An introduction to the picante package

Steven Kembel (skembel@uoregon.edu)

April 2010

Contents

1 Installing picante **1**

2 Data formats in picante **2**
2.1 Phylogenies 2
2.2 Community data 3
2.3 Trait data 4

3 Visualizing trees and data **5**

4 Community phylogenetic structure **7**
4.1 Phylogenetic diversity 7
4.2 MPD, MNTD, SES_{MPD} and SES_{MNTD} 8
4.3 Phylogenetic beta diversity 11

5 Comparative analyses **12**
5.1 Phylogenetic signal 12

6 Literature cited **14**

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.

Once the package has been installed, it can be loaded by typing:

```
> library(picante)
```

2 Data formats in picante

Most analyses in picante work with one of three types of data. The picante package includes a dataset that contains examples of each of these data types. Loading the dataset creates an object named phylocom that include some artificial data included with the Phylocom (Webb et al. 2008) software. We'll create new objects containing each data type so we don't have to type the full name every time:

```
> data(phylocom)
> names(phylocom)

[1] "phylo"  "sample" "traits"

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

2.1 Phylogenies

Picante uses the `phylo` format implemented in the `ape` package to represent phylogenetic relationships among taxa. The format itself is documented at the `ape` homepage (`http://ape.mpl.ird.fr/`). If you have a phylogeny in Newick or Nexus format it can be imported into R with the `read.tree` or `read.nexus` functions.

```
> phy

Phylogenetic tree with 32 tips and 31 internal nodes.

Tip labels:
        sp1, sp2, sp3, sp4, sp5, sp6, ...
        Node labels:
        A, B, C, D, E, F,...

Rooted; includes branch lengths.
```

2.2 Community data

Picante uses the same community data format as the `vegan` package - a matrix or data.frame with sites/samples in the rows and taxa in the columns. The elements of this data frame should be numeric values indicating the abundance or presence/absence (0/1) of taxa in different samples.

One important thing to note is that most functions in picante will use the labels on columns in the community data set to match community and phylogenetic data together. You need to make sure your column names are present and match the tip labels of the phylogeny or your analysis may not work. Most functions in picante do basic error checking and will report when there are mismatches between the data present in the community and phylogenetic data sets. Similarly, your communities/sites/samples can be given informative names, and these should be contained in the row labels, not in a column of the data.frame.

```
> comm
```

```
        sp1 sp10 sp11 sp12 sp13 sp14 sp15 sp17 sp18 sp19 sp2 sp20 sp21
clump1    1    0    0    0    0    0    0    0    0    0   1    0    0
clump2a   1    2    2    2    0    0    0    0    0    0   1    0    0
clump2b   1    0    0    0    0    0    0    2    2    2   1    2    0
clump4    1    1    0    0    0    0    0    2    2    0   1    0    0
even      1    0    0    0    1    0    0    1    0    0   0    0    1
random    0    0    0    1    0    4    2    3    0    0   1    0    0
        sp22 sp24 sp25 sp26 sp29 sp3 sp4 sp5 sp6 sp7 sp8 sp9
clump1     0    0    0    0    0   1   1   1   1   1   1   0
clump2a    0    0    0    0    0   1   1   0   0   0   0   2
clump2b    0    0    0    0    0   1   1   0   0   0   0   0
clump4     0    0    2    2    0   0   0   0   0   0   0   1
even       0    0    1    0    1   0   0   1   0   0   0   1
random     1    2    0    0    0   0   0   2   0   0   0   0
```

```
> class(comm)
```

```
[1] "matrix"
```

```
> colnames(comm)
```

```
 [1] "sp1"  "sp10" "sp11" "sp12" "sp13" "sp14" "sp15" "sp17" "sp18"
[10] "sp19" "sp2"  "sp20" "sp21" "sp22" "sp24" "sp25" "sp26" "sp29"
[19] "sp3"  "sp4"  "sp5"  "sp6"  "sp7"  "sp8"  "sp9"
```

```
> rownames(comm)

[1] "clump1"  "clump2a" "clump2b" "clump4"  "even"    "random"
```

2.3 Trait data

Trait data include any kind of data associated with the taxa present in a phylogeny. Most functions in picante work with trait data represented as a vector (for individual traits) or data.frame. The documentation for individual functions will explain which data format is expected.

When trait data are contained in a data.frame, taxa are in rows and different traits are in columns. As with community data, rows must be labelled and the labels must match taxa names in the phylogeny. If your trait data.frame does not have row labels, it is assumed that the rows are sorted in the same order as the tip labels of the phylogeny.

When trait data are contained in a vector, the vector must be labelled, or it will be assumed that the elements of the vector are in the same order as the tip labels of the phylogeny. There is a utility function in picante that will turn columns of a data frame into vectors maintaining taxa labels that you can use if you have a data.frame and need a vector for a particular analysis:

```
> head(traits)

    traitA traitB traitC traitD
sp1      1      1      1      0
sp2      1      1      2      0
sp3      2      1      3      0
sp4      2      1      4      0
sp5      2      2      1      0
sp6      2      2      2      0

> traitA <- df2vec(traits, "traitA")
> traitA

 sp1  sp2  sp3  sp4  sp5  sp6  sp7  sp8  sp9 sp10 sp11 sp12 sp13 sp14
   1    1    2    2    2    2    2    2    1    1    2    2    2    2
sp15 sp16 sp17 sp18 sp19 sp20 sp21 sp22 sp23 sp24 sp25 sp26 sp27 sp28
   2    2    1    1    2    2    2    2    2    2    1    1    2    2
sp29 sp30 sp31 sp32
   2    2    2    2
```

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 (`comm`), phylogeny (`phy`), and trait (`traits`) data sets.

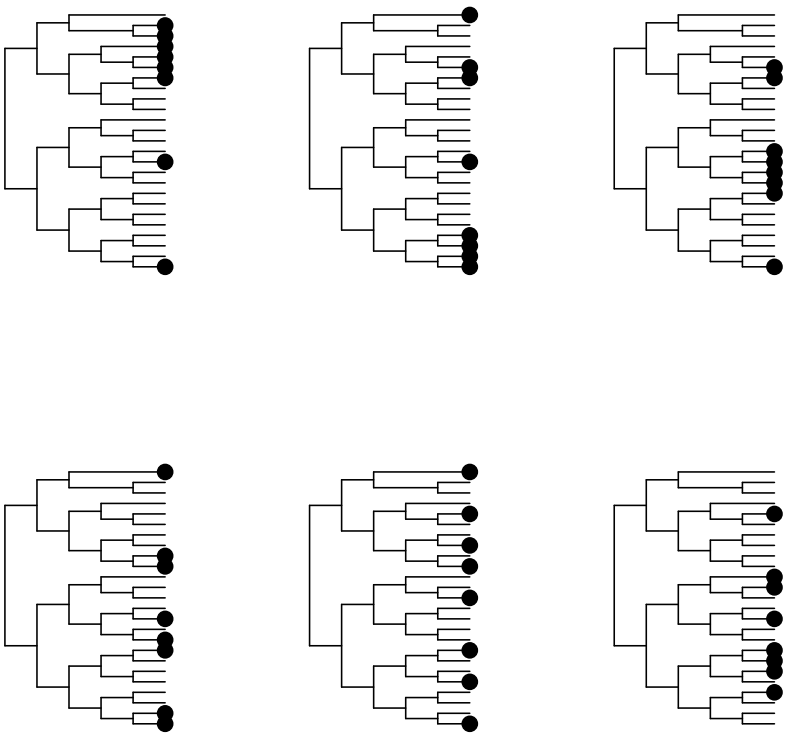
Most functions in `picante` assume that the community, trait and phylogeny data sets contain the same taxa arranged in the same order. Functions that are affected by this assumption will attempt to automatically match the taxa labels in different data sets, and will report taxa that are not present in both data sets.

Let's see how taxa from the six communities in the Phylocom example data set are arranged on the tree. To do this, we first need to prune the phylogeny to include only the species that actually occurred in some community.

```
> prunedphy <- prune.sample(comm, 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.
```

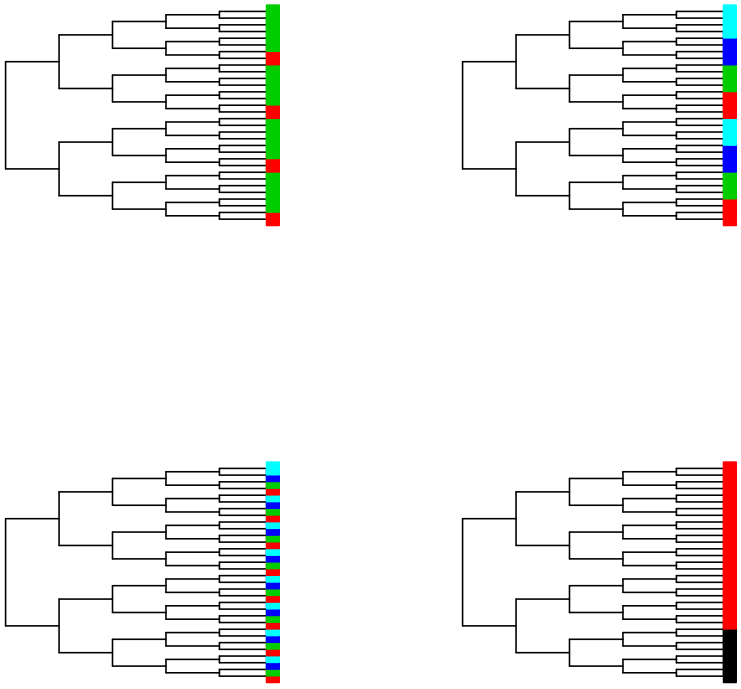
The following commands set up the layout of the plot to have 2 rows and 3 columns, and then plot a black dot for the species present in each of the six communities:

```
> par(mfrow = c(2, 3))
> for (i in row.names(comm)) {
+     plot(prunedphy, show.tip.label = FALSE, main = i)
+     tiplabels(tip = which(comm[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(comm, 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 a 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 random-

ization 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(comm, 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      8.295094   0.3462849          1.0
clump2a     8 6.000000      8.322511   0.3060142          1.0
clump2b     8 7.142857      8.328283   0.3464762          1.0
clump4      8 8.285714      8.330447   0.3297965         33.0
even        8 8.857143      8.375180   0.2602228        100.0
random      8 8.428571      8.370851   0.3035053         50.5
         mpd.obs.z mpd.obs.p runs
clump1  -9.9281004     0.010   99
clump2a -7.5895523     0.010   99
clump2b -3.4213773     0.010   99
clump4  -0.1356383     0.330   99
even     1.8521149     1.000   99
random   0.1901781     0.505   99
```

```
> ses.mntd.result <- ses.mntd(comm, 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       4.744949    0.6398539             1
clump2a     8        2       4.739899    0.5812978             1
clump2b     8        2       4.797980    0.6192869             1
clump4      8        2       4.722222    0.6437384             1
even        8        6       4.742424    0.6938639           100
random      8        5       4.765152    0.6016752            63
        mntd.obs.z mntd.obs.p runs
clump1  -4.2899629       0.01   99
clump2a -4.7134171       0.01   99
clump2b -4.5180670       0.01   99
clump4  -4.2287711       0.01   99
even     1.8124242       1.00   99
random   0.3903244       0.63   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 (= mpd.obs.rank / 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 eveness, 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

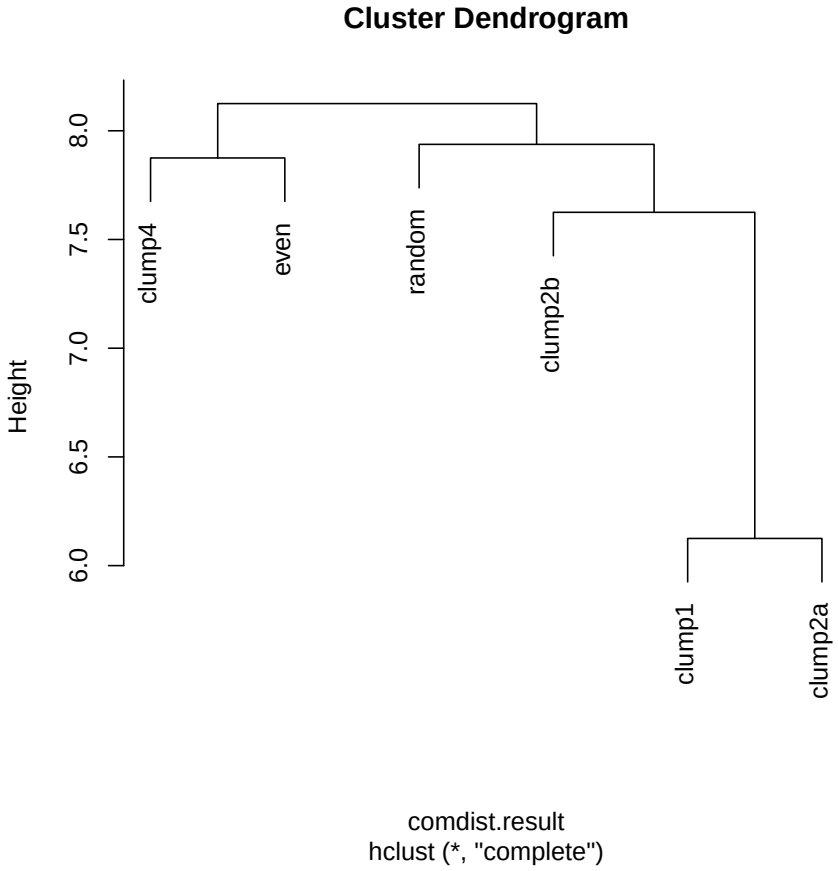
We can measure patterns of phylogenetic relatedness among communities in a manner similar to the within-community measures described above. The `comdist` and `comdistnt` functions measure the among-community equivalent of MPD and MNTD, the mean phylogenetic distance or mean nearest taxon distance between 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(comm, phydist)

[1] "Dropping taxa from the distance matrix because they are not present in the commu
[1] "sp16" "sp23" "sp27" "sp28" "sp30" "sp31" "sp32"

> 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             0.1241637          0.001
traitB 2.9340184       0.10920635             0.8295830          0.001
traitC 0.5149502       0.62222222             0.8341615          0.056
traitD 4.3536696       0.01103943             0.1241840          0.001
       PIC.variance.Z
traitA      -3.509772
traitB      -5.288247
traitC      -1.580329
traitD      -5.593002
```

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. McPeek, 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. Bioinformatics 18:2098-2100.