5.7 Comparing the Statistical Fit of Different Evolutionary Models in R

Comparing the Statistical Fit of Different Evolutionary Models in R

The 'geiger' package in R allows one to test many of the evolutionary models for continuous character evolution that we have discussed so far. First, download R and install the geiger package along with its dependencies. Next, download the tree for this exercise and a data table of female body weights for species in this tree, which consists of the guenons and papionins with a few colobines for outgroups.

First, load the 'geiger' package:

```
library("geiger")
```

Next, load the phylogenetic tree and the data:

```
tree_primates = read.nexus("57.Tree.txt")
```

```
data_primates = read.table("57.MassData.txt",header=T,sep="\t")
```

We are now ready to estimate the likelihood of the data on this tree given a variety of evolutionary models. For example, we can test for a directional trend in body size by comparing the likelihood of the "trend" model to the likelihood of the "BM" (Brownian motion) model. Brownian motion is the simplest (fewest parameter) model of trait evolution in which variance in a continuous trait accumulates at a constant rate in random directions (addition to or subtraction from the trait mean). Because of this constant variance assumption, it is common to logarithmically transform body size data so that proportional rather than absolute changes in size are modeled on the tree. You can very easily transform your data in R as follows (note that lines starting with # are ignored by R).

```
#convert kilograms to grams
data_primates$Mass_gm = data_primates$Mass_kg*1000
#take the natural logarithm of grams
data_primates$Log_Mass_gm = log(data_primates$Mass_gm)
```

Take a look at the two new columns of data you added to your dataframe:

```
data_primates
hist(data_primates$Log_Mass_gm)
```

You can see that the structure of the data table is preserved but the weights are all now the natural logarithm of grams. Now calculate the likelihood of the data under Brownian motion by typing:

```
brownian = fitContinuous(tree_primates, data_primates$Log_Mass_gm,model="BM")
```

Then call "brownian". R will print the following

```
$Trait1
$Trait1$lnl
[1] -30.98123
$Trait1$beta
[1] 0.02371125
$Trait1$aic
[1] 65.96246
$Trait1$aicc
[1] 66.27825
$Trait1$k
[1] 2
```

R is first saying these are the values from the table for trait 1 (there is only one trait in this case). After the second $ sign is says which parameter or result is printed on the next line. So, the "$Trait1$lnl" means the next line shows the log likelihood of the data under Brownian motion, which is -30.98. We can calculate the comparable likelihood under model of Brownian motion with a directional trend by typing:

```
trend = fitContinuous(tree_primates, data_primates$Log_Mass_gm,model="trend")
```

If you type "trend," R will return to the command line:

```
$Trait1
$Trait1$lnl
[1] -30.86872
$Trait1$mean
[1] 9.867361
$Trait1$beta
[1] 0.02358446
$Trait1$mu
[1] -0.05444671
$Trait1$aic
[1] 67.73744
$Trait1$aicc
[1] 68.38609
$Trait1$k
[1] 3
```

Here R gives the likelihood of the data under the trend model is -30.87, the mean log body size is 9.87, and the directional trend is for lineages to evolve larger body sizes over time (mu = 0.024). Remember, though, that more complex models will always have higher (less negative) likelihoods than simpler models like Brownian motion. We will use the likelihood ratio test to decide if the additional parameter for a trend (mu) is justified for this data set. This is easily done in R by calculating the probability of the chi square value under one degree of freedom, where the chi square is 2 times the difference in log likelihoods. So, the likelihood ratio statistic is 2(-30.87-(-30.98))=0.22. We can find the chance probability of this degree of improvement in likelihood from adding one parameter by referring to the chi square distribution. Type:

```
pchisq(0.22,1,lower.tail=F)
```

R will print

```
[1] 0.6390399
```

So, the p-value associated with this difference in log likelihoods is not significant, and we do not have evidence that a Brownian model of evolution should be rejected in favor of one with a directional trend.

Let us look at another important model of trait change - Pagel's lambda value for phylogenetic signal. Lambda is a multiplier of the internal branches of a phylogeny. These internal branches directly express the degree of expected correlation among the traits of species in the study. Thus, when lambda = 1, the model is equivalent to Brownian change in the traits values on the untransformed branch lengths. When lambda is greater than zero but less than one, phylogenetic signal is present, but there is less of a phylogenetic effect than we would expect from the original branch lengths. You can calculate the likelihood of the data under the lambda transformation by typing:

```
lambda = fitContinuous(tree_primates, data_primates$Log_Mass_gm,model="lambda")
```

When you call "lambda" you will see that the likelihood is reported at -30.42. When you calculate the difference between this value and the one for Brownian motion 2*(-30.42-(-30.98)) = 1.12, and apply the "pchisq" to calculate the probability under the chi square distibution, the reported p-value is 0.29. So, the lambda transform in this case does not make the data significantly more likely under as compared to the simpler Brownian motion model on the untransformed branch lengths. When you called "lambdatest", R will have printed among the results the following:

```
$Trait1$lambda
[1] 0.811984
```

The estimate of lambda that makes the data most likely is 0.81. If applying the additional lambda parameter had significantly improved the model, then a value equal to 0.8 essentially says that there is less covariance due to ancestry than the untransformed branch lengths imply under Brownian motion (lambda = 0 is a star phylogeny with no phylogenetic signal). Although the maximum likelihood estimate of lambda is not 0 in this case, the small improvement of likelihood does not justify the additional lambda parameter.

Another popular model is the Ornstein-Uhlenbeck model, also known as an OU model. The simple Brownian motion model assumes that the trait can continue to get larger or smaller indefinitely; that is, it assumes variance can continue to accumulate no matter how small or large the trait value becomes. This is probably unrealistic for many biological traits, like body mass, that have a necessary minimum value of 0 and likely an upper physical constraint as well. These lower and upper 'walls' stop variance from continuing to accumulate in one direction at extreme values. The OU model attempts to reflect this reality by fitting a central tendency for the trait, and a parameter that creates a trend for variance to accumulate toward this central tendency. This model is often described as Brownian motion, but with a rubber band that tethers the traits values to a particular point. Like a rubber band, the further values get from this central tendency, the greater the pull back to the center.

You can fit the OU model by typing:

```
OU = fitContinuous(tree_primates, data_primates$Log_Mass_gm,model="OU")
```

Type "OU" and you will see the likelihood for this model is -30.7. Although the OU model may seem more realistic, it does not significantly improve the likelihood of the data (its improvement in the likelihood is less than for the lambda parameter). Essentially, the body mass data is such that the very simple Brownian motion model adequately describes how change accumulated in this character on the phylogeny.

Several other evolutionary models and branch transformations can be applied through geiger. For example, you could fit an early burst model (also called the ACDC model) in which trait evolution is rapid early in a radiation and slower later on by setting "model="EB"". If you try this one you will see that it does not significantly improve the likelihood.

'Geiger' also outputs AIC values for all models. You can also use those to assess significance instead of the log likelihoods. This is typically done when comparing models that are not nested. Models are nested when the simpler of two models (the one with fewer parameters) can be derived from the more complex one simply by eliminating some parameters. In such a case, the simpler model is 'nested' within the more complex one, because it is just a special case of the complex one where some parameters are set to 0. Models are not nested when the simpler model has parameters that do not appear in the more complex model. The likelihood ratio test is not valid for comparing likelihood scores from non-nested models. It is this situation in which people use AIC as one criteria for model selection. AIC is essentially the likelihood after it has been penalized for the number of parameters in a model. For example, the OU and lambda models have the same number of parameters, but different parameters.

References

Harmon, L., J. Weir, C. Brock, R. Glor, W. Challenger, and G. Hunt (2009). geiger: Analysis of evolutionary diversification. R package version 1.3-1. http://CRAN.R-project.org/package=geiger

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Contributed by Luke Matthews and Charlie Nunn