Ecology, 91(6), 2010, pp. 1850–1859

On goodness-of-fit measure for dendrogram-based analyses

Bastien Mérigot,[1] Jean-Pierre Durbec, and Jean-Claude Gaertner

Centre d'Océanologie de Marseille, UMR-CNRS 6117 Laboratoire de Microbiologie, Géochimie et Ecologie Marines (LMGEM), Université de la Méditerranée, Campus de Luminy, FR-13 009 Marseille, France

Abstract. Clustering methods are widely used tools in many aspects of science, such as ecology, medicine, or even market research, that commonly deal with dendrogram-based analyses. In such analyses, for a given initial dissimilarity matrix, the resulting dendrogram may strongly vary according to the selected clustering methods. However, numerous dendrogram-based analyses require adequate measurement for assessing of which of the clustering methods preserves most faithfully the initial dissimilarity matrix. While cophenetic correlation coefficient-based measures have been widely used for this purpose, we emphasize here that it is not always a suitable approach. We thus propose a measure based on a matrix norm, the 2-norm, to adequately check which of the resulting ultrametric distance matrices related to the dendrograms is the closest to the initial dissimilarity matrix. In addition, we also propose an objective way to define a benchmark value (threshold value) in order to assess whether the degree of conformity between the ultrametric distance matrix selected and the initial dissimilarity matrix is satisfactory. Our proposal may notably be incorporated within a recently proposed approach that involves the use of clustering methods in environmental science and beyond. In ecology, various functional diversity indices based on clustering species from their functional dissimilarities may benefit from this overall approach.

Key words: *clustering; dendrogram; functional classification; functional diversity; matrix norm.*

Introduction

Clustering methods have been applied to a wide variety of research fields, such as ecology, systematic and evolutionary biology, medicine, or even market research (Everitt et al. 2001). Beside their original use for describing multivariate structures, they constitute the preliminary step for dendrogram-based analyses which are developing apace (Petchey and Gaston 2007, 2009, Podani and Schmera 2007, Mouchet et al. 2008). For instance in ecology, various popular functional diversity indices are based on clustering species from their functional dissimilarities (Petchey et al. 2004, Petchey and Gaston 2006, Mouchet et al. 2008, Mouillot et al. 2008, Walker et al. 2008, Cianciaruso et al. 2009). Dendrogram-based analyses require different steps, including the choice of a (dis)similarity or distance index (e.g., Euclidean, Gower distances, Bray-Curtis dissimilarity, and so on) and a particular clustering method (e.g., UPGMA, Ward, and so on; see Petchey and Gaston 2002, 2006, 2007, Podani and Schmera 2006, 2007, Mouchet et al. 2008, Poos et al. 2009). The choice of the (dis)similarity/distance index is clearly essential in the building of a classification (Mouchet et al. 2008, Poos et al. 2009). In choosing a relevant index, the user may be guided by the aims of the study and the nature of the variables considered (e.g., Pavoine et al. 2009). Here we focused our study on the next step which is to determine the clustering algorithm giving an ultrametric matrix the most in agreement with the initial (dis)similarity/distance matrix provided by the preliminary index chosen. This step is also of major concern because different clustering algorithms can generate different dendrograms. This may translate into different results, such as different grouping of the species, different functional diversity index values as supported by the quantitative study on the FD index sensitivity to methodological choices performed by Poos et al. (2009), and/or different interpretation of the effects of anthropogenic and environmental factors. Thus, we feel that it is important to achieve a coherent classification relative to the respective intrinsic information stated by the initial matrix appeared important to be reached. In view of this, proposing a methodology to improve the way of selecting the clustering algorithm is clearly of importance. Indeed, in numerous research areas such as community structuring (Cao et al. 1997, Le Bec et al. 1997, Nicol et al. 2003, Vilchis et al. 2009), functional diversity (Petchey and Gaston 2006, Podani and Schmera 2006, Mouchet et al. 2008, Flynn et al. 2009) or genetics (Peeters and Martinelli 1989, Kantety et al. 1995, Mohammadi and Prasanna 2003, Gonçalves et al. 2008), to be relevant, the dendrogram, which is associated with an ultrametric distance matrix **U** (Legendre and Legendre 1998), has to correspond as

Manuscript received 30 July 2009; revised 25 September 2009; accepted 30 September 2009. Corresponding Editor: J. Franklin.

[1] Present address: UMR 212 Ecosystèmes Marine Exploités, Centre de Recherche Halieutique, Avenue Jean Monnet, BP 171, 34203 Sète Cedex, France.
E-mail: bastienmerigot@yahoo.fr

closely as possible to the initial inter-object dissimilarity matrix **D** (Sokal and Rohlf 1962, Mohammadi and Prasanna 2003, Podani and Schmera 2006, Mouchet et al. 2008). To investigate the goodness of fit of **D** by each ultrametric **U** provided by different clustering methods—i.e., to measure how well the initial dissimilarities are preserved by the dendrogram—most of the previous studies used the cophenetic correlation coefficient c (Sokal and Rohlf 1962, Blackburn et al. 2005, Podani and Schmera 2006, Petchey and Gaston 2009, Viketoft et al. 2009) or a derivative form ($\mathbf{D}_M = 1 - c^2$ in Mouchet et al. 2008), where c corresponds to the Pearson product moment correlation coefficient between the initial dissimilarities and the calculated ultrametric distances. However, measures based on the cophenetic correlation coefficient are not always suitable for that purpose (Farris 1969, Holgersson 1978, May 1999). In this context, we therefore propose the use of a more suitable goodness-of-fit measure based on a matrix norm.

Is cophenetic correlation suitable to measure the goodness of fit?

Cophenetic correlation coefficient c (Sokal and Rohlf 1962) is expected to highlight which of various dendrograms provided by different clustering methods (e.g., unweighted pair group method using arithmetic averages [UPGMA], weighted pair group method using arithmetic averages [WPGMA], Ward, and so on) gives the most faithful representation of the initial inter-object dissimilarities (see Mouchet et al. 2008). However, c does not explicitly measure the overall distance between elements d_{ij} of **D** and corresponding elements u_{ij} of **U** resulting from a clustering method (Holgersson 1978, May 1999), but only measures the intensity of linear relationship between the respective terms of **D** and **U**. Thus, while a threshold value of c equal to 0.80 is usually considered for a relatively low distortion of **D** by **U** (see Mouchet et al. 2008), it only involves a high linear relationship between u_{ij} and d_{ij} (May 1999). Because it is based on the latter criteria, c is not always relevant for properly assessing to what extent **U** is close to **D**.

First, c is very sensitive to a few relatively high values of u_{ij} and d_{ij} that are distinct from the majority of the other values. Because these few values (outliers) are capable of considerably changing the value of the correlation (Koopmans 1987), they can be entirely responsible for a high c even though the relationship between **U** and **D** is weak. For instance, c will be very sensitive to some rare highly different "objects" (e.g., species in some ecological studies) to which correspond high initial dissimilarity values and therefore high calculated ultrametric distances. Likewise, if an object is highly distant from the others and is aggregated at the set of the other objects at the last level of the hierarchical tree, then $(n - 1)$ ultrametric dissimilarities will be high and equal, and will have a high influence on c.

Second, because it is restricted to the quantification of the linear association between elements u_{ij} and d_{ij}, c might be not suitable to quantify the proximity between **D** and **U** when a non-monotonic relationship between u_{ij} and d_{ij} occurs. It can happen that the agglomerative scheme of the clustering algorithm leads to some ultrametric distances that are not all in the same order as the corresponding initial dissimilarities. Such a situation may result in a c value that provides a too optimistic view of the fit of **D** by **U**.

Third, a perfect match between **D** and **U** (i.e., $\mathbf{D} - \mathbf{U} = 0$) would require that points displaying u_{ij} as a function of d_{ij} in a matrix plot are on the line of the equation $y = x$ (first bissectrix). This is the baseline to adequately compare the two matrices and to check the goodness of fit of **D** by **U** (i.e., to find to which degree the u_{ij} are close to d_{ij}). Consequently, in the case where $\mathbf{D} = \mathbf{U}$ and points are on the first bissectrix, the coefficient of correlation is clearly equal to 1. However, even when **D** and **U** are not equal, a value of c equal to 1 can be obtained whereas the points are not on the first bissectrix, but linked by an affine relationship ($u_{ij} = a \times d_{ij} + b$, where a and b are real constants) with parameters a and/or b not close to 1 and 0 respectively. Consequently a high and same value of the cophenetic correlation coefficient c can be associated to very different cases, even where a relatively high deviation occurs between **U** and **D**.

For example, for an initial dissimilarity matrix **D**, in which only three objects A, B, C are considered for illustration (Table 1a), and two ultrametric distance matrix $\mathbf{U}_1$ (Table 1b) and $\mathbf{U}_2$ (Table 1c), the value of the cophenetic correlation coefficient c in both cases is 0.87, although the deviations from **D** are greater in $\mathbf{U}_2$ than in $\mathbf{U}_1$. This can be shown by the matrix plot (Fig. 1) representing the ultrametric distances in function of the

TABLE 1. (a) Dissimilarity matrix **D** and (b and c) ultrametric distance matrices $\mathbf{U}_1$ and $\mathbf{U}_2$, respectively.

	A	B	C
a) **D**			
A	0	2	6
B	2	0	4
C	6	4	0
b) $\mathbf{U}_1$			
A	0	2	5
B	2	0	5
C	5	5	0
c) $\mathbf{U}_2$			
A	0	2	15
B	2	0	15
C	15	15	0

Notes: $\mathbf{U}_1$ is obtained using unweighted pair group method using arithmetic averages (UPGMA) from **D**, and $\mathbf{U}_2$ is obtained by multiplying by three AC and BC distances of $\mathbf{U}_1$. Cophenetic correlation coefficient values (c) between **D** and $\mathbf{U}_1$ and between both **D** and $\mathbf{U}_2$ are the same ($c = 0.87$), implying that both ultrametric matrices are equivalent in respect to goodness of fit of **D**. However, $\mathbf{U}_2$ presents more distortion of **D** than does $\mathbf{U}_1$ as stated by respective 2-norm values (Eq. 1; 2-norm = 14.21 and 1.41, respectively). See also the difference of deviation from the first bissectrix in the matrix plot of Fig. 1.

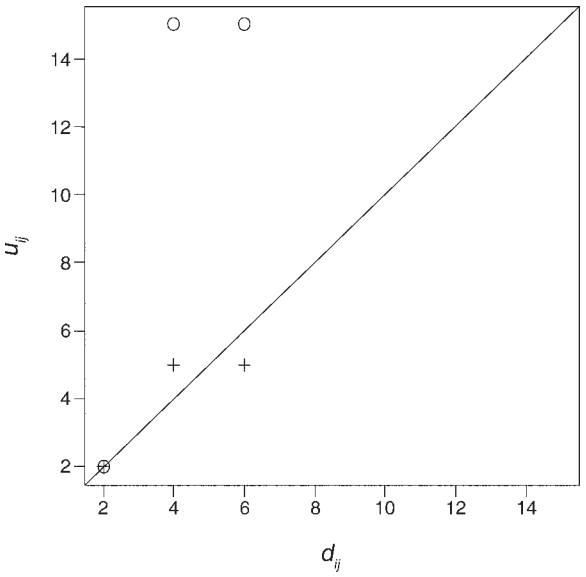

FIG. 1. Matrix plot in which the d_{ij} of the initial dissimilarities matrix **D** (Table 1a) is plotted vs. the u_{ij} of the ultrametric distances matrix. Points marked with plus signs show where the initial ultrametric matrix $\mathbf{U}_1$ is considered (Table 1b); points marked with circles show where $\mathbf{U}_2$ is considered (Table 1c). The line corresponds to the line of equation $y = x$ (first bissectrix), i.e., when the goodness of fit of **D** by **U** is perfect (**D** = **U**). The same value of the cophenetic correlation coefficient c of 0.87 is obtained in both cases, while they provided different goodness of fit of **D**, with a relatively high deviation between **D** and $\mathbf{U}_2$.

initial dissimilarities with the full line corresponding to the line of equation $y = x$ (first bissectrix), i.e., when **D** = **U**. While elements of **D** are relatively well fitted by those of $\mathbf{U}_1$ (plus signs in Fig. 1 are close to the bissectrix), most of those of $\mathbf{U}_2$ are far from being close to those of **D**, leading to points far from the first bissectrix (circles in Fig. 1).

To summarize, in the cases where the relationship between the initial dissimilarities and the ultrametric distances is not linear, the ranges of their values are different, outliers are present, or the parameter of the regression line a and/or b are far from 1 and 0 respectively (both leading to departure from the first bissectrix in a matrix plot), cophenetic correlation coefficient cannot give reliable information on the deviation between **D** and **U**.

The alternative candidates formerly proposed as goodness-of-fit measure were unsuitable for application in comparing **D** and **U** due to the dependent feature of u_{ij} and d_{ij} (Rohlf 1974). In particular, they did not provide a way to objectively find a threshold value under which **U** and **D** are considered to be close, i.e., to check the degree of distortion between the two matrices. To our knowledge, while Rohlf (1974) provided discussions, no such measure has been proposed. Likewise, some existing measures are scaled (i.e., the values are divided by their standard deviations [new variance = 1] and often are centered by their means [new mean = 0]) or are based on rank correlation coefficients (i.e., the order of the values and not the value themselves are considered), such as some listed by Rohlf (1974) or the RV coefficient (Escouffier 1973), prohibiting their use in the context of the present study. Indeed, when comparing an initial dissimilarity-distance matrix **D** to an ultrametric **U** to check the goodness of fit, it is essential to keep the differences of the value levels, i.e., the scale of the differences, between the two matrices. A scaled or a rank correlation-based measure will reject this information, which prevents objective quantification of the proximity between the two matrices.

METHODS

A goodness-of-fit index

To fill the lack of c in providing adequate assessment of the goodness of fit of the initial dissimilarity matrix **D** by **U**, we propose a measure based on a matrix norm of **D** − **U**. It quantifies more adequately the discrepancy between **D** and **U**. A norm allows us to define a distance between **D** and **U** which verifies the general properties of nonnegativity, symmetry, definiteness, triangle inequality (Mardia et al. 1992). If the distance between **D** and **U** is close to zero, then **U** can be considered as a good approximation of **D**.

Among the various norms defined on the set of all matrices (Golub and Van Loan 1996), we have chosen as distance the 2-norm of **D** − **U**. It is the greatest singular value $\lambda_{\max}$ of the **D** − **U** matrix (Golub and Van Loan 1996):

$$2\text{-norm}^2 = ||\mathbf{D} - \mathbf{U}||_2^2 = \lambda_{\max}^2. \quad (1)$$

For this norm, it can be shown that

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij} - u_{ij})^2 \leq \lambda_{\max}^2 \quad (2)$$

$$\max|d_{ij} - u_{ij}| \leq \lambda_{\max} \quad (3)$$

with n the number of objects.

Thus, as a consequence of the inequalities in Eqs. 2 and 3, if the 2-norm $\lambda_{\max}$ can be considered to be "low," then the matrices **D** and **U** can be considered as close to each other ("weak" distance). In this case, the mean squared errors will be "weak" and the maximum deviation between the terms of **D** and **U** will be relatively "small" (see *Threshold value* to define a threshold value under which 2-norm is considered as "low"). Consequently 2-norm quantifies the proximity between **D** and **U** and it is equal to zero if and only if **D** = **U**. Then 2-norm can be used as a goodness-of-fit measure to select the **U** among the different ultrametrics obtained from various clustering algorithms that most faithfully preserves **D**.

It is worth noting that the advantages of using the singular value of the matrix **D** − **U** as we proposed

instead of other norms are multiple: (1) it allows us to bound the mean squared differences (mean squared error) between d_{ij} and u_{ij}; (2) all the matrix terms are taken into account in its calculation; (3) it allows us to bound the absolute value of the maximum difference (absolute maximum error); (4) if the maximum singular value of a random symmetric matrix is close to zero, the other singular values are also close to zero, then its spectrum is flat (this being often considered as characteristic of a pure multivariate matrix error); and (5) the probability distribution of the maximum singular value of a symmetric random matrix of zero expectation can be computed under weak distributional hypotheses (allows calculation of a benchmark (threshold) value of 2-norm in order to assess whether the degree of adequacy between the ultrametric distance matrix chosen and the initial dissimilarity matrix is satisfactory). All these benefits would not be offered by other measures, such as the max norm or even with the sum of squared absolute values of the deviations of the *d*s and *u*s. For such measures, obtaining a threshold value to quantify the quality of the adjustment of **D** by **U** would only be possible on the basis of purely empirical considerations (cf. Kruskal stress for multi-dimensional scaling [MDS]). In addition, it is worth noting that the 2-norm of a matrix is currently used in many multivariate data analysis methods, such as regression, principal component analysis or multidimensional scaling (Krzanowski 2000).

In the first example above (Table 1, Fig. 1), it is shown that a same value of c (i.e., 0.87) can be reached for different cases that differ regarding goodness of fit of **D**. It is worth noting that the 2-norm is equal to 1.41 between **D** and $\mathbf{U}_1$, and 14.21 between **D** and $\mathbf{U}_2$, i.e., 2-norm has been multiplied by about 10. This allows to detect the relatively better fit of **D** by $\mathbf{U}_1$ in contrast to those provided by $\mathbf{U}_2$ (see matrix plot in Fig. 1), while c considers both cases as equal and giving relatively high fits with $c = 0.87$.

Threshold value

Once the closest ultrametric **U** to **D** is adopted on the basis of 2-norm among the various matrices resulting from the different clustering algorithms, an additional interesting step would be to select a benchmark value for the 2-norm beyond which the approximation of **D** by the ultrametric **U** will be rejected. This allows us to measure the distortion between the initial dissimilarity matrix **D** and the selected ultrametric distance matrix **U** to check whether the best of a set of hierarchical classifications leads to a **U** far from **D**.

Unfortunately, a precise formal statistical procedure is very difficult to implement for that purpose. However, all the same, we can consider that the goodness of fit, as measured by 2-norm, is sufficiently high to accept **U** as close to **D** when 2-norm is inferior to the greatest singular value in absolute value of a n symmetric random matrix of which the terms are independent, with zero means and variances of terms inferior to a positive number σ^2 (σ being standard deviation), and with arbitrary distributions. The 2-norm of this random matrix is asymptotically of order $2\sigma\sqrt{n}$ (Achlioptas 2004). Consequently, we will admit that **D** is close to **U** if

$$||\mathbf{D} - \mathbf{U}||_2 = \lambda_{\max} \leq 2\sigma\sqrt{n}. \quad (4)$$

Clearly, the terms of the $\mathbf{D} - \mathbf{U}$ matrix are not independent, but here we seek only an approximated bound for rejecting the approximation of **D** by **U**. σ^2 is a bound of the variances of the terms of the random error matrix $\mathbf{D} - \mathbf{U}$, that is to say a finite positive value greater than all these variances. It can empirically be considered as the maximum acceptable mean squared deviation between the initial dissimilarities and the calculated ultrametric distances when they are considered as satisfyingly close. If σ^2 has a large value, **U** will be considered near to **D**, even though one or some wide deviations from zero are present in the errors ($e_{ij} = d_{ij} - u_{ij}$, $i, j = 1, 2, \ldots n$; $i < j$), because the probabilities of wide deviations are not negligible for such a σ^2. That bound is a measure of scale simply attempting to estimate the maximum acceptable variability of the errors under satisfying adjustment of **D** by **U** hypothesis.

Several choices are possible for estimating the σ^2 bound. We have chosen the sum of the empirical variance of the initial dissimilarities and the empirical variance of the ultrametric distances. Assuming that the correlation between dissimilarities and ultrametric distances is positive, the variance of the differences between dissimilarities and ultrametic distances is inferior to that sum. It is preferable to use "robustified" calculations for the computation of variances. Indeed, presence of outliers could be responsible for a very large value of the variances and consequently for a large value of σ^2. One or some important errors could be observed without the hypothesis of good adjustment of **D** by **U** being rejected. However, large errors are not acceptable when **U** was considered near to **D**. A "robustifying" method can be to discard from calculations of variances the observed errors the farthest from zero. This can be achieved by an iterative procedure discarding the values the farthest from zero, one after the other until variances do not vary any more (Koopmans 1987).

Simulation study

Simulations have been performed to study the behavior of the goodness-of-fit measures c and 2-norm. A total of 5000 assemblages of 20 species were computed from five simulated functional traits randomly sampled among (1) a normal (symmetric) and (2) a log-normal (asymmetric) distributions with a mean of 0 and a standard deviation of 1. Then for each species assemblages, a Euclidean distance was applied on the species functional traits data to provide the initial distance matrix from which the different clustering methods were

computed. The clustering methods chosen are widely used in ecology, particularly those presented in the methodological framework of Mouchet et al. (2008) which consider the family of hierarchical agglomerative classifications (i.e., Ward method, single and complete linkages, unweighted pair group method using arithmetic averages [UPGMA], weighted pair group method using arithmetic averages [WPGMA], weighted pair group method using centroids [WPGMC], unweighted pair group method using centroids [UPGMC], and consensus method; see Mouchet et al. 2008 for details on respective clustering methods). The consensus method applied by Mouchet et al. (2008) is computed from the n ultrametrics obtained by the different clustering algorithms previously cited. It is designed to provide an ultrametric $\mathbf{U}_c$ the "closest" to a set of n ultrametrics ($\mathbf{U}_i$, $i = 1, 2, \ldots, n$) obtained by the different clustering algorithms computed from the same dissimilarity matrix **D**. The consensus ultrametric $\mathbf{U}_c$ is obtained by minimizing the following criterion:

$$L(\mathbf{U}_c) = \min_{\mathbf{U}} \left(\left[\sum_{i=1}^{n} w_i \mathbf{U}_i \right] - \mathbf{U} \right)^2$$

where **U** is an ultrametric and w_i, the weight given to $\mathbf{U}_i$ ($w_i \geq 0$, $\Sigma_{i=1}^{n} w_i = 1$).

While the above methodology starting from a normal distribution (symmetric) to simulate species functional traits comes from published studies (Petchey and Gaston 2006, Mouchet et al. 2008), we also consider a log-normal distribution (asymmetric) to study the behavior of c and 2-norm.

Computations of 2-norm and the corresponding threshold value are available from the authors with an R script (R statistical environment, R Development Core Team 2009). It can be incorporated easily into the R script previously provided by Mouchet et al. (2008), which also includes Owen Petchey's script for the FD index calculation (Petchey and Gaston 2002, 2006). The computation of the consensus method is freely available from the package Clue (Hornik 2009).

Simulations results

For simulations in which traits were randomly sampled among a normal distribution (Fig. 2a, c), the main trend for c values shown by box plots (Fig. 2a) underlined that Ward method achieved most often the worst fit of **D** (median of $c = 0.54$, Fig. 2a) while UPGMA reached the best fit in most simulations (median of $c = 0.72$, Fig. 2a). However, it is worth noting that values of c are quite close across the different clustering algorithms (see the scale of y-coordinates and the overlap of box plots in Fig. 2a). In contrast, 2-norm distribution values discriminated better which clustering algorithms provided the best fit of **D** over all the 5000 simulations (see the differences in the values dispersions of the 2-norm between clustering algorithms, i.e., the extent of box plots and the degree of their overlaps, in Fig. 2c). According to 2-norm, the Ward method appeared to provide in general the worst fit of **D** (median of 2-norm = 73.48, Fig. 2c), in contrast to UPGMA followed by WPGMA and the consensus method which achieved the best fits in most simulations (median of 2-norm = 6.63, 8.84 and 9.71, respectively, see Fig. 2c), while all the other clustering methods clearly provided relatively worse fit of **D** (see Fig. 2c in contrast to Fig. 2a).

Interestingly, when considering simulations in which traits were sampled among a log-normal distribution (Fig. 2b, d), c values were inflated toward very high values over all the 5000 simulations (see Fig. 2b in comparison to Fig. 2a). Indeed, while the Ward method provided in general the worst fit of **D** according to c with a relatively high median of 0.80 (see Fig. 2b in comparison to Fig. 2a), all the other clustering methods provided very high c values over all the simulations (median up to $c = 0.90$ for most of the methods, see Fig. 2b). In contrast, 2-norm distributions values (Fig. 2d) discriminated much better which clustering algorithms provided the best fit of **D** over all the 5000 simulations, following the quite similar results from 2-norm found for simulations based on normal distribution (Fig. 2c). Indeed, according to 2-norm distributions values, UPGMA reached the best fit in most simulations, followed by WPGMA and consensus method (median of 2-norm = 10.57, 14.57, and 17.26, respectively, see Fig. 2d), while all the other clustering methods clearly provided relatively worse fit of **D** (see Fig. 2d in comparison to Fig. 2b).

As already stated, c cannot necessary give reliable information on the deviation between **D** and **U**. For instance, in Fig. 3, for a given **D**, c is equal to 0.92 for two different **U** and **U**′ related to dendrograms provided by single linkage and UPGMA clustering methods. Thus, using c, both situations would be considered to provide the same fit of **D**. However, Fig. 3 shows that the respective dendrograms related to **U** and **U**′ display notable differences, both in the scale of dissimilarities (see comparisons of y-coordinates of u_{ij} [Fig. 3a, c] and u'_{ij} values [Fig. 3b, d]) and the structure of the dendrogram. Even resulting from the same initial dissimilarity matrix, only one of them is closest to the species dissimilarities stated in **D**: 2-norm provided two different values (34.65 and 9.49, respectively), and matrix plots (Fig. 3c, d) showed difference of deviation from the first bissectrix. Hence in this case, in contrast to the information provided by c, our approach shows that the UPGMA provides in fact the ultrametric distance matrix closest to **D** among the considered ultrametrics.

In addition, the calculation of $2\sigma\sqrt{n}$ for the example of the UPGMA method (Fig. 3b, d) gave the threshold value of 2-norm equal to 29.34. As 2-norm is equal to 9.49 in that case, we can accept the ultrametric matrix obtained with the UPGMA as a reasonable approximation of the initial dissimilarity matrix **D**. Regarding the single linkage method (Fig. 3a, c), that provided a worse approximation

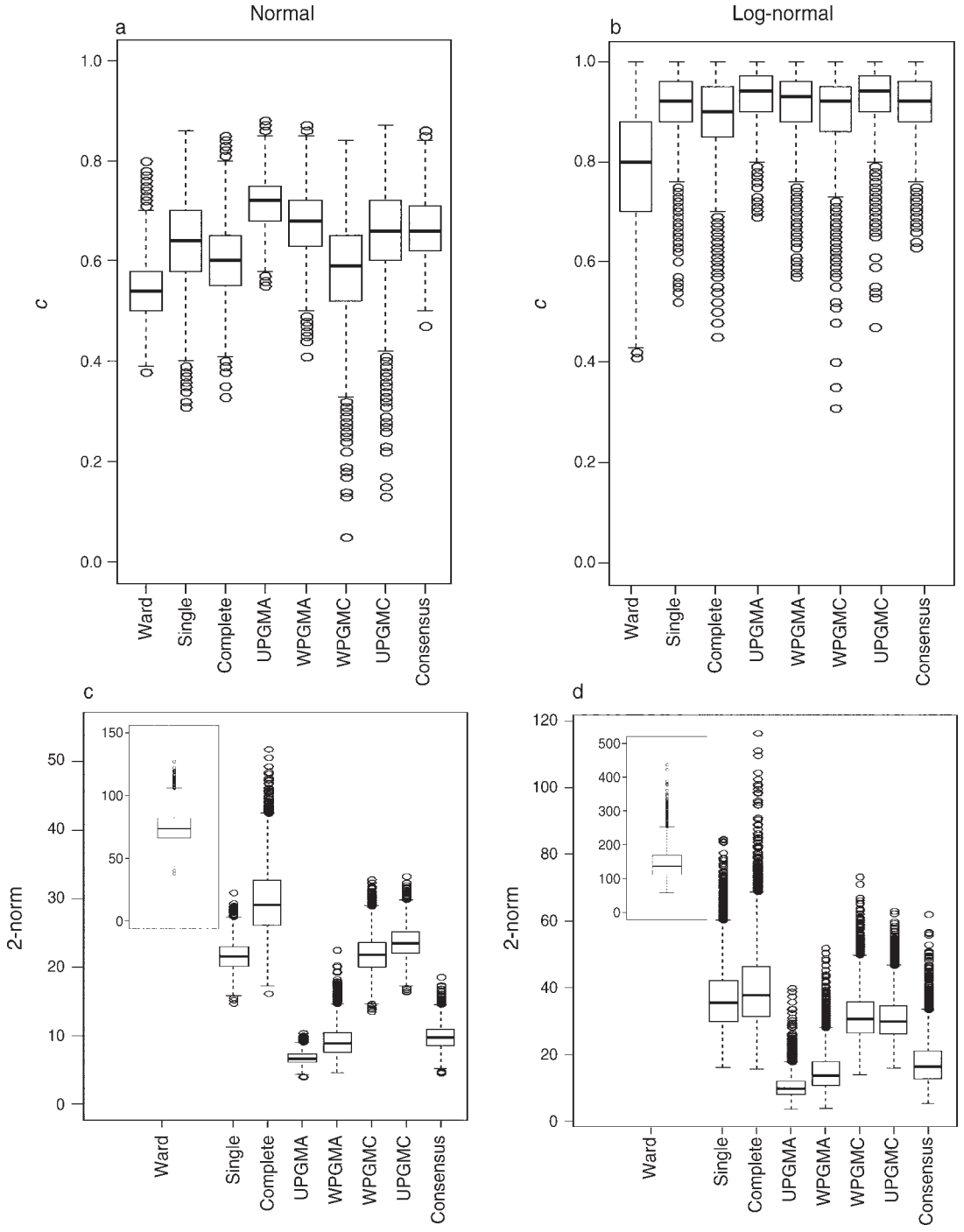

FIG. 2. Box plot of the (a, b) cophenetic correlation coefficient c and (c, d) 2-norm for 5000 simulated initial Euclidean distance matrices and ultrametric distance matrices provided by different clustering algorithms. The five species traits values used to compute the 5000 simulated assemblages of 20 species were randomly sampled among normal (a, c) and log-normal (b, d) distributions with mean 0 and standard deviation 1. Box components are respectively the first, second (i.e., median value), and third quartiles; lower and upper whiskers out from the box are adjacent values, and points are outliers. Clustering algorithm abbreviations are: unweighted pair group method using arithmetic averages (UPGMA), weighted pair group method using arithmetic averages (WPGMA), weighted pair group method using centroids (WPGMC), and unweighted pair group method using centroids (UPGMC).

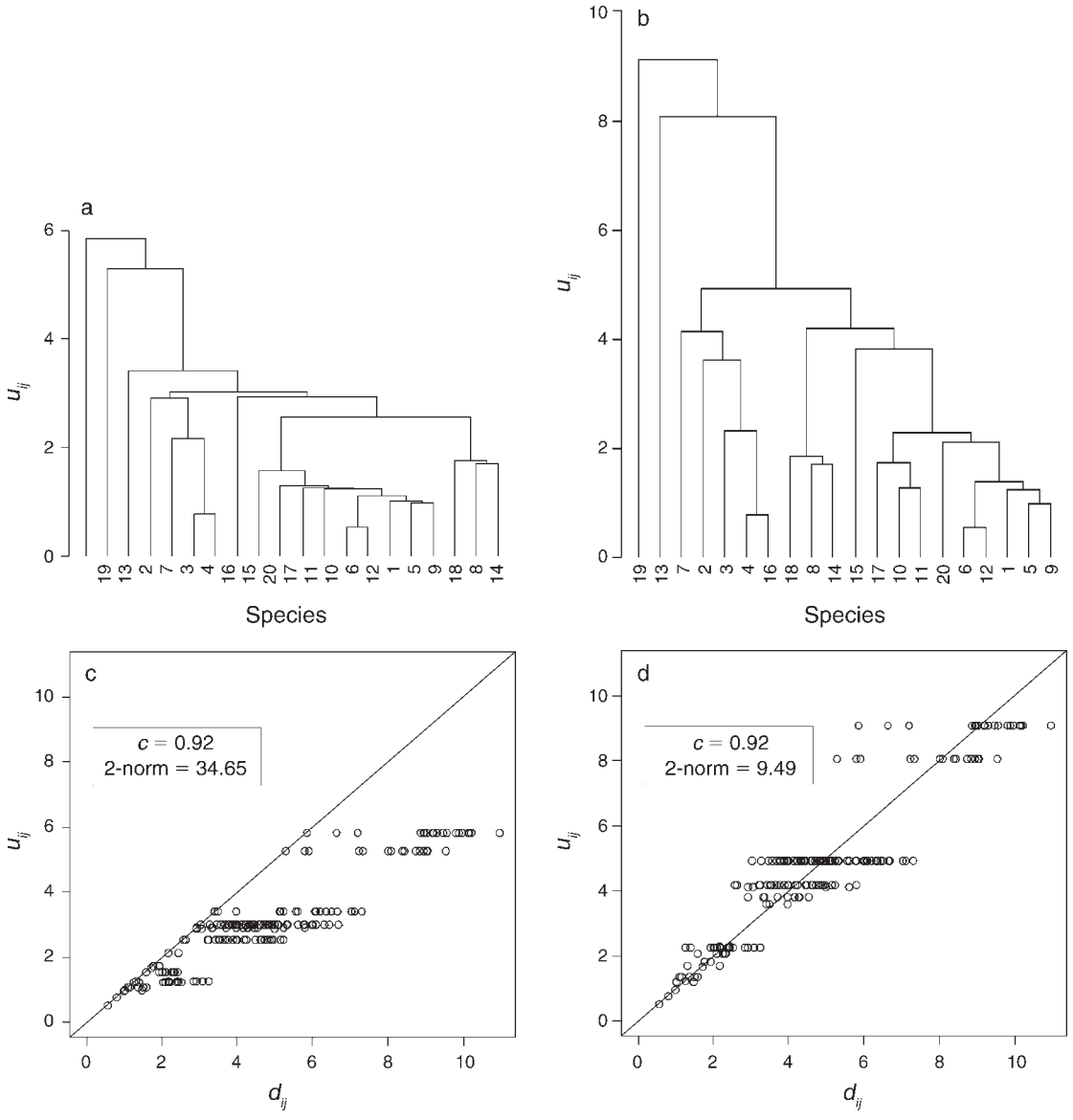

FIG. 3. Hierarchical classifications of a Euclidean distance matrix computed from a simulated assemblage composed of 20 species characterized by five independent functional traits. The species traits values were a random log-normal deviate with mean 0 and standard deviation 1. Clustering methods are (a) single linkage and (b) UPGMA. Matrix plots of **D** and **U** are shown in panels (c) and (d), respectively, with the solid line corresponding to the line of equation $y = x$ (first bissectrix), i.e., when the goodness of fit of **D** by **U** is perfect ($\mathbf{D} = \mathbf{U}$). Data are provided in the Supplement.

of **D** (2-norm = 34.65) than the UPGMA method, the threshold value was 24.96. Consequently in that case, ultrametrics provided by the single linkage method had to be rejected as an adequate approximation of **D**.

Likewise, c can provide other counterintuitive results. Among different ultrametric matrices **U** produced by various clustering algorithms from the same **D** matrix, a particular **U** could be considered as the best fit considering the values of c, while 2-norm adequately highlights which is the closest to **D**. An example of this is displayed in Fig. 4 in which the tree in panel a could be kept as c (a) $>$ c (b), while the tree in panel b is in fact closer to **D** since 2-norm (b) $<$ 2-norm (a) (see also matrix plot in Fig. 4c, d and difference of deviation form the first bissectrix).

CONCLUSION

In many research areas such as functional diversity (e.g., Mouchet et al. 2008, Flynn et al. 2009) or genetics (e.g., Mohammadi and Prasanna 2003, Gonçalves et al. 2008), the relevance of analyses based on a hierarchical clustering method is strongly dependent on the degree

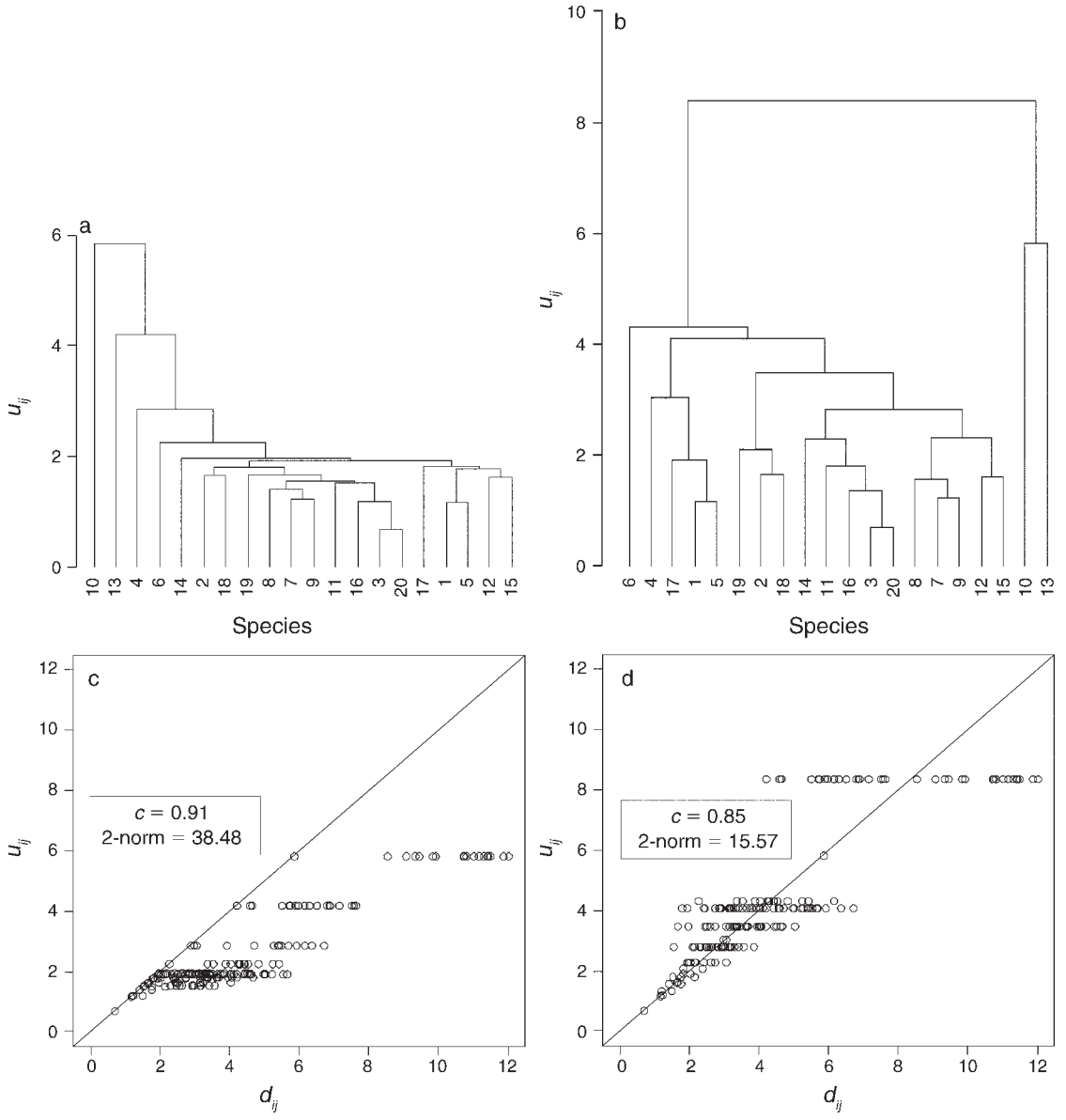

FIG. 4. Hierarchical classifications of a Euclidean distance matrix computed from a simulated assemblage composed of 20 species characterized by five independent functional traits. The species traits values were a random log-normal deviate with mean 0 and standard deviation 1. Clustering methods are (a) single linkage and (b) UPGMA. Matrix plots of **D** and **U** are shown in panels (c) and (d), respectively, with the solid line corresponding to the line of equation $y = x$ (first bissectrix), i.e., when the goodness of fit of **D** by **U** is perfect ($\mathbf{D} = \mathbf{U}$). Data are provided in the Supplement.

to which the resulting ultrametric distance matrix **U** is close to the initial dissimilarity matrix **D** between pairs of objects (Gonçalves et al. 2008, Mouchet et al. 2008). Here we underline that the most traditionally used approach, based on the cophenetic correlation coefficient c is not always a suitable method to be used for goodness-of-fit measures. We advocate that, for general use, a matrix norm-based measure (2-norm) would be more relevant for explicitly quantifying the discrepancy between the ultrametric **U** resulting from a clustering method and its initial dissimilarity matrix **D**. Our contribution might improve and enrich in a constructive way the approach proposed recently by Mouchet et al. (2008) that concerns the use of clustering methods in many fields of science. Overall, it consists in stating the objectives of the study, checking the nature of the variables considered and choosing accordingly the most suitable (dis)similarity/distance measure (Pavoine et al. 2009). Then checking which clustering algorithm should be chosen among the various possibilities by

means of 2-norm (lowest value). The degree of faithfulness between **D** and the chosen **U** can be then checked by comparing the 2-norm value to the respective benchmark value, with a 2-norm value below this threshold for a satisfactorily approximation of **D.** In ecology, various functional diversity indices based on clustering species from their functional dissimilarities may benefit from this approach. This is notably the case of the popular FD index (Petchey and Gaston 2002, 2006, 2009) and its new extended version (Cianciaruso et al. 2009), or the recent conservation of biological originality index (Mouillot et al. 2008) which are computed from ultrametric distance matrix related to a dendrogram. More broadly, this approach might be of value for any analyses that involve clustering algorithms when it is necessary to keep in the resulting dendrogram the most faithful representation of the initial biological, ecological or medical dissimilarity data sets.

Acknowledgments

We thank O. van Tongeren and an anonymous referee for their constructive comments on an earlier version of the manuscript. We are grateful to M. Paul for correcting the English of the paper.

Literature Cited

Achlioptas, D. 2004. Random data matrix in data analysis. Lecture Notes in Computer Science 3202:1–7.

Blackburn, T. M., O. L. Petchey, P. Cassey, and K. J. Gaston. 2005. Functional diversity of mammalian predators and extinction in island birds. Ecology 86:2916–2923.

Cao, Y., A. W. Bark, and W. P. Williams. 1997. A comparison of clustering methods for river benthic community analysis. Hydrobiologia 347:25–40.

Cianciaruso, M. V., M. A. Batalha, K. J. Gaston, and O. L. Petchey. 2009. Including intraspecific variability in functional diversity. Ecology 90:81–89.

Escouffier, Y. 1973. The treatment of the vectoriel variables. Biometrics 29:751–760.

Everitt, B. S., S. Landau, and M. Leese. 2001. Cluster analysis. Hodder Arnold Publication, London, UK.

Farris, J. S. 1969. On the cophenetic correlation coefficient. Systematic Biology 18:279–285.

Flynn, D. F. B., M. Gogol-Prokurat, T. Nogeire, N. Molinari, B. Trautman Richers, B. B. Lin, N. Simpson, M. M. Mayfield, and F. DeClerck. 2009. Loss of functional diversity under land use intensification across multiple taxa. Ecology Letters 12:22–33.

Golub, G. H., and C. F. Van Loan. 1996. Matrix computations. Johns Hopkins University Press, Baltimore, Maryland, USA.

Gonçalves, L. S. A., R. Rodrigues, A. T. Amaral, M. Karasawa, and C. P. Sudre. 2008. Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions. Genetics and Molecular Research 7:1289–1297.

Holgersson, M. 1978. The limited value of cophenetic correlation as a clustering criterion. Pattern Recognition 10: 287–295.

Hornik, K. 2009. CLUE: cluster ensembles. R package version 0.3-27. ⟨http://CRAN.R-project.org/package=clue⟩

Kantety, R. V., X. P. Zeng, J. L. Bennetzen, and B. E. Zehr. 1995. Assessment of genetic diversity in dent and popcorn (*Zea mays* L.) inbred lines using inter-simple sequence repeat (ISSR) amplification. Molecular Breeding 1:365–373.

Koopmans, L. H. 1987. An introduction to contemporary statistics. Duxbury Press, Boston, Massachusetts, USA.

Krzanowski, W. J. 2000. Principles of multivariate analysis: a user's perspective. University Press, Oxford, UK.

Le Bec, C., C. Belin, J. C. Gaertner, B. Beliaeff, B. Raffin, and F. Ibanez. 1997. Time series of the French phytoplankton monitoring network (REPHY). Study of two zones of the west Mediterranean coast. Oceanologica Acta 20:101–108.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Elsevier Science BV, Amsterdam, The Netherlands.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1992. Multivariate analysis. Academic Press, New York, New York, USA.

May, A. C. W. 1999. Towards a more meaningful hierarchical classification of amino acid scoring matrices. Protein Engineering Design and Selection 12:707–712.

Mohammadi, S. A., and B. M. Prasanna. 2003. Analysis of genetic diversity in crop plants: salient statistical tools and considerations. Crop Science 43:1235–1248.

Mouchet, M., F. Guilhaumon, S. Villeger, N. W. H. Mason, J. A. Tomasini, and D. Mouillot. 2008. Towards a consensus for calculating dendrogram-based functional diversity indices. Oikos 117:794–800.

Mouillot, D., J. M. Culioli, D. Pelletier, and J. A. Tomasini. 2008. Do we protect biological originality in protected areas? A new index and an application to the Bonifacio Strait Natural Reserve. Biological Conservation 141:1569–1580.

Nicol, G. W., L. A. Glover, and J. I. Prosser. 2003. Spatial analysis of archaeal community structure in grassland soil. Applied and Environmental Microbiology 69:7420–7429.

Pavoine, S., J. Vallet, A. B. Dufour, S. Gachet, and H. Daniel. 2009. On the challenge of treating various types of variables: application for improving the measurement of functional diversity. Oikos 118:391–402.

Peeters, J. P., and J. A. Martinelli. 1989. Hierarchical cluster analysis as a tool to manage variation in germplasm collections. Theoretical and Applied Genetics 78:42–48.

Petchey, O. L., and K. J. Gaston. 2002. Functional diversity (FD), species richness and community composition. Ecology Letters 5:402–411.

Petchey, O. L., and K. J. Gaston. 2006. Functional diversity: back to basics and looking forward. Ecology Letters 9:741–758.

Petchey, O. L., and K. J. Gaston. 2007. Dendrograms and measuring functional diversity. Oikos 116:1422–1426.

Petchey, O. L., and K. J. Gaston. 2009. Dendrograms and measures of functional diversity: a second installment. Oikos 118:1118–1120.

Petchey, O. L., A. Hector, and K. J. Gaston. 2004. How do different measures of functional diversity perform? Ecology 85:847–857.

Podani, J., and D. Schmera. 2006. On dendrogram-based measures of functional diversity. Oikos 115:179–185.

Podani, J., and D. Schmera. 2007. How should a dendrograms based measure of functional diversity function? A rejoinder to Petchey and Gaston. Oikos 116:1427–1430.

Poos, M. S., S. C. Walker, and D. A. Jackson. 2009. Functional diversity indices can be driven by methodological choices and species richness. Ecology 90:341–347.

R Development Core Team 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ⟨http://www.R-project.org⟩

Rohlf, F. J. 1974. Methods of comparing classifications. Annual Review of Ecology and Systematics 5:101–113.

Sokal, R. R., and F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. Taxon 9:33–40.

Viketoft, M., J. Bengtsson, B. Sohlenius, M. P. Berg, O. Petchey, C. Palmborg, and K. Huss-Danell. 2009. Long-term

effects of plant diversity and composition on soil nematode communities in model grasslands. Ecology 90:90–99.
Vilchis, L. I., L. T. Ballance, and W. Watson. 2009. Temporal variability of neustonic ichthyoplankton assemblages of the eastern Pacific warm pool: can community structure be linked to climate variability? Deep-Sea Research I 56:125–140.
Walker, S. C., M. S. Poos, and D. A. Jackson. 2008. Functional rarefaction: estimating functional diversity from field data. Oikos 117:286–296.

SUPPLEMENT

Simulated data sets of traits and initial and ultrametric distances used for Figs. 3 and 4 (*Ecological Archives* E091-122-S1).