Journal of Phylogenetics & Evolutionary Biology

Brocchieri, J Phylogen Evolution Biol 2015, 3:1
DOI: 10.4172/2329-9002.1000e109

Editorial · Open Access

Phylogenetic Diversity and the Evolution of Molecular Sequences

Luciano Brocchieri*

Department of Molecular Genetics & Microbiology and Genetics Institute, University of Florida, Gainesville, FL 32606, USA

Diversity and its measures is a long-standing and widely explored concept in ecology, physics, economics, and sociology, among others. In ecology, the concept of diversity is tightly connected with the idea of a "healthy" community and of conservation. During the last sixty years a deluge of different indices has been developed to represent diversity of ecological communities [1]. Diversity within a community of species, or between different communities, is affected by the phylogenetic relations among species. Intuition suggests that of two communities composed by the same number and abundance of species, the community composed of more distantly related species is more diverse since distantly related species are likely to exhibit a greater number of unique features.

Different indices have been developed to represent the effect of phylogenetic relatedness on diversity. Faith's *Phylogenetic diversity* [2] measures diversity based on a phylogenetic tree, as the sum of the lengths of its branches L_i, i.e., as the size of the tree. Faith's phylogenetic diversity takes into consideration the phylogenetic relations among the species in the community but not the relative abundance of species. In contrast, Rao's quadratic entropy $Q=\Sigma_{i,j}p_ip_jd_{ij}$ incorporates frequencies evaluating diversity as the average pair-wise dissimilarity d_{ij} of pairs of individuals randomly sampled from the community [3]. *Phylogenetic entropy* $H_p=-\sum_i L_ia_i\ln a_i$ evaluates diversity generalizing Shannon entropy based on a rooted phylogenetic tree of species associating with each branch *i* of length L_i a frequency a_i corresponding to the sum of the frequencies of all the species descended from that branch [4].

In a seminal paper of 2006, Lou Jost advocated the idea that standard diversity indices generally cannot be considered direct measures of diversity and do not show properties expected from "true" diversities [5]. However, all standard indices correspond to and can be transformed into "true diversities" of similar functional form (Table 1), described in different fields as "effective number" or "numbers equivalent" of species, and defined as the number of equally-frequent species necessary to obtain the same diversity-index value of the community under consideration. "True diversities" are in the form of Hill numbers [6] of some order *q* (Table 1). The order of diversity emphasizes differently the most or least frequent species. When $q=0$ all positive frequencies are flattened to 1.0 and thus are counted as occurrences, as in the Species Richness Index. When $0<q<1$ rare species are inflated in calculating diversity. When $q>1$, the most frequent species are favored and when $q\to\infty$ only the frequency of the most abundant species contributes to the calculation of diversity (corresponding to the inverse Berger Parker Index $=\max_i p_i=1/{}^{\infty}D$). Shannon entropy uniquely corresponds to the special case of diversity ${}^1D$ that does not favor any frequency. Contrary to most of the diversity indices, numbers equivalents behave as would be expected from true measures of diversity upon compositional changes (see [5,6] for examples) and I will refer to them as "true diversities".

Using the conceptual unifying perspective advocated by Jost, Faith's phenotypic diversity (PD) can be generalized to any *phylogenetic diversity* ${}^qPD$ of order *q*, considering at once the underlying phylogenetic tree and species frequencies [7]. Considering an ultrametric tree whose branches represent amounts of evolution proportional to time (Figure 1), to each branch of the tree can be assigned an abundance corresponding to the sum of the frequencies of all species derived from that branch. To any specific time within any given interval (e.g., from the time -*T* of the root to present), corresponds a virtual community composed of "species" identified by all branches present at that time with frequencies assigned as described (Figure 1). True diversities can be calculated for each of these communities and the average diversity of all communities within a chosen time interval can be calculated as an *alpha* diversity of any order *q* [8]. In the example of Figure 1, the same communities are conserved within time intervals T_1, T_2, and T_3 and their mean (*alpha*) diversity is calculated as:

$$^{q}\bar{D}(T)=\left(\frac{T_1}{T}\sum_{i=1}^{S_1=2}s_i^q+\frac{T_2}{T}\sum_{i=1}^{S_2=3}q_i^q+\frac{T_3}{T}\sum_{i=1}^{S_3=4}p_i^q\right).$$

With some rearrangement and substitutions, in the general case this averaging is equivalent to the general formulation of *mean diversity of order q over time T* [7]:

$$^{q}\bar{D}(T)=\frac{1}{T}\left\{\sum_{i\in B_T}L_i\left(\frac{a_i}{T}\right)^q\right\}^{1/(1-q)}$$

where L_i represent branch lengths and a_i the corresponding species frequencies. Multiplied by the length of the time interval *T*, mean diversities give phylogenetic diversities ${}^qPD(T)=T\cdot{}^q\bar{D}(T)$ of the same order of ${}^q\bar{D}(T)$. When $q=0$ these correspond to Faith's phylogenetic diversity. In the case of $q=1$, ${}^q\bar{D}(T)$ is not defined by the expression above but its limit ${}^1\bar{D}(T)$ exists and is:

$$^{1}\bar{D}(T)=\exp\left[-\sum_{i\in B_T}\frac{L_i}{T}a_i\ln a_i\right].$$

Similar diversities can be calculated for a rooted non-ultrametric tree [7] substituting *T* with the weighted average tree-depth

$$\bar{T}=\Sigma_{i\ B_{\bar{T}}}L_ia_i\,.$$

Note that ${}^q\bar{D}$ is insensitive to the scale of the tree. Given a tree topology and branch lengths, rescaling the tree so that the new tree has the same topology but branches *k*-fold the original lengths, produces over the time interval *kT* the same mean diversity ${}^q\bar{D}$ than the original tree over time *T*. In contrast, phylogenetic diversity is rescaled to $T.\,kT.{}^q\bar{D}$ and thus increases linearly with *k*. As a consequence, phylogenetic diversity of any order tends to infinity as more and more new features are added to species.

Mean diversities are what their name suggests, an average of the species diversities of different communities. In this application communities are sampled in time rather than in space, from $-T$ to present day. This "historical" average is relevant to describing diversity in a present-day community because the amount of evolution reflected by the length of each branch is assumed to directly reflect in the number of features that are in common and unique to all species descending from that branch. Thus, the length of a branch relative to the size of the tree (sum of all branch lengths), should correspond to

***Corresponding author:** Luciano Brocchieri, Cancer and Genetics Research Complex 2033 Mowry Rd, Gainesville, FL 32606, USA, Tel: 352-273-8131; E-mail: lucianob@ufl.edu

Received December 18, 2014; **Accepted** December 19, 2014; **Published** December 28, 2014

Citation: Brocchieri L (2015) Phylogenetic Diversity and the Evolution of Molecular Sequences. J Phylogen Evolution Biol 3: e109. doi:10.4172/2329-9002.1000e109

the proportion of all features that are shared by the corresponding clade. This assumption is verified if and only if each feature develops only once among all lineages and if features are never lost, a type of parsimonious evolution known as Camin-Sokal parsimony [9]. Also implicit in this measure of community diversity is the description of a species as the set of all features developed along its lineage from time $-T$. Only within this framework mean diversities correctly answer to the question: what is the effective number of species in the community?

The special assumptions on the evolutionary process and definition of species implicit in the above derivation give rise to unexpected results if they are not taken into account. Consider for example a sample of two equally frequent and phylogenetically related species (Figure 2). At the time $-T$ of speciation, the two species and their last common ancestor are identical. Intuition suggests that the true diversity (effective number of species) of a community composed of those two identical species should be 1.0. As the two species gradually differentiate, intuition suggests that an infinitesimal amount of differentiation should result in true diversity infinitesimally greater than 1.0. As differentiation increases, true diversity of zero order (or assuming equal frequencies) in the community should correspondingly gradually increase, and it should approach 2.0 as the two species become so diversified that no common character can be recognized between them. It is interesting to consider instead the behavior of mean diversity and phylogenetic diversity calculated for this system in the time intervals $(-T,x]$, where x identifies either the progression of time, or species evolving at different rates, or from different times (Figure 2). At $x=0$ (time $-T$) mean diversity is not defined but its right-hand limit exists with a value $^{0}\bar{D}(0)=2.0$. In the case of $x>0$ the mean diversity is also $^{0}\bar{D}(x)=2.0$ for any x. Thus, no matter the amount of evolution, i.e., the degree of differentiation between the two species, mean diversity remains constant, depending only on the number of lineages. Phylogenetic diversity instead grows proportionally to x ($^{0}PD(x)=2x$). However, it grows to infinity as x grows to infinity, measuring the total amount of evolution. From this simple example, it becomes clear that mean and phylogenetic diversities do not measure what one might expect them to measure. Phylogenetic diversities do not have the same dimensionality of a "true diversity", the effective number of species in a community. To be compared with the effect of unequal frequencies, phylogenetic diversities need to be divided by T, transforming them into mean diversities. Mean diversities on the other hand depend on the choice of $-T$ and on a definition of "species" that changes over time by the addition of more and more features. Thus two species each with one new unique feature or with 100 new unique features, are both 100% different. This behavior of mean diversitites can be corrected by setting a time $-T$ including in the calculation a root-branch of some length (branch L_7 in Figure 1). The length of this branch will define the "size" of the common ancestral species, and will result in a gradual increase in mean diversity when speciation occurs. It is unclear however how long this branch should be (how many characters should define the phenotype of the last common ancestral species?).

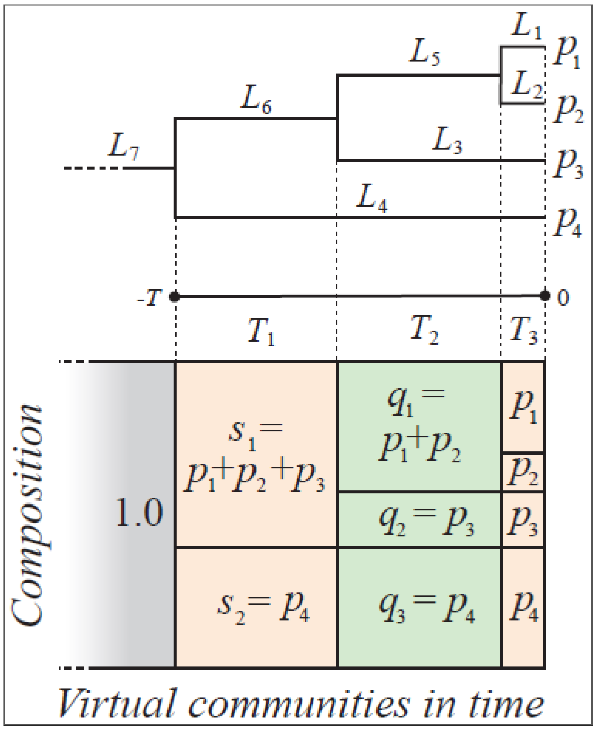

Figure 1: An ultrametric phylogenetic tree representing amounts of evolution proportional to time and corresponding "virtual communities" made of as many species as the number of branches, each with a frequency equal to the sum of the frequencies of all species descended from the branch.

Are Mean and Phylogenetic Diversities Applicable to the Evolution of Molecular Sequences?

Substitution of "features", in the form of nucleotide or amino acid types at each alignment position, and hence loss of features by substitution rather than *accumulation* of features, is the essence of how molecular sequence evolution occurs by point substitutions, and of how evolution is generally modeled in molecular sequence phylogenetics. Branches of molecular-sequence evolutionary trees represent the number of *state* substitutions that occur during evolution over a *fixed* number of characters, rather than to accumulation of new characters. A "species", as represented by a sequence, is not an empty set at the time $-T$ of the common ancestor, and its two incipient descendants in a binary tree are not 100% different but almost identical. The possibility of non-parsimonious evolution in the form of multiple substitutions and back substitutions at each sequence site is implicit in probabilistic models of sequence evolution. As a consequence, branch lengths are not proportional to the frequency of shared or unique states in present-time sequences. Because of these differences, the amount of evolution represented by branch lengths (or by their transformation in phenotypic dissimilarity), cannot be averaged to obtain effective diversity between sequences the same way it can assuming parsimonious evolution.

Mean and phylogenetic diversities depend on the rooting of the tree. Although it is not possible to re-root an ultrametric tree and to obtain a different ultrametric tree, a non-ultrametric tree can be re-rooted on any branch. Rooting the tree in different positions will generally result in different mean and phylogenetic diversities. The dependence on rooting of phylogenetic diversity is consistent with an intrinsic directionality of the evolutionary process, by which trees sharing the same topology and branch lengths but rooted in different positions are not equivalent. The dependence on rooting however is not intrinsic and not natural to phylogenetic trees constructed on the assumption of non-parsimonious time-reversible evolution, such as most trees derived by continuous-time Markov models of state-substitutions applied to multiple alignments of nucleic acid or protein sequences. The amount of diversity implicit in these phylogenetic trees is independent from the direction of time and from the position of the root on the tree.

The elegant unifying perspectives on community diversity

Index H name	H	$D(H)$	$D(p_i)$	*Order*
Species Richness	$\sum_{i=1}^{S} p_i^0$	H	$\sum_{i=1}^{S} p_i^0$	0
Shannon entropy	$-\sum_{i=1}^{S} p_i \ln p_i$	e^H	$\exp\left(-\sum_{i=1}^{S} p_i \ln p_i\right)$	1
Simpson Index	$\sum_{i=1}^{S} p_i^2$	$1/H$	$1\big/\sum_{i=1}^{S} p_i^2$	2
Gini-Simpson Index	$1-\sum_{i=1}^{S} p_i^2$	$1/(1-H)$	$1\big/\sum_{i=1}^{S} p_i^2$	2
Inverse Simpson Index	$1\big/\sum_{i=1}^{S} p_i^2$	H	$1\big/\sum_{i=1}^{S} p_i^2$	2
Tsallis entropy	$\left(1-\sum_{i=1}^{S} p_i^q\right)\big/(q-1)$	$\left[1-(q-1)H\right]^{1/(1-q)}$	$\left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}$	q
Rényi entropy	$\left(-\ln\sum_{i=1}^{S} p_i^q\right)\big/(q-1)$	e^H	$\left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}$	q
Berger-Parker Index	$\max\limits_i p_i$	$1/H$	$1\big/\max\limits_i p_i$	∞

Table 1: Indices of diversity and True diversities (modified from [5])

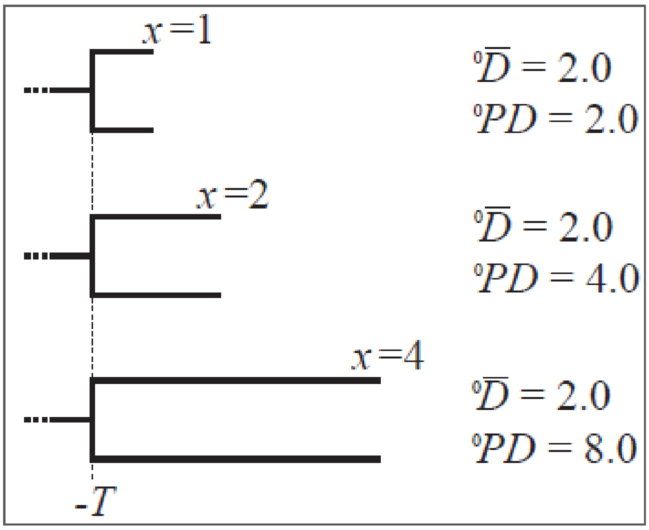

Figure 2: Three examples of trees, each composed of two species evolving at the same rate (ultrametric tree) but with different rates for each tree (or, equivalently, at different times). Mean and phylogenetic diversities of order zero are indicated for each tree.

envisioned by Lou Jost [5,8] have powerfully contributed to the generalization of Faith's [2] concept of phylogenetic diversity within a frame of parsimonious evolution [7]. Likely Jost's true diversities will also open new frontiers for characterizing diversity of metagenomic samples of molecular sequences and for using them as markers of the diversity of ecological communities, such as environmental metagenomic or microbiome samples. I believe that this will be achieved when true diversities will be combined with probabilistic models of sequence evolution and corresponding estimates of genetic relatedness [10].

Acknowledgment

This work is supported by NIH Grant 5R01GM87485-2.

References

1. Magurran AE and McGill BJ (2011) Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press, Oxford, UK.
2. Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61: 1-10.
3. Rao CR (1982) Diversity and dissimilarity coefficients: a unified approach. Theor Popul Biol 21: 24-43.
4. Allen B, Kon M, Bar-Yam Y (2009) A new phylogenetic diversity measure generalizing Shannon index and its application to phyllostomid bats. Am Nat 174: 236-243.
5. Jost L (2006) Entropy and diversity. OIKOS 113: 2.
6. Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54: 427-432.
7. Chao A, Chiu CH, Jost L (2010) Phylogenetic diversity measures based on Hill Numbers. Phil Trans R Soc B 365: 3599-3609.
8. Jost L (2007) Partitioning diversity into independent alpha and beta components. Ecology 88: 2427-2439.
9. Camin JH, Sokal RR (1965) A method for deducing branching sequences in phylogeny. Evolution 19: 31l-326.
10. Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56: 330-338.