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 species-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 gathers 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 it 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 (2015) 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 phylentropy 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’s entropy of a community (Ricotta and Szeidl 2006) and similarity-based diversity (Leinster and Cobbold 2012).

The entropart package for R (R Core Team 2015) enables calculation of all these measures of diversity and entropy and their partitioning.

Diversity partitioning means that, in a given area, the γ diversity $D_\gamma$ of all individuals found may be split into within (α diversity, $D_\alpha$) and between (β diversity, $D_\beta$) 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 $\gamma$ entropy into its $\alpha$ and $\beta$ components, generalized to phylodiversity (Marcon and Hérault 2015) and similarity-based diversity (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). $\alpha$ and $\gamma$ diversities are underestimated by naive estimators (Chao and Shen 2003; Dauby and Hardy 2012). $\beta$ 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 (2015). Estimators of similarity-based diversity were derived by Marcon et al. (2014b). The package includes them all (Marcon and Hérault 2015).

In summary, the framework supported by the package is as follows. First, an information function is chosen to describe the amount of surprise brought by the observation of each individual. In the simplest case of species-neutral diversity, it is just a decreasing function of probability: observing an individual of a rarer species brings more surprise. Various information functions allow evaluating species-neutral, phylogenetic or functional entropy. Surprise is averaged among all individuals of a community to obtain its entropy. Entropy is systematically transformed into diversity for interpretation. Diversity is an effective number of species, \textit{i.e.} the number of equally-different and equally-frequent species that would give the same entropy as the data. The average entropy of communities of an assemblage is $\alpha$ entropy, while the entropy of the assemblage is $\gamma$ entropy. Their difference is $\beta$ entropy. After transformation, $\beta$ diversity is the ratio of $\gamma$ to $\alpha$ diversity. It is an effective number of communities, \textit{i.e.} the number of equally-weighted communities with no species in common (and where species are maximally distinct between communities) necessary to obtain the same diversity as the data. Estimation-bias correction is more easily applied to entropy before transforming it into diversity.

This framework is somehow different from that of Chao, Chiu, and Jost (2014) who define $\alpha$ diversity in another way (see Marcon and Hérault 2015, for a detailed comparison), such that $\alpha$ entropy is not the average surprise of an assemblage. They also propose a definition of functional diversity (Chiu and Chao 2014) based in the information brought by pairs of individuals that is not supported in the package.

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 an assemblage of communities called 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 $\$Ns_i$, a matrix containing the species abundances whose lines are species, columns are communities and $\$W_i$, a vector containing community weights. Creating a MetaCommunity object is the purpose of the MetaCommunity function. Arguments are a data frame containing the number of individuals
Figure 1. Plot of a MetaCommunity. Communities (named C1, C2 and 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.)

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")
(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
w <- c(1, 2, 1)
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 compose the meta-community, \( n_{s,i} \) the number of individuals of species \( s \) sampled in the local community \( i \), \( n_s = \sum_i n_{s,i} \) the total number of individuals of species \( s \), \( n = \sum_s \sum_i n_{s,i} \) the total number of sampled individuals. Within each community \( i \), the probability \( p_{s,i} \) for an individual to belong to species \( s \) is estimated by \( \hat{p}_{s,i} = n_{s,i}/n_i \). The same probability for the meta-community is \( p_s \). Communities have a weight \( w_i \), satisfying \( p_s = \sum_i w_ip_{s,i} \). The commonly-used \( w_i = n_i/n \) is a possible weight, but the weighting may be arbitrary (e.g. the sampled areas). The component \$Ps \) of a MetaCommunity object contains the probability of occurrence of each species in the meta-community, calculated this way:

```r
MC$Ps
```

```r
##     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, Héralt, Baraloto, and Lang 2012):

```r
data("Paracou618")
summary(Paracou618.MC)
```

---

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 (Héralt and Honnay 2007; Marcon and Héralt 2015). 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.

Numeric vectors containing species abundances (such as the $Ns component of MetaCommunity) or probabilities (such as $Ps) may be converted to abundance vectors (AbdVector) or probability vectors (ProbaVector) to clarify their content. By default, the as.AbdVector function transforms abundance values into integer if they are not (the $Ns components of a MetaCommunity is typically not an integer vector if community weights are not proportional to their numbers of individuals):

```r
data("Paracou618")
PAbd <- as.AbdVector(Paracou618.MC$Ns)
```

The as.ProbaVector function transforms abundances to probabilities if necessary:

```r
PProba <- as.ProbaVector(Paracou618.MC$Ps)
```

AbdVector and ProbaVector objects both are SpeciesDistribution objects which can be plotted (Figure 2).

### 2.2. Utilities

The deformed logarithm formalism (Tsallis 1994) is very convenient to manipulate entropies.
Figure 2. Plot of a ProbaVector. Rank-Abundance curve (Whittaker plot) of the Paracou species distribution.

Figure 3. Curves of $\ln_q x$ for different values of $q$ between 0 and 4 ($\ln_1 x = \ln x$).

The deformed logarithm of order $q$ is defined as:

$$\ln_q x = \frac{x^{1-q} - 1}{1 - q}$$  \hspace{1cm} (1)$$

It converges to $\ln$ when $q \to 1$, see figure 3.

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

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

Functions of the packages are $\text{lnq}(x, q)$ and $\text{expq}(x, q)$. 
3. Species-neutral diversity

3.1. Community functions

**HCDT entropy**

Species-neutral HCDT entropy of order \( q \) of a community is defined as:

\[
qH = \frac{1 - \sum sp_s^q}{q - 1} = -\sum sp_s^q \ln p_s = \sum s p_s \ln \frac{1}{p_s}
\]  

where \( q \) 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:

```
Tsallis(PProba, q = 1)
## None
## 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 \( q = 2 \).

```
Shannon(PProba)
## None
## 4.736023
```

Entropy values have no intuitive interpretation in general, except for the number of species \( 0H \) and Simpson entropy \( 2H \) which is the probability for two randomly chosen individuals to belong to different species.

**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):

\[
\hat{C} = 1 - \frac{S^1}{n}
\]  

where \( S^1 \) 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):

```
Coverage(PAbd)
## ZhangHuang
## 0.9226675
```
The sample coverage cannot be estimated from probability data: abundances are required. Its interpretation is straightforward: some species have not been sampled. Their number is unknown but their total probability of occurrence can be estimated accurately. Here, it is a bit less than 8%. From another point of view, the probability for an individual of the community to belong to a sampled species is \( C \): 8% of them belong to missed species. The number of missed species may be estimated by \textbf{Richness} but this is not the point here. The sample coverage is the foundation of many estimators of entropy.

\textit{Bias corrected estimators}

Correction of estimation bias is used to improve the estimation of entropy despite unobserved species and also mathematical issues (Bonachela, Hinrichsen, and Muñoz 2008). Bias-corrected estimators (often relying on sample coverage) are returned by functions whose names are prefixed by \texttt{bc}, such as \texttt{bcTsallis}. They are similar to the non-corrected ones but they use abundance data and propose several bias-correction techniques to select in the \texttt{Correction} argument. A “Best” correction is calculated by default, detailed in the help file of each function.

\begin{verbatim}
bcTsallis(PAbd, q = 1)
\end{verbatim}

The best correction for Tsallis entropy follows Chao and Jost (2015). It combines an unbiased estimator previously derived by Zhang and Grabchak (2014) and an estimate of the remaining bias.

All community functions such as \texttt{Tsallis} are actually generic methods that can handle several types of data the appropriate way: if the first argument of the function is a \texttt{ProbaVector} (or a numeric vector summing to 1), no bias correction is applied. If it is an \texttt{AbdVector} (or an integer vector), the bias-corrected estimator is used (\textit{e.g.} \texttt{bcTsallis}). Numeric vectors summing to more than 2 are considered as abundances but some bias corrections do not allow non-integer values, so they will round them with a warning.

The different ways to use the functions are a matter of personal preference. \texttt{bcTsallis} is equivalent to \texttt{Tsallis} with an abundance vector:

\begin{verbatim}
Tsallis(PAbd, q = 1)
\end{verbatim}

whilst \texttt{Tsallis} with a probability vector does not allow bias correction:

\begin{verbatim}
Tsallis(PProba, q = 1)
\end{verbatim}

Bias-corrected entropy is ready to be transformed into explicit diversity.
Effective numbers of species

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

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

```
expq(Tsallis(PAbd, q = 2), q = 2)
## ChaoWangJost
## 73.13163
Diversity(PAbd, q = 2)
## ChaoWangJost
## 73.13163
```

The effective number of species of the Paracou dataset is estimated to be 73 after bias correction (rather than 69 without it). It means that a community made of 73 equally-frequent species has the same Simpson entropy as the actual one. This is much less than the actual 425 sampled species. Simpson’s entropy focuses on dominant species.

Hurlbert’s diversity

Hurlbert’s index of diversity (Hurlbert 1971) of order $k$ is the expected number of species observed in a sample of size $k$.

$$kS = \sum_s \left[ 1 - (1 - p_s)^k \right]$$  \hspace{1cm} (6)

Greater values of $k$ give more importance to rare species.

An unbiased estimator of $kS$ has been provided by Hurlbert, for values of $k$ up to the sample size $n$:

$$\hat{kS} = \sum_s \left[ 1 - \binom{n - n_s}{k} / \binom{n}{k} \right]$$  \hspace{1cm} (7)

The effective number of species $kD$ can be found by solving the following equation (Dauby and Hardy 2012):

$$kS = kD \left[ 1 - \left( 1 - \frac{1}{kD} \right)^k \right]$$  \hspace{1cm} (8)

Hurlbert’s index is calculated by the Hurlbert function. Its unbiased estimator is obtained by bcHurlbert (implicitly if an abundance vector is used). Its effective number of species is calculated by HurlbertD or bcHurlbertD.
Hurlbert’s diversity of order 2 is identical to Simpson’s diversity.

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. $\alpha$ entropy of a meta-community is defined as the weighted average of community entropy, following Routledge (1979):

$$qH_\alpha = \sum_i w_i q_i H_\alpha$$

$q_i H_\alpha$ is the entropy of community $i$:

$$q_i H_\alpha = \frac{1 - \sum_s p_{s,i}^q}{q - 1} = -\sum_s p_{s,i}^q \ln_q p_{s,i} = \sum_s p_{s,i} \ln_q \frac{1}{p_{s,i}}$$

Jost’s (2007) definition of $\alpha$ 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.

$\gamma$ entropy of the meta-community is defined as $\alpha$ entropy of a community. $\beta$ entropy, the difference between $\gamma$ and $\alpha$, is the generalized Jensen-Shannon divergence between the species distribution of the meta-community and those of communities (Marcon et al. 2014a):

$$qH_\beta = qH_\gamma - qH_\alpha = \sum_s p_{s,i}^q \ln_q \frac{p_{s,i}}{p_s} = \sum_s p_{s,i} \ln_q \frac{p_s}{p_{s,i}}$$

$\beta$ entropy should be transformed into diversity, i.e. an effective number of communities:

$$qD_\beta = e^{\frac{-qH_\beta}{q(1-q)}}$$

Basic meta-community functions

These values can be estimated by the meta-community functions named 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)
summary(e)
```

The Shannon $\alpha$ entropy of the meta-community is 4.52. It is the weighted average entropy of communities.

### 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)
summary(p)
```

The $\alpha$ diversity of communities is 92 effective species (it is the exponential of the entropy calculated previously). This is more than Simpson diversity (73 species, calculated above) because less frequent species are taken into account. $\gamma$ diversity of the meta-community is 133 effective species. $\beta$ diversity is 1.45 effective communities, i.e. the two actual communities are as different from each other as 1.45 ones with equal weights and no species in common.

### Diversity Estimation of a metacommunity

The `DivEst` function decomposes diversity and estimates confidence interval of $\alpha$, $\beta$ and $\gamma$
Figure 4. Plot of the diversity partition of the meta-community Paracou618.MC. The long rectangle of height 1 represents $\gamma$ diversity, equal to 133 effective species. The narrower and higher rectangle has the same area: its horizontal size is $\alpha$ diversity (92 effective species) and its height is $\beta$ diversity (1.45 effective communities).

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.

de <- DivEst(q = 1, Paracou618.MC, Biased = FALSE, Correction = "Best", Simulations = 100)

## ============================================================
summary(de)
## Diversity partitioning of order 1 of MetaCommunity
## Paracou618.MC
## with correction: Best
## Alpha diversity of communities:
## P006 P018
## 80.20125 110.66525
## Total alpha diversity of the communities:
## [1] 92.04908
## Beta diversity of the communities:
## ChaoWangJost
## 1.447498
## Gamma diversity of the metacommunity:
## ChaoWangJost
## 133.2409
## Quantiles of simulations (alpha, beta and gamma diversity):
## 0% 1% 2.5% 5% 10% 25% 50% 75% 90% 95% 97.5% 99% 100%
## 82.11812 84.84627 85.46065 86.00815 87.91901 90.25764 92.39559 94.21335 95.95217 97.35151 97.62791 97.94198 101.32800
## 0% 1% 2.5% 5% 10% 25% 50% 75% 90% 95% 97.5% 99%
## 1.391390 1.394757 1.397429 1.404654 1.414269 1.428899 1.448780 1.463647 1.477888 1.487461 1.494408 1.497410
## 100%
## 1.502463
## 0% 1% 2.5% 5% 10% 25% 50% 75% 90% 95% 97.5% 99%
## 118.8323 123.3996 125.8570 126.6803 127.4761 129.8273 97.35151 97.62791 97.94198 101.32800
## 0% 1% 2.5% 5% 10% 25% 50% 75% 90% 95% 97.5% 99%
##
Figure 5. Plot of the diversity estimation of the meta-community Paracou618.MC. \( \alpha \), \( \beta \) and \( \gamma \) diversity probability densities are plotted, with a 95% confidence interval.

The result is a Divest object which can be summarized and plotted (Figure 5). The uncertainty of estimation is due to sampling: the distribution of the estimators corresponds to the simulated repetitions of sampling in the original multinomial distribution of species. It ignores the remaining bias of the estimator, which is unknown. Yet, except for \( q = 2 \), the corrected estimators are biased (even though much less than the non-corrected ones), especially when \( q \) is small. New estimators to reduce the bias are included in the package regularly.

**Diversity Profile of a metacommunity**

DivProfile calculates diversity profiles, *i.e.* the value of diversity against its order (Figure 6). The result is a DivProfile object which can be summarized and plotted.

```r
dp <- DivProfile(seq(0, 2, 0.2), Paracou618.MC, Biased = FALSE, NumberOfSimulations = 10)
```

```r
summary(dp)
```

```
## Diversity profile of MetaCommunity Paracou618.MC
```
Figure 6. Diversity profile of the meta-community Paracou618.MC. Values are the number of effective species ($\alpha$ and $\gamma$ diversity) and the effective number of communities ($\beta$ diversity). Community P006 is represented by the solid line and community P018 by the dotted line. $\alpha$ and $\gamma$ diversity decrease from $q = 0$ (number of species) to $q = 2$ (Simpson diversity) by construction.

<table>
<thead>
<tr>
<th>Order</th>
<th>Alpha Diversity</th>
<th>Beta Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChaoWangJost 0.0</td>
<td>238.36351</td>
<td>1.321238</td>
</tr>
<tr>
<td>ChaoWangJost 0.2</td>
<td>204.17972</td>
<td>1.323271</td>
</tr>
<tr>
<td>ChaoWangJost 0.4</td>
<td>170.72117</td>
<td>1.335708</td>
</tr>
<tr>
<td>ChaoWangJost 0.6</td>
<td>139.89024</td>
<td>1.361171</td>
</tr>
<tr>
<td>ChaoWangJost 0.8</td>
<td>113.37058</td>
<td>1.399674</td>
</tr>
<tr>
<td>ChaoWangJost 1.0</td>
<td>92.04908</td>
<td>1.447498</td>
</tr>
<tr>
<td>ChaoWangJost 1.2</td>
<td>75.81819</td>
<td>1.498557</td>
</tr>
<tr>
<td>ChaoWangJost 1.4</td>
<td>63.87372</td>
<td>1.546905</td>
</tr>
<tr>
<td>ChaoWangJost 1.6</td>
<td>55.18979</td>
<td>1.588960</td>
</tr>
<tr>
<td>ChaoWangJost 1.8</td>
<td>48.84485</td>
<td>1.623736</td>
</tr>
<tr>
<td>ChaoWangJost 2.0</td>
<td>44.13559</td>
<td>1.651781</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gamma Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChaoWangJost 314.93464</td>
</tr>
<tr>
<td>ChaoWangJost 270.18504</td>
</tr>
<tr>
<td>ChaoWangJost 228.03372</td>
</tr>
<tr>
<td>ChaoWangJost 190.41457</td>
</tr>
<tr>
<td>ChaoWangJost 158.68180</td>
</tr>
<tr>
<td>ChaoWangJost 133.24089</td>
</tr>
<tr>
<td>ChaoWangJost 113.62013</td>
</tr>
<tr>
<td>ChaoWangJost 98.80658</td>
</tr>
<tr>
<td>ChaoWangJost 87.69438</td>
</tr>
<tr>
<td>ChaoWangJost 79.31112</td>
</tr>
<tr>
<td>ChaoWangJost 72.90234</td>
</tr>
</tbody>
</table>
**Figure 7.** 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.

Small orders of diversity give more weight to rare species. P018 can be considered more diverse than P006 because their profiles (Figure 6, top right) do not cross (Tothmeresz 1995): its diversity is systematically higher. The shape of the $\beta$ diversity profile shows that the communities are more diverse when their dominant species are considered.

The bootstrap confidence intervals of the values of diversity (Marcon et al. 2012, 2014a) are calculated if `NumberOfSimulations` is not 0.

**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 $\beta$ entropy of the first community of Paracou618 and the meta-community.

```
ShannonBeta(Paracou618.MC$Psi[, 1], PProba)
## None
## 0.3499358
```

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 in general.

**4. Phylogenetic diversity**
Phylogenetic or functional diversity generalizes HCDT diversity, considering the distance between species (Marcon and Hérald 2015). Here, all species take place in an ultrametric phylogenetic or functional tree (Figure 7). 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 $k$:

$$q_k H = - \sum_l u_{k,l}^q \ln u_{k,l} \quad (13)$$

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:

$$q \Pi(T) = \sum_{k=1}^K \frac{T_k}{T} q_k H \quad (14)$$

Unnormalized values are multiplied by the height of the tree, 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:

$$qD(T) = e^{q \Pi(T)} \quad (15)$$

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 phylo, 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 species-neutral diversity functions Diversity and bcDiversity.

Results are either a PhyloDiversity or a PhyloEntropy object, which can be plotted (Figure 8) and summarized.

```
phd <- bcPhyloDiversity(PAbd, q = 1, Tree = Paracou618.Taxonomy, Normalize = TRUE)
summary(phd)
```

## alpha or gamma
## phylogenetic or functional diversity of order 1
## of distribution PAbd
## with correction: Best
## Phylogenetic or functional diversity was calculated according to the tree
## Paracou618.Taxonomy
## Diversity is normalized
## Diversity equals: 54.78324
Figure 8. Plot of the $\gamma$ 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. The dotted line represents the value of phylodiversity.

The phylogenetic diversity of order 1 of the Paracou dataset is 55 effective species: 55 totally different species (only connected by the root of the tree) with equal probabilities would have the same entropy. It can be compared to its species-neutral diversity: 133 species. The latter is the diversity of the first slice of the tree. When going up the tree, diversity decreases because species collapse. On Figure 8, the diversity of the second slice, between $T = 1$ and $T = 2$, is that of genera (63 effective genera) and the last slice contains 20 effective families. The phylogenetic entropy of the community is the average of the entropy along slices, weighted by the slice lengths. Diversity cannot be averaged the same way.

A less trivial phylogeny would contain many slices, resulting in as many diversity levels with respect to $T$.

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 species-neutral ones if a tree is provided in the arguments.

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

```r
## HCUT 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: 36.66108
- P018: 51.98951

Total alpha diversity of the communities:

- [1] 42.87235

Beta diversity of the communities:

- [1] 1.286827

Gamma diversity of the metacommunity:

- [1] 54.78324

The decomposition is interpreted as the species-neutral one: $\gamma$ diversity is 55 effective species, made of 1.3 effective communities of 43 effective species.

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))
```

---

5. Similarity-based diversity

Leinster and Cobbold (2012) introduced similarity-based diversity of a community $qD^Z$. A matrix $Z$ describes the similarity between pairs of species, defined between 0 and 1. A species ordinarity 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 ordinarity.

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 $Z$. The bcDqz function allows estimation-bias correction (Marcon et al. 2014b).
This example calculates the $\gamma$ 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)
Z <- 1 - DistanceMatrix/max(DistanceMatrix)
bcDqz(PAbd, q = 2, Z)
## Best
## 1.48295
```

If $Z$ is the identity matrix, similarity-based diversity equals HCDT diversity:

```r
Dqz(PProba, q = 2, Z = diag(length(PProba)))
## None
## 68.7215
```

Functional diversity of order 2 is only 1.48 effective species, which is very small compared to 69 effective species for Simpson diversity. 1.48 equally-frequent species with similarity equal to 0 would have the same functional diversity as the actual community (made of 425 species). This means that species are very similar from a functional point of view. The very low values returned by $qD^Z$ are questioned by Chao et al. (2014) and discussed in depth by Marcon et al. (2014b): the choice of the similarity matrix is not trivial.

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 \texttt{Hqz} function calculates it:

```r
Hqz(PProba, q = 2, Z)
## None
## 0.3208152
```

As species-neutral entropy, $qH^Z$ has no straightforward interpretation beyond the average surprise of a community.

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)
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:

<table>
<thead>
<tr>
<th>Species</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>P006</td>
<td>0.3945541</td>
</tr>
<tr>
<td>P018</td>
<td>0.3934725</td>
</tr>
</tbody>
</table>

Average entropy of the communities:

<table>
<thead>
<tr>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3940912</td>
</tr>
</tbody>
</table>

The $\alpha$ functional entropy of the meta-community is the average entropy of communities.

### 6. Advanced tools

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

#### 6.1. Random communities

The `rCommunity` function allows creating random communities. Their species probability distribution can be drawn in a well-known distribution (such as a log-normal one) or obtained from the data, just by dividing abundances by the total number of individuals (Marcon et al. 2012), or derived from a more sophisticated model by Chao and Jost (2015). Finally, the specified number of communities are drawn in a multinomial distribution.

The log-normal (Preston 1948), the log-series (Fisher, Corbet, and Williams 1943), the geometric (Motomura 1932), and the broken-stick (MacArthur 1957) distributions can be simulated.

This example code draws a single community of 1000 individuals according to a log-normal distribution with 300 species. Many species are not observed in the 1000-individual sample: the observed number of species is shown, with an estimation of the actual number (which should be 300). The simulated community is plotted, with its log-normal distribution fitted from the data (Figure 9). Estimated parameters can be compared to the original ones.

```r
NsRef <- rCommunity(n = 1, size = 1000, S = 300, Distribution = "lnorm", sd = 1)
Richness(as.ProbaVector(NsRef))
```

<table>
<thead>
<tr>
<th>Richness</th>
</tr>
</thead>
<tbody>
<tr>
<td>235</td>
</tr>
</tbody>
</table>

```r
Richness(NsRef)
```

| Chao1 |
| 270.2588 |

```r
plot(NsRef, Distribution = "lnorm")
```

| $\mu$ |
| 1.070049 |

| $\sigma$ |
| 0.8542127 |
6.2. Entropy of Monte-Carlo simulated communities

The `EntropyCI` function is a versatile tool to simplify simulations. Simulated communities are obtained by random draws in a multinomial distribution of species and their entropy is calculated. The arguments of `EntropyCI` 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, the observed species frequencies and the method to obtain probabilities for the multinomial distribution (the same as that of `rCommunity`).

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 $\gamma$ HCDT entropy of order 1 (Shannon entropy) of the 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 = Tsallis, Simulations = 100, Ns = PAbd, q = 1), q = 1)

## ============================================================
## Diversity(PAbd, q = 1)
## ChaoWangJost
## 133.2409
## quantile(SimulatedDiversity, probs = c(0.025, 0.975))
## 2.5% 97.5%
## 124.9170 142.9578
```

These results are identical to those of the `DivEst` function but a single community can be addressed (`DivEst` requires a `MetaCommunity`).
6.3. 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 can compute a bootstrap confidence envelope of the estimation of the profile, like EntropyCI. It returns a CommunityProfile object which can be plotted. Profiles can be added to an existing plot by the CEnvelope function.

This example evaluates bias correction on the diversity profile of the Paracou dataset. First, diversity profiles are calculated with and without bias correction. The corrected profile is calculated with its confidence envelope:

```r
bcProfile <- CommunityProfile(Diversity, PAbd, NumberOfSimulations = 10)
Profile <- CommunityProfile(Diversity, PProba)
```

Then, they can be plotted altogether (Figure 10):

```r
plot(bcProfile)
CEnvelope(Profile, lty = 3)
legend("topright", c("Bias Corrected", "Biased"), lty = c(1, 3), inset = 0.02)
```

6.4. Applying a Function over a Phylogenetic Tree

The PhyloApply function is used to apply an entropy community function (generally bcTsallis) along a tree, the same way lapply works with a list.

This example shows how to calculate Shannon entropy along the tree containing the taxonomy to obtain species, genus and family entropy shown on figure 8:

```r
pa <- PhyloApply(Tree = Paracou618.Taxonomy, FUN = bcTsallis,
                 NorP = PAbd)
summary(pa)
```

## bcTsallis applied to PAbd along the tree:
## Paracou618.Taxonomy

## Results are normalized

## The average value is: 4.003384

## Values along the tree are:
## 1 2 3
## 4.892159 4.144688 2.973306

\[ \exp(pa\$Cuts) \]

## 1 2 3
## 133.24089 63.09794 19.55647

\[ \exp(pa\$Total) \]

## [1] 54.78324

### 6.5. 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 considers the aggregated data of each meta-community as a community and builds an upper-level meta-community with them. The $\alpha$ entropy of the new meta-community is the weighted average $\gamma$ entropy of the original meta-communities.

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

This example shows how to do this. A first meta-community is created, weights of communities are proportional to their number of individuals:

```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")))
```

```r
w <- colSums(df)
MC1 <- MetaCommunity(Abundances = df, Weights = w)
```

Then a second one:

```r
(df <- data.frame(C1 = c(10, 4), C2 = c(3, 4), row.names = c("sp1", "sp5")))
```

```r
w <- colSums(df)
MC2 <- MetaCommunity(Abundances = df, Weights = w)
```
They can be merged to obtain a single meta-community containing all original communities:

```r
mergedMC1 <- MergeC(list(MC1, MC2))
mergedMC1$Nsi
```

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sp1</td>
<td>10</td>
<td>0</td>
<td>25</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>sp2</td>
<td>10</td>
<td>20</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sp3</td>
<td>10</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sp4</td>
<td>10</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sp5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

They can also be merged considering each of them as a community of a higher-level meta-community:

```r
mergedMC2 <- MergeMC(list(MC1, MC2), Weights = sapply(list(MC1, MC2), function(x) (x$N)))
mergedMC2$Nsi
```

<table>
<thead>
<tr>
<th></th>
<th>MC1</th>
<th>MC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp1</td>
<td>35</td>
<td>13</td>
</tr>
<tr>
<td>sp2</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>sp3</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>sp4</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>sp5</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

Hierarchical diversity partitioning can then be achieved:

```r
dpAll <- DivPart(q = 1, MC = mergedMC2)
summary(dpAll)
```

```
## HCDT diversity partitioning of order 1 of metaCommunity mergedMC2
##
## Alpha diversity of communities:
## MC1  MC2
## 3.772161 1.943574
## Total alpha diversity of the communities:
## [1] 3.463277
## Beta diversity of the communities:
## None
## Gamma diversity of the metacommunity:
## None
## 4.281826
```

The $\gamma$ diversity of the top assemblage (MC1 and MC2) is 4.28 effective species, made of 1.24 effective meta-communities of 3.46 effective species. The $\alpha$ diversity of each meta-community of the top assemblage is their $\gamma$ diversity when it is partitioned in turn:

```r
dpMC1 <- DivPart(q = 1, MC = MC1)
summary(dpMC1)
```

```
## HCDT diversity partitioning of order 1 of metaCommunity MC1
##
## Alpha diversity of communities:
## C1  C2  C3
## 4.000000 2.429521 2.273918
## Total alpha diversity of the communities:
## [1] 2.741671
```

```r
dpMC2 <- DivPart(q = 1, MC = MC2)
summary(dpMC2)
```

```
## HCDT diversity partitioning of order 1 of metaCommunity MC2
##
## Alpha diversity of communities:
## C1  C2  C3
## 4.000000 2.429521 2.273918
## Total alpha diversity of the communities:
## [1] 2.741671
```
The $\gamma$ diversity of MC1 is 3.77 effective species, made of 1.38 effective meta-communities of 2.74 effective species. The same decomposition can be done for MC2.

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; Chao and Jost 2015) and its transformation into equivalent numbers of species (Hill 1973; Jost 2006; Marcon et al. 2014a). Phylogenetic or functional diversity (Marcon and Hérault 2015) can be estimated, considering phyloentropy as the average species-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 similarity or dissimilarity matrix between species without requiring building a dendrogram and thus preserving the topology of species (Pavoine, Ollier, and Dufour 2005; Podani and Schmera 2007).

The classical diversity estimators (Shannon and Simpson entropy) can be found in many R packages. vegetarian (Charney and Record 2009) allows calculating Hill numbers and partitioning them according to Jost’s framework. Bias correction is never available except in the EntropyEstimation (Cao and Grabchak 2014) package which provides the Zhang and Grabchak’s estimators of entropy and diversity and their asymptotic variance (not included in entropart). Phylodiversity and similarity-based diversity are not available in any package as far as we know. So we believe entropart is a useful toolbox for ecologists who need to estimate the diversity of actual, undersampled communities and to partition it.

References


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


**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