Journal of
Plant Ecology

VOLUME 8, NUMBER 1,
PAGES 12–16

FEBRUARY 2015

doi:10.1093/jpe/rtu008

Advance Access publication
30 July 2014

available online at
www.jpe.oxfordjournals.org

A cautionary note on some phylogenetic dissimilarity measures

Carlo Ricotta[1,*], *Giovanni Bacaro*[2,3] *and Sandrine Pavoine*[4,5]

[1] *Department of Environmental Biology, University of Rome 'La Sapienza', Piazzale Aldo Moro 5, 00185 Rome, Italy*
[2] *CNR-IRPI, Istituto di Ricerca per la Protezione Idrogeologica, Via Madonna Alta 126, 06128 Perugia, Italy*
[3] *Department of Life Science, BIOCONNET, Biodiversity and Conservation Network, University of Siena, Via P. A. Mattioli 4, 53100 Siena, Italy*
[4] *Département Ecologie et Gestion de la Biodiversité, Muséum National d'Histoire Naturelle, UMR 7204 CNRS UPMC, 55–61 Rue Buffon, 75005 Paris, France*
[5] *Department of Zoology, Mathematical Ecology Research Group, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*

*Correspondence address. Department of Environmental Biology, University of Rome 'La Sapienza', Piazzale Aldo Moro 5, 00185 Rome, Italy. Tel: +390649912405; E-mail: carlo.ricotta@uniroma1.it

Abstract

Aims
Measures of plot-to-plot phylogenetic dissimilarity and beta diversity are providing a powerful tool for understanding the complex ecological and evolutionary mechanisms that drive community assembly.

Methods
Here, we review the properties of some previously published dissimilarity measures that are based on minimum or average phylogenetic dissimilarity between species in different plots.

Important Findings
We first show that some of these measures violate the basic condition that for two identical plots the measures take the value zero. They also violate the condition that the dissimilarity between two identical plots should always be lower than that between two different plots. Such erratic behavior renders these measures unsuitable for measuring plot-to-plot phylogenetic dissimilarity. We next propose a new measure that satisfies these conditions, thus providing a more reasonable way for measuring phylogenetic dissimilarity.

Keywords: abundance-weighted dissimilarity measures, index symmetry, patristic distances, presence/absence dissimilarity measures, nearest-neighbor metrics

Received: 3 March 2014, Revised: 21 May 2014, Accepted: 7 June 2014

INTRODUCTION

There are many different measures for expressing the dissimilarity between two species assemblages (or communities, samples, plots, etc.). Most of these measures attempt to summarize different aspects of plot-to-plot dissimilarity based either on species presences and absences within plots or on species abundances, thus implicitly assuming that all species are equally and maximally distinct from one another. However, it has been recently understood that more valuable measures of pairwise plot-to-plot dissimilarities should also summarize interspecies differences. As noted by Nipperess *et al.* (2010): 'given the central role of evolution in the generation of biological diversity, and the strong link between phylogeny and variation in morphological, functional and other traits, phylogeny represents the most fundamental (but not the only) basis for measuring the distinctness of organisms'. Accordingly, several indices have been recently proposed that take evolutionary relationships into account when measuring the dissimilarity of ecological assemblages.

As well as providing a starting point for community-level exploratory data analysis, dissimilarity coefficients between plots can be also interpreted as expressions of an important ecological phenomenon, such as beta diversity (see Koleff *et al.* 2003; Podani and Schmera 2011). In this framework, Webb *et al.* (2008) and Swenson *et al.* (2011) proposed two measures for summarizing plot-to-plot dissimilarity and beta diversity that are based on the average phylogenetic overlap between each species in the first plot and all species in the second plot, and *vice versa*. Unfortunately, in spite of their simplicity, these measures do not meet the foremost requirement for a dissimilarity coefficient (see Clarke *et al.* 2006): for two identical plots, the measures do not take the value zero. Also, the dissimilarity between two identical plots, A and B, is not always lower than that between two different plots A and C. Therefore, due to the widespread use of these

measures in the ecological literature, we feel compelled to bring attention to this problem and to propose a possible solution.

AN OVERVIEW ON SELECTED MEASURES OF PHYLOGENETIC DISSIMILARITY

Given two plots A and B, together with a matrix Δ of pairwise phylogenetic dissimilarities d_{ij} between species i and j (with $d_{ij} = d_{ji}$ and $d_{ii} = 0$), a simple measure of plot-to-plot phylogenetic dissimilarity based on presence and absence data can be calculated as the average minimum dissimilarity between any two species in different plots. Two such presence/absence measures were defined by Clarke and Warwick (1998) and Izsak and Price (2001) as follows:

$$D_{CW} = \frac{1}{2}\left(\frac{\sum_i^{S_A} \min d_{iB}}{S_A} + \frac{\sum_j^{S_B} \min d_{jA}}{S_B}\right) \quad (1)$$

and

$$D_{IP} = \frac{\left(\sum_i^{S_A} \min d_{iB} + \sum_j^{S_B} \min d_{jA}\right)}{S_A + S_B} \quad (2)$$

where min d_{iB} is the minimum phylogenetic dissimilarity (usually measured as the patristic distance or linking path determined from a phylogenetic tree) between species i in plot A and all species in plot B, min d_{jA} is the minimum phylogenetic dissimilarity between species j in plot B and all species in plot A and S_A and S_B denote the species richness of plots A and B, respectively (Ricotta and Bacaro 2010).

D_{IP} summarizes the mean dissimilarity between each species in one plot and its 'phylogenetic nearest neighbor' in the second plot, whereas D_{CW} separately calculates the mean phylogenetic dissimilarity between all species in plot A and their phylogenetic nearest neighbors in plot B, and *vice versa*. Then, it averages the two means. Note that D_{IP} was independently proposed by Webb *et al.* (2008) under the name *COMDISTNN*. The R script (R Development Core Team 2013) for this index is available in the package picante (Kembel *et al.* 2010). Note also that if both plots contain the same number of species, then $D_{CW} = D_{IP}$. For a thorough discussion of the relationship of D_{CW} and D_{IP} with the Bray-Curtis dissimilarity family, see Clarke *et al.* (2006).

Ricotta and Burrascano (2008) generalized Equation (1) to allow the inclusion of species' relative abundances in the calculation of phylogenetic dissimilarity:

$$D_{RB} = \frac{1}{2}\left(\sum_i^{S_A} p_i \min d_{iB} + \sum_j^{S_B} q_j \min d_{jA}\right), \quad (3)$$

where p_i is the relative abundance of species i in plot A, q_j is the relative abundance of species j in plot B and the summation $\sum_i^{S_A} p_i \min d_{iB}$ is the expected minimum phylogenetic dissimilarity between plot A and plot B if one individual is chosen randomly from plot A (Ricotta and Bacaro 2010).

Alternatively, other indices rely on mean phylogenetic dissimilarities, rather than on minimums. Rao (1982) introduced a measure of the expected (phylogenetic) dissimilarity between one individual chosen at random from plot A and one individual chosen at random from plot B as follows:

$$Q_{AB} = \sum_i^{S_A} p_i \sum_j^{S_B} q_j d_{ij} \quad (4)$$

For presence and absence data (i.e. for $p_i = 1/S_A$ and $q_j = 1/S_B$), the same index was independently proposed by Webb *et al.* (2008) for measuring plot-to-plot phylogenetic dissimilarity under the name *COMDIST*. Webb *et al.* (2008) also independently proposed an abundance-weighted version of *COMDIST*, which is mathematically identical to Q_{AB}.

Referring to Webb *et al.* (2008), the COMDIST index was reproposed by Swenson (2011) as a measure of 'presence–absence weighted pairwise phylogenetic dissimilarity' D_{pw}. Using a slightly different, less compact notation from the original formulation of Swenson (2011), this measure can be defined as follows:

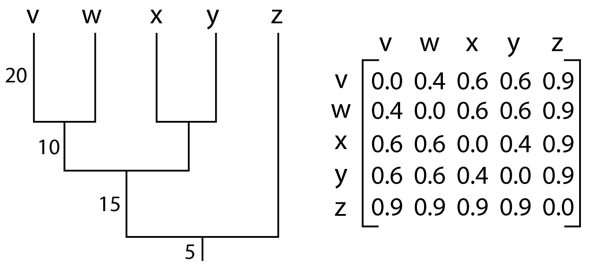
$$D_{pw} = \frac{1}{2}\left(\sum_i^{S_A} \frac{1}{S_A}\bar{d}_{iB} + \sum_j^{S_B} \frac{1}{S_B}\bar{d}_{jA}\right) \quad (5)$$

where $\bar{d}_{iB} = \sum_j 1/S_B \times d_{ij}$ is the mean phylogenetic dissimilarity between species i in plot A and all species in plot B and $\bar{d}_{jA} = \sum_i 1/S_A \times d_{ij}$ is the mean phylogenetic dissimilarity between species j in plot B and all species in plot A.

Finally, Swenson *et al.* (2011) also proposed an abundance-weighted version of the above pairwise phylogenetic dissimilarity, D_{pw}':

$$D_{pw}' = \frac{1}{2}\left(\sum_i^{S_A} p_i \bar{d}_{iB} + \sum_j^{S_B} q_j \bar{d}_{jA}\right) \quad (6)$$

Unfortunately, Q_{AB}, D_{pw} and its abundance-weighted generalization D_{pw}' do not take the value zero for two identical plots. Take the index D_{pw} and the artificial phylogenetic tree in Fig. 1 as example. Given two identical plots, A and B, both composed of species x and y, we have: $D_{pw} = \left(\frac{1}{2}\bar{d}_{xB} + \frac{1}{2}\bar{d}_{yB} + \frac{1}{2}\bar{d}_{xA} + \frac{1}{2}\bar{d}_{yA}\right)/2 = 0.2$ with $\bar{d}_{xB} = \bar{d}_{xA} = \left(d_{xy} + d_{xx}\right)/2$ and $\bar{d}_{yB} = \bar{d}_{yA} = \left(d_{yx} + d_{yy}\right)/2$.



	v	w	x	y	z
v	0.0	0.4	0.6	0.6	0.9
w	0.4	0.0	0.6	0.6	0.9
x	0.6	0.6	0.0	0.4	0.9
y	0.6	0.6	0.4	0.0	0.9
z	0.9	0.9	0.9	0.9	0.0

Figure 1: example of an artificial phylogenetic tree composed of five species, together with its species distance matrix. The patristic distance (linking path) between two species is equal to the total branch length separating both species in the phylogeny. For an ultrametric tree, this distance is twice the distance from the most recent common ancestor or node (i.e. branch length from species i to the most recent common ancestor plus branch length from the most recent common ancestor to species j). The species distances in the distance matrix are rescaled in the range [0, 1] by dividing the patristic distances by the maximum distance between two theoretical species $d_{max} = 100$ (i.e. twice the distance from the tips to the root of the tree).

If we substitute species x with species z in both plots, we obtain D_{pw} = 0.45. Hence, D_{pw} itself is not a statistically valid index of dissimilarity, as for any two identical plots, it is possible to have different index values depending on the number and the identity of species within plots. From this, it also follows that the dissimilarity among two identical plots can be higher than the dissimilarity between two different plots. For example, for two identical plots with species x, y and z, we have D_{pw} = 0.488, whereas if we compare one plot composed of the same species x, y and z with a second plot composed of species x and y, we obtain D_{pw} = 0.433. This paradox renders these measures unsuitable for measuring plot-to-plot phylogenetic dissimilarity.

A NEW MEASURE OF PHYLOGENETIC DISSIMILARITY

A number of different phylogenetic dissimilarity indices have been already proposed that correctly take the value zero for two identical plots. Examples are the phylogenetic Sørensen measure 'PhyloSor' (Bryant *et al.* 2008) and its abundance-weighted version (Nipperess *et al.* 2010), or the phylogenetic fuzzy weighting approach of Pillar and Duarte (2010; see also Duarte 2011). As an alternative, we propose here a new plot-to-plot dissimilarity coefficient that is based on index symmetry. For a given species assemblage, together with an interspecies dissimilarity matrix $\Delta = [d_{ij}]$ with d_{ij} in the range 0–1, Leinster and Cobbold (2012) defined the average ordinariness of the assemblage Z_A as the expected similarity between two individuals chosen at random with replacement from the assemblage:

$$Z_A = \sum_i^{S_A} p_i \sum_j^{S_A} p_j s_{ij} \tag{7}$$

where s_{ij} is the *similarity* between species i and j ($s_{ij} = 1 - d_{ij}$). For a thorough discussion on the relationship between Z_A and the Rao (1982) quadratic diversity, see Leinster and Cobbold (2012). Of course, for any dissimilarity measure with an upper bound $d_{\max} > 1$, division by $d_{\max}$ gives a standardized dissimilarity measure in the range [0, 1]. On the other hand, for dissimilarities that do not have a fixed upper bound, such as patristic distances, it is still possible to locally normalize all d_{ij} values in the range [0, 1] by dividing each term by the highest dissimilarity value found in the data set.

The quantity $Z_i = \sum_j^{S_A} p_j s_{ij}$ in Equation (7) is the expected similarity between an individual of species i and an individual chosen at random from the assemblage. Z_i, therefore, measures the ordinariness of species i within the assemblage. Leinster and Cobbold (2012) called Z_i the abundance of species similar to i. As $s_{ii} = 1$ ($d_{ii} = 0$) by definition, we always have $Z_i \geq p_i$ meaning that the relative abundance of all species similar to i is at least as great as the relative abundance of i itself.

Starting from index Z_A in Equation (7), we note that, for two identical plots A and B, the quantities $\sum_j^{S_{AB}} p_j s_{ij}$ in one plot should be equal to the corresponding quantities $\sum_j^{S_{AB}} q_j s_{ij}$ in the other plot, where i, j = 1, 2, … , S_{AB}, p_j is the relative abundance of species j in plot A, q_j is the relative abundance of species j in plot B, and S_{AB} is the species richness of the pooled pair of plots. In other words, the abundance of species similar to i in plot A should be equal to the abundance of species similar to i in plot B.

Accordingly, we can get an index of plot-to-plot phylogenetic dissimilarity, D_{AB}, by taking the sum of the absolute differences $\left|\sum_j^{S_{AB}} p_j s_{ij} - \sum_j^{S_{AB}} q_j s_{ij}\right|$ over all species in the pooled pair of plots, and normalizing this sum by the total species ordinariness (sum of Z_i's) in plots A and B:

$$D_{AB} = \frac{\sum_i^{S_{AB}} \left|\sum_j^{S_{AB}} p_j s_{ij} - \sum_j^{S_{AB}} q_j s_{ij}\right|}{\sum_i^{S_{AB}} \left(\sum_j^{S_{AB}} p_j s_{ij} + \sum_j^{S_{AB}} q_j s_{ij}\right)} \tag{8}$$

with D_{AB} = 0 for two identical plots and D_{AB} = 1 for maximally distinct plots with no species in common and with zero similarities among their species. For details on the definition of maximally distinct plots, see Pavoine and Ricotta (2014). Figure 2 contains a worked example to clarify how the proposed index works, and online supplementary Appendix S1 shows R script for calculating the new measure.

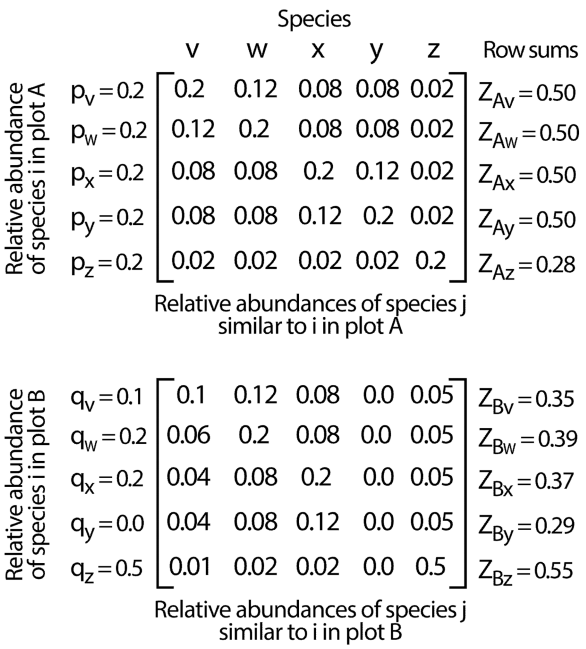


Figure 2: worked example to demonstrate how to calculate the dissimilarity index D_{AB}. For two artificial plots, A and B, and the phylogenetic tree in Fig. 1, the figure shows the species relative abundances in both plots, together with the resultant matrices with elements $p_j \times s_{ij}$ and $q_j \times s_{ij}$. The pairwise species similarities s_{ij} are calculated from the species dissimilarities d_{ij} in the distance matrix of Fig. 1 as $s_{ij} = 1 - d_{ij}$. D_{AB} is then obtained from the row sums Z_{Ai} and Z_{Bi} of both matrices as follows:

$$D_{AB} = \frac{|Z_{Av} - Z_{Bv}| + |Z_{Aw} - Z_{Bw}| + |Z_{Ax} - Z_{Bx}| + |Z_{Ay} - Z_{By}| + |Z_{Az} - Z_{Bz}|}{Z_{Av} + Z_{Aw} + Z_{Ax} + Z_{Ay} + Z_{Az} + Z_{Bv} + Z_{Bw} + Z_{Bx} + Z_{By} + Z_{Bz}} = 0.206$$

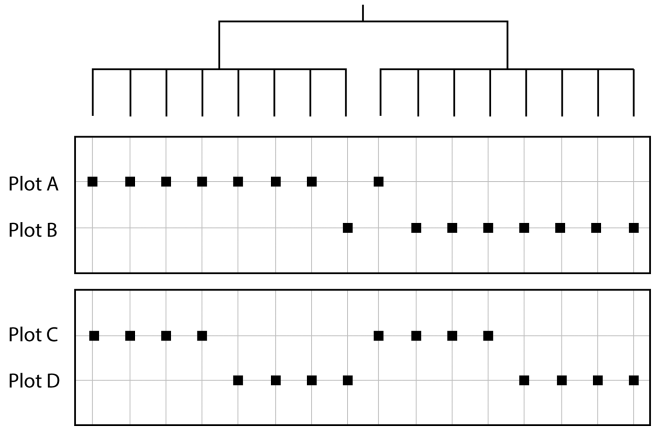


Figure 3: theoretical example showing that nearest-neighbor dissimilarity measures may lead to unexpected results. Given four plots, each composed of eight equally abundant species (filled square), together with the corresponding phylogenetic tree, the phylogenetic dissimilarity between plots A and B calculated with the measures D_{CW}, D_{IP} or D_{RB} is equal to the dissimilarity between plots C and D irrespective of the length of the branches in the phylogeny (P Legendre, personal communication to S.P.).

CONCLUSION

Recent advances in ecological theory have been made possible through the integration of phylogenetic information into studies of community ecology. As a result, in the last two decades, several measures of plot-to-plot phylogenetic dissimilarity have been independently discovered and rediscovered by different authors. In this short note, we showed that some of these measures, such as Q_{AB}, D_{pw} and D_{pw}', do not take the value zero for two identical plots. This renders them unsuitable for calculating plot-to-plot phylogenetic dissimilarity and beta diversity. This is not to say that nearest-neighbor indices, such as D_{CW}, D_{IP} or D_{RB} are perfect measures of plot-to-plot dissimilarity. For instance, in Fig. 3, an example is shown where these indices lead to unexpected results. Accordingly, nearest-neighbor indices should be used only in cases where the calculation of a minimum phylogenetic distance between species is based on strong biological reasons.

To fix these shortcomings, we suggested a new measure (D_{AB}) of phylogenetic dissimilarity between a pair of plots. The proposed measure takes the value zero for identical plots and its maximum value when two plots, *A* and *B*, have no species in common and $s_{ij} = 0$ ($d_{ij} = 1$) for one species in plot *A* and one species in plot *B*. Although we discussed the calculation of D_{AB} in the framework of patristic distances only, the proposed index calculates the dissimilarity between a pair of plots based on any dissimilarity measure of choice. Accordingly, the same index may be applied to any other ecologically relevant measure of interspecies dissimilarity, such as genetic or functional dissimilarities, and does not need to be necessarily restricted to phylogenetic differences between species.

Overall, although further research is needed to explore the properties of the proposed measure in deeper detail, we think, D_{AB} is a promising tool for summarizing plot-to-plot dissimilarity in a meaningful way. Nonetheless, we would like to stress, once again, that there is no 'magic' measure that is able of uniquely characterizing all aspects of plot-to-plot dissimilarity. Dissimilarity coefficients condense multivariate ecological data of high dimension into single univariate measures. Therefore, information is necessarily lost and a perfect dissimilarity measure does not exist. Rather, several 'tailored' measures may be developed and their specific relevance must be evaluated based on their ability to answer the particular ecological question under scrutiny.

SUPPLEMENTARY MATERIAL

Supplementary appendix is available at Journal of Plant Ecology online.

FUNDING

University of Rome 'La Sapienza' (C26A13JZ89).
Conflict of interest statement. None declared.

REFERENCES

Bryant JA, Lamanna C, Morlon H, *et al*. (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proc Natl Acad Sci USA* **105**:11505–11.

Clarke KR, Somerfield PJ, Chapman MG (2006) On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. *J Exp Mar Biol Ecol* **330**:55–80.

Clarke KR, Warwick RM (1998) Quantifying structural redundancy in ecological communities. *Oecologia* **113**:278–89.

Duarte LDS (2011) Phylogenetic habitat filtering influences forest nucleation in grasslands. *Oikos* **120**:208–15.

Izsak C, Price RG (2001) Measuring β-diversity using a taxonomic similarity index, and its relation to spatial scale. *Mar Ecol Prog Ser* **215**:69–77.

Kembel SW, Cowan PD, Helmus MR, *et al.* (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**:1463–4.

Koleff P, Gaston KJ, Lennon JJ (2003) Measuring beta diversity for presence-absence data. *J Anim Ecol* **72**:367–82.

Leinster T, Cobbold CA (2012) Measuring diversity: the importance of species similarity. *Ecology* **93**:477–89.

Nipperess DA, Faith DP, Barton K (2010) Resemblance in phylogenetic diversity among ecological samples. *J Veg Sci* **21**:809–20.

Pavoine S, Ricotta C (2014) Functional and phylogenetic similarity among communities. *Methods Ecol Evol*, 10.1111/2041-210X.12193

Pillar VD, Duarte LDS (2010) A framework for metacommunity analysis of phylogenetic structure. *Ecol Lett* **13**:587–96.

Podani J, Schmera D (2011) A new conceptual and methodological framework for exploring and explaining pattern in presence–absence data. *Oikos* **120**:1625–38.

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao CR (1982) Diversity and dissimilarity coefficients: a unified approach. *Theor Popul Biol* **21**:24–43.

Ricotta C, Bacaro G (2010) On plot-to-plot dissimilarity measures based on species functional traits. *Community Ecol* **11**:113–9.

Ricotta C, Burrascano S (2008) Beta diversity for functional ecology. *Preslia* **80**:61–71.

Swenson NG (2011) Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PLoS One* **6**:e21264.

Swenson NG, Anglada-Cordero P, Barone JA (2011) Deterministic tropical tree community turnover: evidence from patterns of functional beta diversity along an elevational gradient. *Proc Biol Sci* **278**:877–84.

Webb CO, Ackerly DD, Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics* **24**:2098–100.