

Community phylogenetic analysis with picante

Steven Kembel (skembel@uoregon.edu)

August 2009

Contents

1	Installing picante	1
2	Reading and writing data	1
3	Visualizing trees and data	1
4	Community phylogenetic structure	4
4.1	Phylogenetic diversity	4
4.2	MPD, MNTD, SES_{MPD} and SES_{MNTD}	5
4.3	Phylogenetic beta diversity	8
5	Comparative analyses	9
5.1	Phylogenetic signal	9
6	Literature cited	10

1 Installing picante

The picante homepage is located at <http://picante.r-forge.r-project.org>. From within R, you can install the latest version of picante by typing "install.packages(picante, dependencies=TRUE)". Typing "help(functionName)" will display documentation for any function in the package.

2 Reading and writing data

Had to excise this section for easier compilation (will put back later). For now we can simply load the Phylocom example data set included with `picante`.

```
> data(phylocom)
```

3 Visualizing trees and data

One of the main advantages of using R is that a suite of graphical and statistical tools are included. Now that we've loaded our data sets, we can use some of those tools to visualize them. Remember that we have three objects containing the community (`samp`), phylogeny (`phy`), and trait (`traits`) data sets.

```
> samp <- phylocom$sample
> phy <- phylocom$phylo
> traits <- phylocom$traits
```

As a prelude to the analyses we're about to run, let's look at how species in some of the communities in the example data set we just loaded are distributed across the tips of the phylogeny. To do this, we first need to prune the phylogeny to include only the species that actually occurred in some community.

```
> prunedphy <- prune.sample(samp, phy)
> prunedphy
```

Phylogenetic tree with 25 tips and 24 internal nodes.

Tip labels:

sp1, sp2, sp3, sp4, sp5, sp6, ...

Node labels:

A, B, C, D, E, F,...

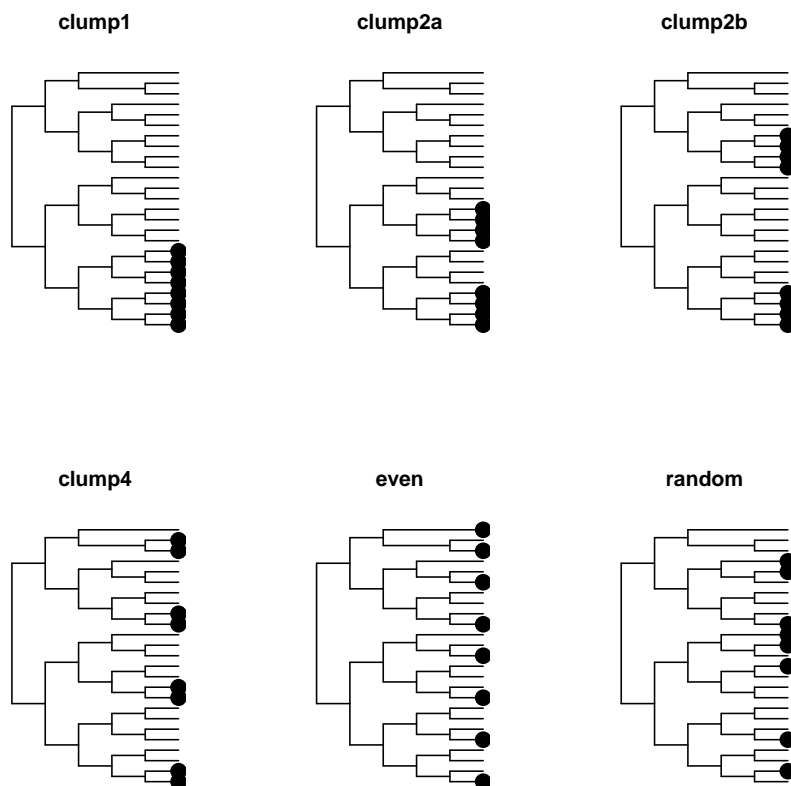
Rooted; includes branch lengths.

We also need to make sure the species are arranged in the some order in the community data and the phylogeny. This is an important step - several functions in `picante` assume that the community or trait data and phylogeny data have species arranged in the same order, so it's good to always make sure we've done so before running any analysis. The following command sorts the columns of `samp` to be in the same order as the tip labels of the phylogeny:

```
> samp <- samp[, prunedphy$tip.label]
```

Now let's see how taxa from the six communities in the Phylocom example data set are arranged on the tree. The following commands set up the layout of the plot to have 2 rows and 3 columns, and then plot a black dot for the species present in each of the six communities:

```
> par(mfrow = c(2, 3))
> for (i in row.names(samp)) {
+   plot(prunedphy, show.tip.label = FALSE, main = i)
+   tiplabels(tip = which(samp[i, ] > 0), pch = 19, cex = 2)
+ }
```

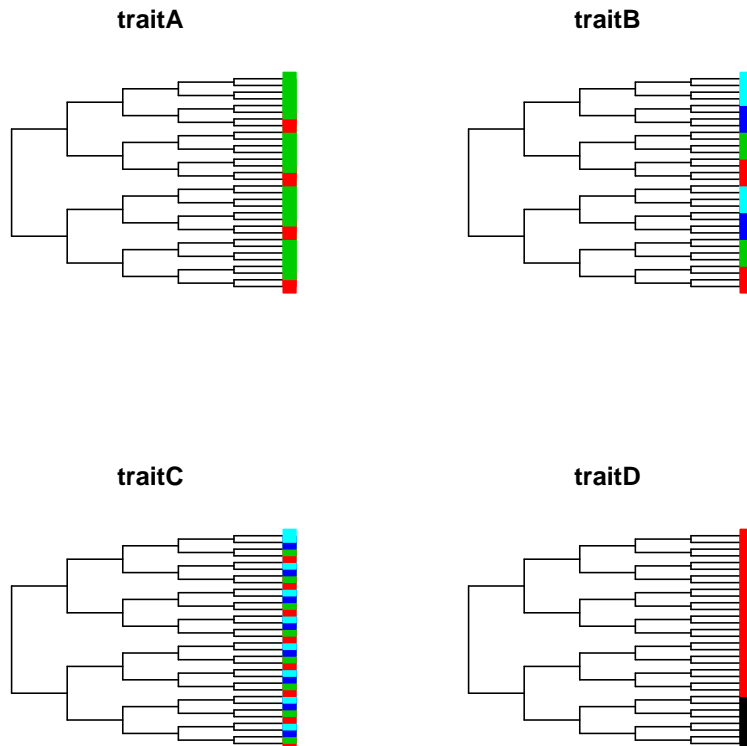


Similarly, let's visualize the trait values on the trees by plotting a different color for each trait value. The arguments to the `tiplabels` function give each trait value a unique color and adjust the size of the trait symbols.

```

> par(mfrow = c(2, 2))
> for (i in names(traits)) {
+   plot(phy, show.tip.label = FALSE, main = i)
+   tiplabels(pch = 22, col = traits[, i] + 1, bg = traits[,
+     i] + 1, cex = 1.5)
+ }

```



Looking at these plots, which communities would you expect to be phylogenetically clustered or phylogenetically even? Which of the four traits do you think has the greatest amount of phylogenetic signal?

4 Community phylogenetic structure

4.1 Phylogenetic diversity

One of the earliest measures of phylogenetic relatedness in ecological communities was the phylogenetic diversity (PD) index proposed by Faith (1992). Faith's PD is defined as the total branch length spanned by the tree including all species in a local community. The `pd` function returns two values for each community, the PD and the species richness (SR).

```
> pd.result <- pd(samp, phy, include.root = TRUE)
> pd.result
```

	PD	SR
clump1	16	8
clump2a	17	8
clump2b	18	8
clump4	22	8
even	30	8
random	27	8

Looking at these results, we can see that the communities where taxa are clumped on the phylogeny tend to have a lower PD, because the species in these communities capture only a small part of the total phylogenetic diversity present in the phylogeny.

4.2 MPD, MNTD, SES_{MPD} and SES_{MNTD}

Another way of thinking about the phylogenetic relatedness of species in a community is to ask 'how closely related are the average pair of species or individuals in a community?', and relate the patterns we observe to what we'd expect under various null models of evolution and community assembly. These types of questions are addressed by the measures of community phylogenetic structure such as MPD, MNTD, NRI and NTI described by Webb et al. (2002) and implemented in Phylocom (Webb et al. 2008).

The function `mpd` will calculate the mean pairwise distance (MPD) between all species in each community. Similarly, the `mntd` function calculates the mean nearest taxon distance (MNTD), the mean distance separating each species in the community from its closest relative. The `mpd` and `mntd` functions differs slightly from the `pd` function in that they take a distance matrix as input rather than a phylogeny object.

A `phylo` object can be converted to an interspecific phylogenetic distance matrix using the `cophenetic` function. Since the `mpd` and `mntd` functions can use any distance matrix as input, we could easily calculate trait diversity measures by substituting a trait distance matrix for the phylogenetic distance matrix.

Measures of ‘standardized effect size’ of phylogenetic community structure can be calculated for MPD and MNTD by compared observed phylogenetic relatedness to the pattern expected under some null model of phylogeny or community randomization. Standardized effect sizes describe the difference between phylogenetic distances in the observed communities versus null communities generated with some randomization method, divided by the standard deviation of phylogenetic distances in the null data:

$$SES_{metric} = \frac{Metric_{observed} - mean(Metric_{null})}{sd(Metric_{null})}$$

Phylocom users will be familiar with the measures NRI and NTI; SES_{MPD} and SES_{MNTD} are equivalent to -1 times NRI and NTI, respectively, when these functions are run with a phylogenetic distance matrix.

Several different null models can be used to generate the null communities that we compare observed patterns to. These include randomizations of the tip labels of the phylogeny, and various community randomizations that can hold community species richness and/or species occurrence frequency constant. These are described in more detail in the help files, as well as in the Phylocom manual. Let’s calculate some of these measures of community phylogenetic structure for our example data set. We will use a simple null model of randomly shuffling tip labels across the tips of the phylogeny. For a ‘real’ analysis we’d want to use a much higher number of runs:

```
> phydist <- cophenetic(phy)
> ses.mpd.result <- ses.mpd(samp, phydist, null.model = "taxa.labels",
+   abundance.weighted = FALSE, runs = 99)
> ses.mpd.result
```

	ntaxa	mpd.obs	mpd.rand.mean	mpd.rand.sd	mpd.obs.rank
clump1	8	4.857143	8.318182	0.3693642	1.0
clump2a	8	6.000000	8.375180	0.2541499	1.0
clump2b	8	7.142857	8.341270	0.3290014	1.0
clump4	8	8.285714	8.373737	0.3098125	33.0
even	8	8.857143	8.305916	0.4372351	99.5
random	8	8.428571	8.282107	0.3360187	58.5

	mpd.obs.z	mpd.obs.p	runs
clump1	-9.3702615	0.010	99
clump2a	-9.3455868	0.010	99
clump2b	-3.6425760	0.010	99
clump4	-0.2841173	0.330	99
even	1.2607097	0.995	99
random	0.4358825	0.585	99

```
> ses.mntd.result <- ses.mntd(samp, phydist, null.model = "taxa.labels",
+   abundance.weighted = FALSE, runs = 99)
> ses.mntd.result
```

	ntaxa	mntd.obs	mntd.rand.mean	mntd.rand.sd	mntd.obs.rank
clump1	8	2	4.825758	0.5708195	1.0
clump2a	8	2	4.563131	0.7163750	1.0
clump2b	8	2	4.717172	0.6630234	1.0
clump4	8	2	4.654040	0.6659609	1.0
even	8	6	4.780303	0.6270606	98.5
random	8	5	4.669192	0.6641592	68.0

	mntd.obs.z	mntd.obs.p	runs
clump1	-4.9503524	0.010	99
clump2a	-3.5779184	0.010	99
clump2b	-4.0981533	0.010	99
clump4	-3.9852796	0.010	99
even	1.9451022	0.985	99
random	0.4980855	0.680	99

The output includes the following columns:

- `ntaxa` Number of taxa in community
- `mpd.obs` Observed mpd in community
- `mpd.rand.mean` Mean mpd in null communities
- `mpd.rand.sd` Standard deviation of mpd in null communities
- `mpd.obs.rank` Rank of observed mpd vs. null communities
- `mpd.obs.z` Standardized effect size of mpd vs. null communities (equivalent to -NRI)

- `mpd.obs.p` P-value (quantile) of observed mpd vs. null communities ($= \text{mpd.obs.rank} / \text{runs} + 1$)
- `runs` Number of randomizations

Positive *SES* values (`mpd.obs.z` > 0) and high quantiles (`mpd.obs.p` > 0.95) indicate phylogenetic evenness, or a greater phylogenetic distance among co-occurring species than expected. Negative *SES* values and low quantiles (`mpd.obs.p` < 0.05) indicate phylogenetic clustering, or small phylogenetic distances among co-occurring species than expected. MPD is generally thought to be more sensitive to tree-wide patterns of phylogenetic clustering and evenness, while MNTD is more sensitive to patterns of evenness and clustering closer to the tips of the phylogeny. For example, community 'clump4' contains species that are spread randomly across the entire tree (SES_{MPD} close to zero) but phylogenetically clustered towards the tips (negative SES_{MNTD} and `mntd.obs.p` in the low quantiles of the null distribution).

All of these measures can incorporate abundance information when available using the `abundance.weighted` argument. This will change the interpretation of these metrics from the mean phylogenetic distances among species, to the mean phylogenetic distances among individuals.

4.3 Phylogenetic beta diversity

We can measure patterns of phylogenetic relatedness among communities in a manner similar to the within-community measures described above. The `comdist` and `comdistnt` functions measure the among-community equivalent of MPD and MNTD, the mean phylogenetic distance or mean nearest taxon distance between pairs of species drawn from two distinct communities.

Phylogenetic beta diversity measures can be used with any method based on measuring among-community distances. For example, they could be used in a cluster analysis or phyloordination to group communities based on their evolutionary similarity, or they could be compared with spatial or environmental distances separating communities using a Mantel test. The code below calculates MPD between pairs of communities, and uses these phylogenetic distances to cluster communities based on their phylogenetic similarity:

```
> comdist.result <- comdist(samp, phydist)
> comdist.result

      clump1 clump2a clump2b  clump4    even
clump2a 6.12500
```



```

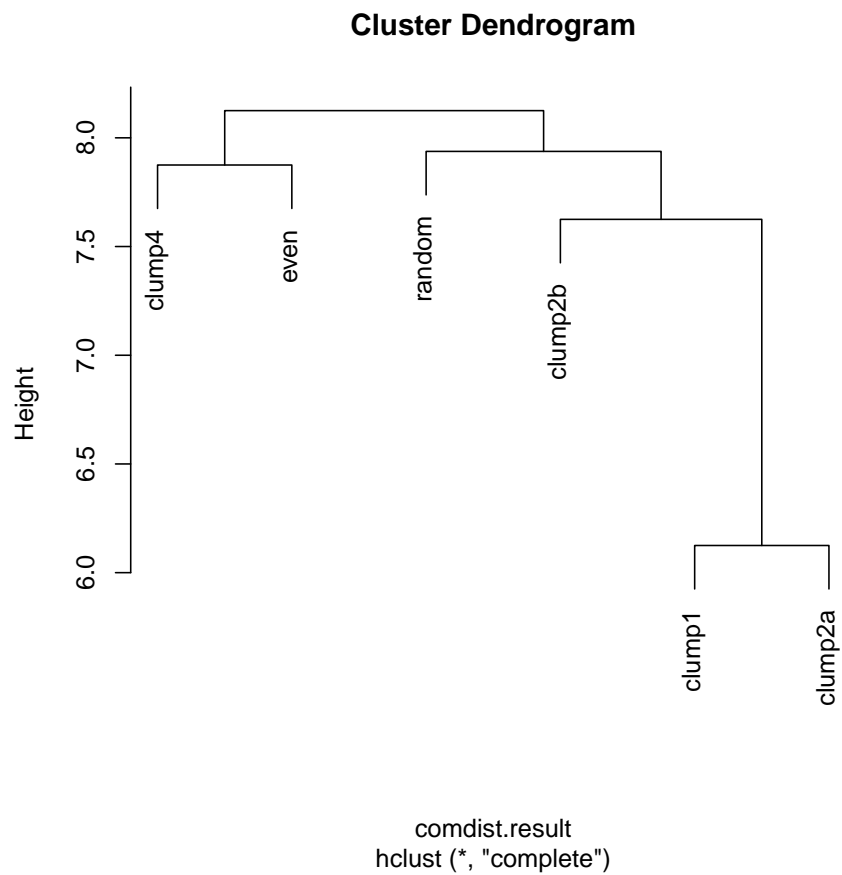
clump2b 7.12500 7.62500
clump4  8.06250 7.62500 7.62500
even    8.06250 8.06250 8.06250 7.87500
random  7.81250 7.68750 7.93750 8.12500 8.03125

```

```

> library(cluster)
> comdist.clusters <- hclust(comdist.result)
> plot(comdist.clusters)

```



5 Comparative analyses

5.1 Phylogenetic signal

The idea of phylogenetic niche conservatism (the ecological similarity of closely related species) has attracted a lot of attention recently, for example in the widely used framework of inferring community assembly processes based on knowledge of community phylogenetic structure plus the phylogenetic conservatism of traits. (Webb et al. 2002).

Phylogenetic signal is a quantitative measure of the degree to which phylogeny predicts the ecological similarity of species. The K statistic is a measure of phylogenetic signal that compares the observed signal in a trait to the signal under a Brownian motion model of trait evolution on a phylogeny (Blomberg et al. 2003). K values of 1 correspond to a Brownian motion process, which implies some degree of phylogenetic signal or conservatism. K values closer to zero correspond to a random or convergent pattern of evolution, while K values greater than 1 indicate strong phylogenetic signal and conservatism of traits. The statistical significance of phylogenetic signal can be evaluated by comparing observed patterns of the variance of independent contrasts of the trait to a null model of shuffling taxa labels across the tips of the phylogeny.

These tests are implemented in the `Kcalc`, `phylosignal`, and `multiPhylosignal` functions. All of these functions assume the trait data are in the same order as the phylogeny tip.labels. Let's make sure the Phylocom trait data are in this order and then measure phylogenetic signal in these data.

```
> traits <- traits[phy$tip.label, ]  
> multiPhylosignal(traits, phy)
```

	K	PIC.variance.obs	PIC.variance.rnd.mean	PIC.variance.P
traitA	0.8905609	0.05396825	0.1253741	0.001
traitB	2.9340184	0.10920635	0.8323205	0.001
traitC	0.5149502	0.62222222	0.8270207	0.062
traitD	4.3536696	0.01103943	0.1259874	0.001

	PIC.variance.Z
traitA	-3.604988
traitB	-5.159166
traitC	-1.537023
traitD	-5.837426

The higher the K statistic, the more phylogenetic signal in a trait. `PIC.variance.P` is the quantile of the observed phylogenetically independent contrast variance versus the null distribution, which can be used as a 1-tailed P-value to test for greater phylogenetic signal than expected. Traits with `PIC.variance.P` < 0.05 have non-random phylogenetic signal.

6 Literature cited

- Blomberg, S. P., T. Garland, Jr., and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57:717-745.
- Faith, D.P. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61:1-10.
- Webb, C., D. Ackerly, M. McPeck, and M. Donoghue. 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics* 33:475-505.
- Webb, C.O., Ackerly, D.D., and Kembel, S.W. 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Version 4.0.1. <http://www.phylodiversity.net/phylocom/>