COMMUNITY ECOLOGY 3(1): 95-99, 2002
1585-8553 © Akadémiai Kiadó, Budapest

Measuring taxonomic diversity with parametric information functions

C. Ricotta

Department of Plant Biology, University of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185 Rome, Italy. Tel.: +39-06-499 12408, Fax: +39-06-445 7540, E-mail: carlo.ricotta@uniroma1.it

Keywords: Diversity profiles, Information theory, Rényi's generalized information, Schur-concavity, Set monotonicity.

Abstract: This paper discusses the measurement of taxonomic diversity of a given community or set of species. First, given **D**, the matrix of pairwise distances between species, I propose to summarize taxonomic diversity through Rényi's information-theoretical parametric formalism on **D**. In this way, a family of taxonomic diversity indices is obtained which shows different sensitivity to the presence of species with different levels of taxonomic distinctness. The adequacy of information-theoretical indices for quantifying taxonomic diversity is discussed.

Introduction

Traditional ecological diversity indices summarize information about the relative abundances of species within a community or sample without regard to differences between constituting species. Nevertheless, for large-scale environmental protection purposes, data on species abundances are generally unknown and the only available information relates to the number of species. In addition, when focusing on conservation problems, species abundances are mostly irrelevant and the common treatment of species abundances is largely meaningless in case of systematically remote organisms, such as oaks and orchids (Izsák and Papp 2000). In this view, Vane-Wright et al. (1991) were the first to suggest that, for conservation purposes, we should quantify the relative values we assign to different species such that their relative abundances are ignored. Based on the seminal work of May (1990), they proposed a measure of species "taxonomic distinctness" based on phylogenetic relationships amongst species. The proposal of Vane-Wright et al. (1991) is based only on the topology of cladistic classifications and is appropriate when branch lengths are unknown, whereas Faith (1992, 1995) suggested to measure taxonomic diversity based on known branch lengths. The resulting measure of phylogenetic diversity (PD) is simply the cumulative branch length of the full phylogenetic tree.

Unfortunately, this literature has been largely ignored in environmental monitoring research, where emphasis is not on selecting species to conserve but rather on assessing whether sampled communities exhibit some changes in biodiversity following environmental degradation or remediation efforts (Clarke and Warwick 1998). Also, detailed, fully resolved cladograms are not available for most groups of organisms, and the basic information on species relatedness is often just the set of pairwise distances between species. These distances (not necessarily fulfilling the distance axioms) can be based on morphological or functional differences (Izsák and Papp 1995), on Linnaean taxonomy (Izsák and Papp 1995, Clarke and Warwick 1998, Rogers et al. 1999), or on more refined molecular biological methods (Solow et al. 1993, Shimatani 2001).

Finally, since the ultimate aim of any summary statistics is to provide a manageable tool for characterizing and comparing different multivariate sets based on distinct objectives and motivations, it is generally understood that different indices may inconsistently rank a given pair of sets. The main reason for this confusion is that, by mapping the structure of a multidimensional set such as a biological community with scalars, information is necessarily lost, and there is no ideal function capable of uniquely characterizing all aspects of taxonomic diversity. Quoting from Patil and Taillie (1979, p. 15): "Such inconsistencies are inevitable whenever one attempts to reduce a multidimensional concept to a single number [...]. For example, the mean and median are inconsistent measures of central tendency; likewise, the standard deviation, mean absolute

deviation, and range are inconsistent measures of spread". A paradigmatic example within the context of conservation biology is Faith's (1992) criticism on the measure of taxonomic distinctness proposed by Vane-Wright et al. (1991). Faith (1992) notes that in re-examining reserve-selection scenarios based on a phylogeny of bumble bees (Apidae), the index PD produces different priorities for species conservation relative to the measure proposed by Vane-Wright et al. (1991).

A more complete summarization of taxonomic diversity would require a parametric family of indices whose members have different sensitivity to the presence of species with different levels of taxonomic distinctness. In this paper, I propose to summarize taxonomic diversity by Rényi's (1961) formalism on the matrix of pairwise distances between species.

Background

Let us define a species distance matrix **D**, the elements d_{ij} of which represent the taxonomic distances between the *i*-th and the *j*-th species such that $d_{ii} = 0$, and $d_{ij} = d_{ji}$ for any $i \neq j$. As an application for demonstration, I used a small artificial community composed of the following five species: *Ostrya virginiana, Populus grandidentata, Prunus serotina, Quercus rubra,* and *Ulmus americana*. Based on the pairwise genetic distances d_{ij} proposed by Shimatani (2001, Appendix 2), I constructed the genetic distance matrix **D** of Table 1. Once the genetic distance matrix of the analyzed species set has been constructed, a straightforward way to collapse its structure into a summary statistic Σ_D is to sum the (off-diagonal) elements d_{ij} in **D**:

$$\Sigma_{\mathbf{D}} = \sum_{i,j \in \mathbf{D}} d_{ij} \qquad (1)$$

Within the context of taxonomic diversity, this measure was independently proposed by Warwick and Clarke (1995) and Izsák and Papp (2000) for quantifying the structure of Linnaean taxonomic trees. However, a slightly different formulation of Σ_D termed the "Wiener index" has been known in chemometrics since the late 1940s for summarizing the topology of molecular structures (Wiener 1947, Ricotta et al. 2000). Since **D** is symmetric with zeros in its main diagonal, one could reduce the calculation of Σ_D to the upper triangular submatrix without loss of any information. It is easily demonstrated that total species distance Σ_D satisfies set monotonicity, which is a desirable property for biodiversity measures (Solow and Polasky 1994). That is, the value of Σ_D will increase by adding a new species *x* to a given species set S. Formally, $\Sigma_D(S \cup \{x\}) > \Sigma_D(S)$. Nevertheless, Σ_D is not a species richness index insofar as it is not a monotone increasing function of the number of species in the sample plot. Instead, its values are jointly determined by the number of species and the structural complexity of the taxonomic tree (Izsák and Papp 2000). For example, for the genetic distance matrix of Table 1, $\Sigma_D = 60.64$.

Table 1. Genetic distance matrix for an atrificial five-species community composed of *Ostrya virginiana*, *Populus grandidentata*, *Prunus serotina*, *Quercus rubra* and *Ulmus americana*. The pairwise genetic distances are those proposed by Shimatani (2001, Appendix 2).

	Ov	Pg	Ps	Qr	Ua
Ostrya virginiana (Ov)	0	4.31	3.01	1.89	1.89
Populus grandidentata (Pg)	4.31	0	3.66	4.31	2.85
Prunus serotina (Ps)	3.01	3.66	0	3.34	2.21
Quercus rubra (Qr)	1.89	4.31	3.34	0	2.85
Ulmus americana (Ua)	1.89	2.85	2.21	2.85	0

A different approach for summarizing taxonomic diversity consists in the application of information-theoretical formalism to **D**. Let us consider a system composed of *N* different sets where p_i is the relative abundance of the *i*-th set ($i = 1, 2, 3, \ldots, N$) such that $0 \leq p_i \leq 1$ and $\sum_{i=1}^{N} p_i = 1$. For a distribution function characterized by its relative abundance vector $\mathbf{p} = (p_1, p_2, \ldots, p_N)$, Rényi (1961) extended the concept of Shannon's (1948) entropy by defining a generalized information (or entropy in information-theoretical sense) of order α as:

$$H^{(\alpha)} = \frac{1}{1-\alpha} \ln \sum_{i=1}^{N} p_i^{\alpha} \qquad (2)$$

where α makes mathematical sense for $-\infty \geq \alpha \geq \infty$. Nevertheless, as explained in the remainder, for negative values of α, the resulting index has some undesirable properties that render it inadequate for summarizing taxonomic diversity. Therefore, non-negative values of α (i.e., $\alpha \geq 0$) should be preferably used.

Rényi proved that $H^{(\alpha)}$ satisfies certain axioms that entitled it to be regarded as a measure of generalized information (Beck and Schlögl 1993). Mathematically, the various measures obtained by varying α are in fact different moments of the same basic Rényi's information function. Notice also that in Rényi's original definition, logarithm to the base 2 is used to measure information content in bits, while in ecological applications the natural logarithm is traditionally used (Tóthmérész 1995).

The statistical information of a given system is basically a measure of uncertainty in predicting the relative abundance of the different sets. The maximum value of Rényi's entropy is obtained in case of equiprobability (i.e., if $p_i = 1/N$ for $i = 1, 2, \ldots, N$). Minimum entropy is obtained if there is a set having its relative abundance ap-

proaching 1 (the abundances of all other sets being zero), which implies $H_{\min}^{(\alpha)} = 0$. Since uncertainty is maximum when entropy is the highest, the entropy concept forms one of the basic foundations of ecological diversity theory (Orlóci 1991).

Within the context of ecological diversity theory, Hill (1973) showed that the generalized information function $H^{(\alpha)}$ has many desirable properties as a diversity index. One particularly convenient property is that a number of traditional diversity indices computed from species relative abundances p_i are special cases of $H^{(\alpha)}$. For $\alpha = 1$, Equation (2) is defined in the limiting sense using l'Hospital's rule of calculus, and $H^{(1)}=\sum^{N}_{i=1} p_i \ln p_i$ (i.e., Shannon's entropy). For $\alpha = 0$, $H^{(0)} = \ln N$, where N is species richness; for $\alpha = 2$, $H^{(2)} = \ln 1/D$, where D is Simpson's (1949) dominance index $D=\sum^{N}_{i=1} p_i^2$, and for $\alpha = \infty$, $H^{(\infty)} = \ln 1/d = \ln 1/p_{\max}$, where d is the dominance index of Berger and Parker (1970) and $p_{\max}$ is the proportional abundance of the most frequent species. While traditional indices supply point descriptions of community diversity, according to Rényi's formulation, there is a continuum of possible diversity measures that differ in their sensitivity to the rare and abundant species, becoming increasingly dominated by the most common species for increasing values of the parameter α. For a given community, $H^{(\alpha)}$ is a decreasing function of α. From Equation (2) it follows that for a given community $\ln N \geq H \geq \ln 1/D \geq \ln 1/d$, where equality holds for equiprobable distributions. In other words, traditional species diversity can be described by its diversity profile of $H^{(\alpha)}$ vs. α (Patil and Taillie 1979, 1982).

New taxonomic diversity measures

The application of Rényi's formalism to different systems is based on the possibility of partitioning all system elements into N sets, so that a finite probability scheme is obtained. Since the criterion for partitioning the elements of a given system is generally not unique, it is always possible to select for any system several information measures that represent statistical characteristics of that system (Bonchev 1993). In this view, a simple way to obtain a finite probability scheme from the taxonomic distances d_{ij} of a given species distance matrix, such as the genetic distance matrix of Table 1, is to add all distances d_{ij} along row i (or column i) of **D**. This results in a vector **v** = (11.1, 15.13, 12.22, 12.39, 9.8) whose elements v_i represent the (cumulative) taxonomic distance between species i and all other species. The corresponding parametric taxonomic diversity measure is:

$$H_{v(i)}^{(\alpha)} = \frac{1}{1-\alpha} \ln \sum_{i=1}^{N} \left(\frac{v_i}{\sum_{\mathbf{D}}} \right)^{\alpha} \qquad (3)$$

From Equation (3), it follows that for $\alpha = 0$, $H_{v(i)}^{(0)} = \ln N$ (a monotonic function of species richness), whereas for $\alpha = \infty$, $H_{v(i)}^{(\infty)} = \ln(1/p_{\max}) = \ln\left(\sum_{\mathbf{D}}/v_{\max}\right)$, where $v_{\max}/\sum_{\mathbf{D}}$ is the cumulative genetic distance of the taxonomically most distinct species transformed to a finite probability space. That is, by selecting $\alpha >> 1$, the sensitivity of $H_{v(i)}^{(\alpha)}$ is tuned in the domain of the taxonomically most distinct species. Conversely, if one analyzes community structure considering a wider range of species (i.e., including taxonomically less remote species), then lower values of α should be selected.

Similarly, one can introduce a parametric information function on pairwise genetic distances $H_{d(ij)}^{(\alpha)}$ by transforming the elements d_{ij} of the upper triangular submatrix of **D** in a finite probability scheme (Bonchev 1993):

$$H_{d(ij)}^{(\alpha)} = \frac{1}{1-\alpha} \ln \sum_{i>j} \left(\frac{2d_{ij}}{\sum_{\mathbf{D}}} \right)^{\alpha} \qquad (4)$$

Naturally, $H_{v(i)}^{(\alpha)}$ and $H_{d(ij)}^{(\alpha)}$ conform to the usual diversity axiom that maximal diversity arises for an equiprobable distribution of pairwise species distances (see Pielou 1975). In both cases, in analogy to ecological diversity theory, one can define a taxonomic diversity ordering of a given set of communities. In this view, community A is taxonomically more diverse than community B (written as A > B) if the taxonomic diversity profile $H_{v(i)}^{(\alpha)}$ (or $H_{d(ij)}^{(\alpha)}$) vs. α of A lies everywhere above the taxonomic diversity profile of B (Tóthmérész 1995). Notice that the taxonomic diversity ordering is only partial in that two diversity profiles may intersect. In this case, A and B cannot be unambiguously ordered according to their taxonomic diversity as different moments of $H_{v(i)}^{(\alpha)}$ (or $H_{d(ij)}^{(\alpha)}$) rank them in contradictory ways. That is, it is not necessarily true that for every A and B, either A > B or B > A.

Further, two parametric indices of taxonomic evenness that measure the degree to which taxonomic distinctness is divided equitably among species can be derived in the usual way from Equations (3) and (4) as $\exp H_{v(i)}^{(\alpha)} / N$ and $\exp H_{d(ij)}^{(\alpha)} / N$, respectively. Being based on Rényi's parametric information, both evenness indices conform to the Lorenz partial order that has been proposed by several authors as the foundation of the ecological notion of evenness (Taillie 1979, Gosselin 2001, Ricotta and Avena 2002).

Discussion and conclusion

Thus far, I suggested that Rényi's formalism might be used to describe quantitatively the distribution of the pairwise distances between species, to express taxonomic diversity. I assumed a certain degree of conceptual analogy

between traditional ecological diversity theory developed for summarizing species relative abundance structure and taxonomic diversity. In traditional diversity theory, it is generally agreed that diversity measures combine in a non-standard way two components: species richness and evenness. High species richness and evenness are both equated with high diversity such that community B is considered intrinsically more diverse than community A without reference to indices, provided A leads to B by a finite sequence of forward transfers of species abundances from one species to another strictly less abundant species (Patil and Taillie 1979, 1982, Solomon 1979). Formally, let A and B be communities with respective species abundance vectors $\mathbf{p}^{(A)}$ and $\mathbf{p}^{(B)}$. We say that A leads to B by a forward transfer of abundances if there are positive integers i and j such that $p_i^{(A)} > p_j^{(A)} \geq 0$ and

$$p_k^{(B)} = \begin{cases} p_k^{(A)} & \text{if } k \neq i, j \\ p_i^{(A)} - h & \text{if } k = i \\ p_j^{(A)} + h & \text{if } k = j \end{cases} \quad (5)$$

where $0 \leq h \leq p_i^{(A)} - p_j^{(A)}$. Such a transfer increases species richness when $p_j^{(A)} = 0$, and increases evenness when $p_j^{(A)} > 0$ (Patil and Taillie 1979, 1982). In ecological diversity theory, diversity measures δ that satisfy this property are termed "Schur-concave". Given a Schur-concave diversity index, $\delta^{(A)} \leq \delta^{(B)}$ whenever community A leads to community B by a forward transfer of species relative abundances. This requirement that transferring abundances should increase the index is known in econometrics as Dalton's (1920) "principle of transfers" and was originally proposed in connection with the measurement of income inequality (see Patil and Taillie 1982). Applying Dalton's principle of transfers to the computation of taxonomic diversity, it is easily shown that Rényi's generalized entropy function $H^{(\alpha)}$ is Schur-concave in the interval $0 \geq \alpha \geq \infty$. Nevertheless, it is also easily shown that Schur-concave parametric indices such as $H^{(\alpha)}$ satisfy set monotonicity only for the trivial case $\alpha = 0$. That is, by adding a new species x to a given species set S (i.e., increasing the richness component of the community), the values of $H_{v(i)}^{(\alpha)}$ and $H_{d(ij)}^{(\alpha)}$ will not necessarily increase. This effect may be a serious drawback for those who believe that taxonomic diversity measures must possess set monotonicity.

To conclude, diversity research is one of the fields where relevant biological problems meet sophisticated mathematical tools that evolved at the crossroad with other statistical disciplines such as econometrics or chemometrics. For instance, information-theoretical measures belong to the standard apparatus for quantifying the topological structure of chemical compounds (Bonchev 1993, Basak et al. 2000). Also, the proposal of applying information-theoretical measures for summarizing taxonomic diversity is not entirely new. Pielou (1975) formalized this idea in a modified version of the Shannon index where, besides species diversity, generic and familial diversity is also considered, whereas Ricotta (2002) proposed a generalization of Shannon's entropy that takes into account a taxonomic weighting factor based on the cumulative taxonomic distances v_i. Nonetheless, to the best of my knowledge, the idea of mapping the structure of the species distance matrix with parametric information is proposed here for the first time.

Finally, it is again worth noting that diversity measures are merely numbers and their relevance to ecological problems must be judged on the basis of observed correlations with other environmental variables (Molinari 1989). Quoting from Magurran (1988, p. 113): "diversity measures are valuable, but are only a means to an end. That end is that ecologists should be able to ask the questions and formulate the hypotheses to help them understand, and sensibly manage, the natural world".

Acknowledgements: I wish to thank Dan Faith and an anonymous referee for the stimulating comments on a previous version of this paper.

References

Basak, S.C, A.T. Balaban, G.D. Grunwald and B.D. Gute. 2000. Topological indices: Their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.* 40: 891-898.

Beck, C. and F. Schlögl. 1993. *Thermodynamics of Chaotic Systems.* Cambridge University Press, Cambridge.

Berger, W.H. and F.L. Parker. 1970. Diversity of planctonic Foraminifera in deep sea sediments. *Science* 168: 1345-1347.

Bonchev, D. 1993. Information-Theoretic Indices for Characterization of Chemical Structure. Research Studies Press, Letchworth, UK.

Clarke, K.R. and R.M. Warwick.1998. A taxonomic distinctness index and its statistical properties. *J. Appl. Ecol.* 35: 523-531.

Dalton, H. 1920. The measurement of the inequality of incomes. *Econ. J.* 30: 348-361.

Faith, D.P. 1992. Conservation evaluation and phylogenic diversity. *Biol. Conserv.* 61: 1-10.

Faith, D.P. 1995. Phylogenic pattern and the quantification of organismal biodiversity. In: D.L. Hawksworth (ed.), *Biodiversity Measurement and Estimation.* Chapman & Hall, London, pp. 45-58.

Gosselin, F. 2001. Lorenz partial order: the best known logical framework to define evenness indices. *Community Ecol.* 2: 197-207.

Hill, M.O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54: 427-431.

Izsák, J. and L. Papp. 1995. Application of the quadratic entropy index for diversity studies on drosophilid species assemblages. *Environ. Ecol. Stat.* 2: 213-224.

Izsák, J. and L. Papp. 2000. A link between ecological diversity indices and measures of biodiversity. *Ecol. Model.* 130: 151-156.

Magurran, A. 1988. *Ecological Diversity and its Measurement*. Princeton University Press, Princeton, NJ.

May, R.M. 1990. Taxonomy as destiny. *Nature* 347: 129-130.

Molinari, J. 1989. A calibrated index for the measurement of evenness. *Oikos* 56: 319-326.

Orlóci, L. 1991. *Entropy and Information*. SPB Academic Publishing, The Hague, NL.

Patil, G.P. and C. Taillie. 1979. An overview of diversity. In: J.F. Grassle, G.P. Patil, W. Smith and C. Taillie (eds.), *Ecological Diversity in Theory and Practice*. International Co-operative Publishing House, Fairland, MD, pp. 3-27.

Patil, G.P. and C. Taillie. 1982. Diversity as a concept and its measurement. *J. Am. Stat. Ass.* 77: 48-567.

Pielou, E.C. 1975. *Ecological Diversity*. Wiley Interscience, New York.

Rényi, A. 1961. On measures of entropy and information. In: G. Neyman (ed.), *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, California, Vol. 1, pp. 547-561.

Ricotta, C. 2002. Bridging the gap between ecological diversity indices and measures of biodiversity with Shannon's entropy: comment to Izsák & Papp. Ecol. Model., in press.

Ricotta, C. and G.C. Avena. 2002. On the information-theoretical meaning of Hill's parametric evenness. *Acta Biotheoretica* 50: 63-71.

Ricotta, C., A. Stanisci, G.C. Avena and C. Blasi. 2000. Quantifying the network connectivity of landscape mosaics: a graph-theoretical approach. *Community Ecol.* 1: 89-94.

Rogers, S., K.R. Clarke and J.D. Reynolds. 1999. The taxonomic distinctness of coastal bottom-dwelling fish communities of the North-east Atlantic. *J. Anim. Ecol.* 68: 769-782.

Shannon, C. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27: 379-423.

Shimatani, K. 2001. On the measurement of species diversity incorporating species differences. *Oikos* 93: 135-147.

Simpson, E.H. 1949. Measurement of diversity. *Nature* 163: 688.

Solomon, D.L. 1979. A comparative approach to species diversity. In: J.F. Grassle, G.P. Patil, W. Smith and C. Taillie (eds.), *Ecological Diversity in Theory and Practice*. International Co-operative Publishing House, Fairland, MD, pp. 29-35.

Solow, A.R. and S. Polasky. 1994. Measuring biological diversity. *Environ. Ecol. Stat.* 1: 95-107.

Solow, A., S. Polasky and J. Brodaus. 1993. On the measurement of biological diversity. *J. Environ. Econ. Manage.* 24: 60-68.

Taillie, C. 1979. Species equitability: a comparative approach. In: J.F. Grassle, G.P. Patil, W. Smith and C. Taillie (eds.), *Ecological Diversity in Theory and Practice*. International Co-operative Publishing House, Fairland, MD, pp. 51-62.

Tóthmérész, B. 1995. Comparison of different methods for diversity ordering. *J. Veg. Sci.* 6: 283-290.

Vane-Wright R.I., C.J. Humphries and P.M. Williams. 1991. What to protect: systematics and the agony of choice. *Biol. Conserv.* 55: 235-254.

Warwick, R.M. and K.R. Clarke. 1995. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Mar. Ecol. Prog. Ser.* 129: 301-305.

Wiener, H. 1947. Structural determination of paraffin boiling point. *J. Amer. Chem. Soc.* 69: 17-20.