

LTET Ecophylogenetics Workshop Series

Instructions for preparing data for phylogenetic analysis of community structure and species traits

EcoPhyl version 2.0

J. Cavender-Bares and C. Lehman, May 16, 2007

I. Raw Data

To start, we need four raw data matrices: 1) a matrix of species abundance or presences within plots, 2) a matrix of environmental data within plots, 3) a matrix of trait data for each species, and 4) a list of taxa giving the current scientific names for the family, genus and species. These raw matrices will be transformed into several derived matrices, including a species co-occurrence matrix, an ecological niche overlap matrix, a trait difference matrix, and a phylogenetic distance matrix. (An interactive interface allows you to choose the analyses you wish to run. It can also be run using a command line inside Cygwin.)

1. Community data

The community data are in the form of either the presence/absence of species within vegetation plots (T1) or their relative abundances (T2), based on percent cover, basal area, biomass or other measures. For each species, relative abundances sum to one across all plots. These $n \times n$ matrices (T1 and T2) can then be converted to a 3 column matrix (T3). (*Module “[matrixtr](#)” converts the presence/absence or relative abundance matrix to a three-column matrix. It can also take a 3 column matrix and convert it to an $n \times n$ matrix.*) The species names can be given in any form, e.g., as a genus and species name connected with an underscore (Andropogon_repens) or separated by a space (Andropogon repens), a one word abbreviation (andre), or as a number. The data table with species then plots (T1 or T2) should be formatted with tabs separating the columns and saved as a text file.

T1) Presence/absence (PA) species by plot matrix

Spp	plot 1	plot 2	plot 3
Agropyron_repens	0	1	1
Quercus_elipsoidalis	1	0	1
Quercus_macrocarpa	1	0	1

T2) Relative abundance (RA) species by plot matrix

Spp	plot 1	plot 2	plot 3
Agropyron_repens	0	0.25	0.75
Quercus_elipsoidalis	0.5	0	0.5
Quercus_macrocarpa	0.3	0	0.7

T3) Three-column species-plot matrix

Plot	Spp	RA
plot1	Agropyron_repens	0
plot2	Agropyron_repens	0.25
plot3	Agropyron_repens	0.75
plot1	Quercus_elipsoidalis	0.5
plot2	Quercus_elipsoidalis	0
plot3	Quercus_elipsoidalis	0.5
plot1	Quercus_macrocarpa	0.3
Quercus_macrocarpa	plot2	0
Quercus_macrocarpa	plot3	0.7

These data matrices are converted to species “co-occurrence” (T5) or “niche overlap” (T6) matrices (see next section) that give the degree of co-occurrence or environmental similarity between pairs of species. These are three column matrices that include two columns of species and one column with the similarity index (T5). EcoPhyl calculates Schoener’s proportional similarity index (Schoener 1970) or “co-occurrence index,” CI_{ih} . This indicates the degree to which species pairs co-occur within plots:

$$CI_{ih} = 1 - 0.5 \times \sum |p_{ij} - p_{hj}| \quad \text{eq. 1}$$

where CI_{ih} is the co-occurrence of species i and h , and p_{ij} is the proportion of total abundance or the proportion of occurrences of the i th species in the j th plot. (*Module “pwco” creates a co-occurrence matrix from the raw data matrix using this index.*) The similarity matrix is output as a three column matrix in which the first two columns are species pairs and the third column is the similarity of those two species (T5).

T5) Species co-occurrence matrix

Spp1	Spp2	CI
Agropyron_repens	Quercus_elipsoidalis	0.5
Agropyron_repens	Quercus_macrocarpa	0.7
Quercus_elipsoidalis	Quercus_macrocarpa	0

2. Environmental data

In conjunction with environmental data, the community data can also be used to calculate matrices of niche overlap or of ecological similarity of species using Pianka’s niche overlap index (Pianka 1973). This requires environmental variables within plots (T6).

T6) Environmental variables within plots

Plot #	Ave				
	PAR	Min T	Max T	PPT	N Min
1	2000	-5	38	10	5
2	1000	-5	38	10	4
3	500	-3	30	10	3
4	50	0	25	10	1

Pianka's niche overlap index indicates the degree to which species overlap in their preferences for resources (e.g., % soil moisture) or environmental regulators (e.g., temperature).

$$O_{jk} = O_{kj} = \frac{\sum_i^n E_{ij} E_{ik}}{\sqrt{\sum_i^n (E_{ij}^2) \sum_i^n (E_{ik}^2)}} \quad \text{eq. 2}$$

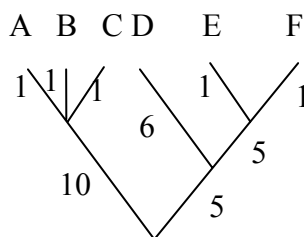
where O_{jk} is the overlap of species j on species k , i is the resource level, n is the number of resource levels, and E_{ij} is the proportion of abundance or occurrences of species j in level i divided by the number of plots that fall within level i . (*This module is under development.*)

3. Phylogenetic Data

Phylogenetic data originates as a phylogeny showing the relationships between species (Fig. 1). To generate the phylogeny, a series of preliminary steps are necessary using various other programs and tools. One possibility is to use [Phyloomatic](#) (Webb & Donoghue 2005), which generates a phylogenetic tree using family, genus and species names based on the literature <http://www.phylodiversity.net/phyloomatic/>. In Phyloomatic, output the phylogenetic tree (Fig. 1) in Newick format. Examples are shown without (T7) or with branchlengths (T8). The taxa in the example phylogeny in Fig. 1 are shown as species A,B,C,D,E and F. The numbers represent branch length distances between nodes that can be summed to get the branch length distances between pairs of species. The branch length distances between species are determined either from molecular data or by extrapolating minimum ages since divergence from a common ancestor using fossil records. This can be done using an algorithm in [Phylocom](#) called "BLADJ" for branch length adjustment (Webb & Donoghue 2005).

The phylogenetic tree is converted to a distance matrix (T9), which is an $n \times n$ matrix that includes the branch length distances between species pairs. If branchlengths are included in the phylogeny, these are used to calculate distances, otherwise nodal distances are assigned values of 1. The $n \times n$ matrix can then be transformed to a 3 column matrix as shown below (T10). (*Module "treedist" converts a Newick formatted phylogeny into an $n \times n$ formatted phylogeny.*) The program allows polytomies, e.g., (A,B,C) and the use of taxon names in any format. It is best to use the same taxon names that are used for the community data. The "simple" command in [Phylocom](#) (Webb et al. 2004) also converts the string formatted phylogeny with branch lengths to an $n \times n$ phylogenetic distance matrix.

Fig. 1



T7 Newick format without branchlengths:
((A,B,C),((E,F),D))

T8 Newick format including branch lengths:
((A:1,B:1,C:1):10,((E:1,F:1):5,D:6):5)

T9) Phylogenetic distance matrix

	A	B	C	D	E	F
A	0	2	2	22	22	22
B		0	2	22	22	22
C			0	22	22	22
D				0	12	12
E					0	2
F						0

T10) Three-column phylogenetic distance matrix

A	B	2
A	C	2
A	D	22
A	E	22
A	F	22
B	C	1
B	D	22
B	E	22
B	F	22
C	D	22
C	E	22
C	F	22
D	E	12
D	F	12
E	F	2

4. Trait data

In the raw matrix, there is a column of species followed by multiple columns of continuous trait values (T11). Each trait needs to be transformed into a “trait difference matrix” (T12) which has the absolute value of the trait difference between species pairs as shown below. The multivariate Euclidean trait distance among species for a designated number of traits is also calculated. (This is the square root of the sum of the squares of the Δ trait values). Optionally, variables can first be log-transformed and subsequently standardized to have a mean of zero and a standard deviation of one. (*Module « [traitdif](#) » converts a trait matrix into a trait difference matrix.*)

T11) Trait matrix

Spp	Trait1	Trait2	Trait3
Agropyron_repens	0.2	2	372
Quercus_elipsoidalis	0.5	15	250
Quercus_macrocarpa	0.7	17	199
Prunus_caroliniana	0.8	16	200

T12) Trait difference matrix

Spp1	Spp2	Δ Trait1	Δ Trait2	Δ Trait3
Agropyron_repens	Quercus_elipsoidalis	0.3	13	122
Agropyron_repens	Quercus_macrocarpa	0.5	15	173
Agropyron_repens	Prunus_caroliniana	0.6	14	172
Quercus_elipsoidalis	Quercus_macrocarpa	0.2	2	51
Quercus_elipsoidalis	Prunus_caroliniana	0.3	1	50
Quercus_macrocarpa	Prunus_caroliniana	0.1	1	1

II. Statistical analyses**1. Correlation between co-occurrence and phylogenetic distance**

Two approaches are used to determine whether the co-occurrence of species or their niche similarity is associated with their phylogenetic relationships. Both use Mantel methods that

compare the co-occurrence (or niche overlap) matrix to the phylogenetic distance matrix (Fig. 2). Using the first procedure, a correlation coefficient based on least squares regression is determined for the relationship between the co-occurrence of species pairs and the phylogenetic distance between them. The second procedure uses quantile regression which minimizes the absolute distance between the data and a fitted line. The minimized distance can be weighted to focus on the upper bound of the data (e.g., 75th or 90th quantile) rather than the median (50th quantile). Quantile regression to find the median is most similar to least squares regression which finds the mean.

Both procedures require the species co-occurrence matrix and a phylogenetic distance matrix. The community data are randomly permuted using various constraints a designated number of times (e.g., 999) to generate a null model of simulated species co-occurrence matrices. The observed correlation coefficient or quantile regression slope is compared to the same test statistics generated from randomized permutations of the raw community data matrix from which the simulated co-occurrence matrices are calculated. Alternatively, the observed test statistics are compared to null model values in which species phylogenetic relationships are permuted by randomizing the tips of the phylogeny. A description of the randomization procedures is given in section III below. (*The module “qslope” calculates the correlation quantile regression slope for the fitted line between an x and y variable; the module “corr” calculates the correlation coefficient. The test statistics are compared to one of several user-determined null models. Currently, the co-occurrence matrix needs to be in a 3-column format and the distance matrix in the $n \times n$ format.*)

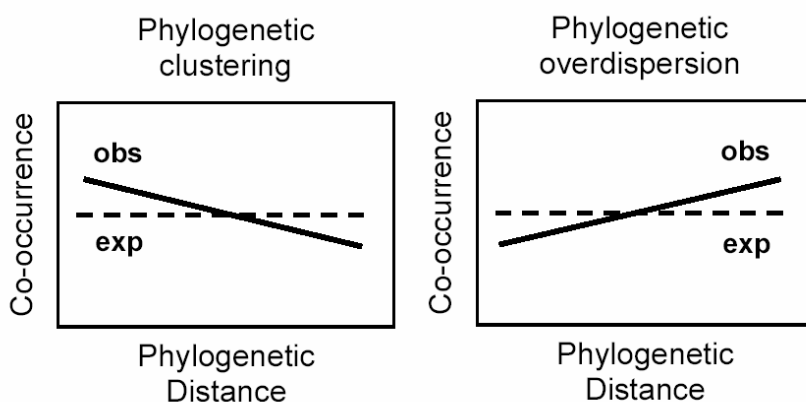


Fig. 2. A) An observed relationship (obs) between the co-occurrence of species and their phylogenetic distances that is more negative than expected (exp) indicates that close relatives co-occur more often than expected (phylogenetic clustering). B) An observed relationship that is more positive than expected indicates that close relatives co-occur less often than expected.

2. Correlation between co-occurrence and trait difference

This test finds a correlation coefficient for the relationship between the co-occurrence of species pairs and the single or multivariate Euclidean trait distance between the species (Fig. 3). This procedure requires the species co-occurrence matrix and a trait difference matrix. The raw community data are randomly permuted to generate a null model of simulated species co-occurrence matrices. The observed correlation coefficient is compared to the expected correlation coefficient from these simulations. (*The module “corr” calculates the correlation coefficient for these two variables, but the trait differences matrix must be in $n \times n$ format.*)

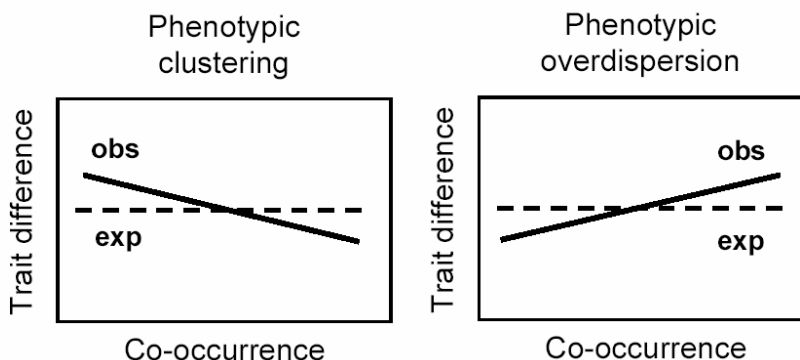


Fig. 3. A) An observed relationship (obs) between the trait differences of species and their co-occurrence that is more negative than expected (exp) indicates that species with similar phenotypes co-occur more often than expected (phenotypic clustering). B) An observed relationship that is more positive than expected indicates that species with similar phenotypes co-occur less often than expected (phenotypic overdispersion). From (Cavender-Bares et al. 2006)

3. Conservatism of trait evolution

Several programs exist to determine whether there is phylogenetic signal in traits, such that close relatives have more similar trait values than distant relatives. One of these is a procedure developed by David Ackerly, a module in [Phylocom](#) called “aot” (analysis of traits) (see Moles et al. 2005). The procedure determines the degree to which close relatives share similar trait values.

4. Relationship between trait evolution and phenotypic similarity within communities

The final step is to relate the degree to which traits are conserved and the degree to which traits are clustered or overdispersed within communities. This relationship should explain the outcome of analysis (1) above. This is simply a visual test at this point. An example is shown below (Fig. 4) (taken from Cavender-Bares et al. 2004):

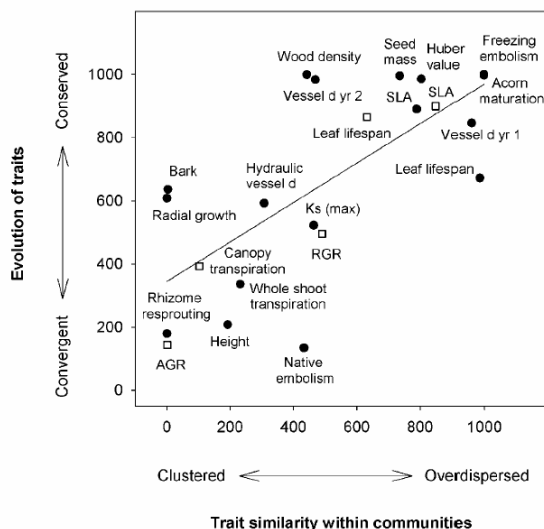


Fig. 4. The phylogenetic signal in traits is compared to how similar species traits are within communities. Exploring this relationship helps to understand the underlying mechanisms for the phylogenetic structure of communities.

III. Null model randomization procedures

There are several null models that can be used to compare observed correlation coefficients with simulated correlation coefficients using randomization procedures. For a thorough discussion of null models and caveats in their use, see Gotelli (1996).

1. Randomization of community data

Start with a matrix of species relative abundance species by plots. A null model can be generated by randomizing relative abundance within rows, keeping row totals constant, but not attempting to keep column totals constant (null model 1). (*The module “[shuffle1](#)” does this randomization routine.*) For presence/absence data, a better null model is to randomize rows and columns, keeping row and column totals constant (null model 2). This should maintain the same biomass within a plot in the simulation compared to the observed and, hence, is biologically more realistic. There are several algorithms for randomizing the data to achieve these constraints, and care should be taken that an unbiased procedure is used. The knights tour and the sequential swap algorithms are used here (Gotelli & Entsminger 2001).

2. Randomization of phylogenetic relationships

Alternatively, the phylogenetic data can be permuted rather than the community data. This can be done by keeping the tree topology constant and randomizing the taxa across the phylogeny (null model 3). (*The module “[nullphylo](#)” does this randomization routine.*) Finally, randomization procedures are available in Mesquite (Maddison & Maddison 2000) that allow randomization of the tree topology, keeping total evolutionary distance between the base of the tree and the extant taxa constant (null model 4).

Literature cited

- Cavender-Bares, J., D. D. Ackerly, D. A. Baum, and F. A. Bazzaz. 2004. Phylogenetic overdispersion in Floridian oak communities. *American Naturalist* **163**:823-843.
- Cavender-Bares, J., A. Keen, and B. Miles. 2006. Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. *Ecology* **87**:S109-S122.
- Gotelli, N. J., and G. L. Entsminger. 2001. Swap and fill algorithms in null model analysis: rethinking the knight’s tour. *Oecologia* **129**:281–291.
- Gotelli, N. J., and G. R. Graves. 1996. *Null Models in Ecology*. Smithsonian Institution Press, Washington, D.C.
- Maddison, W., and D. Maddison. 2000. Mesquite: A modular programming system for evolutionary analysis. *in*. University of Arizona.
- Moles, A. T., D. D. Ackerly, C. O. Webb, J. C. Tweddle, J. B. Dickie, and M. Westoby. 2005. A brief history of seed size. *Science* **307**:576-580.
- Pianka, E. R. 1973. The structure of lizard communities. *Annual Review of Ecology and Systematics* **4**:53-74.
- Schoener, T. W. 1970. Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology* **51**:408-418.
- Webb, C., D. Ackerly, and S. Kembel. 2004. Phylocom: phylogenetic analysis of communities and characters. www.phylodiversity.net/phylocom. *in*.
- Webb, C. O., and M. J. Donoghue. 2005. Phylomatic: tree assembly for applied phylogenetics. *Molecular Ecology Notes* **5**:181.