HETEROGENEOUS DATA CHALLENGES
COMBINING TREES WITH OTHER DATA

Susan Holmes
http://www-stat.stanford.edu/~susan/

Bio-X and Statistics, Stanford University

SISMID14-Lecture 11, 2011

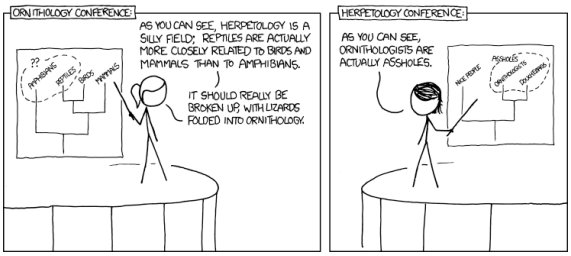
'Homogeneous data are all alike;

all heterogeneous data are heterogeneous

in their own way.'

Part I

Using Trees

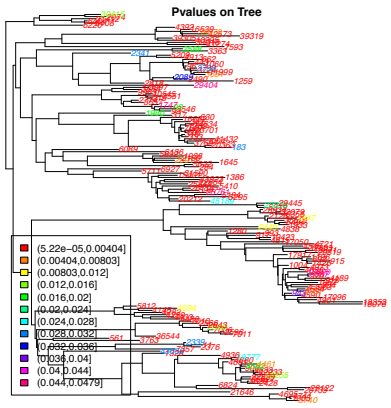


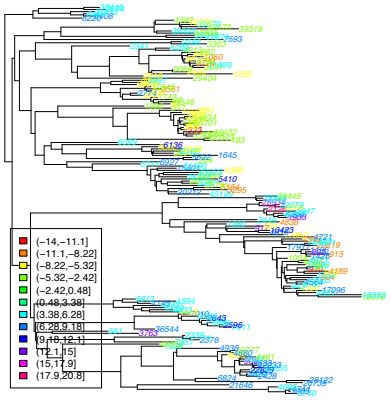
Manipulating Trees

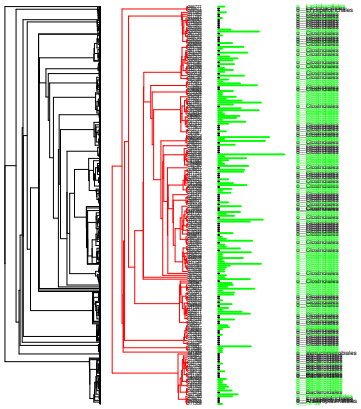
```
Main tool:  ape.
 Example:
> treeall=read.tree("/Users/susan/Dropbox/Yana/gg.tr
> todrop=which(!(treeall$tip.label %in% otu.id))
>
> todrop=which(!(treeall$tip.label %in% otu.id))
> dropthese=treeall$tip.label[todrop]
> treespecific=drop.tip(treeall,dropthese)
>layout(matrix(c(1,1,1,2,2,2,3,4,5,5,5,5),ncol=12,n
>subtree = drop.tip(tree, which(!tree$tip.label %in%
>plot(tree, use.edge.length=F, no.margin=T, edge.col
        show.tip.label=F)
###Relevant subtree
>plot(subtree, edge.color='red', show.tip.label=T, n
       use.edge.length=F)
```

Mapping Variables onto Phylogenies

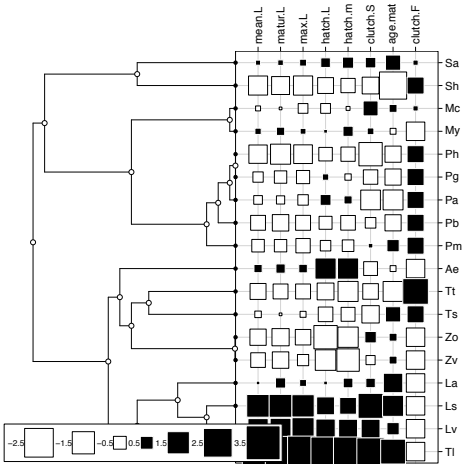
Example using ape and picante packages
color.plot.phylo(phyl,df,"tstat","otus")







Example of using the table.phylog function




```
data(lizards)
#Variables of lizards$traits are the following
#mean.L(mean length (mm)), matur.L (length at maturit
#max.L(maximum length (mm)), hatch.L (hatchling lgth
#hatch.m (hatchling mass (g)), clutch.S (Clutch size)
#age.mat (age at maturity (number of months of activit
# clutch.F (clutch frequency).
$hprA
[1] "((Sa:17,Sh:17):16,((((Tl:17,(Mc:1,My:1):16):1,
((((Ph:0.02,Pg:0.02):0.98,Pa:1):2,Pb:3):2,Pm:5):13)
:4,(Ae:20,(Tt:15,Ts:15):5):2):2,((Zo:0.1,Zv:0.1):21
(La:10,(Ls:5,Lv:5):5):12):2):9);"
w <- data.frame(scalewt(log(lizards$traits)))
par(mfrow = c(1,2))
wphy <- newick2phylog(lizards$hprA)
table.phylog(w, wphy, csi = 3)
```

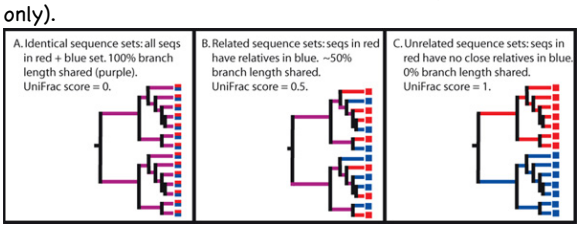
Phylogenetic Information

A distance on a tree is still a distance, unifrac is a distance, these can still be decomposed and ``correlated'' to other factors in the same way.
Simple tests in `picante`, `vegan` and `ape` using the phylogenetic distance between communities.
Phylogenetic diversity is a sort of inertia, it can be decomposed in the same way.

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

Using the Tree to provide the distances between samples

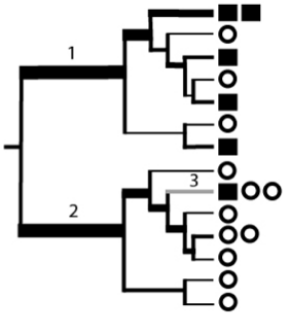
The idea between unifrac [9].
Suppose we have the OTUs present in sample 1 (blue) and in sample 2(red).
Question: Do the two samples differ phylogenetically?
We will start with the unweighted version (presence absence only).



Unweighted Unifrac

The distance is calculated between pairs of samples (each sample represents a bacterial community). All taxa found in one or both samples are placed on a phylogenetic tree. A branch leading to taxa from both samples is marked as "shared" and branches leading to taxa which appears only in one sample are marked as "unshared".
The distance between the two samples is then calculated as (the sum of "unshared" branch lengths)/(the sum of all tree branches (= shared+unshared)), i.e. the fraction of total branch length.
If there are several different samples (different parts of the body), a distance matrix can be created, by making a tree for each pair of samples and calculating their UniFrac measure.

Weighted Unifrac



Weighted Unifrac

$$d_{wu}(A,B) = \sum_{i\ \text{branch}}^{n} b_i \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

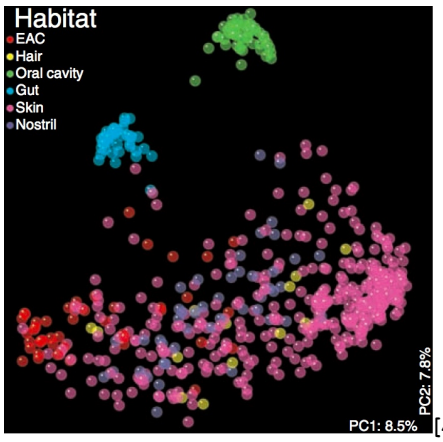
Here, n is the total number of branches in the tree, b_i is the length of branch i, A_i and B_i are the number of descendants of branch i from communities A and B respectively, and A_T and B_T are the total number of sequences from communities A and B respectively. In order to control for unequal sampling effort, A_i and B_i are divided by A_T and B_T.
[7] or can be seen as by probabilists as the Wasserstein distance (earth movers) [6].

Normalized Weighted Unifrac

$$d_{nwu}(A,B) = \frac{\sum_i^n \text{branch}\ b_i \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|}{\sum_j^s \text{seq tip}\ d_j \left| \frac{A_j}{A_T} + \frac{B_j}{B_T} \right|}$$

Where d_j is the distance from the root to the sequence j.
This accounts for problems where some the branches of the tree are very long (higher resolution) causing the tree not to be ultrametric.

Multidimensional Scaling of Unifrac Distances



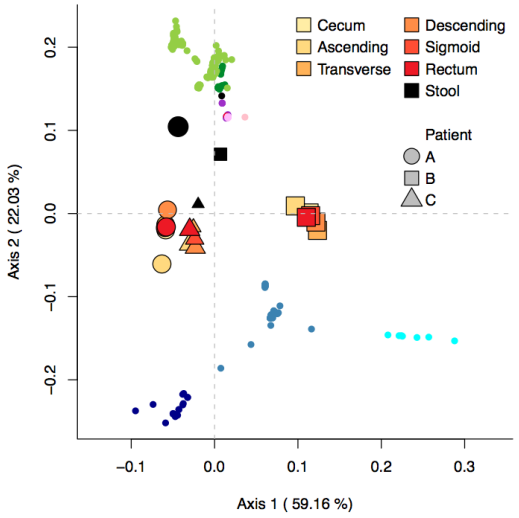
[4]

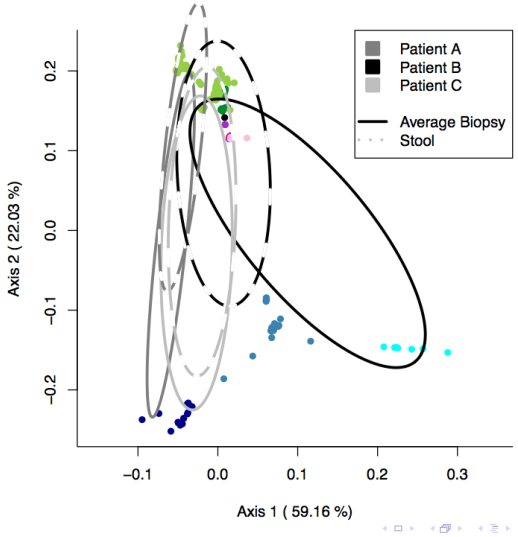
More refined use of distances on a Tree

DPCO: double coordinate analysis.
Pavoine et al. (2004) invented Double Principal Coordinates Analysis (DPCoA), to take into account a supplementary distance known the same variables as the columns of a matrix X. This distance is called Q_p.
So we suppose we are given a triplet (X, Q_p, Q_n).
Before performing a PCA type analysis a little care is taken in centering the matrix X
Q_p is used in a MDS analyses with weights providing a new set of coordinates of the original p variables in which X was measured.
The same transformation is given for the new basis say by XZ on which a PCA is done.
Available as the dpcoa function in `ade4`.

Application to Gastric Ecosystem

This was used to explore intersubject variability of the gastric ecosystem by Purdom, 2010, Annals of Applied Statistics [10]. Bik et al 2006 [1] and Eckburg et al.[5]





Permutation Tests from Trees

- Monte Carlo distribution of unifrac under reshuffling of labels.
- Mantel test for two distances.

Over-representation of certain phyla

Set	Over-represented	Universe
Microbiome	Families/Phyla	Species Present
Gene Expression	Ontological groups	Filtered Genes

Test in both cases: hypergeometric / Fisher's exact test.
We define the set of prefiltered species (**species universe**) as those that passed the threshold test of being present ($>$ 6000) in at least 31 of the arrays.
This method is especially relevant here as the tree does not show equal representation of different families and phyla.

Type of Results(at a rougher level than tree co-inertia)

- IBS higher group had significantly more Bacteriodetes
- overrepresentation of Firmicutes in the healthy controls.
- At the family level, the results showed that the families of Oxalobacteraceae, Prevotellaceae, Burkholderiaceae, Sphingobacteriaceae were significantly overrepresented in IBS.
- Conversely, the most significantly enriched family in control rats were Lachnospiraceae, including Ruminococcus sp., followed by Erysipelotrichaeceae and Clostridiaceae.

Elisabeth M Bik, Paul B Eckburg, Steven R Gill, Karen E Nelson, Elizabeth A Purdom, Fritz Francois, Guillermo Perez-Perez, Martin J Blaser, and David A Relman.
Molecular analysis of the bacterial microbiota in the human stomach.
Proc Natl Acad Sci USA, 103(3):732--7, Jan 2006.

J. Chakerian and S. Holmes.
Computational methods for evaluating phylogenetic trees, 2010.
arXiv.

J. Chakerian and S. Holmes.
distory:Distances between trees, 2010.

E.K. Costello, C.L. Lauber, M. Hamady, N. Fierer, J.I. Gordon, and R. Knight.
Bacterial community variation in human body habitats across space and time.
Science, 326(5960):1694, 2009.

Paul B Eckburg, Elisabeth M Bik, Charles N Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R Gill, Karen E Nelson, and David A Relman.
Diversity of the human intestinal microbial flora.
Science, 308(5728):1635--8, Jun 2005.

Steven N Evans and Frederick A Matsen.
The phylogenetic kantorovich-rubinstein metric for environmental sequence s amples.
arXiv, q-bio.PE, Jan 2010.

M Hamady, C Lozupone, and R Knight.
Fast unifrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data.
The ISME Journal, Jan 2009.

Susan Holmes.
Multivariate analysis: The French way.
In D. Nolan and T. P. Speed, editors, Probability and Statistics: Essays in Honor of David A. Freedman, volume 56 of IMS Lecture Notes--Monograph Series. IMS, Beachwood, OH, 2006.

C. Lozupone and R. Knight.
Unifrac: a new phylogenetic method for comparing microbial communities.
Applied and environmental microbiology, 71(12):8228, 2005.

Elizabeth Purdom.
Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree.
Annals of Applied Statistics, Jul 2010.

C. R. Rao.
The use and interpretation of Principal Component Analysis in applied research.
Sankhya A, 26:329--359., 1964.