…The problem

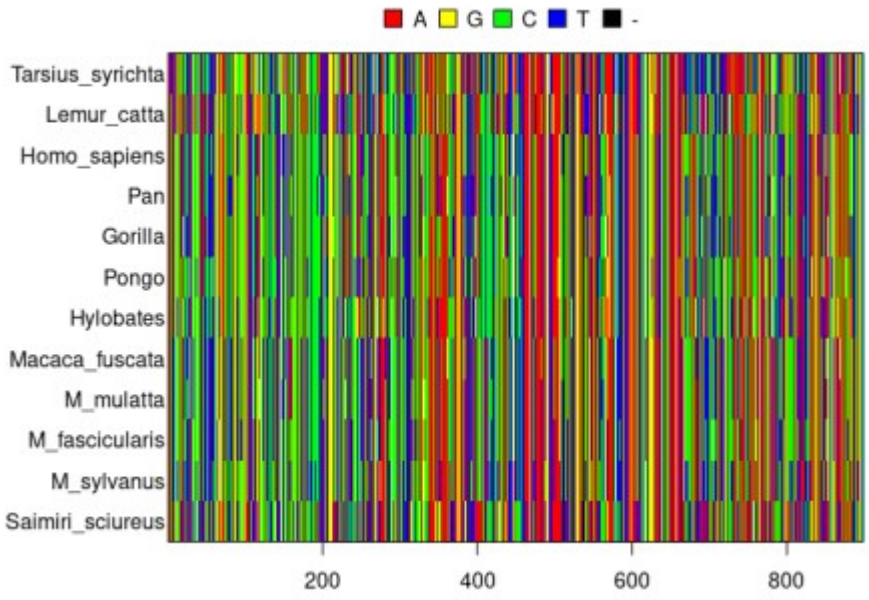
Imagine you are a field biologist. All around the world, you captured multiple bird species of which you ob blood sample. From the blood, you have extracted the DNA. Using DNA, one can determine how these s are evolutionarily related. The problem is, which model of evolution do you assume for your birds?

To illustrate the problem of picking the right model of evolution, we start from the DNA sequences of prim (we will abandon birds here). To be more precise, we will be using a DNA *alignment*, which are DNA sequ that are arranged in such a way that similar parts of the DNA sequences are at the same position in the sequences. The DNA alignment we use needs first to be converted from NEXUS to FASTA format:

```
library(beastier) # beastier is part of babette
fasta_filename <- tempfile("primates.fasta")
save_nexus_as_fasta(
get_beast2_example_filename("Primates.nex"),
fasta_filename
)
```

DNA consists of a long string of four different elements called nucleotides, resulting in a four letter alphab encoding for the proteins a cell needs. In our case, we do not have the full DNA sequence of all primates ‘only’ 898 nucleotides. Here I show the DNA sequences:

```
library(ape)
par0 <- par(mar = c(3, 7, 3, 1))
dna_sequences <- read.FASTA(fasta_filename)
image.DNAbin(dna_sequences, mar = c(3, 7, 3, 1))
```



DNA alignment of primates

```
par(par0)
```

From this DNA alignment, we can use the R package babette [23] to estimate the evolutionary history of th species.

First, we'll load babette:

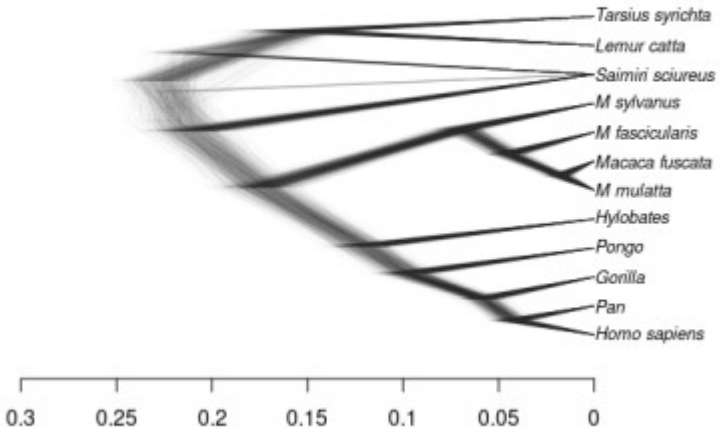
```
library(babette, quietly = TRUE)
```

Here, we estimate the evolutionary history of these species:

```
out <- bbt_run_from_model(fasta_filename)
```

An evolutionary history can be visualized by a tree-like structure called a phylogeny. babette, however, cr multiple phylogenies of which the more likelier ones show up more often. This results in a visualization th shows the uncertainty of the inferred phylogenies:

```
plot_densitree(
out$primates_trees[9000:10000],
alpha = 0.01
)
```



the estimated evolutionary history of primates

☐ The problem?

As we have observed, inferring the evolutionary history from DNA sequences is easy. The open question have we used the best evolutionary model?

This is where mcbette can help out. mcbette is an abbreviation of 'Model Comparison using babette' and to pick the best evolutionary model, where 'best' is defined as 'the evolutionary model that has been mos to have generated the alignment, from a set of models'. The addition of 'from a set of models' is importan there are infinitely many evolutionary models to choose from.

So far in this example we have used babette's default evolutionary model. An evolutionary model consist among others, three most important parts, which are the site, clock and tree model. The site model encompasses the way the (in our case) DNA sequence changes through time. The clock model embodie rate of change over the different (in our case) species. The tree model specifies the (in our case) speciati model, that is how the branches of the trees are formed.

Let's figure out what a default babette evolutionary model assumes.

```
default_model <- create_inference_model()
print(
paste0(
"Site model: ", default_model$site_model$name, ". ",
"Clock model: ", default_model$clock_model$name, ". ",
"Tree model: ", default_model$tree_prior$name
)
)
```

```
[1] "Site model: JC69. Clock model: strict. Tree model: yule"
```

Apparently, the default site model embeds a Jukes-Cantor nucleotide substitution model (i.e. all nucleotid mutations are equally likely), the default clock model is strict (i.e. all DNA sequences change at the same and the speciation model is Yule (i.e. speciation rates are constant and extinction rate is zero). These def settings are picked for a reason: these are the simplest site, clock and tree model.

The question is if this default evolutionary model is the most likely to have actually resulted in the original alignment. It is easy to argue that the site, clock and tree model are overly simplistic (they are!).

☐ The competing model

In this example, I will let the default evolutionary model compete with only one other evolutionary model. there are plenty of options! Tip: to get an overview of all inference models, view the `inference models` vignette of the beautier package (which is part of babette), or go to URL https://cran.r-project.org/web/pac /beautier/vignettes/inference_models.html.

Here, I create the competing model:

```
competing_model <- create_inference_model(
clock_model = create_rln_clock_model()
)
```

The competing model has a different clock model: 'rln' stands for 'relaxed log-normal', which denotes tha different species can have different mutation rates, where these mutation rates follow a log-normal distrib

☐ Getting the results

We must modify our inference model first, to prepare them for model comparison:

```
default_model$mcmc <- create_ns_mcmc(particle_count = 16)
competing_model$mcmc <- create_ns_mcmc(particle_count = 16)
```

Increasing the number of particles improves the accuracy of the marginal likelihood estimation. Because t
accuracy is also estimated, we can also see how strongly to believe a model is better.

Now, we load mcbette, 'Model Comparison using babette' to do our model comparison:

```
library(mcbette)
```

Then, we let mcbette estimate the marginal likelihoods of both models. The marginal likelihood is the likel
to observe the data given a model, which is exactly what we need here. Also note that this approach to c
models has no problems to honestly compare models with a different amount of parameters; there is a na
penalty for more models with more parameters.

```
marg_liks <- est_marg_liks(
fasta_filename = fasta_filename,
inference_models = list(
default_model,
competing_model
)
)
```

Note that this calculation takes quite some time!

Here we show the results as table:

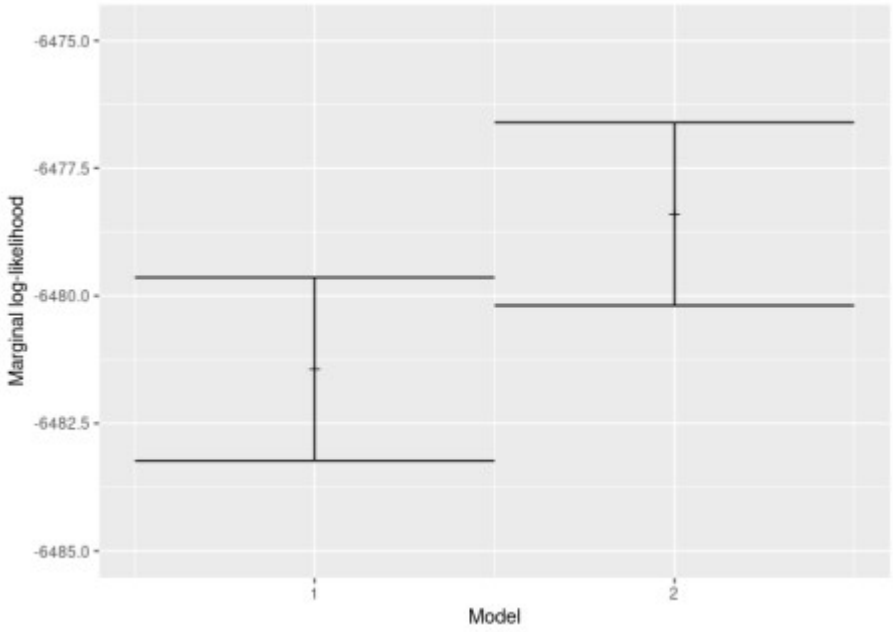
```
knitr::kable(marg_liks)
```

site_model_name	clock_model_name	tree_prior_name	marg_log_lik	marg_log_lik_sd	weight
JC69	strict	yule	-6481.435	1.794633	0.0457542 1
JC69	relaxed_log_normal	yule	-6478.397	1.792379	0.9542458 2

The most important column to look at here is the `weight` column. All (two) weights sum up to one. A mo
weight is its relative chance to have observed the alignment given the model. As can be seen, the weight
more complex (relaxed log-normal) clock model is higher (i.e. 0.9343919).

We can also visualize which model is the best, by plotting the estimated marginal likelihoods and the erro
estimation:

```
plot_marg_liks(marg_liks)
```



the estimated marginal likelihoods

Note that marginal likelihoods can be very close to zero. Hence, mcbette use log values. The model with lowest log value, thus has the highest marginal likelihood and is thus more likely to have resulted in the d

...