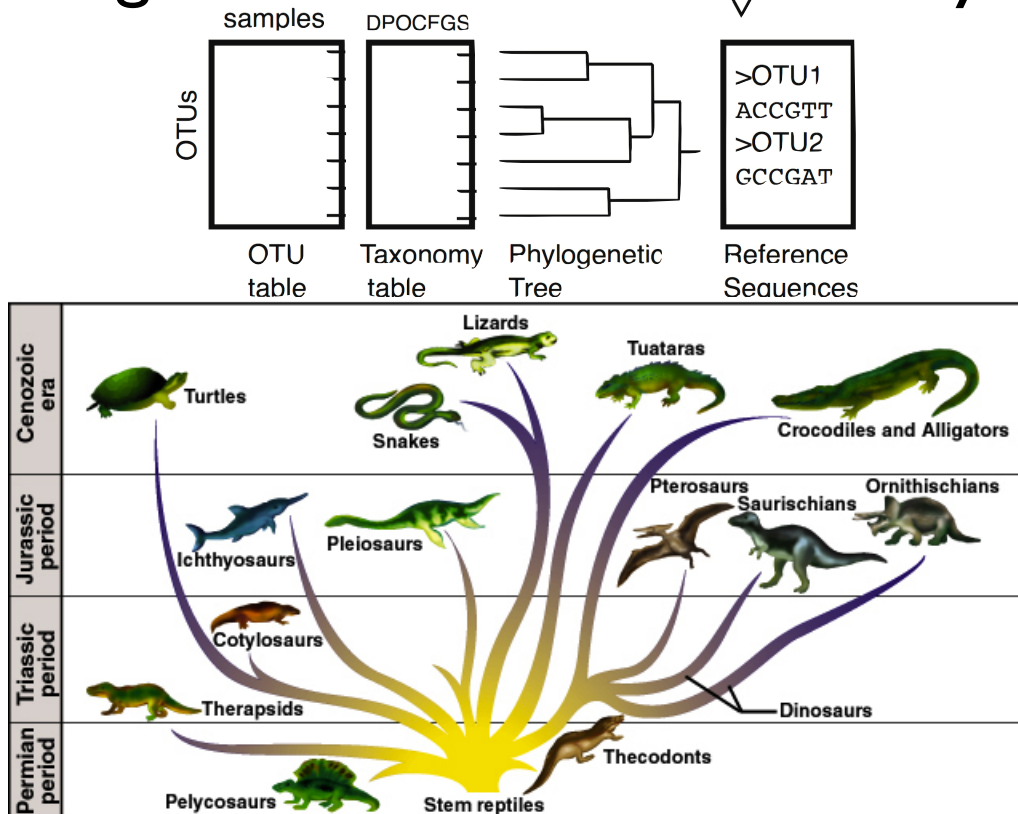


Using Trees in Microbiome Analysis



1

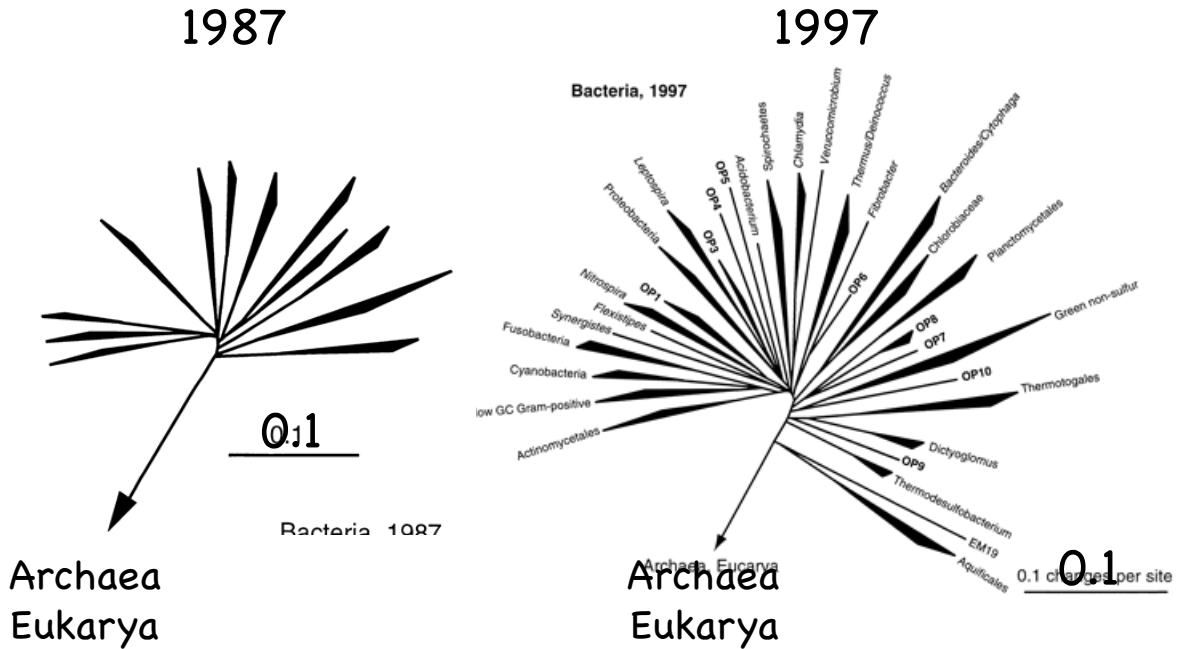
Using Trees in Microbiome Analysis

- Phylogenetic (Evolutionary) Trees
 - Tree-Building (“quick” overview)
 - Tree formats (Newick, Ape’s “phylo”)
 - Manipulating Trees in phyloseq/ape
 - Tree plots (Examples, how to interpret)
- Using Trees and contingency tables together
 - UniFrac and variants
 - DPCoA

2

Phylogenetic Trees

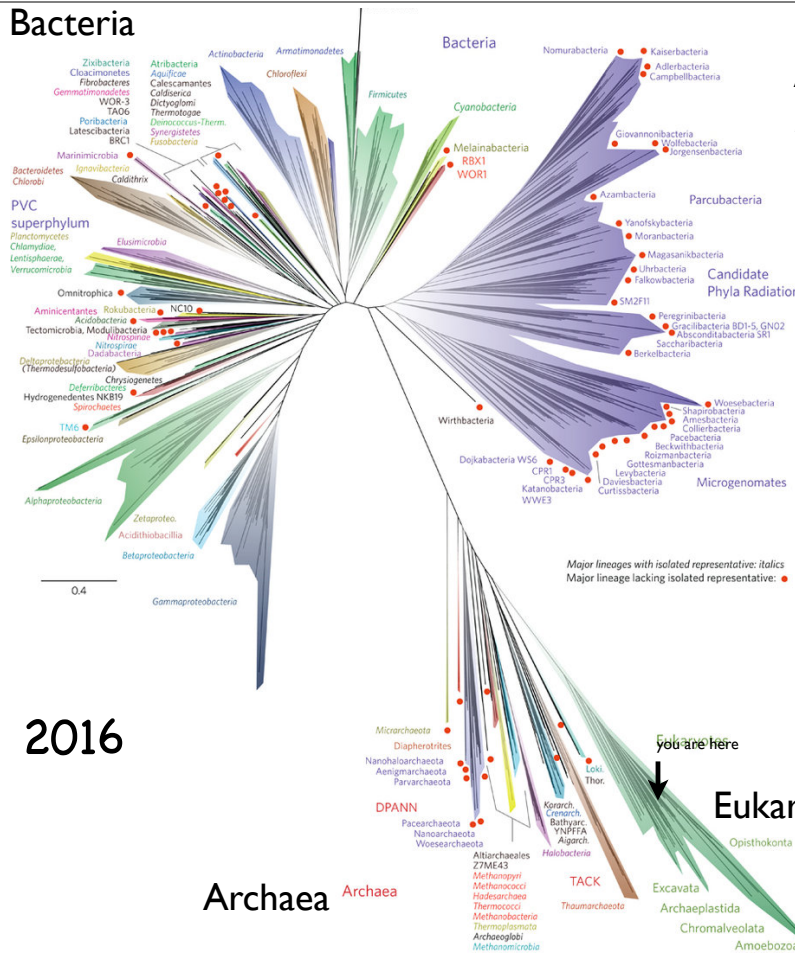
Evolutionary Tree, Known Bacteria



Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313), 734–740.

3

Bacteria



Hug & Banfield (2016)
A new view of the tree of life.
Nature Microbiology

Candidate
Phyla
Radiation
(mostly
uncultivated)

Overwhelming majority of
evolutionary diversity in
bacteria; highly correlated
with metabolic/functional
diversity

4

Phylogenetic Trees

Motivation:

- (1) Reconstructing evolutionary history from incomplete information
- (2) Robust summary of the similarity of related biological sequences (a lot like hclust)

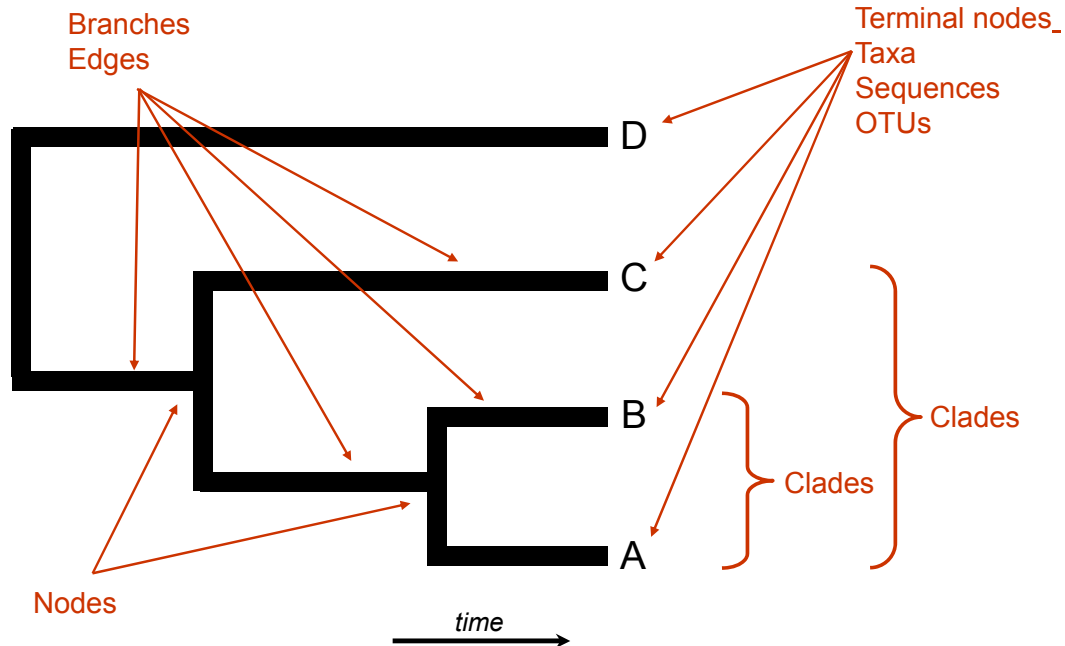
The data - biological sequences

- mostly proteins, sometimes DNA/RNA (16S rRNA), etc.

5

Phylogenetic Trees Nomenclature

the anatomy of a tree



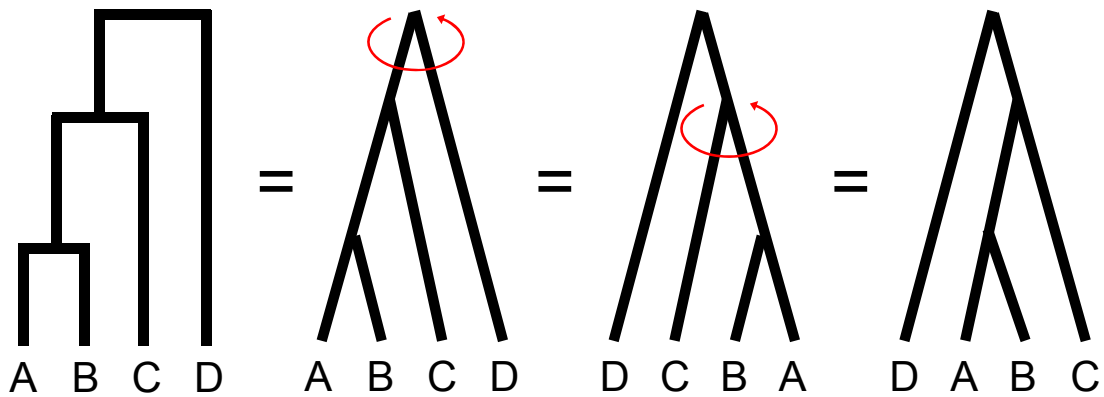
Most Recent Common Ancestor (MRCA) of A, B, C; but not D

Adapted from N. Provart & D. Guttman

6

Phylogenetic Trees

Rotating internal nodes is not meaningful:



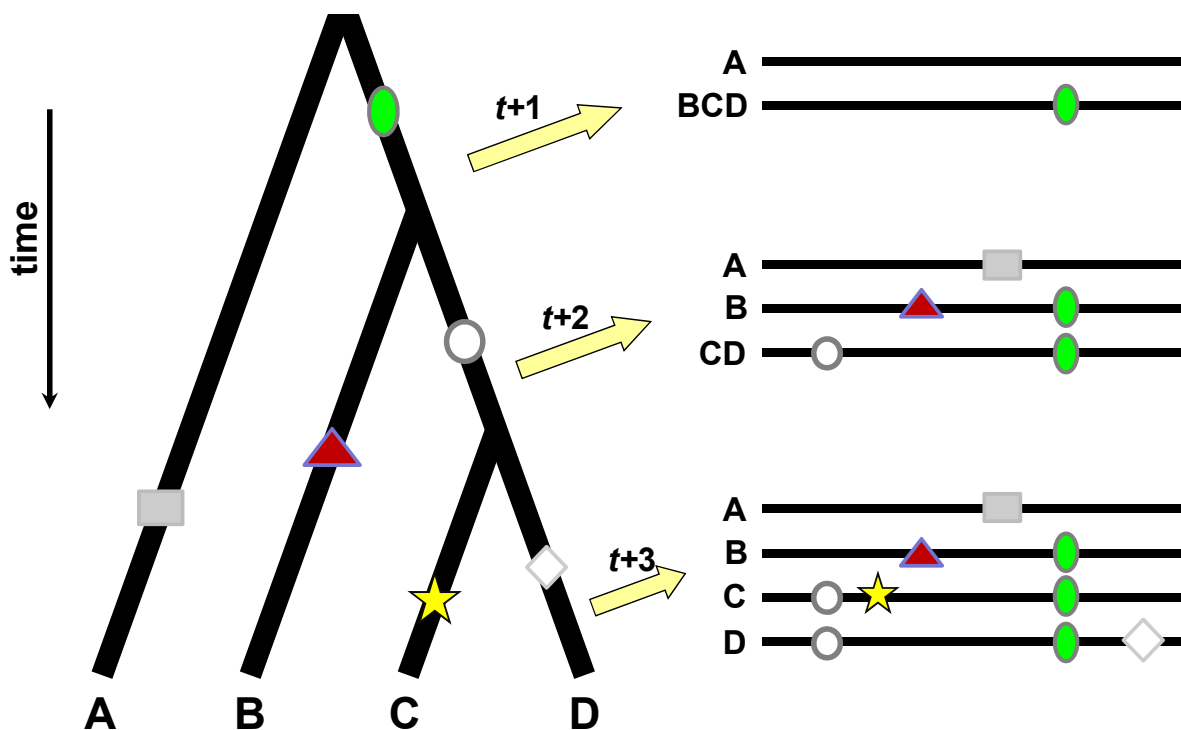
2^{N-1} possible arrangements for a particular rooting

Adapted from N. Provart & D. Guttman

7

Phylogenetic Trees

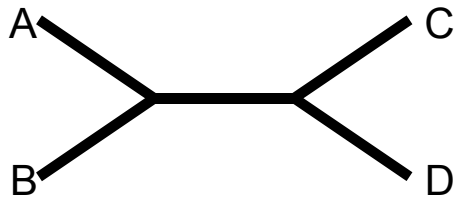
example



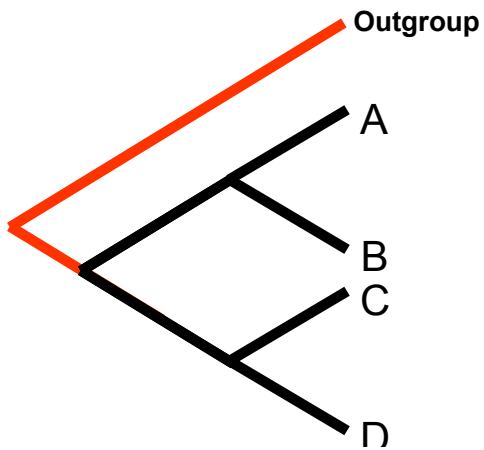
Adapted from N. Provart & D. Guttman

8

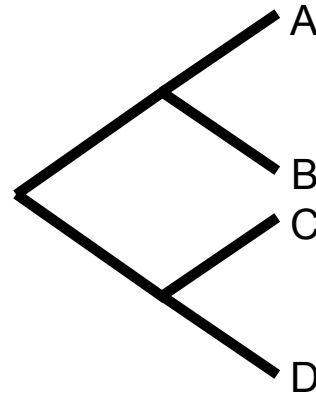
Using an "Outgroup"



Outgroup Rooting



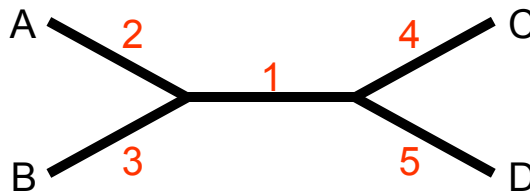
Midpoint Rooting



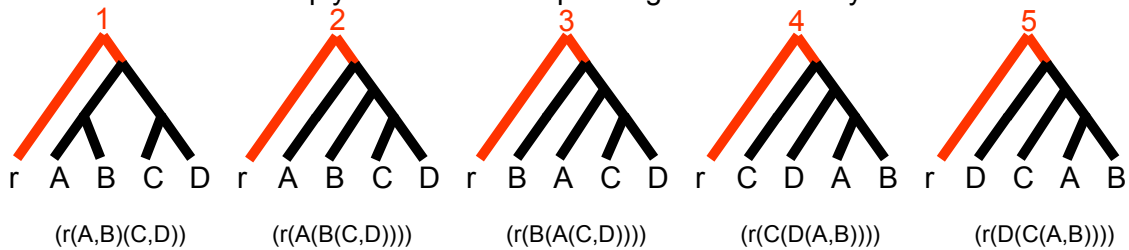
Adapted from N. Provart & D. Guttman

Rooting Trees

Unrooted Tree

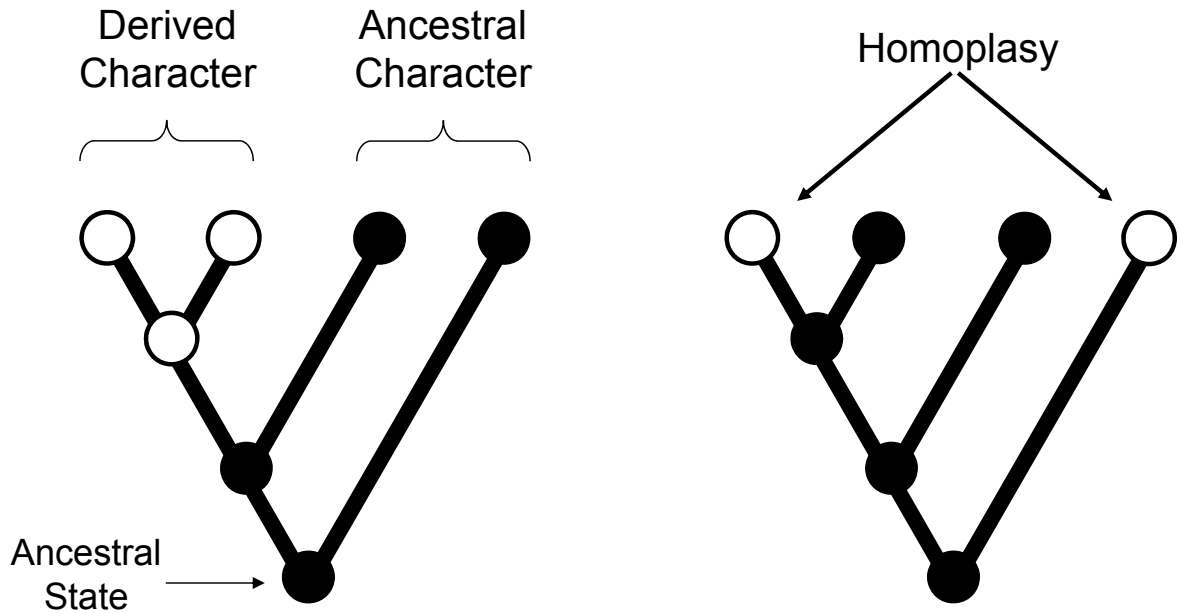


Rooted Trees – have one node from which all other nodes descend
 – imply direction corresponding to evolutionary time



Adapted from N. Provart & D. Guttman

More Terminology



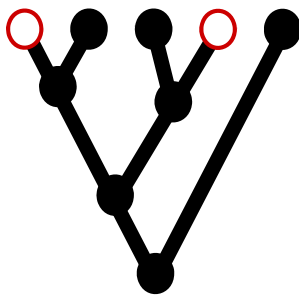
Homology - Similarity due to common ancestry

Homoplasy - Similarity due to parallel evolution, convergent evolution, or secondary loss

Adapted from N. Provart & D. Guttman

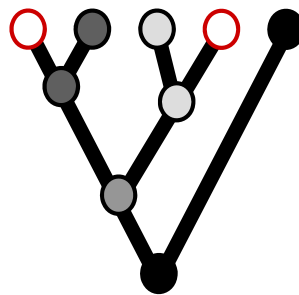
11

Forms of homoplasy...



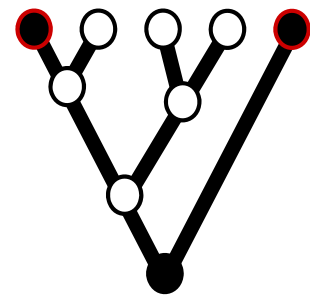
Parallel Evolution

Independent evolution of same character from same ancestral state



Convergent Evolution

Independent evolution of same character from different ancestral state



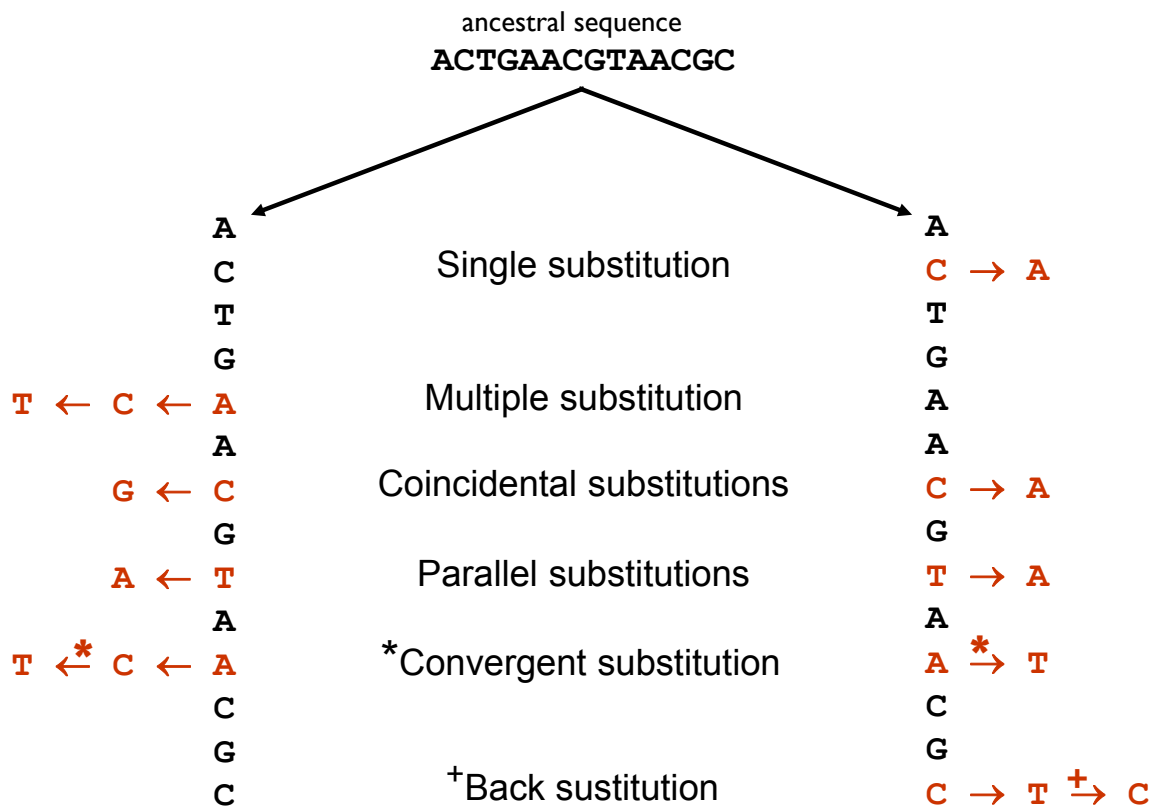
Secondary Loss

Reversion to ancestral state

E.g. Ni-Fe and Fe-only hydrogenases:
highly-similar enzymatic activity, no
detectable shared ancestry

Adapted from N. Provart & D. Guttman

12

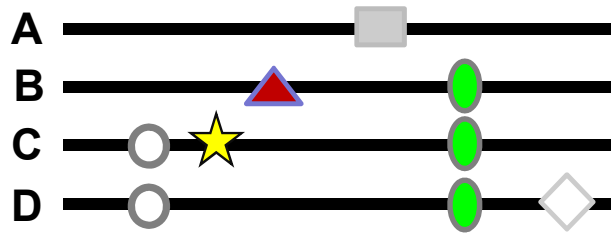


Adapted from N. Provart & D. Guttman

Phylogenetic Tree Construction Methods

Multiple Sequence Alignment:

All tree-building begins with multiple-alignment



Naïve multiple sequence alignment is NP-complete. No students indicated they wanted to learn about it this quarter, so Susan forbade me from spending any time on it. Just read about / use one of the following multiple-alignment algorithms:

ClustalW Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment...* Nucleic Acids Research, 22(22), 4673–4680.

Muscle Edgar, R. C. (2004). *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Research, 32(5), 1792–1797.

MAFFT Katoh, Misawa, Kuma, Miyata 2002 ([Nucleic Acids Res. 30:3059-3066](#))
MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.

Mauve, Lagan, etc. Whole genome alignment...

NOTE: You will not create a meaningful tree from a meaningless alignment. Spending time selecting the appropriate alignment tools and checking your alignment is probably a worthwhile thing to do.

15

Phylogenetic Tree Construction Methods

Distance Methods

UPGMA Bad, don't use. Implemented as guesses in better, more complex algorithms for m-alignment / tree construction

Neighbor-Joining Also not very good, only use if other methods intractable, or use as initial guess for parsimony or ML tree.

Character-based (discrete) Methods

Maximum Parsimony

Maximum Likelihood

Bayesian Methods

16

Phylogenetic Tree Construction Methods

Distance Methods

Relationships based upon sequence similarity.

Advantages

- Computationally fast.
- Single “best tree” found.

Disadvantages

- Assumptions
 - additive distances (always)
 - molecular clock (sometimes)
- Information loss occurs due to data transformation
- Uninterpretable branch lengths
- Single “best tree” found.



17

Phylogenetic Tree Construction Methods

UPGMA

Not much point in discussing. Not very good. You know how to do it from clustering lecture(s).

Details:

- * Assumes rates of evolution are same among different lineages (severely unrealistic)
- * Very sensitive to unequal evolutionary rates
- * Tends to be reliable only if data/phylogeny is essentially ultrametric (severely unrealistic)

18

Phylogenetic Tree Construction Methods

Neighbor Joining

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4), 406–425.

1. Calculate pairwise distances
2. Create distance matrix
3. Determine net divergence for each terminal node
4. Create rate-corrected distance matrix
5. Identify taxa with minimum rate-corrected distance
6. Connect taxa with minimum rate-corrected distance via a new node, and determine their distance from this new node
7. Determine the distance of new node from rest of taxa or nodes
8. Regenerate distance matrix
9. Return to step 2

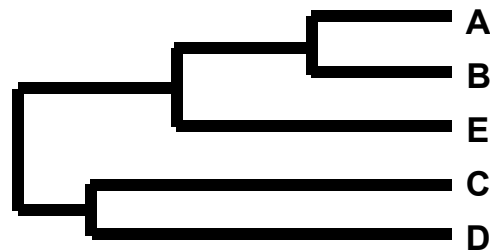
Adapted from N. Provart & D. Guttman

19

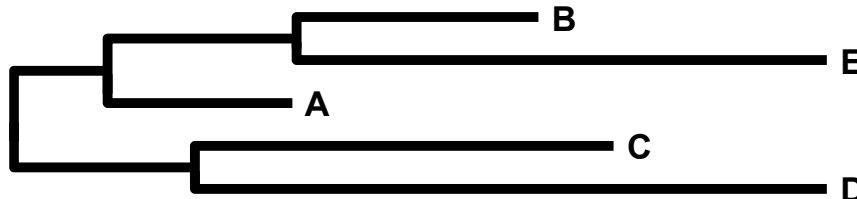
UPGMA vs. Neighbour-Joining

	A	B	C	D	E
A	-	17	21	31	23
B		-	30	34	21
C			-	28	39
D				-	43
E					-

UPGMA



NJ



Adapted from N. Provart & D. Guttman

20

Phylogenetic Trees

Character-based (discrete) Methods

Maximum Parsimony

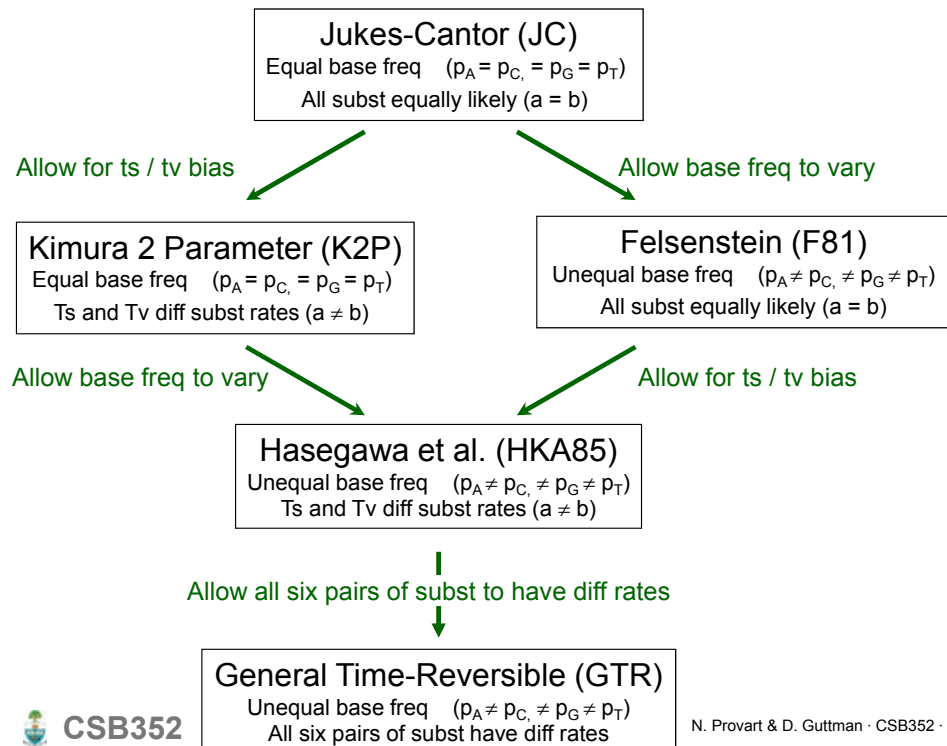
Maximum Likelihood

Bayesian Methods

These methods attempt to map the history of gene sequences onto a tree.
(And decide what the tree looks like)

21

Models of Sequence Evolution



Adapted from N. Provart & D. Guttman

22

Phylogenetic Trees

Maximum Parsimony

Works under the principle of “Occam's razor”

Farris (1983), has a justification for parsimony:
“minimizes requirements of ad hoc hypotheses of homoplasy”.

Analogy is made between homoplasies and residuals, (part of the data that the tree does not explain), minimizing homoplasies is akin to minimizing residuals in regression.

Based on the assumption that “evolution is parsimonious” which means that there should be no more evolutionary steps than necessary.

The best tree(s) minimize the number of changes between ancestors and descendants.

Under independence of each of the characters, this has a clear combinatorial translation.

23

Phylogenetic Trees

Maximum Parsimony

Implementation:

- In parsimony, the score is simply the minimum number of mutations that could possibly produce the data.
- There are fast algorithms that guarantee that any tree can be scored correctly
- There are lots of possible trees to choose between...

Math people:

If you take it in terms of distance on a graph the inner points are what are known as Steiner points and the problem of finding the tree is equivalent to the Steiner tree problem...

Drawbacks:

- the score of a tree is completely determined by the minimum number of mutations among all of the reconstructions of ancestral sequences.
- fails to account for the fact that the number of changes is unlikely to be equal on all branches in the tree.
- As a result, susceptible to “long-branch attraction”, in which two long branches that are not adjacent on the true tree are inferred to be closest relatives
- in practice this is still pretty good... ML/Bayesian better

24

Phylogenetic Trees

Maximum Likelihood

Attempts to answer the question:

- What is the probability of observing the data, given a particular model of evolution and evolutionary history?
 - data = MSA
 - model = transition probabilities, base frequencies, rate heterogeneity...
 - evolutionary history = phylogenetic tree

Evaluates the likelihood of every substitution of every possible tree.

All possible trees are considered, and the number of substitutions that must have occurred are calculated.

The tree with the highest likelihood is assumed to be the correct tree.

Adapted from N. Provart & D. Guttman

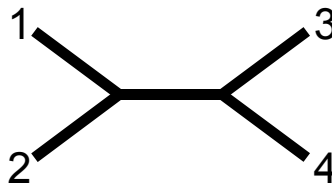
25

Phylogenetic Trees

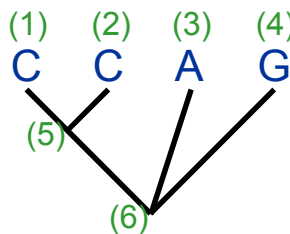
Maximum Likelihood

1.....*j*.....N
1 C...GGACACGTTTA...C
2 C...AGACACCTCTA...C
3 C...GGATAAGTTAA...C
4 C...GGATAGCCTAG...C

Unrooted tree for
the 4 taxa



Arbitrarily rooted tree
for site *j*



Adapted from N. Provart & D. Guttman

26

Phylogenetic Trees

Maximum Likelihood

$$L_j = P \begin{pmatrix} C & C & A & G \\ A & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ G & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ C & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ T & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix}$$

$$* P \begin{pmatrix} C & C & A & G \\ A & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ G & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ C & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ T & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix}$$

$$* P \begin{pmatrix} C & C & A & G \\ A & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ G & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ C & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ T & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix}$$

$$* P \begin{pmatrix} C & C & A & G \\ A & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ G & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ C & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix} * P \begin{pmatrix} C & C & A & G \\ T & \diagdown & \diagup & \\ & \diagdown & \diagup & \\ & & & \end{pmatrix}$$

Adapted from N. Provart & D. Guttman

27

Phylogenetic Trees

Maximum Likelihood

Likelihood of the tree = product of the likelihoods for each site.

$$L = L_1 \times L_2 \times \dots \times L_N = \prod_{j=1}^N L_j$$

Usually evaluated as the sum of the log likelihoods.

$$\ln L = \ln L_1 + \ln L_2 + \dots + \ln L_N = \sum_{j=1}^N \ln L_j$$

ML evaluates:

- all possible ancestral states
- at all variable site
- in all possible tree topologies

→The most likely (best) tree is the topology that has the highest overall likelihood.

Adapted from N. Provart & D. Guttman

28

Phylogenetic Trees

Maximum Likelihood

Advantages of ML methods

- Based on explicit evolutionary models.
- Permits statistical evaluation of the likelihood of specific tree topologies.
- Often returns many equally likely trees.
- Usually outperforms other methods.

Disadvantages

- Computationally very intensive.
- Often returns many equally likely trees.

Adapted from N. Provart & D. Guttman

29

Bayesian Approach to Phylogeny Estimation

Approach:

Uses the likelihood function

Normally implemented using same models of evolutionary change used in ML

Metropolis-Hastings - Metropolis-Coupled Markov Chain Monte Carlo (MC³)

Assumptions:

Same set of parameter choices for evolutionary model as for ML

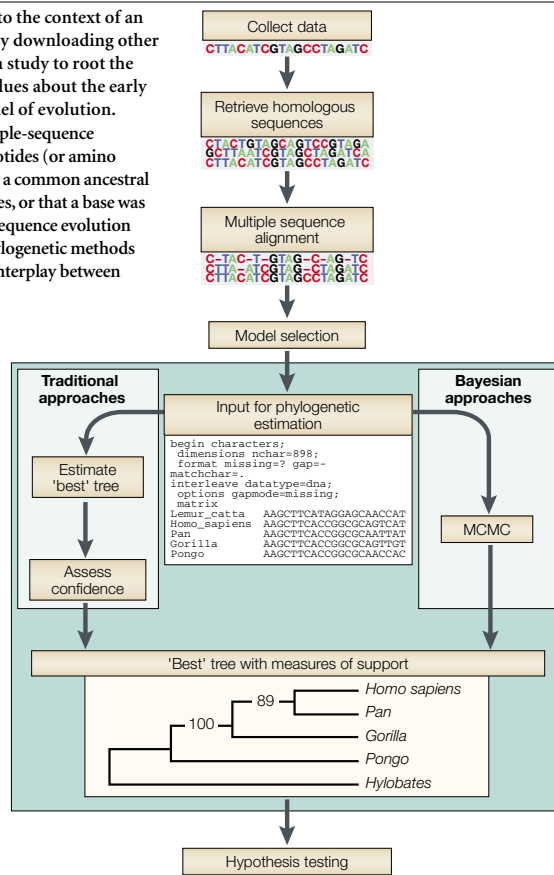
Must also choose initial set of prior probabilities.

Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews Genetics*, 4(4), 275–284.

Ronquist, F. and J.P. Huelsenbeck. (2003) MrBayes3: Bayesian phylogenetic inference... *Bioinformatics*, 19, 1572–1574.

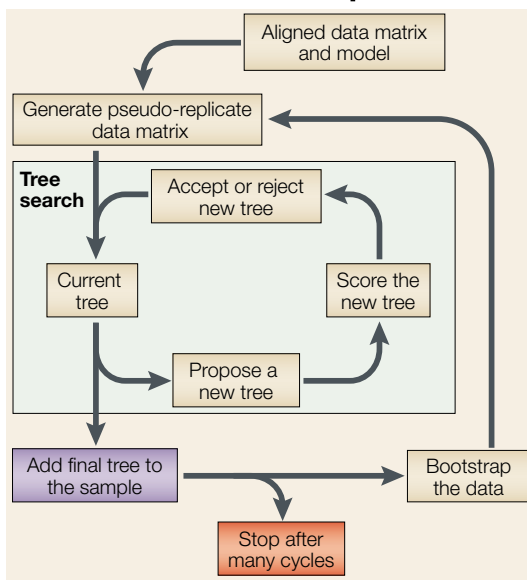
30

to the context of any downloading other study to root the trees about the early el of evolution. ple-sequence stides (or amino a common ancestral es, or that a base was equence evolution logenetic methods terplay between

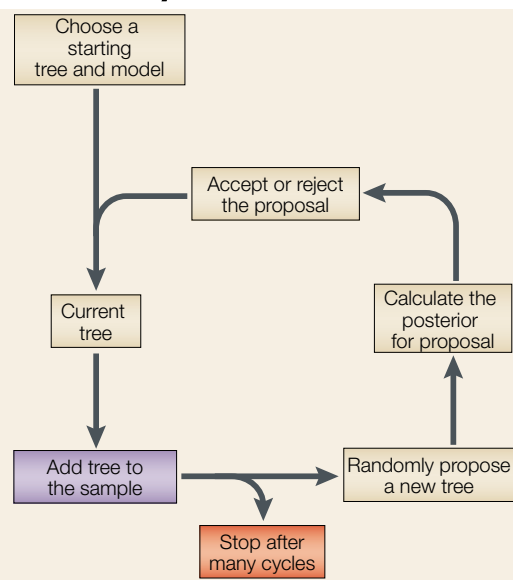


Holder, M., & Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews Genetics*, 4(4), 275–284.

ML-bootstrap



Bayesian MC³



Holder, M., & Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews Genetics*, 4(4), 275–284.

Phylogenetic Tree Construction Methods

Recommended Software



phangorn - MP, ML, and Bayesian tree estimation
ape - tree-handling in R, tree-build, graphics
picante -
phyloseq - integrated tree-abundance ana./graphics
ggdendro - ggplot2 hclust graphics



geneIOUS
(CMGM has a site-license)

NJ, UPGMA, PAUP*, PhyML, MrBayes
(including "cloud" MrBayes)

RAxML

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, 22(21), 2688–2690.

MrBayes

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, 17(8), 754–755.

BEAUti / BEAST 1.7

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*. <http://beast.bio.ed.ac.uk/>

http://en.wikipedia.org/wiki/List_of_phylogenetics_software

33

Phylogenetic Tree Construction Methods

But we're not going to spend any time worrying about building trees ourselves in this course...

Why we won't:

- There are many manually-curated public trees
- Optimal tree is not really known, lots to argue over
- For our purposes small differences should not matter

Why you might want to calculate a new tree:

- You have counts from non-16S rRNA gene
- Have concatenated whole genome sequence data
- Basically any time you have new biological sequence data for which a public reference tree is not available

34

Tree file format, data representation: Newick

Green Genes Tree in Newick format:

```
((((((((836:0.06877,
((549322:0.00892,522457:0.01408)1.000:0.
314761:0.09977)0.161:0.01566)0.882:0.00924,
(((311539:0.0484 ((174835:0.01627,
(34207:0.00082,45996:0.00334)0.863:0.00433
1.000.3:0.09792)1.000.4:0.04652,((((945:0.08077,
(178877:0.01342,
(29928:0.00726,35548:0.00187)0.748:0.01216)
1.000.5:0.05924)0.975:0.01729, ... ;
```

A simple Newick tree with branch lengths is noted:

```
((1:1,4:1):3,(2:1,3:1),5:2):1);
```

http://evolution.genetics.washington.edu/phylip/newick_doc.html

35

Tree file format, data representation: phylo (ape)

Terminology and Notations:

branch: edge, vertex

node: internal node

degree: the number of edges that meet at a node

tip: terminal node, leaf, node of degree 1

n: number of tips

m: number of nodes

http://ape-package.ird.fr/misc/FormatTreeR_24Oct2012.pdf

36

Tree file format, data representation: phylo (ape)

Definition of the Class "phylo"

The class "phylo" is used to code "acyclical" phylogenetic trees. These trees have no reticulations, and all their internal nodes are of degree 3 or more, except the root (in the case of rooted trees) which is of degree 2 or more. An object of class "phylo" is a list with the following mandatory elements:

1. A numeric matrix named `edge` with two columns and as many rows as there are branches in the tree;
2. A character vector of length `n` named `tip.label` with the labels of the tips;
3. An integer value named `Nnode` giving the number of (internal) nodes;
4. An attribute class equal to "phylo".

In the matrix `edge`, each branch is coded by the nodes it connects: tips are coded `1, ..., n`, and internal nodes are coded `n+1, ..., n+m` (`n+1` is the root). Both series are numbered without gaps.

`edge.length`, `node.label`, `root.edge` are optional annotation slots in "phylo" list

http://ape-package.ird.fr/misc/FormatTreeR_24Oct2012.pdf

37

Tree file format, data representation: phylo (ape)

The "ape::phylo" edge-matrix has the following properties:

1. The first column has only values greater than `n` (thus, values less than or equal to `n` appear only in the second column).
2. All nodes appear in the first column at least twice.
3. The number of occurrences of a node in the first column is related to the nature of the node: twice if it is dichotomous (i.e., of degree 3), three times if it is trichotomous (degree 4), and so on.
4. All elements, except the root `n + 1`, appear once in the second column.

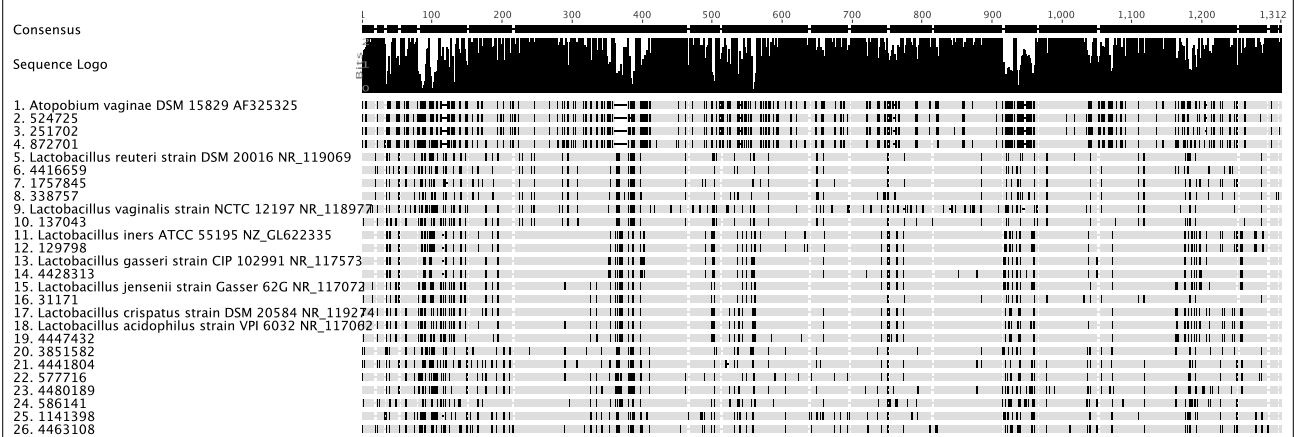
38

Example Tree Plots: “How to Read a Tree”

Example I: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database

39

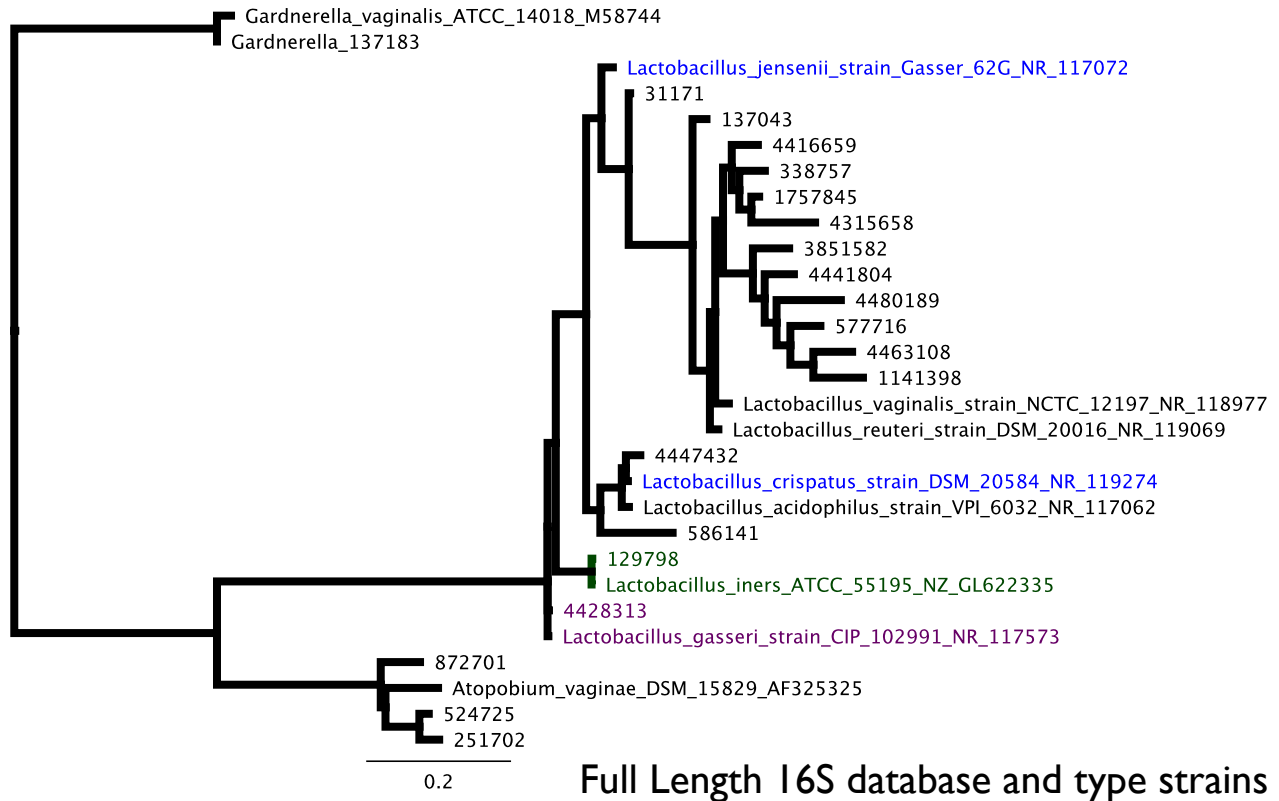
Example I: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database



Full Length 16S database and type strains

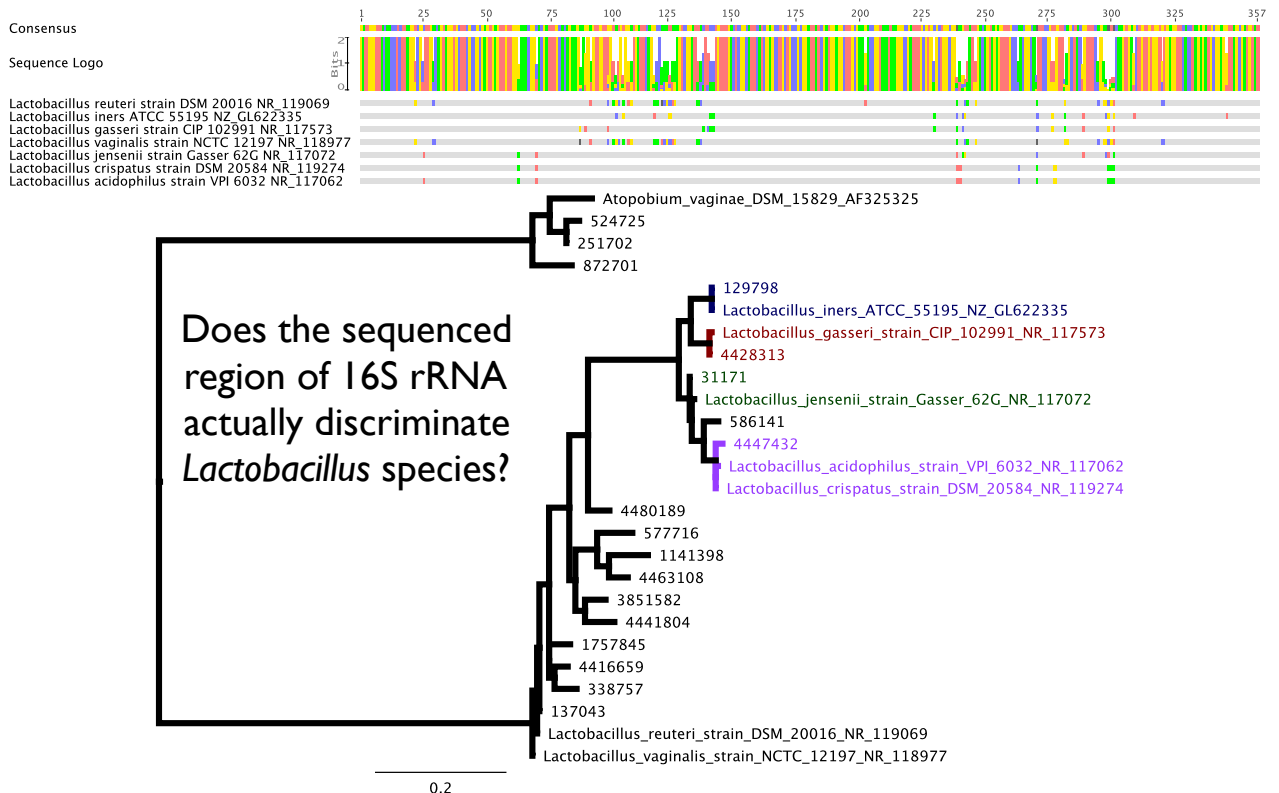
40

Example I: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database



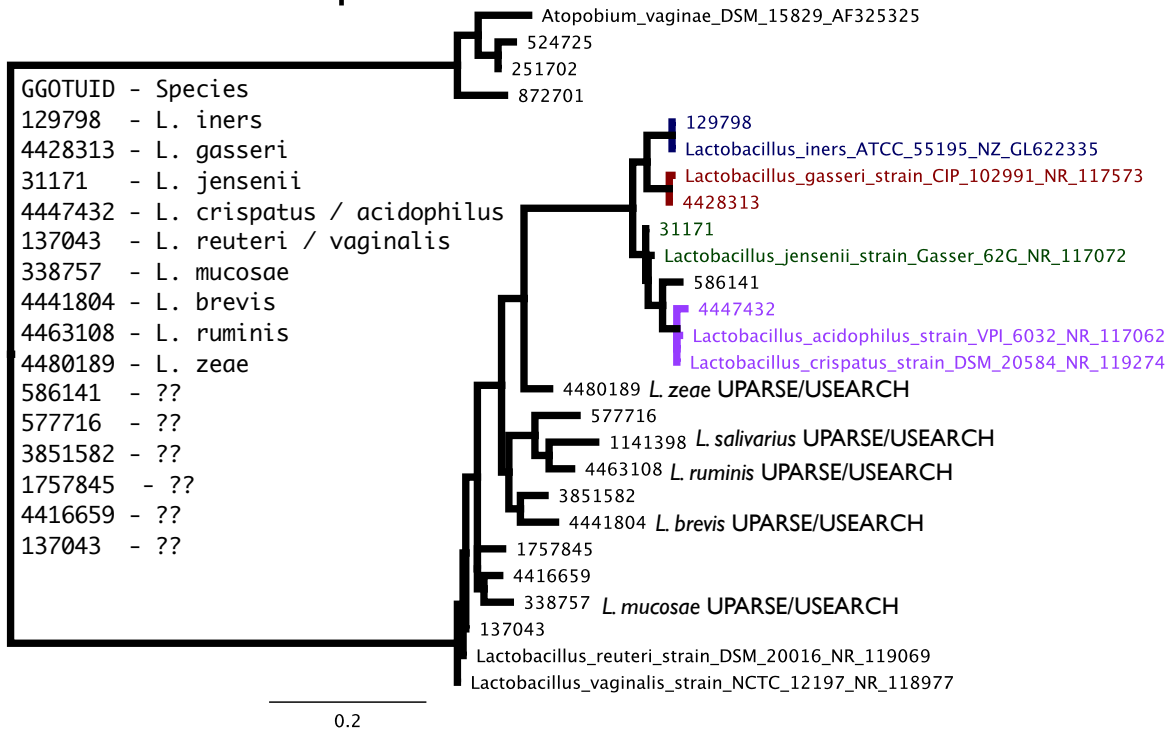
41

Example I: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database



42

Example I: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database



43

Manipulating Trees in phyloseq/ape

- Use standard OTU/species functions
 - `prune_taxa()`, `filter_taxa()`, `subset_taxa()`
- `tip_glom()`, `tax_glom()`
- ape functions after accession:
 - `plot.tree(phy_tree(physeq))`
 - `root(phy_tree(physeq), ...)`

44

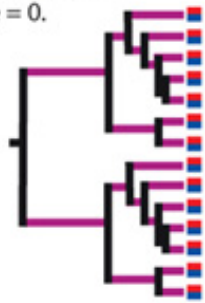
Tree Method: UniFrac

45

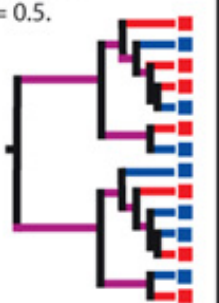
(Unweighted) UniFrac Distance

A proposal for using the phylogenetic tree and OTU table

A. Identical sequence sets: all seqs in red + blue set. 100% branch length shared (purple).
UniFrac score = 0.



B. Related sequence sets: seqs in red have relatives in blue. ~50% branch length shared.
UniFrac score = 0.5.



C. Unrelated sequence sets: seqs in red have no close relatives in blue. 0% branch length shared.
UniFrac score = 1.



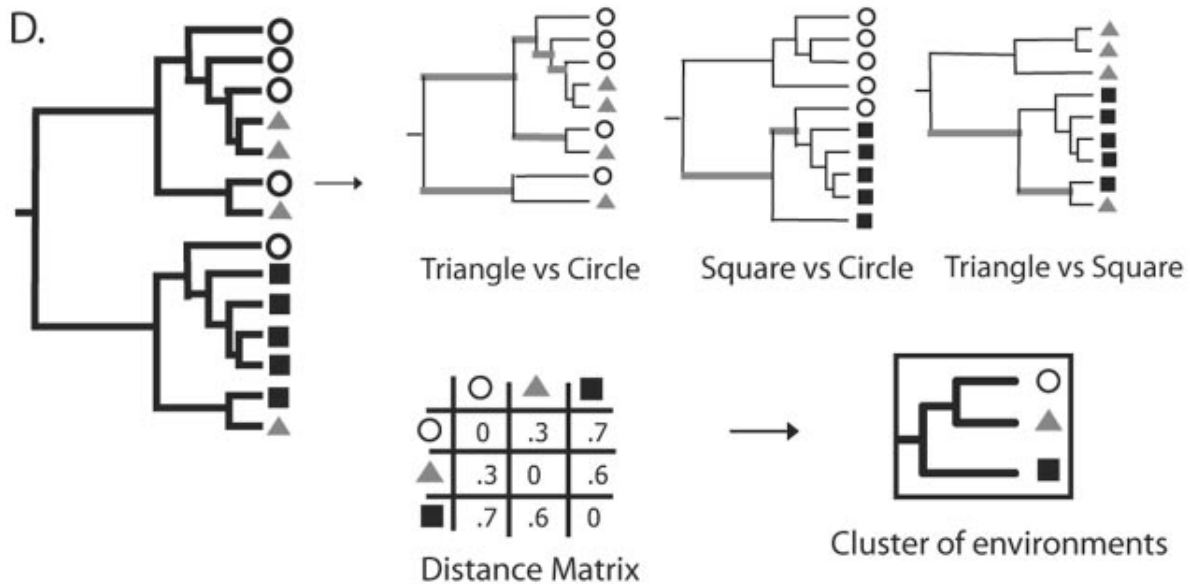
“Since we compared environments on a large scale, the ability of particular lineages of organisms to survive in each environment is more likely to represent the relevant aspects of similarity between environments than the relative abundance of each surviving lineage”

Lozupone & Knight (2005) *Applied and Environmental Microbiology*

46

(Unweighted) UniFrac Distance

A proposal for using the phylogenetic tree and OTU table



Lozupone & Knight (2005) *Applied and Environmental Microbiology*

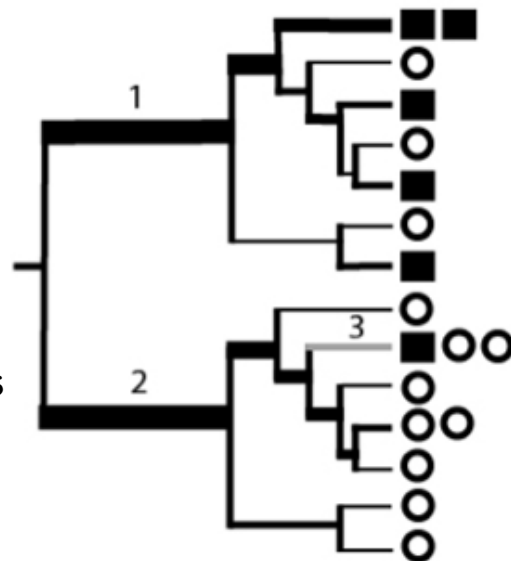
47

Weighted UniFrac Distance

A modification of (unweighted) UniFrac

$$\sum_{i=1}^n b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

- n = number of branches in the
- b_i = length of the i th branch
- A_i = number of descendants of i th branch in group A
- A_T = total number of sequences in group A



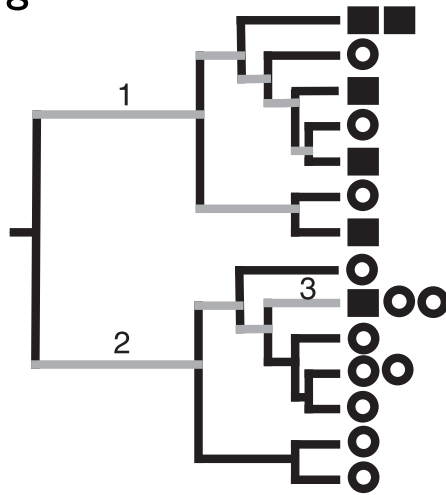
Lozupone et al., 2007

48

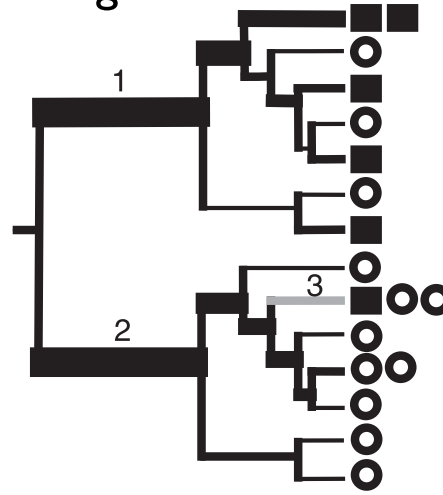
UniFrac Comparison

(Fraction of branch lengths not shared)

Unweighted



Weighted



gray branches have no weight

Lozupone, et al (2007). Quantitative and qualitative... *Applied and Environmental Microbiology*

49

Tree Method: UniFrac

Ordination Examples:

- PCoA/MDS (very common)
- NMDS

Exploratory Analyses often

rarefy - UniFrac - PCoA - Write Paper

(Not that we recommend this approach)

50

Tree Method: DPCoA

“Double Principle Coordinates Analysis”, DPCoA, implemented in ade4

Suppose we have n species in p locations and a (euclidean) matrix Δ giving the squares of the pairwise distances between the species. Then we can

- Use the distances between species to find an embedding in $(n - 1)$ -dimensional space such that the euclidean distances between the species is the same as the distances between the species defined in Δ .
- Place each of the p locations at the barycenter of its species profile. The euclidean distances between the locations will be the same as the square root of the Rao dissimilarity between them. (Rao 1986)
- Use PCA to find a lower-dimensional representation of the locations.
- Gives the species and sample coordinates such that the inertia decomposes the same way the diversity does.
- Note: Don't have to use patristic distance. Could use other D for species

Pavoine, et al. (2004). From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology*, 228(4), 523–537

51

Comparison of UniFrac and DPCoA

Original description	New formula	Properties
square root of Rao's distance based on the square root of the patristic distances	$[\sum_i b_i (A_i/A_T - B_i/B_T)^2]^{1/2}$	Most sensitive to outliers, least sensitive to noise, upweights deep differences, gives OTU locations
$\sum_i b_i A_i/A_T - B_i/B_T $	$\sum_i b_i A_i/A_T - B_i/B_T $	Less sensitive to outliers/more sensitive to noise than DPCoA
fraction of branches leading to exactly one group	$\sum_i b_i 1\{\frac{A_i/A_T - B_i/B_T}{A_i/A_T + B_i/B_T} \geq 1\}$	Sensitive to noise, upweights shallow differences on the tree

Summary of the methods under consideration. “Outliers” refers to highly abundant OTUs, and noise refers to noise in detecting low-abundance OTUs (see Fukuyama and Holmes, 2012)

52

Comparison of UniFrac and DPCoA

Microbiome Example: Antibiotic Timecourse

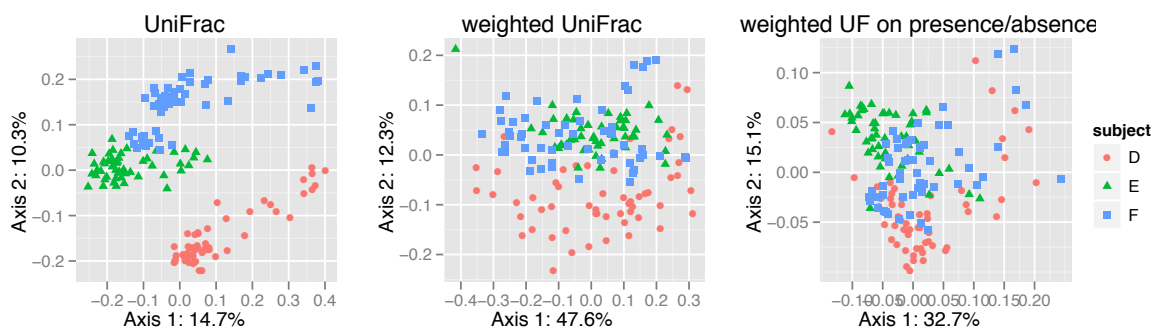
Measurements of about 2500 different bacterial OTUs from stool samples of three patients (D, E, F)

Each patient sampled ~ 50 times during the course of treatment with ciprofloxacin (an antibiotic).

Times categorized as Pre Cp, 1st Cp, 1st WPC (week post cipro), Interim, 2nd Cp, 2nd WPC, and Post Cp.

53

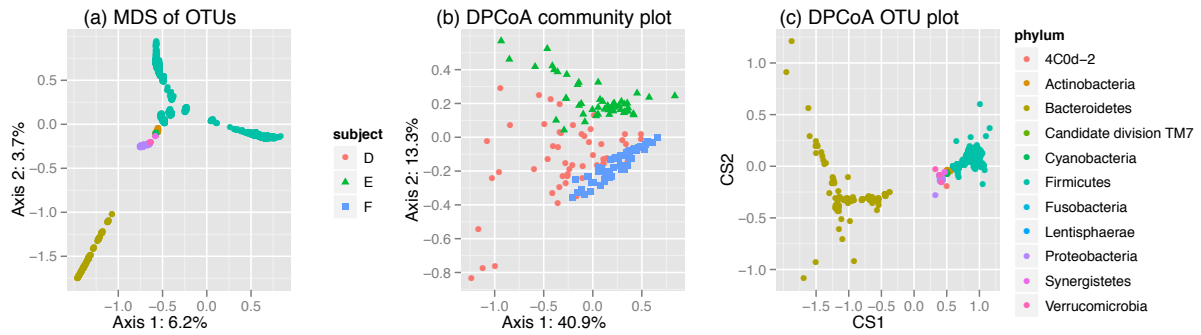
Comparison of UniFrac and DPCoA



Comparing the UniFrac variants. From left to right: PCoA/MDS with unweighted UniFrac, with weighted UniFrac, and with weighted UniFrac performed on presence/absence data extracted from the abundance data used in the other two plots

54

Comparison of ~~UniFrac~~ and DPCoA



- (a) PCoA/MDS of the OTUs based on the patristic distance
- (b) community and
- (c) species points for DPCoA after removing two outlying species.