**Generalized phylogenetic fuzzy weighting: unravelling phylogenetic imprints on species distribution across space**

Leandro Duarte, Gabriel Nakamura, Lucas Jardim, José Alexandre F. Diniz-Filho

*Introducing Phylogenetic Fuzzy Weighting*

Given a phylogenetic tree, Duarte et al. (2016) derived matrix **S** of phylogenetic similarities (*sij*) between every pair of species in the tree from matrix **D** of species pairwise patristic distances (*dij*), as

Where max *dij* is the largest phylogenetic distance in **D**.

Matrix **S** expresses phylogenetic divergences for each pair of species in the phylogeny. Standardizing each *sij* value in matrix **S** (including the main diagonal) by column totals, breaks its symmetry, generating the elements *qij* of matrix **Q**, as follows:

This procedure is the basis of phylogenetic fuzzy weighting (PFW, Pillar & Duarte 2010; Duarte et al. 2016), which builds upon previously described methods that apply fuzzy set theory to ecological analysis (Dale 1977; Feoli & Zuccarello 1986; Richards 1986; Pillar & Orlóci 1991; Pillar et al. 2009). Each diagonal element in **Q** describes self-belonging of a given species, which is a measure of phylogenetic uniqueness (Duarte et al. 2016). Off-diagonal elements in **Q** describe pairwise degree of phylogenetic cross-belonging of each species *i* to any other species *j*. Phylogenetic cross-belonging captures, in a single value, the amount of phylogenetic divergence between *i* and *j*, and also the rate of diversification between *i* and the ancestral node connecting *i* to *j*. If the path linking species *i* to the ancestral node presents a higher rate of diversification than the path connecting *j* to the ancestral node, then species *i* will show lower degree of belonging to *j* than *j* to *i*. Matrix **Q** expresses, simultaneously, symmetric phylogenetic covariances among species and also asymmetric diversification trajectories connecting them, which implies that the topological arrangement of the phylogenetic tree is also captured by matrix **Q** (Pillar & Duarte 2010; Duarte et al. 2016).

In PFW, the abundances/frequencies of species *j* across a set of sites *k* (matrix **W**) are weighted based on matrix **Q** by the matrix multiplication **P** = **QW**. The resulting matrix **P** describes each site by phylogenetically weighted species composition (Pillar & Duarte 2010; Duarte et al. 2016), and has been used to perform phylogenetic gradient analysis (sensu Peres Neto & Kembel 2015) across local communities (e.g. Gianuca et al. 2012; Seger et al. 2017; Viany et al. 2017; Iserhard et al. 2018; Spaniol et al. 2019; Drose et al. 2019), within biomes (e.g. Brum et al. 2012; Duarte et al. 2014a; Portillo et al. 2019) and broader biogeographical scales (e.g. Brum et al. 2014; Maestri et al. 2015; Duarte et al. 2014b; Carlucci et al. 2017). Each element in **P** (*pik*) characterizes each site *k* by phylogeny-weighted abundance/frequency of each species *i*, and were originally computed as follows:

Pairwise dissimilarities between sites computed from matrix **P** has been demonstrated to be an appropriate measure of phylogenetic beta diversity (e.g. Duarte et al. 2014a), with higher statistical robustness when compared to other available measures (Duarte et al. 2016).

*PFW and Grafen’s branch length transformation*

As seen above, phylogenetic similarities used in PFW framework are computed as the complement of *dij*/max *d*ij ratio, which is equivalent to the height *h* defined by Grafen (1989) as the node value where a given pair of species diverge in the phylogenetic tree. Accordingly, for any pair of species diverging in the root of the tree, *dij*/max *d*ij = 1, which is the maximum possible value for *h*. For any tip in the phylogenetic tree, *h* = 0. This implies that **S** is equivalent to the standardized phylogenetic variance/covariance matrix **V**.Thus, each off-diagonal element in **S** describes the phylogenetic correlation between a given pair of species in the tree, and can be computed based on Grafen (1989) as

Where ρ is a coefficient that accelerates (ρ < 1) or decelerates (ρ > 1) the rate of phylogenetic diversification along the phylogeny (Grafen 1989). As ρ tends to zero, phylogenetic covariance between species also tends to zero, and **S** approximates to the identity matrix **I**. In such case, every species varies independently of all others in the phylogeny along evolutionary history. In such case, any phenotypic trait evolving under ρ < 1 will show lower phylogenetic signal than expected by Brownian motion (sensu Blomberg et al. 2002; see Diniz-Filho et al. 2012). Otherwise, when ρ > 1, phylogenetic covariance gets concentrated towards the tips of the tree. That is to say that closely related species tend to show increasing phylogenetic covariance as ρ increases, while far related species remain showing low covariance. In such case, any phenotypic trait evolving under ρ >1 will show higher phylogenetic signal than expected by Brownian motion (Diniz-Filho et al. 2012). When ρ = 1, *h* equals to the original node values in the phylogeny. Not surprisingly, any phenotypic trait evolving under ρ = 1 will behave as expected by Brownian motion model of trait evolution (Diniz-Filho et al. 2012). The PFW framework developed so far uses the original *h* values of the phylogenetic tree in order to derive **S** and **Q** matrices (ρ = 1).

Matrix **Q** can be generalized in order to allow different ρ values. We call this generalized matrix as **X** (read chi), whose elements (χij) are computed as follows

Matrix **X** allows for different rates of evolutionary change along phylogenetic nodes. For instance, for the set of Grafen’s ρ values = (0.01, 0.1, 1, 1.5, 2), the respective set of matrices **P** = (**P**0.01, **P**0.1, **P**1, **P**1.5, **P**2) can be computed for any species by site matrix **W** via the matrix multiplication **X**ρ**W**, or

*GPFW and James Cheverud’s phylogenetic autoregressive method*

The seminal article of Cheverud (1985) defined the basis of phylogenetic autoregressive models (phyARM) as a tool for analysis of phylogenetic signal in variation of a phenotypic trait *y*. Cheverud’s framework is based on estimating phylogenetic signal as

where ρC is the autoregressive coefficient that expresses the phylogenetic effect size on trait variation. The subscript C refers to Cheverud and intends to differentiate the autoregressive coefficient from Grafen’s ρ previously defined. **Q**C (C also refers to Cheverud) is a matrix derived from the phylogenetic variance/covariance matrix **S**, which is akin to matrices **Q** and **X** previously described. But differently from **Q/X**, **Q**C is standardized based only on the off-diagonal elements of **S** (Cheverud 1985). The product **Q**C*y* provides a vector of predicted trait values given phylogenetic covariances among species. The vector **s** expresses the trait variation not explained by phylogeny, also called specific component.

At this point the reader likely noticed a common theoretical basis for both PFW and phyARM. Actually, matrix **P** and **Q**C*y* express both expected values of species abundances/frequencies per site and species traits, respectively, given phylogenetic covariances among species. Since **P** = **QW** (or = **XW**), we can express (G)PFW as a special case of phyARM in which ρ = 1. Since we have now a generalized fuzzy set matrix **X**, we are able to estimate the Grafen’s ρ coefficient that best adjusts to species distribution across sites. By doing so, we are able to estimate the evolutionary imprint on species distribution across sites.

For this, a first step consists in rewriting the generalized version of phylogenetic fuzzy weighting framework (GPFW) according to the autoregressive framework, as follows:

The elements of each matrix are thereby computed as

We can compute the effect size of phylogeny on species distribution across sites as a R2, which is defined as a ratio between the expected (MS**XW**) and the observed (MS**W**) mean squares (MS)

Those mean squares can be computed using the method proposed by Legendre & Cáceres (2013). The residual matrix **Ε** is describes species abundances/frequencies not explained by phylogeny and is therefore equivalent to the specific component **s** of phyARM. Each element of the residual matrix **Ε** is computed as

As the importance of phylogenetic relationships to drive species distribution across sites increases, predicted species abundances/frequencies tend to approximate the observed values, and εik tends to zero. Thus, the mean square error (MS**Ε**) is expected to decrease. At this point, we must remember that as ρ value used to define **X** tends to zero, **X** approximates identity matrix **I** (see *PFW and Grafen’s branch length transformation*), which leads MS**E** to tend to zero, and therefore the respective R2 to tend to unity. This point will be relevant in further considerations on statistical approaches.

*Statistical tests*

For any ρ value used to define matrix **X**, a null model-based hypothesis test can be performed to evaluate the influence of phylogenetic relationships on the distribution of species across the sites. The null hypothesis H0 of no effect of phylogeny on species distribution across sites can be tested as follows: 1) Compute an observed F statistic (FObs)for a model built using a given ρ value to define a **X** matrix ass

2) Freely permute species names in the phylogenetic tree a number of times (say 999 permutations) (see Kembel *et al.* 2010). This procedure follows the null model adopted in Pillar & Duarte (2010) and Peres-Neto, Leibold & Dray (2012).

3) At each permutation, compute Fnull.

4) Define the probability of obtaining the observed statistic by chance (H0 = FObs ≤ FNull), as the proportion of permutations in which the null test statistics (FNull) exceeded the observed one.

Note that alternative ρ values can be used to define matrix **X**, which means that for any set of ρ values, a respective set of hypothesis tests can be performed.