entropart: An R Package to Measure and Partition Diversity

Eric Marcon
AgroParisTech
UMR EcoFoG

Bruno Hérault
Cirad
UMR EcoFoG

Abstract

entropart is a package for R designed to estimate diversity based on HCDT entropy or similarity-based entropy. It allows calculating neutral, phylogenetic and functional entropy and diversity, partitioning them and correcting them for estimation bias.

Keywords: biodiversity, entropy, partitioning.

1. Introduction

Diversity measurement can be done through a quite rigorous framework based on entropy, *i.e.* the amount of uncertainty calculated from the frequency distribution of a community (Patil and Taillie 1982; Jost 2006; Marcon, Scotti, Hérault, Rossi, and Lang 2014a). Tsallis entropy, also known as HCDT entropy (Havrda and Charvát 1967; Daróczy 1970; Tsallis 1988), is of particular interest (Jost 2006; Marcon *et al.* 2014a) namely because it includes the number of species, Shannon (1948) and Simpson (1949) indices of diversity into a single framework. Interpretation of entropy is not straightforward but one can easily transform into Hill numbers (Hill 1973) which have many desirable properties (Jost 2007): mainly, they are the number of equally-frequent species that would give the same level of diversity as the data.

Marcon and Hérault (2014) generalized the duality of entropy and diversity, deriving the relation between phylogenetic or functional diversity (Chao, Chiu, and Jost 2010) and phylogenetic or functional entropy (we will write *phylodiversity* and *phyloentropy* for short), as introduced by Pavoine, Love, and Bonsall (2009). Special cases are the well-known PD (Faith 1992) and FD (Petchey and Gaston 2002) indices and Rao's (1982) quadratic entropy. The same relation holds between Ricotta and Szeidl entropy of a community (Ricotta and Szeidl 2006) and similarity-based diversity (Leinster and Cobbold 2012).

The **entropart** package for R (R Development Core Team 2014) enables calculation of all these measures of diversity and entropy and their partitioning.

Diversity partitioning means that, in a given area, the γ diversity D_γ of all individuals found may be split into within (α diversity, D_α) and between (β diversity, D_β) local assemblages. α diversity reflects the diversity of individuals in local assemblages whereas β diversity reflects the diversity of the local assemblages. Marcon *et al.* (2014a) derived the decomposition of Tsallis γ entropy into its α and β components, generalized to phyloentropy by Marcon and Hérault (2014) and similarity-based diversity by Reeve, Matthews, Cobbold, Leinster,

Thompson, and Brummitt (2014) and Marcon, Zhang, and Hérault (2014b).

Estimators of diversity are biased because of unseen species and also because they are not linear functions of probabilities (Marcon *et al.* 2014a). α and γ diversities are underestimated by naive estimators (Chao and Shen 2003; Dauby and Hardy 2012). β diversity is severely biased too when sampling is not sufficient (Beck, Holloway, and Schwanghart 2013). Bias-corrected estimators of phylodiversity have been developed by Marcon and Hérault (2014). Estimators of similarity-based diversity were derived by Marcon *et al.* (2014b). The package includes them all.

The successive sections of this paper presents the package features, illustrated by worked examples based on the data included in the package.

2. Package organization

2.1. Data

Most functions of the package calculate entropy or diversity of a community or of a meta-community. Community functions accept a vector of probabilities or of abundances for species data. Each element of the vector contains the probability or the number of occurrences of a species in a given community. Meta-community functions require a particular data organization in a `MetaCommunity` object described here.

A `MetaCommunity` is basically a list. Its main components are `$Nsi`, a matrix containing the species abundances whose lines are species, columns are communities and `$Wi`, a vector containing community weights. Creating a `MetaCommunity` object is the purpose of the `MetaCommunity` function. Arguments are a dataframe containing the number of individuals per species (lines) in each community (columns), and a vector containing the community weights. The following example creates a `MetaCommunity` made of three communities of unequal weights with 4 species. The weighted average probabilities of occurrence of species and the total number of individuals define the meta-community as the assemblage of communities.

```
R> library("entropart")
R> (df <- data.frame(C1 = c(10, 10, 10, 10), C2 = c(0, 20,
+      35, 5), C3 = c(25, 15, 0, 2), row.names = c("sp1",
+      "sp2", "sp3", "sp4")))

    C1 C2 C3
sp1 10  0 25
sp2 10 20 15
sp3 10 35  0
sp4 10  5  2

R> w <- c(1, 2, 1)
R> MC <- MetaCommunity(Abundances = df, Weights = w)
```

A meta-community is partitioned into several local communities (indexed by $i = 1,\ 2, \ldots, I$). n_i individuals are sampled in community i. Let $s = 1,\ 2, \ldots, S$ denote the species that

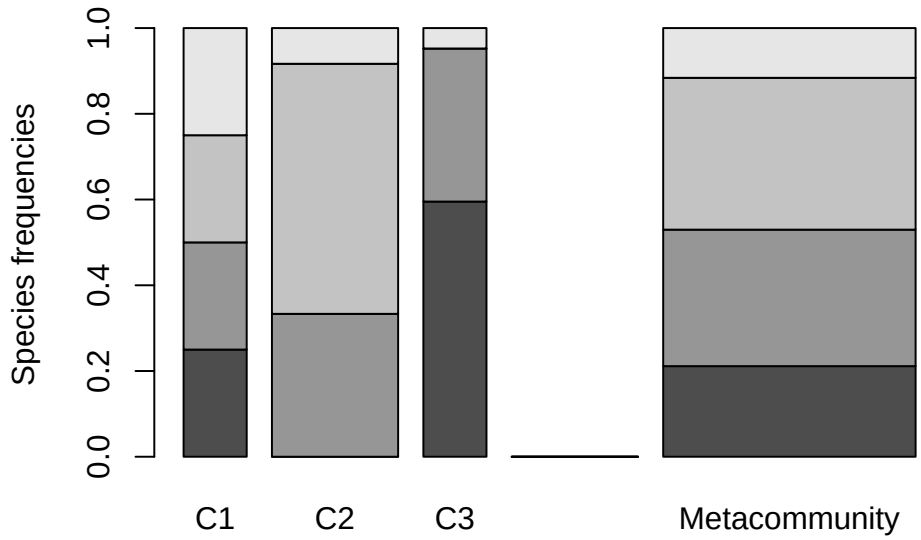

Figure 1. Plot of a `MetaCommunity`. Communities (named C1, C2 ad C3) are represented in the left part of the figure, the metacommunity to the right. Bar widths are proportional to community weights. Species abundances are represented vertically (4 species are present in the meta-community, only 3 of them in communities C2 and C3.

compose the meta-community, n_{si} the number of individuals of species s sampled in the local community i, $n_s = \sum_i n_{si}$ the total number of individuals of species s, $n = \sum_s \sum_i n_{si}$ the total number of sampled individuals. Within each community i, the probability p_{si} for an individual to belong to species s is estimated by $\hat{p}_{si} = n_{si}/n_i$. The same probability for the meta-community is p_s. Communities have a weight w_i, satisfying $p_s = \sum_i w_i p_{si}$. The commonly-used $w_i = n_i/n$ is a possible weight, but the weighting may be arbitrary (*e.g.* the sampled areas). The components of a `MetaCommunity` object satisfy these conditions: `$Ps` for example contains the probability of occurrence of each species in the meta-community:

```
R> MC$Ps
```

```
      sp1       sp2       sp3       sp4
0.2113095 0.3184524 0.3541667 0.1160714
```

A `MetaCommunity` can be summarized and plotted (Figure 1).

The package contains an example dataset containing the inventory of two 1-ha tropical forest plots in Paracou, French Guiana (Marcon, Hérault, Baraloto, and Lang 2012):

```
R> data("Paracou618")
R> summary(Paracou618.MC)
```

```
Meta-community (class 'MetaCommunity') made of  1124 individuals in 2
communities and 425 species.

Its sample coverage is  0.92266748426447
```

```
Community weights are:
[1] 0.5720641 0.4279359
Community sample numbers of individuals are:
P006 P018
 643  481
Community sample coverages are:
     P006      P018
0.8943859 0.8463782
```

`Paracou618.MC` is a meta-community made of two communities named "P006" and "P018", containing 425 species (their name is *Family_Genus_Species*, abbreviated to 4 characters). The values of the abundance matrix are the number of individuals of each species in each community. Sample coverage will be explained later.

The dataset also contains a taxonomy and a functional tree. `Paracou618.Taxonomy` is an object of class `phylog`, defined in **ade4** (Dray and Dufour 2007), namely a phylogenetic tree. This example data is only a taxonomy, containing family, genus and species levels for the sake of simplicity. `Paracou618.Functional` is an object of class `hclust` containing a functional tree based on leaf, height, stem and seed functional traits (Hérault and Honnay 2007; Marcon and Hérault 2014). The package accepts any ultrametric tree of class `phylog` or `hclust`. `Paracou618.dist` is the distance matrix (actually a `dist` object) used to build the functional tree.

2.2. Utilities

The deformed logarithm formalism (Tsallis 1994) is very convenient to manipulate entropies. The deformed logarithm of order q is defined as:

$$\ln_q x = \frac{x^{1-q} - 1}{1 - q} \tag{1}$$

It converges to ln when $q \to 1$, see figure 2.

The inverse function of $\ln_q x$ is the deformed exponential:

$$e_q^x = [1 + (1 - q)x]^{\frac{1}{1-q}} \tag{2}$$

Functions of the packages are `lnq(x, q)` and `expq(x, q)`.

3. Neutral diversity

3.1. Community functions

HCDT entropy

Neutral HCDT entropy of order q of a community is defined as:

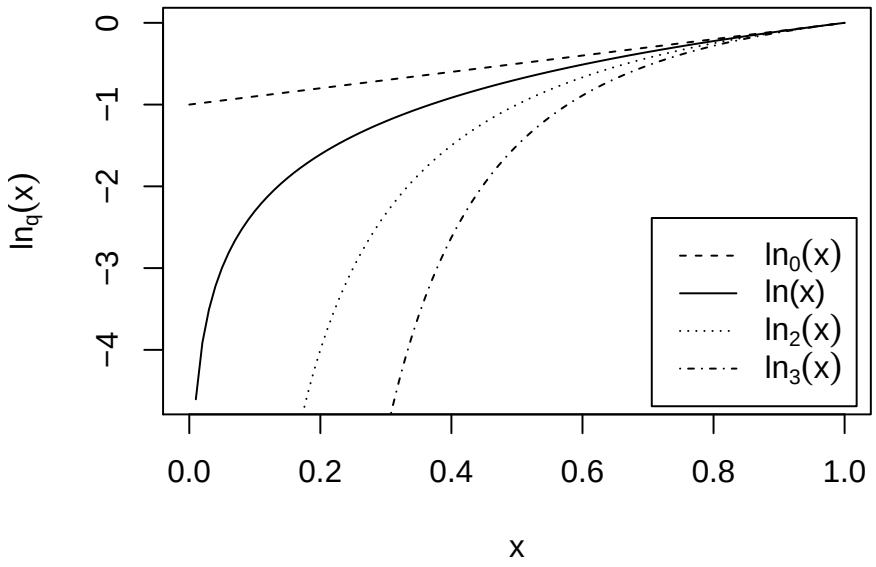

Figure 2. Curves of $\ln_q x$ for different values of q between 0 and 4 ($\ln_1 x = \ln x$).

$$ {}^qH = \frac{1 - \sum_s p_s^q}{q - 1} = -\sum_s p_s^q \ln_q p_s \tag{3} $$

`q` is the order of diversity (*e.g.*: 1 for Shannon). Entropy can be calculated by the **Tsallis** function. Paracou meta-community entropy of order 1 is:

```
R> Tsallis(Ps = Paracou618.MC$Ps, q = 1)
```

```
[1] 4.736023
```

For convenience, special cases of entropy of order q have a clear-name function: `Richness` for $q = 0$, `Shannon` for $q = 1$, `Simpson` for $q = 2$.

```
R> Shannon(Ps = Paracou618.MC$Ps)
```

```
[1] 4.736023
```

Sample coverage

A useful indicator of sampling quality is the sample coverage (Good 1953; Chao, Lee, and Chen 1988; Zhang and Huang 2007), that is to say the probability for a species of the community

to be observed in the actual sample. It equals the sum of the probability of occurrences of all observed species. Its historical estimator is (Good 1953):

$$\hat{C} = 1 - \frac{S^1}{n} \tag{4}$$

S^1 is the number of singletons (species observed once) of the sample, and n is its size. The estimator has been improved by taking into account the whole distribution of species (Zhang and Huang 2007). The `Coverage` function calculates it, allowing to choose the estimator (Zhang and Huang's by default):

```
R> Coverage(Ns = Paracou618.MC$Ns)

[1] 0.9220438
```

The sample coverage cannot be estimated from probability data: abundances are required.

Estimation-bias corrected estimators

Estimation-bias correction is used to improve the estimation of entropy despite unobserved species. Bias-corrected estimators (often relying on sample coverage) are returned by functions whose names are prefixed by `bc`, such as `bcTsallis`. They are similar to the non-corrected ones but they use abundance data and propose several bias-correction techniques to select in the `Correction` argument. A "Best" correction is calculated by default, detailed in the help file of each function.

```
R> bcTsallis(Ns = Paracou618.MC$Ns, q = 1)

[1] 4.898061
```

The best correction for Tsallis entropy follows Marcon *et al.* (2014a). `bcSimpson` returns Lande's correction (Lande 1996) and `bcShannon` returns the very efficient correction by Chao, Wang, and Jost (2013), so their results are different (and more accurate) than those of the general `bcTsallis` function.

```
R> bcShannon(Ns = Paracou618.MC$Ns)

[1] 4.892159
```

Effective numbers of species

Entropy should be converted into "true diversity" (Jost 2007), *i.e.* effective number of species equal to Hill (1973) numbers:

$${}^qD = \left(\sum_s p_s^q\right)^{\frac{1}{1-q}} \tag{5}$$

This can be done by the deformed exponential function, or using directly the `Diversity` or `bcDiversity` functions (equal to the deformed exponential of order q of `Tsallis` or `bcTsallis`)

```
R> expq(Simpson(Ps = Paracou618.MC$Ps), q = 2)

[1] 68.7215

R> Diversity(Ps = Paracou618.MC$Ps, q = 2)

[1] 68.7215

R> expq(bcTsallis(Ns = Paracou618.MC$Ns, q = 2), q = 2)

[1] 73.19676

R> bcDiversity(Ns = Paracou618.MC$Ns, q = 2)

[1] 73.19676
```

3.2. Meta-community functions

Meta-community functions allow partitioning diversity according to Patil and Taillie's concept of diversity of a mixture (Patil and Taillie 1982), *i.e.* α entropy of a meta-community is defined as the weighted average of community entropy, following Routledge (1979):

$$ {}^{q}H_{\alpha} = \sum_{i} w_i \, {}^{q}_{i}H_{\alpha} \tag{6} $$

${}^{q}_{i}H_{\alpha}$ is the entropy of community i:

$$ {}^{q}_{i}H_{\alpha} = \frac{1 - \sum_s p_{si}^q}{q - 1} = -\sum_s p_{si}^q \ln_q p_{si} \tag{7} $$

Jost's (2007) definition of α entropy is not supported explicitly in the package since it only allows partitioning of equally weighted communities. In this particular case, both definitions are identical.

γ entropy of the meta-community is defined as α entropy of a community. β entropy, the difference between γ and α, is the generalized Jensen-Shannon divergence between the species distribution of the meta-community and those of communities (Marcon *et al.* 2014a):

$$ {}^{q}H_{\beta} = {}^{q}H_{\gamma} - {}^{q}H_{\alpha} = \sum_s p_{si}^q \ln_q \frac{p_{si}}{p_s} \tag{8} $$

β entropy should be transformed into diversity, *i.e.* an effective number of communities:

$$ {}^{q}D_{\beta} = e_q^{\frac{{}^{q}H_{\beta}}{1-(q-1){}^{q}H_{\alpha}}} \tag{9} $$

Basic meta-community functions

These values can be estimated by the meta-community functions `AlphaEntropy`, `AlphaDiversity`, `BetaEntropy`, `BetaDiversity`. They accept a `Metacommunity` and an order of diversity q as arguments, and return an `MCentropy` or `MCdiversity` object which can be summarized and plotted. `GammaEntropy` and `GammaDiversity` return a number. Estimation-bias corrections are applied by default:

```
R> e <- AlphaEntropy(Paracou618.MC, q = 1)
R> summary(e)

Neutral alpha entropy of order 1 of metaCommunity Paracou618.MC
with correction: Best

Entropy of communities:
    P006     P018
4.403435 4.673620
Average entropy of the communities:
[1] 4.519057
```

Diversity Partition of a metacommunity

The `DivPart` function calculates everything at once. Its arguments are the same but bias correction is not applied by default. It can be, using the argument `Biased = FALSE`, and the correction chosen by the argument `Correction`. It returns a `DivPart` object which can be summarized (entropy is not printed by `summary`) and plotted:

```
R> p <- DivPart(q = 1, MC = Paracou618.MC, Biased = FALSE)
R> summary(p)

HCDT diversity partitioning of order 1 of metaCommunity Paracou618.MC
 with correction: Best
Alpha diversity of communities:
     P006      P018
 81.73115 107.08473
Total alpha diversity of the communities:
[1] 91.74905
Beta diversity of the communities:
[1] 1.460828
Gamma diversity of the metacommunity:
[1] 134.0296

R> p$CommunityAlphaEntropies

    P006     P018
4.403435 4.673620
```

Diversity Estimation of a metacommunity

The `DivEst` function decomposes diversity and estimates confidence interval of α, β and γ diversity following Marcon *et al.* (2012). If the observed species frequencies of a community are assumed to be a realization of a multinomial distribution: they can be drawn again to obtain a distribution of entropy.

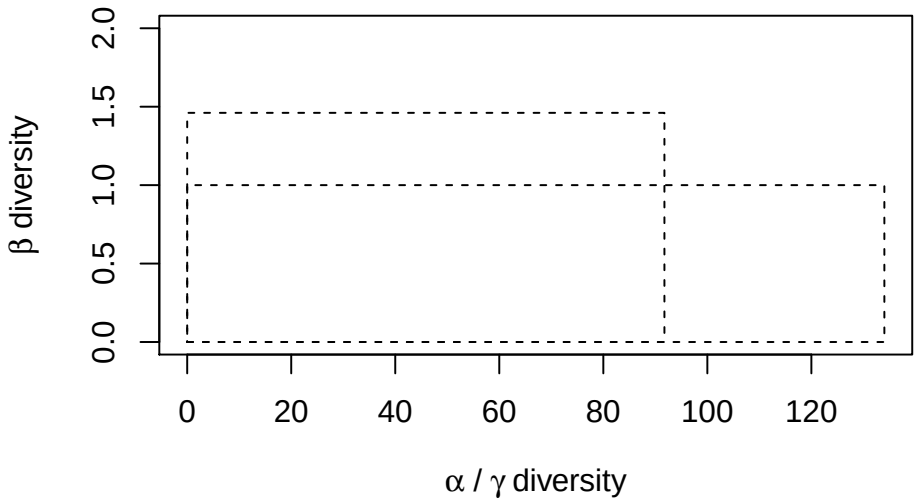

Figure 3. Plot of the diversity partition of the meta-community `Paracou618.MC`. The long rectangle of height 1 represents γ diversity, equal to 134 effective species. The narrower and higher rectangle has the same area: its horizontal size is α diversity (92 effective species) and its height is β diversity (1.46 effective communities).

```
R> de <- DivEst(q = 1, Paracou618.MC, Biased = FALSE, Correction = "Best",
+      Simulations = 1000)

======================================================================

R> summary(de)

Diversity partitioning of order 1 of MetaCommunity MC
 with correction: Best
Alpha diversity of communities:
     P006      P018
 81.73115 107.08473
Total alpha diversity of the communities:
[1] 91.74905
Beta diversity of the communities:
[1] 1.460828
Gamma diversity of the metacommunity:
[1] 134.0296
Quantiles of simulations (alpha, beta and gamma diviersity):
       0%       10%       50%       10%       25%       50%       75%
 78.78268  87.19635  91.74626  87.19635  89.46648  91.74626  94.26293
      90%       95%       99%      100%
 96.35708  97.72680 100.21950 103.65521
      0%      10%      50%      10%      25%      50%      75%
1.381186 1.431527 1.460950 1.431527 1.444351 1.460950 1.477487
```

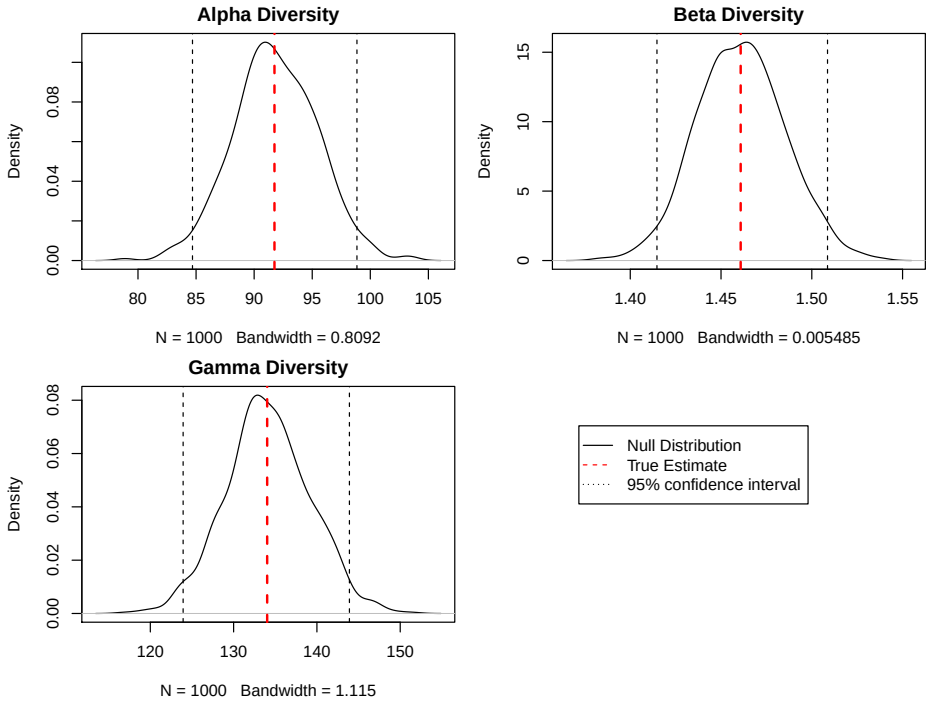

Figure 4. Plot of the diversity estimation of the meta-community `Paracou618.MC`. α, β and γ diversity probability densities are plotted, with a 95% confidence interval.

```
     90%      95%      99%     100%
1.492361 1.501829 1.518393 1.538482
      0%      10%      50%      10%      25%      50%      75%
116.7834 127.7285 133.9441 127.7285 130.8865 133.9441 137.4975
     90%      95%      99%     100%
140.7890 142.4859 146.4895 151.5443
```

The result is a `Divest` object which can be summarized and plotted (Figure 4).

Diversity Profile of a metacommunity

`DivProfile` calculates diversity profiles, *i.e.* the value of diversity against its order (Figure 5). The result is a `DivProfile` object which can be summarized and plotted.

```
R> dp <- DivProfile(seq(0, 2, 0.2), Paracou618.MC, Biased = FALSE)
R> summary(dp)

Diversity profile of MetaCommunity MC
 with correction: Best
Diversity against its order:
      Order Alpha Diversity Beta Diversity Gamma Diversity
 [1,]   0.0       205.84226       1.441996       296.82368
 [2,]   0.2       181.63811       1.424471       258.73825
 [3,]   0.4       157.35277       1.413780       222.46224
```

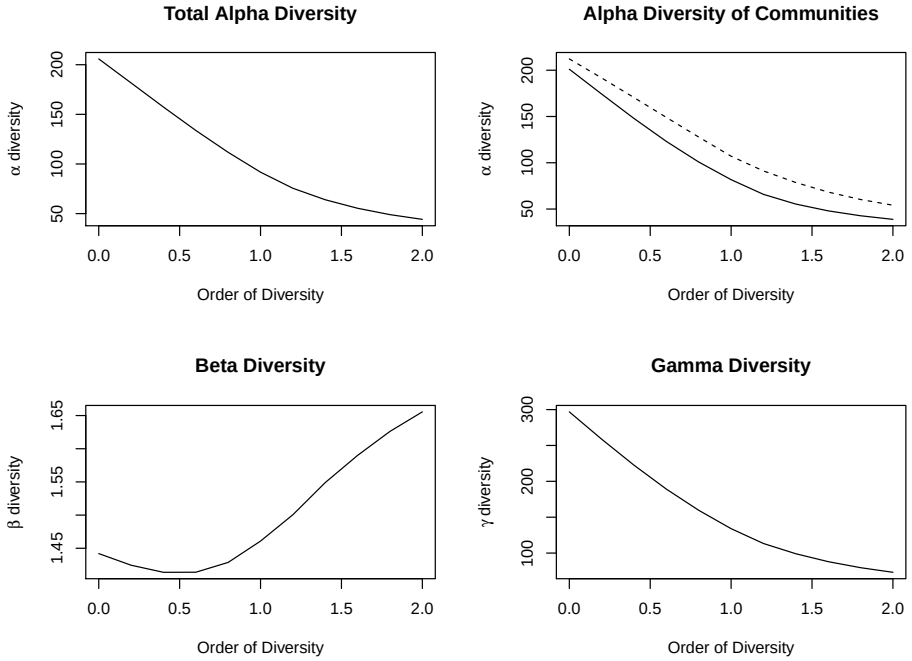

Figure 5. Diversity profile of the meta-community `Paracou618.MC`. Values are the number of effective species (α and γ diversity) and the effective number of communities (β diversity). α and γ diversity decrease from $q = 0$ (number of species) to $q = 2$ (Simpson diversity) by construction.

```
 [4,]   0.6       133.77507       1.413903       189.14504
 [5,]   0.8       111.70847       1.428705       159.59848
 [6,]   1.0        91.74905       1.460828       134.02961
 [7,]   1.2        75.51773       1.500587       113.32093
 [8,]   1.4        63.95522       1.549024        99.06819
 [9,]   1.6        55.37376       1.590012        88.04495
[10,]   1.8        48.97244       1.626123        79.63520
[11,]   2.0        44.21244       1.655569        73.19676
```

Alternative functions

Beta entropy can also be calculated by a set of functions named after the community functions, such as `TsallisBeta`, `bcTsallisBeta`, `SimpsonBeta`, etc. which require two vectors of abundances or probabilities instead of a `MetaCommunity` object: that of the community and the expected one (usually that of the meta-community). Bias correction is currently limited to Chao and Shen's correction. The example below calculates the Shannon β entropy of the first community of Paracou618 and the meta-community.

```
R> ShannonBeta(Paracou618.MC$Psi[, 1], Paracou618.MC$Ps)

[1] 0.3499358
```

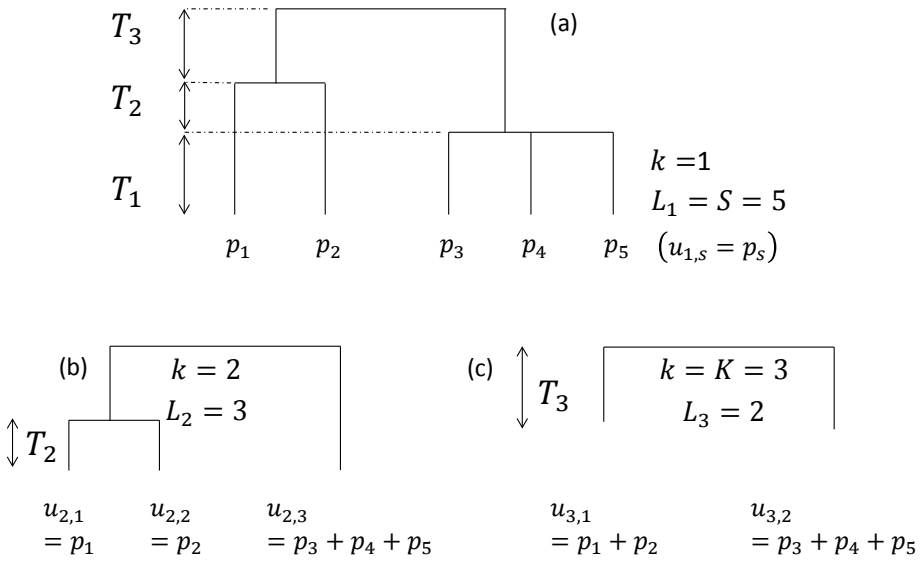

Figure 6. Hypothetical ultrametric tree. (a) The whole tree contains three slices, delimited by two nodes. The length of slices is T_k. (b) Focus on slice 2. The tree without slice 1 is reduced to 3 leaves. Frequencies of collapsed species are $u_{k,l}$. (c) Slice 3 only.

These functions are available for particular uses, when a `MetaCommunity` is not available or not convenient to use (*e.g.* simulations). Meta-community functions are preferred for current use.

4. Phylogenetic diversity

Phylogenetic or functional diversity generalizes HCDT diversity, considering the divergence between species (Marcon and Hérault 2014). Here, all species take place in an ultrametric phylogenetic or functional tree (Figure 6). The tree is cut into slices, delimited by two nodes. The first slice starts at the bottom of the tree and ends at the first node. In slice k, L_k leaves are found. The probabilities of occurrence of the species belonging to branches that were below leaf l in the original tree are summed to give the grouped probability $u_{k,l}$. HCDT entropy can be calculated in slice k:

$$ {}^q_k H = -\sum_l u_{k,l}^q \ln_q u_{k,l} \tag{10} $$

Then, it is summed over the tree slices. Phyloentropy can be normalized or not. We normalize it so that it does not depend on the tree height:

$$ {}^q\overline{H}(T) = \sum_{k=1}^{K} \frac{T_k}{T} {}^q_k H \tag{11} $$

Unnormalized values are multiplied by the tree height, such as ${}^qPD(T)$ (Chao *et al.* 2010).

Phyloentropy is calculated as HCDT entropy along the slices of the trees applying possible estimation-bias corrections, summed, possibly normalized, and finally transformed into diversity:

$$ {}^q\overline{D}\left(T\right) = e_q^{{}^q\overline{H}(T)} \tag{12}$$

4.1. Community functions

`PhyloEntropy` and the estimation-bias-corrected `bcPhyloEntropy` are the phylogenetic analogs of `Tsallis` and `bcTsallis`. They accept the same arguments plus an ultrametric tree of class `hclust` or `phylog`, and `Normalize`, a boolean to normalize the tree height to 1 (by default).

Phylogenetic diversity is calculated by `PhyloDiversity` or `bcPhyloDiversity`, analogous to the neutral diversity functions `Diversity` and `bcDiversity`.

Results are either a `PhyloDiversity` or a `PhyloEntropy` object, which can be plotted (Figure 7) and summarized.

```
R> phd <- bcPhyloDiversity(Paracou618.MC$Ns, q = 1, Tree = Paracou618.Taxonomy,
+      Normalize = TRUE)
R> summary(phd)

alpha or gamma phylogenetic or functional diversity of order 1
of distribution Paracou618.MC$Ns
 with correction: Best
Phylogenetic or functional diversity was calculated according to the tree
Paracou618.Taxonomy

Diversity is  normalized

Diversity equals: 55.13383
```

The `AllenH` function is close to `PhyloEntropy`: it also calculates phyloentropy but the algorithm is that of Allen, Kon, and Bar-Yam (2009) for $q = 1$ and that of Leinster and Cobbold (2012) for $q \neq 1$. It is much faster since it does not require calculating entropy for each slice of the tree but it does not allow estimation-bias correction. `ChaoPD` calculates phylodiversity according to Chao *et al.* (2010), with the same advantages and limits compared to `PhyloDiversity`.

For convenience, `PDFD` and `Rao` functions are provided to calculate unnormalized phyloentropy of order 0 and 2.

4.2. Meta-community functions

`DivPart`, `DivEst` and `DivProfile` functions return phylogenetic entropy and diversity values instead of neutral ones if a tree is provided in the arguments.

```
R> dp <- DivPart(q = 1, Paracou618.MC, Biased = FALSE, Correction = "Best",
+      Tree = Paracou618.Taxonomy)
R> summary(dp)
```

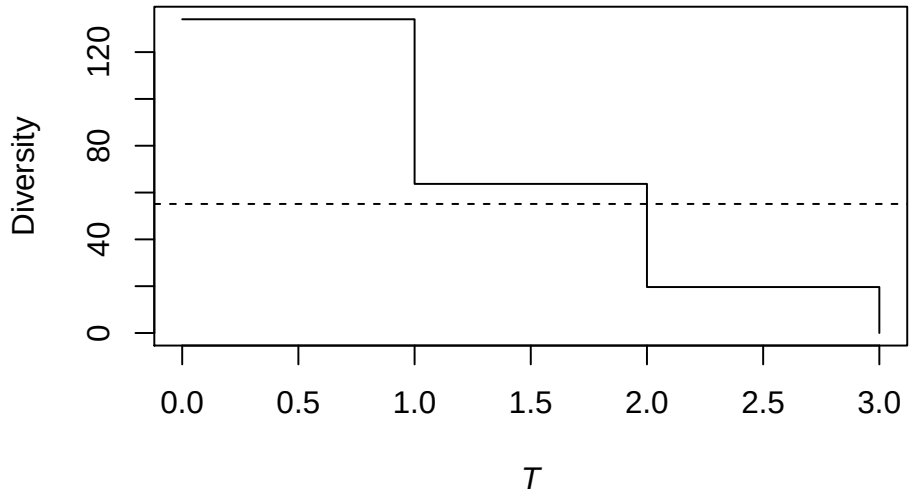

Figure 7. Plot of the γ phylodiversity estimation of the meta-community `Paracou618.MC`. The effective number of taxa of Shannon diversity is plotted against the distance from the leaves of the phylogenetic tree. Here, the tree is based on a rough taxonomy, so diversity of species, genera and families are the three levels of the curve.

```
HCDT diversity partitioning of order 1 of metaCommunity Paracou618.MC
 with correction: Best
Phylogenetic or functional diversity was calculated according to the tree
Paracou618.Taxonomy

Diversity is  normalized

Alpha diversity of communities:
    P006     P018
37.22132 51.31045
Total alpha diversity of the communities:
[1] 42.70238
Beta diversity of the communities:
[1] 1.291119
Gamma diversity of the metacommunity:
[1] 55.13383
```

Other meta-community functions, such as `AlphaEntropy` behave the same way:

```
R> summary(BetaEntropy(Paracou618.MC, q = 2, Tree = Paracou618.Taxonomy,
+      Correction = "None", Normalize = FALSE))

HCDT beta entropy of order 2 of metaCommunity Paracou618.MC
with correction: None
```

```
Phylogenetic or functional entropy was calculated according to the tree
Paracou618.Taxonomy

Entropy is not normalized
Entropy of communities:
      P006       P018
0.04117053 0.02325883
Average entropy of the communities:
[1] 0.03350547
```

Compare with Rao's `divc` computed by **ade4**:

```
R> library("ade4")
R> divc(as.data.frame(Paracou618.MC$Wi), disc(as.data.frame(Paracou618.MC$Nsi),
+      Paracou618.Taxonomy$Wdist))

                  diversity
Paracou618.MC$Wi 0.03350547
```

5. Similarity-based diversity

Leinster and Cobbold (2012) introduced similarity-based diversity of a community ${}^qD^Z$. A matrix $\mathbf{Z}$ describes the similarity between pairs of species, defined between 0 and 1. A species ordinariness is its average similarity with all species (weighted by species frequencies), including similarity with itself (equal to 1). Similarity-based diversity is the reciprocal of the generalized average of order q (Hardy, Littlewood, and Pólya 1952) of the community species ordinariness.

The `Dqz` function calculates similarity-based diversity. Its arguments are the vector of probabilities of occurrences of the species, the order of diversity and the similarity matrix $\mathbf{Z}$. The `bcDqz` function allows estimation-bias correction.

This example calculates the γ diversity of the meta-community Paracou. First, the similarity matrix is calculated from the distance matrix between all pairs of species as 1 minus normalized dissimilarity.

```
R> DistanceMatrix <- as.matrix(Paracou618.dist)
R> Z <- 1 - DistanceMatrix/max(DistanceMatrix)
R> bcDqz(Paracou618.MC$Ns, q = 2, Z)

[1] 1.483027
```

If $\mathbf{Z}$ is the identity matrix, similarity-based diversity equals HCDT diversity:

```
R> Dqz(Paracou618.MC$Ps, q = 2, Z = diag(length(Paracou618.MC$Ps)))

[1] 68.7215
```

```
R> Diversity(Paracou618.MC$Ps, q = 2)

[1] 68.7215
```

The similarity-based entropy of a community ${}^qH^Z$ (Leinster and Cobbold 2012; Ricotta and Szeidl 2006) has the same relations with diversity as HCDT entropy and Hill numbers. The `Hqz` function calculates it:

```
R> Hqz(Paracou618.MC$Ps, q = 2, Z)

[1] 0.3208152

R> lnq(Dqz(Paracou618.MC$Ps, q = 2, Z), q = 2)

[1] 0.3208152
```

All meta-community functions can be used to estimate similarity-based diversity: argument `Z` must be provided:

```
R> e <- AlphaEntropy(Paracou618.MC, q = 1, Z = Z)
R> summary(e)

Similarity-based alpha entropy of order 1 of metaCommunity
Paracou618.MC with correction: Best

Phylogenetic or functional entropy was calculated according to the similarity matrix
Z

Entropy of communities:
     P006      P018
0.3945541 0.3934725
Average entropy of the communities:
[1] 0.3940912
```

6. Advanced tools

The package comes with a set of tools to realize frequents tasks: running Monte-Carlo simulations on a community, quickly calculate its diversity profile, applying a function to a species distribution along a tree, and manipulation of meta-communities.

6.1. Entropy of Monte-Carlo simulated communities

The `EntropyCI` function is a versatile tool to simplify these simulations. Its arguments are an entropy function (any entropy function of the package accepting a vector of species abundances, such as `bcTsallis`), the number of simulations to run and the observed species

frequencies. The result is a numeric vector containing the entropy value of each simulated community. Entropy can be finally transformed into diversity but it is not correct to use a diversity function in simulations because the average simulated value must be calculated (and only entropy can be averaged).

This example shows how to use the function. First, the distribution of the γ HCDT entropy of order 2 (Simpson entropy) of Paracou meta-community is calculated and transformed into diversity. Then, the actual diversity is calculated and completed by the 95% confidence interval of the simulated values.

```
R> SimulatedDiversity <- expq(EntropyCI(FUN = bcTsallis,
+      Simulations = 1000, Ns = Paracou618.MC$Ns, q = 2),
+      q = 2)
```

```
======================================================================
```

```
R> bcDiversity(Paracou618.MC$Ns, q = 2)
```

```
[1] 73.19676
```

```
R> quantile(SimulatedDiversity, probs = c(0.025, 0.975))
```

```
     2.5%     97.5%
 63.72941 102.84703
```

6.2. Diversity or Entropy Profile of a community

This function is used to calculate diversity or entropy profiles based on community functions such as `Tsallis` or `ChaoPD`. It is similar to `DivProfile` but does not require a `Metacommunity` for argument. It returns a list which can be plotted, see the help page of the function for an example.

6.3. Applying a Function over a Phylogenetic Tree

The `PhyloApply` function is used to apply an entropy community function (generally `bcTsallis`) along a tree.

6.4. Manipulation of meta-communities

Several meta-communities, combined in a list, can be merged two different ways: the `MergeMC` function simplifies hierarchical partitioning of diversity: it creates a new meta-community whose communities are the original meta-communities aggregated data. The α entropy of the new meta-community is the weighted average γ entropy of the original meta-communities.

`MergeC` combines the communities of several meta-communities to create a single meta-community containing them all. Last, `ShuffleMC` randomly shuffles communities accross meta-communities to allow simulations to test differences between meta-communities.

7. Conclusion

The entropart package allows estimating biodiversity according to the framework based on HCDT entropy, the correction of its estimation-bias (Grassberger 1988; Chao and Shen 2003) and its transformation into equivalent numbers of species (Hill 1973; Jost 2006; Marcon *et al.* 2014a). Phylogenetic or functional diversity (Marcon and Hérault 2014) can be estimated, considering phyloentropy as the average neutral diversity over slices of a phylogenetic or functional tree (Pavoine *et al.* 2009). Similarity-based diversity Leinster and Cobbold (2012) can be used to estimate (Marcon *et al.* 2014b) functional diversity from a similartiy or dissimilarity matrix between species without requiring building a dendrogram and thus preserving the topology of species.

We believe it is a complete toolbox for ecologists who need to estimate the diversity of actual, undersampled communities and to partition it.

8. Acknowledgments

This work has benefited from an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-0025).

References

Allen B, Kon M, Bar-Yam Y (2009). "A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats." *American Naturalist*, **174**(2), 236–243.

Beck J, Holloway JD, Schwanghart W (2013). "Undersampling and the Measurement of Beta Diversity." *Methods in Ecology and Evolution*, **4**(4), 370–382.

Chao A, Chiu CH, Jost L (2010). "Phylogenetic Diversity Measures Based on Hill Numbers." *Philosophical Transactions of the Royal Society B*, **365**(1558), 3599–609.

Chao A, Lee SM, Chen TC (1988). "A generalized Good's Nonparametric Coverage Estimator." *Chinese Journal of Mathematics*, **16**, 189–199.

Chao A, Shen TJ (2003). "Nonparametric Estimation of Shannon's Index of Diversity When There Are Unseen Species in Sample." *Environmental and Ecological Statistics*, **10**(4), 429–443.

Chao A, Wang YT, Jost L (2013). "Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species." *Methods in Ecology and Evolution*, **4**(11), 1091–1100.

Daróczy Z (1970). "Generalized Information Functions." *Information and Control*, **16**(1), 36–51.

Dauby G, Hardy OJ (2012). "Sampled-Based Estimation of Diversity Sensu Stricto by Transforming Hurlbert Diversities into Effective Number of Species." *Ecography*, **35**(7), 661–672.

Dray S, Dufour AB (2007). "The ade4 Package: Implementing the Duality Diagram for Ecologists." *Journal of Statistical Software*, **22**(4), 1–20.

Faith DP (1992). "Conservation Evaluation and Phylogenetic Diversity." *Biological Conservation*, **61**(1), 1–10.

Good IJ (1953). "On the Population Frequency of Species and the Estimation of Population Parameters." *Biometrika*, **40**(3/4), 237–264.

Grassberger P (1988). "Finite Sample Corrections to Entropy and Dimension Estimates." *Physics Letters A*, **128**(6–7), 369–373.

Hardy G, Littlewood J, Pólya G (1952). *Inequalities*. Cambridge University Press.

Havrda J, Charvát F (1967). "Quantification Method of Classification Processes. Concept of Structural alpha-Entropy." *Kybernetika*, **3**(1), 30–35.

Hill MO (1973). "Diversity and Evenness: A Unifying Notation and Its Consequences." *Ecology*, **54**(2), 427–432.

Hérault B, Honnay O (2007). "Using Life-History Traits to Achieve a Functional Classification of habitats." *Applied Vegetation Science*, **10**(1), 73–80.

Jost L (2006). "Entropy and Diversity." *Oikos*, **113**(2), 363–375.

Jost L (2007). "Partitioning Diversity into Independent Alpha and Beta Components." *Ecology*, **88**(10), 2427–2439.

Lande R (1996). "Statistics and Partitioning of Species Diversity, and Similarity Among Multiple Communities." *Oikos*, **76**, 5–13.

Leinster T, Cobbold C (2012). "Measuring Diversity: the Importance of Species Similarity." *Ecology*, **93**(3), 477–489.

Marcon E, Hérault B (2014). "Decomposing Phylodiversity." *HAL*, **hal-00946177**(version 1), 1–22.

Marcon E, Hérault B, Baraloto C, Lang G (2012). "The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity." *Oikos*, **121**(4), 516–522.

Marcon E, Scotti I, Hérault B, Rossi V, Lang G (2014a). "Generalization of the Partitioning of Shannon Diversity." *PLOS One*, **9**(3), e90289.

Marcon E, Zhang Z, Hérault B (2014b). "The Decomposition of Similarity-Based Diversity and its Bias Correction." *HAL*, **hal-00989454**(version 1), 1–12.

Patil GP, Taillie C (1982). "Diversity as a Concept and its Measurement." *Journal of the American Statistical Association*, **77**(379), 548–561.

Pavoine S, Love MS, Bonsall MB (2009). "Hierarchical Partitioning of Evolutionary and Ecological Patterns in the Organization of Phylogenetically-Structured Species Assemblages: Application to Rockfish (Genus: Sebastes) in the Southern California Bight." *Ecology Letters*, **12**(9), 898–908.

Petchey OL, Gaston KJ (2002). "Functional Diversity (FD), Species Richness and Community Composition." *Ecology Letters*, **5**, 402–411.

R Development Core Team (2014). "R: A Language and Environment for Statistical Computing."

Rao C (1982). "Diversity and Dissimilarity Coefficients: a Unified Approach." *Theoretical Population Biology*, **21**, 24–43.

Reeve R, Matthews L, Cobbold C, Leinster T, Thompson J, Brummitt N (2014). "How to partition diversity." *arXiv*, **1404.6520**(v1), 1–9. 1404.6520.

Ricotta C, Szeidl L (2006). "Towards a Unifying Approach to Diversity Measures: Bridging the Gap Between the Shannon Entropy and Rao's Quadratic Index." *Theoretical Population Biology*, **70**(3), 237–243.

Routledge R (1979). "Diversity Indices: Which Ones are Admissible?" *Journal of Theoretical Biology*, **76**(4), 503–515.

Shannon CE (1948). "A Mathematical Theory of Communication." *The Bell System Technical Journal*, **27**, 379–423, 623–656.

Simpson EH (1949). "Measurement of Diversity." *Nature*, **163**(4148), 688.

Tsallis C (1988). "Possible Generalization of Boltzmann-Gibbs Statistics." *Journal of Statistical Physics*, **52**(1), 479–487.

Tsallis C (1994). "What are the Numbers that Experiments Provide?" *Química Nova*, **17**(6), 468–471.

Zhang Z, Huang H (2007). "Turing's Formula Revisited." *Journal of Quantitative Linguistics*, **14**(2-3), 222–241.

Affiliation:

Eric Marcon
AgroParisTech
Campus agronomique, BP 316
97310 Kourou, French Guiana
E-mail: eric.marcon@ecofog.gf

Bruno Hérault
Cirad
Campus agronomique, BP 316
97310 Kourou, French Guiana
E-mail: bruno.herault@ecofog.gf