp` analysis to those chains gives much better convergence statistics:

```
name                effsize     rel_diff

loglik              1586        0.00714185
length              942         0.10133
alpha               505         0.180445
Nmode               2400        0
statent             389         0.139412
statalpha           2400        0
rrent               311         0.0898343
rrmean              1877        0.0254995
```

Running `bpcomp` on those chains gives a `maxdiff` of 0, indicating that they have stabilized on one single tree. Importantly, this tree differs from the ones obtained under CAT-F81 or CAT-GTR, in that Microsporidia now appear as being the sister-group to all other eukaryotes (Figure 4). In fact, this tree differs more specifically from the tree obtained by CAT-GTR by rotating the eukaryotic ingroup, so as to root it on Microsporidia. Equivalently, if we ignore the rooting, we might see this tree as the result of Archaea moving up in the tree and becoming the sister-group to Microsporidia. The same tree topology, with maximal support,

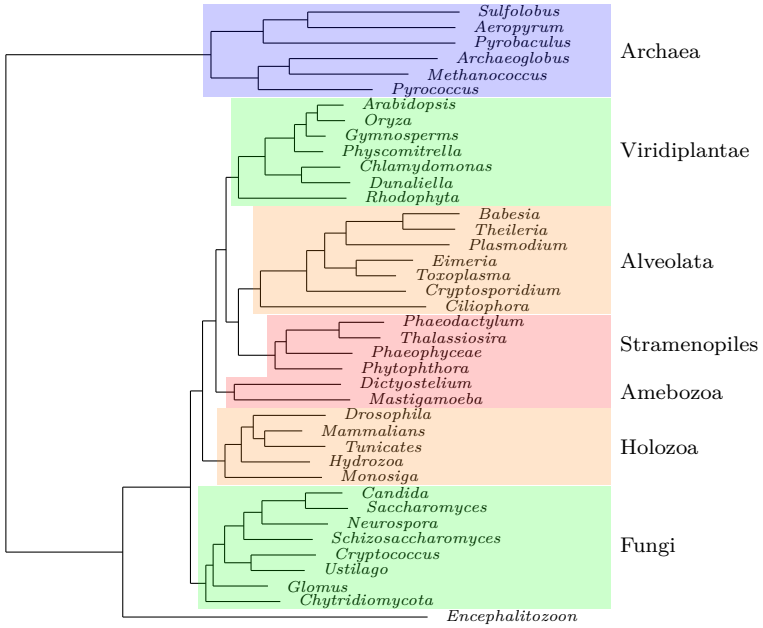

Figure 4 Consensus tree under the GTR model

is also obtained under LG (not shown). Of note, this result is confirmed by maximum likelihood analyses –such as conducted using RaxML (Stamatakis et al., 2005) for the LG model, or IQTree (Nguyen et al., 2015) for LG and GTR– which all give the topology obtained here using PhyloBayes under these 2 site-homogeneous models (not shown). This illustrates the fact that what determines the outcome of the analysis is not, in itself, the choice between maximum likelihood or Bayesian inference, but instead, the choice of the model of sequence evolution.

2.3 A typical case of long-branch attraction artifact

The analysis conducted above represents a situation where different models give different tree topologies based on the same dataset. On one side, site-homogeneous models (GTR, LG) give Microsporidia sister-group to all other eukaryotes (Figure 4); on the other side, site-heterogeneous models (CAT-F81, CAT-GTR) give Microsporidia sister-group to Fungi (Figures 2 and 3). Thus, at least one category of models is subject to a systematic error.

In the present case, we happen to have a relatively good independent knowledge of which of the two estimates obtained above is more likely to be correct. The position of Microsporidia in the eukaryotic tree is a well-known phylogenetic problem, having received a lot of attention since the early days of molecular phylogenetics. Based on the first universal trees reconstructed using 16S rRNA (Woese et al., 1990), it was at originally believed that Microsporidia were “early-emerging” eukaryotes, as also suggested here by GTR or by LG. However, there are now quite a few lines of evidence suggesting that Microsporidia are in fact closely related to Fungi (Brinkmann et al., 2005), as suggested by CAT-F81 and CAT-GTR.

In most practical situations, however, independent knowledge about the true tree is lacking. A case in point is the position of Ctenophores in the metazoan tree, which has received a lot of attention recently, and for which alternative hypotheses are still the subject

of some controversy (Telford et al., 2016). In such situations, independent and objective arguments are needed, in order to determine which of the alternative models is more likely to return an accurate estimate of the phylogeny.

In the case of Microsporidia, and assuming no independent knowledge, it can be noted that the two longest branches in the tree are, first, the branch leading to Microsporidia, and second, the branch between Eukaryotes and Archaea. The tree inferred by GTR is thus exactly what one would expect as the result of an artifactual attraction between these two long branches. Such arguments in terms of long-branch attraction artifacts are often useful. In many situations, they offer a good heuristic for making sense of the observed incompatibilities between the trees returned by alternative models, or the instability in tree estimation caused by varying the taxonomic sampling. These heuristic arguments, however, should be complemented by more formal model evaluation. This can be done along two main directions: model comparison and model checking.

2.4 Model comparison

Model comparison consists of measuring how well each model fits the data at hand. Choosing the best-fitting model (and the corresponding phylogenetic estimate) usually defines a good decision procedure, in particular if the difference in fit between alternative models is very large. In Bayesian inference, a classical measure of the fit of a model is its marginal likelihood, which is the likelihood averaged over the prior. The higher the marginal likelihood, the more likely it is that the model would produced the observed data, upon drawing a parameter configuration from its prior and then simulating the sequence evolutionary process. When comparing two models, it is customary to define the Bayes factor between these two models as the ratio of their marginal likelihoods (Jeffreys, 1935; Kass and Raftery, 1995; Gelman et al., 2004).

Marginal likelihoods have several drawbacks, however. First, on a theoretical ground, they are sensitive to the prior, and much more so than the posterior distribution itself. Second, on a more practical note, marginal likelihoods are notoriously difficult to numerically evaluate.

For those reasons, an alternative approach, used in PhyloBayes, is cross-validation. The idea of cross-validation is relatively simple: the data D are split into two subsets of unequal size, D_1 (typically, 90% of the original dataset) and D_2 (10% of the original dataset). The model is trained on D_1, and then the trained model is used to "predict" dataset D_2. In a Bayesian framework, this translates into averaging the likelihood of D_2 under the posterior distribution obtained by conditioning the model on D_1. This procedure is replicated over random splits of the data into D_1 and D_2. Of note, cross-validation automatically accounts for the dimensional penalty: a model that is overfitting (i.e. capturing non-reproducible random fluctuations) on D_1 is not expected to generalize well on unseen data D_2.

The entire procedure for cross-validation is computationally intensive and is thus not shown in detail here (see the manual available from the package for the detailed procedure). The results for the Microsporidia dataset are shown in Table 1. We see that the score of the CAT-GTR model (relative to LG) is much higher than that of CAT, which in turns has a better fit than GTR, and then finally LG. More generally, cross-validation on most empirical datasets over broad evolutionary scales (metazoans, eukaryotes, archaea, eubacteria, angiosperms, etc) invariably shows that site-heterogeneous models are much better fitting than one-matrix models, thus confirming the common-sense intuition that pattern heterogeneity is prevalent in empirical coding sequences.

Model	CV-score	standard deviation
GTR	288	28
CAT-F81	1154	122
CAT-GTR	2337	66

Table 1 Cross-validation scores (over 10 replicates) for the GTR, CAT-F81 and CAT-GTR models, relative to the LG model

2.5 Model checking by posterior predictive resampling

Measuring the empirical fit of alternative models represents a first fundamental principle for guiding the inference and the decision. On the other hand, it shows two main limitations. First, the fit of a model is a *global* measure of how well a model fits all aspects of the data, and not just those aspects that are relevant for the specific question being asked (here, the phylogenetic relationships). Thus, it can never be totally excluded that a lesser fitting model is in the end more accurate for phylogenetic reconstruction. Second, the fit is a *relative* measure, allowing one to determine the best among a series of models. Yet, even the best among the currently available models may not provide a good *absolute* fit to the data. For those reasons, it is essential to complement model comparison with model *checking* (i.e. using goodness-of-fit tests). Unlike model comparison, goodness-of-fit tests offer an absolute measure, by implementing a rejection test for each model taken individually. In addition, the test can be targeted, via the choice of summary statistics, to those aspects of the data that are deemed particularly relevant for the question being asked (Meng, 1994; Gelman et al., 2004).

In Bayesian inference, model checking is done using posterior predictive simulations. This method has been extensively used in Bayesian phylogenetics (Bollback, 2002; Lewis et al., 2014; Höhna et al., 2018). Posterior predictive checks can be seen as the Bayesian analogue of the parametric bootstrap: once the model has been conditioned on empirical data, the parameter configurations sampled from the posterior distribution are used to re-simulate replicates of the original dataset. Then, the value of some summary statistic of interest is computed on the simulated replicates, thus yielding a null distribution for the statistic under the fitted model. The value of the statistic computed on the original data is then compared to this null distribution. If it deviates significantly, this means that there is something in the empirical data which is not reproduced in the simulated replicates – and which is thus missed by the model.

The summary statistic used for the test should be designed so as to capture key features of the data that are deemed to be important in the context of the specific question under consideration. In the present case, we want to test each of the four models considered above, LG, GTR, CAT and CAT-GTR, concerning their ability to account for site-specific restrictions imposed by selection on amino acid usage. To this end, a simple statistic we can use is the amino acid *diversity*, i.e. the mean number of distinct amino acids per column across the sequence alignment. Under strong site-specific amino acid preferences, we expect a low diversity (a small subset of amino acids observed at each column).

Running a posterior predictive analysis can be done with the `readpb_mpi`, using the `-div` option for the diversity statistic (more general posterior predictive checks can be conducted using the `-ppred` option). In the case of the GTR model, the command is:

```
mpirun -np 8 readpb_mpi -x 500 1 -div gtrmicro1
```

and the program returns the following output:

```
diversity test
obs div : 4.07397
mean div: 4.67027 +/- 0.00923482
z-score : 64.5712
pp      : 0
```

The mean number of amino acids per column on the original alignment is 4.07. In contrast, the datasets simulated under the GTR model show on average 4.67 distinct amino acids per site. This difference is highly significant: the p-value is indistinguishable from 0, and the observed diversity is more than 64 standard deviations away from the mean of the posterior predictive null distribution (a p-value of 0.5 would approximately correspond to only 2 standard deviations away from the mean). In other words, the spectrum of amino acids present at each site in datasets simulated under GTR is too broad, compared to original sequence alignment, which indicates that the GTR model does not correctly reproduce (and thus, does not correctly capture) positional biochemical constraints. A similar result is obtained with LG (observed diversity is 72 standard deviations off the null distribution).

Doing the same experiment with CAT-F81 leads to a clearly different outcome:

```
diversity test
obs div : 4.07397
mean div: 4.07294 +/- 0.00810093
z-score : -0.126924
pp      : 0.5625
```

The diversity observed in data simulated under CAT-F81 is very close to the diversity of the original alignment (in fact, a bit smaller), and well within the posterior predictive null distribution ($p = 0.56$). This suggests that CAT-F81 adequately models site-specific amino acid propensities.

Finally, the CAT-GTR model, it is formally rejected by the test ($p < 0.01$):

```
diversity test
obs div : 4.07397
mean div: 4.0939 +/- 0.00850677
z-score : 2.3433
pp      : 0.00833333
```

However, the deviation between observed and posterior predictive diversity is much less than what was obtained above under the GTR or the LG model: the observed diversity (4.07) is now only 2.3 standard deviations away from the mean of the posterior predictive null distribution (4.09).

2.6 Pattern heterogeneity across sites and phylogenetic accuracy

We can now summarize the analysis and propose a global interpretation. Essentially two types of models were considered. On one side, LG and GTR assume pattern homogeneity across sites (i.e. invoke a single amino acid replacement process across all sites); on the other side, CAT and CAT-GTR explicitly account for site-specific amino acid preferences. Strikingly, the two models assuming pattern homogeneity give Microsporidia sister-group to all other eukaryotes, whereas the two models accounting for pattern heterogeneity give instead Microsporidia sister-group to Fungi.

Ignoring contextual knowledge about eukaryotic evolution, several independent lines of evidence suggest that site-heterogeneous models are more accurate in the present case. First, the tree produced by LG and GTR is a typical long-branch attract tree. Second, the much better relative fit of CAT-F81 and CAT-GTR, compared to LG and GTR, combined with the posterior predictive goodness-of-fit test using the diversity statistic, both show that site-specific amino acid preferences represent an important aspect of the true evolutionary process, which is not correctly captured by one-matrix models such as LG or GTR.

Why should the incorrect modeling of site-specific selective constraints make classical one-matrix models particularly sensitive to tree reconstruction errors? One main reason is that sites that are under strong biochemical constraints may nevertheless evolve rapidly – it is just that they stay within a small range of biochemically similar amino acids, which they keep repeatedly visiting. However, this fast evolution among a small number of possible amino acid states then makes it very likely for distantly related species to display the same amino acid at that site just by chance. Models ignoring site-specific amino acid preferences will underestimate this effect and will instead tend to incorrectly interpret the resulting identity by state as indicative of shared ancestry. As a result, they will underestimate sequence saturation and evolutionary distances (Halpern and Bruno, 1998) and will be more sensitive to long-branch attraction (Lartillot et al., 2007).

In the present case, we can get a rough estimate of the effective number of allowed amino acids per site by taking the exponential of the mean site entropy (8th. column of the trace file). Under the CAT-GTR model, this gives around 6.3 accepted amino acids per site, to be contrasted with 16.7 amino acids per site, according to the GTR model. Such a large discrepancy in the expectations under the two types of models as to the long-term probability of convergent evolution suggests that the effect of accounting for site-specific amino acid preferences on phylogenetic accuracy is likely to be substantial, and therefore represents a plausible explanation for the observed discrepancy between the two classes of models in the case of Microsporidia.

A similar phenomenon has been observed in several other well-characterized phylogenetic problems (e.g. Lartillot et al., 2007). More generally, there is now a broad array of empirical analyses showing that site-specific amino acid preferences represent an important aspect of the sequence evolutionary process, which is likely to negatively impact phylogenetic accuracy if not properly modeled. This is particularly true in the context of deep phylogenies (i.e. over broad evolutionary scales), for which sequence saturation is the rule and the risk of systematic errors in tree reconstruction is always an important concern. In the face of these problems, the site-heterogeneous models presented here certainly represent an important option to consider in a typical phylogenomic analysis.

In terms of practical recommendations, the best model is, by far, CAT-GTR, although the computational cost is high. A reasonable and computationally more efficient alternative is offered by CAT-F81. In spite of its rather crude approximations, CAT-F81 has generally proven more robust against long-branch attraction than one-matrix models in many situations (which is consistent with its good absolute fit under the posterior predictive diversity test). Therefore, a good procedure would be to always start with CAT-F81, so as to get a first idea of how much time it takes to obtain reasonable chains and to obtain a first series of useful results for the the dataset of interest. Then, if deemed affordable, a CAT-GTR analysis can be conducted (given that it would take about 6 to 8 times longer). A GTR (and possibly LG) analysis can also be conducted (possibly, with maximum likelihood implementations, such as RaxML [Stamatakis et al. 2005] or IQTree [Nguyen et al. 2015]). If the results are the same between site-homogeneous and site-heterogeneous models, then they can be considered

as robust. Conversely if the inferred tree topology turns out to depend on the model, then, a more thorough analysis along the lines proposed here can be conducted, using posterior predictive checks to make a stronger case about which model is likely to give a more accurate tree.

3 Challenges and perspectives

In summary, there is now good empirical evidence showing that accounting for pattern heterogeneity across sites makes an important difference when reconstructing deep phylogenies. The site-heterogeneous models implemented in PhyloBayes (CAT and CAT-GTR) currently represent the most radical approach – and perhaps the most accurate thus far – for capturing the modulations across sites in amino acid preferences. Over the recent years, the use of these models has proven instrumental for accurate inference on multiple practical cases.

All this comes at a cost, however. As it stands, the non-parametric random-effect models implemented in PhyloBayes are computationally very demanding. If the current implementation strikes a reasonable compromise between computational efficiency, model adequacy and phylogenetic accuracy for datasets of intermediate size (50 to 100 taxa, 10,000 to 30,000 positions), it does not scale up well to larger datasets, such as those that are currently contemplated in phylogenomics. For instance, resolving the relationships at the base of the metazoan tree, and addressing particularly difficult questions such as the position of ctenophores, seems to require datasets of the order of at 100,000 to 500,000 aligned positions (Simion et al., 2017). For such large datasets, the current implementation is not usable in practice, at least not directly on the full dataset.

In the face of these limitations, pragmatic solutions have been considered. A first simple approach is to use jackknife resampling, i.e. repeating the Bayesian analysis on subsets of genes or sites drawn without replacement from the original dataset and then averaging out the results over the replicates (Delsuc et al., 2008; Simion et al., 2017). Of note, this approach is not purely Bayesian and is admittedly a work-around. However, because of the additional layer of non-parametric resampling, jackknife should in fact produce robust estimates of the statistical support for the phylogeny.

Alternatively, other non strictly Bayesian approaches are currently being explored, most of which aim at proposing reasonable approximations explicitly accounting for site-specific amino acid preferences, while preserving the benefits of the increased robustness against tree reconstruction errors: posterior mean site frequency approximations (Wang et al., 2018), better empirical finite mixtures (Susko et al., 2018), penalized likelihood (Tamuri et al., 2014) or variational approaches (Dang and Kishino, 2019). These represent promising developments.

References

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387(6632):489–493.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, 19(7):1171–1180.

Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., and Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.*, 54(5):743–757.

Dang, T. and Kishino, H. (2019). Stochastic Variational Inference for Bayesian Phylogenetics: A Case of CAT Model. *Mol. Biol. Evol.*, 36(4):825–833.

Delsuc, F., Tsagkogeorga, G., Lartillot, N., and Philippe, H. (2008). Additional molecular support for the new chordate phylogeny. *Genesis*, 46(11):592–604.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15(7):910–917.

Höhna, S., Coghill, L. M., Mount, G. G., Thomson, R. C., and Brown, J. M. (2018). P3: Phylogenetic Posterior Prediction in RevBayes. *Mol. Biol. Evol.*, 35(4):1028–1034.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.*, 31:203–222.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *J. Am. Stat. Assoc.*, 90(430):773–795.

Koshi, J. M. and Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins*, 32(3):289–295.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.*, 7 Suppl 1:S4.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21(6):1095–1109.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, 62(4):611–615.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, 25(7):1307–1320.

Le, S. Q., Lartillot, N., and Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 363(1512):3965–3976.

Lewis, P. O., Xie, W., Chen, M.-H., Fan, Y., and Kuo, L. (2014). Posterior predictive Bayesian phylogenetic model selection. *Syst. Biol.*, 63(3):309–321.

Meng, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.*, pages 1142–1160.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32(1):268–274.

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, 53(4):571–581.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.*, 9(3):e1000602.

Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.*, 22(5):1246–1253.

Quang, L. S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.*, 67(5):901–904.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, E., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., and Manuel, M. (2017). A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol.*, 27(7):958–967.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463.

Susko, E., Lincker, L., and Roger, A. J. (2018). Accelerated Estimation of Frequency Classes in Site-Heterogeneous Profile Mixture Models. *Mol. Biol. Evol.*, 35(5):1266–1283.

Tamuri, A. U., Goldman, N., and Dos Reis, M. (2014). A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics*, 197(1):257–271.

Telford, M. J., Moroz, L. L., and Halanych, K. M. (2016). Evolution: A sisterly dispute. *Nature*, 529(7586):286–287.

Wang, H.-C., Li, K., Susko, E., and Roger, A. J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.*, 8:331.

Wang, H.-C., Minh, B. Q., Susko, E., and Roger, A. J. (2018). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.*, 67(2):216–235.

Wang, H.-C., Susko, E., and Roger, A. J. (2014). An amino acid substitution-selection model adjusts residue fitness to improve phylogenetic estimation. *Mol. Biol. Evol.*, 31(4):779–792.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, 18(5):691–699.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA*, 87(12):4576–4579.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol*, 8(2):357–366.