Phylogenetic signal and diversity

Simon Joly

BIO 6008 - Fall 2015

Contents

Phylogenetic signal

It is frequent to find studies that estimate phylogenetic signal of a given chartacter trait. By that, it is often meant how much the variation in a trait fits the expectation of an evolutionary model, generally the Brownian Motion model.

There are four statistics that are often used for this purpose: Moran's I, Abouheif's c, Pagel's λ, and Blomberg's K, although the last two are by far the most popular. We will define each statistic and apply them on the seedplant data. If you want more information on the different methods, you could read the review paper of Münkemüller et al. (2012) that describes and compare these different methods.

We will also apply the different methods on the seed plants dataset. Let's load it.

```
require(ape)
seedplantstree <- read.nexus("./data/seedplants.tre")
seedplantsdata <- read.csv2("./data/seedplants.csv")
# Remove species for which we don't have complete data
seedplantsdata <- na.omit(seedplantsdata)
# Remove species in the tree that are not in the data matrix
species.to.exclude <- seedplantstree$tip.label[!(seedplantstree$tip.label %in%
                                                 seedplantsdata$Code)]
seedplantstree <- drop.tip(seedplantstree,species.to.exclude)
rm(species.to.exclude)
# Name the rows of the data.frame with the species codes used as tree labels
rownames(seedplantsdata) <- seedplantsdata$Code
seedplantsdata <- seedplantsdata[,-1]
# Order the data in the same order as the tip.label of the tree. In the present
# example, this was already the case.
seedplantsdata <- seedplantsdata[seedplantstree$tip.label,]
# Remove non continuous variables
seedplantsdata <- seedplantsdata[,-c(1,2,8,9)]
```

Moran's I

Moran's I was originally introduced as a measure of spatial autocorrelation (Moran 1950). Gittleman & Kot (1990) suggested it could also be used in phylogenetic analyses. They refer to it as an autocorrelation coefficient describing the relation of cross-taxonomic trait variation to phylogeny. The estimator is given as:

$$\hat{I} = \frac{n}{S_0} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} v_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where y_i is the trait value of species i and $\bar{y}$ the average trait value. The heart of this statistic is the weighting matrix $\mathbf{V} = [v_{ij}]$ where v_{ij} describes the phylogenetic proximity between species i and j. The sum of all pairwise weights is S_0. Moran's I is related to Pearson's correlation coefficient: the numerator is a covariance that compare the values of all pairs of values, while the denominator is the maximum likelihood estimator of the variance. For more information, see Legendre and Legendre (1998).

Moran's I is very flexible, because different types of proximity matrices can be used to describe the phylogenetic information (e.g. Pavoine et al. 2008). Very often, proximities are computed as the inverse of the patristic distances, with v_{ii} equal to zero.

Moran's I can be estimated with the function `Moran.I` of the `ape` package.

The null hypothesis of no phylogenetic correlation is tested assuming normality of I under this null hypothesis. If the observed value of I is significantly greater than the expected value, then the values of x are positively autocorrelated, whereas if $I_{observed} < I_{expected}$, this will indicate negative autocorrelation.

Example

Let's calculate Moran's I to the seed plant data.

```
require(ape)
# Compute a matrix of weights, here the inverse of the patristic distance
w <- 1/cophenetic(seedplantstree)
# Set the diagonal to w[i,i] = 0 (instead of Inf)
```

```r
diag(w) <- 0
# Compute Moran's I for Wood density
Moran.I(seedplantsdata$Wd, w)
```

```
## $observed
## [1] 0.3612186
##
## $expected
## [1] -0.01785714
##
## $sd
## [1] 0.08029841
##
## $p.value
## [1] 2.348756e-06
```

```r
# Compute Moran's I for all characters
res <- apply(seedplantsdata,2,function(x) Moran.I(x,w))
MoranI <- matrix(unlist(res),ncol=4,byrow=TRUE,
            dimnames=list(colnames(seedplantsdata),c("observed","expected","sd","p.value")))
MoranI
```

```
##        observed    expected         sd      p.value
## maxH  0.2232017 -0.01785714 0.08011839 2.622970e-03
## Wd    0.3612186 -0.01785714 0.08029841 2.348756e-06
## Sm    0.4159100 -0.01785714 0.06739615 1.225968e-10
## Shade 0.2234985 -0.01785714 0.08020264 2.618306e-03
## N     0.4773086 -0.01785714 0.07979762 5.460337e-10
```

You can see that all variables are positively autocorrelated.

Abouheif's c

This is one of the first statistic described to estimate the phylogenetic signal in a parameter (Abouheif 1999). Pavoine et al. (2008) have shown that Abouheif's c is actually a Moran's I statistic with a special phylogenetic proximity estimate between tips. It can be estimated with the function `abouheif.moran` of the `adephylo` package.

Example

```r
require(adephylo)
# Compute the Abouheif matrix of weights for the tree
w <- proxTips(seedplantstree,method="oriAbouheif")
# Let's compute abouheif stat for all characters
abouheif.moran(seedplantsdata, W=w)
```

```
## class: krandtest
## Monte-Carlo tests
## Call: as.krandtest(sim = matrix(res$result, ncol = nvar, byrow = TRUE),
```

```
##     obs = res$obs, alter = alter, names = test.names)
##
## Number of tests:   5
##
## Adjustment method for multiple comparisons:   none
## Permutation number:   999
##    Test       Obs  Std.Obs   Alter Pvalue
## 1  maxH 0.3297438 3.612764 greater  0.001
## 2    Wd 0.5209647 5.774519 greater  0.001
## 3    Sm 0.6076848 7.782505 greater  0.001
## 4 Shade 0.3916910 4.140135 greater  0.001
## 5     N 0.6526499 7.107602 greater  0.001
##
## other elements: adj.method call
```

Again, the results show that all parameter have significant phylogenetic signal.

Pagel's λ

Pagel's approach to estimate phylogenetic signal is relatively simple. It is based on the idea that under the Brownian Motion model of evolution, the expected covariance matrix between traits is perfectly defined by the phylogenetic tree. Pagel (1999) has introduced a branch transformation that can downweight the importance of the expected Brownian phylogenetic covariances to fit the observed ones. The λ parameter defines this weight. The modified branch lenght d_i^* for branch i is:

$$d_i^* = \lambda d_i,$$

where d_i is the original length of branch i. In general, λ can take a value between 0 and 1. When $\lambda = 1$, it means that the branch lengths are unaffected and the model correspond to the Brownian model. At the opposite, $\lambda = 0$ means that branch lengths will equal 0, resulting in a more star-like phylogeny. The value of λ is fitted by maximum likelihood. In `R`, this can be done using the `fitContinuous` function of the `geiger` package.

To test for the presence of phylogenetic signal, one can use model comparison. That is, you can compare the fit of the model with the optimmized λ value to a model with a fixed λ. For instance, if you want to test if there is phylogenetic signal in your dataset, you can compare with a model where $\lambda = 0$, and compare the results with a likekihood ratio test or the AIC.

Example

```
library(geiger)
# Let's start by estimating lambda for all traits
lambdafit <- fitContinuous(seedplantstree, seedplantsdata, model="lambda")
# To print a more detail results table, it is useful to make a loop
traits <- colnames(seedplantsdata)
lambda <- numeric(length(traits))
lnL <- numeric(length(traits))
aicc <- numeric(length(traits))
params <- numeric(length(traits))
for (i in 1:length(traits)) {
  res <- fitContinuous(seedplantstree, seedplantsdata[traits[i]], model="lambda")
```

```
  lambda[i]=res$opt$lambda
  lnL[i]=res$opt$lnL
  aicc[i]=res$opt$aicc
  params=res$opt$k
}
(results <- data.frame(traits=traits,lambda=lambda,lnL=lnL,AICc=aicc,df=params))
```

```
##   traits    lambda        lnL       AICc df
## 1   maxH 0.4915358 -206.66282  419.77848  3
## 2     Wd 0.7630357   58.99352 -111.53422  3
## 3     Sm 0.9084594 -521.90987 1050.27258  3
## 4  Shade 0.9483664  -83.06946  172.59175  3
## 5      N 0.6661551  -18.06712   42.58707  3
```

This is nice, but it doesn't say if it is significant or not. To test this, we need to compare the model fit with that of a model in which λ is fixed to 0. To do this, we will reshape the phylogeny as we saw in the last lecture. The trick is to reshape the phylogeny using a λ model with value 0, and then fit a BM model on this phylogeny. This will git the fit of the model with $\lambda = 0$.

```
library(geiger)
# Reshape the tree with lambda = 0
tree.lambda0 <- rescale(seedplantstree,model="lambda",0)
# Now fit a BM on this tree to get the fit for lambda = 0
lnL.0 <- numeric(length(traits))
aicc.0 <- numeric(length(traits))
params.0 <- numeric(length(traits))
for (i in 1:length(traits)) {
  res <- fitContinuous(tree.lambda0, seedplantsdata[traits[i]], model="BM")
  lnL.0[i]=res$opt$lnL
  aicc.0[i]=res$opt$aicc
  params.0=res$opt$k
}
# Results with lamnda=0
(results.0 <- data.frame(traits=traits,lnL=lnL.0,AICc=aicc.0,df=params.0))
```

```
##   traits        lnL       AICc df
## 1   maxH -206.95993  418.14208  2
## 2     Wd   52.21186 -100.20149  2
## 3     Sm -531.18357 1066.58935  2
## 4  Shade  -88.30486  180.83195  2
## 5      N  -41.10639   86.43501  2
```

```
#
# Perform statistical tests
#
# Compute likelihood ratio test for all variables
pval <- apply(matrix(c(lnL,lnL.0),ncol=2),1,function(x) {1-pchisq(2*(x[1]-x[2]),1)})
# Compute delta AICc (negative value support lambda=0)
d.aicc <- aicc.0 - aicc
# Print the results
data.frame(traits=traits,lambda=lambda,p.value=round(pval,4),delta.AICc=round(d.aicc,2))
```

```
##   traits    lambda p.value delta.AICc
## 1   maxH 0.4915358  0.4408      -1.64
## 2     Wd 0.7630357  0.0002      11.33
## 3     Sm 0.9084594  0.0000      16.32
## 4  Shade 0.9483664  0.0012       8.24
## 5      N 0.6661551  0.0000      43.85
```

This table give the estimated λ parameter, the p-value in favor of the more complex model (that is $\lambda \neq 0$), and the $\Delta AICc$ where a positive value supports the more complex model (that is $\lambda \neq 0$). In this case, the two approaches give the same conclusions. All paramters are found to have significant phylogenetic signal, except for maximum height.

Blomberg's K

Blomberg et al. (2003) suggested to estimate the phylogenetic signal as the ratio of MSE_0, the *mean squared error* of the tip data to the phylogenetically corrected mean (the value at the root of the tree), to MSE, the *mean squared error* of the data to the values expected under Brownian Motion model. When trait values is well predicted by the phylogeny, MSE will be small and thus $\frac{MSE_0}{MSE}$ large.

To make the resulting value comparable to other trees with different sizes and shapes, this ratio is standardized by the analytically derived expectation for the ratio under BM evolution. K is computed as:

$$K = \frac{\text{observed}\frac{MSE_0}{MSE}}{\text{expected}\frac{MSE_0}{MSE}}$$

Blomberg's K can be estimated using the function `phylosignal` of the `picante` package.

Example

```
library(picante)
# With the phylosignal function, it is important that the order of
# matrix rows is the same as for the tip labels.
#
# Example with one character
phylosignal(seedplantsdata$Wd, seedplantstree)
```

```
## Warning in match.phylo.data(phy, x): Data set lacks taxa names, these are
## required to match phylogeny and data. Data are returned unsorted. Assuming
## that data and phy$tip.label are in the same order!
```

```
##           K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P
## 1 0.0469481         2.339215              10.00329          0.001
##   PIC.variance.Z
## 1      -2.486523
```

```
# Let's compute Blomberg's K for all characters
res <- apply(seedplantsdata,2,function(x) phylosignal(x,seedplantstree))
K <- matrix(unlist(res),ncol=5,byrow=TRUE,
            dimnames=list(colnames(seedplantsdata),
              c("K","PIC.variance.obs","PIC.variance.rnd.mean",
                "PIC.variance.P","PIC.variance.Z")))
K
```

```
##                K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P
## maxH  0.01885410     4.573630e+04          8.986838e+04          0.035
## Wd    0.04694810     2.339215e+00          1.041145e+01          0.001
## Sm    0.07088576     1.126973e+09          8.062127e+09          0.008
## Shade 0.03222634     4.134031e+02          1.415080e+03          0.001
## N     0.07701311     4.231345e+01          2.757694e+02          0.001
##       PIC.variance.Z
## maxH       -1.471372
## Wd         -2.411594
## Sm         -1.116067
## Shade      -2.158602
## N          -2.422743
```

The results show that all paramter have significant phylogenetic signal. Maximum height is only marginally significant though, siggesting a weeker phylogenetic signal.

Moran's I correlogram

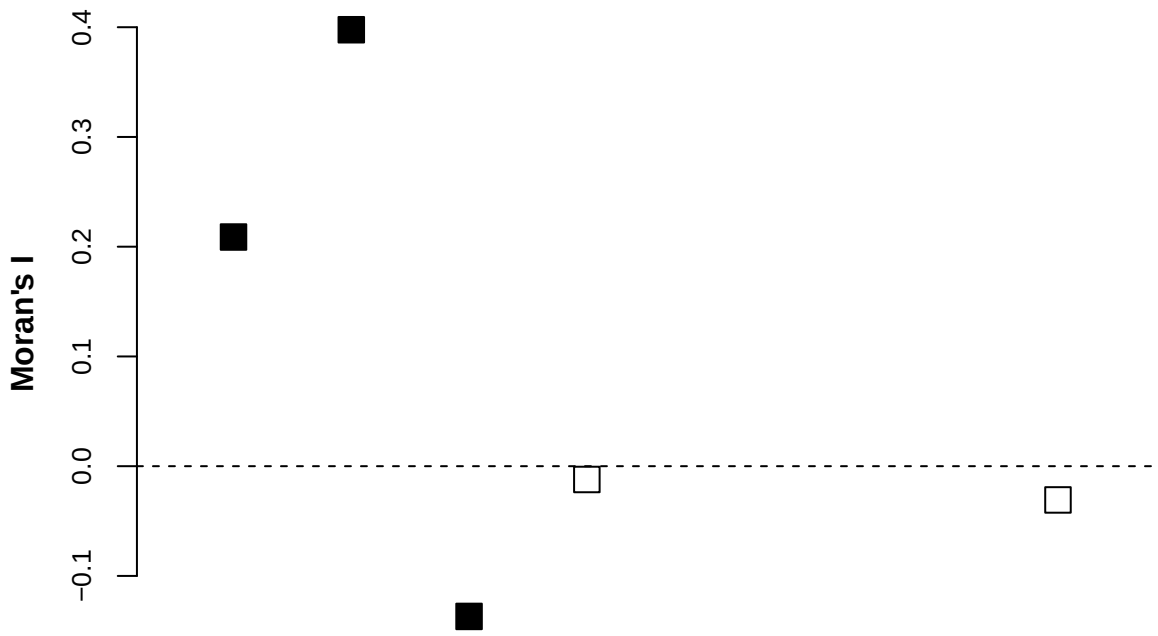
Instead of getting one value of phylogenetic signal per trait for the whole phylogeny, it might be of interest to test the phylogenetic signal at different depths of the tree. This can be done using Moran's I correlograms (Paquette et al. 2015). This is essentially the same idea as correlograms used in landscape ecology to describe the spatial structure of a trait (Legendre and Legendre 1998, chapter 13). The idea is to divide the tree into slices of different distances and evaluate Moran's I for only the species grouping within a given slice at a time. Basically, you first compare the species that have diverged very recently, then species that diverged for a sightly longer time and so on... This allows to check if species are more or less similar than expected at a given distance class (Paquette et al. 2015).

To perform a Moran's I correlogram, it is important to first load the function "moran.phylo.cor".

```
source("./moran.phylo.cor.R")
```

Once this is done, we can make a Moran's I autocorrelogram for seed mass.

```
# Matrix of phylogenetic weights
w <- cophenetic(seedplantstree)
# Morran's I autocorrelogram
moran.phylo.cor(seedplantsdata$Sm,w,breaks=8,plot=TRUE)
```




```
## $values
## [1]  0.20881886  0.39770741 -0.13688168 -0.01210471          NA          NA
## [7]          NA -0.03075563
##
## $significance
## [1] 9.018141e-03 6.823112e-08 1.458902e-02 9.107028e-01 1.000000e+00
## [6] 1.000000e+00 1.000000e+00 6.031775e-01
##
## $sd
## [1] 0.08680362 0.07701676 0.04873214 0.05129136 1.00000000 1.00000000
## [7] 1.00000000 0.02481259
##
## $breaks
## [1] 0.00000000 0.03766363 0.07532726 0.11299089 0.15065452 0.18831815
## [7] 0.22598177 0.26364540 0.30130903
##
## $mids
## [1] 0.01883181 0.05649544 0.09415907 0.13182270 0.16948633 0.20714996
## [7] 0.24481359 0.28247722
```

The plot shows Moran's I for the different distance classes. The black squares indicate a significant correlation (p-value <0 0.05), either positive or negative. The correlogram indicates that species within small phylogenetic distance classes (recently diverged) are significantly positively correlated, whereas species of the third distance class are negatively correlated, indicating that they are more different than expected by chance. At large distance classes, seed mass between species is not correlated, showing an absence of phylogenetic signal.

These correlograms are interesting because they can show that there can be phylogenetic signal at some distance classes, but not at other. This information is impossible to get when estimating a single value for the whole phylogeny.

Parametric bootstrapping

We saw in a previous lecture that Bayesian approaches can give an idea of the amount of information present in the data to estimate the parameter of interest. This is also possible with a frequentist approach using parametric bootstrapping. The concept is to first fit the model on the data. Once the ML parameter estimates are obtained, you then simulate new datasets similar to the empirical one using these parameter values. Then, you re-estimate the parameters of the model from the simulated datasets. If the parameter estimated from the simulated data are close the the values used in the simulation, it means that the model and data is adequate to estimate these parameters. If the estimated values are far from the values used in the simulations, then it means that either the model is wrong or that there insufficient information in the data (too few species, for intance) to properly estimate de parameters of the model.

Let's look at an example from the seed plant data. We will first estimate lambda for the wood density data, and then simulate 250 new datasets similar to the empirical one (i.e., same number of species on the phylogeny). Ideally, 1000 replicates would give a more precise result, but we will use only 250 here because it is a bit long to fit λ on all replicates (you should also try on fewer replicates (ca. 50) to test the function).

```
Wdfit <- fitContinuous(seedplantstree, seedplantsdata['Wd'], model="lambda")
# Fitted parameters
Wdfit$opt
```

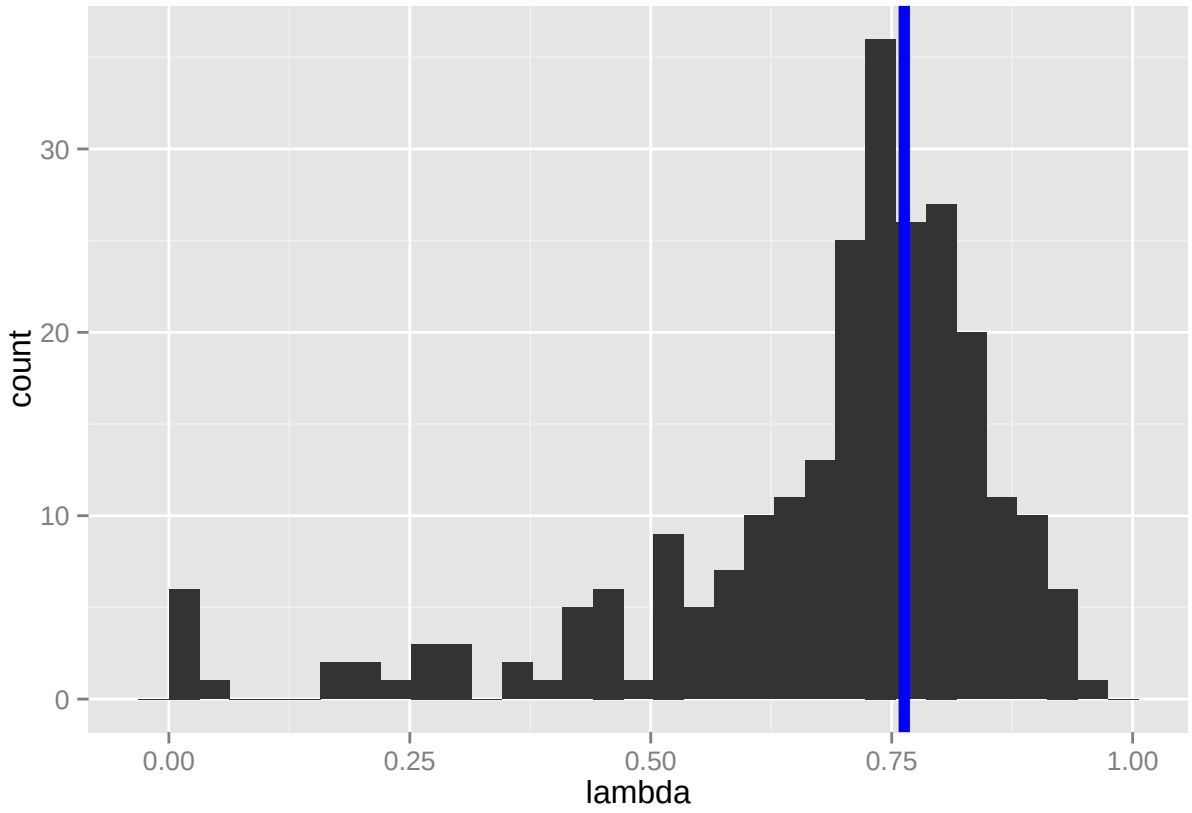
```
## $lambda
## [1] 0.7630403
##
## $sigsq
## [1] 0.122474
##
## $z0
## [1] 0.4348003
##
## $lnL
## [1] 58.99352
##
## $method
## [1] "subplex"
##
## $k
## [1] 3
##
## $aic
## [1] -111.987
##
## $aicc
## [1] -111.5342
```

```
#
# Simulate new datasets
require(phytools)
# Reshape the tree according to the fitted lambda that will be used in the
# simulationd.
lambdatree <- rescale(seedplantstree,model="lambda",Wdfit$opt$lambda)
# Use the lambdatree to simulate new data using the fitted values at the
# root of the tree and the fitted sigma squared value. We will make
```

```
# 250 simulations
simdata <- fastBM(lambdatree,a=Wdfit$opt$z0,sig2=Wdfit$opt$sigsq,nsim=250)
```

Now that we have simulated data, we will estimate λ for all of these datasets and we will compare the values estimated from the simulations from the original ones.

```
lambdas <- numeric(250)
for (i in 1:250){
  # We will focus only on the lambda value, so we will only extract this value.
  # This is why we all the '$opt$lambda' at the end of the fitContinuous
  # command
  lambdas[i] <- fitContinuous(seedplantstree, simdata[,i],
                              model="lambda")$opt$lambda
}
# Plot the distribution of values
require(ggplot2)
lambdadata <- data.frame(lambda=lambdas)
ggplot(lambdadata,aes(x=lambda)) + geom_histogram() + geom_vline(xintercept = Wdfit$opt$lambda, col="blu
```



With this specific example, you can see that the lambda values estimated from the simulated data are most often similar to the original λ, but that there is nevertheless considerable uncertainty. Indeed, in some cases, you can obtain values of 0 for λ. Paramtetric bootstrapping is commonly used to estimate confidence intervals for the estimation of the parameter. Thus we can look at the 2.5% and 97.5% quantiles of the distribution to obtain a 95% confidence interval.

```
# Get 95% confidence inbtervals
quantile(lambdas,probs=c(0.025,0.5,0.975))
```

```
##       2.5%        50%      97.5%
## 0.08275267 0.73519478 0.91277460
```

From this, we can conclude that there is 95% of probability that the true value of lambda falls between 0.083 and 0.913, which as you can see is pretty large.

Phylogenetic diversity

Phylogenetic diversity is a measure of biodiversity that incorporates phylogenetic difference between species. There are several different definitions of phylogenetic diversity and several ways to estimate it. Here we will only see a few method and instead focus on the concept.

The concept of phylogenetic diversity is to provide more information than a simple species diversity index. For instance, two communities might have exactly the same number of species, but one might have species that are more evolutionary distant from one another than the other community. For example, a given field might have three species of plants, three maple species, whereas the other also has three species, but in that case these consist of a maple, a beech and an ash tree. Clearly, the latter possess species that represent a more evolutionary diverse community. PD was proposed to incorporate these notions in conservation biology. Since then, these concepts have also been used in community ecology (see below).

Faith's PD

The original definition of phylogenetic diversity (PD) was proposed by Faith (1992). Formally, Phylogenetic diversity is defined as the *Sum of all branch lengths in the portion of a phylogenetic tree connecting the focal set of species.*

Faith's PD can be calculated using the `picante` package in R. `picante` uses community matrices to calculate PD. These standardly consist in a matrix in which species are in columns and communities (or region for biogeography applications) are in rows. To show an example of application, we will use the example provided with the `picante` package.

```
require(picante)
# Example data set
data(phylocom)
# Look at what it contains
str(phylocom)
```

```
## List of 3
##  $ phylo :List of 5
##   ..$ edge       : num [1:62, 1:2] 33 34 35 36 37 37 36 38 38 35 ...
##   ..$ tip.label  : chr [1:32] "sp1" "sp2" "sp3" "sp4" ...
##   ..$ Nnode      : int 31
##   ..$ node.label : chr [1:31] "A" "B" "C" "D" ...
##   ..$ edge.length: num [1:62] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "class")= chr "phylo"
##  $ sample: num [1:6, 1:25] 1 1 1 1 1 0 0 2 0 1 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:6] "clump1" "clump2a" "clump2b" "clump4" ...
##   .. ..$ : chr [1:25] "sp1" "sp10" "sp11" "sp12" ...
##  $ traits:'data.frame':  32 obs. of  4 variables:
##   ..$ traitA: int [1:32] 1 1 2 2 2 2 2 2 1 1 ...
##   ..$ traitB: int [1:32] 1 1 1 1 2 2 2 2 3 3 ...
```

```
##   ..$ traitC: int [1:32] 1 2 3 4 1 2 3 4 1 2 ...
##   ..$ traitD: int [1:32] 0 0 0 0 0 0 0 0 1 1 ...
```

```
# Here is how the community matrix look like
head(phylocom$sample,n=10)
```

```
##         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
```

```
# Now calculate PD
pd(phylocom$sample, phylocom$phylo, include.root=TRUE)
```

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

The PD results in a two column matrix. The first column is Faith's PD and the second is the species richness, that is the number of species in the sample. In that example, all species have the same number of species, but different PD.

> By default, PD includes the root in the calculations. If you want to exclude the root, you can indicate `include.root=FALSE` in the function. Also, note that the PD values are function of the tree length. If you change the units, for instance from nucleotide substitions per site per year to years, the results will be different.

Phylogenetic species variability (PSV)

Helmus et al. (2007) have proposed new metrics for estiamting PD. Of these, an interesting one is the Phylogenetic Species Variability (PSV), because it is unrelated to species richness. It estimates how phylogenetic relatedness decreases the variance of a hypothetical trait shared by all species in a community. PSV varies between 0 and 1. A value of 1 means that species in a community are as unrelated as possible (maximum diversity) whereas a value closer to 0 means that species are more closely related.

```
psv(phylocom$sample, phylocom$phylo)
```

```
##              PSVs SR        vars
## clump1  0.4857143  8 0.001055303
## clump2a 0.6000000  8 0.001055303
## clump2b 0.7142857  8 0.001055303
## clump4  0.8285714  8 0.001055303
## even    0.8857143  8 0.001055303
## random  0.8428571  8 0.001055303
```

Other measures of PD

Other measures of PD have been proposed, many of which have been described by Cadotte et al. (2010). Some are more interesting for phylogenetic (conservation) studies whereas others are more interesting for community studies.

Evolutionary distinctiveness

Phylogenetic diversity is very present in the conservation literature, but there are other ways phylogenies can be incorporated into statistic that could be useful for conservation. One statistic that is relatively popular is Evolutionary Distinctiveness (ED: Redding and Mooers, 2006). ED represents the evolutionary history that is unique to a species in a given sample. Practically, it is the length of the branch on the tree that links the species to the rest of the tree (i.e., the length of the terminal branch). So species that are on longer branches have higher evolutionary distinctiveness. Another way to look at it is that *living fossils* such as gingkos or coleocanths will have high ED compared to dandelions or finches, respectively.

```
evol.distinct(seedplantstree)
```

```
##    Species           w
## 1     ABBA 0.029174140
## 2     ACNE 0.030261596
## 3     ACNI 0.003488247
## 4     ACPE 0.006411584
## 5     ACPL 0.006599221
## 6     ACRU 0.008731284
## 7     ACSA 0.003488247
## 8     ACSI 0.008731284
## 9     ACSP 0.006411584
## 10    ALCR 0.007522701
## 11    ALRU 0.007522701
## 12    AMSP 0.005294702
## 13    BEAL 0.003544232
## 14    BEPA 0.003544232
## 15    BEPO 0.006818631
## 16    CACA 0.009995641
## 17    CACO 0.007823476
## 18    CAOV 0.007823476
## 19    COAL 0.064461672
## 20    CRSP 0.005294702
## 21    FAGR 0.027728007
```

```
## 22    FRAM 0.016285318
## 23    FRNI 0.032435578
## 24    FRPE 0.016285318
## 25    JUCI 0.008278839
## 26    JUNI 0.008278839
## 27    JUVI 0.061626058
## 28    LALA 0.067293233
## 29    MASP 0.008974957
## 30    OSVI 0.009995641
## 31    PIAB 0.004417619
## 32    PIBA 0.013330569
## 33    PIGL 0.016906566
## 34    PIMA 0.008691446
## 35    PIRE 0.013330569
## 36    PIRU 0.004417619
## 37    PIST 0.021357273
## 38    PLOC 0.111236528
## 39    POBA 0.012785160
## 40    PODE 0.020209672
## 41    POGR 0.007734641
## 42    POTR 0.007734641
## 43    PRPE 0.007591076
## 44    PRSE 0.014934847
## 45    PRVI 0.007591076
## 46    QUAL 0.007792581
## 47    QUBI 0.003947227
## 48    QUMA 0.003947227
## 49    QURU 0.015063189
## 50    SASP 0.034292872
## 51    SOAM 0.015405301
## 52    THOC 0.061626058
## 53    TIAM 0.055606375
## 54    TSCA 0.029174140
## 55    ULAM 0.010754483
## 56    ULRU 0.021345906
## 57    ULTH 0.010754483
```

Phylogenetic beta diversity

Diversity can be divided in three components: α diversity, which represents the diversity ofberved at a given site, β diversity, which represents the diversity among sites, and γ diversity, which represents the total diversity. The PD statistic can provide a measure of the α or γ diversity. However, it is sometimes of interest to measure the amount of variation between sites, that is the β diversity.

Several measures of β phylogenetic diversity have been proposed (Swenson 2011), although several of these are very similar (Sewnson 2011). We will see a popular statistic here, that is *Phylosor*, which converge to the Sorensen Index when there is no phylogenetic information. *Phylosor* is defined as

$$Phylosor = \frac{BL_{k_1 k_2}}{(BL_{k1} + BL_{k2}) \times \frac{1}{2}}$$

where $BL_{k_1 k_2}$ is the total length of the branches shared between community k_1 and k_2, BL_{k_1} and BL_{k_2} are the total branch lengths found in communities k_1 and k_2 respectively.

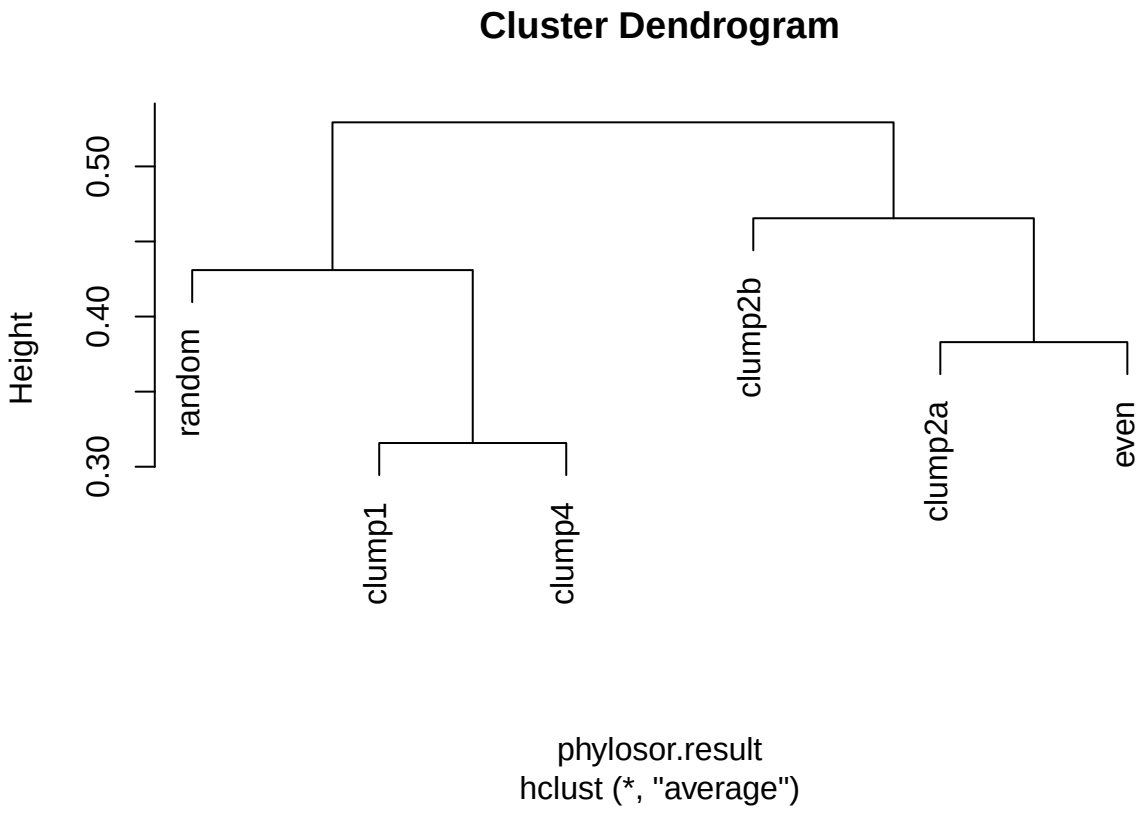
It can be estimated using the `phylosor` function in the `picante` package.

```
phylosor.result <- phylosor(phylocom$sample,phylocom$phylo)
phylosor.result
```

```
##            clump1   clump2a   clump2b    clump4      even
## clump2a 0.5454545
## clump2b 0.5294118 0.5142857
## clump4  0.3157895 0.5641026 0.6000000
## even    0.3478261 0.3829787 0.4166667 0.6923077
## random  0.3720930 0.4090909 0.4444444 0.4897959 0.6315789
```

This similarity index can be used to represent the similarity between communities using a phenogram.

```
library(cluster)
#UPGMA clustering
phylosor.clusters <- hclust(phylosor.result,method="average")
plot(phylosor.clusters)
```



Phylogenies in community ecology

In the last 15 years, phylogenetic information has been increasingly used in the field of community ecology. This whole field has really started up with the seminal publication of Webb et al. (2002). The concept is to apply some of the methods we saw in this lecture and infer possible processes that could have shape the community composition. The following table whos how patterns are generally inferred according to the phylogenetic clustering of traits and communities.

Dominant ecological force	Traits conserved	Traits labile
Habitat filterig (phenotypic attraction)	Clustered community	Overdispersed community
Competitive exclusion (phenotypic repulsion)	Overdispersed community	random community

To determine the dominant ecological force, you need to first need to evaluate how related are the average pair of species or individuals in a community and compare this value to what we would expect to have under null models of evoluion or community assembly. In community ecology, Faith's PD is rarely used. More frequently, the measures of community phylogenetic structure used are MPD or MNTD (Webb et al. 2002).

MPD is the *mean pairwise distance* between all species in each community, whereas MNTD is the *mean nearest taxon distance*, that is the the mean distance separating each species in the community from its closest relative. 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. They can be computed with the functions `MPD` and `mntd` of the `picante` package. These functions take a community matrix and a phylogenetic distance matrix. This distance matrix can be obtained from the patristic distances between species along the tree, obtained using the `cophenetic` function.

Once these statistics are estimated, we need to compare them with that expected with some null model of evolution or community randomization. These *Standardized effect sizes* (SES) 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:

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

Two very similar statistics can be found in the literature, *NRI* and *NTI*. They can be obtained from SES_{MPD} and SES_{MNTD} using the following formulas:

$$SES_{MPD} = -1 \times NRI$$

$$SES_{MNTD} = -1 \times NTI$$

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 of `picante`.

Here is an example:

```
phydist <- cophenetic(phylocom$phy)
ses.mpd.result <- ses.mpd(phylocom$sample, phydist, null.model = "taxa.labels",
abundance.weighted = FALSE, runs = 999)
ses.mpd.result
```

```
##         ntaxa  mpd.obs mpd.rand.mean mpd.rand.sd mpd.obs.rank   mpd.obs.z
## clump1      8 4.857143      8.320749   0.3164771          1.0 -10.9442558
## clump2a     8 6.000000      8.317746   0.3299487          1.0  -7.0245647
## clump2b     8 7.142857      8.330831   0.3333197         10.5  -3.5640672
## clump4      8 8.285714      8.333762   0.3110539        369.0  -0.1544685
## even        8 8.857143      8.311526   0.3282980        996.0   1.6619566
## random      8 8.428571      8.323967   0.3231812        542.0   0.3236717
##         mpd.obs.p runs
```

```
## clump1     0.0010  999
## clump2a    0.0010  999
## clump2b    0.0105  999
## clump4     0.3690  999
## even       0.9960  999
## random     0.5420  999
```

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 11
- 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. In the previous example, the communities `clump1`, `clump2a` and `clump2b` a significantly overdispersed ($p<0.05$), whereas community `clump4` is slightly clustered, but not significantly ($p=0.35$).

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.

> Note that the `comdist` function in `picante` estimate a beta phylodiversity index that is the equivalent to the MPD statistic.

References

Abouheif E. 1999. A method for testing the assumption of phylogenetic independence in comparative data. *Evol Ecol Res.* 1:895–909.

Blomberg S.P., Garland T., Ives A.R. 2003. Testing for phylogenetic signal in phylogenetic comparative data: behavioral traits are more labile. *Evolution.* 57:717–745.

Boettiger C., Coop G., Ralph P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. *Evolution.* 66:2240–2251.

Cadotte M.W., Jonathan Davies T., Regetz J., Kembel S.W., Cleland E., Oakley T.H. 2010. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecology Letters.* 13:96–105.

Faith D.P. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation.* 61:1–10.

Gittleman, J. L. and Kot, M. (1990) Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, 39, 227–241.

Helmus M.R., Bland T.J., Williams C.K., Ives A.R. 2007. Phylogenetic measures of biodiversity. *The American Naturalist.* 169:68–83.

Legendre P., Legendre L. 1998. *Numerical Ecology, second english edition.* Amsterdam: Elsevier.

Münkemüller T., Lavergne S., Bzeznik B., Dray S., Jombart T., Schiffers K., Thuiller W. 2012. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*. 3:743–756.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature*. 401:877–884.

Paquette A., Joly S., Messier C. 2015. Explaining forest productivity using tree functional traits and phylogenetic information: two sides of the same coin over evolutionary scale? *Ecol Evol(. 5:1774–1783.

Pavoine S., Ollier S., Pontier D., Chessel D. 2008. Testing for phylogenetic signal in phenotypic traits: New matrices of phylogenetic proximities. *Theoretical Population Biology*. 73:79–91.

Redding, D.W. and Mooers, A.O. (2006). Incorporating evolutionary measures into conservation prioritisation. *Conservation Biology*, 20, 1670-1678.

Swenson N.G. 2011. Phylogenetic Beta Diversity Metrics, Trait Evolution and Inferring the Functional Beta Diversity of Communities. *PLoS ONE*. 6:e21264.

Webb C.O., Ackerly D.D., McPeek M.A., Donoghue M.J. 2002. Phylogenies and community ecology. *Annu. Rev. Ecol. Syst.* 33:475–505.