Methods in Ecology and Evolution

Methods in Ecology and Evolution 2014

Functional and phylogenetic similarity among communities

Sandrine Pavoine^{1,2*} and Carlo Ricotta³

¹Department of Ecology and Biodiversity Management, UMR 7204 CNRS UPMC, Muséum National d'Histoire Naturelle, 75005 Paris, France; ²Mathematical Ecology Research Group, Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK; and ³Department of Environmental Biology, University of Rome 'La Sapienza', Piazzale Aldo Moro 5, 00185 Rome, Italy

Summary

1. Ecological studies often rely on coefficients of intercommunity (dis)similarity to decipher effects of ecological, evolutionary and human-driven mechanisms on the composition of communities. Yet, two main criticisms have been levelled at (dis)similarity coefficients. First, few developments include information on species' abundances, and either phylogeny or functional traits. Secondly, some (dis)similarity coefficients fail to always provide maximum dissimilarity between two completely distinct communities, that is, communities without common species and with zero similarities among their species.

2. Here, we introduce a new family of similarity coefficients responding to these criticisms. Within this family, we concentrate on four coefficients and compare them to Rao's dissimilarity on macroinvertebrate communities, and simulated data.

3. Our new coefficients correctly treat maximally dissimilar communities: similarities are always zero between two completely distinct communities. The originality of these new coefficients is even more profound as the existence of maximally dissimilar communities was not a requirement for the new coefficients to behave differently than Rao's dissimilarity coefficient.

4. Our new family of similarity coefficients relies on the abundances or occurrences of species within communities and on phylogenetic, taxonomic or functional similarities among species. We demonstrate that this new family embeds many of the recent developments in both functional and phylogenetic diversity. It provides a unique framework for comparing traditional compositional turnover with functional or phylogenetic similarities among communities.

Key-words: beta diversity, biodiversity, choice of coefficient, community ecology, community phylogenetics, compositional turnover, principle of maximum dissimilarity, quadratic entropy

Introduction

There are many different coefficients for expressing the (dis)similarity between two communities (or plots, stations, samples, assemblages, etc.). The large majority of these measures attempts to summarize different aspects of community-to-community dissimilarity based either on species presences and absences within communities or on species abundances. However, the utility of dissimilarity measures that incorporate information about the degree of ecological differences between the species in both communities is becoming increasingly recognized (Pavoine, Dufour & Chessel 2004; Lozupone & Knight 2005; Ferrier *et al.* 2007; Bryant *et al.* 2008; Graham & Fine 2008; Webb, Ackerly & Kembel 2008; Ricotta & Szeidl 2009; Ives

& Helmus 2010; Nipperess, Faith & Barton 2010; Chiu, Jost & Chao 2014). Such interspecies differences can be based either on phylogenetic or on functional relationships among species, as ecological differences between species are believed to be reflected in both of them (Webb *et al.* 2002).

To summarize mean interspecies differences within single communities, Rao (1982) proposed a diversity index, termed quadratic diversity (Q), that is defined as the expected dissimilarity between two individuals of a given community randomly drawn with replacement:

$$Q = \sum_{ij} p_i p_j \delta_{ij} \qquad \text{eqn 1}$$

where p_i is the relative abundance of species i (i = 1, 2, ..., N) with $p_i \ge 0$ and $\sum_i p_i = 1$, and $\Delta = (\delta_{ij})$, where i, j = 1, 2, ..., N, is a symmetric matrix of pairwise (functional or phylogenetic) dissimilarities among all species i and j. Given two

© 2014 The Authors. Methods in Ecology and Evolution © 2014 British Ecological Society

^{*}Correspondence author. E-mail: pavoine@mnhn.fr

communities with relative abundance vectors $\mathbf{p} = (p_1 \dots p_i \dots p_N)^t$ and $\mathbf{q} = (q_1 \dots q_i \dots q_N)^t$, where *t* is the transpose, Rao (1982) defined a dissimilarity coefficient:

$$D_Q = \sum_{ij} p_i q_j \delta_{ij} - \frac{1}{2} \sum_{ij} p_i p_j \delta_{ij} - \frac{1}{2} \sum_{ij} q_i q_j \delta_{ij} \qquad \text{eqn } 2$$

where $\sum_{ij} p_i q_j \delta_{ij}$ is the expected dissimilarity between two randomly drawn individuals, one from each community. D_Q is thus obtained by subtracting the mean of within-community diversities from among-community diversity. If $\mathbf{\Delta} = (\delta_{ij})$ is squared Euclidean (a mathematical property described in the Glossary in Table 1) and if $0 \le \delta_{ij} \le 1$, for all *i* and *j*, D_Q is bounded between 0 and 1 due to the concavity of function $Q(\mathbf{p}) = \sum_{ij} p_i p_j \delta_{ij}$ (i.e. $Q(\frac{\mathbf{p}+\mathbf{q}}{2}) \ge \frac{1}{2}Q(\mathbf{p}) + \frac{1}{2}Q(\mathbf{q})$, with $Q(\frac{\mathbf{p}+\mathbf{q}}{2})$ being Rao's quadratic entropy index computed from vector $(\mathbf{p} + \mathbf{q})/2$, Champely & Chessel 2002). Concavity here means that diversity increases by mixing so that the diversity in a pool of communities is always higher than (or equal to) the average diversity within the communities.

Referring to Jost's (2006) observations on more traditional dissimilarity indices, Ricotta and Szeidl (2009) observed that Rao's dissimilarity coefficient fails to always provide maximum dissimilarity between two completely distinct communities. If two communities are completely distinct, that is, if they have no species in common and $\delta_{ij} = 1$ for species *i* belonging to the first community and species *j* to the second one, we expect the average dissimilarity between the species of the first and the species of the second community, that is, $\sum_{ij} p_i q_j \delta_{ij}$, to be equal to unity. Yet, in that case, D_Q can be low if $\sum_{ij} p_i p_j \delta_{ij}$ or $\sum_{ij} q_i q_j \delta_{ij}$, which measure the diversity within each community, is high. The main objective of this study was thus to introduce new (dis)similarity coefficients that provide maximum dissimilarity (and thus zero similarity) between two completely distinct communities.

Methods

A NEW FAMILY OF SIMILARITY INDICES

A way of providing maximum dissimilarity between two completely distinct communities as recommended by Jost (2006) and Ricotta and Szeidl (2009) is to standardize the dissimilarity coefficient D_Q by

Table 1. Glossary

dividing it by the value expected for two completely distinct theoretical communities with the same quadratic diversity as the real communities (see Meirmans 2006 for the use of a related standardization process in genetics). The standardized coefficient would thus be:

$$D_{st} = \frac{\sum_{ij} p_i q_j \delta_{ij} - \frac{1}{2} \sum_{ij} p_i p_j \delta_{ij} - \frac{1}{2} \sum_{ij} q_i q_j \delta_{ij}}{1 - \frac{1}{2} \sum_{ij} p_i p_j \delta_{ij} - \frac{1}{2} \sum_{ij} q_i q_j \delta_{ij}}$$
eqn 3

Because D_{st} is bounded between 0 and 1, an associated similarity coefficient can be defined as $S_{st} = 1 - D_{st}$. Using interspecies similarities instead of dissimilarities, the expression of S_{st} simplifies to:

$$S_{st} = \frac{\sum_{ij} p_i q_j \sigma_{ij}}{\frac{1}{2} \sum_{ij} p_i p_j \sigma_{ij} + \frac{1}{2} \sum_{ij} q_i q_j \sigma_{ij}} \qquad \text{eqn 4}$$

where $\sigma_{ij} = 1 - \delta_{ij}$ for all *i*, *j* is a measure of pairwise (functional or phylogenetic) similarity among species (Appendix S1, section 1.1). This paper will show that S_{st} is a special case of a more general formula, which contains well-known indices such as the Sørensen (1948) and Horn (1966) coefficients, together with more special cases.

The standardization of D_Q opens the way for constructing a new family of similarity coefficients. First, dealing with interspecies similarities σ_{ij} , the relative abundance vectors **p** and **q** can be replaced by more general vectors **x** = $(x_1 \dots x_i \dots x_N)'$ and **z** = $(z_1 \dots z_i \dots z_N)'$ where any nonnegative value is allowed (i.e. the quantities x_i and z_i are not necessarily required to sum to one). Vectors **x** and **z** can contain either presence/absence (1/0) scores, absolute abundance values, such as individual counts or biomass data, or relative abundance data that sum to one over all species in a given community. Next, several similarity indices (Table 2) can be developed by considering different combinations of the between-community component $\sum_{ij} x_i z_j \sigma_{ij}$ and the withincommunity components $\sum_{ij} x_i x_j \sigma_{ij}$ and the withincommunity components $\sum_{ij} x_i x_j \sigma_{ij}$ and the withincommunity components (dis)similarity (Podani 2000; Legendre & Legendre 2012), the index

$$S_{Sokal-Sneath} = \frac{\sum_{ij} x_i z_j \sigma_{ij}}{2 \sum_{ij} x_i x_j \sigma_{ij} + 2 \sum_{ij} z_i z_j \sigma_{ij} - 3 \sum_{ij} x_i z_j \sigma_{ij}}$$
(5)

is a generalization of an index proposed by Sokal and Sneath (1963) for the presence/absence scores. Indeed, using species presence/absence data and setting $\sigma_{ij} = 0$ for $i \neq j$ and $\sigma_{ii} = 1$, then $\sum_{ij} x_i z_i \sigma_{ij} = a$ is the number of species shared by the two communities, $\sum_{ij} x_i x_j \sigma_{ij} = a + b$ is the total number of species in the first community (with *b* the number of species in the first community that are absent from the second community), and $\sum_{ij} z_i z_j \sigma_{ij} = a + c$ is the total number of species in the second community (with *c* the number of species in the second com-

	Definition						
Dissimilarity	Here dissimilarity between two entities <i>i</i> and <i>j</i> (e.g. two species or two communities) denotes any nonnegative value d_{ij} that measures any functional or phylogenetic difference between the two entities, with $d_{ii} = 0$. In this paper, the discussion is limited to symmetric dissimilarities bounded between 0 and 1						
Similarity	In this paper, the discussion is limited to symmetric similarities bounded between 0 and 1 so that $s_{ii} = 1 - d_{ii}$ for all i and j						
Distance matrix	A matrix of dissimilarities $\mathbf{D} = (d_{ij})$, for all $i, j = 1,, N$ is a distance matrix if $d_{ij} \leq d_{ik} + d_{ki}$ for all $i, j, k = 1,, N$						
Euclidean matrix	A matrix of dissimilarities $\mathbf{D} = (d_{ij})$, for all $i, j = 1,, N$, is Euclidean if one can find N points $M_1,, M_N$ in a Euclidean space, so that the Euclidean distance between any two points M_i , M_i is d_{ij} . Euclidean matrices are distance matrices						
Squared Euclidean matrix	A matrix of dissimilarities $\mathbf{D} = (d_{ij})$, for all $i, j = 1,, N$, is squared Euclidean if the matrix $(\sqrt{d_{ij}})$, for all $i, j = 1,, N$, is Euclidean						
Positive semi-definite matrix	Let A be a square matrix (a_{ij}) , for all $i, j = 1,, N$. A is positive semi-definite (=non-negative definite) if, for any real vector $\mathbf{x} = (x_1x_N)^t, \sum_{i=1}^N \sum_{j=1}^N x_i x_j a_{ij} \ge 0$						

		Special cases: maximally dis	tinct species [†]		
	General formula (x , z positive)*	x, z positive*	$\operatorname{Reference}^{\ddagger}$	Presence/absence only [§]	$ m Reference^{3}$
Sokal-Sneath	$\frac{\sum_{ij} x_i z_j \sigma_{ij}}{2 \sum_{ij} x_i z_j \alpha_{ij} + 2 \sum_{ij} z_i z_j \sigma_{ij} - 3 \sum_{ij} x_i z_j \sigma_{ij}}$	$\frac{2\sum_{i} x_{i}^{2} + 2\sum_{i} z_{i}^{2} - 3\sum_{i} x_{i} z_{i}}{z_{i}^{2} - 3\sum_{i} x_{i} z_{i}}$		$\frac{a}{a+2b+2c}$	Sokal & Sneath (1963)
SJaccard	$\frac{\sum_{ij} x_i z_j \sigma_{ij}}{\sum_{ij} x_i z_j \sigma_{ij} + \sum_{ij} z_i z_j \sigma_{ij} - \sum_{ij} x_i z_j \sigma_{ij}}$	$\sum_i x_i^2 + \sum_i z_i^2 - \sum_i x_i z_i$	Wishart (1969)	$\frac{a}{a+b+c}$	Jaccard (1901)
Ssørensen	$\frac{\sum_{ij} x_{i,j} \sigma_{ij}}{2} \sum_{ij} x_{i,ij} \sigma_{ij} + \frac{1}{2} \sum_{ij} z_{ij} \sigma_{ij}}$	$\frac{\sum_i x_i z_i}{2\sum_i x_i^2 + \frac{1}{2}\sum_i z_i^2}$	Morisita (1959) and Horn (1966)	$\frac{2a}{2a+b+c}$	Dice (1945) and Sørensen (1948
Sochiai	$\frac{\sum_{ij} x_{l} z_{j} \sigma_{ij}}{\sqrt{\sum_{ij} x_{l} x_{l}} \sigma_{lj} \sqrt{\sum_{ij} z_{l} z_{j} \sigma_{ij}}}$	$\frac{\sum_{i} x_{i} z_{i}}{\sqrt{\sum_{i} x_{i}^{2}} \sqrt{\sum_{i} z_{i}^{2}}}$		$\sqrt{a+b}\sqrt{a+c}$	Ochiai (1957)
$\mathbf{x} = (x_1, \dots, x_i, \dots$	$(z_N)^t$ and $\mathbf{z} = (z_1 \dots z_i, \dots z_N)^t$ contain positive	e values attributed to N species wi	thin two communities: for example presence,	(absence (1/0), absolute values (e.g.	abundance, biomass) or pr

munity that are absent from the first community). In this special case, the coefficient $S_{Sokal-Sneath}$ reduces to

$$\frac{a}{a+2b+2c} \qquad \qquad \text{eqn 6}$$

an index introduced by Sokal and Sneath (1963) as a measure of species turnover.

Likewise, the index

$$S_{Jaccard} = \frac{\sum_{ij} x_i z_j \sigma_{ij}}{\sum_{ij} x_i x_j \sigma_{ij} + \sum_{ij} z_i z_j \sigma_{ij} - \sum_{ij} x_i z_j \sigma_{ij}} \qquad \text{eqn 7}$$

turns out to be a generalization of the index developed by Jaccard (1901) for the presence/absence data and that of Wishart (1969) for species abundances (Table 2), while the index

$$S_{Sorensen} = \frac{\sum_{ij} x_i z_j \sigma_{ij}}{\frac{1}{2} \sum_{ij} x_i x_j \sigma_{ij} + \frac{1}{2} \sum_{ij} z_i z_j \sigma_{ij}}$$
eqn 8

is a generalization of the index S_{st} introduced above. $S_{Sorensen}$ is also a generalization of the Dice–Sørensen index for the presence/absence data and Morisita–Horn index for species abundances (Dice 1945; Sørensen 1948; Morisita 1959; Horn 1966; Table 2). When applied to ultrametric phylogenetic similarities among species, this index is related to the phylo-Morisita–Horn index of Chiu, Jost and Chao (2014; see Appendix S1, section 1.2). The index $S_{Sorensen}$ provides, however, more flexibility in the types of similarities among species that can be used. Finally, if, in the denominator of $S_{Sorensen}$, the arithmetic mean $\frac{1}{2}S_A + \frac{1}{2}S_B$ is replaced by the geometric mean $\sqrt{S_A}\sqrt{S_B}$, we obtain

$$S_{Ochiai} = \frac{\sum_{ij} x_i z_j \sigma_{ij}}{\sqrt{\sum_{ij} x_i x_j \sigma_{ij}} \sqrt{\sum_{ij} z_i z_j \sigma_{ij}}}$$
eqn 9

which is an extension of the well-known index developed by Ochiai (1957) to incorporate interspecies measures of functional or phylogenetic similarity (see Table 2). *S_{Ochial}* is also related to the chord distance introduced in ecological studies by Orlóci (1967; see also Burt 1948 and Tucker 1951).

All these indices are bounded between 0 and 1 (section 1.3 in Appendix S1) and lead to positive semi-definite (p.s.d.) matrices $\mathbf{S} = (s_{ij})$ of intercommunity similarities if the interspecies similarity matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ is also p.s.d. (see Glossary and Appendix S1 section 1.4). This property implies that the dissimilarity matrices $\mathbf{D} = (\sqrt{1 - s_{ij}})$ and $\Delta = (\sqrt{1 - \sigma_{ij}})$ are Euclidean so that they can be associated with clouds of points in Euclidean space (Gower & Legendre 1986; see Glossary).

The following inequalities exist among the new indices: $0 \le S_{So-kal-Sneath} \le SJaccard \le S_{Sorensen} \le S_{Ochiai} \le 1$ (section 1.5 in Appendix S1). Some properties of the new indices depend on their extrema. When two communities are completely distinct with no species in common and $\sigma_{ij} = 0$ for species *i* occurring in the first community and species *j* in the second one, all indices in Table 2 equal zero. A difference among the new similarity coefficients is that coefficients $S_{Sokal-Sneath}$, $S_{Jaccard}$ and $S_{Sorensen}$ all equal 1 (perfect similarity) when $\sum_{ij} x_i z_j \sigma_{ij}$ equals $\frac{1}{2} \sum_{ij} x_i x_j \sigma_{ij} + \frac{1}{2} \sum_{ij} z_i z_j \sigma_{ij}$ (arithmetic mean), whereas S_{Ochiai} equals 1 when $\sum_{ij} x_i z_j \sigma_{ij}$ equals $\sqrt{\sum_{ij} x_i x_j \sigma_{ij}} \sqrt{\sum_{ij} z_i z_j \sigma_{ij}}$ (geometric mean). In both cases, perfect similarity includes the situation where $\mathbf{x} = \mathbf{z}$ but not exclusively. Perfect similarity among species within communities than between communities.

See Gleason (1920) in addition to Dice (1945) and Sørensen (1948). The formula $\sum_i x_i z_i / \sqrt{\sum_i x_i^2 \sum_i z_i^2}$ is related to the chord distance introduced in ecology by Orlóci (1967).

Here $\sigma_{ij} = 0$ for all $i \neq j$ and $\sigma_{ii} = 1$. Species are thus all maximally different.

CONSTRAINTS FOR DEVELOPING THE NEW FAMILY OF SIMILARITY INDICES

When combining components $\sum_{ij} x_i z_j \sigma_{ij}$, $\sum_{ij} x_i x_j \sigma_{ij}$ and $\sum_{ij} z_i z_j \sigma_{ij}$, the first criterion is of course that the combination leads to a similarity index. Our second criterion was to restrict our family to relative similarity indices bounded between 0 and 1. Without this restriction, the component $\sum_{ij} x_i z_j \sigma_{ij}$ itself could be used as an index of similarity among communities. Indeed, in the same line, Webb, Ackerly and Kembel (2008) suggested the following formula, designated as COM-DIST in the software Phylocom, for measuring the dissimilarity between two communities: $\sum_{ij} p_i q_j \delta_{ij}$, with p_i the relative abundance of species *i* in community A, q_i the relative abundance of species *j* in community B and δ_{ii} the phylogenetic dissimilarity between *i* and *j*. Like $\sum_{ij} x_i z_j \sigma_{ij}$, COMDIST is an absolute index that does not consider how diverse communities are. In the next sections, we illustrate with a simple theoretical data set, the consequences the use of COMDIST can have when measuring the dissimilarity between two communities. On the other hand, index standardization between 0 and 1, like for all indices of the new family, is not without consequences. For instance, one has first to fix a value (generally unity) for maximum similarity among species, such that interspecies similarities are bounded between 0 and this maximum. If the dissimilarities among species are multiplied by 2, the resulting values of COMDIST and D_O are also multiplied by 2. Such multiplicity does not hold for similarity indices bounded between 0 and 1, like Ssokal-Sneath, Ssocard, Ssorensen and Sochiai. Also, if the similarity values among distinct species are divided by a constant (leaving the similarity between individuals of the same species equal to unity, i.e. $\sigma_{ii} = 1$), then the values of $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and Sochiai will depend on how this division modifies the index components $\sum_{ij} x_i z_j \sigma_{ij}, \sum_{ij} x_i x_j \sigma_{ij}$, and $\sum_{ij} z_i z_j \sigma_{ij}$, and how these components are combined in the formulation of the different indices. Altogether, the results obtained with any index applied to different data sets will be comparable only if the (dis)similarities among species have been defined in the same way (see Appendix S2 for a short discussion on the definition of taxonomic, phylogenetic and functional similarities among species).

Given the high number of possibilities to combine the three index components $\sum_{ij} x_i z_j \sigma_{ij}$, $\sum_{ij} x_i x_j \sigma_{ij}$, and $\sum_{ij} z_i z_j \sigma_{ij}$ into an index of similarity, we further restricted our discussion to indices that lead to p.s.d. similarity matrices. Although this property might not be critical for many ecological studies, it is an interesting property when (dis)similarities among communities have to be visualized graphically. Indeed, as mentioned above, when the matrix $\mathbf{S} = (s_{ij})$ of pairwise similarities among communities is p.s.d., then the associated dissimilarity matrix $\mathbf{D} = (\sqrt{1 - s_{ij}})$ is Euclidean. Being Euclidean, the dissimilarities among communities can be associated with clouds of points using principal coordinate analysis (Gower & Legendre 1986). With this methodology, the cloud of points can be projected in a finite number of dimensions that best express how (functionally or phylogenetically) different communities are. The observed patterns can then be interpreted in terms of environmental gradients, geographical distributions, etc.

Note that the combination of the components $\sum_{ij} x_i z_j \sigma_{ij}$, $\sum_{ij} x_i x_j \sigma_{ij}$ and $\sum_{ij} z_i z_j \sigma_{ij}$ into a similarity index does not necessarily lead to p.s.d. matrices. For example, Ricotta and Szeidl (2009) defined the dissimilarity among a collection of communities as $\hat{\beta} = \hat{\gamma}/\hat{\alpha}$. When applied to a pair of communities with their respective vectors of species' proportions **p** and **q**, then

$$\hat{\gamma} = 1/\left[1 - Q\left(\frac{\mathbf{p} + \mathbf{q}}{2}\right)\right]$$
 eqn 10

$$\hat{\alpha} = 1 / \left[1 - \frac{1}{2} \mathcal{Q}(\mathbf{p}) - \frac{1}{2} \mathcal{Q}(\mathbf{q}) \right]$$
eqn 11

If $\Lambda = (\delta_{ij})$ is squared Euclidean, the coefficient $\hat{\gamma}/\hat{\alpha}$ is not bounded between 0 and 1 but between 1 and 2. This is because the objective of Ricotta and Szeidl (2009) was to obtain an effective number of communities: if community A is identical to community B, we actually have only one community, such that $\hat{\gamma}/\hat{\alpha} = 1$. On the other hand, if both communities are completely distinct $\hat{\gamma}/\hat{\alpha} = 2$. Therefore, a simple solution to obtain a dissimilarity coefficient bounded between 0 and 1, within the Ricotta and Szeidl framework, is to define $D_{\beta} = \hat{\gamma}/\hat{\alpha} - 1$, which can be written as:

$$D_{\beta} = \frac{\frac{1}{2} \sum_{ij} p_i q_j \delta_{ij} - \frac{1}{4} \sum_{ij} p_i p_j \delta_{ij} - \frac{1}{4} \sum_{ij} q_i q_j \delta_{ij}}{1 - \frac{1}{2} \sum_{ij} p_i q_j \delta_{ij} - \frac{1}{4} \sum_{ij} p_i p_j \delta_{ij} - \frac{1}{4} \sum_{ij} q_i q_j \delta_{ij}}$$
eqn 12

Because D_{β} is bounded between 0 and 1, an associated similarity coefficient can be defined as $S_{\beta} = 1 - D_{\beta}$. Using interspecies similarities instead of dissimilarities, the expression of S_{β} simplifies to:

$$S_{\beta} = \frac{4 \sum_{ij} p_i q_j \sigma_{ij}}{2 \sum_{ij} p_i q_j \sigma_{ij} + \sum_{ij} p_i p_j \sigma_{ij} + \sum_{ij} q_i q_j \sigma_{ij}} \qquad \text{eqn 13}$$

where $\sigma_{ij} = 1 - \delta_{ij}$ for all *i*, *j* are pairwise (functional or phylogenetic) similarities among species. Formula S_{β} thus combines $\sum_{ii} p_i q_j \sigma_{ij}$, $\sum_{ij} p_i p_j \sigma_{ij}$ and $\sum_{ij} q_i q_j \sigma_{ij}$ into an index of similarity bounded between 0 and 1. However, this index does not lead to intercommunity similarity matrices that are p.s.d. (a counter-example is given in the demonstration of the use of the R scripts in Appendix S3). Here, note that, while the similarity counterpart of the index D_{β} , $S_{\beta} = 1 - D_{\beta}$ is generally not p.s.d., nonetheless we have $S_{\beta} = \theta_i/(\theta_i - 1 + 1/S_i)$, with $\theta_1 = 8$ and $S_1 = S_{Sokal-Sneath}$, $\theta_2 = 4$ and $S_2 = S_{Jaccard}$, $\theta_3 = 2$ and $S_3 = S_{Sorensen}$. Note also that S_{β} is equivalent to the phylo-regional-overlap index of Chiu, Jost and Chao (2014) applied to two communities only and extended to any type of (taxonomic, phylogenetic or functional) similarities among species (section 1.6 in Appendix S1). Preliminary analysis suggests that the behaviour of S_{β} is similar to $S_{Sokal-Sneath}$, SJaccard, SSørensen and SOchiai, although it tends to provide higher similarity values (as illustrated in the examples for the use of the R scripts in Appendix S3). We have thus not included S_{β} in our case studies.

CASE STUDIES

We first compared indices $S_Q = 1 - D_Q$, $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} with COMDIST and the associated similarity index, 1-COMDIST, on a small theoretical data set as described in Fig. 1. After having illustrated the strong difference between COM-DIST and the behaviour of the other indices, we then concentrated on the latter ones. Note that as S_Q and COMDIST depend on species' proportions, vectors x and z in the equations of $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} were expressed as proportions in all our case studies.

To compare indices S_Q , $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} , we considered the data set analysed in Ivol *et al.* (1997) and Pavoine and Dolédec (2005) where a total of 40 macroinvertebrate species (here Coleoptera and Trichoptera) were sampled in 38 stations distributed in the Loire River (France) from the spring to 200 km upstream of the mouth. Stations have been sampled in July 1989 and 1991, and in March and May 1993, in rubble riffle habitats with a hand-net for about 10 min per station. Individuals were identified at the species level and counted. The objective here was not to re-analyse an old data set but rather to show where the indices of similarity proposed in Table 2



Fig. 1. Behavior of COMDIST for a small data set composed of four communities (Com1–Com4). Top left: theoretical phylogenetic tree with equal branch lengths and species as tips. The height of the tree (sum of branch lengths on the smallest path between tips and root) was considered to be unity. An open circle indicates the root node of the phylogeny. Top right: compositions of the four communities. Close squares represent species presences. Bottom: Index values of COM-DIST and six similarity measures among communities (all measures are calculated with species proportions equal to $1/N_i$, where N_i is the number of species in the *i*th community). As COMDIST is bounded between 0 and 1 (because interspecific dissimilarities among species, themselves varied between 0 and 1), we also included 1-COMDIST among the similarity measures.

behave similarly and where they do not. We performed two distinct analyses.

In the first analysis, the index $S_Q = 1 - D_Q$ was compared to $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} using the real data set. First, similarities among stations were evaluated by considering only the relative number of individuals from each species collected in each station. Then, similarities among species were considered using taxonomy and feeding habits. The taxonomic tree was considered with unit branch length and the root placed at the class level (Insecta) assuming that species of the order Coleoptera have zero taxonomic similarity with Trichoptera species. The index used to calculate the taxonomic similarities among stations. For example, for calculating the index $S_{Sokal-Sneath}$, we used the following index of interspecific similarity: for any two species *i* and *j*, $\sigma_{ij} = a/(a + 2b + 2c)$ where a = sum of

branch lengths from the nearest common ancestor of the two species *i* and *j* to the root of the tree, b = sum of branch lengths from species *i* to this nearest common ancestor and c = sum of branch lengths from species *j* to this nearest common ancestor. Using the same notation, the indices of interspecific similarity a/(a + b + c), 2a/(2a + b + c) and $a/\sqrt{(a + b)(a + c)}$ were used for calculating $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} , respectively. In general, we suggest the use of consistent indices for measuring the similarity among species and among communities since this is allowed by our methodology (Appendix S2). The indices $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} were then compared with S_Q . For comparing S_Q with, say, $S_{Sokal-Sneath}$, we calculated S_Q using the same measure of interspecies similarity used for calculating $S_{Sokal-Sneath}$. The same procedure was used for all pairwise comparisons between S_Q and one of the newly proposed measures.

Regarding feeding habits, the affinity of each species to each feeding category (engulfers, shredders, scrapers, deposit-feeders, active filter-feeders, passive filter-feeders and piercers) was quantified using a fuzzy coding approach (Chevenet, Dolédec & Chessel 1994). The species affinity for each feeding category was estimated by expert opinion on an ordinal scale ranging from 0 (no affinity) to 3 (high affinity; Ivol *et al.* 1997). The similarity between two species was then calculated using Table 2, second column, by replacing the vectors of species' proportions in each community with vectors showing the species relative affinities for each feeding category. For example, the index $S_{Sokal-Sneath}$ was calculated using the following index of interspecific similarity: for any two species *i* and *j*,

$$\sigma_{ij} = \sum_{k} a_{ik} a_{jk} / (2 \sum_{k} a_{ik}^{2} + 2 \sum_{k} a_{jk}^{2} - 3 \sum_{k} a_{ik} a_{jk}) \qquad \text{eqn 14}$$

where a_{ik} and a_{jk} are the relative affinities of species *i* and *j* to the *k*th feeding category. Using the same notation, the indices of interspecies similarity $\sigma_{ij} = \sum_k a_{ik}a_{jk}/(\sum_k a_{ik}^2 + \sum_k a_{jk}^2 - \sum_k a_{ik}a_{jk}), \quad \sigma_{ij} = 2\sum_k a_{ik}a_{jk}/(\sum_k a_{ik}^2 + \sum_k a_{jk}^2), \text{ and } \sigma_{ij} = \sum_k a_{ik}a_{jk}/\sqrt{\sum_k a_{ik}^2 \sum_k a_{jk}^2}$ were used for calculating $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} , respectively. These indices were then compared with S_Q . For the calculation of S_Q , we used the same measure of interspecies similarity used for calculating the corresponding similarity measures $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} .

In the second analysis, the behaviour of $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} was analysed in more details using the theoretical data in Fig. 2. We also calculated four different versions of S_Q , each with a different measure of interspecies similarity (for details see Table 3).

We first used the same taxonomy as in the real data set. Next, we distorted it by attributing different branch lengths to the taxonomic levels. The theoretical communities associated with the taxonomy in Fig. 2 included only the presence/absence of species. Two of them contained only Coleoptera (communities 1 and 2), while the other two communities contained Trichoptera only (communities 3 and 4). Each community contained a species per family and, when possible (i.e. when more than one species was represented per family), different species were attributed to different communities. Taxonomic similarities among species and among communities were calculated with the same procedure as for the real data set. Only communities 3 and 4 have some species in common. We thus expected their taxonomic similarity to be the highest. Because communities 1 and 2 contain Coleoptera only and communities 3 and 4 Trichoptera only, we expected the similarity of community 1 (or 2) with community 3 (or 4) to be the lowest. The data set was designed so that the similarities between communities with Coleoptera only and communities with Trichoptera only (i.e. communities 1×3 , 1×4 , 2×3 , and 2×4) were all equal.



Fig. 2. Theoretical data set. (a) The taxonomy of the real data set with equal branch lengths. The first split in the taxonomy is at the order-level (top: Coleoptera, bottom: Trichoptera). Subsequent splits represent families, genera and species (tips). (b–d) The taxonomy was then deformed by adjusting the branch lengths so that the similarities among species are increased (b) and decreased (c). In (d) equal branch lengths are considered, together with an additional taxonomic level (*e.g.* phylum = Arthropoda) common to all species. This added taxonomic level increases the similarities among all species. Open circles indicate the root node of the taxonomies. The taxonomic height of all trees (sum of branch lengths on the shortest path between tips and root) was considered equal to unity. These taxonomies were associated with four theoretical communities defined in (e). Close squares indicate species presences within each community.

Results

Using COMDIST, the same value of dissimilarity was found between communities 1 and 2, between communities 1 and 3 and between communities 1 and 4 in Fig. 1. In contrast, the indices $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$, S_{Ochiai} and S_Q correctly identify that communities 1 and 2 are identical and that community 1 shows decreasing similarity with communities 2, 3 and 4 in that order (Fig. 1).

The real data set showed discrepancies between S_Q and the other indices when species were considered to be completely distinct from each other (i.e. $\sigma_{ij} = 0$ for all $i \neq j$; Fig. 3a). These differences decreased when taxonomic and functional similarities among species were added (Fig. 3b,c). Taken together, despite some fundamental differences between S_Q and the new indices, like the shape of the relationships and the spread of points in Fig. 3, in general, S_Q had high Spearman correlation with the other indices. S_Q and the new indices would thus tend to rank, at least with this data set, the similarities among communities in consistent ways. Situations where S_Q was positive whereas the other indices equal zero were observed only between a few sites when taxonomic similarities among species were used (Fig. 3b).

As expected, with the theoretical data set in Fig. 2, communities 3 and 4 had the highest taxonomic similarity (Table 3). Also, for all scenarios of taxonomic branch lengths, all indices $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} , attributed the lowest values of taxonomic similarity between communities with Coleoptera only and communities with Trichoptera only (named C × T in Table 3). However, with the index S_Q , we obtained higher similarity values between two communities composed of species from different orders than between two communities composed of Coleoptera only. This happened when S_Q was calculated using the index of interspecies similarity associated with $S_{Sokal-Sneath}$. It also happened, whatever the index of interspecies similarity, when the similarities among species were low. Overall, the analysis revealed that the impact of the index chosen for summarizing pairwise community similarity can be drastic. For example, using the phylogeny with equal branch lengths (Fig. 2a), the similarity between communities 1 and 2 is equal to 0.080 for the index $S_{Sokal-Sneath}$ and 0.600 for S_{Ochiai} .

Discussion

DIFFERENCES BETWEEN 1-COMDIST, S_Q , AND THE NEW INDICES

Our results first highlighted a difference in behaviour between 1-COMDIST and the other indices. A criticism that can be raised towards COMDIST (and thus 1-COMDIST, Webb, Ackerly & Kembel 2008) is that it would provide equal levels of dissimilarity between two identical communities as between two communities with distinct species. This unexpected behav-

Table 3. Taxonomic similarities among the theoretical communities in Fig. 2

	Similarity indices among communities (S) and among species (σ)									
	$S_{\it Sokal-Sneath}$	S_Q calculated with $\sigma_{Sokal-Sneath}$	S _{Jaccard}	S_Q calculated with $\sigma_{Jaccard}$	S _{Sørensen}	S_{Ochiai}	S_Q calculated with $\sigma_{Sorensen} (= \sigma_{Ochiai})$			
Uniform model										
$Com 1 \times 2$	0.080	0.600	0.263	0.667	0.600	0.600	0.750			
$\text{Com 3} \times 4$	0.403	0.954	0.719	0.963	0.915	0.915	0.973			
$\text{Com}C\timesT$	0	0.646	0	0.600	0	0	0.525			
High interspecific	similarity model									
$Com 1 \times 2$	0.914	0.981	0.978	0.990	0.995	0.995	0.995			
$\text{Com } 3 \times 4$	0.989	0.998	0.998	0.999	0.999	0.999	0.999			
$\operatorname{Com} C \times T$	0	0.215	0	0.127	0	0	0.070			
Low interspecific s	similarity model									
$Com 1 \times 2$	0.001	0.501	0.003	0.503	0.011	0.011	0.505			
$\text{Com } 3 \times 4$	0.146	0.940	0.258	0.941	0.420	0.420	0.942			
$\text{Com}C\timesT$	0	0.700	0	0.700	0	0	0.700			
Uniform model ar	nd additional taxo	nomic level								
$Com 1 \times 2$	0.125	0.636	0.373	0.714	0.714	0.714	0.800			
$Com 3 \times 4$	0.531	0.959	0.821	0.968	0.952	0.952	0.978			
$\operatorname{Com} C \times T$	0.041	0.659	0.132	0.636	0.345	0.352	0.620			

Interspecific similarities were defined as, $\sigma_{Sokat-Sneath} = a/(a + 2b + 2c)$, $\sigma_{Jaccard} = a/(a + b + c)$, $\sigma_{Sorensen} = 2a/(2a + b + c)$, $\sigma_{Ochiai} = a/\sqrt{(a + b)(a + c)}$, with a = sum of branch lengths from the nearest common ancestor of two species to the root of the tree, b = sum of branch lengths from the first species to this nearest common ancestor and c = sum of branch lengths from the second species to this nearest common ancestor. Given the ultrametric tree in this particular example, $\sigma_{Sorensen} = \sigma_{Ochiai}$. Also, wherever the average similarities among species within compared communities are equal, $S_{Sorensen} = S_{Ochiai}$. Com 1 × 2 is the similarity among the Coleoptera communities (Com1 and Com 2) and Com 3 × 4 is the similarity among the Trichoptera communities (Com3 and Com4). Com C × T is the similarity between a community composed of Coleoptera only (either Com1 or Com2) and a community composed of Trichoptera only (either Com3 or Com4). Bold values indicate situations where Com 1 × 2 or Com 3 × 4 < Com C × T. Note that the data set in Fig. 2 was designed so that the similarities between communities with Coleoptera only and communities with Trichoptera only (i.e. communities 1 × 3, 1 × 4, 2 × 3, and 2 × 4) were all equal.

iour for an index of dissimilarity among communities is due to the fact that COMDIST is an absolute index. Indeed it calculates how different species from two distinct communities are without considering how different species from the same community are. When measuring (dis)similarities among communities, it is thus important to compare how (dis)similar species from different communities are with the level of (dis)similarity among species from the same community.

In addition, the real data set confirmed that S_Q behaves differently from the other indices when species are treated as maximally different (*i.e.* zero similarity among species). We also found very few maximally dissimilar stations (absence of similarities between species from distinct stations). Such scenarios of maximally dissimilar communities are likely to be infrequent in ecological studies that incorporate similarities among species, at least at local scales.

However, our simulations showed that the existence of maximally dissimilar species (and thus of maximally dissimilar sites) did not increase the differences between S_Q and the other indices. For instance, all indices tended to have close behaviour when the similarities among species were artificially and drastically increased. In contrast, S_Q showed distinct values when the interspecific similarities were decreased. By adding a taxonomic level, we increased similarities among species and eliminated the existence of maximally dissimilar species, but in spite of that, S_Q still provided high taxonomic similarity between stations with only Coleoptera and stations with only Trichoptera. In contrast, the new indices acknowledged low similarities between these stations. The values of S_Q also depended to some extent on the coefficients used to calculate the taxonomic and functional similarities among species. This does not mean that S_Q is meaningless. As highlighted by Pavoine, Dufour and Chessel (2004), D_Q (= 1 - S_Q) can be viewed as the distance between the centroids of two communities in a multidimensional space where species are positioned according to their functional or phylogenetic distances. It is now well-established that a single index cannot summarize all aspects of biodiversity. The same conclusion holds for (dis)similarity indices.

Comparing the new indices $S_{Sokal-Sneath}$, $S_{Jaccard}$, $S_{Sorensen}$ and S_{Ochiai} , we showed that all indices tend to rank communities similarly. However, the index values can be very different. This might be annoying if one interprets the index values in an intuitive way, from 0 meaning no similarity to 1 meaning complete similarity. Two data sets can thus be compared only if the same index is used. On the other hand, as the different indices have slightly different properties, they allow calculating community similarity from different viewpoints and perspectives. The differences among the indices only depend on how the three components $\sum_{ij} x_i z_j \sigma_{ij}$ (similarity among the species in different communities), $\sum_{ij} x_i x_j \sigma_{ij}$ (similarity among the species in the first community) and $\sum_{ij} z_i z_j \sigma_{ij}$ (similarity among the species in the second community) are combined into an index of similarity. For example, both indices S_{Ochiai} and



Fig. 3. Scatterplots between the new indices and the index S_Q applied to the real data set considering (a) minimum (zero) similarity among species; (b) taxonomic similarity; and (c) similarities in feeding habits. Spearman correlations between the new indices and S_Q are indicated on each panel.

 $S_{Sorensen}$ compare the similarity among the species in different communities (numerator) with the average similarity among the species in the same community (denominator). However, by substituting the geometric mean for the arithmetic mean of $\sum_{ij} x_i x_j \sigma_{ij}$ and $\sum_{ij} z_i z_j \sigma_{ij}$ in $S_{Sorensen}$ compared with Sochiai, the contribution of the species similarity within two communities to the value of the index is decreased just because the geometric mean leads to lower values than the arithmetic mean. From this point of view, the denominator of S_{Sokal-Sneath} and S_{Jaccard} does not contain solely the average similarities among species within communities but rather the difference between these similarities and the similarity among the species in different communities. Accordingly, in Ssokal-Sneath and $S_{Jaccard}$, the relevance of the similarities among the species in different communities is decreased compared to the similarities within communities.

Altogether, the interests of the new family of indices are thus (i) to extend traditional similarity measures to taxonomic, phylogenetic or functional similarities among species and communities; (ii) to be flexible in the choice of the species' weights (presence/absence, individual counts/biomass, relative abundance, etc.) and in the definition of similarities among species that can be computed in many different ways from taxonomy, phylogeny, functional dendrograms and functional data sets that can contain nominal, quantitative, binary, proportional and fuzzy variables (Appendix S2); and (iii) to provide a particular treatment of maximally dissimilar communities.

A last remark can be made on the development of this family of indices. Indices in columns 4 of Table 2 can be applied to vectors of species presences/absences within communities but also to any set of elements including evolutionary units (which would lead to the indices of Lozupone and Knight 2005 and Ferrier *et al.* 2007) for measuring the phylogenetic similarity between communities, and volumes in functional spaces (which would embed the index of Villéger, Novack-Gottshall & Mouillot 2011) for measuring the functional similarity between communities.

THE NEW FAMILY AS A UNIFIED FRAMEWORK FOR FUNCTIONAL SIMILARITY AND PHYLOGENETIC SIMILARITY INDICES

A previous paper showed that, when comparing functional diversity with phylogenetic diversity, the same mathematical indices should be used to avoid the risk of misinterpreting differences in functional and phylogenetic diversity by biological processes when the differences are just mathematical artefacts (Pavoine *et al.* 2013). This remark is also true for functional and phylogenetic similarity indices. This is why we developed

here a methodological framework that can apply to functional similarities or to (taxonomic) phylogenetic similarities among species and communities.

Among the indices of intercommunity similarity developed in the literature, some have been designed for functional traits (Villéger, Novack-Gottshall & Mouillot 2011), and others for phylogenies (Lozupone & Knight 2005; Ferrier *et al.* 2007; Pavoine, Love & Bonsall 2009; Ives & Helmus 2010; Nipperess, Faith & Barton 2010; Chiu, Jost & Chao 2014). On the contrary, Rao's D_Q has been suggested to measure any type of dissimilarity including taxonomic, genetic, phylogenetic and functional similarity (Nei & Li 1979; Rao 1982; Izsák & Papp 1995; Pavoine & Dolédec 2005; Ricotta 2005).

Many indices developed within the functional or phylogenetic context can be easily transposed to the other type of data (Pavoine & Bonsall 2011). While the development of traditional diversity indices was mainly interdisciplinary, the addition of functional vs. phylogenetic (dis)similarities among species has split research on diversity and (dis)similarity measures. The main difficulty when comparing functional and phylogenetic data is that phylogenetic data intrinsically imply a tree-shaped structure among species. For example, the indices developed by Lozupone and Knight (2005), Ferrier et al. (2007), Pavoine, Love and Bonsall (2009), Nipperess, Faith and Barton (2010), and Chiu, Jost and Chao (2014) depend on tree-shaped structures among species (phylogenies). In general, functional data might thus be used with these indices only if they are artificially transformed into functional dendrograms using clustering approaches as suggested by Petchey and Gaston (2002). Such approaches add methodological choices and the distortion of the data might be high when one or a few functional traits only are considered. This is because a tree is a multidimensional object (e.g. Nabben & Varga 1994), whereas a quantitative trait can be displayed in one only dimension. In contrast, our family of similarity indices can be used with many different data types provided similarity among species is bounded between 0 and 1.

Some of the solutions proposed to develop functional indices implied to transform functional data into the form of data traditionally used to measure compositional similarity among communities. For example, Robertson, McAlpine and Maron (2013) used a *species × trait* matrix to define functional groups. Then, they applied the Bray–Curtis index (Bray & Curtis 1957) on log-transformed summed densities of all members of functional groups. In a phylogenetic context, this solution could be adapted by measuring the Bray–Curtis index on log-transformed summed densities/abundances/biomasses of all members of clades. However, in comparison with our indices, this solution considers all functional groups (or clades) as maximally dissimilar, which is not always the case.

In conclusion, we have introduced a new family of similarity indices that is very flexible in the type of data used. By designing this family, we advised, wherever possible, for the use of coherent indices for measuring similarities among species and similarities among communities; for the use of relative similarity indices that integrate the diversity of the community and that are bounded between 0 and 1; and for the use of p.s.d. indices. This family embeds a large variety of similarity indices developed so far for measuring species, functional or phylogenetic diversity and (dis)similarity.

Acknowledgements

We thank János Podani and an anonymous reviewer for their comments. They both contributed to improve our paper.

Data accessibility

Data are available in the R package ade4 (Chessel, Dufour & Thioulouse 2004; R Development Core Team 2013). An R script can be uploaded as online supporting information (Appendix S4). The R script for the index COMDIST is available in the R package picante (Kembel *et al.* 2010).

References

- Bray, J.R. & Curtis, J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27, 325–349.
- Bryant, J.A., Lamanna, C., Morlon, H., Kerkhoff, A.J., Enquist, B.J. & Green, J.L. (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences* of the United States of America, 105, 11505–11511.
- Burt, C. (1948) The factorial study of temperamental traits. *British Journal of Psychology* (Statistical Section), 1, 178–203.
- Champely, S. & Chessel, D. (2002) Measuring biological diversity using Euclidean metrics. *Environmental and Ecological Statistics*, 9, 167–177.
- Chessel, D., Dufour, A.B. & Thioulouse, J. (2004) The ade4 package-I- One-table methods. *R News*, 4, 5–10.
- Chevenet, F., Dolédec, S. & Chessel, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology*, **31**, 295– 309.
- Chiu, C.-H., Jost, L. & Chao, A. (2014) Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs*, 84, 21–44.
- Dice, L.R. (1945) Measures of the amount of ecologic association between species. *Ecology*, 26, 297–302.
- Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity & Distributions*, 13, 252–264.
- Gleason, H.A. (1920) Some applications of the quadrat method. Bulletin of the Torrey Botanical Club, 47, 21–33.
- Gower, J.C. & Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3, 5–48.
- Graham, C.H. & Fine, P.V.A. (2008) Phylogenetic beta diversity: linking ecological and evolutionary processes across space and time. *Ecology Letters*, 11, 1265–1277.
- Horn, H.S. (1966) Measurement of "overlap" in comparative ecological studies. *The American Naturalist*, **100**, 419–424.
- Ives, A.R. & Helmus, M.R. (2010) Phylogenetic metrics of community similarity. *The American Naturalist*, **176**, E128–E142.
- Ivol, J.M., Guinand, B., Richoux, P. & Tachet, H. (1997) Longitudinal changes in Trichoptera and Coleoptera assemblages and environmental conditions in the Loire River (France). *Archiv fur Hydrobiologie*, 138, 525–557.
- Izsák, J. & Papp, L. (1995) Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. *Environmental and Ecological Statistics*, 2, 213–224.
- Jaccard, P. (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles, 37, 547–579.
- Jost, L. (2006) Entropy and diversity. Oikos, 113, 363-375.
- Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., Blomberg, S.P. & Webb, C.O. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26, 1463–1464.
- Legendre, P. & Legendre, L. (2012) Numerical Ecology. Elsevier, Amsterdam.
- Lozupone, C.A. & Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71, 8228–8235.

10 S. Pavoine & C. Ricotta

- Meirmans, P.G. (2006) Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution*, 60, 2399–2402.
- Morisita, M. (1959) Measuring of interspecific association and similarity between communities. *Memoirs of the Faculty of Science, Kyushu University Series E* (Biology), 3, 65–80.
- Nabben, R. & Varga, R.S. (1994) A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant Stieltjes matrix. *SIAM Journal of Matrix Analysis and Applications*, 15, 107–113.
- Nei, M. & Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76, 5269–5273.
- Nipperess, D.A., Faith, D.P. & Barton, K. (2010) Resemblance in phylogenetic diversity among ecological assemblages. *Journal of Vegetation Science*, 21, 809–820.
- Ochiai, A. (1957) Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of the Japanese Society of Scientific Fisheries*, 22, 526–530.
- Orlóci, L. (1967) An agglomerative method for classification of plant communities. *Journal of Ecology*, 55, 193–206.
- Pavoine, S. & Bonsall, M. (2011) Measuring biodiversity to explain community assembly: a unified approach. *Biological Reviews*, 86, 792–812.
- Pavoine, S. & Dolédec, S. (2005) The apportionment of quadratic entropy: a useful alternative for partitioning diversity in ecological data. *Environmental and Ecological Statistics*, 12, 125–138.
- Pavoine, S., Dufour, A.B. & Chessel, D. (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology*, **228**, 523–537.
- Pavoine, S., Love, M. & Bonsall, M.B. (2009) Hierarchical partitioning of evolutionary and ecological patterns in the organization of phylogenetically-structured species assemblages: application to rockfish (genus: *Sebastes*) in the Southern California Bight. *Ecology letters*, **12**, 898–908.
- Pavoine, S., Gasc, A., Bonsall, M.B. & Mason, N.W.H. (2013) Correlations between phylogenetic and functional diversity: mathematical artefacts or true ecological and evolutionary processes? *Journal of Vegetation Science*, 24, 781– 793.
- Petchey, O.L. & Gaston, K. (2002) Functional diversity (FD), species richness and community composition. *Ecology Letters*, 5, 402–411.
- Podani, J. (2000) Introduction to the Exploration of Multivariate Biological Data. Backhuys, Leiden.
- R Development Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rao, C.R. (1982) Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, 21, 24–43.
- Ricotta, C. (2005) Through the jungle of biological diversity. Acta Biotheoretica, 53, 29–38.
- Ricotta, C. & Szeidl, L. (2009) Diversity partitioning of Rao's quadratic entropy. *Theoretical Population Biology*, **76**, 299–302.

- Robertson, O.J., McAlpine, C. & Maron, M. (2013) Influence of interspecific competition and landscape structure on spatial homogenization of avian assemblages. *PLoS ONE*, 8, e65299.
- Sokal, R.R. & Sneath, P.H.A. (1963) Principles of Numerical Taxonomy. W. H. Freeman, San Francisco, California.
- Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter*, 5, 1–34.
- Tucker, L.R. (1951) A Method for Synthesis of Factor Analysis Studies (Personnel Research Section Report No. 984). Department of the Army, Washington, District of Columbia.
- Villéger, S., Novack-Gottshall, P.M. & Mouillot, D. (2011) The multidimensionality of the niche reveals functional diversity changes in benthic marine biotas across geological time. *Ecology Letters*, 14, 561–568.
- Webb, C.O., Ackerly, D.D. & Kembel, S.W. (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, 18, 2098–2100.
- Webb, C.O., Ackerly, D.D., McPeek, M.A. & Donoghue, M.J. (2002) Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, 33, 475–505.
- Wishart, D. (1969) An algorithm for hierarchical classification. *Biometrics*, 25, 165–170.

Received 27 November 2013; accepted 1 April 2014 Handling Editor: David Orme

Supporting Information

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

Appendix S1. Proofs.

Appendix S2. How to calculate similarities among species.

Appendix S3. Manual for R functions and examples of their use.

Appendix S4. R scripts for functions.

Appendix S5. R scripts for examples.