Forum

Ecography 36: 001–006, 2013
doi: 10.1111/j.1600-0587.2013.00528.x

Subject Editor: Nathan J. Sanders. Accepted 26 July 2013

Phylogenetic imputation of plant functional trait databases

Nathan G. Swenson

N. G. Swenson (swensonn@msu.edu), Dept of Plant Biology, Michigan State Univ., East Lansing, MI 48824, USA.

Continental-scale maps of plant functional diversity are a fundamental piece of data of interest to ecosystem modelers and ecologists, yet such maps have been exceedingly hard to generate. The large effort to compile global plant functional trait databases largely for the purpose of mapping and analyzing the spatial distribution of function has resulted in very sparse data matrices thereby limiting progress. Identifying robust methodologies to gap fill or impute trait values in these databases is an important objective. Here I argue that existing statistical tools from phylogenetic comparative methods can be used to rapidly impute values into global plant functional trait databases due to the large amount of phylogenetic signal often in trait data. In particular, statistical models of phylogenetic signal in traits can be generated from existing data and used to predict missing values of closely related species often with a high degree of accuracy thereby facilitating the continental-scale mapping of plant function. Despite the promise of this approach, I also discuss potential pitfalls and future challenges that will need to be addressed.

Theoretical and empirical ecological investigations suggest a strong linkage between plant functional diversity and ecosystem function (Tilman et al. 1997, Diaz and Cabido 2001). The distribution of functional diversity across a variety of spatial scales is therefore of fundamental interest to ecosystem modelers and ecologists (Reich 2005, Westoby and Wright 2006, Elser et al. 2010). Quantifying the continental-scale distribution of plant functional diversity has, however, been particularly challenging due to the patchiness inherent in global scale plant functional trait databases (ter Steege et al. 2006, Swenson and Weiser 2010, Swenson et al. 2012). Indeed even the largest and most comprehensive plant functional trait database compiled to date has an extraordinarily sparse data matrix (Kattge et al. 2011). This lack of species-level functional information has generally limited ecosystem modelers to characterizing the vegetation within an entire map grid cell using only a few plant functional types. The reduction of an entire flora in a map grid cell to a few functional types is clearly undesirable and may result in uncertain projections regarding the fate of ecosystems under global climate change (Moorcroft 2006, Purves and Pacala 2008).

The most obvious obstacle in estimating the continental-scale distribution of plant functional diversity is that it requires species-level functional trait data for hundreds or thousands of species distributed across a vast area. In well-studied temperate floras, species-level functional trait values and geographic distributions for most species may have already been reported in the literature and databased making large-scale mapping of plant function more feasible (Swenson and Weiser 2010), but in other systems critical for our global carbon cycle (e.g. tropical forests) strikingly little species-level trait information is available. Such data may require many years or decades to collect for diverse tropical floras. Given the current pressures on our ecosystems and our desire to model their future, an alternative solution to facilitate present research is necessary.

A potentially powerful and more easily employed alternative approach for imputing plant functional trait values in global databases is to take advantage of the phylogenetic signal in functional trait data to estimate species-level function. Phylogenetic signal is a term used to signify the degree to which closely related species tend to have similar trait values (Blomberg et al. 2003). A high degree of phylogenetic signal indicates closely related species tend to be more similar in their trait values than distantly related species, while a low degree of phylogenetic signal indicates closely related species are not similar. Plant ecologists have demonstrated a remarkably high degree of phylogenetic signal in global-scale plant functional trait databases (Moles et al. 2005, Donoghue 2008, Chave et al. 2009, Swenson 2011) suggesting that estimating the trait values based on their phylogenetic position may be reasonable. This has led some to predict that phylogenies may prove to be powerful tools in future global change and ecosystem research (Edwards et al. 2007, Cavender-Bares et al. 2009).

Here I discuss the potential of phylogenetically-informed imputation methods for filling in sparse global plant functional trait databases. Specifically, I discuss four phylogenetic

The review and decision to publish this paper has been taken by the above noted SE. The decision by the handling SE was shared by a second SE.

approaches for imputing plant trait data in global databases with special focus being placed on two methods – phylogenetic generalized linear models and phylogenetic eigenvectors. The ultimate goal of this discussion is to determine the most effective and pragmatic way to gap fill global trait databases that ecologists and vegetation modelers can utilize for mapping and modeling plant functional diversity, ecosystem function and ecosystem response to global change (Diaz and Cabido 2001, Baker et al. 2004, 2009, Reich 2005, Mahli et al. 2006, Moorcroft 2006, ter Steege et al. 2006, Purves and Pacala 2008, Kattge et al. 2011).

Imputation of trait values in databases and phylogenetic signal

Immediately upon compiling these large trait databases it became clear to many that the trait matrices would be exceedingly sparse particularly for taxa in tropical latitudes that are the most crucial to model given their central role in the biogeochemical cycles (Swenson et al. 2012). One way to fill in these sparse matrices is to conduct massive trait inventories on continental scales (Baker et al. 2004, Reich 2005), but this approach will likely prove difficult in tropical regions particularly given the urgent need to produce stronger vegetation models. Thus plant functional ecologists and ecosystem modelers are forced to explore alternative methods for dealing with super sparse trait matrices. One approach could be to not use trait data that are sparse, but this returns modelers to the problem of using big green leaves. A second approach that is beginning to be increasingly used by ecosystem ecologists is to estimate or impute missing trait values. For example, the RainFor research group, that is responsible for the majority of the in depth investigations into carbon stocks and fluxes in the Amazon Basin, has imputed trait values for species based on the mean trait value for the genus or family of that species (Baker et al. 2004, 2009, Mahli et al. 2006, ter Steege et al. 2006). In general, they have shown that this imputation method may not introduce too much bias into their analyses. Additional work by Shan et al. (2012) has utilized hierarchical models based on taxonomic ranks and trait co-variance to predict missing trait values in global databases. Importantly, such an approach acknowledges and makes use of phylogenetic signal in trait data through the use of mean trait values for a taxonomic rank. The downside of such a taxonomic approach though is that it lacks information regarding branch lengths and could therefore be refined using phylogenetic trees.

An alternative approach used by other functional ecologists in the Amazon (Baraloto et al. 2010) is multivariate imputation using chained equations (MICE), which uses a Gibb's sampler and multivariate trait relationships (Rubin 1987, 1996). A benefit of using an approach such as MICE is that one can impute trait values for individuals and not species as a whole thereby potentially refining predictions particularly if other trait data for that individual exist. A potential weakness of the MICE approach is that it requires multiple traits in the database and a strong co-variance structure. Further it does not take advantage of the potentially high degree of phylogenetic signal in trait data and therefore will likely not provide as reliable estimates as phylogenetic methods that incorporate both phylogenetic signal and trait co-variance.

Here I suggest an alternative imputation approach that explicitly incorporates phylogenetic information to impute trait values could be used. Specifically, a framework already exists to build a statistical model of how well the phylogenetic tree fits the trait matrix and to use this model to impute missing trait values given their phylogenetic placement (Fig. 1). These phylogenetic imputation methods build off of well-established and tested phylogenetic methods in comparative biology, but they have not been transferred to the realm of imputing trait values in large global plant functional trait databases built for the purposes of mapping and modeling species and ecosystem function on the globe. In the following sections I will describe in detail two such phylogenetic imputation approaches, phylogenetic generalized linear models and phylogenetic eigenvectors, and two additional approaches in less detail. My goal here is to simply outline phylogenetic methods that could be utilized for imputation to promote discussion, testing and the promotion of certain methods and not to make a judgment regarding their relative merits at this point in time.

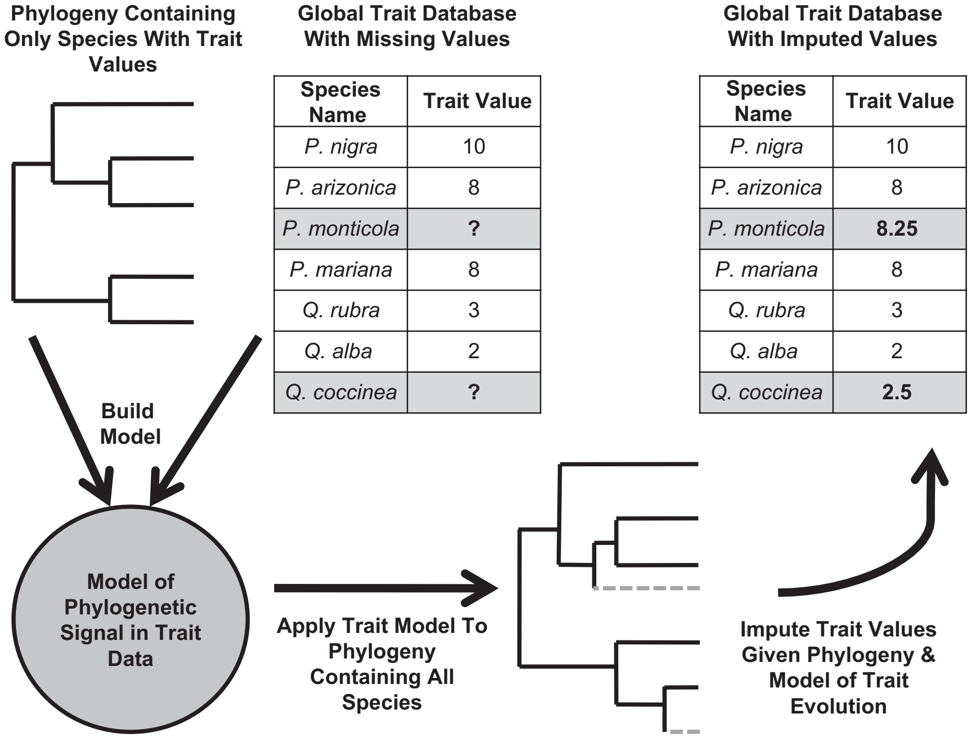
Phylogenetic generalized linear models

The first phylogenetic imputation method that I will discuss utilizes phylogenetic generalized linear models (pGLMs) (Martins and Hansen 1997, Garland and Ives 2000), which should produce identical results to the re-rooting procedure of Garland and Ives (2000). This generalized linear model can be written as:

$$y = \beta x + e$$

where trait y is regressed onto trait x with a slope of β and an error structure (e) defined by a phylogenetic variance–covariance matrix. The phylogenetic variance–covariance matrix is determined by a model of trait evolution. In the simplest scenario trait evolution is expected to follow a random walk or Brownian Motion (Felsenstein 1985). Here the expected co-variance in a trait for two species is equal to the total shared branch length and the expected variance is the branch length from root to tip on the phylogeny (i.e. the time for evolution). In more sophisticated analyses no model of trait evolution is assumed, rather a model of trait evolution is fit to the data using maximum likelihood and this fit model is used to inform the phylogenetic variance–covariance matrix (Freckleton et al. 2002, 2011).

In both of the pGLM approaches above, with a Brownian Motion or a fitted model of trait evolution, one trait is regressed onto another. Using this approach one could impute a value for trait y given a value for trait x (e.g. another functional trait or perhaps the mean climate of a species) and the phylogenetic variance–covariance matrix. Thus if one had a trait database with two correlated traits such as specific leaf area and leaf %N or the average climate of a species, but missing trait values for one trait for a species, the pGLM model could be used to predict or impute the value for the second trait (Garland and Ives


Phylogeny Containing Only Species With Trait Values
Global Trait Database With Missing Values
Global Trait Database With Imputed Values
Build Model
Model of Phylogenetic Signal in Trait Data
Apply Trait Model To Phylogeny Containing All Species
Impute Trait Values Given Phylogeny & Model of Trait Evolution

Species Name	Trait Value
P. nigra	10
P. arizonica	8
P. monticola	?
P. mariana	8
Q. rubra	3
Q. alba	2
Q. coccinea	?

Species Name	Trait Value
P. nigra	10
P. arizonica	8
P. monticola	**8.25**
P. mariana	8
Q. rubra	3
Q. alba	2
Q. coccinea	**2.5**

Figure 1. A workflow for imputing missing functional trait values using phylogenetic information. The workflow begins with a trait database with missing values and a phylogenetic tree containing all species with known trait values. Next this data is used to construct a phylogenetic generalized linear model, a phylogenetic eigenvector model or some other statistical model of phylogenetic signal in trait data using a single trait or the known co-variance between multiple traits. This model is then applied to a phylogenetic tree containing all species with known and unknown trait values to predict or impute the missing trait values. The imputed values are then inserted into the global database at which point they can be joined to spatial information to facilitate the continental-scale mapping of plant function.

2000). This helps to hone the prediction by adding in additional constraints regarding known trait co-variance and trait–environment relationships. To do this a phylogenetic variance–covariance matrix is estimated for the dataset of observed values and then used to inform a phylogenetic variance–covariance matrix including all species with and without trait values. This larger variance–covariance matrix and the known trait values for *x* can then be used to predict or impute the missing *y* trait values. In those instances where the researcher desires to impute a single trait (*y*), while not relying on its correlation with a second trait (*x*), the single trait could be regressed onto the phylogenetic error structure (i.e. the variance–covariance matrix) to generate a model that could predict or impute missing values for that trait and an estimated error in the predicted value. This estimated error can be utilized in downstream analyses when estimating values such as the mean trait value or the functional diversity of all species in a map grid cell. Another benefit of the pGLM approach is that it has been expanded to incorporate known intraspecific variation in trait data making it feasible to not just utilize a species means for imputation (Ives et al. 2007).

Phylogenetic eigenvectors

The second phylogenetic imputation method that I will discuss utilizes phylogenetic eigenvector regression models (pERMs) (Diniz-Filho et al. 1998, Kuhn et al. 2009). This method generates phylogenetic eigenvectors by conducting a principal components analysis (PCA) on a distance matrix representing the phylogenetic branch length separating each species in the phylogeny. This results in one PCA axis per species in the phylogeny. The location of species on the PCA axes (i.e. phylgenetic eigenvectors) are then used as predictors in a multiple linear regression with a trait as the response variable:

$$y \sim PEV + e$$

where *PEV* are the phylogenetic eigenvectors. Not all PEVs are utilized in the linear model to avoid over-fitting. The broken stick method for selecting eigenvectors (Frontier 1976, Legendre and Legendre 1998) is used to select the PEVs to be used in the model where the number of eigenvectors used is equal to point at which the variance explained by the observed PEVs is less than that expected under a broken stick model (i.e. 50%, 75%, 87.5%, etc. of the variance explained for PC1, PC2, PC3, etc.) (Diniz-Filho et al. 1998). The pERM method could be expanded to include a second trait as:

$$y \sim x + PEV + e$$

where the second trait or mean climate, *x*, simply becomes an additional predictor variable in the multiple linear model with the PEVs. Thus the pERM approach can also be honed by constraining the prediction using additional independent variables. Intraspecific trait variable could also be included into the pERM approach by re-running the regression analysis, for example, 1000 times each time drawing for the known distributions of trait values for each species in the database. Similar re-sampling approaches have been used in the past for maximum likelihood reconstruction of the ancestral climatic niche on a phylogenetic tree using the probability of occurrence of species along climatic axes (Evans et al. 2009)

This phylogenetic eigenvector approach has generally been used to quantify phylogenetic signal in datasets, effectively the R^2 of the model, but it can also be utilized to predict or impute values. To accomplish this the user must first generate a phylogenetic tree containing all species of interest including those with trait values and those without. Next, PEVs are generated from this entire phylogeny, but the linear model is constructed only using the PEVs for species with known trait values. The model derived is then used to predict or impute the missing trait values by feeding the model the PEVs of the species lacking trait values. It is important to note that unlike the pGLM approach described in the previous subsection, the pERM approach described here does not fit or assume a model of trait evolution. This lack of a model of trait evolution and the requirement to not use all PEVs in the regression model have been criticized and suggest that pERMs may not perform as well as pGLMs (Rohlf 2001, Freckleton et al. 2011). Thus, care must be taken when using pERMs particularly when the goal is to quantify the phylogenetic signal itself. That said, if the goal is to simply impute traits without want for making particular inferences about the tempo and mode of trait evolution itself, then it might be an effective method. Further, recent research has demonstrated that spatial autocorrelation may also be inserted into pERM approaches (Kuhn et al. 2009) thereby providing an additional level of information to help refine imputation.

Other existing approaches

The two approaches described above will likely become the most widely employed phylogenetic methods for imputing functional trait data in global databases, but here I will discuss two additional methods that have been developed and are already implemented through web interfaces. The first builds statistical models regressing the probability of sharing a character between two species and their phylogenetic distance. For example, what is the probability that two species share the same pathogen or herbivore and how does that probability relate to phylogenetic distance? To address this exact question Gilbert et al. (2012) utilized a global database of plant–pest interactions and quantified the degree of overlap in pest interactions between all plant species in the database. This overlap was then regressed onto the phylogenetic distance between plant species. The research by Gilbert et al. (2012) showed that there was a strong and significant decline in shared pests with phylogenetic distance. The linear model that was used to describe this relationship is now the basis for a web application into which a user can input a list of species to determine the probability that the plants in the list will share pests. This approach utilizes no specific model of trait or interaction evolution and is still being cross-validated, but it represents an easy and clever way to utilize phylogenetic signal in data to inform those working in the applied realm making management decisions.

The second approach that is already implemented as a web service largely builds off of the conceptual and statistical foundations of phylogenetic generalized linear models. Specifically, Bruggeman et al. (2009) begin with the assumption that the characters evolve under a Brownian Motion model. This assumption and the corresponding phylogenetic variance–covariance matrix are the same components as those used in a phylogenetic generalized linear model using Brownian Motion. However, the advance that Bruggeman et al. (2009) implement is the introduction of predicting the intraspecific variation in traits such that the imputed trait values have a mean expectation with an associated error. This advance is useful in that it allows the user to draw from a probability distribution for sensitivity analyses in any downstream analyses (Bruggeman 2011) whereas other methods provide a single predicted value with no reported error. The web service for this approach, PhyloPars, takes a trait matrix containing one or many traits and a phylogenetic tree and reports out the imputed trait matrix (< www.ibi.vu.nl/programs/phylopars/ >). The clear downside of this approach is that to date it rests solely on the assumption of Brownian Motion presumably for mathematical expediency and therefore doesn't allow the user to compete alternative models of trait evolution. Though I expect this limitation to be short-lived and that the application will consequently become even more widely used.

Conclusions, future challenges and prospects

There has been a considerable amount of time and money spent collecting functional trait data and compiling it into global databases and this effort has facilitated a considerable amount of strong research. Despite this effort and continuing trait collection efforts, the databases will undoubtedly remain very sparse particularly in tropical ecosystems. Thus, imputing or estimating trait values will be necessary if, in the near term, ecologists and modelers want to refine their analyses beyond functional groups and big green leaf approaches. In this article I have suggested that phylogenetic imputation methods represent a pragmatic and potentially very powerful approach for filling in sparse global functional trait databases. In every case, there will be error in the estimates and it will be up to the users to determine whether that degree of error is tolerable, but in most cases I would argue that when generating a vegetation model that uses a map with a distribution of functional traits with some error in each map grid cell is far better that using a map with only a few functional types per map grid cell particularly in hyper-diverse ecosystems.

If ecologists and modelers do continue to increase their usage of phylogenetic imputation methods to fill in their trait databases a number of challenges will have to be overcome. Many of these challenges concern the magnitude of error in the phylogenetic and trait data and how this error is potentially propagated into downstream analyses. First, the generation of large phylogenetic trees that are reliable will become a challenge. Methods currently exist for generating large phylogenetic trees from species lists (Webb and Donoghue 2005), existing systematic treatments (Beaulieu et al. 2012) or super-sparse sequence matrices (Smith et al. 2009) and all have associated weaknesses and are not always easily implemented by a non-phylogeneticist. As the quality of the imputation critically relies on the quality of the phylogeny, great care will be needed when generating the phylogenetic trees for this work and the interpretation of the results. For example, the lack of resolution within genera that is commonplace in a Phylomatic phylogeny (Webb and Donoghue 2005) will generally result in little-to-no variation in trait values within genera using many imputation methods and will therefore artificially reduce the functional diversity of assemblages containing con-geners.

Second, any taxonomic biases, sampling artifacts or measurement error in existing trait databases will increase imputation errors and propagate error in downstream analyses. For example, one would not want to impute the trait values for all ~ 250 000 angiosperms from a 'global' database of ~ 100 species particularly if those species were from only a handful of eudicot families. Importantly, traditional cross-validation approaches may reveal little error in such a situation. Thus, prior to any imputation there should be a careful consideration of the taxonomic sampling. This highlights the importance of the continued compilation of large trait databases such as TRY (Kattge et al. 2011) and the importance of additional field campaigns to collect trait data from undersampled floras.

Third, although many traits of interest to ecologists have been found to have a fair deal of phylogenetic signal in global databases, many traits will not have signal and phylogenetic imputation will be a pointless exercise particularly without constraining the imputation by information about additional traits or climate. For example, traits related to perhaps reproductive isolation or defense may not have enough phylogenetic signal to be imputed with a high degree of confidence. Conversely, it may be that the variation in these traits within lineages may be low enough when compared to the global variation to permit reasonable estimates of trait values. This raises another point of concern – if a researcher is interested in fine scale differences in trait values and perhaps relating those differences to species-to-species resource partitioning or species diversification an imputed trait data matrix may be worthless. That is, all of the interesting finer scale biological differentiation between closely related species will be effectively 'washed out' using imputation methods that will likely capture more 'basal' signal and use that information to impute trait values. Research will therefore be needed to determine the magnitude of error at the species-level and whether this range of error may encapsulate the magnitude of the effect size a researcher is interested in detecting. In those cases whether the effect size is small relative to the imputation error – imputed trait values should not be used for the study. In sum, the user should be clear on the limitations of this approach and utilize the imputed information for situations where the loss of fine scale information has fewer consequences.

Lastly, if phylogenetic imputation methods are to really spur the mapping of plant function on global scales a great deal of spatial information will be needed. In other words, an imputed trait matrix is only useful if there is spatial information to which the trait data can be pinned. Unfortunately as with trait data matrices, spatial data matrices are exceedingly sparse for most non-vertebrates. Sparse spatial data matrices have spurred ecologists to implement spatial imputation methods (e.g. species distribution models) and it is likely that in order to map the continental-scale distribution of plant function we will need to combine heavily imputed trait and spatial databases. While this will propagate error, I argue that this level of error will likely pale in comparison to the error generated by the assumption that an entire region is composed of one or a few functional types.

In sum, the imputation of global functional trait databases is a needed next step in order to facilitate the mapping of the continental-scale function. Research over the last decade has demonstrated there is often a high degree of phylogenetic signal in functional trait data on global scales. I argue that this signal can be utilized in many cases to impute global trait databases to permit the mapping of plant function given the great advances recently made in compiling large trait databases (Kattge et al. 2011) and generating large phylogenetic trees. Future work will need to cross-validate imputation models and conduct simulation-based research to determine when, where and why phylogenetic imputation methods are useful or not useful, but I argue that a robust framework already exists for ecologists to begin to fill in the gaps in global databases on the basis of phylogenetic information.

Acknowledgements – I would like to thank J. C. Svenning for discussions regarding mapping the distribution and diversity of plant function of temperate zone trees. I would like to thank R. Freckelton for discussions regarding the calculation of phylogenetic generalized least squares regression and trait imputation. I would like to thank G. Gilbert for discussions about predicting plant–pest interactions from phylogenetic signal in large datasets. Michigan State Univ. has supported my work on mapping the geographic distribution of plant function.

References

Baker, T. R. et al. 2004. Variation in wood density determines spatial patterns in Amazonian forest biomass. – Global Change Biol. 10: 545–562.

Baker, T. R. et al. 2009. Do species traits determine patterns of wood production in Amazonian forests? – Biogeosciences 6: 297–307.

Baraloto, C. et al. 2010. Decoupled leaf and stem economics in rain forest trees. – Ecol. Lett. 13: 1338–1347.

Beaulieu, J. M. et al. 2012. Synthesizing phylogenetic knowledge for ecological research. – Ecology 93: S4–S13.

Blomberg, S. P. et al. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. – Evolution 57: 717–745.

Bruggeman, J. 2011. A phylogenetic approach to the estimation of phytoplankton traits. – J. Phycol. 47: 52–65.

Bruggeman, J. et al. 2009. PhyloPars: estimation of missing parameter values using phylogeny. – Nucl. Acids. Res. 37: W179–W184.

Cavender-Bares, J. et al. 2009. The merging of community ecology and phylogenetic biology. – Ecol. Lett. 12: 693–715.

Chave, J. et al. 2009. Towards a worldwide wood economics spectrum. – Ecol. Lett. 12: 351–366.

Diaz, S. and Cabido, M. 2001. Vive la difference: plant functional diversity matters to ecosystem processes. – Trends Ecol. Evol. 16: 646–655.

Diniz-Filho, J. A. F. et al. 1998. An eigenvector method for estimating phylogenetic inertia. – Evolution 52: 1247–1262.

Donoghue, M. J. 2008. A phylogenetic perspective on the distribution of plant diversity. – Proc. Natl Acad. Sci. USA 105: 11549–11555.

Edwards, E. J. et al. 2007. The relevance of phylogeny to studies of global climate change. – Trends Ecol. Evol. 22: 243–249.

Elser, J. J. et al. 2010. Biological stoichiometry of plant production: metabolism, scaling, and ecological response to global change. – New Phytol. 186: 593–608.

Evans, M. E. K. et al. 2009. Climate, niche evolution, and diversification of the "bird-cage" evening primroses (Oenothera, Sections Anogra and Kleinia). – Am. Nat. 173: 225–240.

Felsenstein, J. 1985. Phylogenies and the comparative method. – Am. Nat. 125: 1–15.

Freckleton, R. P. et al. 2002. Phylogenetic analysis of comparative data: a test and review of evidence. – Am. Nat. 160: 712–726.

Freckleton, R. P. et al. 2011. Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. – Am. Nat. 178: E10–E17.

Frontier, S. 1976. Etude de la decroissance des valeurs propres dans une analyse en composantes principales: comparaison avec le modele du baton brise. – J. Exp. Mar. Biol. Ecol. 25: 67–75.

Garland, T. Jr and Ives, A. R. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. – Am. Nat. 155: 346–364.

Gilbert, G. S. et al. 2012. Evolutionary tools for phytosanitary risk analysis: phylogenetic signal as a predictor of host range of plant pests and pathogens. – Evol. Appl. 5: 869–878.

Ives, A. R. et al. 2007. Within-species measurement error in phylogenetic comparative methods. – Syst. Biol. 56: 252–270.

Kattge, J. et al. 2011. TRY – a global database of plant traits. – Global Change Biol. 17: 2905–2935.

Kuhn, I. et al. 2009. Combining spatial and phylogenetic eigenvector filtering in trait analysis. – Global Ecol. Biogeogr. 18: 745–758.

Legendre, P. and Legendre, L. 1998. Numerical ecology, 2nd English ed. – Elsevier.

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

Moles, A. T. et al. 2005. A brief history of seed size. – Science 307: 576–580.

Moorcroft, P. R. 2006. How close are we to a predictive science of the biosphere? – Trends Ecol. Evol. 21: 400–407.

Purves, D. W. and Pacala, S. W. 2008. Predictive models of forest dynamics. – Science 320: 1452–1453.

Reich, P. B. 2005. Global biogeography of plant chemistry: filling in the blanks. – New Phytol. 168: 263–268.

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

Rubin, D. B. 1987. Multiple imputation for nonresponse in surveys. – Wiley.

Rubin, D. B. 1996. Multiple imputation after 18 + years. – J. Am. Stat. Assoc. 94: 473–489.

Shan, H. et al. 2012. Gap filling in the plant kingdom – trait prediction using hierarchical probabilistic matrix factorization. – Proceedings on the 29th International Conference on Machine Learning.

Smith, S. A. et al. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. – BMC Evol. Biol. 9: 37.

Swenson, N. G. 2011. The role of evolutionary processes in producing biodiversity patterns, and the interrelationships between taxonomic, functional and phylogenetic biodiversity. – Am. J. Bot. 98: 472–480.

Swenson, N. G. and Weiser, M. D. 2010. Plant geography upon the basis of functional traits: an example from eastern North America. – Ecology 91: 2234–2241.

Swenson, N. G. et al. 2012. The biogeography and filtering of woody plant functional diversity in North and South America. – Global Ecol. Biogeogr. 21: 798–808.

ter steege, H. et al. 2006. Continental-scale patterns of canopy tree composition and function across Amazonia. – Nature 443: 444–447.

Tilman, D. et al. 1997. The influence of functional diversity and composition on ecosystem processes. – Science 277: 1300–1302.

Webb, C. O. and Donoghue, M. J. 2005. Phylomatic: tree assembly for applied phylogenetics. – Mol. Ecol. Not. 5: 181–183.

Westoby, M. and Wright, I. J. 2006. Land-plant ecology on the basis of functional traits. – Trends Ecol. Evol. 21: 261–268.