Journal of Theoretical Biology 264 (2010) 693–701

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

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

ARTICLE INFO

Article history:
Received 20 January 2010
Received in revised form
25 March 2010
Accepted 25 March 2010
Available online 31 March 2010

Keywords:
Phylogenetic principal component analysis
pPCA
Multivariate
Comparative method
Phylogenetic signal

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 *adephylo* package for the free software R.

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

* Corresponding author.
E-mail address: t.jombart@imperial.ac.uk (T. Jombart).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.03.038

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 (1948, 1950) I, an index originally used to measure spatial autocorrelation, has been proposed for measuring phylogenetic autocorrelation (Gittleman and Kot, 1990). Adapting the former definition (Cliff and Ord, 1973, p. 13) to the phylogenetic context, I is defined as

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

where $\mathbf{x}$ is the centred vector of a trait observed on n taxa, 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:

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

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-standardising 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 (Eq. (1)). The essential difference between the two approaches is that autoregressive models perform the regression of $\mathbf{x}$ onto $\mathbf{Wx}$, while I computes the inner product between both vectors (numerator of Eq. (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 centred spatial weighting matrix (Dray et al., 2006). Applied to a matrix of phylogenetic proximity $\mathbf{W}$, this method yields uncorrelated variables modelling 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