Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 1120–1131 doi: 10.1111/2041-210X.12111

Phylogenetic eigenvector maps: a framework to model and predict species traits

Guillaume Guénard[1]*, Pierre Legendre[1] and Pedro Peres-Neto[2]

[1]*Département de sciences biologiques, Université de Montréal, CP 6128, Succ. Centre-Ville, Montréal, QC H3C 3J7, Canada; and* [2]*Canada Research Chair in Spatial Modeling and Biodiversity, Département des sciences biologiques, Universitédu Québec à Montréal, CP 8888, Succ. Centre-Ville, Montréal, QC H3C 3P8, Canada*

Summary

1. Phylogenetic signals are the legacy related to evolutionary processes shaping trait variation among species. Biologists can use these signals to tackle questions related to the evolutionary processes underlying trait evolution, estimate the ancestral state of a trait and predict unknown trait values from those of related species (i.e. 'phylogenetic modelling'). Approaches to model phylogenetic signals rely on quantitative descriptors of the structures representing the consequences of evolution on trait differences among species.

2. Here, we propose a novel framework to model phylogenetic signals: Phylogenetic Eigenvectors Maps (PEM). PEM are a set of eigenfunctions obtained from the structure of a phylogenetic graph, which can be a standard phylogenetic tree or a phylogenetic tree with added reticulations. These eigenfunctions depict a set of potential patterns of phenotype variation among species from the structure of the phylogenetic graph. A subset of eigenfunctions from a PEM is selected for the purpose of predicting the phenotypic values of traits for species that are represented in a tree, but for which trait data are otherwise lacking. This paper introduces a comprehensive view and the computational details of the PEM framework (with calculation examples), a simulation study to demonstrate the ability of PEM to predict trait values and four real data examples of the use of the framework.

3. Simulation results show that PEM are robust in representing phylogenetic signal and in estimating trait values.

4. The method also performed well when applied to the real-world data: prediction coefficients were high (0·76–0·88), and no notable model biases were found.

5. Phylogenetic modelling using PEM is shown to be a useful methodological asset to disciplines such as ecology, ecophysiology, ecotoxicology, pharmaceutical botany, among others, which can benefit from estimating trait values that are laborious and often expensive to obtain.

Key-words: comparative method, cross-validation, evolutionary models, graph theory, Ornstein–Uhlenbeck process, phylogenetic eigenvectors, phylogenetic modelling, phylogenetic signal, statistical modelling, trait values

Introduction

The morphological, physiological and behavioural characteristics of organisms (i.e. species traits) are in varying degrees the evolutionary legacy of common ancestry (i.e. they are autocorrelated). The result of these patterns of common variation in the phenotypic values of traits is hereafter referred to as 'phylogenetic signal'. The fact that species traits often have a strong phylogenetic signal has long been recognized as an issue because it means that statistical tests of trait correlation among species are biased in comparison with those obtained from independent observations (Felsenstein 1985). Reducing such bias has become perhaps the primary purpose of phylogenetic comparative methods (Freckleton 2009). However, the development of comparative methods has unravelled opportunities extending much beyond this primary purpose. These comparative frameworks allow one, among other goals, to explore the evolutionary processes underlying the evolution of a trait (Hansen 1997; Butler & King 2004; Bartoszek *et al.* 2012; Slater *et al.* 2012; Hernández *et al.* 2013), estimate ancestral trait values (Felsenstein 1985; Garland & Ives 2000; Zheng *et al.* 2009) and predict unknown trait values (Martins & Hansen 1997; Guénard *et al.* 2011; Fagan *et al.* 2013). The present study will focus on that latter purpose, which we refer to as 'phylogenetic modelling'. There are two important components in current comparative frameworks that allow us to model missing traits based on incomplete trait information. The first one is the modelling of phylogenetic signal which is currently performed by a variety of methods such as generalized least-squares (Garland & Ives 2000; Freckleton *et al.* 2002; Blomberg *et al.* 2003), autocorrelation / autoregression

*Correspondence author. E-mail: guillaume.guenard@gmail.com

(Gittleman & Kot 1990; Martins 1996), Bayesian inference (Zheng *et al.* 2009) and eigenfunctions (Diniz-Filho *et al.* 1998; Desdevise *et al.* 2003; Pavoine *et al.* 2008). The second component is the ability to estimate the processes underlying trait evolution (Boettiger *et al.* 2012). Given the current available genetic data and our ability to estimate phylogenies for large number of species, phylogenetic modelling becomes particularly compelling.

The significance of phylogenetic modelling is grounded in the fact that traits can be very laborious and often expensive (e.g. physiological traits) to estimate while being crucial to empirical research as well as conservation and management practices. The ability to predict whether organisms can tolerate new environments, for example, is particularly important given current trends in global changes. However, estimating trait that can inform us about their capacity to tolerate these changes for all species in a phylogeny may not be possible as distributional, and associated environmental information is often not available for a large number of species in any given phylogeny. In such circumstances, a robust and well-implemented phylogenetic model is a useful alternative to estimate trait values from those known from other species in the phylogeny (Zheng *et al.* 2009; Guénard *et al.* 2011).

In this paper, we use graph theory within a phylogenetic modelling context as the approach allows a generic, flexible and robust framework for representing trees as well as candidate evolutionary processes underlying trait variation and divergence among species. Trees are a particular kind of directed graph called an acyclic directed graph (West 2001). Following the terminology of graph theory, which is adapted from that of geometry, a graph is a set of objects, called vertices, that are interconnected by edges. Each edge of a graph connects a pair of vertices, thereby representing their relationship to one another and altogether defining the topology of the graph. In addition to the classical view whereby a phylogeny is represented as having a strict tree structure, phylogenetic graphs can also accommodate any other kind of phylogenetic network originating, for instance, from species hybridization or horizontal gene transfer in bacteria (Makarenkov *et al.* 2004). In a phylogenetic graph, vertices represent extant species (tips) as well as hypothetical ancestors (referred to as 'nodes' or 'root' in the case of the last common ancestor), while edges represent species affiliation. Moreover, since affiliation is unidirectional, going from ancestors to their descendants, phylogenetic graphs are intrinsically directed (i.e. their edges are unidirectional). Graph theory is closely related to linear algebra, and it follows that spectral decomposition of relational matrices obtained from graphs (e.g. pairwise phylogenetic matrices, spatial neighbourhood matrices, time-series matrices) can be used to quantify and represent (i.e. map) the influence of vertices (species, sampling sites, time steps) on one another (Dray *et al.* 2006).

An asymmetric spectral decomposition approach in which the influences of vertices on each other are asymmetric (AEM: asymmetric eigenvector maps; Blanchet *et al.* 2008) has been proposed in ecology as a modelling tool able to accommodate spatial neighbourhood matrices among sampling sites (vertices) that have asymmetric relationships in the sense that the connectivity (influence) of any two given sites may be different on one another (e.g. site A is connected to site B, but site B is not connected to site A, or both sites are connected but with different connectivity weights). This matrix representation and associated spectral decomposition in a modelling framework is then able to cast spatial variability representing spatial latent processes influencing species distributions in space (or time). In addition to the topological information on site locations with respect to a latent spatial process (or a combinations of multiple spatial processes), the AEM framework also allows one to integrate information about the spatial processes influencing variation between adjacent sites (vertices) by applying weights to the graph edges. This approach has great potential value for modelling phylogenetic variation in trait as it allows one to represent information not only on phylogenetic topology (e.g. speciation and hybridization events) but also information associated with different evolutionary processes occurring along the edges (e.g. evolutionary rates). Indeed, different evolutionary models can be represented by giving different weights to the edges in a graphic representation of a phylogenetic tree (e.g. Pagel's lambda, Ornstein–Uhlenbeck process; see Butler & King 2004 and Boettiger *et al.* 2012 for a recent overview). Finally, there is a conceptual similarity between the particular context in which a directed process could drive a temporal (and certain spatial) signal and that where successive speciation events and trait evolution in time could structure phylogenetic signal: both kinds of signals are generated by processes acting in a single direction on a set of interconnected objects or vertices (i.e. sampling sites, time periods or species).

It follows that the evolutionary processes influencing a trait can be modelled from the edges of the phylogenetic graph. The evolutionary forces acting on traits are often seen as a gradient whose end points are neutral (e.g. genetic drift, random mutations leading to gradual changes over generations) and selective (e.g. stabilizing or directed natural selection; Hansen 1997; Butler & King 2004; leading to either lack of change or abrupt changes in trait trajectories in contrast to neutral variation) processes, respectively. Although a phylogenetic signal may occur regardless of the absence or presence of natural selection, the nature of the intervening evolutionary processes may influence the strength and the structure of the phylogenetic signal. For instance, on average, a trait evolving neutrally will change gradually along edges, whereas the same trait may either remain conserved or change steeply after speciation (i.e. at nodes) in the presence of strong stabilizing or directional natural selection. With a graph-based approach to model phylogenetic signal, accommodating these two extreme scenarios of trait evolution is straightforward, in addition to a broad range of practical situations where both neutral processes and natural selection may have driven trait evolution.

The goal of this paper is to propose a general and flexible phylogenetic modelling framework based on graph theory, namely Phylogenetic Eigenvectors Maps (PEM). The proposed framework allows one to model trait variation based on the information obtained on both the graph structure (topology and branch length) and the dynamics of trait evolution, that is, whether phenotypes tend to change gradually over

generations or abruptly at nodes. The greatest advantage of the proposed approach is its flexibility in the sense that, unlike other approaches, PEM can be used within any statistical fitting procedure (e.g. Griffith & Peres-Neto 2006). In order to demonstrate its strengths and robustness, we use a combination of computer simulations and real data sets from the literature, representing different groups of organisms and potentially evolutionary processes.

Materials and methods

CALCULATION OF PHYLOGENETIC EIGENVECTOR MAPS

As established earlier, any phylogenetic tree can be regarded as a phylogenetic graph. In this section, we describe the algebra underlying PEM, whereas a complete numerical example is given in the next subsection.

PEM work on a similar basis as (centred) principal component analysis (PCA; Pearson 1901; Legendre & Legendre 2012). In PEM, one uses a matrix containing the graphic structure of a phylogeny to calculate eigenvectors that are later used as descriptors (predictors) in predictive modelling procedures such as multiple regression (see Griffith & Peres-Neto 2006 for an overview). Because the number of eigenvectors is usually large (number of tips minus one), a subset of eigenvectors is selected and used, alone or with other descriptors (e.g. other traits), to model the variation in the trait of interest.

The calculation of a PEM from a phylogenetic graph begins by coding the topology of the phylogenetic graph in an influence matrix $\mathbf{B} = [b_{i,j}]$. The influence matrix is a $n \times m$ binary matrix whose rows and columns represent the n vertices and m edges of a directed graph, respectively. The term 'influence' here means the inheritance of ancestors on their descendants. An element $b_{i,j}$ is set to 1 when vertex i is under the direct or indirect influence of edge j; an indirect influence is when a non-pendant edge j is in the path from the root to vertex i. Element $b_{i,j}$ is set to 0 when vertex i is assumed not to be influenced by edge j. The information that is not associated with the topology of the graph but, instead, with details about the dynamics of trait change along the edges, is represented using edge weights. The estimation of these weights is presented in details in the next subsection 'Estimating weighting function parameters'.

While trees having the same topology share the same influence matrix, edge weighting is used to represent a particular evolutionary model (e.g. Pagel's lambda, Ornstein–Uhlenbeck) influencing the variation of a given trait. Therefore, by considering appropriate edge weighting, one can obtain a covariance structure that is best at modelling a particular trait. As such, we propose that an edge is assigned a weight $w_{a,\psi}$ proportional to the extent of the change that is expected to occur along that edge based on the following monotonic function:

$$w_{a,\psi}(\phi_j) = \begin{cases} \psi\phi^{\frac{1-a}{2}} & \phi_j > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad \text{eqn 1}$$

where a ($0 \leq a \leq 1$) is the steepness parameter describing how abrupt the changes in trait values occur with time following branching (steepness is related to Pagel (1999) κ, with $a = 1 - \kappa$), ψ ($0 < \psi < \infty$) is the relative evolution rate of the trait being modelled, and ϕ_j is the length (phylogenetic distance) of edge j (Fig. 1). For any given trait, a tree can be assigned a single pair of parameters a and ψ, thereby assuming that the trait evolved in a steady manner throughout the phylogeny. It is also possible to relax this assumption (i.e. one single evolutionary rate for the entire phylogeny) by assigning different parameters to different subordinate phylogenies (e.g. use different pairs of parameters a and ψ for sub-trees; Revell *et al.* 2011; Beaulieu *et al.* 2012), but obviously at the cost of increasing model complexity (i.e. more parameters). Following that model, a describes the initial steepness of the relationship between the extent of the changes in the trait values along the edges of the phylogenetic graph and their (phylogenetic) lengths ϕ_j. Under purely neutral evolution, $a = 0$ and the expected trait changes along edges are proportional to the square root of the phylogenetic distances, with ψ being the proportionality constant, whereas when $a = 1$, changes occur at a fixed rate ψ whenever species diverge; the phylogenetic variation (distances) due to trait evolution subsequently along the edges are irrelevant to the size of the changes. Hence, changes occur more gradually with time for traits evolving neutrally than for traits under natural selection, where adaptive forcing either prevents trait changes (i.e. stabilizing selection) or induces very drastic trait changes (i.e. directional selection).

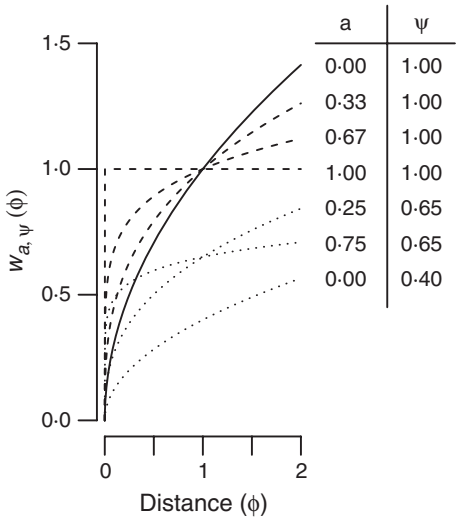

Fig. 1. Profiles of the edge weighting function (eqn. 1) for phylogenetic eigenvectors maps for six different combinations of selection strengths a and evolution rates ψ. Solid line: purely neutral evolution with $\psi = 1$; dashed lines: different selection strengths with $\psi = 1$; dotted lines: different evolution rates and selection strengths.

Phylogenetic Eigenvector Maps (PEM) are then obtained by weighting (eqn. 1) and centring the reduced influence matrix $\mathbf{B}^*$ (the subset of the rows of $\mathbf{B}$ corresponding to the tips), and extracting the singular values of the following product:

$$\mathbf{Q}_{(n)}\mathbf{B}^*\mathbf{D}_w = \mathbf{U}\mathbf{D}_\Sigma\mathbf{V}^{\mathrm{T}} \qquad \text{eqn 2}$$

where $\mathbf{Q}_{(n)}$ is an order n centring matrix (obtained as $\mathbf{Q}_{(n)} = \mathbf{I}_{(n)} - [1/n]_{n\times n}$, where $\mathbf{I}_{(n)}$ is an order n identity matrix and $[1/n]_{n\times n}$ is an n by n matrix whose elements are $1/n$), $\mathbf{D}_\mathbf{w}$ and $\mathbf{D}_\Sigma$ are diagonal matrices of edge weights (i.e. the j^{th} element in the diagonal of $\mathbf{D}_\mathbf{w}$ contains edge weight $w_{a,\psi}(\phi_j)$; matrix has order m) and singular values (order n), respectively, while $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vector matrices, respectively. Due to centring (through the operation $\mathbf{Q}_{(n)}\mathbf{B}^*$ of eqn. 2), the $n-1$ left singular vectors have a mean of 0 and are orthonormal (i.e. their scalar products and column correlations are both zero); $\mathbf{U}$ can be then used as a design matrix (predictors) in modelling procedures. The Phylogenetic Eigenvector Maps (PEM) are the columns of matrix $\mathbf{U}$, that is, the principal components of $\mathbf{Q}_{(n)}\mathbf{B}^*\mathbf{D}_\mathbf{w}$ and each

element $u_{i,j}$ of **U** is the loading of species i on a phylogenetic eigenfunction j. As such, PEM is the orthogonal basis obtained from the structure of the phylogeny (both in terms of topology and trait change dynamic, which is inferred from the edge weights; Dray *et al*. 2006; Blanchet *et al*. 2008). Singular values associated with PEM are proportional to the extent of the variation of $\mathbf{Q}_{(n)}\mathbf{B}^*\mathbf{D}_\mathbf{w}$ in which the largest singular values represent the broadest patterns of phylogenetic variation (i.e. those involving many species) and the smallest values being associated with the narrowest patterns (i.e. those involving only a few species; see Fig. 2). How gradually or abruptly trait changes occur is then modelled as the influence of edge lengths (phylogenetic distances) on trait variation; see example in Fig. 2. Phylogenetic signal in response traits can be then modelled using subsets of the columns of PEM as explanatory variables in fitting procedures.

ESTIMATING WEIGHTING FUNCTION PARAMETERS

The purpose of eqn. 1 is to obtain the phylogenetic model that is the most suitable to represent a phylogenetic signal. The weighting function parameters are estimated on the basis of the among-species phylogenetic covariance matrix **C**. It is interesting to note that there is a direct relationship between PEM (columns of **U**, eqn. 2) and **C**, which can be modelled as a function of **U** as follows:

$$\mathbf{C} = \{\mathbf{Q}_{(n)}\mathbf{B}^*\mathbf{D}_\mathbf{w}\}\{\mathbf{Q}_{(n)}\mathbf{B}^*\mathbf{D}_\mathbf{w}\}^\mathrm{T} = \mathbf{Q}_{(n)}\mathbf{B}^*\mathbf{D}_{\mathbf{w}^2}\mathbf{B}^{*\mathrm{T}}\mathbf{Q}_{(n)} = \mathbf{U}\mathbf{D}_{\Sigma^2}\mathrm{U}^\mathrm{T}. \quad \text{eqn 3}$$

Given that **C** here is produced already centred, its elements can be interpreted as the relative degrees of resemblance between any two species with respect to the covariances among the other species. Positive and negative covariance values indicate species pairs that have greater or lesser resemblance, respectively, than that observed, on average, among other species. Here, we proposed that PEM be extracted using a weighting function parameters a and ψ that best represent the among-species phylogenetic covariance structures associated with a species trait **y**. To do so, we begin by assuming that trait values y_i for each species i follow a multivariate normal probability distribution with

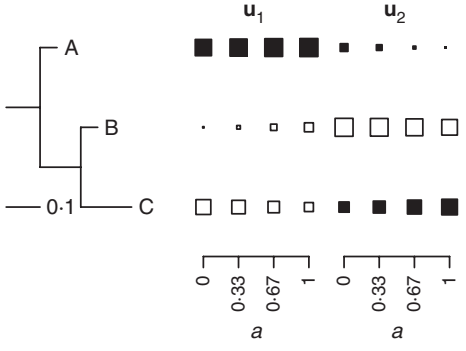

Fig. 2. Impact of increasing selection strength (a) on a centred phylogenetic eigenvector map (PEM) having two eigenvectors ($\mathbf{u}_1$ and $\mathbf{u}_2$ computed for a minimal tree of three species). The open and closed markers are negative and positive species scores on the eigenvectors, respectively, while the sizes of the markers are proportional to the absolute values. For a trait evolving neutrally (i.e. when a is close to 0), evolution occurs progressively along the edges and phylogenetic distances play a more important role on the structure of PEM, whereas under strong natural selection (i.e. when a is close to 1), the structure of PEM is mostly driven by the topology.

covariances $\sigma^2\mathbf{C}$, where σ^2 is the variance of **y**, and **C** is the phylogenetic covariance structure as calculated above (eqn. 3).

Under normality assumptions, the joint probability density function of **y** is:

$$f_\mathbf{C}(\mathbf{y}) = (2\pi)^{\frac{\mathrm{rank}(\mathbf{C})}{2}}|\sigma^2\mathbf{C}|_*^{-\frac{1}{2}}e^{-\frac{1}{2\sigma^2}[y_i-\mathbf{X}\beta]^\mathrm{T}\mathbf{C}^{-*}[y_i-\mathbf{X}\beta]}, \quad \text{eqn 4}$$

where $|\mathbf{C}|_*$ and $\mathbf{C}^{-*}$ denote the pseudo-determinant and the (Moore-Penrose) pseudo-inverse of **C**, respectively; *rank*(**C**) denotes the rank of **C**, **X** is a matrix whose columns are auxiliary traits (i.e. traits that can also serve as predictors of y), $\boldsymbol{\beta}$ is a vector containing the slope estimates of the relationships between the auxiliary traits and **y**, and superscript T denotes matrix transpose. The maximum likelihood estimate of $\boldsymbol{\beta}$ is then obtained as:

$$\hat{\beta} = (\mathbf{X}^\mathrm{T}\mathbf{C}^{-*}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{C}^{-*}\mathbf{y}, \quad \text{eqn 5}$$

whereas the maximum likelihood of σ^2 is obtained as:

$$\hat{\sigma}^2 = \frac{[y_i - \mathbf{X}\beta]^\mathrm{T}\mathbf{C}^{-*}[y_i - \mathbf{X}\beta]}{n}. \quad \text{eqn 6}$$

Because **C** is modelled as a centred matrix (its rows and columns have means of 0), the density function in eqn. 4 is not affected by adding or subtracting a constant value to **y**. Therefore, when no auxiliary traits are used (i.e. when the goal is uniquely predict the response trait solely on the basis of phylogenetic variation contained in **C**), the term $\mathbf{X}\hat{\beta}$ can simply be removed from eqns 4 and 6. The parameters of the weighting function are then estimated as the values that minimize the objective (deviance) function corresponding to eqn. 4 as follows:

$$-2\log L = n + \mathrm{rank}(\mathbf{C})\log(2\pi) + n\log(\hat{\sigma}^2) + \log|\mathbf{C}|_* \quad \text{eqn 7}$$

where L is the likelihood of observing trait **y** given **C**. We propose to use a box-constrained optimization method for this purpose (e.g. Byrd *et al*. 1995; Nocedal & Wright 1999). It is noteworthy that it is not possible to estimate ψ when it is assumed to be constant for the whole phylogeny because it would conflict with the estimation of $\hat{\sigma}^2$ (eqn. 6). In the latter case, we propose to use $\psi = 1$ as a standard value. Moreover, when multiple values of ψ are used, a value needs to be considered a constant (e.g. $\psi_1 = 1$) so that the other values (e.g. $\psi_2, \psi_3, \psi_{\ldots}$) are interpreted relative to ψ_1.

CALCULATION EXAMPLE, PART 1 – MODELLING PHYLOGENETIC SIGNALS

Here, we illustrate how to obtain a PEM using a fictional phylogenetic tree containing seven species labelled A–G (Fig. 3a) and associated made-up trait values (Fig. 3b). The influence matrix $\mathbf{B}^*$ of the example tree is:

$$\mathbf{B}^* = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \\ \mathrm{E} \\ \mathrm{F} \\ \mathrm{G} \end{array} \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

where the columns of $\mathbf{B}^*$ are, from left to right, edges E1−E12 of Fig. 3a. In that example, species A to C are influenced by edge E1, and species D to G by edge E6. The vector of phylogenetic distances of the edges is:

$$\phi = \{0{\cdot}25\ 0{\cdot}15\ 0{\cdot}15\ 0{\cdot}20\ 0{\cdot}35\ 0{\cdot}10\ \ 0{\cdot}30\ 0{\cdot}25\ 0{\cdot}10\ 0{\cdot}30\ 0{\cdot}15\ 0{\cdot}20\}.$$

The steepness parameter computed from the trait values (Fig. 3b) is estimated as $a = 0{\cdot}52$. ψ is assumed constant for the whole tree with a standard value of 1. The vector of edge weights is therefore:

$$\mathbf{w} = \{0{\cdot}72\ 0{\cdot}64\ 0{\cdot}64\ 0{\cdot}68\ 0{\cdot}78\ 0{\cdot}58\ \ 0{\cdot}75\ 0{\cdot}72\ 0{\cdot}58\ 0{\cdot}75\ 0{\cdot}64\ 0{\cdot}68\},$$

the centred and weighted influence matrix for the tips is:

$$\mathbf{Q}_{(n)}\mathbf{B}^*\mathbf{D}_\mathbf{w} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \\ \mathrm{E} \\ \mathrm{F} \\ \mathrm{G} \end{array} \begin{Bmatrix} 0{\cdot}41 & 0{\cdot}45 & 0{\cdot}55 & -0{\cdot}10 & -0{\cdot}11 & -0{\cdot}33 & -0{\cdot}21 & -0{\cdot}10 & -0{\cdot}08 & -0{\cdot}21 & -0{\cdot}09 & -0{\cdot}10 \\ 0{\cdot}41 & 0{\cdot}45 & -0{\cdot}09 & 0{\cdot}58 & -0{\cdot}11 & -0{\cdot}33 & -0{\cdot}21 & -0{\cdot}10 & -0{\cdot}08 & -0{\cdot}21 & -0{\cdot}09 & -0{\cdot}10 \\ 0{\cdot}41 & -0{\cdot}18 & -0{\cdot}09 & -0{\cdot}10 & 0{\cdot}67 & -0{\cdot}33 & -0{\cdot}21 & -0{\cdot}10 & -0{\cdot}08 & -0{\cdot}21 & -0{\cdot}09 & -0{\cdot}10 \\ -0{\cdot}31 & -0{\cdot}18 & -0{\cdot}09 & -0{\cdot}10 & -0{\cdot}11 & 0{\cdot}25 & 0{\cdot}54 & 0{\cdot}62 & -0{\cdot}08 & -0{\cdot}21 & -0{\cdot}09 & -0{\cdot}10 \\ -0{\cdot}31 & -0{\cdot}18 & -0{\cdot}09 & -0{\cdot}10 & -0{\cdot}11 & 0{\cdot}25 & 0{\cdot}54 & -0{\cdot}10 & 0{\cdot}50 & -0{\cdot}21 & -0{\cdot}09 & -0{\cdot}10 \\ -0{\cdot}31 & -0{\cdot}18 & -0{\cdot}09 & -0{\cdot}10 & -0{\cdot}11 & 0{\cdot}25 & -0{\cdot}21 & -0{\cdot}10 & -0{\cdot}08 & 0{\cdot}54 & 0{\cdot}55 & -0{\cdot}10 \\ -0{\cdot}31 & -0{\cdot}18 & -0{\cdot}09 & -0{\cdot}10 & -0{\cdot}11 & 0{\cdot}25 & -0{\cdot}21 & -0{\cdot}10 & -0{\cdot}08 & 0{\cdot}54 & -0{\cdot}09 & 0{\cdot}58 \end{Bmatrix},$$

and the centred covariance matrix $\mathbf{C}$ obtained from eqn. 3 is:

$$\mathbf{C} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \\ \mathrm{E} \\ \mathrm{F} \\ \mathrm{G} \end{array} \begin{Bmatrix} 0{\cdot}93 & 0{\cdot}52 & 0{\cdot}21 & -0{\cdot}43 & -0{\cdot}40 & -0{\cdot}41 & -0{\cdot}42 \\ 0{\cdot}52 & 0{\cdot}97 & 0{\cdot}20 & -0{\cdot}43 & -0{\cdot}41 & -0{\cdot}42 & -0{\cdot}43 \\ 0{\cdot}21 & 0{\cdot}20 & 0{\cdot}90 & -0{\cdot}34 & -0{\cdot}31 & -0{\cdot}32 & -0{\cdot}33 \\ -0{\cdot}43 & -0{\cdot}43 & -0{\cdot}34 & 0{\cdot}96 & 0{\cdot}47 & -0{\cdot}11 & -0{\cdot}12 \\ -0{\cdot}40 & -0{\cdot}41 & -0{\cdot}31 & 0{\cdot}47 & 0{\cdot}83 & -0{\cdot}08 & -0{\cdot}09 \\ -0{\cdot}41 & -0{\cdot}42 & -0{\cdot}32 & -0{\cdot}11 & -0{\cdot}08 & 0{\cdot}88 & 0{\cdot}46 \\ -0{\cdot}42 & -0{\cdot}43 & -0{\cdot}33 & -0{\cdot}12 & -0{\cdot}09 & 0{\cdot}46 & 0{\cdot}92 \end{Bmatrix}.$$

From matrix $\mathbf{Q}_{(n)}\mathbf{B}^*D_\mathbf{w}$, a PEM with six (left) singular vectors $\mathbf{U}$ is obtained using eqn. 2 (Fig. 4). A model of the phylogenetic signal (trait values) is obtained by regressing $\mathbf{y}$ against $\mathbf{U}$. Note, however, that the number of PEMs (i.e. columns on $\mathbf{U}$) is too large for the number of species (i.e. 6 PEMs are produced), and as such, we need to use a model selection procedure to retain the most important eigenvectors mapping (explaining) trait variation (see Diniz-Filho *et al.* 2012 for a recent review). The selection procedure used in our simulation and real examples are explained later on. For this small example, however, we did not use a selection and simply picked the set that explained the largest amount of trait variation (adjusted coefficient of determination). As a result, the phylogenetic signal presented in Fig. 3b was modelled as the following combination of the three PEMs, among those shown in Fig. 4:

$$y = -0{\cdot}10 - 3{\cdot}33\mathbf{u}_2 + 2{\cdot}20\mathbf{u}_3 + 0{\cdot}67\mathbf{u}_5.$$

It is noteworthy that the intercept of that regression model is not the estimated trait value at the root of the phylogeny (see *Discussion* for details).

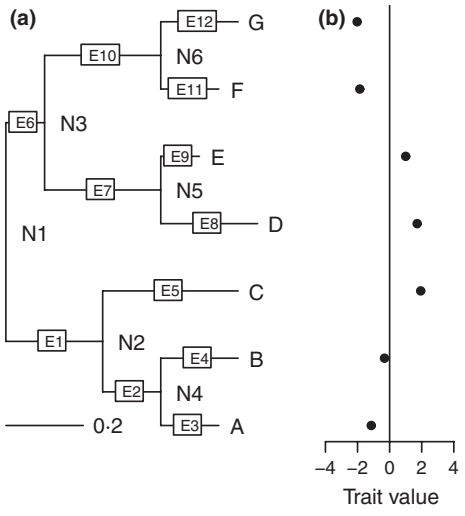

Fig. 3. Example of (a) a phylogenetic tree with seven tips labelled A–G, six nodes labelled N1–N6, and 12 edges labelled E1–E12, and (b) a phylogenetic signal with values of a hypothetical trait among the species.

ESTIMATING PREDICTED VALUES

Phylogenetic modelling, in the context of this paper, deals with incomplete trait information. Its goal is to predict unknown trait values for one or several species (referred here as the 'target' species) in phylogenies for which we have trait values already estimated for a reduced set of species (referred here as the 'model' species), based on their relative phylogenetic positions. In order to estimate trait values for the target species, one must know where the target species are located on the phylogeny in relation to the model species, that is, on which edge and where on that edge. That information is used to obtain vectors of species coordinates describing the relative position of any target species in the model's weighted influence matrix ($\mathbf{s}_{\text{target}}$; see example in the next section). These coordinates are obtained based on the following three steps: (i) determining the identity of (a) the 'split edge', that is, the edge on which the target species is located and (b) the 'parent vertex', that is, the vertex located immediately downward (i.e. closer to the root) from that edge; (ii) copy the coordinates of the parent vertex in the model's influence matrix; and (iii) assign a weight to the split edge. That weight is calculated on the basis of eqn. 1 using the parameters a and ψ estimated from the model species and the phylogenetic distance ϕ between the parent vertex and the location of the target species on the split edge.

Equation 2 can be rearranged as follows:

$$\mathbf{U} = \{\mathbf{B}^*\mathbf{D}_\mathbf{w} - 1_{n\times 1}[1/n]_{1\times n}\mathbf{B}^*\mathbf{D}_\mathbf{w}\}\mathbf{V}\mathbf{D}_\Sigma^{-1}, \qquad \text{eqn 8}$$

where $[1/n]_{1\times n}$ is a row vector whose n elements have values $1/n$, and $1_{n\times 1}$ is a column vector whose n elements have values 1. The term $[1/n]_{1\times n}\mathbf{B}^*\mathbf{D}_\mathbf{w}$ contains the means of the columns of $\mathbf{B}^*\mathbf{D}_\mathbf{w}$. To calculate the scores of the target species ($\mathbf{u}_{\text{target}}$) on a map of the model species, we substitute the relative position of the target species ($\mathbf{s}_{\text{target}}$) in place of the weighted influence matrix ($\mathbf{B}^*\mathbf{D}_\mathbf{w}$) in eqn. 8:

$$\mathbf{u}_{\text{target}} = \{\mathbf{s}_{\text{target}} - [1/n]_{1\times n}\mathbf{B}^*\mathbf{D}_\mathbf{w}\}\mathbf{V}\mathbf{D}_\Sigma^{-1}. \qquad \text{eqn 9}$$

PEM predictions work on the principle that $\mathbf{s}_{\text{target}}$ contains the closest substitutes, in terms of phylogenetic topology and dis-

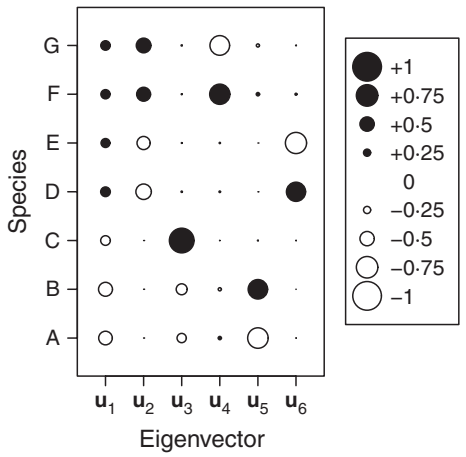

Fig. 4. A phylogenetic eigenvector map composed of seven eigenvectors obtained for the seven species whose phylogeny is illustrated in Fig. 3; trait evolution is assumed to be neutral: $a = 0$. The open and closed markers are negative and positive species scores on the eigenvectors, respectively, while marker sizes are proportional to the absolute values.

tances, for the lines of $\mathbf{B}^*\mathbf{D}_\mathbf{w}$ and that $\mathbf{u}_{\text{target}}$, by analogy, contains also the closest substitutes for the lines of $\mathbf{U}$. To predict trait values for the 'target species', these scores are used as predictors in a model that is obtained using the loadings of the 'model species'. It is important to note that target species are not added to, but rather projected (i.e. 'mapped') on, the PEM.

CALCULATION EXAMPLE, PART 2 – PREDICTING TRAIT VALUES

In order to illustrate how to estimate predicted values using our proposed PEM framework, let us add three target species (called X, Y and Z) to the previous example (Fig. 5a). Target species X intersects the phylogeny on edge E1, at a phylogenetic distance $\phi_X = 0{\cdot}2$ from parent vertex (node N1). The coordinates of node N1 in the model's weighted influence matrix are:

$$\begin{array}{c} \quad E1\ E2\ E3\ E4\ E5\ E6\ E7\ E8\ E9\ E10\ E11\ E12 \\ s_{N1} = \{\ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \} \end{array}$$

note that in this example, N1 is neither influenced directly nor indirectly by any of the edges of the graph thus receiving values 0 everywhere. Given the evolution parameters obtained in the first part of this example ($a = 0{\cdot}52$ and $\psi = 1$), the weight (or coordinate) of target species X on edge E1 is 0·68 (eqn. 1: $w_{0{\cdot}52,1}(0{\cdot}2) = 0{\cdot}68$) so that the location of X in the model's weighted influence matrix is:

$$\begin{array}{c} \quad E1\ \ E2\ E3\ E4\ E5\ E6\ E7\ E8\ E9\ E10\ E11\ E12 \\ s_x = \{\ \mathbf{0{\cdot}68}\ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \}. \end{array}$$

Likewise, target species Y and Z are located upwards from nodes N3 and N2, respectively, whose locations in the model's weighted influence matrix are:

$$\begin{array}{c} \quad E1\ \ E2\ E3\ E4\ E5\ \ E6\ \ E7\ E8\ E9\ E10\ E11\ E12 \\ s_{N3} = \{\ 0\ \ \ 0\ \ 0\ \ 0\ \ 0\ 0{\cdot}58\ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \}. \\ s_{N2} = \{\ 0{\cdot}72\ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \} \end{array}$$

Target species Y intersects the phylogeny on edge E7 at distance $\phi_Y = 0{\cdot}25$ from N and has a coordinate of $w_{0{\cdot}52,1}(0{\cdot}25) = 0{\cdot}72$ on edge E7, whereas target species Z, which intersects the phylogeny on edge E5 at distance $\phi_Z = 0{\cdot}1$ from N2, has a coordinate of $w_{0{\cdot}52,1}(0{\cdot}10) = 0{\cdot}58$ on edge E5. Hence, their locations in the model's weighted influence matrix are:

$$\begin{array}{c} \quad E1\ \ E2\ E3\ E4\ \ E5\ \ E6\ \ E7\ E8\ E9\ E10\ E11\ E12 \\ s_y = \{\ \ 0\ \ \ 0\ \ 0\ \ 0\ \ \ 0\ 0{\cdot}58\ \mathbf{0{\cdot}72}\ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \}. \\ s_z = \{\ 0{\cdot}72\ 0\ \ 0\ \ 0\ \mathbf{0{\cdot}58}\ \ 0\ \ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \} \end{array}$$

From these coordinates and eqn. 9, the projected scores of target species X, Y and Z on the Phylogenetic Eigenvector Map (Fig. 4) are:

$$\begin{array}{c} \quad\quad \mathbf{u}_1\ \ \ \mathbf{u}_2\ \ \ \mathbf{u}_3\ \ \ \mathbf{u}_4\ \ \ \mathbf{u}_5\ \ \ \mathbf{u}_6 \\ \mathbf{u}_x = \{\ -0{\cdot}25\ \ 0{\cdot}00\ \ 0{\cdot}20\ \ 0{\cdot}01\ -0{\cdot}01\ -0{\cdot}01\ \} \\ \mathbf{u}_y = \{\ \ 0{\cdot}27\ -0{\cdot}35\ -0{\cdot}01\ \ 0{\cdot}01\ \ 0{\cdot}00\ -0{\cdot}15\ \}. \\ \mathbf{u}_z = \{\ -0{\cdot}32\ \ 0{\cdot}00\ \ 0{\cdot}69\ -0{\cdot}01\ \ 0{\cdot}02\ \ 0{\cdot}01\ \} \end{array}$$

By applying the regression model built using the PEM of the target species in our previous example using the scores presented above, we obtain the following predictions: $y_x = 0{\cdot}328$, $y_y = 1{\cdot}019$, and $y_z = 1{\cdot}436$.

In our example, the target species X diverged from a common ancestor shared with model species A, B and C, while species Z diverged later; sharing its most recent ancestry with species C only (Fig. 5). It follows that estimated trait values for the target species indicate that species X is intermediate in relation to species A, B and C, whereas the trait estimates for species Z are much closer to the value observed for species C. Along similar lines, trait estimates for target species Y are similar to values observed for model species D and E, with whom Y has its last common ancestor somewhere along edge E7. However, because Y is phylogenetically closer to E than to D, and given that trait evolution contains a neutral component (i.e. $\hat{a} < 1$), the estimated trait value for species Y is more similar to that of E than to D.

SIMULATION STUDY

In order to demonstrate the robustness and accuracy of PEM in predicting trait values, we designed a numerical experiment in which we repeatedly applied PEM at predicting simulated trait values. Quantitative traits were simulated based on the Ornstein–Uhlenbeck (OU) process (Uhlenbeck & Ornstein 1930; Hansen 1997). This process describes the instantaneous rate of change in a trait value (dy) as the sum of a deterministic component forcing the trait value towards an optimum θ and a random component allowing the trait value to spread around the optimum:

$$\mathrm{d}y = \alpha(y - \theta)\mathrm{d}\phi + \sigma \mathrm{d}W\phi, \qquad \text{eqn 10}$$

where α and σ are the rates guiding the deterministic (i.e. selection) and random (i.e. diffusion) components of the trait change, respectively, $\mathrm{d}\phi$ is an infinitesimal amount of phylogenetic distance, and $\mathrm{d}W\phi \sim N(0,\mathrm{d}\phi)$ is a deviate from a normal distribution with 0 mean and variance $\mathrm{d}\phi$. To simulate quantitative trait evolution, we began by drawing random phylogenies with 50, 100, 200 and 400 species (100 random trees for each number of species) using the function 'rtree' in R package ape (Paradis *et al.* 2004). We then defined a set of seven optima ($\theta \in \{-3,-2,-1,0,1,2,3\}$) for which the trait is selected towards. For each phylogeny, we let these optimal values shift from one vertex to the next, in the direction going from the root to the tips, following a Markov chain process. The chain allowed the optima to shift from 0 (the value set at the root) towards either −3 or 3 in either in an ascending or descending step of 1 with a transition probability of 0·1 for individual steps. Trait values were simulated node by node, by let-

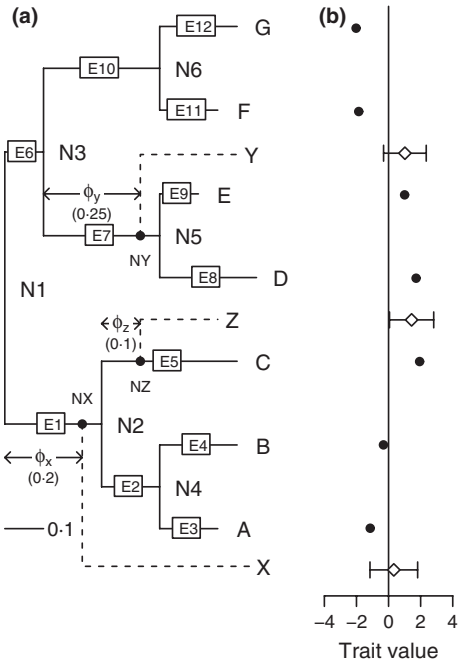

Fig. 5. (a) Intersection of three target species (X, Y and Z) with the phylogeny of the model species A–G (Fig. 3). The intersection vertices, node NX, NY and NZ, are located at distances ϕ_X, ϕ_Y and ϕ_Z from nodes N1, N3 and N2 along edges E1, E7 and E5, respectively. The information about where a target species intersects the phylogeny of the model species can be used to predict its trait values. (b) Predicted trait values for the target species obtained using PEM (open diamond; obtained using three eigenvectors $\mathbf{u}_2$, $\mathbf{u}_3$ and $\mathbf{u}_5$, shown in Fig. 4), both with bars showing the limits of 95% prediction intervals. The close circles are the trait values for the model species (as in Fig. 3).

ting the trait shift from a starting value of 0 at the root, once again from one vertex to the next and in the direction going from the root towards the tips. For each vertex k, we calculated the mean value of the trait μ_k as:

$$\mu_k = y_i e^{-\alpha\phi_j} + \theta_k(1 - e^{-\alpha\phi_j}), \qquad \text{eqn 11}$$

where y_i is the trait value at vertex i, the immediate ancestor of k, ϕ_j is the length (i.e. phylogenetic distance) of edge j joining vertices i and k, and θ_k is the optimal trait value at vertex k. The diffusion σ_j^2 expected to occur along edge j is:

$$\sigma_j^2 = \sigma^2 \frac{1 - e^{-2\alpha\phi_j}}{2\alpha}. \qquad \text{eqn 12}$$

The simulated trait value at vertex k, y_k, is a number drawn from a random normal distribution with mean μ_k and variance σ_j^2. For every single tree, we chose four selection rates: $\alpha = 0$ (pure diffusion), $\alpha = 0{\cdot}5$ (weak selection), $\alpha = 1$ (medium selection) and $\alpha = 10$ (strong selection), with the diffusion rate kept constant at $\sigma = 1$. Using these parameter combinations, we generated 1600 simulation runs. Finally, for each simulation run, we generated 100 sample traits.

For the sake of comparing the simulated values with model predictions, we split each data set in two equal parts; the first half being the model species used to fit regression models on the basis of PEMs and the second half being the target species for which predictions were to be made. Note that because we know the expected values for the target species (i.e. we simulated them along with the model species), by comparing estimated values from the PEM models and the simulated ('true') values, we can estimate prediction error for each parameter combination. The model species were then used to estimate parameter a of eqn. 1 (a single value for the whole phylogeny). We then computed PEMs and regressed trait values against them. PEM were selected in order to minimize information loss using a forward stepwise procedure. Information loss was estimated using the Akaike information criterion with correction for small sample size (AICc; Hurvich & Tsai 1993).). The model with the lowest AICc was retained. We then calculated the scores of the target species (eqn. 9) and used them within the regression model to estimate trait values for the target species.

We used the following prediction coefficient to assess the predictive power of PEM models:

$$P^2 = 1 - \frac{\text{MSPE}}{s_y^2} \qquad \text{eqn 13}$$

where $MPSE$ is the mean square prediction error. $MPSE$ is estimated as $MSPE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n$, where $\hat{y}_i$ is the trait value estimated by the PEM model for target species i, y_i is the simulated 'true' value for target species i (but obviously not used in the fitting process), and s_y^2 is the sample variance of the 'true' trait values. As such, P^2 can be interpreted similarly as Ezekiels' (1930) adjusted coefficient of determination. $P^2 = 1$ when all predictions perfectly match the observations, whereas values below 1 indicate imperfect predictions. Values P^2 close to 0 (negative or positive) indicate that predictions have poor accuracy, being no better than what would be expected from chance alone. We also used the simulation results to explore how the selection rate (α) used to generate the traits influenced the estimates of the steepness parameter (a) and highlight possible methodological behaviours regarding prediction biases as a function of these estimates.

REAL EXAMPLES

As a demonstration of the accuracy of the directed graph approach described here, we applied our PEM framework to four cases taken from published studies (Table 1; Purvis & Rambaut 1995; Isaac *et al.* 2005; Lislevand & Thomas 2006). In all cases, we first selected a single response trait to be modelled. We applied the following iterative procedure: for each species i, we took i as the target species and the remaining $n-1$ species as the model species and estimated the steepness parameter a (eqn. 1); we then used the correlation structure obtained, together with that between the target and the model species, to predict the trait value for i using the same procedure as in our simulation study. Although our method (eqn. 1) allows one to ascribe values of parameter a to specific sub-trees, for the sake of simplicity, we did not use that functionality and assumed instead a unique value for parameter a for all species in data sets. As in the simulation study, we compared the observed values against the predictions to estimate prediction power and also highlight possible methodological behaviours regarding prediction biases.

COMPUTER SOFTWARE

Computer software to perform the analysis and simulations described in the present study is available as the contributed R package MPSEM (i.e. *Modelling Phylogenetic Signals using Eigenvector Maps*) from the Comprehensive R Archive Network (CRAN; http://cran.r-project.org/). The R language scripts and data used to generate all the examples and figures in this paper are available as supplementary electronic material.

Table 1. Phylogenetic modelling of traits observed on four groups of n organisms. Model parameters $\hat{a}$ are the estimated steepness parameter of the trait evolution model (eqn. 1, $\psi = 1$), P^2 is the prediction coefficient of the trait; b_0 and b_1 are the regression intercepts and slopes [with their 95% confidence limits, respectively]

#	Organism – trait	n	$\hat{a}$	P^2	b_0	b_1	References
1	*Anolis* lizards – body mass	85	0·34	0·76	0·18 [−0·31,0·66]	0·95 [0·84,1·07]	Rich Glor*
2	Shorebirds – log egg mass	71	0·08	0·83	−0·06 [−0·39,0·27]	1·02 [0·91,1·13]	Lislevand & Thomas 2006
3	Odd-toed ungulates – log Neonatal mass	13	0·00	0·88	−0·01 [−0·38,0·37]	1·01 [0·75,1·26]	Purvis & Rambaut 1995
4	Primates – gestation period	63	0·38	0·80	−3·33 [−25·65,18·99]	1·02 [0·89,1·15]	Isaac *et al.* 2005

*Source: http://bodegaphylo.wikispot.org/Phylogenetics_and_Comparative_Methods_in_R?action=Files&do=view&target=anolis_mtDNA.mrb.con [first tree, accessed 13-04-19].

Table 2. Synthetic results of the simulations (P^2: prediction coefficients, $\hat{a}$: estimates of the steepness parameter) performed to illustrate the predictive power of PEM for different sample sizes (n) and natural selection strengths (α). Upper and lower 95% confidence limits are the 2·5 and 97·5 percentiles of the 10 000 (100 phylogenies × 100 iterations per phylogeny) signals generated for every combination of n and α that we explored. Format: mean [lower CL, upper CL]

n	α	P^2	$\hat{a}$
25	0	0·56 [0·03, 0·87]	0·28 [0·00, 1·00]
	0·5	0·45 [−0·16, 0·81]	0·40 [0·00, 1·00]
	1	0·56 [0·01, 0·86]	0·48 [0·00, 1·00]
	10	0·79 [0·46, 0·95]	0·64 [0·00, 1·00]
50	0	0·66 [0·33, 0·87]	0·19 [0·00, 0·94]
	0·5	0·56 [0·17, 0·81]	0·39 [0·00, 1·00]
	1	0·66 [0·33, 0·86]	0·51 [0·00, 1·00]
	10	0·85 [0·68, 0·95]	0·75 [0·00, 1·00]
100	0	0·73 [0·52, 0·88]	0·14 [0·00, 0·65]
	0·5	0·62 [0·33, 0·82]	0·33 [0·00, 0·97]
	1	0·69 [0·39, 0·87]	0·46 [0·00, 1·00]
	10	0·86 [0·67, 0·96]	0·69 [0·00, 1·00]
200	0	0·78 [0·64, 0·89]	0·09 [0·00, 0·45]
	0·5	0·68 [0·49, 0·83]	0·30 [0·00, 0·78]
	1	0·75 [0·56, 0·87]	0·45 [0·00, 0·95]
	10	0·89 [0·80, 0·96]	0·71 [0·04, 1·00]

Results

SIMULATION EXPERIMENT

Our simulation experiments showed that PEM phylogenetic models had prediction coefficients (power) ranging from −1·66 to 0·98, with 95% of the cases being between 0·23 and 0·93 (Table 2). Note that prediction coefficients closer to 1 indicate perfect predictability, whereas coefficients of zero and negative indicate complete absence of prediction power. The prediction power tended to increase with sample size ($F_{1,158400} = 46705$; $p < 0·0001$). When $\alpha = 0$, for instance, P^2 increases by 0·00045, on average, for each new species added within the range of conditions that we simulated ($25 \leq n \leq 200$, i.e., half of the species for the simulated phylogenies $50 \leq n \leq 400$). The different selection rates used to generate the phylogenetic signals also influenced the percentage of variation that could effectively be predicted using PEM ($F_{3,158400} = 40327$; $p < 0·0001$). We found that for signals generated with $\alpha = 0·5$, P^2 were, on average, 0·0479 lower than those generated with $\alpha = 0$, whereas signals generated with $\alpha = 1$ and $\alpha = 10$ had P^2 higher by 0·0333 and 0·253 than those with $\alpha = 0·5$, respectively. The results of the simulation study also evidenced an interaction between sample size and selection rate on the prediction power of PEM-based regression models ($F_{3,158400} = 1117$; $p < 0·0001$). The effect of sample size on the predictive power was lower when signals were simulated with some amount of selection (by −0·000200, −0·000200 and −0·000358 for $\alpha = 0·5$, $\alpha = 1$, and $\alpha = 10$, respectively) then with pure diffusion. We also found that the combined effects of sample size and selection rate varied among the phylogenies ($F_{1592,158400} = 62·40$; $p < 0·0001$), indicating that differences in the structure of phylogenies also affect the power of our proposed modelling framework. Finally, we found that the selection rate used for simulations impacted the estimation of the steepness parameter ($F_{3,158400} = 26424$; $p < 0·0001$); the mean estimates were 0·17, 0·35, 0·48 and 0·70 when α was 0, 0·5, 1 and 10, respectively.

REAL EXAMPLES

With the exception of example 1 (body mass of *Anolis*), all published phylogenetic trees were ultrametric; hence, the original authors assumed that trait evolution occurred at constant rates along all edges of their graphs (Table 1, Fig. 6). Examples presented sample sizes varying between 13 and 85 species, and estimates of the steepness parameter (a) ranged from 0 to 0·38. Models were quite accurate, having prediction coefficients (P^2) ranging from 0·76 (case #1: body mass of *Anolis* lizards) to 0·88 (case #3: neonatal mass of odd-toed ungulates). We found neither absolute nor relative prediction biases, that is, in all four cases, the value 0 was inside the (95%) confidence interval of the regression intercepts and the value 1 was inside the confidence interval of the regression slopes.

Discussion

In the present study, we developed a framework for phylogenetic modelling based on graph theory that allows one to predict trait values for species lacking this information. Phylogenetic modelling has been around for some time (Felsenstein 1985; Martins & Hansen 1997), yet its applications are still sparse in spite of the far-reaching empirical and applied potentials (Guénard *et al.* 2011). Here, we provide an approach that is simple yet versatile enough to be applied jointly with other fitting procedure that can be used to describe and predict traits across multiple species (e.g. multiple regres-

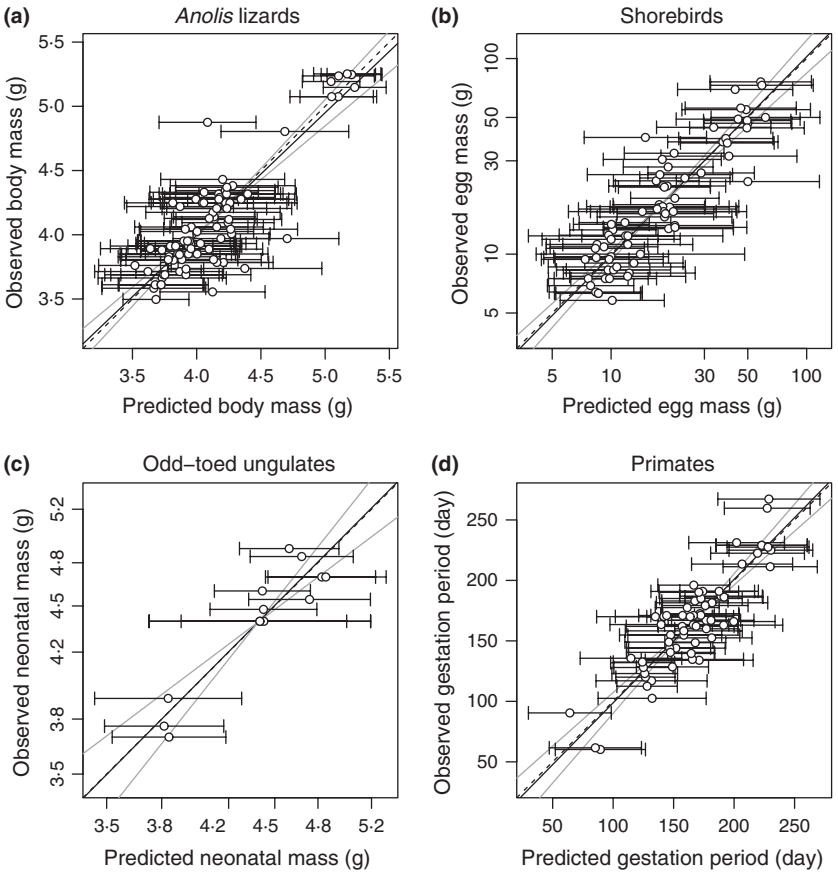

Fig. 6. Observed and predicted trait values obtained following leave-one-out cross-validation for four exemplary cases involving (a) genus *Anolis* lizards, (b) shorebirds, (c) odd-toed ungulates and (d) primates. The dashed lines are 1:1 lines, the solid black lines are regression lines of observed values as a function of predictions, and the solid grey lines are 95% confidence limits of the regression lines. Vertical bars are limits of the 95% prediction intervals.

sion, bilinear regression, neural networks, GLMs). In addition to its versatility, PEM is computationally far less intensive than estimating optimal trait values using Markov chain algorithm (e.g. Butler & King 2004), the latter being a non-polynomial complete problem.

There are many modelling methods that rely on knowledge about species traits, and it is rather common that biologists face the issue posed by the absence of information on such traits when performing modelling exercises. For example, a bioenergetic model requires that several traits describing the physiological response of organisms to the physical and chemical characteristics of their habitat be known in order to predict their growth response and their impact on resources, among other issues (Adams & Breck 1990). Because the collection of trait information can sometimes be logistically challenging or unacceptable (e.g. performing lethal assays on rare or endangered species), onerous, and/or time-consuming, researchers and practitioners (e.g. conservation biologists) require appropriate and robust methods to estimate them. Our simulation results cover a wide variety of situations encompassing cases where the method produced accurate models (P^2 close to 1) and other cases where models yielded low predictive power (positive P^2 near 0 or, in some case, negative P^2). Using a wide range of scenarios, we were able to capture the essences of the predictive power of PEM. Not all traits may be equal with respect to their capacity for being modelled solely based on their phylogenetic relationships. Similarly, simulation results also pointed out that not all phylogenies appeared to have the same potential for phylogenetic modelling. While a particular trait may be clearly phylogenetically structured and thus relatively easy to model from the phylogeny even with relatively modest sample size (i.e. with few model species), patterns of phylogenetic signal may be rather weak for other traits. In the latter case, phylogenetic modelling attempts are unlikely to be successful. However, we showed that it is straightforward to assess fit as well as the level of accuracy related to a phylogenetic model using a cross-validation exercise.

The present paper focused on describing the PEM and the simulation study was primarily concerned about assessing how the method generally performs. The prediction coefficient increased with sample size was lower for $0 < \alpha < 1$ than for the other selection rates simulated, and we found interactions between sample size and selection rate. We also found that the selection rate has a positive influence on the steepness parameter. We expected such a relationship because α controls the rate at which a trait tends towards an optimum, whereas a describes whether a trait tends either to change progressively along edges or to shift abruptly at nodes. Further simulation studies

regarding the properties of PEM could be performed, but they are beyond the scope of the present paper. One issue, not considered here, is how trait coverage across the phylogeny may affect predictive power. For instance, we may have disproportionally larger numbers of trait values for some clades in contrast to other clades, and this uneven distribution is likely to affect phylogenetic modelling. It would also be useful that future studies investigate the relationship between metrics of prediction power, model accuracy and properties of phylogenetic trees like resolution, and balance, among other. Moreover, an assessment of the limits of applicability of phylogenetic modelling by PEM and other methods such as generalized least square regression (Goldberger 1962) would be a valuable asset for applied biology. Such a study could be performed by creating signals in large phylogenies and subsampling small numbers of tips from them. In that way, it would be possible to evaluate the accuracy of phylogenetic model predictions for target species with many of their relatives being model species, compared to that of target species whose common ancestry with the model species is more remote.

A key feature of PEM is versatility. Freckleton *et al.* (2002) have proposed a phylogenetic comparative method to consider different evolutionary models by multiplying the off-diagonal elements of the phylogenetic covariance matrix by a factor between 0 and 1 (e.g. Pagel's lambda). In that framework, selection (i.e. the change in trait optima) is regarded to be intrinsically independent of phylogeny and resulting from processes that are entirely driven by the environment. It therefore makes sense to take out that part of the trait variation to assess the relationship among trait values. In phylogenetic modelling, one is primarily concerned with estimating the relative patterns of trait covariance among species in order to obtain the eigenvectors that are the most suitable to explain trait variation. Hence, we proposed to reach that goal by treating topology and edge length separately in order to fine-tune eigenfunctions with details about the dynamics of trait change. The graph-based method described in the present study is a straightforward implementation of this approach, allowing one to fine-tune trait predictions by incorporating selection (both stabilizing and directional) with diffusion in the form of a steepness parameter (a). To this end, our simulation results have shown that there is a relationship between a and the OU selection strength (α), at least under the range of simulation parameters we used. The method also has the additional benefit of being flexible from an algorithmic standpoint. Hence, phylogenetic graphs have the ability to represent a broader array of potential evolutionary relationships than classical phylogenetic trees, for example, hybridization, or lateral gene transfer. This feature will likely facilitate the adaptation of our framework to address future phylogenetic modelling challenges such as to the modelling of traits under reticulated evolution (Makarenkov *et al.* 2004). Because PEM actually consist in building a set of descriptor variables, the framework is adaptable to a broad array of possible modelling methods in addition to multiple regression. Hence, PEM can be used in, redundancy analysis, generalized linear models, bilinear models (Gabriel 1998), artificial neural network models, regression trees among others (see Griffith & Peres-Neto (2006) for a discussion of flexibility of eigenvector predictors in the context of spatial modelling).

The applications to real cases presented here have shown that our PEM framework is accurate enough and potentially useful in many other cases. This should not be assumed to be always the case, however, as predictions may sometimes be too inaccurate to have a practical value. Although one may be inclined to use a test of phylogenetic signal prior to phylogenetic modelling (e.g. Abouheif 1999; Blomberg *et al.* 2003; see Münkemüller *et al.* 2012 for a review), even when such a preliminary test is found to be significant, there is no guaranty that the trait value is estimable with sufficient accuracy to be useful in practice. Instead, we propose that predictive performance under phylogenetic modelling (P^2) be used instead. A potential additional inference test of the statistical significance of a phylogenetic signal for modelling purposes can be obtained, if deemed necessary, by comparing observed P^2 values under cross-validation against cross-validated values obtained from permuted sets (i.e. trait values are permuted across the phylogeny). In this way, significance of predictive power could be estimated by a one-tail permutation test of P^2.

As mentioned earlier, the intercept of regression models estimating trait values via PEM should not be confused with the mean trait value at the root of the phylogeny, as would be the case using generalized least-squares. The coordinates of the root being $\mathbf{s}_{\text{root}} = 0$, its scores ($\mathbf{u}_{\text{root}}$) are computed as:

$$\mathbf{u}_{\text{root}} = -[1/n]_{1\times n}\mathbf{B}^{*}\mathbf{D}_{\mathbf{w}}\mathbf{V}\mathbf{D}_{\Sigma}^{-1} \qquad \text{eqn 14}$$

and the PEM regression estimate of the trait value for the root would be obtained by estimating the regression equation for these scores. The regression intercept is the trait value when $\mathbf{u}_{\text{target}} = \mathbf{0}$ and whose location in the graph is therefore:

$$\mathbf{s}_{\text{intercept}} = [1/n]_{1\times n}\mathbf{B}^{*}\mathbf{D}_{\mathbf{w}}. \qquad \text{eqn 15}$$

In practice, these coordinates are not represented in the phylogenetic graph (neither at tips, nodes nor along edges). The regression intercept should therefore be regarded as the mean trait value in a case where no eigenvector is used (i.e. as a null model).

We mentioned earlier that it is wise to assume that not all traits are equal in terms of their predictability by phylogenetic models. For instance, an OU process will drive a wide range of phenotypic variation when σ is large, and/or when α is large and optimal trait value (θ) tends to shift by large extents at nodes (eqn. 10). In the latter scenario, traits for which optimum shifts occurred only recently (i.e. near the tips) will be not be as predictable as those whereby shifts occurred earlier in the phylogeny (i.e. near the root). In the situation where α is small, variation is mostly driven by random fluctuations that is phylogenetically structured. Traits showing high variability and that have been sampled over a sufficiently broad phylogeny are likely to be predictable by phylogenetic modelling. Such

condition is more likely to be the case among species from different classes or even phyla than among closely related species (e.g. from the same genus: Senior & Nakagawa 2013). We expect that phylogenetic modelling will be particularly useful in fields for which knowledge about traits for numerous species, spread over a wide taxonomic range, is available, as is often the case when assessing species sensitivity distribution (SSD, De Zwart, 2002; De Zwart & Posthuma 2005) or ecological risk (Faggiano *et al*. 2010; Carafa *et al*. 2011), for example.

Acknowledgements

We are thankful to our colleagues Steven C. Walker and Marie-Hélène Ouellette for their advice during several stages of the present work, and two anonymous reviewers whose advice are instrumental in improving the quality of our manuscript. G. Guénard received support from the *Fond Québécois de Recherche sur la Nature et la Technologie* (FQRNT). That project was enabled through NSERC grants no. 7738 to P. Legendre and support from the Canada Research Chair program to P. Peres-Neto.

References

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

Adams, S.M. & Breck, J.E. (1990) *Bioenergetics*, volume Methods for Fish Biology, chapter 12, pp. 389–415. American Fisheries Society, Bethesda, Maryland, USA.

Bartoszek, K., Pienaar, J.P.M., Andersson, S. & Hansen, T.F. (2012) A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Population Biology*, **314**, 204–215.

Beaulieu, J.A., Jhwueng, D.-C., Boettiger, C. & O'Meara, B.C. (2012) Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*, **66**, 2369–2383.

Blanchet, F.G., Legendre, P. & Borcard, D. (2008) Modelling directional spatial processes in ecological data. *Ecological Modelling*, **215**, 325–336.

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

Boettiger, C., Coop, G. & Ralph, P. (2012) Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*, **66**, 2240–2251.

Butler, M.A. & King, A.A. (2004) Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist*, **164**, 683–695.

Byrd, R.H., Lu, P., Nocedal, J. & Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, **16**, 1190–1208.

Carafa, R., Faggiano, L., Real, M., Munné, A., Ginebreda, A., Guasch, H., Fio, M., Tirapu, L. & von der Ohe, P.C. (2011). Water toxicity assessment and spatial pollution patterns identification in a mediterranean river basin district. tools for water management and risk analysis. *The Science of the Total Environment*, **409**, 4269–4279.

De Zwart, D. (2002) *Species Sensitivity Distributions in Ecotoxicology*, chapter Observed regularities in SSDs for aquatic species, pp. 133–154. Lewis Publishers, Boca Raton, Florida, USA.

De Zwart, D. & Posthuma, L. (2005) Complex mixture toxicity for single and multiple species: proposed methodologies. *Environmental Toxicology and Chemistry*, **24**, 2665–2676.

Desdevise, Y., Legendre, L., Azouzi, L. & Morand, S. (2003) Quantifying phylogenetically structured environmental variation. *Evolution*, **57**, 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.

Diniz-Filho, J.A.F., Bini, L.M., Rangel, T.F., Morales-Castilla, I., Olalla-Tárraga, M. Á., Rodríguez, M.A. & Hawkins, B.A. (2012) On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography*, **34**, 239–249.

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

Ezekiel, M. (1930) *Methods of Correlation Analysis*. John Wiley and Sons, New--York, USA.

Fagan, W.F., Pearson, Y.E., Larsen, E.A., Lynch, H.J., Turner, J.B., Staver, H., Noble, A.E., Bewick, S. & Goldberg, E. (2013) Phylogenetic prediction of the maximum *per capita* rate of population growth. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20130523.

Faggiano, L., de Zwart, D., Garcia-Berthou, E., Lek, S. & Gevrey, M. (2010) Patterning ecological risk of pesticide contamination at the river basin scale. *The Science of the Total Environment*, **408**, 2319–2326.

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

Freckleton, R.P. (2009) The seven deadly sins of comparative analysis. *Evolutionary Biology*, **22**, 1367–1375.

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

Gabriel, K.R. (1998) Generalized bilinear regression. *Biometrika*, **85**, 689–700.

Garland, T. & Ives, A. (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist*, **155**, 346–364.

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

Goldberger, A. (1962) Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, **57**, 369–375.

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

Guénard, G., von der Ohe, P.C., de Zwart, D., Legendre, P. & Lek, S. (2011) Using phylogenetic information to predict species tolerances to toxic chemicals. *Ecological Applications*, **21**, 3178–3190.

Hansen, T.F. (1997) Stabilizing selection and the comparative analysis of adaptation. *Evolution*, **51**, 1341–1351.

Hernández, C.E., Rodríguez Serrano, E., Avaria-Llautureo, J., Inostroza-Michael, O., Morales-Pallero, B., Boric-Bargetto, D., Canales-Aguirre, C.B., Marquet, P.A. & Meade, A. (2013) Using phylogenetic information and the comparative method to evaluate hypotheses in macroecology. *Methods in Ecology and Evolution*, **4**, 401–415.

Hurvich, C.M. & Tsai, C.-L. (1993) A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis*, **14**, 271–279.

Isaac, N.J.B., Jones, K.E., Gittleman, J.L. & Purvis, A. (2005) Correlates of species richness in mammals: body size, life history, and ecology. *The American Naturalist*, **165**, 600–607.

Legendre, P. & Legendre, L. (2012). *Numerical Ecology*. *Number 24 in Developments in Environmental Modelling*. Elsevier Science B.V., Amsterdam, The Netherlands, third English edition.

Lislevand, T. & Thomas, G.H. (2006) Limited male incubation ability and the evolution of egg size in shorebirds. *Biology Letters*, **2**, 206–208.

Makarenkov, V., Legendre, L. & Desdevise, Y. (2004) Modelling phylogenetic relationships using reticulated networks. *Zoologica Scripta*, **33**, 89–96.

Martins, E.P. (1996) Phylogenies, spatial autoregression, and the comparative method: a computer simulation test. *Evolution*, **50**, 1750–1765.

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

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K. & Thuiller, W. (2012) How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, **3**, 743–756.

Nocedal, J. & Wright, S. (1999) *Numerical Optimization*. Springer, New-York, USA.

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

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 phenotypic traits: New matrices of phylogenetic proximities. *Theoretical Population Biology*, **73**, 79–91.

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**, 559–572.

Purvis, A. & Rambaut, A. (1995) Comparative analysis by independent contrasts (caic): an Apple Macintosh application for analysing comparative data. *Computer Applications in the Biosciences*, **11**, 247–251.

Revell, L.J., Mahler, D.L., Peres-Neto, P.R. & Redelings, B.D. (2011) A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution*, **66**, 135–146.

Senior, A.M. & Nakagawa, S. (2013) A comparative analysis of chemically induced sex reversal in teleosts: challenging conventional suppositions. *Fish and Fisheries*, **14**, 60–76.

Slater, G.J., Harmon, L.J., Wegmann, D., Joyce, P., Revell, L.J. & Alfaro, M.E. (2012) Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate bayesian computation. *Evolution*, **66**, 752–762.

Uhlenbeck, G.E. & Ornstein, L.S. (1930) On the theory of the Brownian motion. *Physical Review*, **36**, 823–841.

West, D.B. (2001) *Introduction to Graph Theory*, 2nd edn. Prentice Hall, Upper Saddle River, New Jersey, USA.

Zheng, L., Ives, A.R., Garland, T., Larget, B.R., Yu, Y. & Cao, K. (2009) New multivariate tests for phylogenetic signal and trait correlations applied to eco-physiological phenotypes of nine Manglietia species. *Functional Ecology*, **23**, 1059–1069.

Received 11 January 2013; accepted 21 August 2013
Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Data S1. R script and data files used obtain the analyses, examples and figures of the present paper.