RevBayes Tutorial – Phylogenetic and comparative models

Nicolas Lartillot

July 28, 2014

1 Overview

1.1 The comparative method

The aim of the comparative method is to reconstruct the evolutionary history of various traits characterizing living organisms and to analyze the processes underlying their evolution at the macroevolutionary scale. Traits can be morphological, behavioral or have some connection with the life-history strategies of the organisms that bear them. They can be discrete (typically, presence of absence of particular morphological structures, number of vertebrae or segments, reproductive strategies, such as autogamy or allogamy) or continuous (body size, or life-history traits such as longevity or age at sexual maturity). 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.

1.2 The role of phylogenies in comparative analyses

In order to correctly formalize all these comparative questions, the underlying phylogeny should always be be explicitly accounted for. This point is clearly illustrated by the problem of establishing whether two traits (say, body mass and longevity in mammals) are correlated. As is well known, just plotting longevity versus body mass and fitting a regression line to this XY-plot will not be a valid approach to assess the significance and the strength of the correlation, since this would not account for the non-independence between data points induced by the evolutionarily relatedness of the species. Instead, one should explicitly model the joint evolution of the two traits of interest along the phylogeny and then estimate the strength of the correlation directly as a parameter of the evolutionary process itself.

Practically speaking, correctly formalizing a comparative question therefore requires to have access to the underlying phylogeny. Traditionally, this has been (and is still most often) done by first estimating the phylogeny and the divergence times using a separate phylogenetic reconstruction software. In a second step, this phylogeny is used as an input

to the comparative method. Doing this, however, raises a certain number of methodological problems.

- the uncertainty about phylogenetic reconstruction and, perhaps even more importantly, about divergence times, is ignored at the level of the comparative analysis. In some cases, this could result in artifactually significant or decepteively strong results.
- the traits themselves may have something to say about the phylogenetic tree or about divergence times: after all, what we call phylogenetic inertia in the context of the comparative method, and which is usually considered as a nuisance, is also a genuine source of information about the evolutionary relatedness of species. As such, it could be used as a primary source of information, along with genetic sequence data, for reconstructing the phylogeny.
- the phylogenetic reconstruction software will often typically assume that the rate of substitution is not constant across the phylogeny (the so-called relaxed molecular clock). The rate of substitution, and more generally the parameters of the substitution process, could therefore also be seen as quantitative traits of sorts. As such, they could also be subjected to a comparative analysis, just as what would have been done with body-size or longevity.

1.3 An integrative comparative approach

All these considerations suggest that, ultimately, phylogenetic reconstruction, molecular dating and the comparative method should all be considered jointly, in the context of one single overarching probabilistic model. The modular nature of RevBayes provides 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. This can also be considered as one first example of the more general perspective of integrative modeling, which can be recruited in many other contexts.

2 Practical exercises on the classical comparative method

2.1 Univariate models: Brownian versus OU evolution of traits

Possible exercises (all this, using fixed, externally provided, time-calibrated phylogenies)

- fitting a simple univariate Brownian model to body mass evolution in some mammalian group; estimate ancestral state, with credible interval. Recover a set of trees from the posterior distribution, with trait values attached to nodes in newick format. Use this to output a color-coded reconstruction of the evolutionary history of the

trait (this requires a special string-output, implemented via a method in RealNodeValContainer). Emphasize uncertainty in ancestral inference.

- trends: put a uniform prior on the trend of the Brownian motion, over a short interval symmetric around 0: show that the posterior distribution is also uniform. Obviously, the model is not identifiable.
- Now use a phylogeny with fossils as tips. In that case, it becomes possible to estimate a trend: thanks to the data-through-time structure. Draw posterior marginal distribution on trend parameter: show that the posterior probability density at 0 can be used to indirectly derive a Bayes factor (Dickey-Savage ratio). More simply, the fact that the posterior distibution is (or is not) concentrated on one single side of the origin is in itself the most straightforward qualitative criterion to assess the empirical support in favor of the presence of a systematic trend.
- fit an OU process: try to find an interesting case where the OU process is indeed supported, and where this makes sense, in terms of the underlying macro-evolutionary process (e.g. that could be interpreted in terms of stabilizing selection).
- in the case of body mass in mammals, the OU process will have a very low rate of return of equilibrium ϕ. Typically, $\phi T \sim 0.1$, where T is the depth of the phylogeny. This means that the mean relaxation time of the process is of the order of $10T$: thus, for all practical means, the process of body-size evolution is effectively Brownian. Present this criterion (compare relaxation time of the process with the depth of the phylogeny) as an indirect method for comparing Brownian and OUP – in fact, more insightful than Bayes factors.

2.2 Multivariate models: correlated evolution of quantitative traits

Exercises (still using a fixed pre-defined phylogeny)

- take the 73-taxon placental dataset, with 3 life-history traits from AnAge (body mass, longevity and sexual maturity).
- fit a Brownian motion of dimension 3
- estimate covariance matrix. Identify diagonal elements with the variance parameter of the one-dimensional Brownian motion fitted previously.
- output posterior distribution on: covariance parameters, correlation coefficients
- output partial correlation coefficient between, say, maturity and longevity when controlling for body size.
- *all this requires outputting the inverse of the covariance matrix, as well as its separate entries, all this on demand (thus, some functions still to be defined)*

2.3 Further directions

Emphasize that what is currently available in revbayes is but a small part of what can be done in the context of the comparative method. In this respect, mention the already existing software programs: BayesTraits, Ape, everything currently existing in the R environment. Also briefly mention the connections with GLS and maximum likelihood.

However, the main advantage of RevBayes is that we can now integrate the comparative method with the phylogenetic models, as will now be done.

3 Accounting for uncertainty in phylogeny and divergence times

The models implemented in the last section are structured as follows:

- a data set of quantitative traits is loaded
- a phylogenetic tree is declared, whose tips correspond to the taxa of the data set
- the phylogenetic tree is clamped to an externally given tree topology with divergence times specified
- a comparative model is declared and conditioned on the quantitative trait data
- running the model gives posterior distributions over the parameters
- marginalization on the parameters of interest (covariance or correlation coefficient) essentially answers to our original question (are traits correlated)

Starting from this model, how could we account for phylogenetic uncertainty? One possibility would be as follows. Typically, the tree given as an input is a consensus tree obtained by running a Bayesian software program (like Beast or RevBayes) on another data set, made of aligned genetic sequences for the same set of taxa. Thus, instead of running the comparative analysis only on the consensus tree, one could instead run it on a series of, say, 100 or 1000 trees sampled from the posterior distribution (after burn-in). Then, the resulting estimates could be pooled and averaged. However, doing this is a bit problematic.

Another possibility is to make a hierarchical model where

- two data sets are loaded: one for sequence data and one for quantitative traits
- A tree is declared, with a prior (birth death or uniform)
- a substitution model is declared, parameterized by the tree
- the substitution model is clamped on the molecular data.

- a comparative model is declared, parameterized by the same tree
- the comparative model is clamped on the quantitative trait data
- the model is run

Here, the important conceptual difference with the previous analysis the tree will be inferred using both sources of empirical data. In practice, this is not likely to make such a big difference (usually, sequence data have much more signal than quantitative traits for informing the phylogeny). But this is conceptually more elegant and, more importantly, will be the basis for further modeling developments.

Would be useful to draw some graphical model representations of the various models alluded to in this section: essentially, how to merge together two disconnected graphical models, one representing the comparative method and another one representing the phylogenetic analysis into one single connected graph.

4 Autocorrelated relaxed molecular clock

Should make a more explicit connection with the tutorial on molecular dating, in which tools were already presented for conducting a robust dating analysis.

In the previous model, no consideration was made for the problem of rate variation among lineages. This is of course problematic, in particular at the phylogenetic scale considered here (mammals), where we know that there is substantial rate variation.

In addition, we know that substitution rates 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, which they tend to belong to the same order, also tend to be characterized by similar substitution rates.

The auto-correlated clock is fundamentally a model where the log of the instant substitution rate is seen as a Brownian motion. Thus, it is exactly like a univariate quantitative trait, such as those that we have modeled in the first section of this tutorial. Since the Brownian motion describes the evolution of the log of the rate, we need to exponentiate this brownian process in order to obtain substitution rates, which can then be plugged into the substitution model.

- load sequence data
- declare the tree
- define a univariate Brownian motion along this tree, using the same tools as for the univariate comparative analysis

- exponentiate the Brownian process, using exponentialBranchTree. This will create a deterministic function of this brownian motion, summarized through branchwise averages.
- plug these rates into the substModel object, as the branchRates parameter
- condition the model and run the program.

You may compare this analysis with the kind of analyses that were done earlier in the context of the molecular dating session. Could also be interesting to reconstruct the evolution of substitution rate as a trait along the phylogeny, and visualize it as a color-coded tree: just to emphasize the conceptual similarity between substitution rate and quantitative traits.

Note that, here, we do not have included any fossil information: we are merely doing *relative* dating. This is not ideal, but we will see at the end of this tutorial how all this can be integrated with fossil information. In fact, this represents one of the most exciting frontiers of the present integrative approach.

5 Rates and traits

5.1 Merging the clock and the comparative analysis

Once you have been able to construct this simple auto-correlated relaxed clock model, and based on what was done in the previous section, it should not be too difficult to run a joint dating and comparative analysis. Thus, you would essentially declare two Brownian motions along the same tree: one univariate motion for the relaxed clock, and one multiariate Brownian motion for the quantitative traits.

Once this is done, we can now ask one further obvious question: why considering the substitution rate and the quantitative traits as separate Brownian motions? Why not considering them as a joint multivariate motion? 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.

6 A comparative analysis of the variation in equilibrium GC content

Time-heterogeneous models: have they been introduced at some point in the previous tutorials?

Some words here about the prevalence of compositional variation, and the possible causes. Discussing the alternative between (1) modeling variation in GC using iid GC

parameters branchwise, and (2) modeling this variation by assuming an explicit time-continuous process for GC evolution along the lineages. The latter is more principled, and one should thoroughly discuss why.

Some exercises:

- mammalian dataset (add the karyotypic quantitative traits here): correlation between GC and recombination rate and body size. The biased gene conversion hypothesis
- in the case of mammals, possible to do some multiple regression analysis on gc, rate and traits.
- archaeal rRNA GC content and temperature.

7 Towards integrative macro-evolution modeling

The integration proposed here is just one example of the integration of multiple domains of macroevolutionary studies that could be done with revbayes.

7.1 Rates and traits with fossil calibrations

Make a calibrated analysis: either using "focal dating", based on Tracy's method. The 5 to 7 few fossil calibrations typically used for mammals could easily be recruited in this context.

7.2 Total evidence

Alternatively, undertake a more ambitious and more extensive total evidence approach. The obvious problem here is that the approach requires a good fossil sampling, with a good matrix of both discrete and continuous traits. Getting those data is probably a fair amount of work in itself. But the outcome could be very exciting: in particular, adding body size information about fossils!

7.3 Testing diversification models

make an integrated diversification studies and comparative analysis: combining phylogenetic estimation, dating, reconstruction of body size evolution and test of a diversification model in the case of placentals: in their case, a good model to test would be a skyline birth death or, perhaps more interestingly, a burst and stasis model, with the burst occurring right at KT.

7.4 Comparative analysis and incomplete lineage sorting

Let the generation time t, the mutation rate per generation u and the effective population size N_e be a joint multivariate (log-normal) Brownian process along the species tree. Then, plug $(N_e t)^{-1}$ as the coalescence rate within the ILS model, and u/t as the substitution rate per calendar unit of time in the substitution model running along gene genealogies. Of course, all this can be correlated with body size and life-history traits. Predictions are: N_e correlates negatively, t positively and u positively (but u/t negatively) with body-size. Nice allometric analysis of ILS...

One nasty thing here is that we need to pre-define classes similar to ExponentialBranchTree, but that would implement the functional relations introduced above between N_e, t and u: since, for the moment, we do not have the kind of templated containers, combined with user-defined functions, that would allow users to do that automatically...

8 Some notes on the overall modeling strategy used here

- emphasize that the model is still approximate here: in particular, in the way we take the average over branches. In this respect, mention recent developments in fine-grained models of Brownian evolution of substitution rates
- ultimately, no need to restrict oneself to Brownian motions: more general processes could be imagined.