*Paleobiology*, 48(2), 2022, pp. 284–301
DOI: 10.1017/pab.2021.33

PALEOBIOLOGY
A PUBLICATION OF THE Paleontological SOCIETY

*Article*

# An abundance- and morphology-based similarity index

*Daniel G. Dick** and *Marc Laflamme*

*Abstract*.—Classic similarity indices measure community resemblance in terms of incidence (the number of shared species) and abundance (the extent to which the shared species are an equivalently large component of the ecosystem). Here we describe a general method for increasing the amount of information contained in the output of these indices and describe a new "soft" ecological similarity measure (here called "soft Chao-Jaccard similarity"). The new measure quantifies community resemblance in terms of shared species, while accounting for intraspecific variation in abundance and morphology between samples. We demonstrate how our proposed measure can reconstruct short ecological gradients using random samples of taxa, recognizing patterns that are completely missed by classic measures of similarity. To demonstrate the utility of our new index, we reconstruct a morphological gradient driven by river flow velocity using random samples drawn from simulated and real-world data. Results suggest that the new index can be used to recognize complex short ecological gradients in settings where only information about specimens is available. We include open-source R code for calculating the proposed index.

*Daniel G. Dick and Marc Laflamme. Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, Ontario, L5L 1C6, Canada. E-mail: daniel.dick@mail.utoronto.ca, marc.laflamme@utoronto.ca*

Accepted: 13 September 2021
*Corresponding author.

## Introduction

Recognizing and interpreting the types of environmental gradients controlling the distribution of taxa in a region is a key goal of ecological research. One of the most popular means of exploring/reconstructing underlying gradients is the use of ordination techniques (Kowalewski et al. 1994; Patzkowsky 1995; Holland et al. 2001; Scarponi and Kowalewski 2004; Holland and Patzkowsky 2007). While ordination techniques offer an effective means of indirect gradient analysis (e.g., Holland and Patzkowsky 2007), the end result is inextricably related to the type of gradient and the choice of similarity index (Legendre and Gallagher 2001; Podani and Miklós 2002; Legendre and Legendre 2012). While commonly used similarity indices (e.g., Jaccard, Sørensen-Dice, Bray-Curtis, Horn-Morisita) can be used to order samples along a gradient with extensive faunal turnover (a "long" gradient), ordering random samples (e.g., quadrats, transects) along a "short" ecological gradient (one primarily characterized by morphological/phenotypic change) is more difficult. Much of this difficulty stems from the fact that conventional measures of compositional (dis)similarity give no weight to intraspecific morphological variation. Furthermore, studies on modern short gradients suggest that only a subset of taxa in a region will show a clear morphological/phenotypic response to the changing conditions (Puijalon and Bornette 2004; Kramer et al. 2020), and the taxa that do respond are often not known a priori. Given this, an ideal similarity index would be one that incorporates information about shifts in incidence, abundance, and intraspecific morphology, allowing researchers to reconstruct short gradients from random samples containing individuals of all species known from a given region—including those that show no obvious response to the change in conditions.

Indirect gradient analysis refers to a suite of techniques that use measures of the dissimilarity between samples to identify latent environmental variables controlling their distribution (Whittaker 1967; Walker 2015). For example, a steady shift in the number of presumed

© The Author(s), 2021. Published by Cambridge University Press on behalf of The Paleontological Society. 0094-8373/22

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

shallow-water taxa relative to the number of deep-water taxa can indicate a depth gradient controlling the distribution of samples (Scarponi and Kowalewski 2004; Holland and Patzkowsky 2007). This approach is contrasted with direct gradient analysis, wherein the environmental variables of interest are measured and directly related to measures of sample composition (Whittaker 1967; Ter Braak 1986; Walker 2015). In direct gradient analysis, measures of the environment (salinity, water-table depth, pH, stable isotope ratios, sedimentary proxies of water depth, velocity, etc.) can be correlated with ordination axes summarizing sample composition (ratios of different species of diatoms, pollen grains, testate amoebae, bivalves, gastropods, etc.) using linear or nonlinear regression approaches (Whittaker 1967; Booth 2001; Walker 2015). For example, Brett et al. (2007) used the correlation between detrended correspondence analysis axes (summarizing fossil biofacies composition) and sedimentary proxies of water depth to show long-term compositional stability and tracking of local environmental changes by common species. In both indirect and direct gradient analyses (when based on ordination axes), the outcome is determined by the choice of similarity measure (Podani and Miklós 2002; Legendre and Legendre 2012). Classic measures of the similarity/dissimilarity between samples (e.g., Jaccard 1901; Bray and Curtis 1957) are, in effect, blind to much of the information needed to reconstruct morphological gradients, while conventional morphology-based indices (e.g., Gower 1971) give no weight to incidence and abundance. Here we show how a concept developed in the field of soft computing (soft similarity) can solve this problem, allowing ecologists to produce ordinations of samples from different localities that incorporate variation in incidence, abundance, and intraspecific morphology. By "softening" existing similarity indices, researchers are able to produce ordination plots, clusters, and classifications that take into account far more information than is usually possible. While much of our discussion focuses on ordering random samples along subtle underlying gradients, the new similarity indices described here serve as generalized measures of community resemblance, incorporating more information than classic similarity indices.

## Methods and Materials

Here we explore how two popular measures of ecological similarity (the Jaccard index [Jaccard 1901] and the Chao-Jaccard index [Chao et al. 2005]) can be "softened" (sensu Jimenez et al. 2010, 2016), increasing the amount of information contained in the final output value. Specifically, we aim to modify these two indices such that the contribution of a given taxon to the numerical similarity of any two samples takes into account the extent to which the morphology of that taxon is similar in both samples. The approach described here can be used to soften any of the most popular ecological similarity indices described earlier—we chose the Jaccard "family" of indices for simplicity.

### Background: Phenotypic Similarity

To soften the Jaccard and Chao-Jaccard indices such that they include additional information about the resemblance of two samples, we first need a separate measure that can be used to establish the morphological similarity of a given species in two different samples. For example, if we know that an arbitrary species $x$ exists in sample $A$ and $B$ but its morphology varies (for example, $x$ could be twice as large in $B$), incorporating that information using the soft similarity approach described necessitates an appropriate measure of morphological similarity. Assume individual taxa can be represented as sets $C$ and $D$ consisting of $k$ morphological measurements ($C = \{c_1, c_2, \ldots c_k\}$ and $D = \{d_1, d_2, \ldots d_k\}$), where $c_k$ and $d_k$ are the mean values of trait $k$ for that species in that sample. Here we are primarily concerned with metric morphological measurements (e.g., body size/volume) and thus primarily use the term "morphology" when describing the set of measurements used for each taxon, but the $k$ measurements could be any sort of phenotypic measurement (fractional trophic level, functional ecospace characters [sensu Novack-Gottshall 2007], etc.). Because we are allowing for mixed data types when describing the morphology/phenotype of a given species, a measure that can handle

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

different types of data is needed. Gower (1971) generalized Jaccard's concept of set-based similarity to include the ability to compare collections of objects/traits in which the types of elements in a set could vary. Given that phenotypic measurements may vary in scale and in the type of measurement (e.g., metric, binary, ordinal), we utilize the Gower index as an approach to quantifying phenotypic similarity. Gower (1971; see also Podani 1999) proposed a suite of sub-indices for comparing different types of characters/elements of a given set (e.g., binary, nominal, numeric, and ordinal values), which would be summed to produce the final similarity index. Given this, we can calculate similarity as:

$$Gower(C, D) = \frac{|C \tilde{\cap} D|}{|C_\delta \cap D_\delta|}. \quad (1)$$

Modifying the general argument described in Gower (1971), the similarity of species $C$ and $D$ is calculated using a membership function that assigns each element $c_k$ a similarity value to its counterpart $d_k$, based on the class of element being compared:

$$|C \tilde{\cap} D| = \sum_{k=1}^{|C|} \mu(c_k, C \tilde{\cap} D) \text{ for all } c_k \in C. \quad (2)$$

In the original formulation of the Gower index, the membership of a given character $c_k$ in the intersection of $C$ and $D$ was calculated using three base cases:

$$\mu(c_k, C \tilde{\cap} D) = \begin{cases} \mu_{Interval}(c_k, C \tilde{\cap} D) \text{ if } c_k \text{ is on the interval or ratio scale} \\ \mu_{Binary}(c_k, C \tilde{\cap} D) \text{ if } c_k \text{ is binary} \\ \mu_{Nominal}(c_k, C \tilde{\cap} D) \text{ if } c_k \text{ is nominal} \end{cases}. \quad (3)$$

In the case of numeric data (e.g., intervals, ratios, ordinal characters, real numbers), the similarity between two traits $c_k$ and $d_k$ was calculated as:

$$\mu_{Interval}(c_k, C \tilde{\cap} D) = 1 - \frac{||c_k - d_k||}{R(k)}, \quad (4)$$

where $R(k)$ is the range of values observed for trait $k$ of the set. Note that there are important features of the Gower index that must be accounted for if it is used as a measurement of phenotypic similarity. The first is the problem of interspecific variation. Most existing implementations of the Gower index use the range of all values for a given trait to calculate the similarity of any two individual trait values. For example, if all species being compared were measured in terms of maximum linear dimension (as a proxy for body size), then the proximity of any two body-size measurements *for the same species* is taken in relation to the range of body sizes *for all species*. What this means is that the relative proximity of any two measurements is driven by the absolute difference between the smallest and largest species in the two samples. Consider, for example, the proximity of two arbitrarily chosen mean body-size measurements ($c_k = 1.2$ mm and $d_k = 2.1$ mm). If, as stated above, the value of trait $k$ is the mean maximum linear dimension for that species, consider the following two cases. In case 1, the smallest maximum linear dimension observed in the pooled assemblage is $min_k = 1$ mm, while the largest maximum linear dimension observed in the pooled assemblage is $max_k = 5$ mm. Here we scale the values between zero and one by dividing by the maximum value. In this case, the proximity of $c_k$ and $d_k$ is 0.78 $(1 - \frac{||0.24-0.42||}{0.8})$. Now consider case 2, where the largest maximum linear dimension value in the pooled assemblage is 10 times greater ($max_k = 50$ mm). In this case, the proximity of $c_k$ and $d_k$ is 0.98 $(1 - \frac{||0.024-0.042||}{0.98})$. This general pattern shows how the relative proximity (and contribution to the final similarity value) of any two measurements increases with the total range of values and decreases as the range decreases. This makes interpreting the proximity of any two values difficult—is

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

the range of interspecies body sizes in each sample a meaningful space to compare the proximity of intraspecies body sizes? We suggest that when quantifying the similarity of a given species to itself in another sample, it is not, and is particularly problematic when interspecies measurements differ by orders of magnitude. For example, if a species of zooplankton is twice as large in one sample, but the proximity of these two measurements is compared on a scale that incorporates body-size measurements for something several orders of magnitude larger (e.g., a fish), the phenotype of the zooplankton will be treated as if it was essentially exactly the same (even though it doubled/halved, depending on your frame of reference). As such, we suggest using the range of values observed *for that species* to compare the proximity of mean values for that species. In these cases, the absolute difference between the mean values for a given trait for each species in each sample is divided by the absolute difference between the minimum and maximum values observed for that species in the pooled assemblage. Note that this suggestion is made in relation to calculating the new indices described here only—in other cases (e.g., measuring functional diversity; Leinster and Cobbold 2012), the proximity of measurements should be made on the interspecific scale.

In addition to this problem of scaling, there exists an issue related to handling binary/logical/nominal characters when they vary in their expression between individuals within a given sample (e.g., a feature like a marking could be present on some individuals but not others). While the notion of a mean value for a metric phenotypic character (that is, one that is expressed as a real number/integer) is readily understood as the usual meaning of the term, it is less clear how we should handle binary/logical/nominal traits. In the standard Gower index, conventionally used to measure the dissimilarity between two phenotypes where the trait has a fixed value of either true/false or present/absent, the case for binary variables is:

$$\mu_{Binary}(c_k, C \tilde{\cap} D) = \begin{cases} 1 \text{ if } c_k = d_k = 1 \\ 0 \text{ if } c_k = 1 \text{ and } d_k = 0 \\ 0 \text{ if } c_k = 0 \text{ and } d_k = 1 \\ 0 \text{ if } c_k = d_k = 0 \end{cases}. \quad (5)$$

Finally, the case for nominal variables is:

$$\mu_{Nominal}(c_k, C \tilde{\cap} D) = \begin{cases} 1 \text{ if } c_k = d_k \\ 0 \text{ if } c_k \neq d_k \end{cases}. \quad (6)$$

When used to describe a given species, Gower's approach to binary/nominal characters assumes they are a fixed component of that species, found on all individuals. When comparing the mean phenotype of a species in two different samples, we suggest converting the binary/logical/nominal characters into frequencies (the probability that a randomly chosen *individual* of a given species in one sample has the feature) and comparing the frequencies using the case for interval data described above. Start by recording the presence/absence of that binary trait *for each individual*. Next, divide the total number of positive cases ("present" values) by the total number of individuals of that species, to attain a frequency for the binary trait. Next, calculate the similarity of the frequency of that trait in each sample using the $\mu_{Interval}(c_k, C \tilde{\cap} D)$ formula described above. In these cases, if the frequency of the binary trait is equal in both samples, it contributes a value of 1. If the frequency of the binary trait is 1 in one sample and 0 in another, the trait contributes a value of 0. In all other cases, it contributes a value between 0 and 1 equal to the similarity of the frequency that the trait occurs at in the two samples. In a practical sense, binary/logical/unordered nominal characters should be converted into columns of 1s and 0s, where 1 indicates the character is present *on that individual* (i.e., it is a 1 for a binary character, TRUE for a logical character, and present for a nominal character). If we consider only the values assigned to individuals of a single species, the mean value of these columns will be the frequency of true positives (the probability that a randomly selected individual of that species has the trait). The included R code (see Supplementary Material and the Appendix) automatically calculates this modified version of the Gower similarity based on the type of data and can handle binary, logical, and nominal traits that vary between individual specimens.

## Soft Jaccard Similarity

The classic Jaccard index uses a *strict cardinality* when counting the number of elements in

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

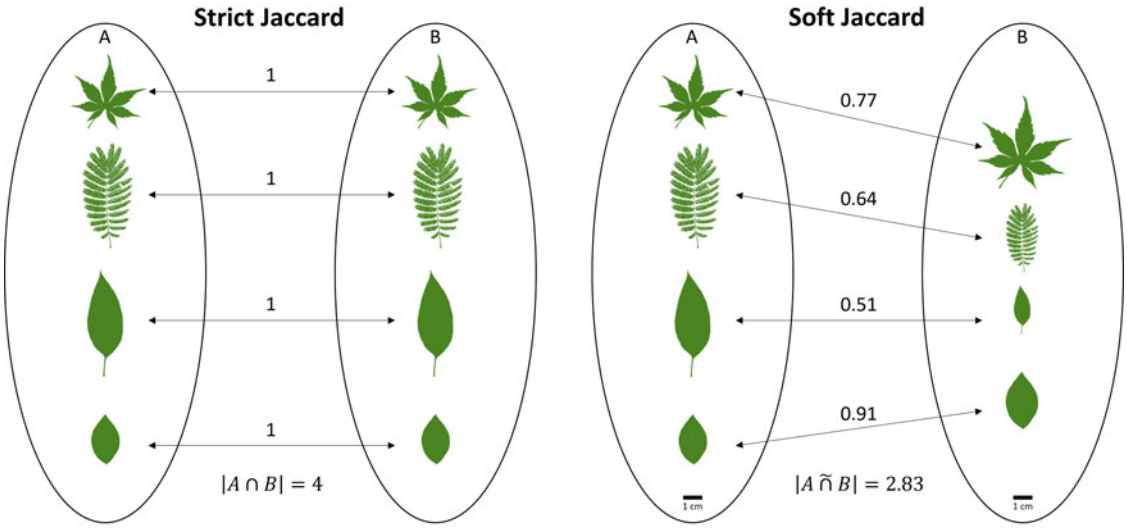



FIGURE 1. Image showing how the cardinality of the intersection is calculated in the strict Jaccard index and the soft Jaccard index. Arrows and their associated values indicate the contribution of that species to the cardinality of the intersection. In the strict Jaccard index (left), shared species contribute a value of 1, regardless of any degree of morphological/phenotypic variation that might exist. In the soft Jaccard index (right), shared species contribute a value on the interval [0,1], depending on their morphological similarity. In this way, the soft Jaccard index incorporates more information than its strict counterpart. $|A \cap B|$ indicates the strict cardinality of the intersection, while $|A \tilde{\cap} B|$ indicates the soft cardinality of the intersection. Values chosen for the soft Jaccard example are arbitrary and used for illustration purposes only.

the intersection—the cardinality of the intersection increases by 1 each time a shared species is identified (Jimenez et al. 2016) (Fig. 1). The soft Jaccard index proposed here uses a *soft cardinality*, wherein shared species increase the cardinality of the intersection by a value equal to their Gower similarity—if the phenotype of an arbitrary shared species pair $x$ is identical in the two samples, they contribute a value of 1 (as in the strict cardinality) (Fig. 1). If their phenotype shows no overlap at all, they contribute a value of 0. When their phenotype is *similar* in the two samples, they contribute a value equal to their Gower similarity, which is expressed on the interval [0,1] (Fig. 1). Note that we are comparing the phenotype of a given species in one sample to the phenotype *of the same species* in another sample—we are not computing pairwise dissimilarities between all taxa in a sample as is frequently done in measurements of trait/functional diversity (see Leinster and Cobbold [2012] for an example of the latter type of calculation). Repeating this process for every species that is found in both samples, we can derive a Jaccard-type index that weights incidence by morphological/phenotypic similarity. Imagine two or more samples taken from communities where $n$ species can be described in terms of metrics associated with morphology/phenotype (height and thickness of sponges/corals, leaf area of plants, body volume, diet proportions, etc.). Begin with two sets of $m$ shared species with associated phenotypic measurements $F = \{C_1, C_2, \ldots C_m\}$ and $G = \{D_1, D_2, \ldots D_m\}$. Assume each shared species (denoted $C_m$ in $F$ and $D_m$ in $G$) has a set of $k$ phenotypic measurements associated with it (that is, $C_m = \{c_1, c_2, \ldots c_k\}$ and $D_m = \{d_1, d_2, \ldots d_k\}$), where each measurement $c_k$ and $d_k$ can be expressed as above. Note that the characters used to measure one species *do not* need to be the same as the characters used to measure another. Our proposed index takes the form of the original Jaccard index, with the numerator replaced by a value equal to the soft cardinality of the intersection of sets $F$ and $G$:

$$Soft\ Jaccard(F, G) = \frac{|F \tilde{\cap} G|}{|F \tilde{\cup} G|}. \quad (7)$$

The soft cardinality of the intersection of $F$ and $G$ is calculated using a membership function that weights the incidence of a given species $m$ by

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

its phenotypic dissimilarity between the two sites (that is, it softens the definition of "shared"):

$$|F \tilde{\cap} G| = \sum_{m=1}^{|F|} \mu(C_m, F \tilde{\cap} G) \text{ for all } C_m \in F. \tag{8}$$

The membership of each shared species $C_m$ in $F$ in the soft intersection of $F$ and $G$ ($F \tilde{\cap} G$) is a value equal to the phenotypic similarity of that species when compared with its counterpart in $G$ ($D_m$):

$$\mu\left(C_m, F \tilde{\cap} G\right) = \frac{\left|C_m \tilde{\cap} D_m\right|}{\left|C_{m_\delta} \cap D_{m_\delta}\right|}, \tag{9}$$

where phenotypic similarity is the Gower similarity of the two taxa, calculated as above. We divide this value by the cardinality of the soft union of $F$ and $G$ ($F \tilde{\cup} G$), calculated as the number of species in $F$ and $G$ minus the species shared by the two samples, where the definition of shared is softened as above:

$$|F \tilde{\cup} G| = |F| + |G| - |F \tilde{\cap} G|. \tag{10}$$

Note that if the phenotype/morphology of all shared species is the same in both samples, the index reduces to the strict Jaccard index.

*Example.*—Assume $X = \{x_1, x_2, x_3\}$ and $Y = \{y_1, y_2, y_3\}$ are communities composed of the same three species (that is, $x_1 = y_1$, $x_2 = y_2$, and $x_3 = y_3$, such that $X = Y$). By definition $|X \cap Y| = 3$. We can represent the intersection of $X$ and $Y$ as an identity matrix $I$ (wherein the diagonal indicates it is TRUE that species 1 is present in $X$ and $Y$, etc.):

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{11}$$

such that $\sum(I) = |X \cap Y|$. When calculating the strict Jaccard similarity of $X$ and $Y$, we take the cardinality of the intersection and divide it by the cardinality of the union ($|X \cup Y| = |X| + |Y| - |X \cap Y|$), such that $Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{3}{3} = 1$. To calculate the soft Jaccard similarity of $X$ and $Y$, we replace the values in the diagonal of the identity matrix $I$ with values equal to the phenotypic similarity of that species as quantified between $X$ and $Y$ (calculated as $Gower(x_i, y_i)$), producing a special diagonal matrix we are calling a *soft identity matrix* $\tilde{I}$:

$$\tilde{I} = \begin{bmatrix} Gower(x_1, y_1) & 0 & 0 \\ 0 & Gower(x_2, y_2) & 0 \\ 0 & 0 & Gower(x_3, y_3) \end{bmatrix}. \tag{12}$$

For example, if $Gower(x_1, y_1) = 0.97$, $Gower(x_2, y_2) = 0.33$, and $Gower(x_3, y_3) = 0.72$, the soft identity matrix is:

$$\tilde{I} = \begin{bmatrix} 0.97 & 0 & 0 \\ 0 & 0.33 & 0 \\ 0 & 0 & 0.72 \end{bmatrix}. \tag{13}$$

We can calculate the soft cardinality of the intersection of $X$ and $Y$ as $\sum(\tilde{I}) = |X \tilde{\cap} Y|$. Following from this, the soft cardinality of the union of $X$ and $Y$ is $|X \tilde{\cup} Y| = |X| + |Y| - |X \tilde{\cap} Y|$. Finally, the soft Jaccard similarity of $X$ and $Y$ is $Soft\ Jaccard(X, Y) = \frac{|X \tilde{\cap} Y|}{|X \tilde{\cup} Y|} = \frac{2.02}{3.98} = 0.51$. Note that the use of the Gower index to calculate the similarity of shared elements is only one way to produce a soft identity matrix (chosen here for its applicability to questions of phenotypic similarity). A generalized form is:

$$\tilde{I} = \begin{bmatrix} S(x_1, y_1) & 0 & 0 \\ 0 & S(x_2, y_2) & 0 \\ 0 & 0 & S(x_3, y_3) \end{bmatrix}, \tag{14}$$

where $S(x_i, y_i)$ can be any similarity index expressed on the interval [0,1] (e.g., see Molodtsov 1999; Chum et al. 2008, Jimenez et al. 2016; Chauvin et al. 2020).

## Soft Chao-Jaccard Similarity

Chao et al. (2005, 2006) noted that classic incidence-based indices like Jaccard and Sørenson-Dice are sensitive to sample size and the presence of rare species. Because of this, rare species and common species are treated equally, despite sound a priori reasons to compare samples in terms of individuals, rather than species/clades (Magurran

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

2004). This lack of appropriate feature weighting can lead the classic Jaccard index to be downward biased when comparing samples taken from populations with rare species (Wolda 1981). Chao et al.'s (2005) index circumvents this issue by working with sets $U = \{u_1, u_2, \ldots u_n\}$ and $V = \{v_1, v_2, \ldots v_n\}$ of relative abundances $u_n$ and $v_n$ of $n$ shared species. Here we describe a method for producing a soft Chao-Jaccard index, wherein variation in abundance and intraspecific morphology is taken into account when comparing samples. As above, we imagine two or more samples taken from communities where $n$ species can be described in terms of metrics associated with phenotype. If each species in each sample also has an associated abundance value, we can derive a soft version of the Chao-Jaccard similarity index. Given two sets $F$ and $G$ of $m$ shared species with associated phenotypic measurements $F = \{C_1, C_2, \ldots C_m\}$ and $G = \{D_1, D_2, \ldots D_m\}$ and sample relative abundances $U = \{u_1, u_2, \ldots u_m\}$ and $V = \{v_1, v_2, \ldots v_m\}$, with a set $W$ of corresponding phenotypic similarity values for each shared species:

$$W = \{Gower(C_1, D_1), Gower(C_2, D_2), \ldots Gower(C_m, D_m)\}, \quad (15)$$

We can derive a soft version of the Chao-Jaccard index. Note that $U$ and $V$ are sets of relative abundances $u_n$ and $v_n$ of $n$ *shared* species. Specifically, if $P$ and $Q$ are sets of absolute abundance values (specimen counts) associated with the $n$ species found in $U$ and $V$, respectively ($P = \{p_1, p_2, \ldots, p_n\}$, $Q = \{q_1, q_2, \ldots, q_n\}$), and $a_n$ and $b_n$ are presence/absence values associated with species $n$ (such that $a_n$ indicates whether species $n$ is present in $P$ [$p_n > 0$] and $b_n$ indicates whether that species is found in $Q$ [$q_n > 0$]), then $u_n = \frac{p_n}{\sum(P)}(a_n \times b_n)$. That is, $u_n$ is equal to the relative abundance of species $n$ in sample $P$ if species $n$ is a shared species, and equal to 0 if species $n$ is not a shared species. The values included in set $V$ are calculated in the same way with appropriate substitutions. To soften the original Chao-Jaccard index, we first weight the probability of selecting an individual representing a shared species from each sample using the phenotypic similarity of that species to its counterpart in the other sample. In doing this we are producing a soft cardinality of the sets $U$ and $V$, denoted $\tilde{U}$ and $\tilde{V}$ and calculated as $\tilde{U} = U \times W$ and $\tilde{V} = V \times W$ respectively. From this, we insert the weighted relative abundances into the original Chao-Jaccard index to produce a measure of similarity sensitive to differences in abundance and mean phenotype of shared species:

$$Soft\ Chao - Jaccard(U, V) = \frac{\sum(\tilde{U}) \times \sum(\tilde{V})}{\sum(\tilde{U}) + \sum(\tilde{V}) - \left(\sum(\tilde{U}) \times \sum(\tilde{V})\right)}. \quad (16)$$

While the soft Jaccard index uses a soft cardinality of the intersection, the soft Chao-Jaccard index uses a soft cardinality of each individual sample. In essence, when counting the number of individuals of a given species in a given sample, we take into account the extent to which that species has a similar morphology in both samples. As above, this comparison is only taking into account the phenotype of the same species in the other samples. If all species have a maximum pairwise similarity of 1, the index reduces to the standard Chao-Jaccard index.

*Example.*—As above, assume $X = \{x_1, x_2, x_3\}$ and $Y = \{y_1, y_2, y_3\}$ are communities composed of the same three species (that is, $x_1 = y_1$, $x_2 = y_2$, and $x_3 = y_3$ such that $X = Y$). Here we assign each species a relative abundance for that sample, and, as above, use $U$ and $V$ to indicate sets of relative abundances $u_i$ and $v_i$. As we did with the intersection, we can represent $U$ and $V$ as diagonal matrices $J$ and $K$ (here the diagonal indicates the *relative abundance* of species 1 in community $X$, etc.). We use $J$ and $K$ to indicate diagonal matrices referring to individual sets, to differentiate them from the symbol $I$ used above to represent the identity matrix for the intersection. For example:

$$J = \begin{bmatrix} u_1 & 0 & 0 \\ 0 & u_2 & 0 \\ 0 & 0 & u_3 \end{bmatrix} \text{ and } K = \begin{bmatrix} v_1 & 0 & 0 \\ 0 & v_2 & 0 \\ 0 & 0 & v_3 \end{bmatrix}. \quad (17)$$

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

Note that as stated in Chao et al. (2005) and described above, $U$ and $V$ include the relative abundances of *shared* species only—thus, there are no zeros on the diagonal of $J$ and $K$. For sake of demonstration, assume $u_1 = v_1 = 0.2$, $u_2 = v_2 = 0.2$, and $u_3 = v_3 = 0.6$:

$$J = K = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.6 \end{bmatrix}. \quad (18)$$

As was done above when calculating the soft Jaccard similarity, we first calculate a soft identity matrix for the intersection of $X$ and $Y$ (restating $W$ as a diagonal matrix):

$$\tilde{I} = \begin{bmatrix} Gower(x_1, y_1) & 0 & 0 \\ 0 & Gower(x_2, y_2) & 0 \\ 0 & 0 & Gower(x_3, y_3) \end{bmatrix}. \quad (19)$$

For sake of demonstration, assume, as above, that $Gower(x_1,\ y_1) = 0.97$, $Gower(x_2,\ y_2) = 0.33$, and $Gower(x_3,\ y_3) = 0.72$, such that:

$$\tilde{I} = \begin{bmatrix} 0.97 & 0 & 0 \\ 0 & 0.33 & 0 \\ 0 & 0 & 0.72 \end{bmatrix}. \quad (20)$$

We calculate the soft cardinality of $U$ ($|\tilde{U}|$) and $V$ ($|\tilde{V}|$) as follows:

$$\begin{aligned} \tilde{U} &= \tilde{I} \times J \\ &= \begin{bmatrix} 0.97 & 0 & 0 \\ 0 & 0.33 & 0 \\ 0 & 0 & 0.72 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.6 \end{bmatrix} \\ &= \begin{bmatrix} 0.194 & 0 & 0 \\ 0 & 0.066 & 0 \\ 0 & 0 & 0.432 \end{bmatrix}, \end{aligned} \quad (21)$$

given that $J = K$, $\tilde{U} = \tilde{V}$. As above, $|\tilde{U}| = \sum(\tilde{U})$ and $|\tilde{V}| = \sum(\tilde{V})$. Following from this, the soft Chao-Jaccard similarity of $X$ and $Y$ is $Soft\ Chao\text{-}Jaccard\ (U, V) = \frac{\sum(\tilde{U}) \times \sum(\tilde{V})}{\sum(\tilde{U}) + \sum(\tilde{V}) - \left(\sum(\tilde{U}) \times \sum(\tilde{V})\right)} = \frac{0.479}{0.905} = 0.53$.

The soft Chao-Jaccard similarity of these two sets is slightly higher than the soft Jaccard similarity, because the most dissimilar species (species 2) is less abundant, and therefore contributes less to the final value.

## Description of Datasets

To explore the new approach, we compared principal coordinates analysis (PCoA) outputs based on the original Chao-Jaccard index (Chao et al. 2005) with ones based on the soft Chao-Jaccard index (proposed here). Note that although we are choosing the Chao-Jaccard index for comparative purposes, we must emphasize that increasing the information provided to *any* index improves a researcher's ability to recognize short gradients using ordination techniques and does not reflect issues with the original index. Random samples were compared from two settings—a simulated river velocity gradient wherein random taxa are affected by flow velocity in different ways (described here), and a real-world example wherein a river velocity gradient was sampled at five different points (original data described in Puijalon and Bornette [2004]).

*Simulated Data: River Velocity Gradient.*—To generate artificial communities simulating the effects of a short gradient, we used a modification of the approach described in Wolda (1981), wherein a hypothetical fauna is generated using a log-series distribution:

$$M = \alpha \ln\left(1 + \frac{P}{\alpha}\right). \quad (22)$$

Here, $M$ is the number of species, $P$ is the number of individuals in the population, and $\alpha$ is a diversity coefficient (Fig. 2). We used an $\alpha$ value of 5 with a $P$ of 100,000, corresponding to 50 species (Fig. 2). The relative abundance of each of the 50 species was generated randomly (common and rare species are distributed randomly within the 50 species). To determine the relative abundance of each species, we generated 50 random integers between 1 and 1000; these integers were divided by the sum of the random values, and the resulting string of 50 values was used as a probability vector for randomly selecting the species label a given individual would receive. Next,

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

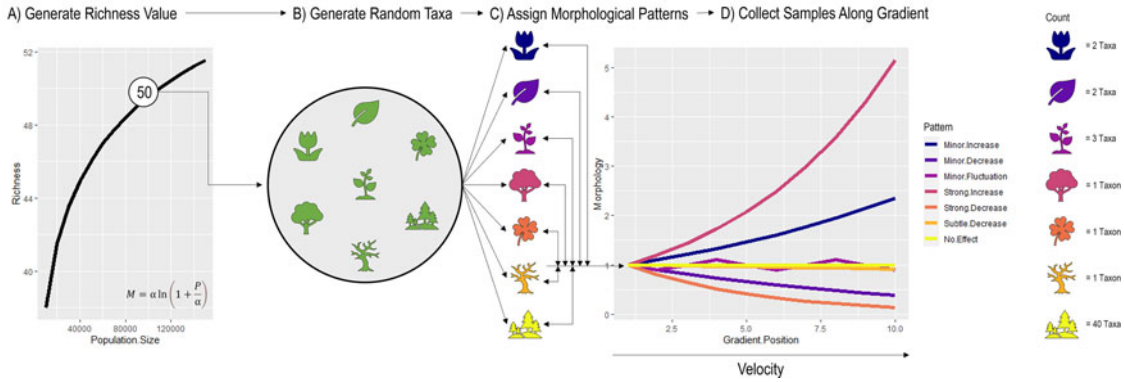



FIGURE 2. Flow diagram illustrating how the hypothetical communities were generated. A, Richness value (here 50 species) is determined based on the size of the underlying population ($P$), a diversity value ($\alpha$), and a formula describing a log-series distribution. B, Taxa are randomly generated and assigned random species labels, with some labels being assigned less often ("rare species"). C, A subset of the taxa ($n = 10$) is randomly chosen to display one of six morphological patterns along the 10 positions of the gradient, while the remaining taxa ($n = 40$) show no morphological change along the gradient. D, Random samples are pulled from the resulting 10 communities and compared using the proposed indices.

individual specimens were assigned a randomly generated set of morphological measurements (10 measurements in total)—measurements were generated using R's built-in *rnorm* function with a mean of 1 and a standard deviation of 3. From this initial community, we generated 9 more communities (representing different points in a hypothetical flow gradient) from which we drew random samples. Each of the 9 new communities was identical to the original, except for the morphological measurements of 10 taxa, selected at random to show one of six morphological trends across the idealized flow gradient (Fig. 2):

1. *Minor increase*: Two taxa were randomly selected to show a minor increase in measured values along the gradient. Specifically, the mean values for these species increase by 10% between each community.
2. *Minor decrease*: Two taxa were randomly selected to show a minor decrease in measured values along the gradient. Specifically, the mean values for these species decrease by 10% between each community.
3. *Minor fluctuation*: Three taxa were randomly selected to show a minor fluctuation in measured values along the gradient. Namely, two were chosen to randomly decrease by 10%, return to the original values, increase by 10%, return to the original values, etc. One taxon was chosen to show the opposite pattern (begin by increasing, etc.).
4. *Strong increase*: One taxon was chosen to randomly show a strong increase in measured values along the gradient. Specifically, the mean values for this species increase by 20% between each community.
5. *Strong decrease*: One taxon was chosen to randomly show a strong decrease in measured values along the gradient. Specifically, the mean values for this species decrease by 20% between each community.
6. *Subtle decrease*: One taxon was chosen to randomly show a very subtle decrease in measured values along the gradient. Specifically, the mean values for this species decrease by 1% between each community.

The remaining 40 taxa were chosen to show no pattern of variation along the gradient. Because the distribution used to generate the random values had a standard deviation of 3, some degree of morphological overlap between communities with close mean values is expected. See R code in the Supplementary Material for more information. The decision to have visible phenotypic effects show up randomly in only a small number of taxa—with a subset showing nonlinear variation—was intended to produce a situation in which randomly choosing a specific taxon that tracks the gradient would be difficult,

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

necessitating the use of an index that works with the entire sample. Furthermore, these underlying populations were designed such that an ordination based on a classic similarity measure (e.g., Jaccard 1901; Bray and Curtis 1957) would be unable to reproduce the gradient, as it is contained in the morphology only (Fig. 3).

*Real-World Data: Rhône River Velocity Gradient.*—The simulated data represent an idealized case and may not exactly reflect how a short gradient appears in reality. Given this, we tested whether our new index could produce an accurate ordination of random samples taken from a natural flow-velocity gradient (using data described in Puijalon and Bornette [2004]). Puijalon and Bornette (2004) explored the relationship between morphology and flow velocity by investigating samples of two distantly related freshwater plant species (water mint [*Mentha aquatica*] and lesser water-parsnip [*Berula erecta*]) collected from different points along the Rhône River (Germany). Their results suggested that only *M. aquatica* shows the kind of clear linear change with increasing velocity that one might expect, while *B. erecta* shows a nonlinear shift in morphology across the gradient. The data presented in Puijalon and Bornette (2004) represent a sort of "worst-case scenario" for recognizing a short gradient from specimens alone—incidence and abundance are the same at all five sampling locations, and the two taxa show different reactions to the underlying gradient. Subjectively deciding to use either *B. erecta* or *M. aquatica* as an indicator species for ordering the samples along the gradient will give different results, while using a classic measure of similarity (e.g., Jaccard, Horn-Morisita) will give the impression that the samples all came from the same location. Note that we excluded the mass-based characters Puijalon and Bornette (2004) used, as they have no obvious analogue when applied to fossil material.

## Results

### Reconstructing Short Gradients Using the Soft Chao-Jaccard Index: Simulated Data

In Figure 4A, when the similarity of the random samples was calculated using the strict Chao-Jaccard index, the underlying gradient was missed (this is the same problem seen in Fig. 3). As can be seen in Figure 4B, when the contribution of shared taxa to the similarity of two samples is weighted by the morphological similarity and relative abundance of the shared taxa, the underlying gradient becomes readily apparent. Specifically, as can be seen in Figure 4B, coordinate 1 captures much of the underlying velocity gradient, and the grouping of random samples in the coordinate space reflects their position along the hypothetical river. Note that the gradient can be recovered even at relatively small samples sizes (e.g., when $N = 200$), but the separation between populations becomes more apparent with increasing sample size and/or repeated sampling (see section "Diversity, Sample Size, and Accuracy"). If the soft Jaccard or soft Chao-Jaccard indices are intended to be used with an unsupervised clustering algorithm (e.g., *k*-means [MacQueen 1967], *k*-medoids [Schubert and Rousseeuw 2019], or fuzzy *c*-means [Dunn 1973]), larger sample sizes are to be preferred, as they will increase the accuracy of any clustering/classification approach. With repeated sampling and/or increasing sample size, the mean values of the phenotypic traits being measured increasingly approach the true mean value for that population (this is a consequence of the law of large numbers; see Dekking et al. [2005] for an overview), resulting in a greater ability to differentiate between samples (see section "Diversity, Sample Size, and Accuracy"). Note in Figure 4B that the points are arranged in a parabolic/arch-shaped pattern—this is an example of the "arch"/"horseshoe" effect (also called the "Guttman effect"; Camiz 2005), which is a well-known consequence of plotting strong linear gradients using ordination techniques (Podani and Miklós 2002). Patterns like this are introduced during the production of an ordination plot from a dissimilarity matrix and will appear regardless of what dissimilarity index is used, provided a linear data structure is contained within the underlying matrix (Podani and Miklós 2002; Legendre and Legendre 2012). Given that there is no linear data structure in the matrix produced using the original Chao-Jaccard index (because the effects of the linear gradient

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

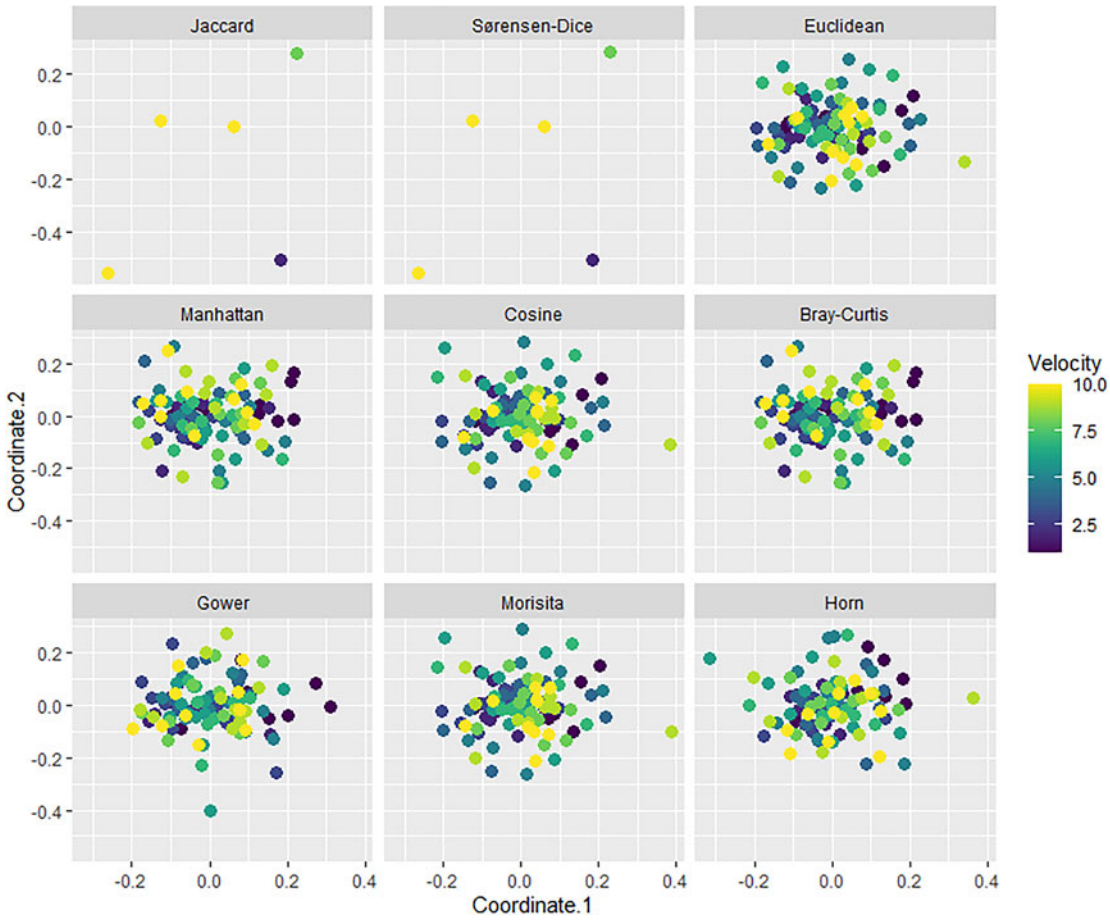



FIGURE 3. Ordination plots (principal coordinates analysis) showing the behavior of nine commonly used similarity indices when they are used to compare random samples pulled from the 10 artificial communities described in the text. Each point represents a random sample of 1000 individuals pulled from the corresponding velocity grade. Because the underlying gradient is contained in the morphology of taxa only, none of the common similarity indices produces a plot illustrating the gradient. Incidence-based indices (Jaccard and Sørensen-Dice) produce a plot with multiple samples coinciding on five points; because of their large size, most samples include all the same species (and thus coincide in the plot), with a few samples missing the rarest species in the population by chance alone. The remaining indices are all sensitive to the number of individuals and show random scattering around a central point driven by the random sampling process.

are only visible when morphology is incorporated), there is no horseshoe/arch effect in Figure 4A. Visual distortions of the kind seen in Figure 4B can only be removed by transforming the data or otherwise manipulating the results (Williamson 1978; Bradfield and Kenkel 1987; Dale 1994; Legendre and Gallagher 2001). It is important to stress that this pattern *is not* a consequence of using the new indices described here—the original Chao-Jaccard index will also produce a horseshoe/arch if the resulting dissimilarity matrix represents a linear data structure (as will any of the commonly used dissimilarity indices; see Williamson 1978; Bradfield and Kenkel 1987; Dale 1994; Legendre and Gallagher 2001; Podani and Miklós 2002; Legendre and Legendre 2012). In all prominent cases discussed in the literature (e.g., Swan 1970; Williamson 1978; Hill and Gauch 1980; Bradfield and Kenkel 1987; Dale 1994; Økland 1999; Legendre and Gallagher 2001; Podani and Miklós 2002; Legendre and Legendre 2012), the horseshoe/arch effect is associated with linear gradients and corresponding linear data structures. Raising the underlying dissimilarity matrix to a power of

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

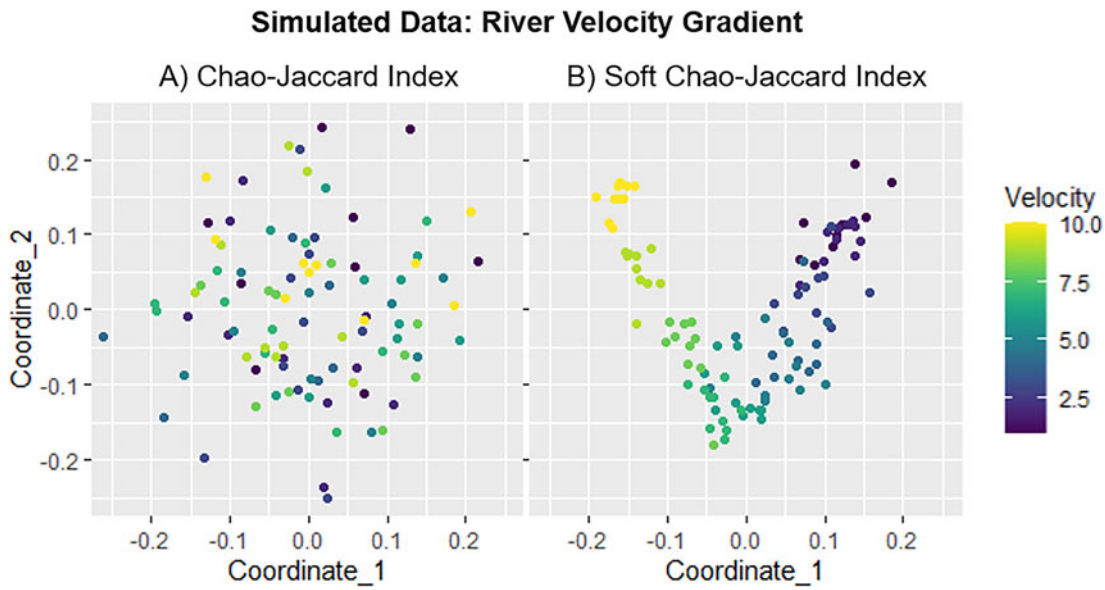



FIGURE 4. Ordination (principal coordinates analysis) of the 100 hypothetical samples described in the text, plotted based on (A) pairwise strict Chao-Jaccard similarity and (B) pairwise soft Chao-Jaccard similarity (presented in this article), wherein the samples are taken along a simulated flow-velocity gradient. Each point represents a random sample of 1000 individuals. When compared with the random distribution of points in A, the new index proposed here produces an accurate ordination of the underlying gradient (B). Note that the parabolic distribution of points in B is an example of the well-known "horseshoe"/"Guttman" effect, which is a mathematical artifact created when one uses an ordination technique to display any strong underlying gradient (Podani and Miklós 2002). It is *not* an effect of the new index proposed here, and techniques exist to remove this distortion if desired (Williamson 1978; Bradfield and Kenkel 1987; Dale 1994; Legendre and Gallagher 2001).

2 or greater can potentially decrease the impact of the horseshoe effect (Podani and Miklós 2002). A thorough discussion of the mathematics behind the horseshoe/arch effect is beyond the scope of this paper, and we refer interested readers to the reviews cited.

### Reconstructing Short Gradients using the Soft Chao-Jaccard Index: Modern Data

As described earlier, Puijalon and Bornette (2004) noted that only *Berula erecta* showed a significant linear morphological trend along the measured flow gradient. Conversely, *Mentha aquatica* showed nonlinear (unimodal) size variation along the gradient, with size increasing initially (such that the largest individuals are found in flow patch 3) before dropping off toward flow patch 5 (Puijalon and Bornette 2004). Despite this variation in the effect of the flow-velocity gradient, PCoA of the five samples for which the distance between points is based on soft Chao-Jaccard dissimilarity corresponds to the underlying gradient (Fig. 5B). When samples are compared using the original Chao-Jaccard index, the five samples are all equally similar and therefore coincide, appearing as a single point on the plot (Fig. 5A). As with the simulated data, the linear nature of the underlying flow gradient produces a horseshoe-shaped distribution of points when plotted (Fig. 5B). Given the congruence between the outputs seen when the index is applied to hypothetical and empirical data (Fig. 4 vs. Fig. 5), our results suggest that the soft Chao-Jaccard index can be used to order random samples along a subtle underlying gradient, even when the impacts of the gradient are nonlinear and distributed across multiple taxa.

## Discussion and Applications in Paleontology

Most short gradients described in the literature can best be described as linear—increasing flow velocity between points A and B, increasing wave energy between points A and B, and so on (Miles and Ricklefs 1984; Price 1991; Li et al. 2006; Einbinder et al. 2009; Soto et al.

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

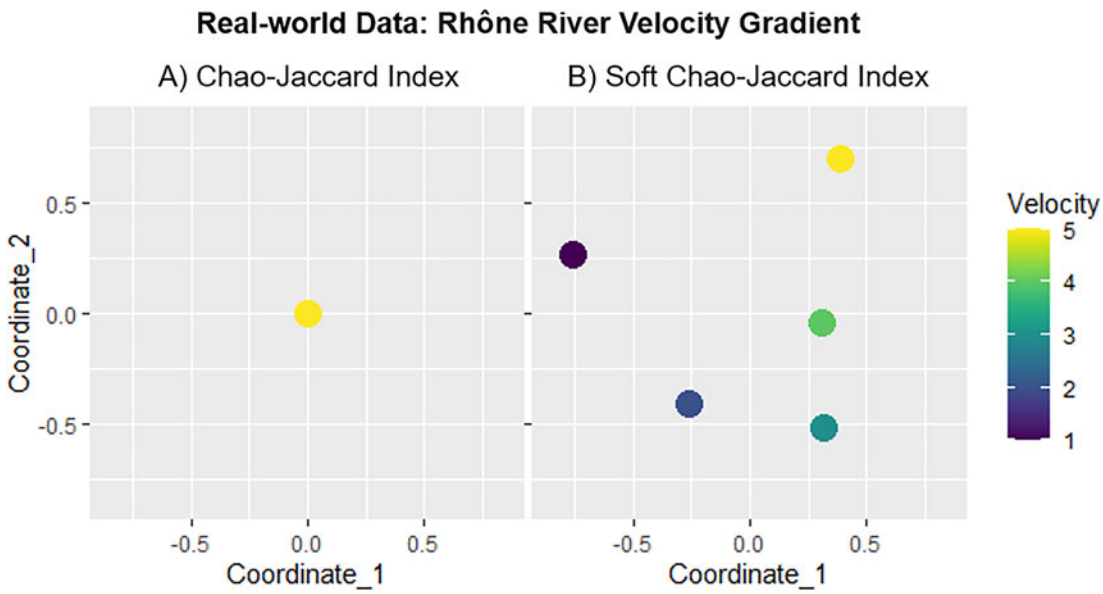



FIGURE 5. Ordination (principal coordinates analysis) of the five samples described in Puijalon and Bornette (2004), plotted based on (A) pairwise strict Chao-Jaccard similarity and (B) pairwise soft Chao-Jaccard similarity (presented in this article). The original article indicates that samples were taken along an increasing flow-velocity gradient at different points along the Rhône River in Germany. Each point represents a random sample of 60 individuals (30 each of *Mentha aquatica* and *Berula erecta*). When compared with the overlapping distribution of points in A, the new index proposed here produces an accurate ordination of the underlying gradient (B). Note that, as in Fig. 4, the parabolic distribution of points in B is an example of the "horseshoe" effect (Podani and Miklós 2002). Here, flow velocity ranges from 1 (0 m $s^{-1}$) to 5 (0.5 m $s^{-1}$).

2018; Kramer et al. 2020). While the underlying physical features of most short gradients may be linear, empirical evidence suggests that only some taxa will show a linear response (e.g., Puijalon and Bornette 2004; Fulton et al. 2005; Kramer et al. 2020). The index proposed here uses a consensus-based approach to sample similarity, and the results seen here suggest it can be used to recognize linear gradients even when some taxa show nonlinear responses (Figs. 4, 5). Because of this flexibility, the proposed index could be used to order and cluster samples for which the response of the taxa being modeled is unknown, such that the nature and effects of the underlying gradient do not need to be stated a priori. For example, the existence of short gradients controlling the distribution of enigmatic fossil taxa (e.g., Ediacaran fronds, archaeocyaths, sessile Burgess Shale–type taxa) could be tested for using this method (e.g., Clapham et al. 2003; Darroch et al. 2013; Nanglu et al. 2020). By collecting random samples (e.g., quadrats) and comparing an ordination of these samples based on a classic similarity index (e.g., the Chao-Jaccard index) to one based on a soft similarity index (e.g., the soft Chao-Jaccard index proposed here), the difference in the resulting plots can be used to explore whether a morphological gradient is present. By using an ordination technique such as nonmetric multidimensional scaling (NMDS), wherein environmental variables can be plotted as a biplot, it is possible to explore whether a soft similarity–based ordination better explains the correlation between the distribution of samples and the underlying environmental conditions than an ordination based on a strict similarity index. For example, when the numeric water velocity data from Puijalon and Bornette (2004) are included as a biplot in a NMDS plot showing the soft Chao-Jaccard similarity of the samples, the ordering of samples along the underlying velocity gradient becomes apparent (Fig. 6).

## Modeling Long Gradients

By virtue of how the index was designed, phenotypic similarity can never contribute more to the similarity of two samples than the presence/absence of a given taxon can. Because

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

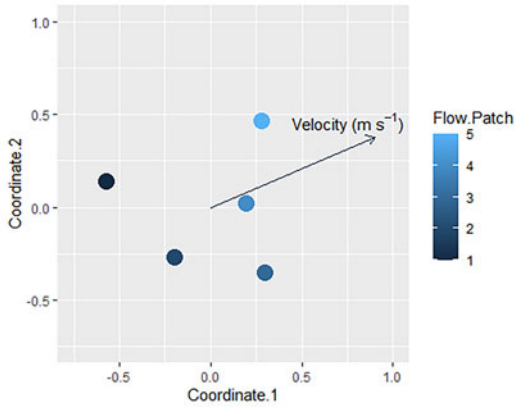



FIGURE 6. Nonmetric multidimensional scaling plot showing the correspondence between the ordering of samples (based on soft Chao-Jaccard similarity) and the underlying velocity gradient. The arrow indicates an increase in water velocity (from 0 m s$^{-1}$ to 0.5 m s$^{-1}$) along that axis (based on data in Puijalon and Bornette [2004]). The samples are ordered correctly along the underlying gradient, despite the presence of a horseshoe/arch.

of this, if the samples being modeled are better understood as coming from along a long gradient (one defined by extensive faunal turnover), the values returned by the index (and the corresponding ordination plot) will automatically reflect this.

## Diversity, Sample Size, and Accuracy

Our ability to recognize cases when samples are distributed along an underlying short gradient is directly related to the signal-to-noise ratio in the resulting ordination plot (the same statement applies to an unsupervised machine learning algorithm working with the data in the plot). The signal-to-noise ratio of a given ordination plot is in turn related to sample size, sampling effort, the diversity of the underlying population (the number of individuals and the number of species), and the proportion of the community showing a response to the

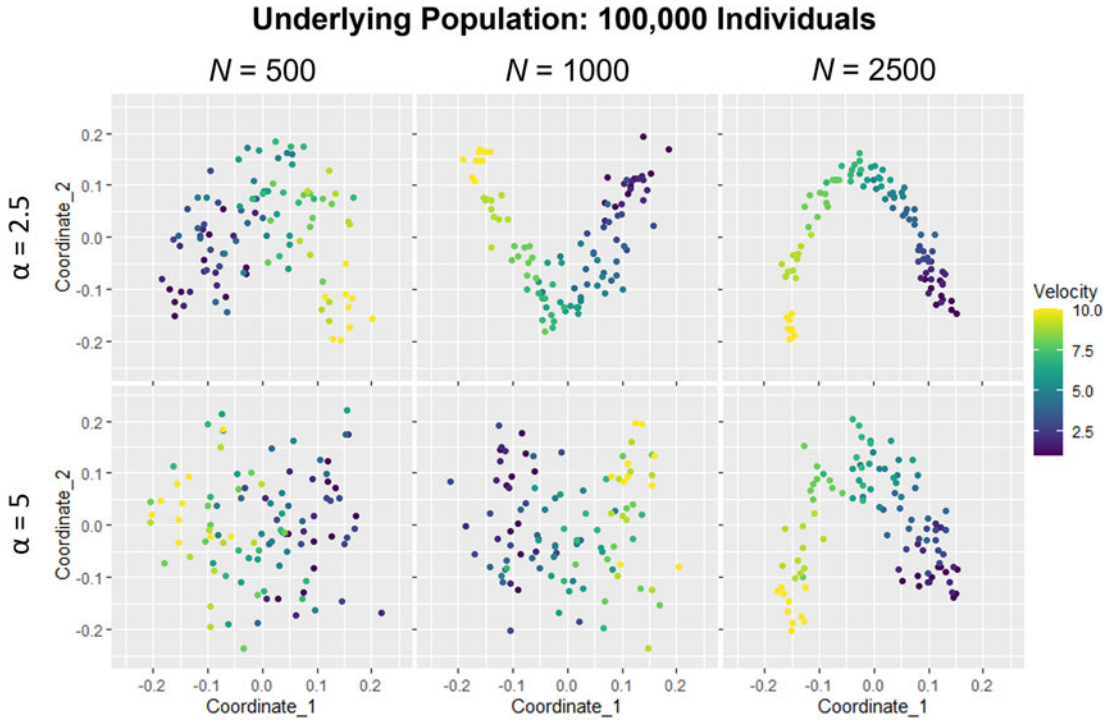



FIGURE 7. Six plots showing the relationship between sample size (where each point represents *N* random individuals) and degree of gradient recovery, when samples are drawn from a large underlying population (100,000 individuals) at two different levels of diversity ($\alpha = 2.5$ and $\alpha = 5$). When combined with the log distribution for generating communities described in the "Introduction," $\alpha = 2.5$ produces 26 species distributed randomly among 100,000 individuals (with each species having a different rarity), while $\alpha = 5$ produces 50 species. As can be seen when moving from left to right, larger sample sizes result in a more high-fidelity representation of the underlying gradient (a greater signal-to-noise ratio). While coordinate 1 always represents position along the gradient to some extent, recognizing the pattern is easier with larger sample sizes.

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

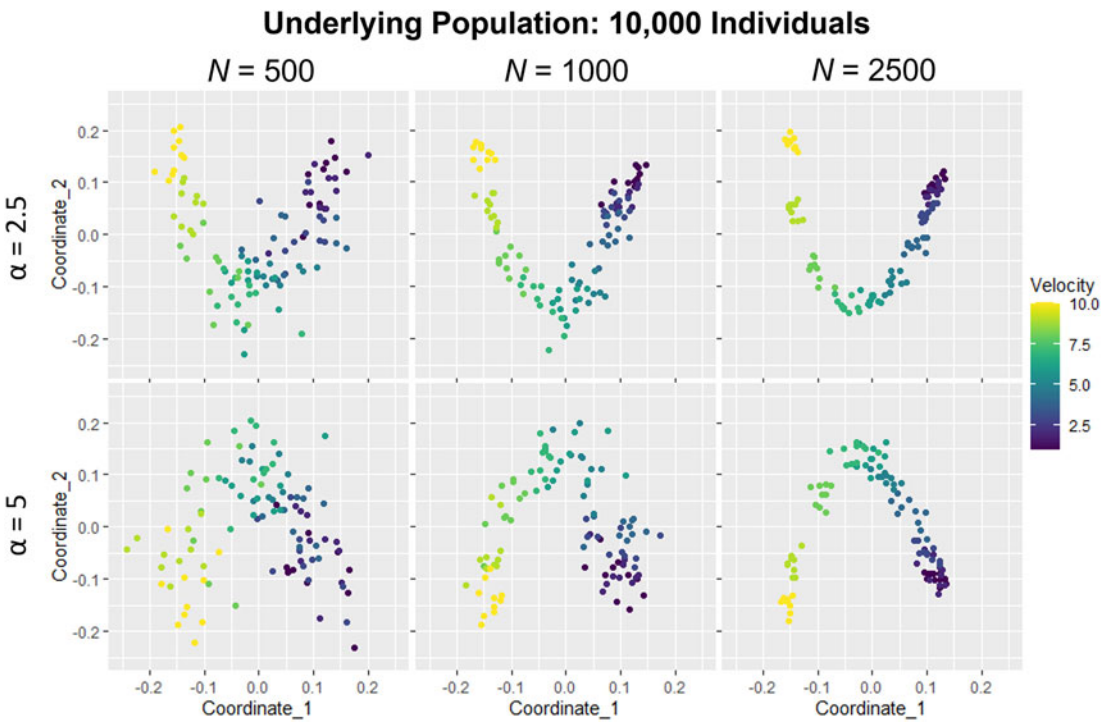



FIGURE 8. Six plots showing the relationship between sample size (where each point represents *N* random individuals) and degree of gradient recovery, when samples are drawn from a smaller underlying population than the one used in Fig. 7 (10,000 individuals) at two different levels of diversity ($\alpha = 2.5$ and $\alpha = 5$). When combined with the log distribution for generating communities described in the introduction, $\alpha = 2.5$ produces 21 species distributed randomly among 10,000 individuals (with each species having a different rarity), while $\alpha = 5$ produces 38 species. As was seen in Fig. 7, when moving from left to right, larger sample sizes result in a more high-fidelity representation of the underlying gradient. While coordinate 1 always represents position along the gradient to some extent, recognizing the pattern is easier with larger sample sizes. As can be seen when comparing Figs. 7 and 8, smaller/less diverse underlying populations result in a greater signal-to-noise ratio.

gradient. Recall that we chose 10 random taxa to show some degree of variation along the gradient, of which only 7 showed a linear response (half of which showed a positive correlation and half showed a negative correlation). In Figure 7, when $\alpha = 5$, only 14% of taxa in the underlying population contribute to our ability to reconstruct the gradient (the remaining 43 taxa either oscillate randomly along the gradient or show no change). Because of this, when sample sizes are small, the position of most samples in the space defined is driven by random fluctuations in shared taxa caused by sampling, making the underlying gradient harder to detect (Fig. 7). If we reduce the diversity in the populations, the proportion of taxa that contribute to our ability to reconstruct the gradient increases (that is, when $\alpha = 2.5$, 27% of taxa show a linear response to the gradient). Because of this, the signal-to-noise ratio is higher than when $\alpha = 5$, such that the underlying gradient is recognizable at smaller sample sizes (Fig. 7). This general pattern continues if we reduce the size of the underlying population (Fig. 8). When the population consists of 10,000 individuals, the gradient is recognizable even at small sample sizes when $\alpha = 5$. When the area being modeled is thought to have very high diversity, or a very small number of taxa are thought to possess sufficient phenotypic plasticity to track the gradient, large sample sizes may be needed to reconstruct the gradient. However, in simpler communities (e.g., the one modeled in Fig. 5), small sample sizes are sufficient to reproduce the gradient.

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

### Potential Issues in Paleontological Settings

As with the application of most numerical methods in paleontological settings, there are potential issues a user should be aware of. In particular, while the soft Chao-Jaccard index is sensitive to shifts in morphology and abundance and can order samples along a short gradient, it is up to the user to determine the ultimate cause of the observed pattern. Specifically, a taphonomically generated pattern (e.g., sorting of specimens in beach rock by wave action) or substantial time averaging could give the impression that individuals are sorted along an in vivo short gradient when modeled using the indices described here. While this may be useful in its own right (e.g., as a tool for describing the taxonomic component of sedimentary facies), it is important to note that a nonrandom distribution of taxa is not, in itself, a guarantee that a subtle underlying shift in the environment explains the pattern. Differentiating between taphonomic and ecological processes controlling the distribution of fossil taxa will have to be done on a case-by-case basis, taking into account the depositional environment being investigated.

## Conclusions

The soft similarity measures described here offer an effective means of ordering random samples along underlying gradients in which the impacts of the gradient may be nonlinear and distributed across multiple taxa. The accuracy of a given ordination based on these indices is dependent on the size and number of samples used, as well as the strength of the underlying gradient. The indices proposed here work backward from the strict similarity of two samples, exploring how morphological/phenotypic dissimilarity observed between pairs of shared species decreases the similarity of the two samples (i.e., they explore how the variable morphology/phenotype of a given species at two different sites can reduce the similarity of those sites). This is the opposite process from what is described for other soft similarity measures (e.g., Leinster and Cobbold 2012; Sidorov et al. 2014; Jimenez et al. 2016)—these indices work with nonmatching elements and explore how their similarity can allow them to be treated as partially matching elements. Our indices expressly work with shared elements—those in the strict cardinality of the intersection. The general approach described here can be used to soften any of the most popular ecological similarity indices (e.g., Sørensen-Dice, Horn-Morisita), allowing them to take phenotypic dissimilarity of shared species into account.

## Acknowledgments

The authors thank G. Brown and two anonymous reviewers for their helpful comments on this paper. The authors would also like to thank J. Socolar and N. Mason for their insightful reviews of a prior version of this paper. This project was supported by NSERC PGS-D scholarship PGSD3—504619—2017 awarded to D.G.D., as well as NSERC Discovery Grant RGPIN 435402 to M.L.

## Data Availability Statement

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.rr4xgxd8x.

## Literature Cited

Booth, R. 2001. Ecology of testate amoebae (Protozoa) in two Lake Superior coastal wetlands: implications for paleoecological environmental monitoring. Wetlands 21:564–576.

Bradfield, G., and N. Kenkel. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distances. Ecology 68:750–753.

Bray, J. R., and J. T. Curtis. 1957. An ordination of the upland forest communities in southern Wisconsin. Ecological Monographs 27:325–349.

Brett, C., A. Bartholomew, and G. Baird. 2007. Biofacies recurrence in the Middle Devonian of New York State: an example with implications for evolutionary paleoecology. Palaios 22:306–324.

Camiz, S. 2005. The Guttman effect: its interpretation and a new redressing method. Data Analysis Bulletin 5:7–34.

Chao, A., R. L. Chazdon, R. K. Colwell, and T-J. Shen. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters 8:148–159.

Chao, A., R. L. Chazdon, R. K. Colwell, and T-J. Shen. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62:361–371.

Chauvin, L., K. Kumar, C. Wachinger, M. Vangel, J. de Guise, C. Desrosiers, W. Wells, M. Toews, and Alzheimer's Disease Neuroimaging Initiative. 2020. Neuroimage signature from salient keypoints is highly specific to individuals and shared by close relatives. NeuroImage 204:116208.

Chum, O., J. Philbin, and A. Zisserman. 2008. Near duplicate image detection: min-Hash and tf-idf weighting. Pp. 50.1–50.10 *in* M. Everingham and C. Needham, eds. Proceedings of the British Machine Conference. BMVA Press. doi: 10.5244/C.22.50.

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

Clapham, M., G. Narbonne, and J. Gehling. 2003. Paleoecology of the oldest known animal communities: Ediacaran assemblages at Mistaken Point, Newfoundland. Paleobiology 29:527–544.
Dale, M. 1994. Straightening the horseshoe: a Riemannian resolution. Coenoses 9:43–53.
Darroch, S., M. Laflamme, and M. Clapham. 2013. Population structure of the oldest known macroscopic communities from Mistaken Point, Newfoundland. Paleobiology 39:591–608.
Dekking, F., C. Kraaikamp, H. Lopuhaä, and L. Meester. 2005. A modern introduction to probability and statistics: understanding why and how. Springer-Verlag, London.
Dunn, J. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3(3):32–57.
Einbinder, S., T. Mass, E. Brokovich, Z. Dubinsky, J. Erez, and D. Tchernov. 2009. Changes in morphology and diet of the coral *Stylophora pistillata* along a depth gradient. Marine Ecology Progress Series 381:167–174.
Fulton, E. A., A. D. M. Smith, and A. E. Punt. 2005. Which ecological indicators can robustly detect effects of fishing? ICES Journal of Marine Science. 62:540–551.
Gower, J. C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–871.
Hill, M., and H. Gauch. 1980. Detrended correspondence analysis: an improved ordination technique. Pp. 47–58 *in* Classification and ordination. Springer, Nijmegen, Netherlands.
Holland, S., and M. Patzkowsky. 2007. Gradient ecology of a biotic invasion: biofacies of the type Cincinnatian series (Upper Ordovician), Cincinnati, Ohio region, USA. Palaios 22:392–407.
Holland, S., A. Miller, D. Meyer, and B. Dattilo. 2001. The detection and importance of subtle biofaces within a single lithofacies: the Upper Ordovician Kope Formation of the Cincinnati, Ohio region. Palaios 16:205–217.
Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37:547–579.
Jimenez, S., F. Gonzalez, and A. Gelbukh. 2010. Text comparison using soft cardinality. Pp. 297–302 *in* International symposium on string processing and information retrieval. Springer, Berlin.
Jimenez, S., F. A. Gonzalez, and A. Gelbukh. 2016. Mathematical properties of soft cardinality: enhancing Jaccard, Dice and cosine similarity measures with element-wise distance. Information Sciences 367–368:373–389.
Kowalewski, M., K. Flessa, and J. Aggen. 1994. Taphofacies analysis of recent shelly cheniers (beach ridges), northeastern Baja California, Mexico. Facies 31:209–242.
Kramer, N., R. Tamir, G. Eyal, and Y. Loya. 2020. Coral morphology portrays the spatial distribution and population size-structure along a 5–100m depth gradient. Frontiers in Marine Science 7:1–13.
Legendre, P., and E. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–280.
Legendre, P., and L. Legendre. 2012. Numerical ecology. 3rd ed. Elsevier, Oxford, UK.
Leinster, T., and C. Cobbold. 2012. Measuring diversity: the importance of species similarity. Ecology 93:477–489.
Li, C., X. Zhang, X. Liu, O. Luukkanen, and F. Berninger. 2006. Leaf morphological and physiological responses of *Quercus aquifolioides* along an altitudinal gradient. Silva Fennica 40:5.
MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1(14):281–297.
Magurran, A. 2004. Measuring biological diversity. Blackwell Science, Oxford.
Miles, D., and R. Ricklefs. 1984. The correlation between ecology and morphology in deciduous forest passerine birds. Ecology 65:1629–1640.
Molodtsov, D. 1999. Soft set theory—first results. Computers and Mathematics with Applications 47(4–5):19–31.
Nanglu, K., J-B. Caron, and R. Gaines. 2020. The Burgess Shale paleocommunity with new insights from Marble Canyon, British Columbia. Paleobiology 46:58–81.
Novack-Gottshall, P. 2007. Using a theoretical ecospace to quantify the ecological diversity of Paleozoic and modern marine biotas. Paleobiology 33:273–294.
Økland, R. 1999. On the variation explained by ordination and constrained ordination axes. Journal of Vegetation Science 10:131–136.
Patzkowsky, M. 1995. Gradient analysis of Middle Ordovician brachiopod biofacies: biostratigraphic, biogeographic, and macroevolutionary implications. Palaios 10:154–179.
Podani, J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. Taxon 48:331–340.
Podani, J., and I. Miklós. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. Ecology 83:3331–3343.
Price, T. 1991. Morphology and ecology of breeding warblers along an altitudinal gradient in Kashmir, India. Journal of Animal Ecology 60:643–664.
Puijalon, S., and G. Bornette. 2004. Morphological variation of two taxonomically distant plant species along a natural flow velocity gradient. New Phytologist 163:651–660.
Scarponi, D., and M. Kowalewski. 2004. Stratigraphic paleoecology: bathymetric signatures and sequence overprint of mollusk associations from upper Quaternary sequences of the Po Plain, Italy. Geology 32:989–992.
Schubert, E., and P. Rousseeuw. 2019. Faster *k*-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. Pp. 171–187 *in* G. Amato, C. Gennaro, V. Oria, and M. Radovanović, eds. Similarity search and applications. Springer, Berlin.
Sidorov, G., A. Gelbukh, H. Gómez-Adorno, and D. Pinto. 2014. Soft similarity and soft cosine measure: similarity of features in vector space model. Computación y Sistemas 18:491–504.
Soto, D., S. De Palmas, M. Ho, V. Denis, and C. Chen. 2018. Spatial variation in the morphological traits of *Pocillopora verrucose* along a depth gradient in Taiwan. PLoS ONE 13(8):e0202586.
Swan, J. 1970. An examination of some ordination problems by use of simulated vegetational data. Ecology 51:89–102.
Ter Braak, C. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179.
Walker, S. 2015. Indirect gradient analysis by Markov-chain Monte Carlo. Plant Ecology 216:697–708.
Whittaker, R. 1967. Gradient analysis of vegetation. Biological Reviews 42:207–264.
Williamson, M. 1978. The ordination of incidence data. Journal of Ecology 66:911–920.
Wolda, H. 1981. Similarity indices, sample size and diversity. Oecologia 50:296–302.

## Appendix: Implementing the Accompanying R Code

The included R code (see Supplementary Material) takes as its input two $n \times p$ matrices $X$ and $Y$, where rows represent $n$ individual specimens coded for $p-1$ morphological/phenotypic traits. The $p$th column must contain a numeric value indicating which species that individual specimen belongs to. The value used to indicate a given species must be consistent across samples (e.g., 1 should refer to the same species [species 1] in $X$ and $Y$). For

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press

example, three individuals representing three taxa with five morphological characters each would be coded as similar to what is seen in Supplementary Table 1. Note that because double-zero values do not contribute to the morphological similarity of taxa (see Gower 1971), taxa can be coded using their own unique set of morphological traits. For example, if an arbitrary plant species $s_1$ is being described in terms of an arbitrary morphological character $p_1$ (e.g., leaf area) and another arbitrary species $s_2$ in the sample is not a plant (such that leaf area cannot be measured), putting a 0 for character $p_1$ when describing $s_2$ ensures this character does not contribute to the morphological similarity of specimens of $s_2$ when comparing $X$ and $Y$. Incorporating at least two specimens of each species included in the analysis will give more informative results—if a species is represented by only a single specimen in both samples, the morphological similarity of the two specimens can only be 1 (if their morphology is identical) or 0 (if their morphology differs in any way). If $p = 1$ (that is, if there are no morphological characters included), the index functions as the original Chao-Jaccard index (Chao et al. 2005). Relative abundance values of each species are calculated automatically from the number of specimens—they do not need to be manually entered.

https://doi.org/10.1017/pab.2021.33 Published online by Cambridge University Press