Latest version (September 2017)

User's Guide for Online program iDIP (Information-based Diversity Partitioning)

Anne Chao and Chun-Huo Chiu

Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043

Overview

iDIP (Information-based Diversity Partitioning) is an R-based interactive program accessible via the link https://chao.shinyapps.io/iDIP/ or Anne Chao's website http://chao.stat.nthu.edu.tw/wordpress/software_download/. Clicking either of these links will take you to the online interface window. Users do not need to learn/understand R to run iDIP Online. The interactive web application was built using Shiny (a web application framework).

Program iDIP allows you to decompose information-based diversity under a specified multi-level hierarchical structure. There are two analyses:

(1) Shannon Diversity: this analysis is used to decompose Shannon diversity (the exponential of Shannon entropy) across multiple levels given specified raw or relative species/allele (rows) abundances in each population/community (columns).

(2) Phylogenetic Diversity: this analysis is used to decompose information-based phylogenetic diversity across multiple levels given specified raw or relative species/allele (rows) abundances in each population/community (columns) and a specified phylogenetic tree (in Newick tree format) spanned by all observed species. Users are referred to Chao et al. (2010, 2014) for pertinent background information.

For each of the two analyses, Program iDIP can be applied to any arbitrary number of hierarchical levels. The output consists of a basic data summary and decomposition results, with the latter including (1) gamma (or total) diversity; alpha and beta diversity at each level; (2) proportion of total beta information found at each level; (3) mean differentiation (dissimilarity) among aggregates at each level. Illustrative examples (based on demo data included in the tool) and their corresponding output interpretations are provided in the User's Guide. Full details about the methodology in the context of a real-world data analysis can be found in the paper by Gaggiotti et al. (2017).

How to cite

If you publish your work based on results from iDIP Online, in addition to the methodology paper by Gaggiotti et al. (2017), you should use the following reference to cite iDIP Online:

> Chao, A. and Chiu, C.-H. (2017) iDIP (Information-based Diversity Partitioning) Online: Software for partitioning Shannon diversity and phylogenetic diversity under multi-level hierarchical structures. Program and User's Guide published at http://chao.stat.nthu.edu.tw/wordpress/software_download/.

Data

Data input files

(1) Abundance Data Matrix: specifying raw or relative species/allele (in rows) abundances or frequencies in each population/community (in columns).

iDIP Online supports two kinds of data: raw or relative species/allele abundances. When raw species/allele abundances (e.g., number of individuals or copies of alleles) are uploaded, size-weights are used by default, i.e., the proportion of the total number of individuals/alleles in a population/community is taken to be the weight of that population/community.

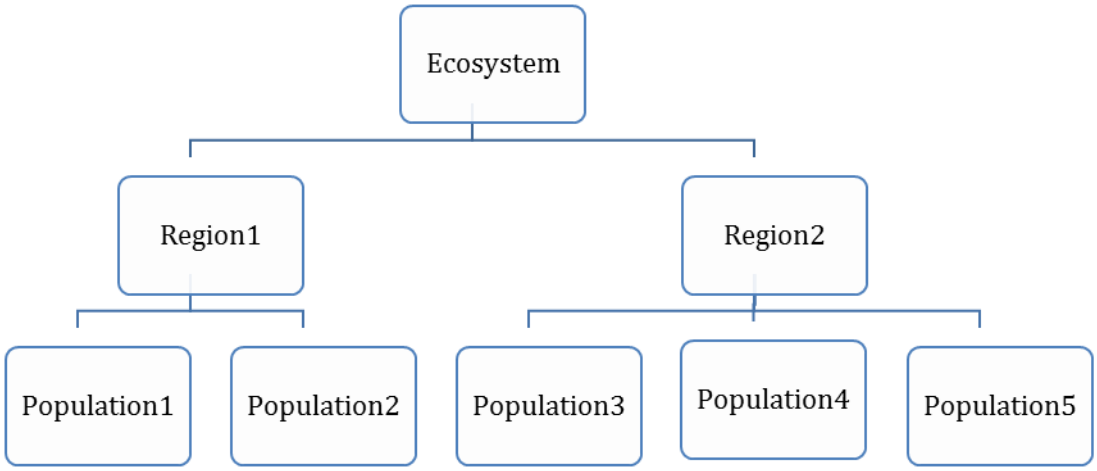
When relative species/allele abundances are uploaded, the "size" for each population/community is unity, leading to equal-weights for all populations or communities. Typically, the two weight functions yield very close results.

(2) Structure Matrix: specifying hierarchical structure matrix; see the simple example below.

(3) Phylogenetic Tree (in Newick tree format): specifying a phylogenetic tree spanned by all observed species or alleles in the study.

A simple example

We use a three-level hierarchical structure to illustrate how to prepare input data files. Consider an ecosystem consisting of two regions (1 and 2). In Region 1, there are two populations, and in Region 2, there are three populations. The hierarchical structure is displayed as follows:



Suppose the raw allele abundances or frequencies are given in the following matrix (alleles in rows and populations in columns). Then the raw abundance input matrix should be arranged as follows:

Table 1. Raw abundance data in the demo data set

	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5
Allele1	1	16	2	10	15
Allele2	0	0	0	5	14
Allele3	7	12	11	1	0
Allele4	0	5	14	1	21
Allele5	2	1	0	11	10
Allele6	0	1	3	2	0

Note that the first row of the uploaded abundance data must specify the labels or names for each population or community. These labels will appear in the output under “Data summary”. Species/allele names are optional for Shannon diversity analysis. If species or allele names are included, they must be placed in the first column as in Table 1. Species or allele names are required for phylogenetic diversity analysis and must match those names in the uploaded phylogenetic tree.

If equal-weights are used for all communities/populations, then you should input within-population relative allele abundance data instead of raw abundance data (for which size-weights are used). In the equal-weight case, the input file for the above example should be arranged as follows:

Table 2. Within-population relative allele abundance data in the demo data set

	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5
Allele1	0.1	0.46	0.06	0.33	0.25
Allele2	0	0	0	0.17	0.23
Allele3	0.7	0.34	0.37	0.03	0
Allele4	0	0.14	0.47	0.03	0.35
Allele5	0.2	0.03	0	0.37	0.17
Allele6	0	0.03	0.1	0.07	0

The hierarchical structure matrix for this simple example should be input as a matrix with levels in rows and populations in columns: (The first row = ecosystem level; the second row = regional level; the third row = population level.) The hierarchical structure of any arbitrary number of levels can be expressed in a similar manner.

Table 3. Hierarchical structure matrix in the demo data set

	Ecosystem	Ecosystem	Ecosystem	Ecosystem	Ecosystem
	Region1	Region1	Region2	Region2	Region2
	Population1	Population2	Population3	Population4	Population5

Based on our demo data for phylogenetic diversity decomposition, a simulated phylogenetic tree for the 6 alleles is described below in Newick tree format: (Species or allele names in the uploaded tree must match those in the abundance data matrix.)

Table 4. Phylogenetic tree of the 6 alleles from the demo data set

```
(((Allele1:16.66254448,Allele2:28.86156926):43.70264926,Allele3:59.19367445):43.4
9065302,(Allele4:9.67060281,Allele5:49.65919121,Allele6:15.361314):54.92297125);
```

Running procedures: three steps

Step 1. Select an analysis type (Shannon Diversity or Phylogenetic Diversity) from the top menu of the iDIP Online window.

Step 2.

(2a) If you select Shannon Diversity, then choose either demo data (raw-abundance data or relative-abundance data) or upload your own data. The latter choice requires that the following two matrices be uploaded as txt files: (i) the abundance data matrix specifying raw or relative species/allele (rows) abundances in each population/community (columns). You can include species names as row names, though it is not necessary to do so for Shannon diversity analysis. (ii) You also need to upload a matrix specifying a hierarchical structure; see the simple example given in the iDIP User's guide for how to input a hierarchical structure.

(2b) If you select Phylogenetic Diversity, then choose either demo data (raw-abundance data or relative-abundance data) or upload your own data. For the latter, three matrices should be uploaded as txt files. In addition to the abundance data matrix and hierarchical structure matrix, as described in Step (2a), a phylogenetic tree (in Newick tree format) spanned by all observed species or alleles must be uploaded. Note that for decomposing phylogenetic diversity, species/allele names (row names) must be included in the abundance data matrix and the species/allele names must match those in the uploaded phylogenetic tree.

Step 3. Press the Run! button. The output is then shown under "Data Summary" and "Decomposition analysis" along the second row menu.

Output and Interpretation

Shannon Diversity Analysis

If the raw abundance matrix in Table 1 and the hierarchical structure in Table 3 are uploaded, you will see the output as shown below. In this case, each population is weighted by its size (n in the output).

(1) In the "Data Summary" tab panel, the output includes the plot of the underlying hierarchical structure and basic data information. The latter part is shown below. You can click "save as csv file" at the bottom of the output to download the output.

	n	S.obs	f1	f2	entropy	diversity
Population1	10	3	1	1	0.8	2.23
Population2	35	5	2	0	1.21	3.34
Population3	30	4	0	1	1.13	3.11
Population4	30	6	2	1	1.44	4.22
Population5	60	4	0	0	1.35	3.87

Notes:

- n = number of observed individuals or allele copies in the sample.
- S.obs = number of observed species or alleles in the sample.
- f1 = number of singletons (species or alleles represented by exactly one individual or one copy).
- f2 = number of doubletons (species or alleles represented by exactly two individuals or copies).
- entropy = empirical (observed) Shannon entropy.
- diversity = empirical (observed) value of exponential of Shannon entropy.

(2) In the "Decomposition Analysis" tab panel, the numerical values for gamma diversity, alpha and beta components at each level, proportion of total beta information found at each level, and mean differentiation (dissimilarity) among sampling units for each level are shown. You can click "save as csv file" at the bottom of the output to download the output.

The output for the demo data is given below.

D_gamma	5.27
D_alpha.2	4.68
D_alpha.1	3.54
D_beta.2	1.13
D_beta.1	1.32
Proportion.2	0.30
Proportion.1	0.70
Differentiation.2	0.20
Differentiation.1	0.31

Notes:

- D_gamma = total diversity in the study area.
- D_alpha.i = alpha diversity at Level i.
- D_beta.i = beta diversity at Level i.
- Proportion.i = total beta information found at Level i.
- Differentiation.i = mean differentiation or dissimilarity among samples at Level i.
- (See the User's Guide for details).

We give simple interpretations for the above output ("effective" is used here in a sense of there being equally abundant alleles/populations/regions.)

(1a) D_gamma = 5.27 is interpreted as the effective number of alleles in the ecosystem (total diversity) being 5.27.

(1b) D_alpha.2 = 4.68 is interpreted as each region containing 4.68 allele equivalents;

D-beta.2 = 1.13 implies that there are 1.13 region equivalents. Thus, 4.68 x 1.13 = 5.27 (=D_gamma).

(1c) D_alpha.1 = 3.54 is interpreted as each population within a region containing 3.54 allele equivalents;

D_beta.1 = 1.32 implies that there are 1.32 population equivalents per region.

Here 1.32 x 3.54 = 4.68 species per region (= D_alpha.2).

(2) Proportion.2 = 0.30 means that the proportion of total beta information found at the regional level is 30%.
Proportion.1 = 0.70 means that the proportion of total beta information found at the population level is 70%.

(3) Differentiation.2 = 0.20 implies that the mean differentiation/dissimilarity among regions is 0.20. This can be interpreted in the following effective sense: the mean proportion of non-shared alleles in a region is around 20%.
Differentiation.1 = 0.31 implies that the mean differentiation/dissimilarity among populations within a region is 0.31, i.e., the mean proportion of non-shared alleles in a population is around 31%.

If the relative abundance matrix in Table 2 and the hierarchical structure in Table 3 are uploaded, the output (including only the observed species/alleles, entropy and diversity) for "Data Summary" is the same as that given above for raw abundances. However the output for "decomposition analysis" given below is (generally slightly) different because all populations are equally weighted for relative abundance data.

D_gamma	5.15
D_alpha.2	4.45
D_alpha.1	3.28
D_beta.2	1.16
D_beta.1	1.36
Proportion.2	0.32
Proportion.1	0.68
Differentiation.2	0.22
Differentiation.1	0.33

All interpretations are similar to those given above for raw abundances and thus we omit the details.

Phylogenetic Diversity Analysis

In the following, we explain the output if the raw allele abundances (Table 1), the hierarchical structure (Table 2) and the phylogenetic tree (Table 4) are uploaded. First, in the "Data Summary" tab panel, the output includes the plot of the underlying hierarchical structure, the plot of the phylogenetic tree and basic data information. You can click "save as csv file" at the bottom of the output to download the basic data summary.

	n	S.obs	f1*	f2*	g1	g2	f1	f2	observed PD	mean_T	phylo-entropy	phylo-diversity
Population1	10	3	2	2	60.37	104.58	1	1	267.63	103.18	70.11	203.55
Population2	35	5	2	0	65.02	0	2	0	293.66	96.91	78.06	216.87
Population3	30	4	0	2	0	60.37	0	1	243	81.75	73.09	199.88
Population4	30	6	2	1	68.86	15.36	2	1	332.53	102.57	92.83	253.56
Population5	60	4	0	0	0	0	0	0	246.97	93.08	83.34	227.87

Notes:

- n = number of observed individuals or allele copies in the sample.
- S.obs = number of observed species or alleles in the sample.
- f1* = number of nodes/branches with abundance = 1 in the observed tree (the tree spanned by the observed species).
- f2* = number of nodes/branches with abundance = 2 in the observed tree.
- g1 = total length of those nodes/branches with sample abundance = 1 in the observed tree.
- g2 = total length of those nodes/branches with sample abundance = 2 in the observed tree.
- observed PD = observed Faith's PD.
- mean_T = weighted (by species abundance) mean of the distance from root node to each of the terminal branch tips.
- phylo-entropy = empirical phylogenetic entropy.
- phylo-diversity = mean_T x exp(empirical phylogenetic entropy divided by mean_T) in unit of branch length.

(3) In the "Decomposition Analysis" tab panel, the numerical values for gamma diversity, alpha and beta components at each level, proportion of total beta information found at each level, and mean differentiation (dissimilarity) among sampling units for each level are shown. You can click "save as csv file" at the bottom of the output to download the output.

Faith's PD	321.53
mean_T	94.17
PD_gamma	274.39
PD_alpha.2	255.19
PD_alpha.1	223.23
PD_beta.2	1.08
PD_beta.1	1.14
Proportion.2	0.35
Proportion.1	0.65
Differentiation.2	0.12
Differentiation.1	0.15

The term "effective" is used here in the sense of there being equally abundant and equally divergent lineages/populations/regions.

(1) The total branch length (Faith's PD) in the phylogenetic tree is 321.53.

(2) The weighted (by species abundance) mean of the distances from root node to each of the tips is 94.17.

(3a) PD_gamma = 274.39 is interpreted as the effective total branch length in the ecosystem (total phylogenetic diversity) being 274.39.

(3b) PD_alpha.2 = 255.19 is interpreted as the effective total branch length per region being 255.19.

PD-beta.2 = 1.08 means that there are 1.08 region equivalents. Thus, 255.19 x 1.08 = 274.39 (= PD_gamma).

(3c) PD_alpha.1 = 223.23 is interpreted as the effective total branch length per population within each region being 223.23.

PD_beta.1 =1.14 implies that there are 1.14 population equivalents per region.

Here 223.23 x 1.14 = 255.19 (= PD_alpha.2).

(4) PD_prop.2 = 0.35 means that the proportion of total phylogenetic beta information found at the regional level is 35%.
PD_prop.1 = 0.65 means that the proportion of total phylogenetic beta information found at the population level within a region is 65%.

(5) PD_diff.2 = 0.12 implies that the mean phylogenetic differentiation among regions is 0.12. This can be interpreted in the following effective sense: the mean proportion of non-shared branch lengths in a region is around 12%.
PD_diff.1 = 0.15 implies that the mean phylogenetic differentiation among communities within a region is 0.15, i.e., the mean proportion of non-shared branch lengths in a population within a region is around 15%.

References

Chao, A., Chiu C.-H., and Jost, L. (2010). Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B.*, 365, 3599-3609.

Chao, A., Chiu, C.-H., and Jost, L. (2014). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity/differentiation measures through Hill numbers. *Annual Reviews of Ecology, Evolution, and Systematics*, 45, 297-324.

Gaggiotti, O. E., Chao, A., Peres-Neto, P., Chiu, C.-H., Edwards, C., Fortin, M.-J., Jost, L., Richards, C. M., and Selkoe, K. A. (2017). Diversity from genes to ecosystems: a unifying framework to study variation across biological metrics and scales. *Evolutionary Applications* Special Issue.