User’s Manual for

Continuous

(copyright M. Pagel)

Mark Pagel
School of Animal and Microbial Sciences
University of Reading
Reading RG6 6AJ
UK
email: m.pagel@rdg.ac.uk
(www.ams.reading.ac.uk/zoology/pagel/)

Continuous

Continuous is a computer program that implements a generalised least squares (GLS) model for the across-species analysis of comparative data. The method is described in two papers (Pagel, 1997, 1999). I would be grateful if users would cite these papers, as appropriate, when they use the method. The application program can be used to

- test for correlated evolution between pairs of characters,
- to find ancestral states,
- to examine random-walk versus directional change models,
- to investigate the tempo and mode of trait evolution,
- and to assess whether and to what extent the a phylogenetic correction is required in the data.

This manual describes how to use **Continuous** and its hypothesis testing capabilities. The manual is intended to provide enough of an introduction to use the program, although it does not attempt to describe all of the things one might use the program for. Some features of Continuous are not yet implemented in the program even though they appear in the menus. The computer interface was designed and programmed by Dr Andrew Rambaut.

The GLS approach makes it possible to analyse and display comparative data across-species, without the need to calculate independent contrasts or other phylogenetic corrections (independent contrasts turn out to be a special case of the GLS approach – see below). In the GLS approach, non-independence among the species is controlled for internally by reference to a matrix of the expected covariances among species. This means that data can be plotted across species and interpreted using the correlations and regressions obtained from **Continuous**. The GLS approach as implemented in **Continuous** also makes it possible to transform and scale the phylogeny to test the adequacy of the underlying model of evolution, to assess whether phylogenetic correction of the data is required, and to test hypotheses about trait evolution itself – for example, is trait evolution punctuational or gradual, is there evidence for adaptive radiation, is the rate of evolution constant.

I. Running Continuous

This version of the program runs on a PowerPc Macintosh and prefers System 8 or later.

II. Data Input Format

Continuous requires a well resolved phylogeny and data on species. It uses its own input format (‘pag-format’) to describe the phylogenetic tree and the data. The attraction of the tree format is its transparency, although it is not as compact as Phylip (Newick) or Nexus formats. We hope to make available software for converting among these three formats.

The input format is simple. Consider the tree below of four species, and three internal nodes, 1, 2, and 3 where node 3 is also the root. Trees must be rooted.

```
 s1         s2       s3      s4
|          |        |       |
|t1        |t2      |       |
|__________|        |       |
      |1            | t3    |
      |t5           |       |t4
      |_____________|       |
              |2            |
              |t6           |
              |_____________|
                      |3
```

The ‘pag’ input format for this tree is:

#PAG
4 3
s1, 1, t1, data1,data2,data3…
s2,1, t2, data1, data2, data3…
s3, 2, t3, data1, data2, data3…
s4,3, t4,data1, data2, data3,…
1, 2, t5
2, 3, t6

The second line ‘4 3’ is similar to Phylip format and describes first the number of species, followed by the number of variables. The variables are the comparative data measured across species and should be kept to about ten or fewer, although this is not a strict requirement.

Read the third line of the input file as "species 1 goes to node "1" over ‘time’ or length t1 with data 1, 2, 3...; the fourth line as species 2 goes to node 1 over time or length t2, with data 1, 2, 3....", and so on. Data points must be real numbers. Missing data are coded as “*”.

Beginning with the seventh line of the file, the connections among the internal nodes are described. , "node 1 goes to node 2 over time or length 4, and node 2 goes to node 3 over time or length t6. Nodes do not have data, and the branch lengths can be any real number. Branch lengths can be any units but units of time and genetic distance (operational time) are especially useful.

Only tips (species) have data, and the tree should be bifurcating. If you have so-called polytomies in your tree, it is best to resolve them to bifurcations, although the program will handle some polytomies. Trees must be rooted, although the root itself is not described but inferred by the program from the input format.

The species names can be any alphanumeric character but should start with a letter. Internal nodes must be integers and must go in ascending order from the tips to the root. Thus, a descendant node must always have a smaller number than its ancestral node.

It may often be easiest to number the species from left to right beginning with species s1 to species s*n*. Then label the nodes as n+1, n+2 and so on until you reach the root. Every node in a bifurcating tree must have two and only two descendants. The root does not "go to" any other node as is not described any further. Items are separated by commas.

At the moment, **Continuous** does not give useful information when it tries to read in a treefile that is wrong in some respect. The most common errors are failing to separate items by commas, or inserting more than one comma, having too few or too many descendants of a node, specifying the wrong descendant, or incorrectly specifying the number of species or variables. The best way to de-bug a treefile that is not working is to print it out and compare it line by line to a picture of the phylogeny.

An example data file is included in the Appendix.

III. Loading and Manipulating the Data

If there is no open window when **Continuous** is started up select **New** from the **File** menu. The treefile is loaded into **Continuous** via the **Import** command in the **File** menu. At present this is the only way to load data into **Continuous**. Input files must be text files in the format as described above. If the input file has been saved by a word processor such as Microsoft Word be sure to choose the text only option. Sometimes word processors can insert invisible characters that can interfere with the input file being read in. If you suspect this, open the file in a word processor and save it as some other file choosing a text only format.

If the treefile loads properly, a picture of the phylogeny will be represented in the open window. The traits are listed on the left of the open window. Traits with checks by them are those that will be analysed. Clicking on a trait holding down the option key selects or de-selects it from analysis.

Double clicking on a trait brings up a box that allows you to view the data and to transform it logarithmically. This box can also be selected from the **Data** menu.

The **Show Data Matrix** option in the **Data** menu shows all of the species and their data points. Unselected traits are in grey. Species with missing data are in grey. Double clicking on a species removes it from any of the analyses that follow. Clicking on it again re-inserts it.

Options in the **Tree** menu are not implemeted.

Additional windows can be opened simultaneously by selecting the **New** command, followed by **Import**.

IV. Analysing Data

Continuous can be used to characterise and test hypotheses about the evolution of single traits, to find ancestral states and to test for evidence of correlations among traits. All phylogenetic correction of the data is done automatically. All parameters are found as maximum likelihood estimators. Some are found analytically, others must be found by searching a likelihood surface for the value that maximises the likelihood of observing the data given the value of the parameter, the phylogenetic tree, and the model of evolution. All likelihoods are expressed as log-likelihoods.

The **Continuous Analysis** menu provides a number of ways to investigate trait evolution and correlations. The next section describes some of the options. How to test these options for significance is described in a later section.

Set Model command

Models of evolution: Two models of evolution are available. Model A corresponds to the standard constant-variance random walk model. It has a single parameter, the (instantaneous) variance of evolution. Model B is a directional random-walk. This model has two parameters, the variance of evolution parameter as in Model A, plus the directional change parameter. This parameter effectively measures the regression of trait values across species against total path length from the root of the tree to the tips. It detects any general trends towards a dominant direction of evolutionary change (have species got bigger, smaller, faster, longer, and so on). The estimates of the parameters for Model A and Model B from a given set of data can be viewed by selecting the **Parameters** option (described below)

Model A and B can be compared via their log-likelihoods. If Model B fits the data better (larger log-likelihood – closer to zero) than Model A, it says that a directional trend exists. This is important in its own right and has implications for other estimators. For example, estimates of ancestral states at the root of the tree will differ greatly between Model A and Model B when Model B is significant, and the Model B estimates are to be preferred. **N.B**. : Model B can only be fitted to trees that have some variation in the total path length from the root to species. Model B cannot be used with ultrametric trees.

Transformations. A useful feature of Continuous are the three scaling parameters that can be estimated for a given data set and phylogeny. They are shown in the Set Model box where it is possible to fix them to a specified value, leave them at their default values, or estimate them by maximum likelihood.

The parameters are denoted kappa (κ), lambda (λ), and delta (δ). These scaling parameters allow tests of the tempo, mode, and phylogenetic associations of trait evolution. All three take the value 1.0 by default. These values correspond to assuming that the phylogeny and its branch lengths accurately describe the constant-variance random walk model A or B. However, if trait evolution has not followed the topology or the branch lengths, these values will depart from 1.0. When they do, incorporating them into the analysis of the data (e.g., when estimating the correlation between two traits) significantly improves the fit of the data to the model.

The **kappa** parameter differentially stretches or compresses individual phylogenetic branch lengths and can be used to test for a punctuational versus gradual mode of trait evolution. Kappa > 1.0 stretches long branches more than shorter ones, indicating that longer branches contribute more to trait evolution (as if the rate of evolution accelerates within a long branch). Kappa < 1.0 compresses longer branches more than shorter ones. In the extreme of Kappa = 0.0, trait evolution is independent of the length of the branch. Kappa = 0.0 is consistent with a punctuational mode of evolution.

The parameter **delta** scales overall path lengths in the phylogeny – the distance from the root to the species, as well as the shared path lengths. It can detect whether the rate of trait evolution has accelerated or slowed over time as one moves from the root to the tips, and can find evidence for adaptive radiations. If the estimate of Delta < 1.0, this says that shorter paths (earlier evolution in the phylogeny) contribute disproportionately to trait evolution – this is the signature of an adaptive radiation: rapid early evolution followed by slower rates of change among closely related species. Delta > 1.0 indicates that longer paths contribute more to trait evolution. This is the signature of accelerating evolution as time progresses. Seen this way, delta is a parameter that detects differential rates of evolution over time and re-scales the phylogeny to a basis in which the rate of evolution is constant.

The parameter **lambda** reveals whether the phylogeny correctly predicts the patterns of covariance among species on a given trait. This important parameter in effect indicates whether one of the key assumptions underlying the use of comparative methods – that species are not independent – is true for a given phylogeny and trait. If a trait is in fact evolving among species as if they were independent, this parameter will take the value 0.0 and indicate that phylogenetic correction can be dispensed with. A lambda value of 0.0 corresponds to the tree being represented as a star or big-bang phylogeny. If traits are evolving as expected given the tree topology and the random walk model, lambda takes the value of 1.0. Values of lambda = 1.0 are consistent with the constant-variance model (sometimes called Brownian motion) being a correct representation of the data. Intermediate values of lambda arise when the tree topology over-estimates the covariance among species.

The value of lambda can differ for different traits on the same phylogeny. If the goal is to estimate the correlation between two traits then lambda should be estimated while simultaneously estimating the correlation. If, on the other hand, the goal is to characterise traits individually, a separate lambda can be estimated for each.

Continuous will automatically incorporate the maximum likelihood values of the three scaling parameters into its calculations, or they can be fixed to predetermined values using the dialog boxes in the **Set Model** box. A maximum of any two parameters can be simultaneously estimated, although this can be time consuming because the likelihood surface is large.

Three scaling parameters and their interpretation when applied to trait evolution on a phylogeny

Parameter	Action	0	<1	1	>1
λ (lambda)	Assess contribution of phylogeny	star phylogeny (species independent)	phylogenetic history has minimal effect	default phylogeny	not defined
κ (kappa)	Scale branch lengths in tree	punctuational evolution	stasis in longer branches	default gradualism	longer branches more change
δ (delta)	Scale total path (root to tip) in tree	not defined	temporally early change important (adaptive radiation)	default gradualism	temporally later change (species-specific adaptation)

Set trait covariances to zero. When checked, this box forces the covariances among pairs of traits to be zero when calculating the likelihood. It is used to test whether the correlation between two traits is greater than zero. The test is described below.

Show Tree Matrix

This command displays a box with a number of options. Most are repeated in the main menu.

The **Traits** option displays the raw data on species.

The **Tree** option displays the tree matrix which is proportional to the default variance-covariance matrix among species as implied by the phylogeny. Elements of this matrix are the sum of the branch lengths on the main diagonal entries (corresponding to a species) and the sum of shared branch lengths on the off-diagonal entries (corresponding to the co-variance between two species). The values in this matrix are proportional to the variances and covariances. It is this matrix that delta acts on. Its effects can be seen by setting delta to different values and viewing the resulting tree matrices.

The **Var-Covar** option shows the calculated variances of the traits and the covariances between traits as estimated from the model of evolution that has been assumed (Model A or B, and assumed or estimated values for kappa, lambda, and delta). If the *set trait covariances to zero* box has been checked the covariances will be zero.

The **Correlations** option displays the correlations among traits, subject to the same model options as described for the **Var-Covar** option.

The **Parameters** option shows the values of up to two parameters and their standard errors. The **alpha** parameter is the estimated root of the tree. The **beta** parameter is the directional change parameter of Model B. It can be interpreted as a regression coefficient of the trait values against total path length. The combination of alpha and beta can be used to predict the values of the trait for any given total path length according to alpha + beta * path length.

V. Hypothesis Testing

Hypotheses are tested using the likelihood ratio statistic. The likelihood ratio statistic compares the log-likelihood of the null hypothesis model to that of the alternative hypothesis model. **Continuous** automatically calculates the log-likelihood of whatever model is chosen in the **Set Model** box, and displays this likelihood in the text window.

The likelihood ratio (LR) test compares the goodness of fit of a model to the data with that of a simpler model that lacks one or more of the parameters. The LR statistic is then defined as

$$LR = -2\log_e\left[H_0/H_1\right],$$

where H_0 represents the simpler (null) model and H_1 the (alternative) model containing the parameters representing the evolutionary processes one wishes to estimate. If the simpler model is a special case of the more complicated one, the LR statistic is asymptotically distributed as a chi-squared variate with degrees of freedom equal to the difference in the number of parameters between the two models, i.e., $LR \sim \chi^2(v)$, where v is the number of degrees of freedom. One test is a special case of another if it is possible to collapse the more complicated model to the simpler model by setting some parameters to zero or to other fixed values. For example, the model in which a correlation is estimated collapses to the null hypothesis model by forcing the correlation to be zero: the zero-correlation or null model is a special case of the alternative hypothesis model. In such circumstances the two models are often referred to as being ‘nested’.

If two different models are chosen in the **Set Models** box in succession their log-likelihoods can be compared by hand to perform a LR test. Alternatively, the **Likelihood Ratio Tests** option in the **Continuous Analysis** menu allows one to define two different models simultaneously and **Continuous** then fits the models and automatically performs a likelihood ratio test on them. A p-value is displayed from the appropriate chi-squared distribution on the assumption that the tests are nested (i.e., subsets of each other). Continuous is reasonably good at recognising tests that are not nested, but may not be infallable.

The **Likelihood Ratio Tests** option displays separate Set Model boxes for the null and alternative hypotheses. By choosing the models appropriately, one can test a range of hypotheses. The list that follows gives an idea about how to perform some general hypothesis tests in the LR framework.

Example Likelihood Ratio tests (Pagel, 1999 gives examples of several of the tests listed below)

Correlations and Regressions: compare a model (A or B) in which the *set trait covariances* to zero box is checked for the Null hypothesis model but not for the Alternative hypothesis model (this is a nested test). This tests the hypothesis of ρ=0.0. The correlation can be found from the **Continuous Analysis** menu. The regression coefficient can be found by calculating the ratio of the covariance between the two traits to the variance of the X-axis character. These values are in the **Var-Covar** display. The regression line can be drawn on a plot of the two variables using the calculated regression coefficient and the formula $y_{\text{int}} = \bar{y} - b\bar{x}$, where “y-bar” and “x-bar” are the means in the raw data, y_{int} is the y-axis intercept, and “b” is the regression coefficient from Continuous. If the correlation is significant so is the regression coefficient and they have identical p-values.

Drift versus Directional Models of Evolution. Compare Model A to Model B using the LR test. If done for a single trait, this test shows whether there is a dominant direction of change to a trait. If Model B is significant the maximum likelihood estimate of the root node can be reconstructed to lie outside of the range of values observed in the data (see Pagel, 1999 for an example). Random-walk or Model A methods always reconstruct the ancestor within this range.

Model B can be used when estimating the correlation between two traits. If Model B is significant, then the correlation between the traits will represent that variation in the two traits that is independent of their respective relationships with path length. In effect, the correlation coefficient estimated under Model B is a correlation of residuals from the line relating each trait to time or distance from the root.

Constant-variance (so-called Brownian motion) model: The adequacy of this model as a descriptor of the data can be tested by asking whether all three scaling parameters are either 1.0 or not significantly different from 1.0 as their ML values.

Molecular clock. If the sampling times are available for the tips of the tree, and the branch lengths are in units of genetic distance, the comparison of Model B versus Model A tests whether time and genetic distance are linearly related. A significant Model B does not imply a strict clock, but if kappa or delta do not differ from 1.0, this interpretation is strengthened. The Model B directional coefficient gives the realtonship of time to genetic distance. Its reciprocal is the rate of evolution (see Pagel, 1999 for an application to Influenza evolution).

Influence of Phylogeny: Compare a model in which lambda is set to 1.0 with one in which lambda is allowed to take its maximum likelihood (ML) value. If the LR test is significant it says that lambda < 1.0. Alternatively, test a model in which lambda = 0.0 to one in which it takes its ML value. This tests whether lambda > 0.0. If it is significant it says that some sort ofphylogenetic correction is required.

Punctuational and Gradual Trait Evolution. Perform a LR test of kappa = 0.0 (null) to kappa = ML value. If kappa is not significantly different from 0.0, then trait evolution is consistent with a punctuational mode of change. Kappa > 0.0 implies some form of gradualism. Test whether kappa < 1.0 to see if default or 'scaled' gradualism is better supported.

Adaptive Radiation versus Species Adaptation. If delta is < 1.0 on a LR test, adaptive radiation is suggested. Delta > 1.0 suggests that later or species-specific adaptation has been dominant.

Other Analysis Issues

Species Regressions. A species regression can be approximated (i.e, result is very nearly equivalent) by setting lambda to 0.0 in the **Set Model** box. Setting lambda to 0.0 and delta to 0.001 (approximate) gives a result virtually indistinguishable from a species regression.

Independent Contrasts. Model A with no scaling transformations returns results equivalent to those obtained from independent contrasts analyses. The estimate of **alpha** (root) in this cased is equivalent to that obtained from squared-change parsimony.

Binary Data. Continuous will accept binary data. If a single binary trait is significantly correlated with a quantitative trait, the interpetation is that the mean value of the quantitative trait for the group coded zero on the binary trait, differs from the mean of the group coded 1. If all of the variables are binary a different model should be employed (see Pagel, 1994 and computer program **Discrete**).

Partial Correlation and Regression. Continuous is not set-up to calculate regressions on more than a pair of traits. However, partial correlations and regressions can be calculated by hand using the correlation matrix and standard formulae for partial correlations. The degrees of freedom for such tests are always a function of "n" where here "n" is the number of species.

Scaling Parameters in other analyses. In a likelihood ratio test the maximum likelihood values of the scaling parameters can differ under the null and alternatuve hypothesis tests of some other feature, such as the correlation. To ensure that tests are nested (see LR tests above) scaling parameters can be separately estimated under thenull and alternative hypothesis, and then fixed in both analyses at the value least supportive of the hypothesis. Lamba values may differ for two traits analysed separately but whose correlation one wishes to estimate. Here it is appropriate to estimate the value of lambda simultaneously with estimating the correlation.

VI. Graphing and Display

Continuous has some basic display and graphing options.

The **Show Likelihood Curves** box allows one to view a plot of the likelihood surface associated with the scaling parameters. These plots are informative by showing both the maximum likelihood value of the parameter, and by showing its 95% confidence intervals. If these intervals exlcude 0 or 1, this is equivalent to having performed an LR test.

The **Set Graph Options** and **Show Graph** option allow one to choose the nature of a display and then view it. Currently one can plot traits against total path length or against each other. These are useful for identifying outliers that may unduly influence the results.

With a likelihood surface or plot displayed, selecting **Export** from the **File** menu produces a text file of the raw data used to make the plot. This file can be used as input into plotting routines and is useful for making graphs and figures.

VII. General Tips and Known Problems

Likelihood solutions can be unstable especially when the number data points per parameter estimated drops below a ratio of about ten (i.e, ten data points per parameter). It is often a good idea to fit a model several times. Outlying data points can exert large effects on the estimates of parameters, especially on delta.

It is best to test nested hypotheses (for discussion see Pagel, 1994 or 1997) as these are approximately chi-squared distributed.

If the program returns an error message about the determinant being zero, it usually means that the phylogeny (either default or as implied by choice of scaling options) has at least two species that are effectively perfectly correlated. Removing one of them or changing the scaling can treat this problem. An effective way to alter the scaling in this circumstance is to change the lambda parameter from its default value of 1.0 to a value of 0.9999.

VIII. References

Pagel, M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society (B)* **255** 37-45 (1994).

Pagel, M. Inferring evolutionary processes from phylogenies. *Zoologica Scripta* **26**, 331-348 (1997).

Pagel, M. Inferring the historical patterns of biological evolution. Nature, 401, 877-884 (1999)

IX. Disclaimer

The routines have been tested and are correct to the best of my knowledge. However, I cannot take responsibility for anyone else's use of the program.

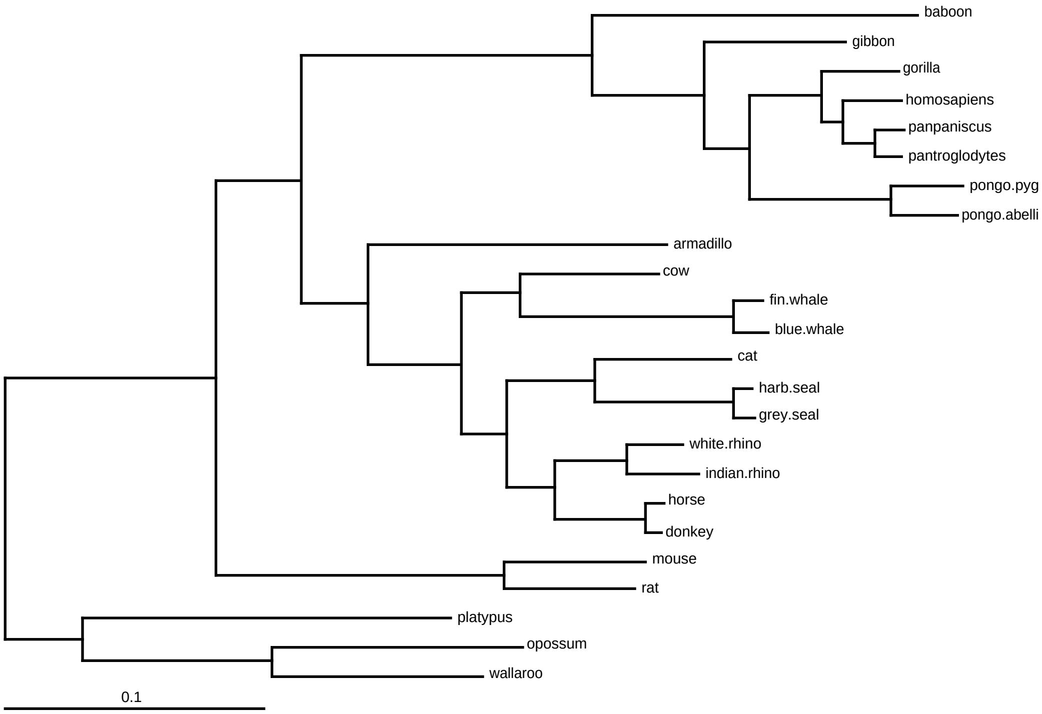
Appendix: Example data set

This data set includes values of brain size (grams) and body size (kg) on 24 mammal species. It is in the correct pag-file input format for Continuous. Some of the trait data are missing. The phylogeny is based on Mt DNA.

pag-format input file for tree shown below. Traits are in order of brain size, body size

```
#PAG
24 2
pan.pan, 1, .011750, 295.500, 33.000
pan.trog, 1, .011040, 410.300, 48.000
h.sap, 2, .022490, 1250.000, 65.000
gorilla, 3, .029830, 505.900, 126.500
pong.pyg, 4, .027020, 413.300, 53.000
pong.abelli, 4, .024970, *,*
gibbon, 6, .054450, 107.700, 5.500
baboon, 7, .126260, 142.500, 15.450
fin.whale, 8, .012300, 8325.000, 59394.000
blue.whale, 8, .014020, 6800.000, 58059.000
cow, 9, .053440, 450.000, 900.000
harb.seal, 10, .007580, 297.200, 84.833
grey.seal, 10, .008250, 320.000, 163.400
cat, 11, .052040, 20.700, 2.213
white.rhino, 12, .021870, 655.000, 764.000
indian.rhino, 12, .028610, *,*
horse, 13, .007820, 712.000, 484.000
donkey, 13, .006680, 405.000, 150.000
armadillo, 20, .115520, 12.000, 3.700
mouse, 14, .055170, .450, .024
rat, 14, .051110, 2.380, .339
platypus, 16, .142190, *,.690
opossum, 15, .096190, 6.420, 1.474
wallaroo, 15, .081150, 33.920, 12.460
1, 2, .011770
2, 3, .008890
3, 5, .027240
4, 5, .055220
5, 6, .017280
6, 7, .043700
7, 21, .111500
8, 9, .082200
9, 19, .022010
10, 11, .053360
11, 18, .033600
12, 17, .027100
13, 17, .034110
14, 22, .110190
15, 16, .073490
16, 23, .029620
17, 18, .018380
18, 19, .017810
19, 20, .036300
20, 21, .025510
21, 22, .032440
22, 23, .081790
```

The phylogeny as described in the (above) tree file.



Analyses of the brain-size and body-size data using log-transformed values.

log-likelihood of

Model A with trait covariances set to zero: -83.78

Model A allowing covariance to be non-zero: -62.87; LR test = 17.84

Model B allowing covariance to be non-zero: -61.42 (no evidence for Model B)

Estimated correlation (brain and body size) = 0.93, slope = 0.62

Estimated ancestral states

brain size: 3.02 (logged data) = 20 grams (anti-logged)

body-size: 1.65 (logged data) = 5.2 kg (anti-logged)

Scaling Parameters (estimated on both traits simultaneously, non-zero covariance))
kappa = 0.85 (95% CI = 0.09-1.5) log-likelihood = -62.77 (not different from 62.87)
delta= 2.07 (0.73-4.05) log-likelihood = -61.75
lambda = 0.99 (0.88-1.0) log-likelihood = -62.86