Putting phylogeny into the analysis of biological traits: A methodological approach

Thibaut Jombart, Sandrine Pavoine, Sébastien Devillard, Dominique Pontier

Putting phylogeny into the analysis of biological traits: A methodological approach

Thibaut Jombart, Sandrine Pavoine, Sébastien Devillard, Dominique Pontier

PII: S0022-5193(10)00173-6
DOI: doi:10.1016/j.jtbi.2010.03.038
Reference: YJTBI 5939

To appear in: *Journal of Theoretical Biology*

Received date: 20 January 2010
Revised date: 25 March 2010
Accepted date: 25 March 2010

Cite this article as: Thibaut Jombart, Sandrine Pavoine, Sébastien Devillard and Dominique Pontier, Putting phylogeny into the analysis of biological traits: A methodological approach, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.03.038

Putting phylogeny into the analysis of biological traits: a methodological approach

Thibaut Jombart[a,], Sandrine Pavoine[b], Sébastien Devillard[c], Dominique Pontier[c]

[a] *MRC Centre for Outbreak Analysis & Modelling, Department of Infectious Disease Epidemiology, Imperial College London, Faculty of Medicine, Norfolk Place, London W2 1PG, UK.*
[b] *Museum National d'Histoire Naturelle; Département Ecologie et Gestion de la Biodiversité; UMR 7204 MNHN-CNRS-UPMC; CRBPO, 61 rue Buffon, 75005, Paris, France*
[c] *Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France.*

Abstract

Phylogenetic comparative methods have long considered phylogenetic signal as a source of statistical bias in the correlative analysis of biological traits. However, the main life-history strategies existing in a set of taxa are often combinations of life history traits that are inherently phylogenetically structured. In this paper, we present a method for identifying evolutionary strategies from large sets of biological traits, using phylogeny as a source of meaningful historical and ecological information. Our methodology extends a multivariate method developed for the analysis of spatial patterns, and relies on finding combinations of traits that are phylogenetically autocorrelated. Using extensive simulations, we show that our method efficiently uncovers phylogenetic structures with respect to various tree topologies, and remains powerful in cases where a large majority of traits are not phylogenetically structured. Our methodology is illustrated using empirical data, and implemented in the free software R.

Keywords:
phylogenetic principal component analysis, pPCA, autocorrelation, multivariate, comparative method, phylogenetic signal

Email address: tjombart@imperial.ac.uk (Thibaut Jombart)

1. Introduction

Phylogeny has long been recognised as a major source of biological variation. For instance, Gregory (1913) and Osborn (1917) considered that species' variability should be partitioned between *heritage* (*i.e.*, phylogenetic inertia) and *habitus* (*i.e.*, adaptation). In their well-known criticism of the adaptationist paradigm, Gould and Lewontin (1979) underlined the importance of the constraints imposed by the phylogeny to the variability observed among organisms. In comparative studies, the effect of phylogeny has merely been perceived as a source of nuisance, since it reveals non-independence among trait values observed in taxa (Dobson, 1985; Felsenstein, 1985), and thus violates one of the basic assumptions required by most statistical tools (Harvey and Pagel, 1991).

Phylogenetic comparative methods (PCM) were especially designed to solve this problem. Various methods have been developed that transform quantitative traits into new variables that are not correlated to phylogeny, according to a given model of evolution. For instance, phylogenetic independent contrasts (PIC, Felsenstein, 1985) transform values observed at the n tips of a phylogeny into $(n-1)$ node values that are not phylogenetically autocorrelated under a Brownian motion model. Generalised least squares (GLS, Grafen, 1989; Rohlf, 2001) is a more general technique that allows specifying the autocorrelation of observations as a component of a linear model. This approach can therefore account for the non-independence among observations using a wide variety of models of evolution (Hansen and Martins, 1996). As stressed by Rohlf (2006), these approaches do not actually remove phylogenetic autocorrelation from the data, but merely take it into account to provide more accurate estimates of model parameters. In fact, PIC, GLS, along with other existing PCM all aim towards the same goal: 'correcting for phylogeny' in the correlative analysis of biological traits at the species level (Harvey and Purvis, 1991; Martins, 2000; Martins et al., 2002; Garland et al., 2005).

Nonetheless, studying the phylogenetic patterns of trait variation allows formation of hypotheses about the evolutionary pathways that led to the trait values of extant species. It also allows shedding light onto the influence of historical and ecological processes on community assembly (Webb et al., 2002). Many biologically meaningful patterns are inherently structured with phylogeny. Indeed, many life-history and ecological strategies are likely to be phylogenetically structured (Webb et al., 2002). Inheritance from a common ancestor and phylogenetic inertia (*i.e.*, constraints to evolution) may cause phylogenetic signal (similar trait values across closely related species) to occur. Other factors leading to phylogenetic signals in traits act at the population level rather than at the species level such as high gene flow, lack of genetic variation, stabilising selection if changes in trait states reduce fitness, or population growth if traits are pleiotropically linked to other traits that reduce fitness (Wiens and Graham, 2005). However, traits might also be affected by variations unrelated to the phylogeny, but relating to ecological conditions experienced by the species. For instance, biotic interactions might drive character displacement and abiotic interactions might lead to trait convergence. From this perspective, phylogenetic signal becomes a source of precious biological information that can be used to identify historical as well as recent evolutionary strategies. Interestingly, a similar paradigm shift occurred in spatial ecology (Legendre, 1993) when it was pointed out that spatial patterns in species' distribution were not only sources of spurious correlations, but also indicators of critical ecological structures such as localised species assemblages and species-environment associations. This paradigm shift proved particularly fecund and still motivates innovative developments in statistical ecology (*e.g.*, Dray et al., 2006; Griffith and Peres-Neto, 2006).

In this paper, we present a method which uses phylogenetic information to uncover the main phylogenetic structures observable in multivariate data associated with a phylogeny. Our approach, *phylogenetic principal component analysis* (pPCA), extends a methodology developed in spatial ecology (Dray

et al., 2008) and in spatial genetics (Jombart et al., 2008) to the analysis of phylogenetic structures in biological features of taxa such as life-history traits. We emphasise that phylogenetic structures can be measured and quantified in the same way as spatial structures, as they are both associated with the concept of autocorrelation. We then define different kinds of phylogenetic structures, and show how pPCA can be used to identify them. After evaluating the ability of pPCA to uncover phylogenetic patterns through extensive simulations, we illustrate our method using an empirical example. pPCA is implemented in the *adephylo* package (Jombart and Dray, 2009) for the free software R (R Development Core Team, 2009).

2. Methods

2.1. Measuring phylogenetic autocorrelation

Phylogenetic autocorrelation is said to occur whenever the values taken by a set of taxa for a given biological trait are not independent of the phylogeny. Frequently, closely related taxa exhibit more similar traits than randomly-chosen taxa. Moran's I, an index originally used to measure spatial autocorrelation (Moran, 1948, 1950), has been proposed for measuring phylogenetic autocorrelation (Gittleman and Kot, 1990). Adapting the former definition (Cliff and Ord, 1973, p13) to the phylogenetic context, I is defined as:

$$I_{\mathbf{W}}(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{W}\mathbf{x}}{n}\frac{1}{\mathrm{var}(\mathbf{x})} \quad (1)$$

where $\mathbf{x}$ is the centred vector of a trait observed on n taxa, $\mathrm{var}(\mathbf{x})$ is the usual variance of $\mathbf{x}$, and $\mathbf{W}$ is a matrix of phylogenetic proximities among taxa ($\mathbf{W} = [w_{ij}]$ with $i, j = 1, \ldots, n$), whose diagonal terms are zero ($w_{ii} = 0$), and rows sum to one ($\sum_{j=1}^{n} w_{ij} = 1$). The null value, *i.e.* the expected value when no phylogenetic autocorrelation arises, is $I_0 = -1/(n-1)$ (Cliff and Ord, 1973). In its initial formulation (Gittleman and Kot, 1990), *i.e.* before row standardisation so that $\sum_{j=1}^{n} w_{ij} = 1$, $\mathbf{W}$ contained binary weights. Before this standardisation, the entry at row i and column j was set to 1 if taxon i shared

a common ancestor with taxon j at a given taxonomic level, and to 0 otherwise. Hence, taxa were considered as either phylogenetically related or not. Moran's I then compared the trait value of a taxon to the mean trait value in related taxa to detect phylogenetic autocorrelation.

Such binary relationships are clearly not sufficient to model the possibly complex structure of proximities among taxa induced by the phylogeny. To achieve better resolution in these comparisons, we propose using as entries of $\mathbf{W}$ any measurement of phylogenetic proximity valued in $\mathbb{R}^+$ verifying:

$$\begin{cases} w_{ij} \geqslant 0 & \forall\ i,j = 1,\ldots,n \\ w_{ii} = 0 & \forall\ i = 1,\ldots,n \\ \sum_{j=1}^{n} w_{ij} = 1 & \forall\ i = 1,\ldots,n \end{cases} \tag{2}$$

Then, Moran's I compares the value of a trait in one taxon (terms of $\mathbf{x}$) to a weighted mean of other taxa states (terms of $\mathbf{Wx}$) in which phylogenetically closer taxa are given stronger weights. This extension gives the index considerable flexibility for quantifying phylogenetic autocorrelation, as phylogenetic proximities can be derived from any model of evolution (including or not branch lengths). For instance, one interesting possibility would be using the covariance matrix estimated in a GLS model (Grafen, 1989) to define phylogenetic proximities. This could be achieved by setting diagonal terms (variances) of the covariance matrix to zero, adding the smallest constant ensuring that all terms are positive, and row-standardizing the resulting matrix.

This formulation of Moran's I also relates the index to other PCM. For instance, the test proposed by Abouheif (1999), initially based on the many possible planar representations of a tree, turned out to be a Moran's I test using a particular measure of phylogenetic proximity for $\mathbf{W}$ (Pavoine et al., 2008).

Moran's I is also related to autoregressive models. In their simplest form,

these models are written as (Cheverud and Dow, 1985; Cheverud et al., 1985):

$$\mathbf{x} = \rho\mathbf{W}\mathbf{x} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e} \tag{3}$$

where ρ is the autocorrelation coefficient, $\mathbf{Z}$ is a matrix of explanatory variables, $\boldsymbol{\beta}$ is the vector of coefficients, and $\mathbf{e}$ is a vector of residuals. The matrix of phylogenetic relatedness $\mathbf{W}$ (Cheverud and Dow, 1985; Cheverud et al., 1985) is exactly the weight matrix of our definition of Moran's I (equation 1). The essential difference between the two approaches is that autoregressive models perform the regression of $\mathbf{x}$ onto $\mathbf{W}\mathbf{x}$, while I computes the inner product between both vectors (numerator of equation 1) to measure phylogenetic autocorrelation.

Lastly, the weighting matrix $\mathbf{W}$ is also the core of another approach producing variables that model phylogenetic structures (Peres-Neto, 2006). Like Moran's I, this approach was initially developed in spatial statistics (Griffith, 1996), and consisted in finding eigenvectors of a doubly centered spatial weighting matrix (Dray et al., 2006). Applied to a matrix of phylogenetic proximity $\mathbf{W}$, this method yields uncorrelated variables modeling different observable phylogenetic patterns, each related to a value of Moran's I. Peres-Neto (2006) performed the regression of a variable $\mathbf{x}$ onto these eigenvectors to partial-out the phylogenetic autocorrelation from $\mathbf{x}$. Alternatively, we suggest using these eigenvectors to simulate what we further call 'global' and 'local' phylogenetic structures.

2.2. Global and local phylogenetic structures

Phylogenetic autocorrelation relates to the non-independence of trait values observed in taxa given their phylogenetic proximity. There are two ways in which this non-independence can arise, depending on whether closely related taxa tend to have more similar, or more dissimilar trait values than expected at random, resulting in *positive* and *negative autocorrelation*, respectively. Positive phylogenetic autocorrelation most often results in global patterns of similarity in related taxa; we thus refer to these patterns as *global structures*. Global patterns reflect the general idea of phylogenetic signal: trait values observed in a set

of taxa are not independent, but tend to be more similar in closely related taxa (*e.g.*, Figure 1A). Most common explanations for this phenomenon are inheritance from a common ancestor, or the conservation of ecological niches (Harvey and Pagel, 1991).Traits whose evolution can be modeled by a Brownian or by an OU process with low stabilising constraint generally display global patterns (Abouheif, 1999; Pavoine et al., 2008). Such phenomenon typically results in close-to-the-root divergence in evolutionary strategies.

Conversely, negative phylogenetic autocorrelation corresponds to dissimilarities among tips localised in specific parts of the tree, which we call *local structures*. A local structure would be observed whenever closely related taxa tend to be more different with respect to a given trait than randomly chosen taxa (*e.g.*, Figure 1E). Local structures correspond to relatively recent events that induced divergence of the evolutionary strategies close to the tips of the phylogenetic tree, such as convergence and character displacement (following past or present biotic interactions). This also occurs when the trait under study has been selected towards different optimal values, resulting in opposed evolutionary strategies being observed in sister taxa.

Both structures can be identified using Moran's index (equation 1). The sign of I depends on how values of a trait ($\mathbf{x}_i$) relate to the values observed on closely related taxa ($\mathbf{W}\mathbf{x}_i$). Moran's I will be greater than (respectively less than) I_0 (value of I in the absence of autocorrelation) when closely related taxa tend to have similar (respectively dissimilar) values for the studied trait. Obviously, the definition of phylogenetic proximities in $\mathbf{W}$ will condition the measurement of global and local structures. As shown by Pavoine et al. (2008), not all phylogenetic proximities are equal in detecting phylogenetic structures. Especially, the phylogenetic proximities underlying Abouheif's test (matrix $\mathbf{A} = [a_{ij}]$ in Pavoine et al. 2008) proved superior to several common phylogenetic proximities for testing phylogenetic inertia in traits simulated under Brownian and Ornstein-Uhlenbeck (OU) processes. More generally, the matrix $\mathbf{W}$ can be derived from any model of evolution which seems appropriate to the data, taking branch lengths into account whenever these are accurately estimated, and rely-

ing only on the topology in other cases.

Ecological and life-history strategies of species require not one, but several traits to be adequately described. Accordingly, ecological and life-history strategies are likely to involve combinations of traits with both global and local phylogenetic structures. In the following, we describe a methodology which explicitly investigates multivariate phylogenetic structures that have barely been considered so far.

2.3. The phylogenetic principal component analysis

Dray et al. (2008) and Jombart et al. (2008) developed a multivariate approach for identifying spatial structures in multivariate data. Essentially, this approach consists in constraining the principal components of a multivariate method to exhibit spatial autocorrelation, as measured by Moran's index. This methodology proved better at detecting autocorrelated patterns than usual multivariate methods such as principal component analysis (Dray et al., 2008; Jombart et al., 2008). Here, we use the same rationale to define the *phylogenetic principal component analysis* (pPCA), a method designed to summarise a set of traits into a few synthetic variables exhibiting global or local phylogenetic structures. Note that while we presented pPCA for the analysis of quantitative traits for the sake of simplicity, this approach can be extended to qualitative traits, or even to mixtures of quantitative and qualitative variables (Dray et al., 2008).

We denote $\mathbf{X} = [x_{ij}]$ ($\mathbf{X} \in \mathbb{R}^{n \times p}$) a matrix containing p quantitative traits measured on n taxa, and $\mathbf{W}$ a matrix of phylogenetic weights used in the computation of Moran's I (equation 1). As in classical PCA, missing data can be set to the mean of the corresponding trait, which does not add artefactual structures to the analyzed traits. Without loss of generality, we assume that traits are centered (*i.e.*, $\sum_i x_{ij} = 0$ with $j = 1, \ldots, p$). The purpose of pPCA is to find linear combinations of traits (columns of $\mathbf{X}$) containing a large variance and displaying global or local phylogenetic structures. Mathematically, this problem

translates into finding the appropriate loadings $\mathbf{u} \in \mathbb{R}^p$ (with $\|\mathbf{u}\|^2 = 1$) that minimise and maximise, respectively, the function:

$$\begin{aligned} f: \quad & \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times n} \times \mathbb{R}^p \quad \longrightarrow \mathbb{R} \\ & (\mathbf{X}, \mathbf{W}, \mathbf{u}) \qquad\quad \longmapsto \mathrm{var}(\mathbf{Xu}) I_{\mathbf{W}}(\mathbf{Xu}) \end{aligned} \tag{4}$$

The solution to this problem is given by the diagonalisation of the matrix $\frac{1}{2n}\mathbf{X}^T(\mathbf{W} + \mathbf{W}^T)\mathbf{X}$ (Dray et al., 2008; Jombart et al., 2008). It results in a set of loadings $\{\mathbf{u_1}, \ldots, \mathbf{u_k}, \ldots, \mathbf{u_r}\}$ with $\mathbf{u_k} \in \mathbb{R}^p$ forming linear combinations of traits ($\mathbf{Xu}_k$, the so-called principal components) associated with decreasing eigenvalues λ_k, so that:

$$\mathrm{var}(\mathbf{Xu}_k) I_{\mathbf{W}}(\mathbf{Xu}_k) = \lambda_k \tag{5}$$

The largest eigenvalues likely correspond to a large variance and a strong positive I, indicating global structures (close-to-root variation in trait states). Conversely, the lowest (*i.e.*, most negative) eigenvalues correspond to a high variance and a large negative I, indicating local structures (close-to-tips variation in trait states). As in other reduced space ordinations, the eigenvalues indicate the amount of structure expressed by each synthetic variable. A sharp decrease in the screeplot is likely to indicate a shift between strong and weak structures. The amount of variance ($\mathrm{var}(\mathbf{Xu}_k)$) and phylogenetic autocorrelation ($I(\mathbf{Xu}_k)$) in each principal component ($\mathbf{Xu}_k$) can be computed for a better interpretation of each structure. Moreover, the loadings $\mathbf{u}_k$ can be used to assess how traits contribute to a given principal component, and thus understand the nature of the corresponding biological structure.

One important choice is that of the phylogenetic weights ($\mathbf{W}$) used in the analysis. Here, we use the measure of phylogenetic proximity underlying the test of Abouheif (1999) to define $\mathbf{W}$, because of its good performances at detecting phylogenetic structures (Pavoine et al., 2008). The phylogenetic proximity a_{ij} among tips i and j is defined as:

$$a_{ij} = \frac{1}{\prod_{p \in P_{ij}} dd_p} \text{ for } i \neq j \tag{6}$$

where P_{ij} is the set of internal nodes on the shortest path from tips i to j, and dd_p is the number of direct descendants from the internal node p. The phylogenetic proximity a_{ij} defines the entries of the off-diagonal terms of $\mathbf{W}$, all diagonal entries being set to 0. As $\mathbf{W}$ is row-standardised, it is defined as:

$$w_{ij} = \frac{a_{ij}}{\sum_{j=1, i \neq j}^{n} a_{ij}} \tag{7}$$

2.4. Sensitivity study

Extensive simulations were conducted to evaluate the sensitivity of the pPCA to various parameters. Datasets were simulated with different characteristics concerning the type of tree, the tree size, the type, strength and numbers of phylogenetically structured traits, and the total number of traits (including structured and unstructured traits). These parameters are summarised in Table 1. Five types of trees of 16, 32, or 128 tips were simulated to encompass a wide range of tree topologies and sizes: completely symmetric trees (Figure 1A), trees obtained by random clustering of tips (as implemented by `rtree` function of the ape package, Paradis et al., 2004, Figure 1B), the Yule model (Yule, 1924, Figure 1C), the biased model (Kirkpatrick and Slatkin, 1993, Figure 1D), and completely asymmetric trees (Figure 1E). Datasets including random traits and phylogenetically structured traits (*i.e.*, displaying global and/or local structures) were obtained for each tree. Random traits were drawn from a normal distribution ($\mathcal{N}(0, 1)$), while 'structured traits' were obtained by adding normally-distributed random noise to phylogenetic eigenvectors of $\mathbf{W}$ (Peres-Neto, 2006). Whenever several structures of the same type (global or local) were simulated in a given dataset, these were derived from the same eigenvector, so that we could evaluate the performance of pPCA when a 'consensus' phylogenetic signal exists in a set of traits (*e.g.*, Figure 1B). This was consistent with the fact that several phylogenetically structured traits are expected to exhibit the same patterns, either because these structures are caused by the same evolutionary process, or because all traits are correlated to another phylogenetically structured trait. 200 datasets were simulated for each of the 810

different combinations of parameters, resulting in a total of 162 000 datasets.

All simulations were performed in the R software (R Development Core Team, 2009). Trees were simulated using the R packages *ape* (Paradis et al., 2004) and *apTreeshape* (Bortolussi et al., 2006), and scripts developed by TJ for the symmetric model. Structured traits were simulated using the *ade4* package (Chessel et al., 2004; Dray et al., 2007), and data were handled using the *phylobase* package (Bolker et al., 2007).

Each dataset was analysed by a pPCA using Equation 7 to define phylogenetic proximities. In each analysis, the structured traits were compared to the first relevant (global and/or local) principal component of pPCA, to assess how the method performed. The strength of the link between the original simulated structures and patterns identified by pPCA was measured using the absolute value of Spearman's rank correlation, $|\rho|$. Whenever the dataset included several distinct structured traits, $|\rho|$ values were averaged by type of structure (*i.e.*, global or local). Hence, we obtained one or two $|\rho|$ per simulated dataset, used as indicator of the performance of pPCA ($|\rho|$ close to one = high performance, $|\rho|$ close to zero = low performance).

Variations in $|\rho|$ according to the different simulation parameters were investigated using a linear model. The relationship between $|\rho|$ and the predictors was linearised using a logit link, *i.e.* using $\text{logit}(|\rho|) = \log\frac{|\rho|}{1-|\rho|}$ as the response variable. When interpreting coefficients of the model, predictions $\hat{\mu}$ were re-transposed onto the $|\rho|$ scale, that is, replacing $\hat{\mu}$ by $\frac{1}{1+e^{-\hat{\mu}}}$. Qualitative variables were modeled using treatment-coded contrasts (Faraway, 2004, p. 173). The explanatory variables were the type of tree (factor 'tree', the biased model being the intercept), the type of structuring (factor 'strutype', with level 'global' at the intercept), the number of tips ('ntips', intercept=16), the total number of traits ('ntraits', intercept=10), the standard deviation of the random noise added to structured variables ('noise', intercept=0.5), and the number of

structures (factor 'nstruc', intercept=1).

2.5. Empirical data analysis

We illustrate pPCA with the data on reproductive and morphometric traits within a small group of species and subspecies from the lizard family Lacertidae published by Bauwens and Diaz-Uriarte (1997). These data are currently available as the dataset *lizards* in the *adephylo* package (Jombart and Dray, 2009) for R (R Development Core Team, 2009). They consist in a molecular phylogeny and 8 life-history traits measured for 16 taxa: mean adult length (in mm, abbreviated *mean.L*), length at maturity (in mm, *matur.L*), maximum length (in mm, *max.L*), hatchling length (in mm, *hatch.L*), hatchling mass (in g, *hatch.m*), clutch size (in number of descendents, *clutch.S*), clutch frequency (in number of events per year, *clutch.F*), and age at maturity (in number of months of activity, *age.mat*). All traits were measured on females. Adult life span and egg size were discarded from the analysis because data were missing for several taxa.

The analyses were conducted in the R software (R Development Core Team, 2009), using the *ade4* package (Chessel et al., 2004; Dray et al., 2007) for factorial analyses and *adephylo* (Jombart and Dray, 2009) to perform the pPCA. As in Bauwens and Diaz-Uriarte (1997), data were log-transformed and regressed onto mean adult female length to partial out the body size effect. As a consequence, mean adult female length was removed from the analysis. We then investigated phylogenetic structures of the transformed life-history traits using pPCA.

3. Results

3.1. Sensitivity study

All explanatory variables had a very significant effect on the response variable (Appendix A, Table A.1), which was trivial because even very low effects might be significant with a large number of observations. All coefficients of

the model (Appendix A, Table A.2) have therefore been interpreted quantitatively to determine the level of effects of the explanatory variables. The model explained satisfyingly 57% of the total variance. Overall, the average $|\rho|$ was relatively high ($CI_{99\%} = [0.667; 0.669]$), showing that phylogenetic structures were well retrieved by pPCA. The strongest effect was by far that of the type of structure: global patterns were more easily retrieved than local structures, with a difference of 0.31 in predicted $|\rho|$ (later denoted $\Delta_{|\hat{\rho}|}$). pPCA performed better in larger trees ($\Delta_{|\hat{\rho}|} = 0.11$ between trees with 16 and 128 tips), suggesting that phylogenetic signal is more easily captured when a large number of taxa is available, which is in line with previous findings for a different PCM (Martins and Hansen, 1997). For a given number of structured traits, the number of random traits slightly lowered the method's ability to retrieve phylogenetic patterns ($\Delta_{|\hat{\rho}|} = 0.10$ between 10 and 50 traits). Phylogenetic structures incorporating larger amounts of random noise were also more difficult to retrieve ($\Delta_{|\hat{\rho}|} = 0.10$ between noise of 0.5 and 1). Lastly, the number of structured traits and the type of tree only marginally affected the ability of pPCA to identify phylogenetic patterns.

3.2. Empirical data analysis

Both global and local phylogenetic structures were found by pPCA in the lacertid lizards data (Figure 2). The first global principal component of pPCA first opposed a lineage with three species (*Lacerta schreiberi*, *L. agilis*, *L. vivipara*) having the largest negative scores to the rest of the tree (taxa with positive scores, or scores closer to zero), with the subspecies *Podarci h. h.* exhibiting the most opposite life histories (Figure 2A). Among the remaining species or subspecies, *Lacerta monticola cantabrica* and *L. m. cyreni* were distinctive by their negative scores. The loadings of the analysis (Figure 2B) provided further insights on the corresponding evolutionary strategies, and showed a trade-off between the frequency of reproductive events per year (clutch.F) and the clutch size (clutch.S). *L. schreiberi*, *L. agilis*, *L. vivipara*, and to a lesser extent *L. m. cantabrica* and *L. m. cyreni*, reproduce less often but produce a larger number

of eggs per reproductive event than other populations. A possible explanation for this structure is that environmental conditions only allow for a few reproduction events in these populations, which then deliver lots of eggs with poor individual survival rates. In contrast, the first local principal component of pPCA (Figure 2A) highlighted a strong opposition between related taxa, especially between *Takydromus tachydromoides* and *Acanthodactylus erythrurus*, but also to a lesser extent between *L. schreiberi* and *L. agilis*. This opposition was also apparent, although weak, within two additional lineages (first *Podarcis muralis*, *P. bocagei* versus *P. h. atrata*, *P. h. hispanica* Asturias; second *L. m. cyreni* versus *L. m. cantabrica*). Figure 2B (vertical axis) shows the meaning of these local variations. The species with positive scores on the first local principal component (especially, *T. tachydromoides* and *L. schreiberi*) produce a large number of small eggs while species with negative scores (especially *A. erythrurus* and *L. agilis*) produce fewer, but larger descendants.

4. Discussion

Phylogenetic autocorrelation has so far been considered as a mere nuisance to the correlative analysis of comparative biological data, when exploring trade-offs as well as in allometric studies. In this paper, we advocate that phylogenetic autocorrelation is a source of relevant biological information for the exploratory analysis of such data. To accomplish this task, we introduced the *phylogenetic principal component analysis* (pPCA), a method that we adapted from existing multivariate spatial statistics (Dray et al., 2008; Jombart et al., 2008) to analyse phylogenetic structures in multivariate sets of traits. Based on the results obtained from simulated and empirical data, we discuss the ability of the pPCA to retrieve phylogenetic signals in a multivariate set of traits, and the impact that this approach could have in evolutionary ecology.

4.1. Performance of the approach - methodological discussion

Preliminary results stemming from our sensitivity study were very promising, and provided guidelines for applications of pPCA. Overall, pPCA performed well to retrieve simulated phylogenetic structures, even in some cases where only 1 out of 50 traits was phylogenetically structured. pPCA seemed to retrieve global phylogenetic structures more easily than local structures. This may be due to the asymmetry of Moran's I distribution, which often has a smaller range of variation in negative values (local structures) than in positive values (global structures) (de Jong et al., 1984). As pPCA seeks principal components with extreme values of I, global structures (associated with large positive I) would be more easily detected than local structures (associated with large negative I). Therefore, local structures may be interpreted even though the corresponding eigenvalue seems negligible compared to global structures, provided it is biologically significant. Other results of our sensitivity study suggest that pPCA performs better in larger trees, although performances on small phylogenies were satisfying. Interestingly, pPCA seemed rather insensitive to the shape of the, indicating that the method can be used with virtually any kind of phylogeny.

Although pPCA will be best appreciated using empirical datasets, further simulation studies may be considered. In this study, we used eigenvectors of a phylogenetic proximity matrix (Peres-Neto, 2006) to simulate phylogenetic structures. This method allows simulation of complex phylogenetic patterns in negligible computational time, which permitted examination of the influence a large number of parameters on pPCA results. The drawback of this approach is that eigenvectors of phylogenetic proximity matrices are not directly related to an model of evolution such as the Brownian motion or the OU models. While current procedures implementing trait simulation under these models are more computer-intensive, it could be possible to study how pPCA behaves under these models, given variation in a few parameters.

The main parameter that should be investigated in further detail is the

phylogenetic proximity used in pPCA. A previous study demonstrated that some phylogenetic proximities were better than others at detecting phylogenetic structures using Moran's index (Pavoine et al., 2008). However, it is likely that the most appropriate measurement of phylogenetic proximity depends on the dataset under scrutiny. To assess whether a given phylogenetic proximity is adapted to a particluar dataset, we advocate to perform Moran's I test using this proximity matrix. Whenever significant structures are detected, one can input this phylogenetic proximity in pPCA to uncover the nature of the underlying phylogenetic structures.

4.2. Potential impacts of the approach in evolutionary and ecological studies

This novel approach should complement nicely the usual PCM toolbox, bringing a new perspective to the analysis of comparative biological data. Contrary to usual PCM, our approach does not attempt to improve estimates of correlations among traits by 'correcting' for phylogenetic dependence among species. Instead, it seeks biologically meaningful combinations of traits that are globally or locally phylogenetically structured, thus allowing us to uncover fundamental evolutionary patterns. As noted by Bauwens and Diaz-Uriarte (1997), *theories of life-history evolution are explicitely micro-evolutionary* [...] *whereas patterns of life-history covariation are most evident when comparisons are made among higher taxonomic levels.* pPCA covers both of these aspects, by providing insights about broad macro-evolutionary patterns (global structures) and more recent, even micro-evolutionary patterns (local structures).

Life histories, for example, are likely to be phylogenetically structured (Gaillard et al., 1989; Pontier et al., 1993; Rochet et al., 2000). In our case study, the pPCA identified phylogenetic patterns in the main life-history tactics adopted by a set of taxa. Our results suggest that the trade-off between clutch frequency and size may have resulted in the ancient divergence of evolutionary strategies. In contrast, the trade-off between hatchling mass on the one hand

and clutch size and frequency on the other hand appears to be more labile, involving more recent character changes in most of the lineages and especially between *T. tachydromoides* and *A. erythrurus*. The pPCA thus allows description of global (close-to-root, phylogenetic signal) versus local (close-to-tips) phylogenetic structures in a multivariate set of traits, and highlights which lineages and which taxa are involved in these structures. Overall, this illustration using an empirical dataset showed that pPCA can bring new insights about evolutionary strategies of a set of taxa. Moreover, whenever a molecular clock is available for the considered phylogeny, it would be possible to estimate the age of the involved lineages and taxa, by dating their most recent common ancestors. This would allow assessing how and when different evolutionary strategies might have appeared and evolved along the history of the considered taxa. Local structures uncovered by pPCA point out more recent evolutionary events, such as speciation caused by diversifying selection or niche separation, and are thus also of fundamental interest. Dating these recent events would be even more interesting as historical information about the considered taxa is more likely to be available for recent speciation events. For instance, we could investigate whether a recent speciation highlighted as local structure would have been preceded by significant modifications of the environment.

A further strength of pPCA lies in its ability to analyse very large sets of traits (*i.e.*, hundreds or thousands of traits) simultaneously. Usual PCM typically rely on pairwise comparisons among traits, which becomes cumbersome when lots of variables are under scrutiny, and often requires discarding traits from the analysis. This issue will be increasingly concerning in the near future as new and large databases of life-history traits will become available. pPCA can be used to explore such data, to unveil evolutionary trade-offs among a large number of traits, without having to make a prior selection of analysed traits. Previous methods have already attempted to analyse phylogenetic signals in a series of traits. These methods determine the proportion of variation in a set of traits correlated with the phylogenetic relatedness among species (Giannini,

2003; Desdevises et al., 2003). They involve factorial analyses, but with the aim of partitioning variation in traits instead of depicting phylogenetic structures. Nevertheless, one step of the Giannini (2003) variation partitioning approach consists of selecting the nodes of the phylogeny that better explain the variation in the trait values. Accordingly, the selected nodes could be used to determine at which depth in the phylogeny the taxa differ in their trait values. However, this selection results from a long series of tests to determine if the differences between the lineages that descend from a given node are significantly responsible for trait variation. The number of tests depends on the size of the phylogenetic tree, with an increasing risk of erroneously significant tests. The pPCA approach thus brings a new optimised way of disentangling the phylogenetic patterns in a set of traits by identifying the lineages and also the combination of traits responsible for global versus local trait variation.

A potential application of the pPCA concerns phylogenetic community ecology (Webb et al., 2002). Phylogenetic clustering in a community (lower phylogenetic diversity than expected by chance in a regional pool of species) merely reflects the simultaneous action of environmental filtering and phylogenetic conservatism. In contrast, distinct, even opposed processes can lead to phylogenetic overdispersion (higher phylogenetic diversity within a community than expected by chance in a regional pool of species). For example, phylogenetically overdispersed communities can arise from (i) limiting similarity and conservative traits or (ii) environmental filters with convergent traits (Kraft et al., 2007). Consequently, knowledge of the evolution of traits is necessary to interpret observed structures in phylogenetic diversity. A difficulty is that the traits involved in environmental filters and those involved in limiting similarity might follow different evolutionary pathways (Emerson and Gillepsie, 2008; Ackerly et al., 2006). Although pPCA does not provide a formal test of phylogenetic conservatism or over-dispersion, it can be used to describe the phylogenetic signal induced by these processes. Therefore, our methodology could be applied to describe the level of phylogenetic signal in sets of traits from labile traits with local close-

to-tips variations to conserved traits with global close-to-root variations. This approach could provide even more details by highlighting lineage-dependent signals. For example, a single trait might exhibit a general pattern of phylogenetic signal (global structure) and also strong localised trait variations in a single lineage (local structure). Moreover, pPCA can be used together with other multivariate methods to relate the combination of traits identified with a given phylogenetic structure (either global or local) to explanatory factors, such as environmental variables. For instance, co-inertia analysis (Dolédec and Chessel, 1994; Dray et al., 2003a,b) could be used to link phylogenetic structures identified by pPCA to descriptors of the ecological niche, so as to assess potential patterns of adaptation. pPCA could therefore complement both existing ecological methods (*e.g.*, co-inertia) and evolutionary approaches (*e.g.*, phylogenetic overdispersion/clustering), providing a link between trait evolution, patterns in phylogenetic diversity, and biotic or abiotic interactions, and giving insights into the historical and ecological processes that underpin community assembles.

To conclude, we illustrate the intersection between issues in spatial and phylogenetic methods. Spatial and phylogenetic patterns are generated by very different processes, but the mathematical tools that can be used to measure and model these patterns may be similar. This is because both rely on the concept of autocorrelation, which can be defined as the non-independence among observations with respect to a set of underlying proximities. Several spatial methods developed in ecology have already been successfully adapted to PCM (Cheverud et al., 1985; Gittleman and Kot, 1990; Diniz-Filho et al., 1998; Desdevises et al., 2003; Giannini, 2003). Originally, spatial autocorrelation was perceived by ecologists as a nuisance that precluded the use of standard statistical tools in correlative studies. However, the study of spatially autocorrelated patterns turned out to be a fecund paradigm, as ecologists realised that these structures were mere indicators of considerable underlying ecological processes. The same may be true of phylogenetically autocorrelated patterns. Rewording Legendre (1993), we can now also ask the question: *is phylogenetic autocorrelation trouble, or a new paradigm?*

5. Acknowledgements

We are grateful to the CCIN2P3 for providing access to their computers, and particularly to Simon Penel for his help. We thank Anne-Béatrice Dufour for useful discussions on a former version of the manuscript. We address many thanks to F. Stephen Dobson and Nigel G. Yoccoz for their insightful comments and thorough review of our manuscript. We finally wish to thank the team developing the phylobase package for easing the handling of phylogenetic comparative data.

Abouheif, E., 1999. A method for testing the assumption of phylogenetic independence in comparative data. Evolutionary Ecology Research 1, 895–909.

Ackerly, D. D., Schwilk, D. W., Webb, C. O., 2006. Niche evolution and adaptive radiation: testing the order of trait divergence. Ecology 87, S50–S61.

Bauwens, D., Diaz-Uriarte, R., 1997. Covariation of life-history traits in lacertid lizards: a comparative study. The American Naturalist 149, 91–111.

Bolker, B., Butler, M., Cowan, P., de Vienne, D., Jombart, T., Kembel, S., Orme, D., Paradis, E., Zwickl, D., 2007. phylobase: base package for phylogenetic structures and comparative data. R package version 0.3.
URL `http://phylobase.R-forge.R-project.org`

Bortolussi, N., Durand, E., Blum, M., François, O., 2006. apTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics 22, 363–364.

Chessel, D., Dufour, A.-B., Thioulouse, J., 2004. The ade4 package-I- one-table methods. R News 4, 5–10.

Cheverud, J. M., Dow, M. M., 1985. An autocorrelation analysis of genetic variation due to lineal fission in social groups of Rhesus macaques. American Journal of Phyisical Anthropology 67, 113–121.

Cheverud, J. M., Dow, M. M., Leutenegger, W., 1985. The quantitative assessment of phylogentic constaints in comparative analyses: sexual dimorphism in body weights among primates. Evolution 39, 1335–1351.

Cliff, A. D., Ord, J. K., 1973. Spatial autocorrelation. Pion, London.

de Jong, P., Sprenger, C., van Veen, F., 1984. On extreme values of Moran's I and Geary's c. Geographical Analysis 16, 17–24.

Desdevises, Y., Legendre, L., Azouzi, L., Morand, S., 2003. Quantifying phylogenetically structured environmental variation. Evolution 57 (11), 2647–2652.

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.

Dobson, F. S., 1985. The use of phylogeny in behavior and ecology. Evolution 39, 1384–1388.

Dolédec, S., Chessel, D., 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology 31, 277–294.

Dray, S., Chessel, D., Thioulouse, J., 2003a. Co-inertia analysis and the linking of ecological data tables. Ecology 84 (11), 3078–3089.

Dray, S., Chessel, D., Thioulouse, J., 2003b. Procrustean co-inertia analysis for the linking of multivariate datasets. Ecoscience 10, 110–119.

Dray, S., Dufour, A.-B., Chessel, D., 2007. The ade4 package - II: Two-table and K-table methods. R News 7, 47–54.

Dray, S., Legendre, P., Peres-Neto, P., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196, 483–493.

Dray, S., Saïd, S., Debias, F., 2008. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. Journal of Vegetation Science 19, 45–56.

Emerson, B., Gillepsie, R., 2008. Phylogenetic analysis of community assembly and structure over space and time. Trends in Ecology and Evolution 23, 619–630.

Faraway, J. J., 2004. Linear Models with R. Chapman et Hall.

Felsenstein, J., 1985. Phylogenies and the comparative method. The American Naturalist 125, 1–15.

Gaillard, J.-M., Pontier, D., Allainé, D., Lebreton, J. D., Trouvilliez, J., Clobert, J., 1989. An analysis of demographic tactics in birds and mammals. Oikos 56, 59–76.

Garland, T., Bennett, A. F., Rezende, E. L., Aug 2005. Phylogenetic approaches in comparative physiology. J Exp Biol 208 (Pt 16), 3015–3035.
URL http://dx.doi.org/10.1242/jeb.01745

Giannini, N. P., 2003. Canonical phylogenetic ordination. Systematic Biology 52, 684–695.

Gittleman, J. L., Kot, M., 1990. Adaptation: statistics and a null model for estimating phylogenetic effects. Systematic Zoology 39, 227–241.

Gould, S. J., Lewontin, R. C., 1979. The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist program. Proceedings of the Royal Society of London, Series B 205, 581–598.

Grafen, A., 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society of London Series B - Biology 326, 119–157.

Gregory, W. K., 1913. Convergence and applied phenomena in the mammalia. Report of the British Association for the Advancement of Science IV, 525–526.

Griffith, D. A., 1996. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. The Canadian Geographer 40, 351–367.

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

Hansen, T. F., Martins, E. P., 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50 (4), 1404–1417.

Harvey, P. H., Pagel, M., 1991. The Comparative Method in Evolutionary Biology. Oxford University Press.

Harvey, P. H., Purvis, A., Jun 1991. Comparative methods for explaining adaptations. Nature 351 (6328), 619–624.
URL http://dx.doi.org/10.1038/351619a0

Jombart, T., Devillard, S., Dufour, A.-B., Pontier, D., 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101, 92–103.

Jombart, T., Dray, S., 2009. adephylo: exploratory analyses for the phylogenetic comparative method.
URL `http://r-forge.r-project.org/projects/adephylo/`

Kirkpatrick, M., Slatkin, M., 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47, 1171–1181.

Kraft, N. J. B., Cornwell, W. K., Webb, C. O., Ackerly, D. D., 2007. Trait evolution, community assembly, and the phylogenetic structure of ecological communities. American Naturalist 170, 271–283.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74, 1659–1673.

Martins, E. P., 2000. Adaptation and the comparative method. Trends in Ecology & Evolution 15 (7), 296–299.

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

Martins, E. P., Hansen, T. F., 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. The American Naturalist 149 (4), 646–667.

Moran, P. A. P., 1948. The interpretation of statistical maps. Journal of the Royal Statistical Society, B 10, 243–251.

Moran, P. A. P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

Osborn, H. F., 1917. Heritage and habitus. Science 45, 660–661.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Pavoine, S., Ollier, S., Pontier, D., Chessel, D., 2008. Testing for phylogenetic signal in life history variable: Abouheif's test revisited. Theoretical Population Biology 73, 79–91.

Peres-Neto, P., 2006. A unified strategy for estimating and controlling spatial, temporal and phylogenetic autocorrelation in ecological models. Oecologica Brasiliensis 10, 105–119.

Pontier, D., Gaillard, J.-M., Allainé, D., 1993. Maternal investment per offspring and demographc tactics in placental mammals. Oikos 66, 424–430.

R Development Core Team, 2009. 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`

Rochet, M. J., Cornillon, P. A., Sabatier, R., Pontier, D., 2000. Comparative analysis of phylogenetic and fishing effects in life history patterns of teleost fishes. Oikos 91, 255–270.

Rohlf, F. J., 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55 (11), 2143–2160.

Rohlf, F. J., 2006. A comment on phylogenetic correction. Evolution 60, 1509–1515.

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.

Wiens, J., Graham, C., 2005. Niche conservatism: integrating evolution, ecology, and conservation biology. Annual Review of Ecology Evolution and systematics 36, 519–539.

667 Yule, G. U., 1924. A mathematical theory of evolution based on the conclusions
668 of Dr J. C. Willis, F.R.S. Philosophical Transactions of the Royal Society of
669 London, A 213, 21–87.

6. Figure legends

Figure 1: pPCA of simulated data. Example of simulated traits (light yellow) for different tree structures (A-E), and structures identified by pPCA. Global components and corresponding eigenvalues are indicated in red, while local components and their eigenvalues are displayed in blue. Positive and negative values of traits and PCs are indicated by black and white circles, respectively. Symbol size is proportional to absolute values. Simulated traits are labelled as G*i*: i^{th} global structure, L*i*: i^{th} local structure, and R*i*: i^{th} random (*i.e.*, non-phylogenetically structured) trait. Principal components (PC) of pPCA are labelled as GPC1: first global PC (*i.e.*, associated with the largest positive eigenvalue). LPC1: first local PC (*i.e.*, associated with the largest negative eigenvalue). (A) Symmetric tree; random noise added structures ('noise') equaled 0.5. (B) Random clustering of tips; noise=1. (C) Yule model; noise=0.5. (D) biased model; noise=0.75. (E) Assymetric tree; noise=1.

Figure 2: pPCA of lizards data. (A) First global (red) and local (blue) principal components of the pPCA of lacertid lizards data, after removal of size effect. Inset barplot displays the corresponding eigenvalues. Positive and negative scores are indicated by black and white circles, respectively. Symbol size is proportional to absolute values. Taxa are labelled as: *Podarcis h. atrata* ('Pa'), *P. h. hispanica* ('Ph'), *Lacerta lepida* ('Ll'), *L. monticola cantabrica* ('Lmca'), *L. m. cyreni* ('Lmcy'), *Podarcis h. hispanica* Asturias ('Phha'), *P. h. h.* Salamanca ('Pha'), *P. bocagei* ('Pb'), *P. muralis* ('Pm'), *Acanthodactylus erythrurus* ('Ae'), *Takydromus tachydromoides* ('Tt'), *T. septentrionalis* ('Ts'), *Lacerta vivipara* ('Lviv'), *L. agilis* ('La'), *L. schreiberi* ('Ls'), and *L. viridis* (''Lvir'). (B) Loadings of the traits for the first global (red) and local (blue) principal components. Inset barplot displays the corresponding eigenvalues. d=0.5 indicates the mesh of the grid. Analysed traits are hatchling length (hatch.L) and mass (hatch.m), clutch frequency (clutch.F) and size (clutch.S), mean and

maximum female length (mean.L, max.L), mean female length and age at sexual maturity (matur.L, age.mat). See text for a more detailed description of analysed traits.

7. Table legends

Table 1: parameters of the simulated data. 200 datasets were simulated for all combinations of these parameters. (1) expressed in number of tips. (2) number of phylogenetically structured traits (global/local). (3) standard deviation of normal variates added to phylogenetically structured traits. (4) total number of traits in the dataset, including phylogenetically structured traits.

Figure 1:

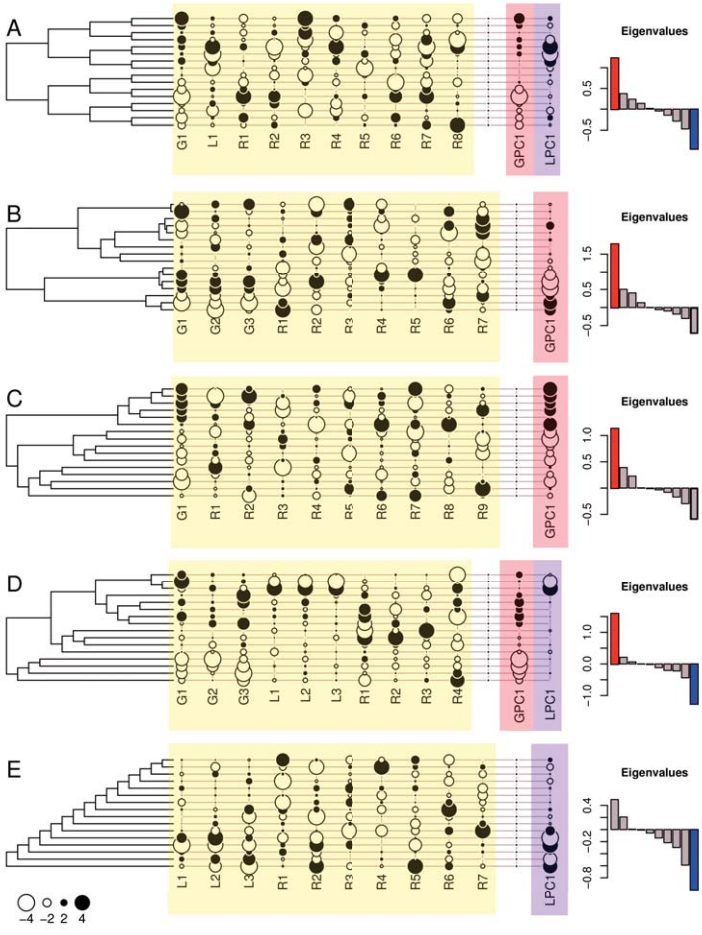
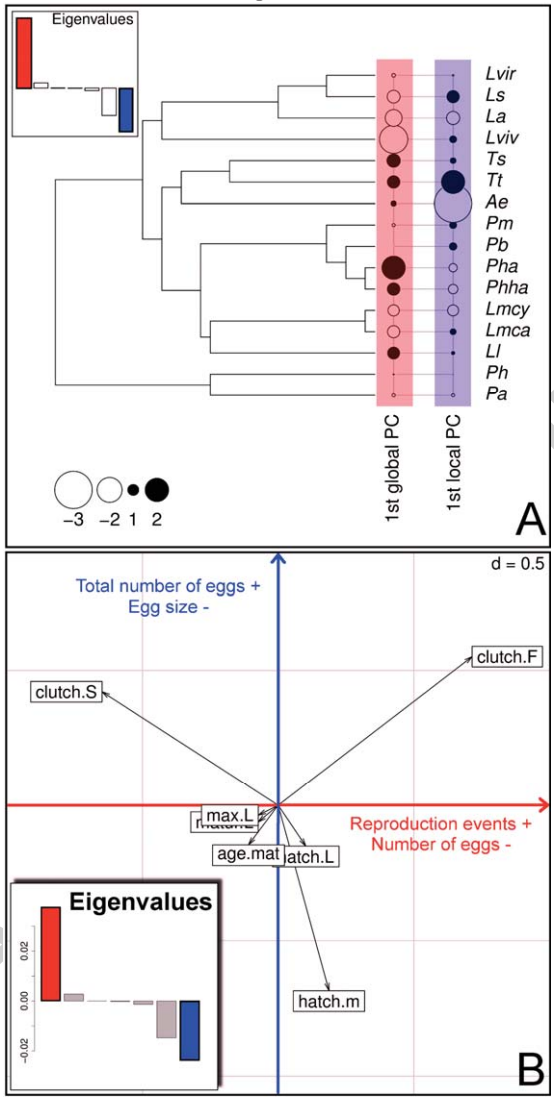


Figure 2:



713 Table 1:

Parameter	Values
Tree model	Symmetric; Random clustering; Yule; Biased; Asymmetric
Tree size[1]	16; 32; 128
Structures[2]	1/0; 0/1; 3/0; 0/3; 1/1; 3/3
Random noise[3]	0.5; 0.75; 1
Number of traits[4]	12; 20; 50

Appendix A. Tables of the analysis of simulations

This appendix presents two tables corresponding to the analysis of simulated data by the linear model described in sections 2.4 and 3.1.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fac.tree[1]	4	6632	1658	2757.68	$< 2.2e^{-16}$
fac.strutype[1]	1	117621	117621	195635.60	$< 2.2e^{-16}$
ntips[3]	1	11412	11412	18981.57	$< 2.2e^{-16}$
ntraits[4]	1	31102.96	31102.96	51732.57	$< 2.2e^{-16}$
noise[5]	1	8156.31	8156.31	13566.13	$< 2.2e^{-16}$
fac.nstruc[6]	1	385.43	385.43	641.08	$< 2.2e^{-16}$
Residuals	215990	129858.80	0.60		

Table A.1: Analysis of variance of the model. Factors are preceded by 'fac'. (1) type of tree. (2) type of structure (global or local). (3) number of tips. (4) total number of traits. (5) number of structured traits (1 or 3). (6) standard deviation of the random noise added to the structured traits.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8474	0.0053	158.99	$< 2e^{-16}$
fac.treeclust[1]	0.0431	0.0053	8.16	$< 2e^{-16}$
fac.treecomb[2]	0.2980	0.0053	56.48	$< 2e^{-16}$
fac.treesym[3]	0.4450	0.0053	84.34	$< 2e^{-16}$
fac.treeyule[4]	0.0410	0.0053	7.78	$< 2e^{-16}$
fac.strutypelocal[5]	-1.4759	0.0033	-442.31	$< 2e^{-16}$
ntips[6]	0.0046	0.0000	137.77	$< 2e^{-16}$
ntraits[7]	-0.0223	0.0001	-227.45	$< 2e^{-16}$
fac.nstruc3[8]	0.0845	0.0033	25.32	$< 2e^{-16}$
noise[9]	-0.9520	0.0082	-116.47	$< 2e^{-16}$

Table A.2: Coefficients of the model. Factors are preceded by 'fac', followed by the levels. (1) trees obtained by random clustering of tips. (2) comb-like model (completely asymmetric trees). (3) completely symmetric trees. (4) Yule model. (5) local phylogenetic structure. (6) number of tips. (7) total number of traits. (8) number of structured traits (1 or 3). (9) standard deviation of the random noise added to the structured traits.