PAM

Phylogenetic Analysis in Macroecology

(v. 09, beta version)

Thiago Fernando Rangel

José Alexandre Felizola Diniz Filho

Universidade Federal de Goiás, Brasil

Richard Field

University of Nottingham, UK

INTRODUCTION

Comparative methods have, since Darwin's time, been the main approach to understanding evolutionary patterns and processes based on current diversity. On the one hand, the processes underlying trait evolution leave a distinct signal that can be measured when correlating species divergence and phylogenetic distances (Hansen & Martins 1996). On the other hand, the implication of this dependence is that species cannot be considered independent units, for statistical analyses, because they share ancestors at different points in time, thus creating autocorrelated structures for trait variation along the phylogeny (Felsenstein 1985; Harvey & Pagel 1991). As a consequence, using species as independent samples invites type I errors when applying standard regression and correlation techniques to analyzing species' traits – whether correlating traits with each other or with environmental variables (Martins & Garland 1991). To solve this problem, many different methods and approaches have been proposed for both analyzing such relationships and measuring the phylogenetic signal in data (Blomberg & Garland 2002; Martins et al. 2002).

Here we present a simple series of exercises to explore our new PAM (Phylogenetic Analysis in Macroecology) software, which was designed to deal with comparative data in a explicit phylogenetic context. The main purpose of PAM is to provide a friendly interface for applying phylogenetic comparative methods to evaluate phylogenetic signal and correlation, as well as phylogenetic patterns of clustering and overdispersion in local assemblages. In this preliminary beta version, only a few techniques are available for data analysis, but hopefully the series of exercises may help users to understand the overall reasoning behind the new software. It allows an explicit evaluation of geographical pattern in the results of these methods, with a direct link with our program for Spatial Analysis in Macroecology (SAM, freely available at www.ecoevol.ufg.br/sam – see Rangel et al. 2006, 2010). PAM was especially designed to deal with phylogenetic uncertainty using a randomization approach, in which species whose exact place in the phylogeny is not known can be included in the analysis (as long as they can confidently be placed within a clade in the phylogeny). This is important when dealing with macroecological data.

THE DATASET

PAM was designed to deal with phylogenetic patterns in traits and assemblages in a explicit spatial context. To use all capacities of the software, you will need four files:

- The **phylogeny,** written in a Newick format, in which phylogenetic relationships are expressed in a series of parenthesis and numbers.

- The **traits**, measured at species level, each trait in a column (the first column contains the species' name, exactly as it appears in the phylogeny). Note that SAM can also use a traits file, formatted in the same way except for one important difference: in SAM all the columns must contain the traits data – thus the species are not named in the traits file, and so they must appear in the same order as in the presence–absence matrix. PAM can function in the same way, but is more flexible in that you can name all the species in the first column and therefore the order of the species can be different from that in the presence–absence matrix.

- A **presence–absence** matrix (0/1) of species in different localities, regions or cells (one row per place) and species (one in each column).

- A **geographical file,** or shape file, with coordinates of spatial units (places – the same localities, regions or cells as in the presence–absence matrix) and associated environmental descriptors of each one. This is the SAM basic format. Note: unlike SAM, PAM can cope with missing values in the traits file and in the geographical file.

The phylogeny file is the MASTER FILE of PAM, and all others can be loaded only after loading the phylogeny. Not all files are necessary for all analyses, but in order to use all the functionality of PAM all of them must be loaded. The species name is used to link all the files in PAM (except the geographic one which links to the presence–absence matrix using the spatial units), so it is important that the species names (and the identities of the spatial units) are consistent between the files.

In addition to these basic files, it is possible for both presence–absence and traits files to contain more species than appear in the phylogeny. This will be common because phylogenetic relationships are not known for many of the species usually available in macroecological datasets. In such a case, a fifth file is necessary to run PAM, containing what we call **PUT**s (**Phylogenetically Unknown Taxa**). Each species in this file will be "linked" to the phylogeny by its **MDCC** (*Most Derived Consensus Clade*), which must be part of the phylogeny.

***Exercise 1** – understanding the format of each file type (especially the PUT file and its link with the phylogeny file).*

- Open each of the 5 'Wolverine' files provided (*.txt (ASCII) files) in Excel or Notepad/Word, as appropriate. Explore each of these files and then close them.

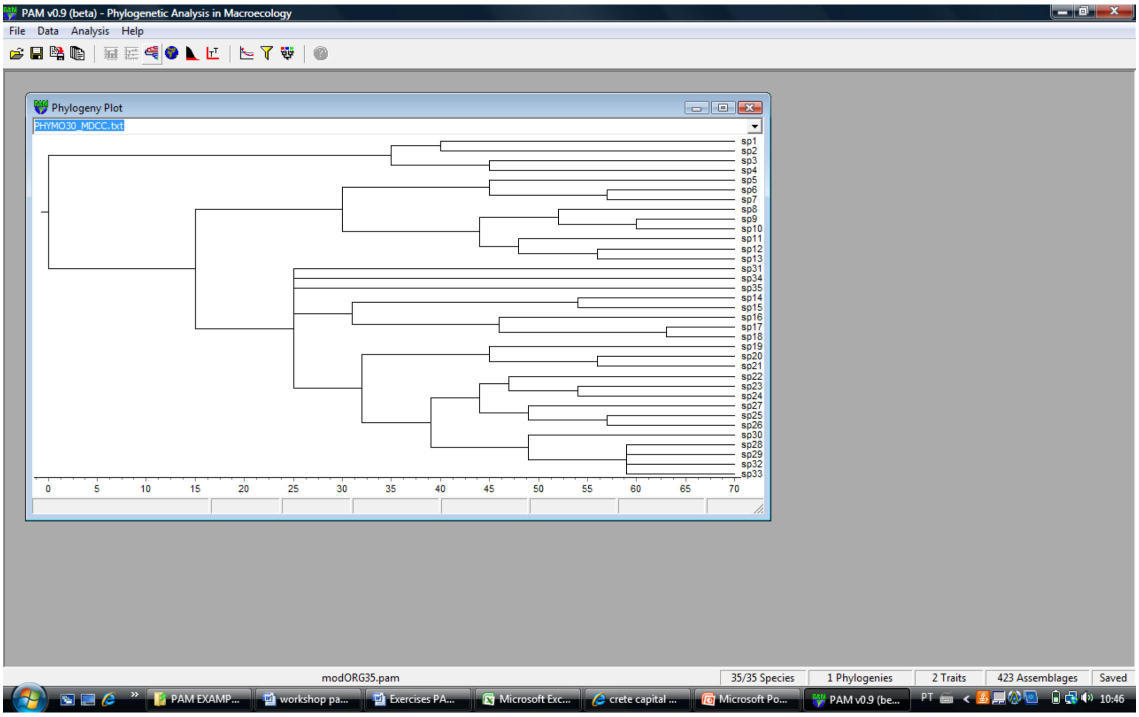
- Open PAM and then open the same 5 files, using the Data Settings dialog (Ctrl-d), starting with the phylogeny. When all five are open, save the whole project as a *.pam file (you can use Ctrl-a).

- Close PAM and open the *.pam file in Notepad or Word. Examine how the five sub-files are organized. You will see that the *.pam file is related to the most commonly used current format for phylogenetic information, the NEXUS format.

In the *.pam file, then, you will find all data you need, combined into a single file, which can be opened using “File/Open”, the 📂 icon or by double-clicking from Windows Explorer.

THE PHYLOGENY

If you click in the icon in PAM (or press Ctrl-p) it will display the phylogeny:



PAM v0.9 (beta) - Phylogenetic Analysis in Macroecology

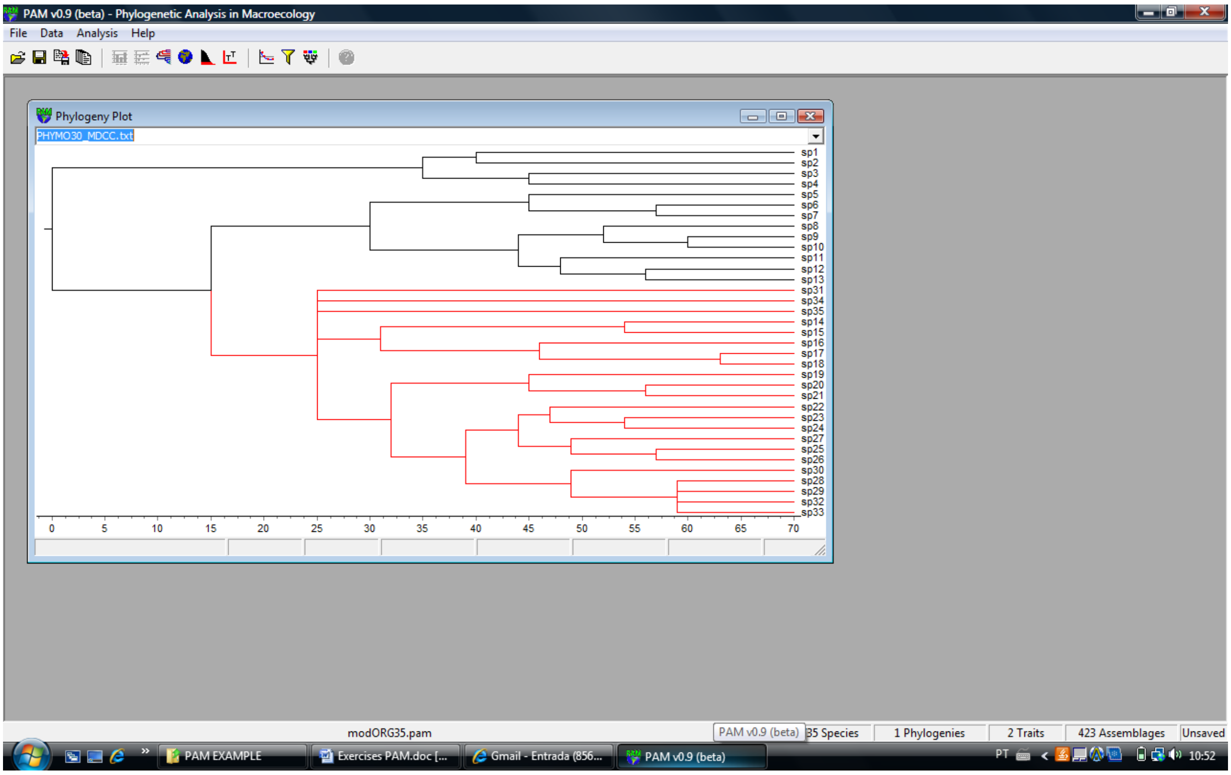
Phylogeny Plot

PHYMO30_MDCC.txt

modORG35.pam | 35/35 Species | 1 Phylogenies | 2 Traits | 423 Assemblages | Saved

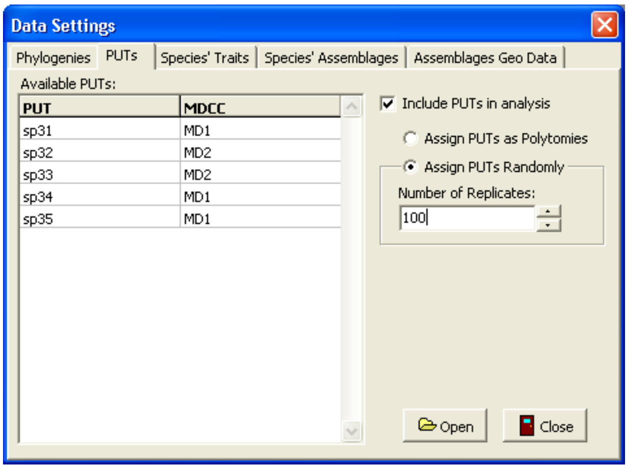
In the bar at the bottom, basic information of how many species, PUTs, traits and assemblages will be displayed. Notice that **you can open different phylogenies for the same set of species at the same time**, and display each one by clicking in the upper right arrow of the phylogeny plot ▼. You can see that, in this phylogeny, there are two sets of PUTs (5 species) that appear as polytomies.

Holding Shift and clicking a clade allows you to select it: it will turn red, as in the screen grab below.



All subsequent analyses and plots will be done <u>only</u> on this clade. You can add to the selection by again holding Shift and left-clicking another clade. Holding Ctrl and left-clicking a clade unselects it. To clear all existing selections and select a clade, hold Alt and left-click the clade. To select the whole phylogeny, Shift-left click the root of the tree, which will turn black to indicate that there is no selection of a subset of the tree (to select an individual clade again after doing this, use Alt). Also notice that when the cursor hovers over a branch, the characteristics of that branch (number of species [tips], level, length, height) appear at the bottom of the window.

If you go to “File/Data Settings” (Ctrl-d) you can see all available files (as in SAM), and if you go to the PUT window you can see the options for PUTs. One choice is whether or not to include the PUTs in analyses. If unticked, all analyses will be done without considering them. If PUTs are to be used, they can be included as polytomies (as shown in the screen shot above) or by randomizing their positions within the MDCC. You specify the number of randomizations; if you choose 100, all analyses will be run 100 times, and in each one the polytomies will be resolved by randomly assigning the PUT within its clade (see below). While this does not affect the display of the phylogeny, it allows computation of the uncertainty that is due to phylogenetically unknown clades. If you disable the PUTs, only the 30 species that are PKTs (phylogenetically known taxa) will be shown, in a fully resolved tree.



Available PUTs:

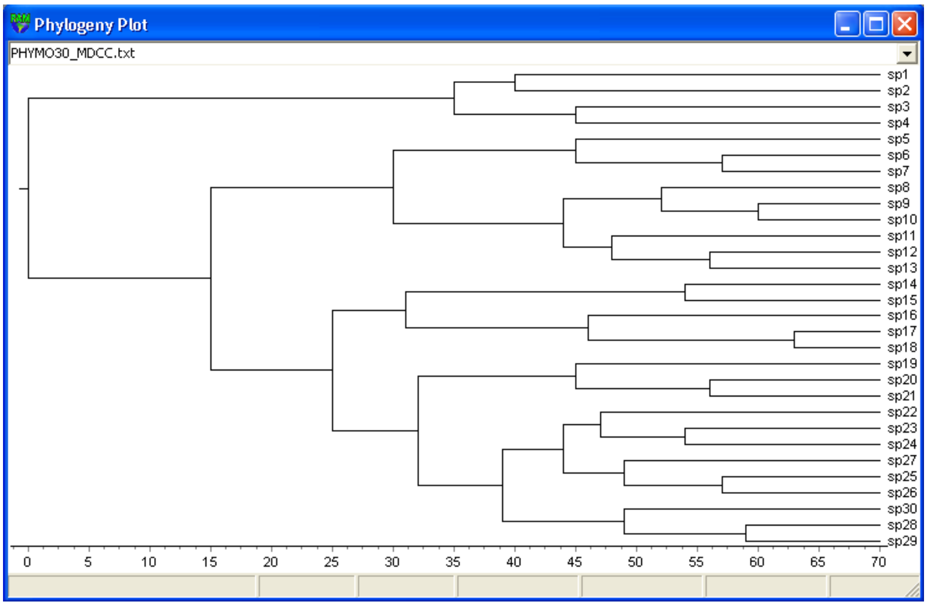
PUT	MDCC
sp31	MD1
sp32	MD2
sp33	MD2
sp34	MD1
sp35	MD1

☑ Include PUTs in analysis

○ Assign PUTs as Polytomies

● Assign PUTs Randomly

Number of Replicates: 100

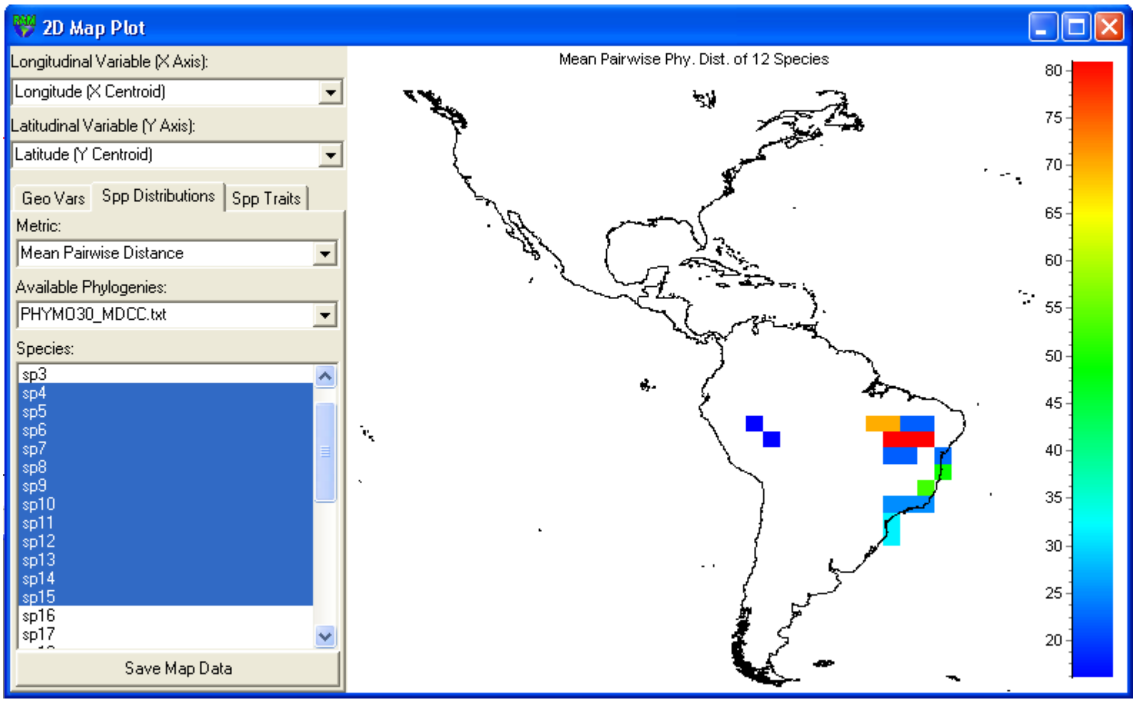


***Exercise 2** – explore the phylogenies, enabling and disabling the PUTs, mark subclades and check their characteristics*

PLOTTING MAPS

Because PAM was designed to deal with phylogenetic patterns in geographical space, we must also explore the mapping facilities. Click the icon 🌍 or press Ctrl-m. You can see the maps within the geographical file (Geo Vars tab), but you can also map different features of species files (Distributions and Traits tabs). This is equivalent to, but much more efficient than, the routines for managing presence–absence and traits data in SAM, and can be useful for exploratory phylogenetic–geographic connections.

In the “Species Distribution” tab, you can see the discrete distribution of each species in the presence–absence matrix, or can combine different species and see their distribution or richness. You can also map Richness, Most Recent Common Ancestor (MRCA) of selected species in each cell, and other metrics of phylogenetic diversity (PSV, Mean Taxonomic Distances). These phylogenetic distances can be also analysed with more details in the “Community Phylogenetics” module (below). Notice also that some of these metrics (PSV and all the community phylogenetic metrics) will be calculated only for cells that contain two or more species.



Longitudinal Variable (X Axis): Longitude (X Centroid)

Latitudinal Variable (Y Axis): Latitude (Y Centroid)

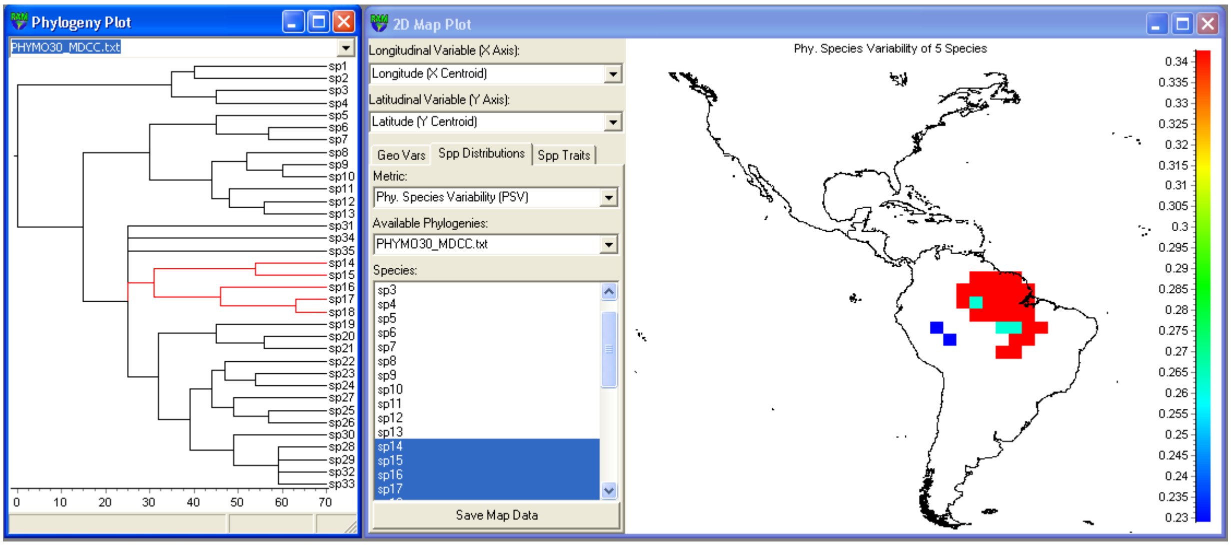
Metric: Mean Pairwise Distance

Available Phylogenies: PHYMO30_MDCC.txt

Mean Pairwise Phy. Dist. of 12 Species

In the “Species Traits” tab, you can map mean, standard deviations, sum, maximum and minimum of traits. Any other species-level outputs from other modules (S-component of PVR, see below) will be available here too, so you will be able to map some of the outputs in later exercises.

Notice that if you select a clade in the phylogeny (as in Exercise 2) this will be reflected automatically in the mapping module as well. You can open both windows at the same time and see this more easily.



Longitudinal Variable (X Axis): Longitude (X Centroid)

Latitudinal Variable (Y Axis): Latitude (Y Centroid)

Metric: Phy. Species Variability (PSV)

Available Phylogenies: PHYMO30_MDCC.txt

Species: sp3, sp4, sp5, sp6, sp7, sp8, sp9, sp10, sp11, sp12, sp13, sp14, sp15, sp16, sp17

***Exercise 3** – Explore the mapping features of PAM, including the geographical variables, the species distributions and the species traits. Do this also after selecting species in the list in the mapping window, and also by selecting clades directly in the phylogeny.*

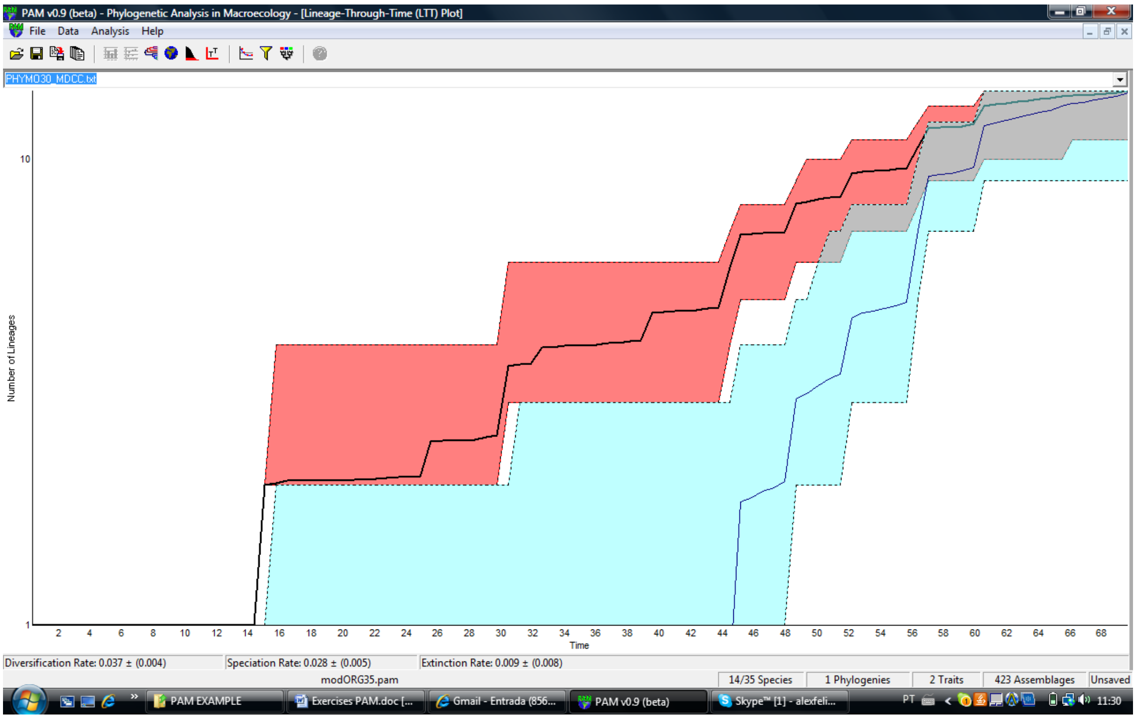
Note for those familiar with SAM: the maps work in the same way, e.g. you can save them or the data behind them, customise them in various ways, etc, by right-clicking on them.

THE ANALYSES

Here we briefly explore five analyses that are currently available in this beta-version of PAM. Further analyses will be available in the future (including more complex ones), but these five will include the basic tools for answering some basic questions in phylogenetic comparative analysis, within the framework of dealing with phylogenetic uncertainty. All example analyses below use 100 randomizations of PUTs in their MDCCs, so the complete outputs are shown. During the exercises, it is also nice to deselect the PUTs and compare the outputs with and without PUTs, and also (when including the PUTs) to compare results from using PUTs as polytomies with those from randomizing them.

LINEAGE-THROUGH-TIME (LTT) PLOTS

The first analysis is a basic description of how the lineages arise through time. This plot can be used for some inferences about clade dynamics. PAM plots the log of number of lineages by sectioning the phylogeny in small steps and, if PUTs are enabled, it shows the variability around the mean curve (randomizing the PUTs on their respective MDCCs). Here is the LTT plot for the clade containing species 14–35.



Diversification Rate: 0.037 ± (0.004)

Speciation Rate: 0.028 ± (0.005)

Extinction Rate: 0.009 ± (0.008)

modORG35.pam | 14/35 Species | 1 Phylogenies | 2 Traits | 423 Assemblages | Unsaved

In the bottom bar of LTT plot, three statistics are shown, and we followed Ricklefs (2006) for computing them. The first is the diversification rate, which is the slope of the linear regression of log(S) against time, using only the middle of the curve (i.e. omitting the points with the top and bottom 20% of the X-values). The slope of the regression is actually $b = \lambda - \mu$, where λ is the speciation rate and μ is the extinction rate. If these parameters are constant through time, it is expected that the actual number of lineages appearing through time is higher than the observed regression line, but possess the same slope. The difference between the ln(current S) and the ln(estimated S) by the curve will give you the a-value, which is the expected difference between the intercepts of the observed and 'real' LTT. Because extinctions tend to be smaller in recent nodes, it is expected in principle that slope will increase towards the present. This allows computing speciation rate as $\lambda = b / \exp(-a)$, and the extinction rate is obtained by subtraction. These are the quantities shown in LTT, and the numbers in parentheses are their standard errors that are due to phylogenetic uncertainty (if you deselect the PUTs, only the rates, and not the standard errors, will be displayed).

Exercise 4 *– Explore the LTT curve, enabling and disabling the PUTs. Link the shape of the curve with the parameters.*

MORAN' I CORRELOGRAMS

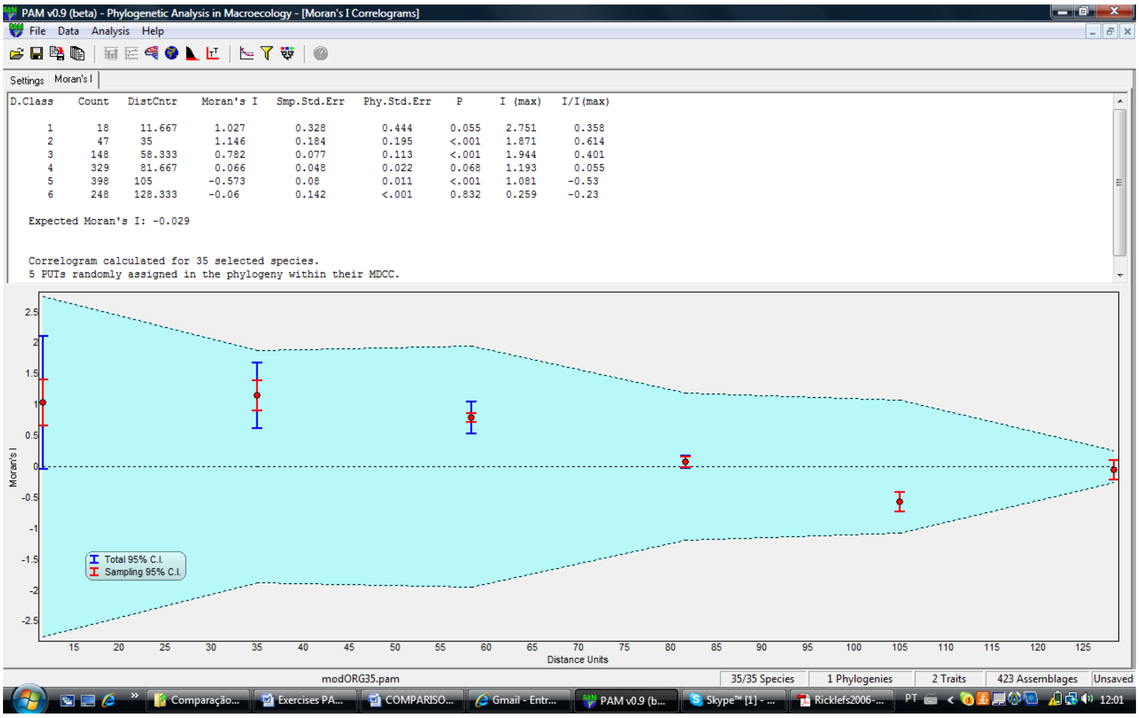
This will allow computing Moran's I autocorrelation coefficients for a series of phylogenetic distance classes, forming a correlogram (see Gittleman & Kot 1990; Diniz-Filho 2001; Pavoine et al. 2008). This is a simple way to define the magnitude and "shape" of phylogenetic signal in data. For those familiar with SAM or any spatial analysis software, using and interpreting this is straightforward. Moran's I tends to vary between -1.0 and 1.0, although much larger values can appear in irregular connections. Moran's I is given by

$$I = \left(\frac{n}{S}\right)\left[\frac{\sum_i \sum_j (y_i - \bar{y})(y_j - \bar{y})w_{ij}}{\sum_i (y_i - \bar{y})^2}\right]$$

where n is the number of species, y_i and y_j are the trait values in the species *i* and *j*, with average $\bar{y}$, and w_{ij} is an element of the weighting matrix **W**, with the pairwise phylogenetic relationship among species. In phylogenetic analyses, the matrix **W** can be given by the phylogenetic correlations, **C**., or generalized to other forms of relationship matrix (see Pavoine et al. 2008). In principle, Moran's I measures the ratio between phylogenetic covariance among species (i.e., product of departures from the mean of each species weighted by their phylogenetic relatedness) and total variance in a trait. **So a high value of Moran's I indicates that closely related species (i.e., higher values in W) are more similar than expected by chance alone**.

It is more common to estimate Moran's I for several classes, forming a phylogenetic correlogram. To build this Moran's I, binary weighting matrices **W** are calculated, whose elements Wij equal to 1 indicate pairs of species that are separated by a given interval in the matrix **C** (see Diniz-Filho, 2001). These phylogenetic distance classes can be defined by regular time steps, or by defining them so that the number of connections is similar in each class (making the errors similar throughout the correlogram). In PAM, there is a default number of classes (that can be changed) and the two ways to define classes, as above. The default is equal distances (unlike in SAM, where the default is equal number of connections).

In PAM, using the PUTs with randomizations will produce two different confidence intervals for each Moran's I coefficient, based on errors (or a combination of errors displayed in the numerical output). The red one is the standard statistical error of the coefficient, whereas the largest one, the blue one, also incorporates the phylogenetic uncertainty that is due to the PUTs. The errors are combined, as suggested by Martins (1996; see also Housworth & Martins 2001), as the square root of the sum of variances that are due to statistical deviation and phylogenetic uncertainty. This is shown below for the variable ‘BS’, using the whole phylogeny:



```
D.Class    Count   DistCntr    Moran's I   Smp.Std.Err   Phy.Std.Err    P      I (max)   I/I(max)

     1       18     11.667      1.027        0.328          0.444      0.055   2.751      0.358
     2       47     35          1.146        0.184          0.195     <.001    1.871      0.614
     3      148     58.333      0.782        0.077          0.113     <.001    1.944      0.401
     4      329     81.667      0.066        0.048          0.022      0.068   1.193      0.055
     5      398    105         -0.573        0.08           0.011     <.001    1.081     -0.53
     6      248    128.333     -0.06         0.142         <.001       0.832   0.259     -0.23

   Expected Moran's I: -0.029


   Correlogram calculated for 35 selected species.
   5 PUTs randomly assigned in the phylogeny within their MDCC.
```

Exercise 5 *– explore the correlogram module, for the two traits available (BS – body size; RANGE – geographic range size), comparing their phylogenetic patterns, calculated with and without the PUTs, as well as changing number of distance classes and ways to define them. Note particularly how counts (second column of numerical output) changes with these options.*

PHYLOGENETIC EIGENVECTOR REGRESSION

A Short Description of PVR

The so-called “partition methods” in phylogenetic comparative analyses were derived in a roughly analogous manner to quantitative genetic models that partition phenotypic variance into genotypic and environmental components. The idea is that the total variation, T, can be partitioned into phylogenetic (P) and specific (S) components (Cheverud *et al.*, 1985). The P-component contains variation among species that is due to phylogenetic effects, whereas S-component expresses the unique variation of each species for the trait, independently of phylogenetic relatedness. This partition of T into P and S can be done in different ways. Cheverud *et al*. (1985) pioneered the use of an autoregressive model (ARM) of the form

$$\mathbf{Y} = \rho\mathbf{WY} + \varepsilon$$

whereas Diniz-Filho *et al.* (1998) later proposed a multiple regression model called Phylogenetic EigenVector Regression (PVR) that is given by

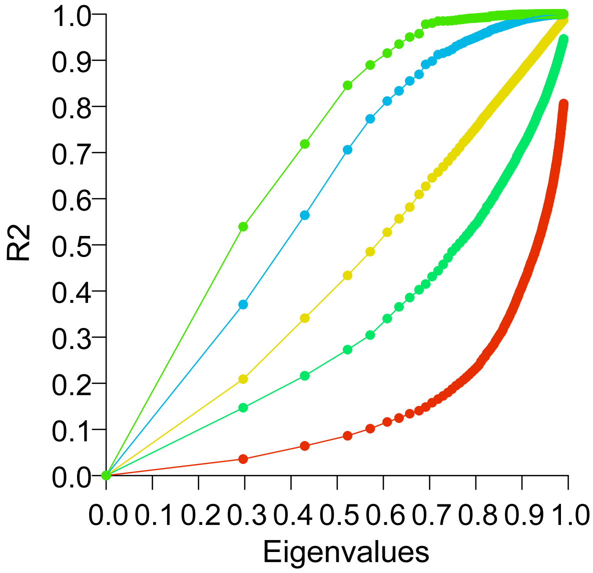
$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

In both cases, the vector ε contains the model residuals and expresses the S-component, whereas the $\rho\mathbf{WY}$ and $\mathbf{X}\beta$ are the vectors of estimated values (the P-components) from both methods. In PVR, the **X** is the matrix with selected eigenvectors extracted from the phylogenetic distances **D** and that express phylogeny in a vector form in a hierarchical way. In both ARM and PVR, the coefficients of determination of the models (R^2) approximate the phylogenetic signal in data.

The S-component thus contains the part of the variation in a trait that is independent of the phylogenetic structure defined by the eigenvectors of **D**, and thus contain only changes after deviations from the ancestor (see Diniz-Filho et al. 2009 for a explicit test of this). On the other hand, the P-component estimates the expected value of the trait under a linear combination of all other species with respect to phylogenetic structure. To correct for Type I error, only correlations between S-components must be considered, and these estimate the “input” correlation, which is the intrinsic change throughout the phylogeny (i.e., the same parameter estimated by Felsenstein’s (1985) popular phylogenetic contrasts method). See Martins et al. (2002) and Diniz-Filho & Torres (2002) for a comparative evaluation of PVR’s statistical performance under diverse evolutionary models.

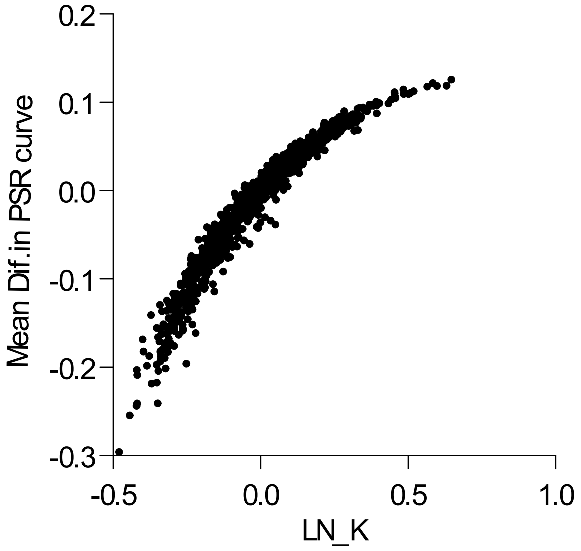
The main problem with PVR, as occurs with any multivariate analysis, is how to select eigenvectors of **D** for further analysis. As pointed out by Rohlf (2001), under a linear evolution process (i.e., Brownian motion – see Hansen & Martins 1996) without error, all eigenvectors of **D** are necessary to fully described the phylogeny and, in this case, the R^2 goes to one (maximum phylogenetic signal). In practice, however, just a few eigenvectors of **D** are needed to explain the variation (see Martins et al. 2002), but the **main critical diagnosis is to check whether PVR residuals (the S-component) are independent**. This can be done using Moran’s I correlograms, as previously discussed (see Gittleman & Kot 1990; see also Diniz-Filho et al. 2012a for a recent evaluaton).

Diniz-Filho et al. (2012b) show that Rohlf’s (2001) criticisms can be used to develop an interesting approach to measuring phylogenetic signal from PVR. Under Brownian motion, we expect that the amount of divergence in a trait (measured by PVR's R^2) along an eigenvector is directly proportional to the representation of this axes in relation to the phylogeny. This leads to a linear relationship between the PVR's R^2 and the cumulative relative eigenvalues of the eigenvectors used in the model (expressed as proportion of the trace), generating what we called the *PSR (Phylogenetic Signal-Representation) curve:*



The PSR curve from PVR under distinct models (the yellow curve is from a Brownian motion evolution).

Different evolutionary models produce different PSR curves, indicating a trait evolving faster or slower than expected under Brownian motion. It is interesting that mean deviations from PSR curve are strongly correlated (in a non-linear way) to Blomberg's et al. (2003) K-statistics (diagram below), recently used to evaluate niche conservatism (e.g., Hof et al. 2010).



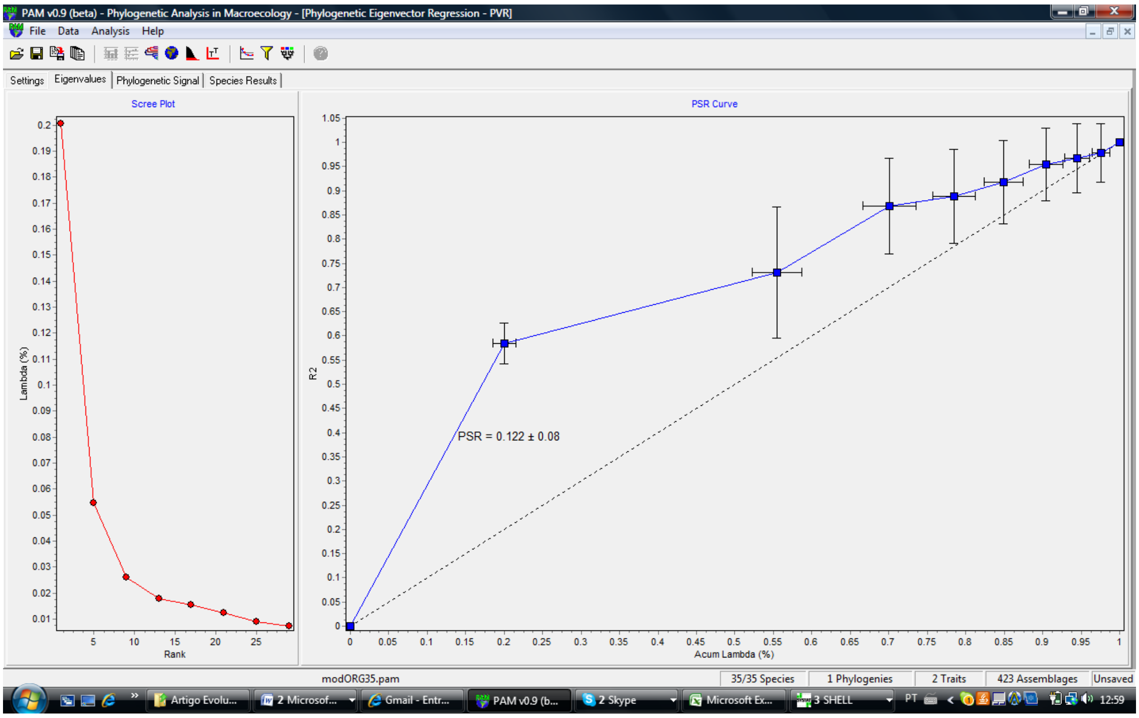
Relationship between Blomberg's et al. (2003) K-statistics, at log-scale, and mean difference between R^2 and eigenvalue in the PSR curve based on 1000 simulations of Brownian motion. Notice that mean under Brownian motion is zero in the two metrics.

Thus, PVR can be used with two slightly different purposes: 1) to estimate Phylogenetic Signal (PVR's R^2 or mean deviation of PSR curve), or 2) to estimate correlated evolution among traits, using the S-component. In the first case, the main issue is that a high value of R^2 indicates a strong phylogenetic signal, but the user must keep in mind the amount of explanation of phylogeny used in PVR (the eigenvectors). More formally, the cumulative eigenvalues are a reference for R^2 measuring deviations from Brownian motion. When using PVR for correlated evolution, the most important issue is to ensure that S-component is independent among species.

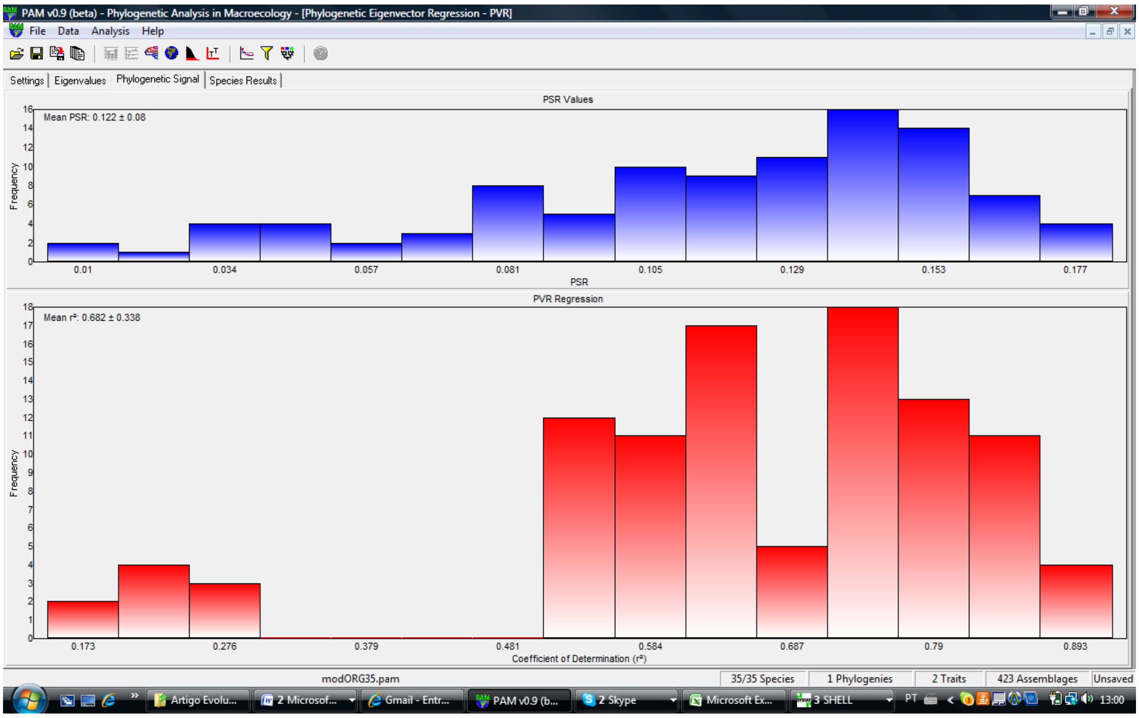
PVR in PAM

The PVR now computed in PAM v. 0.9 (beta) appears in its simplest version. Later, more complex implementations (e.g., Desdevises et al. 2003; Safi & Pettorelli 2010) will be available. The user must select a single trait in the left box, and then define which criterion will be used to select the eigenvectors (e.g., minimizing Moran's I in the first distance class, as used in spatial analyses – see Griffith & Peres-Neto 2006 – or using those eigenvectors whose eigenvalues explain 95% of the variation in the phylogenetic distances, among others – see Diniz-Filho et al. 2012a). The number of steps in the PSR curve defines how many cumulative PVRs will be performed (increasing the number of eigenvectors in the model), but will not affect the final outputs defined by the selected criterion of eigenvector selection (i.e., the R^2 and the S-component).

The first output of PVR in PAM includes the scree plot of eigenvalues and the PSR curve, using the number of steps defined by the user. Note that the 1st eigenvector and the extreme points (0,0 and 1,1 points of PSR curve, that exist by definition) will always be displayed. Notice that bars in the PSR curve indicate the phylogenetic uncertainty in the eigenvalues (X-axis) and in the PVR's R^2 (Y-axis).

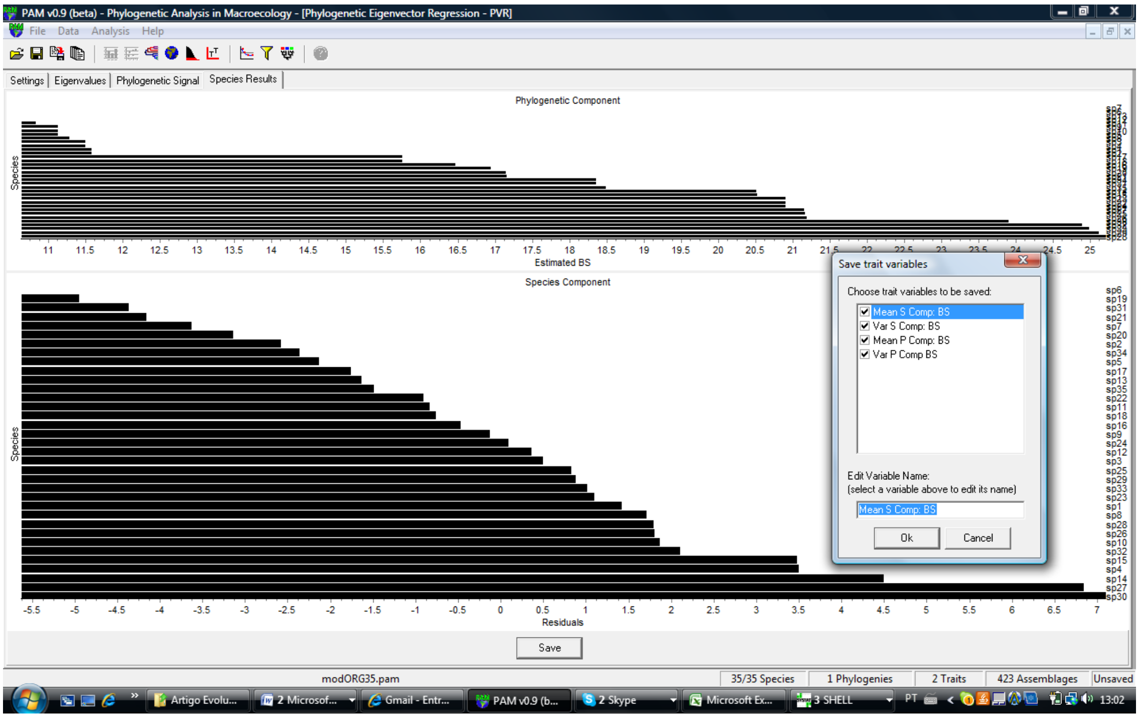


The second set of outputs (shown below) is the statistical distribution of mean differences between R^2 and eigenvalues in the PSR curve and the PVR's R^2 along the randomization of PUTs. Remember that a mean difference of zero suggests a Brownian motion process. If randomized PUTs are not selected, only a single bar will appear.



Distribution of mean differences between R^2 and eigenvalue in PSR curve (blue) and of PVR's R^2, along 100 randomizations of PUTs.

The last output is a plot of P- and S-component values, ordered by species – see screen shot below, and it allows saving these components. (This plot is also available for any species' traits in the Species' Trait Plot of PAM.)


PAM v0.9 (beta) - Phylogenetic Analysis in Macroecology - [Phylogenetic Eigenvector Regression - PVR]

Save trait variables

Choose trait variables to be saved:

☑ Mean S Comp: BS
☑ Var S Comp: BS
☑ Mean P Comp: BS
☑ Var P Comp BS

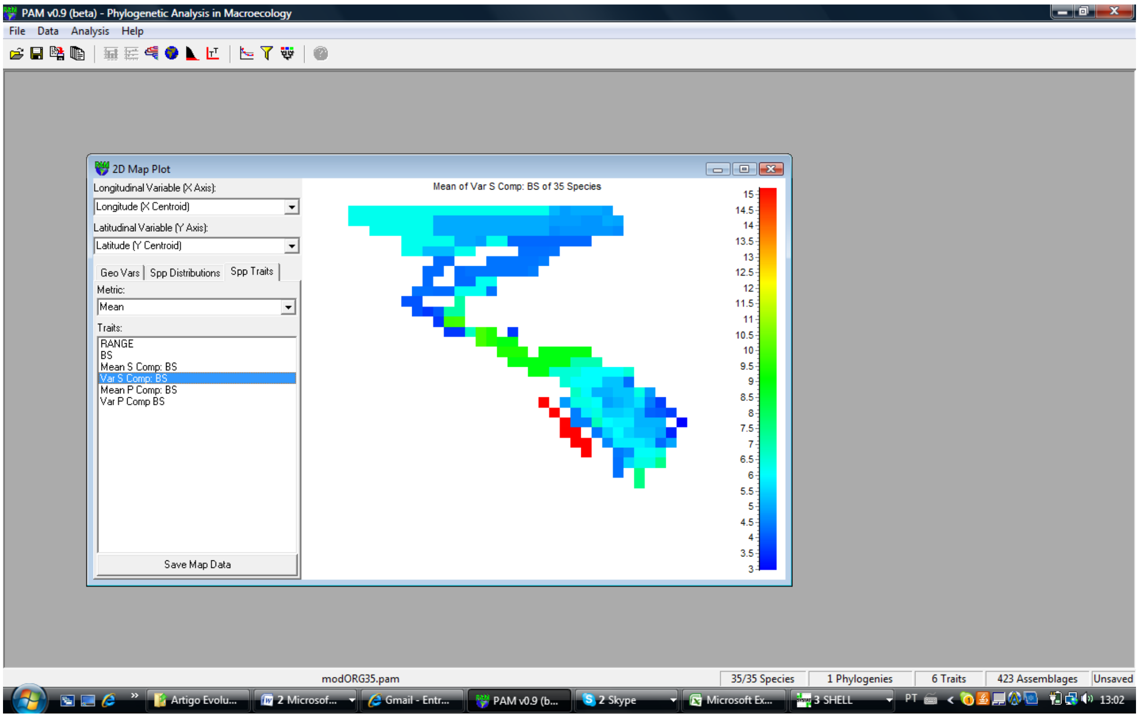
Edit Variable Name:
(select a variable above to edit its name)

Mean S Comp: BS

modORG35.pam | 35/35 Species | 1 Phylogenies | 2 Traits | 423 Assemblages | Unsaved

Values of P- and S-components, ordered by species, and the "Save button" showing the option to save the mean and variances of these components.

If randomized PUTs are enabled, both mean and variances (across randomizations of PUTs) will be saved and become part of the species' traits file, and will later be saved to the *.pam file. You can then go to Moran's I correlogram module to check the phylogenetic structure in the S-component (that must be eliminated) or map the S-component (see Diniz-Filho et al. 2007, 2009):



Longitudinal Variable (X Axis): Longitude (X Centroid)

Latitudinal Variable (Y Axis): Latitude (Y Centroid)

Metric: Mean

Traits:

RANGE
BS
Mean S Comp: BS
Var S Comp: BS
Mean P Comp: BS
Var P Comp BS

modORG35.pam | 35/35 Species | 1 Phylogenies | 6 Traits | 423 Assemblages | Unsaved

Mapping the mean S-component of PVR, using the mapping module (notice the highest positive deviation from phylogenetic expectation in red, in the Andean region).

***Exercise 6**. – Explore the PVR module of PAM, changing the way eigenvectors are selected and comparing the final outputs (the S-component and distribution of R^2). Notice that PSR curve is unaffected by this, but will change when using more steps for improving estimates.*

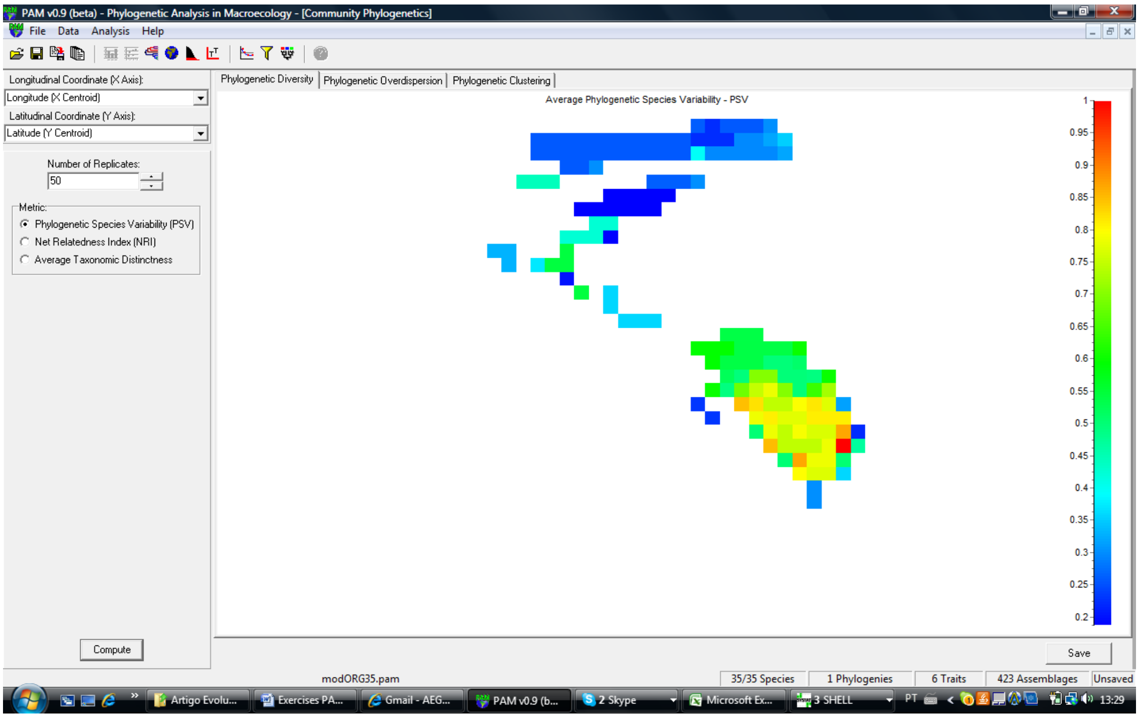
Compute the PVR for the two available traits and compare the patterns. Save the S-components and use them in other modules of PAM, to correlate the two traits and to map their mean variation.

THE COMMUNITY PHYLOGENETICS MODULE

The last analytical module available in PAM v. 0.9 involves some simple metrics developed in the context of community phylogenetics, although of course they can be used at any scale (see Schweiger et al. 2008). In general, the idea is to compute indices for phylogenetic diversity and clustering.

The only option for running these phylogenetic diversity metrics is the number of replicates for testing null hypothesis of randomness in species' patterns within each cell. Notice that this is not the randomization of PUTs, that are already defined in the “Data Settings “ menu. This will be the number of randomizations by which each index is permuted using each of the phylogenies obtained by randomizing the PUTs, for each cell in the map. The output of this module produces maps (for which a simple PSV is also available in the mapping module), both for mean values of phylogenetic diversity, and for the significance test of the patterns. These metrics are calculated only when more than 2 species are present in a cell.

The first metric available is the PSV, Phylogenetic Species Variability (Helmus et al. 2007), which is given by PSV = 1 – *c*, where *c* is the mean pairwise phylogenetic correlation among species found in a cell. A PSV close to 1 is obtained when *c* tends to zero, i.e., species tend to be phylogenetically independent and thus a lot of diversity exists. On the other hand, if species are strongly related and mean *c* is large, this indicates that there is less phylogenetic diversity.



PAM v0.9 (beta) - Phylogenetic Analysis in Macroecology - [Community Phylogenetics]

Longitudinal Coordinate (X Axis): Longitude (X Centroid)

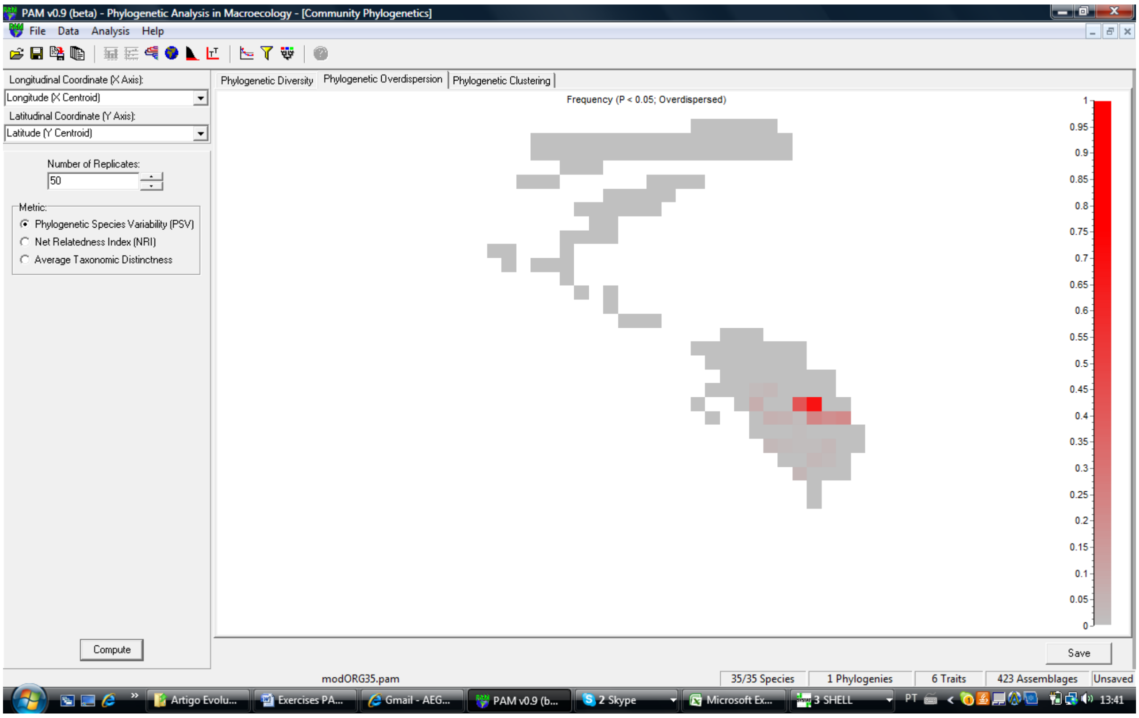
Latitudinal Coordinate (Y Axis): Latitude (Y Centroid)

Number of Replicates: 50

Metric:
● Phylogenetic Species Variability (PSV)
○ Net Relatedness Index (NRI)
○ Average Taxonomic Distinctness

modORG35.pam | 35/35 Species | 1 Phylogenies | 6 Traits | 423 Assemblages | Unsaved

When randomizing the species and recomputing PSV values in each cell, for each phylogeny, it is possible to establish whether PSV is larger or smaller than expected under a random combination of species. If PSV is higher than expected by a random association of species, then species tend to be overdispersed in the phylogeny, for a given cell, whereas a PSV smaller than expected indicates phylogenetic clustering in a cell. Because this test is done for multiple phylogenies (randomizing the PUTs), the other maps show the frequency that a given cell appears as overdispersed or clustered at a 5% level:



Longitudinal Coordinate (X Axis): Longitude (X Centroid)

Latitudinal Coordinate (Y Axis): Latitude (Y Centroid)

Number of Replicates: 50

Metric:
● Phylogenetic Species Variability (PSV)
○ Net Relatedness Index (NRI)
○ Average Taxonomic Distinctness

Frequency of significant PSVs (overdispersion) for each cell.

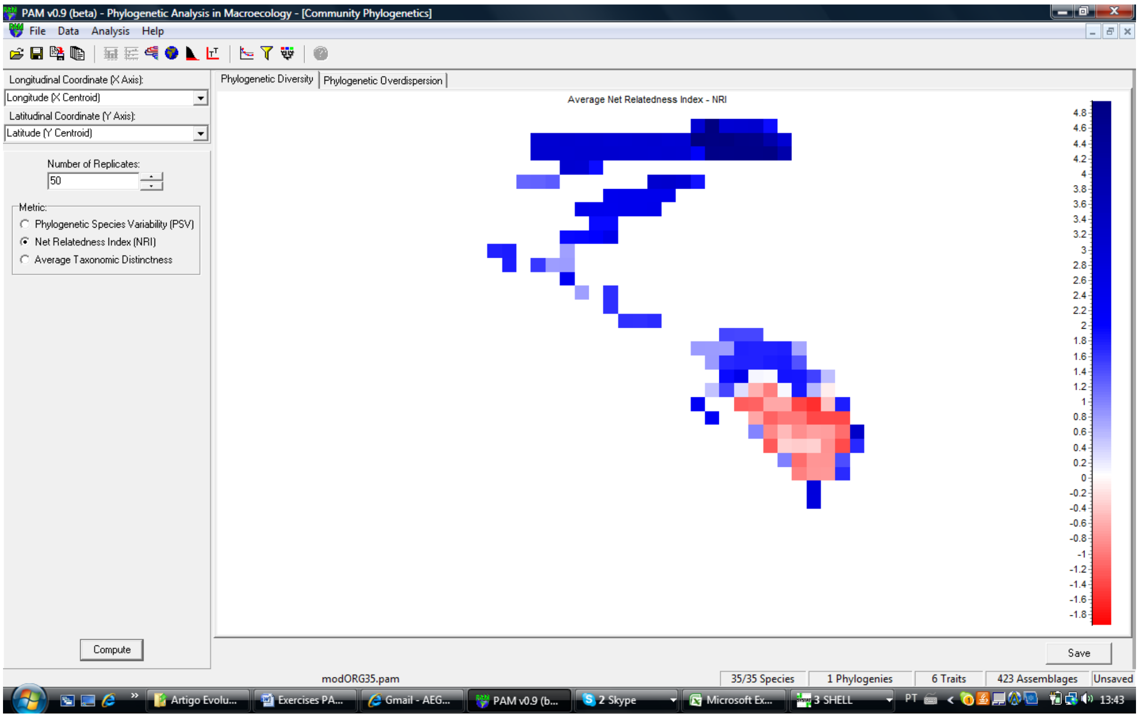
The second metric is the Nearest Relatedness Index (NRI, see Webb et al. 2002), which is actually a Z-statistic based on mean phylogenetic distances in a cell. The NRI is given by

$$-NRI = (M – Mr) / se(Mr)$$

where M is the mean phylogenetic distances observed, Mr is the mean of mean distances after randomly sampling species from the phylogeny and assigning it to a cell, and se(Mr) is the standard error of this mean. Thus, a negative value of NRI indicates that the observed mean is significantly smaller than randomized mean, and thus species are closer phylogenetically than expected. At a 5% significance level, an absolute value of Z greater than 1.96 indicates a significant pattern, with NRI > 1.96 indicating cluster and an NRI <- 1.96 indicating overdispersion.

The third metric is the mean phylogenetic distance, which is the basic metric to compute the NRI, and thus will give exactly the same result of NRI in terms of significance tests.

Notice that the map data can be saved and then used for further geographical analyses, correlating with other variables available in the geographical file. However, for more complex geographical analyses of these patterns, these characteristics of cells must be transferred to more specific software for spatial analyses, like SAM.



Longitudinal Coordinate (X Axis): Longitude (X Centroid)

Latitudinal Coordinate (Y Axis): Latitude (Y Centroid)

Number of Replicates: 50

Metric:

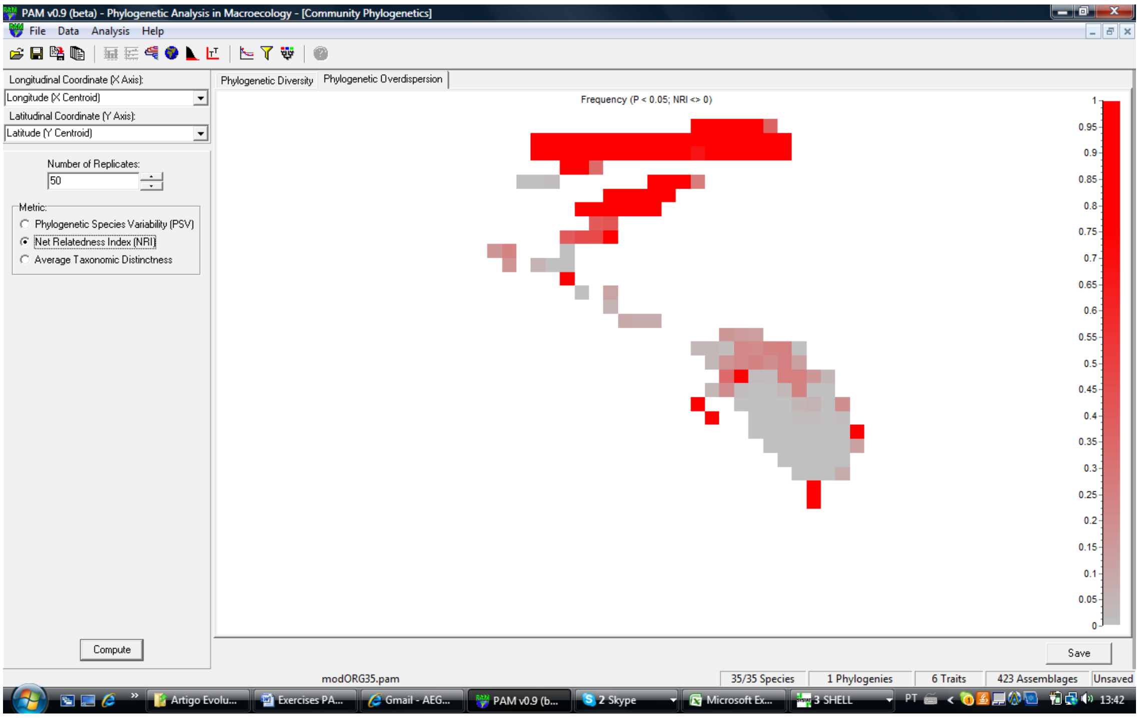
○ Phylogenetic Species Variability (PSV)

● Net Relatedness Index (NRI)

○ Average Taxonomic Distinctness

modORG35.pam | 35/35 Species | 1 Phylogenies | 6 Traits | 423 Assemblages | Unsaved

Average NRI



Longitudinal Coordinate (X Axis): Longitude (X Centroid)

Latitudinal Coordinate (Y Axis): Latitude (Y Centroid)

Number of Replicates: 50

Metric:

○ Phylogenetic Species Variability (PSV)

● Net Relatedness Index (NRI)

○ Average Taxonomic Distinctness

modORG35.pam | 35/35 Species | 1 Phylogenies | 6 Traits | 423 Assemblages | Unsaved

Frequency of significant NRI (|NRI| > 1.96) for each cell. The red part in northern North America shows that cells there have clustered patterns (positive NRI, as show in the previous screenshot) in all randomized phylogenies.

Exercise 7 - *Explore the community phylogenetics metrics, interpreting the outputs in terms of cluster and overdispersed spatial patterns in local assemblages. Save the maps and correlate the metrics with environmental variation.*

THE pGLS MODULE

pGLS models

The phylogenetic Generalized Least Squares (pGLS) is a general and flexible module of PAM that allows calculating phylogenetic signal (by using Blomberg's *et al*. (2003) *K*-statistics and Pagel's λ - Freckleton et al. 2002), comparing the fit of alternative evolutionary models (Brownian and O-U models) by likelihood and using the covariance structure derived from the phylogeny to generate unbiased estimates regression slopes and their statistical significance.

The phylogenetic signal, in the context of pGLS, is usually estimated in respect to a Brownian expectation and is quite useful in a relative context (i.e., comparing phylogenetic signal for the same trait in distinct phylogenies, or comparing groups of species). The idea is that the phylogenetic signal actually depends on tree shape, especially in terms of branch length distribution (stemminess). If TIPs of the phylogeny are too "long", and species only share a small fraction of the overall evolutionary history, the signal will be low regardless of statistics used (i.e., Moran's *I* described above). But if one simulates Brownian motion on this same tree, this would furnish a reference for the amount of phylogenetic signal expected for that particular tree. This Brownian expectation does not actually need to be simulated, being rather derived analytically. Expectations of phylogenetic signal estimated by both *K* and λ, which are too popular metrics for phylogenetic signal based on pGLS, are equal to 1 under Brownian motion.

Blomberg's et al. (2003) *K*-statistics computation starts by computing the ratio between total variance and variance taking phylogeny into account obtained by pGLS (so larger values indicate more signal, because the phylogenetic structure reduces variance). However, this is difficult to interpret (except under a randomization that allows testing the null hypothesis of absence of signal and randomness of trait values), and Blomberg et al. (2003) proposed that this ratio can be scale in respect to the expect ratio under Brownian motion, and the ratio of ratio gives *K*-statistics. *K*-statistics thus larger than 1 indicate that trait is evolving faster than Brownian motion, which as shown before when explaining the PSR curves, occurs for example when two clades diverge a lot (adapting to distinct environments, for example) even if their divergence time is not that large.

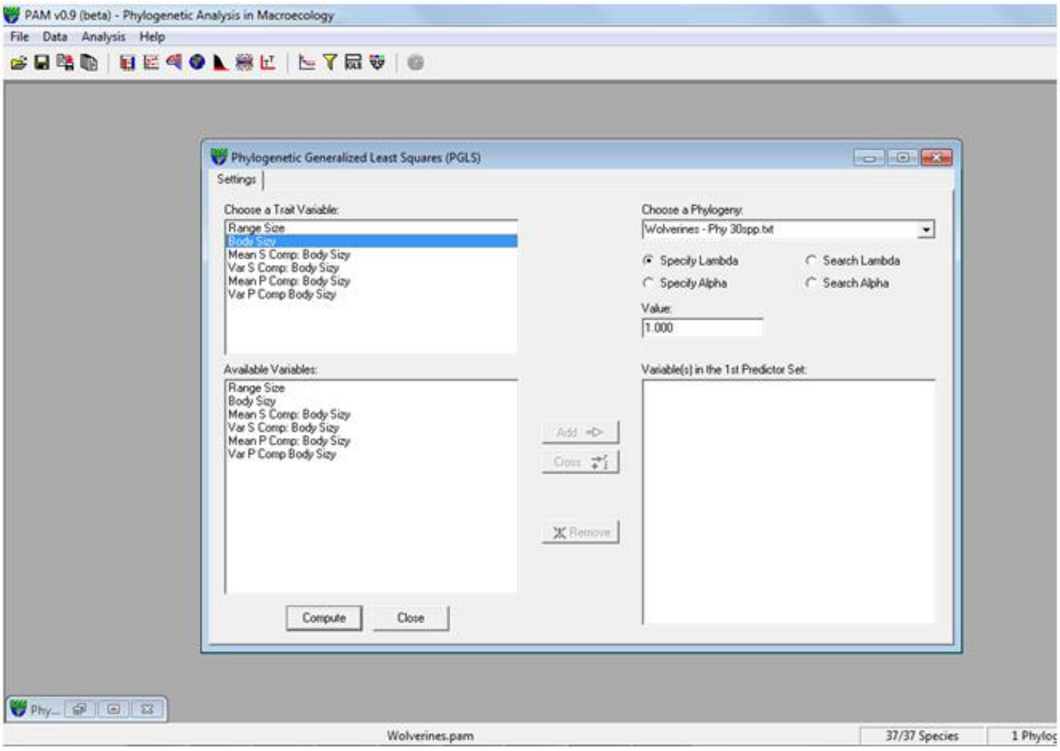
Pagel's λ follows a similar logic, but varies only between 0 and 1. The λ is actually the scalar that multiplies the extra-diagonal covariance or correlation matrices among species, thus also increasing the stemmness of the phylogeny. Several pGLS are then fitted and likelihoods obtained, so the λ fitted to the trait is actually the one that gives the matrix that maximize the likelihood of pGLS.

Finally, to deal with trait variation it is also possible to fit another model, the Ornstein-Uhlenbeck (OU) model, in which there is a exponential relationship between trait divergence and time. The parameter of the model is the α, which is the restraining force pushing the trait values to the adaptive peak. High values of α indicate that restraining force is high and traits are always close to the adaptive peak regardless of the phylogenetic structure (so phylogenetic signal is low).

The pGLS can also be used in a regression context, to evaluate patterns of correlated evolution among traits. As pointed out by Martins & Hansen (1997), if residuals of a regression across species are phylogenetically autocorrelated, two solutions are possible: 1) to incorporate another explanatory variable that "captures" the phylogenetic structure in the model, fixing the problem of residual autocorrelation (this is the strategy used in PVR), or; 2) to directly add the phylogenetic structure in the model residuals and fit the model, using the pGLS, taking this into account. The pGLS is very powerful and produce accurate results, although it requires that phylogenetic structure is known in details to achieve optimal performance (see Martins et al. 2002; Freckleton et al. 2012). Notice that it is possible to use the signal statistical (α,*K* or λ) to fit the two-traits pGLS, which gives a lot of flexibility to regress one trait into another taking into account phylogenetic structure.

pGLS in PAM

PAM gives the option to fit all the above models and calculate the various statistics. If a single trait (a response) is selected, the pGLS is fitted as a intercept-only model, which is the basis for calculating both *K* and λ. In this case, the default option is to estimate or search for these statistics using maximum likelihoods (in the case of λ). However, it is also possible to give a fixed value and obtain the likelihoods. This allows one, for example, to compare the pGLS fit (the likelihood) when λ = 0, or when λ = 1, and compare explicitly with the observed likelihood (and thus deciding if trait behaves more like random or Brownian, respectively). This also allows comparing models with increasingly α values, such as OU and Brownian models.



PAM v0.9 (beta) - Phylogenetic Analysis in Macroecology
File
Data
Analysis
Help
Phylogenetic Generalized Least Squares (PGLS)
Settings
Add
Cross
Remove
Compute
Close
Phy...

Choose a Trait Variable:
Range Size
Body Sizy
Mean S Comp: Body Sizy
Var S Comp: Body Sizy
Mean P Comp: Body Sizy
Var P Comp Body Sizy

Available Variables:
Range Size
Body Sizy
Mean S Comp: Body Sizy
Var S Comp: Body Sizy
Mean P Comp: Body Sizy
Var P Comp Body Sizy

Choose a Phylogeny: Wolverines - Phy 30spp.txt

● Specify Lambda ○ Search Lambda
○ Specify Alpha ○ Search Alpha

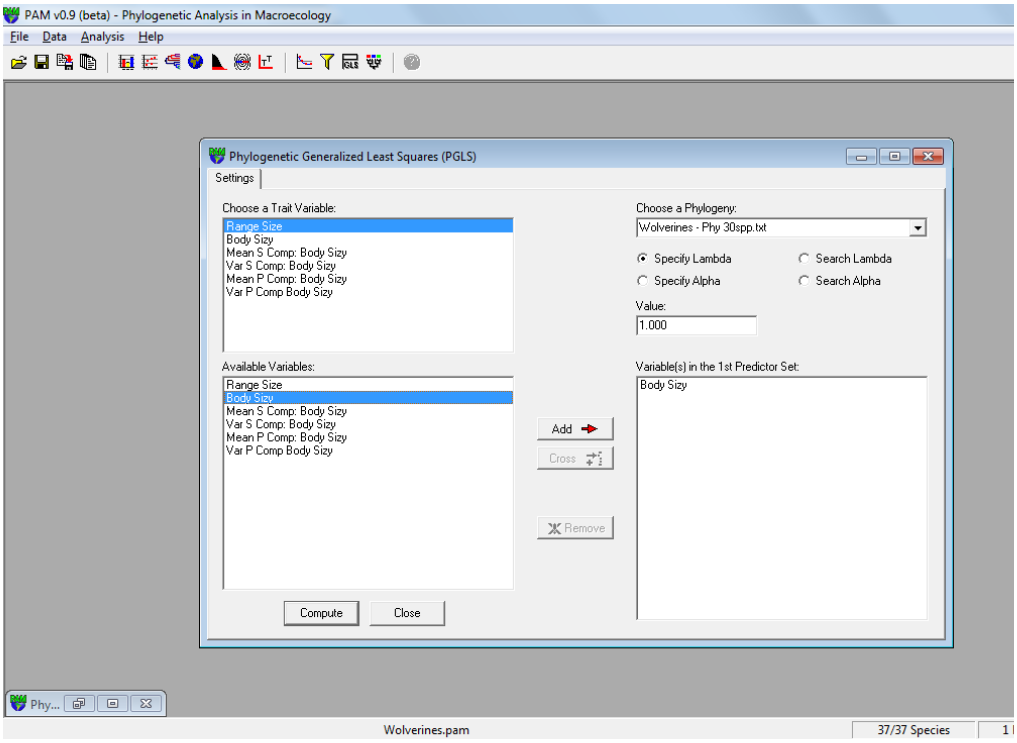
Value: 1.000

Variable(s) in the 1st Predictor Set:

Wolverines.pam | 37/37 Species | 1 Phylog

Notice that, as PUTs are "on" in PAM, there is no single K or λ, but rather a distribution of these statistics, and likelihoods in the case you are fitting a model defined *a priori.*

If two traits are selected, then the pGLS works like a multiple regression and is designed to test the correlated evolution between traits. It is useful to compare this with a "non-phylogenetic" regression (such as obtained in PAM's scatterplots or in any other software). Notice that it is possible to search for the "best" underlying phylogenetic structure, using α, *K* or λ (fixing the values or allowing PAM to estimate them).



PAM v0.9 (beta) - Phylogenetic Analysis in Macroecology
File
Data
Analysis
Help
Phylogenetic Generalized Least Squares (PGLS)
Settings
Add
Cross
Remove
Compute
Close
Phy...
Wolverines.pam
37/37 Species
1

Choose a Trait Variable:
Range Size
Body Sizy
Mean S Comp: Body Sizy
Var S Comp: Body Sizy
Mean P Comp: Body Sizy
Var P Comp Body Sizy

Choose a Phylogeny: Wolverines - Phy 30spp.txt

● Specify Lambda ○ Search Lambda
○ Specify Alpha ○ Search Alpha

Value: 1.000

Available Variables:
Range Size
Body Sizy
Mean S Comp: Body Sizy
Var S Comp: Body Sizy
Mean P Comp: Body Sizy
Var P Comp Body Sizy

Variable(s) in the 1st Predictor Set:
Body Sizy

Exercise 8 - *Explore the several aspects of the pGLS module, including both metrics for phylogenetic signal (K and* λ) and correlated evolution by regressing range size on body size. *One of the most interesting things to do here is to compare the likelihood of the models, either fitting a single trait or regressing one trait into another.*

REFERENCES

Blomberg, S. P. and Garland Jr, T. (2002) Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology*, **15**, 1-12.

Blomberg, S. P., Garland Jr., T. & Ives, A. R. (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution,* **57,** 717-745.

Cheverud, J. M., Dow, M . M. & Leutenegger, W. (1985) The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight in primates. *Evolution*, **39**, 1335–1351.

Desdevises, Y., Legendre, P., Azouzi, L. & Morand, S. (2003) Quantifying phylogenetically structured environmental variation. *Evolution,* **57**, 2647–2652.

Diniz-Filho, J. A. F. (2001) Phylogenetic autocorrelation under distinct evolutionary processes. *Evolution,* **55**, 1104-1109.

Diniz-Filho, J. A. F. & Tôrres, N. M. (2002) Phylogenetic comparative methods and the geographic range size-body size relationship in New World terrestrial Carnivora. *Evolutionary Ecology,* **16**, 351-367.

Diniz-Filho, J. A. F., De Sant'ana, C. E. R. & Bini, L. M. (1998) An eigenvector method for estimating phylogenetic inertia. *Evolution,* **52**, 1247–1262.

Diniz-Filho, J. A. F., Bini, L. M., Rodriguez, M. A., Rangel, T. F. L. V. B. & Hawkins, B. A. (2007) Seeing the forest for the trees: partitioning ecological and phylogenetic components of Bergmann's rule in European Carnivora. *Ecography*, **30**, 598-608.

Diniz-Filho, J.A.F., Rodríguez, M. Á., Bini, L. M., Olalla-Tárraga, M. A., Cardillo, M., Nabout, J. C., Hortal, J. & Hawkins, B. A. (2009) Climate history, human impacts and global body size of Carnivora (Mammalia: Eutheria) at multiple evolutionary scales. *Journal of Biogeography,* **36**, 2222-2236.

Diniz-Filho, J.A.F., Bini, L.M., Rangel, T.F., Morales-Castilla, I., Ollala-Tarraga, M.A., Rodríguez, M.A. & Hawkins, B.A. (2012a) On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography*, **35**, 239-249.

Diniz-Filho, J.A.F., Rangel, T.F., Santos,T. & Bini, L.M. (2012b) Exploring patterns of interspecific variation in quantitative traits using sequential phylogenetic eigenvector regression. *Evolution,* **66**, 1079-1090.

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

Freckleton, R. P., Harvey, P. H. & Pagel, M. (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.*, **160**, 712–726.

Freckleton, R. P., Cooper, N. & Jetz, W (2011) Comparative method as a statistical fix: the dangers of ignoring evolutionary models*. Am.Nat.,* **178**, E10-E17.

Gittleman, J. L. & Kot, M. (1990) Adaptation: Statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, **39,** 227-241.

Griffith, D. A. & Peres-Neto, P. R. (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology,* **87**, 2603–2613.

Harvey, P. & Pagel, M. (1991) *The comparative method in evolutionary biology.* Oxford Univ.Press, Oxford.

Hansen, T. F. & Martins, E. P. (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution,* **50**, 1404–1417.

Helmus, M. R., Bland, T. J., Williams, C. K. & Ives, A. R. (2007) Phylogenetic measures of biodiversity. *American Naturalist*, 169, E68–E83.

Hof, C., Rahbek, C. & Araújo, M. B. (2010) Phylogenetic signals in the climatic niches of world's amphibians. *Ecography,* **33**, 242-250.

Housworth, E. A. and Martins, E. P. 2001. Random sampling of constrained phylogenies: conducting phylogenetic analyses when the phylogeny is partially known. *Systematic Biology*, **50**, 628-639.

Martins, E. P. (1996) Conducting phylogenetic comparative studies when the phylogeny is not known. *Evolution*, **50**,12-22.

Martins EP and Garland Jr T (1991) Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution,* **45,** 534-557.

Martins, E. P. and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporate phylogenetic information into the analysis of interspecific data. Am. Nat. 149: 646-667.

Martins, E. P., Diniz-Filho, J. A. F. & Housworth, E. A. (2002) Adaptive constraints and the phylogenetic comparative method: a computer simulation test. *Evolution,* **56**, 1-13.

Pavoine, S., Ollier, S., Pointier, D. & Chessel, D. (2008) Testing for phylogenetic signal in phenotypic traits: new matrices of phylogenetic proximities. *Theoretical Population Biology*, **73**, 79-91.

Rangel, T.F.L.V.B., Diniz-Filho, J.A.F. & Bini, L.M. (2006) Towards an integrated computational tool for spatial analysis in macroecology and biogeography. *Global Ecology and Biogeography*, **15**, 321-431.

Rangel, T. F. L. V. B., Diniz-Filho, J.A.F. & Bini, L.M. (2010) SAM: A comprehensive application for spatial analysis in macroecology. *Ecography,* **33**, 46-50.

Ricklefs, R. E. (2006) The unified neutral theory of biodiversity: do the numbers add up? *Ecology,* **87**, 1424-1431.

Rohlf, F. J. (2001) Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution*, **55,** 2143–2160.

Safi, K. & Pettorelli, N. (2010) Phylogenetic, spatial and environmental components of extinction risk in carnivores. *Global Ecology and Biogeography*, **19**, 352-362.

Schweiger, O., Klotz, S., Durka, W. & Kuhn, I. (2008) A comparative test of phylogenetic diversity indices. *Oecologia,* **157**, 485-495.

Webb, C. O., Ackerly, D. D., McPeek, M. A. & Donoghue, M. J. (2002) Phylogenies and community ecology. *Annual Review of Ecology and Systematics,* **33**, 475-505.