RevBayes – Phylogenies and the comparative method

Nicolas Lartillot

August 26, 2014

Introduction

The subject of the comparative method is the analysis of trait evolution at the macroevolutionary scale. Traits can be morphological, behavioral or can be related to the life-history strategies (e.g. longevity). In a comparative context, many different questions can be addressed: tempo and mode of evolution, correlated evolution of multiple quantitative traits, trends and bursts, changes in evolutionary mode correlated with major key innovations in some groups, etc (see Harvey and Pagel, 1991, for a good introduction).

In order to correctly formalize comparative questions, the underlying phylogeny should always be explicitly accounted for. This point is clearly illustrated, in particular, by the independent contrast method (Felsenstein, 1985). Practically speaking, the phylogeny and the divergence times are usually first estimated using a separate phylogenetic reconstruction software. In a second step, this time-calibrated phylogeny is used as an input to the comparative method. Doing this, however, raises a certain number of methodological problems:

- the uncertainty about the phylogeny (and about divergence times) is ignored
- the traits themselves may have something to say about the phylogeny
- the rate of substitution, and more generally the parameters of the substitution process, can also be seen as quantitative traits, amenable to a comparative analysis.

All these points are not easily formalized in the context of the step-wise approach mentioned above. Instead, what all this suggests is that phylogenetic reconstruction, molecular dating and the comparative method should all be considered jointly, in the context of one single overarching probabilistic model.

Thanks to its modular structure, RevBayes represents a natural framework for attempting this integration. The aim of the present tutorial is to guide you through a series of examples where this integration is achieved, step by step. It can also be considered as an example of the more general perspective of *integrative modeling*, which can be recruited in many other contexts.

Data and files

In the `data` folder, you will find the following files

- `plac40lhtlog.nex`: 3 life-history traits (age at sexual maturity, body mass, maximum recorded lifespan) for 40 placental mammals (taken from the Anage database, de Magalhaes and Costa, 2009). The traits have been log-transformed.
- `plac40_4fold.nex`: an alignment made of a concatenation of 17 nuclear genes in 40 placental mammals (from Lartillot and Delsuc, 2012), with only the four-fold degenerate third coding positions.
- `chronoplac40.tree`: a time-calibrated phylogeny, which has been obtained by running another software program (PhyloBayes, Lartillot *et al.*, 2009). On the cluster, and if you are logged under an X-terminal, you can visualize this tree using the `njplot` command:
  ```
  njplot chronoplac40.tree
  ```
- `archaeaRNA.nex`: an alignment of rRNA sequences of 33 archaeal species.
- `archaeaTemp.nex`: optimal growth temperatures for the 33 archaeal species.

1 Univariate Brownian evolution of quantitative traits

As a first preliminary exercise, we wish to reconstruct the evolution of body mass in placental mammals and, in particular, estimate the body mass of their last common ancestor. For this, we will assume that the logarithm of body mass follows a simple univariate Brownian motion along the phylogeny. In a first step, we will ignore phylogenetic uncertainty: thus, we will assume that the Brownian process describing body mass evolution runs along a fixed time-calibrated phylogeny (with fixed divergence times), such as specified in the file `chronoplac40.tree`.

You may want to take the time to visualize the tree given in `chronoplac40.tree` as well as the matrix of quantitative traits specified by the `plac40lhtlog.nex` file, before going into the modeling work described below.

The model and the priors

A univariate Brownian motion $x(t)$ is parameterized by its starting value at the root of the phylogeny $x(0)$ and a rate parameter σ. This rate parameter tunes the amplitude of the variation per unit of time. Specifically, along a given time interval $(0, T)$, the value of X at time T is normally distributed, with mean $x(0)$ and variance $\sigma^2 T$:

$$x(T) \quad \sim \quad \text{Normal}\left(x(0), \sigma^2 T\right).$$

Since we have not much prior information about the possible values of σ, we should use a diffuse prior, for instance, an exponential of mean 100:

$$\sigma \sim \text{Exponential}(0.001).$$

Concerning the initial value $x(0)$ of the Brownian process at the root of the phylogeny, the current version of RevBayes only implements a uniform prior. This is done by default (no need to explicitly define it). Note that this uniform prior is invariant by translation. Given that our traits have been log-transformed, this means that the prior is invariant upon changing the units in which the traits are measured: it is insensitive to the fact that, for example, body mass has been given in kilograms, and not in grams.

Finally, the tree topology τ is, as mentioned above, fixed to some externally given phylogeny. However, in order to fully specify the model in RevBayes, we still need to define a prior on this tree: here, we will simply use a uniform prior, both on the topology and on divergence times.

The entire model is now specified: tree τ, variance σ and Brownian process $x(t)$:

$$\begin{aligned}
\tau &\sim \text{Uniform}, \\
\sigma &\sim \text{Exponential}(0.001), \\
x(0) &\sim \text{Uniform}, \\
x(t) &\sim \text{Brownian}\left(x(0),\, \tau,\, \sigma\right).
\end{aligned}$$

Constraining τ based on an externally-provided phylogeny and conditioning the model on empirical data by clamping $x(t)$ at the tips of the phylogeny, we can then run a MCMC to sample from the joint posterior distribution on σ and x. Once this is done, we can obtain posterior means, medians or credible intervals for the value of body mass or other life-history traits for specific ancestors.

Programming the model in RevBayes

We just need to reformulate what has been explained in the last subsection, now in the Rev language:

- load trait data:

```
contData <- readCharacterData("data/plac40lhtlog.nex")
```

- get taxon number and taxon names:

```
nTaxa <- contData.ntaxa()
names <- contData.names()
```

- define a uniformly-distirbuted time-tree:

```
tau ~ uniformTimeTree(originTime = 1.0, taxonNames = names)
```

- set this tree equal to the one defined in `chronoplac40.tree`:

  ```
  treeArray <- readTrees("data/chronoplac40.tree")
  fixedTree <- treeArray[1]
  tau.setValue(fixedTree)
  ```

- define σ:

  ```
  sigma ∼ Exponential(0.001)
  ```

  to accelerate convergence, it can be useful to force initialization of σ to a small value:

  ```
  sigma.setValue(0.1)
  ```

- define the multivariate Brownian process, which we will call `logmass`:

  ```
  logmass ∼ brownian(tau,sigma)
  ```

- condition the Brownian model on empirically observed values for body mass in extant taxa. Here, we need to specify that body mass is the second column of the dataset:

  ```
  logmass.clampAt(contData,2)
  ```

The model is now entirely specified. We can define the moves on its parameters:

- initialize a running index for storing moves:

  ```
  index <- 1
  ```

- push a scaling move on σ:

  ```
  moves[index] <- mScale(sigma, lambda=2.0, tune=true, weight=3.0)
  index <- index + 1
  ```

- a sliding move on the Brownian process

  ```
  moves[index] <- mvRealNodeValTreeSliding(process=logmass, lambda=10, tune=true,
  weight=100)
  index <- index + 1
  ```

- a global translation move on the Brownian process:

  ```
  moves[index] <- mvRealNodeValTreeTranslation(process=logmass,lambda=1,
  tune=true,weight=1)
  index <- index + 1
  ```

- before creating the model, we define summary statistics, to be monitored during the MCMC: the mean and the standard deviation of the trait across the tree:

  ```
  meanlogmass := logmass.mean()
  stdevlogmass := logmass.stdev()
  ```

- as well as the value of the log of body mass for the root:

  ```
  rootlogmass := logmass.rootVal()
  ```

- now, create the model

  ```
  mymodel <- model(sigma)
  ```

- make a screen monitor that tracks the summary statistics of interest:

  ```
  monitors[1] <- screenmonitor(printgen=10, sigma, rootlogmass, meanlogmass, stdevlogmass)
  ```

- a file monitor that does the same thing, but directly into a file:

  ```
  monitors[2] <- filemonitor(filename="output/placmass.trace", printgen=10, separator
  = " ", sigma, rootlogmass, meanlogmass, stdevlogmass)
  ```

- a file monitor for the ancestral reconstruction of traits along the entire tree (in newick format):

  ```
  monitors[3] <- filemonitor(filename="output/placmass.logmass", printgen=10, separator
  = " ", logmass)
  ```

- and a general model monitor:

  ```
  monitors[4] <- modelmonitor(filename="output/placmass.log", printgen=10, separator
  = " ")
  ```

We can finally create a mcmc, and run it for a good 100 000 cycles:

```
mymcmc <- mcmc(mymodel, monitors, moves)
mymcmc.burnin(generations=100,tuningInterval=100)
mymcmc.run(100000)
```

Exercises

- run the model, check convergence and obtain a sample from the posterior distribution
- take the time to explore the files produced by the run: in particular, the `.logmass` file in which the sampled histories of body mass evolution are written in the newick format.
- using `Tracer`, visualize the posterior distribution on ancestral placental body mass
- calculate the credible interval for this ancestral body mass
- sumarize the evolution of body mass along the tree, visualizing it with `njplot`
- calculate a 95% credible interval for the rate of evolution of the logarithm of body mass (σ)

You can of course conduct a similar inference for each of the three life-history traits given in `plac40lhtlog.nex`. You could either run the model separately on each trait or, alternatively, define (within the same model) three independent univariate Brownian motions running along the same phylogeny (each with its own σ parameter).

2 Correlated evolution of multiple traits

Next, we would like to estimate the correlation between the $K = 3$ life-history traits given in the `plac40lhtlog.nex` file, while properly taking into account phylogenetic inertia. To do so, we will assume that the traits jointly evolve along the phylogeny as a *multivariate* Brownian process. We will estimate the *covariance matrix* of this process. The empirical support in favor of positive or negative correlations between pairs of traits will then be formalized in terms of posterior probabilities of having positive or negative entries in this covariance matrix. As a by-product of this correlation analysis, we will also obtain a marginal ancestral reconstruction of the traits along the entire phylogeny, which we will then visualize using graphical software programs. At this stage of the tutorial, we will again ignore phylogenetic uncertainty.

The model and the priors

A multivariate Brownian process $X(t)$, of dimension K (here $K = 3$), is entirely parameterized by its starting value ($X(0)$ at the root of the phylogeny, which a vector of dimension K) and a $K \times K$ symmetric positive matrix (the covariance matrix), which we will call Σ. A positive entry between two traits, say $\Sigma_{12} > 0$, means that when trait 1 increases, trait 2 also tends to increase. Conversely, a negative entry means that the two traits tend to undergo variation in opposite directions. As for the diagonal entries (e.g. Σ_{11}), they represent the variance per unit of time (i.e. the rate of evolution) of each trait considered marginally, thus very much like σ^2 (*not* σ) in the univariate model of the previous section.

On Σ, we will assume an inverse-Wishart prior:

$$\Sigma \quad \sim \quad W^{-1}(\Sigma_0, d).$$

The inverse Wishart distribution has two parameters: a symmetric positive matrix Σ_0, of same dimension as Σ, and a natural number $d \geq K$ (called the number of degrees of freedom). When $d > K + 1$, the inverse Wishart prior has a well defined mean, which is equal to $\frac{\Sigma_0}{d-K-1}$ and is all the more concentrated around this mean than d is large. The choice of this prior is primarily motivated by the fact that it is analytically conjugate to the multivariate normal distribution. It is therefore very convenient when used for the covariance matrix of a Brownian motion.

Since we want a diffuse prior, we will use a small value for d, e.g. $d = K + 2$, which is the smallest value for which the mean of the distribution is well-defined. In addition, we want the prior to be centered on a diagonal matrix (i.e. we want to be a priori *equi-poised*, or indifferent, with respect to either positive or negative correlations among traits). We may also want our prior to be invariant by linear reparameterizations of traits (linear transformations on the logarithm of life-history traits are equivalent to allometric re-parameterizations of the traits considered in natural units). This invariance will hold if and only if Σ_0 is proportional to the identity matrix, i.e. if $\Sigma_0 = \kappa I_K$, for some positive real number κ. This number κ will set the amplitude of the variation per unit of time of the traits. Since we have no idea about the scale of κ, we should use a diffuse prior, e.g. an exponential of mean 100:

$$\kappa \sim \text{Exponential}(0.001).$$

This completes our model:

$$\begin{aligned}
\tau &\sim \text{Uniform}, \\
\kappa &\sim \text{Exponential}(0.001), \\
\Sigma &\sim W^{-1}(\Sigma_0 = \kappa I_K,\, d = K+2), \\
X(0) &\sim \text{Uniform}, \\
X(t) &\sim \text{Brownian}(X(0),\, \tau,\, \Sigma).
\end{aligned}$$

As in the univariate case, we can then constrain τ, clamp X and sample from the joint distribution over the parameters of the model by MCMC. Once this is done, we can estimate marginal posterior probabilities (e.g. for positive or negative covariance among traits) and infer ancestral traits.

Programming the model in RevBayes

We go through the same series of steps as in the univariate case:

- load trait data:

  ```
  contData <- readCharacterData("data/plac40lhtlog.nex")
  ```

- get number of traits, number of taxa and taxon names:

  ```
  nTraits <- contData.nchar()[1]
  nTaxa <- contData.ntaxa()
  names <- contData.names()
  ```

- define a uniformly-distirbuted time-tree:

  ```
  tau ~ uniformTimeTree(originTime = 1.0, taxonNames = names)
  ```

- set this tree equal to the one defined in `chronoplac40.tree`:

  ```
  treeArray <- readTrees("data/chronoplac40.tree")
  fixedTree <- treeArray[1]
  tau.setValue(fixedTree)
  ```

- define κ:

  ```
  kappa ~ Exponential(0.001)
  ```

- define the number of degrees of freedom as $d = K + 2$:

  ```
  df <- nTraits+2
  ```

- define the covariance matrix Σ as inverse Wishart:

  ```
  Sigma ~ invWishart(dim=nTraits, kappa=kappa, df=df)
  ```

  note that we are using a special constructor for the inverse Wishart: when a natural number d and a positive real number κ are given, it is implicitly assumed that the matrix-valued parameter is κ times the identity matrix of dimension d: κI_d.

- define the multivariate Brownian process:

  ```
  X ~ mvtBrownian(tau,Sigma)
  ```

- condition the Brownian model on quantitative trait data. This needs to be done separately for each trait:

  ```
  for (i in 1:nTraits) { X.clampAt(contData,i,i) }
  ```

  Here, we give twice the index `i` to the `clampAt` function: the first corresponds to the entry of the Brownian process, and the second one to the column of the data matrix. In some cases (as we will see below), the Brownian process and the data matrix may not be of same dimension, and therefore, it will be useful to be able to specify arbitrary maps between them.

The model is now entirely specified. We can define the moves on its parameters.

- initialize a running index for storing moves:

  ```
  index <- 1
  ```

- push a scaling move on κ:

  ```
  moves[index] <- mScale(kappa, lambda=2.0, tune=true, weight=3.0)
  index <- index + 1
  ```

- a sliding move on the Brownian process

  ```
  moves[index] <- mvMultivariateRealNodeValTreeSliding(process=X, lambda=10,
  tune=true,weight=100)
  index <- index + 1
  ```

- a global translation move on the Brownian process (component-wise, that is, a random global translation across the entire phylogeny is applied to one trait taken at random):

  ```
  moves[index] <- mvMultivariateRealNodeValTreeTranslation(process=X, lambda=1,
  tune=true, weight=1)
  index <- index + 1
  ```

- finally, a conjugate Gibbs move for Σ: as it turns out, conditional on κ and the Brownian process X, it is possible to directly resample Σ from its conditional posterior distribution (Lartillot and Poujol, 2011). In RevBayes, this is implemented as follows:

  ```
  moves[index] <- mvConjugateInverseWishartBrownian(sigma=Sigma, process=X,
  kappa=kappa, df=df, weight=1)
  index <- index + 1
  ```

Before creating the model, we need to define a few summary statistics, which we want to track during MCMC, either to monitor convergence or for obtaining interesting outputs. First, suppose you are specifically interested in the covariance and the correlation coefficient associated with the joint variation of body-size (trait 2) and longevity (trait 3). You may also be interested in the *partial* correlation coefficient between body mass and longevity, i.e. while controlling for variation in age at sexual maturity. These three quantities can be singled out and named as follows:

- the covariance:
  `cov23 := Sigma.covariance(2,3)`
- the correlation coefficient:
  `cor23 := Sigma.correlation(2,3)`
- the partial correlation:
  `parcor23 := Sigma.partialCorrelation(2,3)`
- note that the variance per unit of time of, say, log body mass is given by the diagonal entry:
  `var2 := Sigma.covariance(2,2)`
- we can also get all correlation coefficients into a single vector:

  ```
  # initialize a running index
  corrindex <- 1
  # loop over all pairs of traits
  for (i in 1:nTraits)    {
      for (j in i+1:nTraits) {
          correl[corrindex] := Sigma.correlation(i,j)
          corrindex <- corrindex + 1
      }
  }
  ```

- we could be interested in tracking several summary statistics also for the Brownian motion, in particular the mean and the standard deviation along the tree, separately for each trait:

  ```
  for (i in 1:nTraits)    {
          meanX[i] := X.mean(i)
          stdevX[i] := X.stdev(i)
  }
  ```

After creating the model, all these new variables (`cor12`, `correl`, `meanX`, etc) can be monitored, along with the other parameters of the model:

- create the model

  ```
  mymodel <- model(kappa)
  ```

- make a screen monitor that tracks correlation coefficients and mean Brownian values:

  ```
  monitors[1] <- screenmonitor(printgen=10, correl, meanX)
  ```

- a file monitor that does the same thing, but directly into a file:

  ```
  monitors[2] <- filemonitor(filename="output/placmass.trace", printgen=10, separator
  = " ", correl, meanX)
  ```

- a file monitor for Σ:

  ```
  monitors[2] <- filemonitor(filename="output/plactraits.cov", printgen=10, separator
  = " ", Sigma)
  ```

- a file monitor for the ancestral reconstruction of traits:

  ```
  monitors[3] <- filemonitor(filename="output/plactraits.traits", printgen=10, separator
  = " ", X)
  ```

- and a general model monitor:

  ```
  monitors[4] <- modelmonitor(filename="output/plactraits.log", printgen=10, separator
  = " ")
  ```

We can finally create the mcmc and run it:

```
mymcmc <- mcmc(mymodel, monitors, moves)
mymcmc.burnin(generations=100,tuningInterval=100)
mymcmc.run(100000)
```

Exercises

- using `Tracer`, visualize the posterior distribution on the correlation coefficient between mass and longevity.

- estimate the posterior mean, median and 95% credible interval for this correlation coefficient.

- does the credible inrerval overlap 0? What does that say about the empirical support for the correlation between body mass and longevity?

- which fraction of the total variance in longevity among placental mammals is explained by body-mass?

- sumarize evolution of body size and longevity along the tree, visualizing it with `njplot`

- calculate the credible interval for body size of the last common ancestor of placental mammals

- as mentioned above, it would be easy to make an *uncorrelated* version of the same model, either by constraining Σ to be a diagonal matrix or by creating three independent Brownian models of dimension 1, one for each trait (the latter is easier under the current version of `RevBayes`, if you want to try it after the workshop). Would this uncorrelated model have a better or a poorer fit than the covariant model constructed above? Why? So, do we need to explicitly compute a Bayes factor in the present case?

3 Accounting for uncertainty in divergence times

Starting from the model implemented in the last section, we now want to account for phylogenetic uncertainty. As first pointed out by Huelsenbeck and Rannala (2003), this can easily be done in a Bayesian framework, through the use of a joint model combining sequence data and quantitative traits. Specifically:

- two data sets are loaded: one for sequence data and one for quantitative traits
- a tree is defined (here, with a uniform prior, but this could be a birth death or anything else)
- a Brownian model is defined over the tree (just as described in the previous section)
- the Brownian model is conditioned on the quantitative trait data
- a substitution model is defined over the same tree
- the substitution model is conditioned on the molecular sequence data.

Instead of remaining fixed to a pre-defined value, the tree should now be moved during the MCMC. Ideally, we would like to move both the toplogy and the divergence times. Mixing over tree topologies under a Brownian model is relatively challenging, however (it works, but it requires rather long MCMC runs). For that reason, in the following, we will mix over divergence times only, under the constraint of a fixed tree topology. The features of the model that would need to be modified in order to also mix over topologies will nevertheless be indicated. You may want try them after the workshop.

Programming the model in RevBayes

Implementing this joint model in RevBayes is just a matter of adding the following features to the model defined in the previous section.

- load a sequence data matrix

  ```
  seqData <- readCharacterData("data/plac40_4fold.nex")
  nSites <- seqData.nchar()[1]
  ```

- create a substitution model, just like what you probably did in previous sessions. Here, we will make a simple GTR model, without any rate variation, neither among sites nor among branches:

  ```
  # equilibrium frequencies
  bf <- v(1,1,1,1)
  pi ~ dirichlet(bf)
  # exchangeabilities
  e <- v(1,1,1,1,1,1)
  er ~ dirichlet(e)
  # rate matrix
  Q := gtr(er,pi)
  # global substitution rate (strict clock)
  clockRate ~ exponential(0.1)
  # sequence evolution model
  seq ~ substModel(tree=tau, Q=Q, branchRates=clockRate, nSites=nSites, type="DNA")
  ```

  (note that the `tau` here refers to the same tree as the `tau` parameter of the Bronwian process)

- condition the model on the sequence data:

  ```
  seq.clamp(seqData)
  ```

- in the moves section, add moves for divergence times:

  ```
  moves[index] <- mSubtreeScale(tau, weight=5.0)
  index <- index + 1
  moves[index] <- mNodeTimeSlideUniform(tau, weight=10.0)
  index <- index + 1
  ```

- you would add topology moves here (again, only in a second step):

  ```
  moves[index] <- mNNI(tau, weight=5.0)
  index <- index + 1
  moves[index] <- mFNPR(tau, weight=5.0)
  index <- index + 1
  ```

Note that, if you do not include topology moves, it is essential that you set the tree equal to an externally-provided reasonable tree topology (e.g. the one defined in `chronoplac40.tree`). In fact, even if you move the tree topology, you may still want to speed up convergence of the MCMC by starting from this externally-provided tree (which is then not anymore a hard constraint, but merely a reasonable starting point).

Note also that the simple strict-clock and the site-homogeneous model considered here are not so good. Although integrating over uncertainty about the phylogeny when doing a comparative analysis is nice in principle, doing it using a poor model, as is done here, may ultimately be worse than using a fixed tree that has itself been separately estimated using a more reasonable model. But of course, nothing prevents us from making a joint analysis in RevBayes that relies on a more reasonable substitution model – in particular, using a relaxed clock.

4 Autocorrelated relaxed molecular clock

In the previous model, no consideration was given to the problem of rate variation among lineages – we bluntly used a strict clock. This is of course problematic, in particular at the phylogenetic scale considered here (placental mammals), where we know that there is substantial rate variation. In addition, we know that substitution rates across branches are *auto-correlated* in the present case: typically, entire orders, such as rodents, are fast evolving, whereas other orders like Cetartiodactyla are slowly evolving. In other words, nearby lineages along the phylogeny tend to be characterized by similar substitution rates.

You have perhaps already seen an autocorrelated relaxed clock model in the molecular dating session (ACLN). You could easily recruit it in the present context (a good exercise to try: modify the model suggested in the previous section so as to replace the strict clock by the ACLN model).

Here, however, we will derive the autocorrelated clock in a slightly different way. This derivation will be less straightforward, but more useful for what we want to do next. Fundamentally, the autocorrelated clock is a model where the logarithm of the instant substitution rate is itself a Brownian motion. Thus, it is exactly like a univariate quantitative trait, such as the logarithm of body mass (section 1). Next, we need to exponentiate this Brownian process and average over branches in order to obtain branch-specific substitution rates, which are finally plugged into the substitution model.

Programming the model in RevBayes

Compared to the model described in the last section, the only new feature is this: instead of defining a simple scalar `clockRate` variable, we now define a univariate Brownian motion along the tree, using the tools introduced in section 1:

- define `nu`, the "rate of evolution of the rate of substitution":

  `nu ~ Exponential(0.001)`

- define the univariate Brownian process describing the instant (log) substitution rate:

  `lograte ~ brownian(tau,nu)`

- exponentiate and average this Brownian process over branches using `expBranchTree`:

  `branchrates := expBranchTree(tree=tau, process=lograte)`

- plug these rates into the substModel object, as the branchRates parameter:

  `seq ~ substModel(tree=tau, Q=Q, branchRates=clockRate, nSites=nSites, type="DNA")`

- condition the model on the sequence and trait data and run the program.

Note that, since `lograte` is, technically, a quantitative trait, you can also reconstruct its evolution along the phylogeny. In particular, you can estimate the substitution rate in the last common ancestor of placental mammals or visualize rate variation over the entire tree using `njplot` or some other graphical display (just as for body mass).

Note also that, here, we do not have included any fossil information: we are merely doing *relative* dating. We will see at the end of this tutorial how fossil information can be integrated.

Exercises

- write the model
- run it on the placental dataset
- estimate the strength of the correlation between traits
- how does that compare with the strength estimated on a fixed topology?

5 Rates, dates and traits

We have just seen that the logarithm of the substitution rate can be seen as a quantitative trait. But then, this raises one further obvious question: why considering the substitution rate and the quantitative traits as separate Brownian motions? Why not instead considering them as a joint multivariate Brownian process? Doing so would have one major advantage: the correlated evolution of rates and traits will be automatically estimated, as a by-product of the model.

To do so, we just need to define a multivariate Brownian process of dimension 4. By convention, we will consider that the first dimension of this process corresponds to the log of the substitution rate, while the other 3 dimensions of the process (2 to 4) will map to the quantitative traits defined by the data matrix (Lartillot and Poujol, 2011).

Programming the model in RevBayes

You now have all the tools to implement this model entirely by yourself, except for one little detail: you now need to exponentiate one specific component of a multivariate process (as opposed to exponentiating a univariate process, as we did in the previous section). Thus, assuming that `X` is your 4-dimensional process, you need to tell the `expBranchTree` function that you want to exponentiate the first component of the process (with the `traitIndex=1` option):

```
branchrates := expBranchTree(tree=tau, process=x, traitIndex=1)
```

Also, be careful with the mapping of the quantitative traits: you need to map trait i to entry $i+1$ of the Brownian process:

```
for (i in 1:nTraits) { X.clampAt(contData,i+1,i) }
```

Exercises

- write the model and run it on the placental example
- investigate the correlation between substitution rate and life-history traits
- multiple regression: controlling for body mass, do you still get some support for a correlation between longevity and substitution rate variation?
- conversely, controlling for longevity, do you get supported correlations of the substitution rate and body mass (or with age at sexual maturity?)
- compare the credible interval obtained here for the body mass of the last common ancestor of placentals with what was obtained in the very first model (section 1).

6 A comparative analysis of variation in GC content

Apart from the overall substitution rate, any other aspect of the substitution process (transition-transversion ratio, dN/dS, equilibrium frequencies, etc) could in principle display variation among lineages. These various aspects of the substitution process could therefore be modeled exactly like the substitution rate, i.e. as Brownian processes – or as components of a multivariate Brownian process. In this section, we will focus on compositional variation, and more particularly variation in equilibirium GC content between species.

We first start with a simple T92 model of sequence evolution:

$$Q = \left(\begin{array}{c|cccc} & A & C & G & T \\ \hline A & - & \frac{\gamma}{2} & \kappa\,\frac{\gamma}{2} & \frac{1-\gamma}{2} \\ C & \frac{1-\gamma}{2} & - & \frac{\gamma}{2} & \kappa\,\frac{1-\gamma}{2} \\ G & \kappa\,\frac{1-\gamma}{2} & \frac{\gamma}{2} & - & \frac{1-\gamma}{2} \\ T & \frac{1-\gamma}{2} & \kappa\,\frac{\gamma}{2} & \frac{\gamma}{2} & - \end{array} \right)$$

The model has two parameters: the transition-transversion rate κ and the equilbrium GC content γ. In the following, we will assume that κ is constant across the tree (although unknown, and thus endowed with a diffuse prior). In contrast, γ will be allowed to vary among lineages, jointly with the overall substitution rate. Technically, since γ is strictly between 0 and 1, its log-it transform

$\ln \frac{\gamma}{1-\gamma}$ will range over the entire real line. Therefore, we could propose that the log-it transform of γ evolves according to a Brownian motion.

Putting everything together, we will therefore propose a multivariate Brownian motion $X(t)$, of dimension $K+2$, where K is the number of quantitative traits, and such that:

$$\begin{aligned} X_1(t) &= \ln r(t) \\ X_2(t) &= \ln \frac{\gamma(t)}{1-\gamma(t)} \\ k = 1..K, \quad X_{k+2}(t) &= \ln C_k(t) \end{aligned}$$

where $r(t)$ is the instant substitution rate and $\gamma(t)$ the instant equilibrium GC composition and $C_k(t)$ is the kth. quantitative trait. Equivalently, we may re-write this as follows:

$$\begin{aligned} r(t) &= e^{X_1(t)} \\ \gamma(t) &= \frac{e^{X_2(t)}}{1+e^{X_2(t)}} \\ &\dots \end{aligned}$$

In other words, the instant rate of substitution $r(t)$ is the exponential of the first component $X_1(t)$ of the Brownian process (as above), while the instant equilbirium GC $\gamma(t)$ is the *hyperbolic tangent* of the second component $X_2(t)$ of the Brownian process.

There is a slight complication here: in a non-homogeneous model, independently of the rate matrices across branches, we also need to specify the nucleotide frequencies from which the sequence at the root of the tree is sampled. We will call this frequency vector π, and we will put a Dirichlet prior on π.

This model has been described in Lartillot (2013).

Programming the model in RevBayes

Assuming that `X` is the multivariate Brownian process:

- as above, define the branch rates as the exponential of the first component:

  ```
  branchrates := expBranchTree(tree=tau, process=X, traitIndex=1)
  ```

- define the branch equilibrium GC as the hyperbolic tangent of the second component:

  ```
  branchGC := tanhBranchTree(tree=tau, process=X, traitIndex=2)
  ```

- for $k = 1..K$, map trait k onto entry $k+2$ of X:

  ```
  for (k in 1:nTraits) { X.clampAt(contData,k+2,k) }
  ```

- define the transition-transversion ratio; usually, this ratio is of the order of 1-10, so we will use an exponential prior of mean 10:

  ```
  tstv ~ exponential(0.1)
  ```

- define a vector of branch-specific T92 substitution matrices:
  `branchMatrices := t92GCBranchTree(tree=tau,branchGC=branchGC,tstv=tstv)`
- create a Dirichlet-distributed vector of equilbirium frequencies over nucleotides:
  `bf <- v(1,1,1,1)`
  `pi ∼ dirichlet(bf)`
- finally, create the substitution model:
  `seq ∼ substModel(tree=tau, Q=branchMatrices, rootFrequencies=pi,`
  `branchRates=branchrates, nSites=nSites, type="DNA")`

Exercises

- program the model in RevBayes
- run the model on the placental dataset
- investigate the correlation between GC and body mass
- how do you explain these correlations?
- run the model on the archaeal rRNA dataset `archaea.nex`, using temperature as the trait. In that case, no phylogeny is provided, so you may try to run the model without any constraint on the topology.
- assess the correlation between GC and temperature
- how much of the variation in GC is explained by temperature?
- what could be the underlying biological cause?

7 Towards integrative macro-evolution modeling

The modeling approach proposed above is just one example of the integration of multiple domains of macroevolutionary studies that could be done with RevBayes. In the following, we outline some possible extensions or variations, based on the integrative modeling philosophy.

Using fossil data

Fossils have much to say about several aspects of the models and questions we have considered thus far. They represent a valuable source of information about divergence times but also about ancestral traits. In particular, fossil calibrations could be used in the context of each of the models that have been considered in sections 3 to 6, thus allowing us to do not just relative, but absolute, dating. In principle, most of the approaches that you have seen during the dating session of the workshop could be adapted to the present situation. It is just a matter of gathering the relevant information about mammalian fossils.

More ambitiously, the *total evidence dating* method (Ronquist *et al.*, 2012) could be extended so as to now include, not just discrete morphological characters, but also quantitative traits, for both extant and extinct taxa. Morphological characters would be modeled using discrete M_k models, while quantitative traits would be described by multivariate Brownian processes, just as in the previous sections.

Beyond Brownian models

Throughout this tutorial, we have exclusively considered undirected Brownian models. However, many other models could be used, and this, both for quantitative traits and for substitution rates or substitution parameters. Right now, there are at least two other models available in RevBayes: the Brownian model with systematic trend and the Ornstein-Uhlenbeck process.

One possible application of the Brownian model with trend would be to test for the existence of a systematic trend in increasing body size (i.e. Cope's rule) during animal, vertebrate or mammalian evolution (Alroy, 1998). Note, however, that systematic trends cannot be estimated using only extant taxa (at least using purely anagenetic processes of evolution, such as considered here): the model would not be identifiable. If we have fossil data, on the other hand, we can estimate a trend: the model will then essentially rely on the average-mass-through-time distribution across the entire geological range.

Technically, to model body size evolution with drift, we would just need to:

- define a drift parameter, with a diffuse prior centered on 0:

  ```
  copestrend ~ norm(0,10)
  ```

- create a (univariate) Brownian motion with drift:

  ```
  logmass ~ brownian(tau,sigma,drift=copestrend)
  ```

- move the trend parameter during the MCMC, using a regular sliding move:

  ```
  moves[index] <- mvSlide(copestrend, delta=2.0, tune=true, weight=3.0)
  index <- index + 1
  ```

After running the model, the posterior distribution on Cope's trend parameter can be visualized and quantified, and the empirical support in favor of Cope's rule can be assessed by estimating the posterior probability that this trend parameter is positive.

References

Alroy, J. (1998). Cope's rule and the dynamics of body mass evolution in North American fossil mammals. *Science*, **280**(5364), 731–734.

de Magalhaes, J. and Costa, J. (2009). A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol*, **22**, 1770–1774.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am Nat*, **125**, 1–15.

Harvey, P. and Pagel, M. (1991). *The comparative method in evolutionary biology*.

Huelsenbeck, J. P. and Rannala, B. (2003). Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution*, **57**(6), 1237–1247.

Lartillot, N. (2013). Interaction between Selection and Biased Gene Conversion in Mammalian Protein-Coding Sequence Evolution Revealed by a Phylogenetic Covariance Analysis. *Mol Biol Evol*, **30**(2), 356–368.

Lartillot, N. and Delsuc, F. (2012). Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, **66**(6), 1773–1787.

Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol*, **28**(1), 729–744.

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

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., and Rasnitsyn, A. P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. *Syst Biol*, **61**(6), 973–999.