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

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

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

Phylogenetic Trees

the anatomy of a tree

- Branches
- Edges
- Nodes
- Clades
- Terminal nodes
- Taxa
- Sequences
- OTUs

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

Adapted from N. Provart & D. Guttman
Phylogenetic Trees
Rotating internal nodes is not meaningful:

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

Adapted from N. Provan & D. Gutman

Phylogenetic Trees

example

Adapted from N. Provan & D. Gutman
Using an "Outgroup"

### Outgroup Rooting

- **A**
- **B**
- **C**
- **D**

### Midpoint Rooting

- **A**
- **B**
- **C**
- **D**

---

**Rooting Trees**

**Unrooted Tree**

**Rooted Trees** — have one node from which all other nodes descend

- Imply direction corresponding to evolutionary time

- **r(A,B)(C,D))**
- **r(A(B(C,D)))**
- **r(B(A(C,D)))**
- **r(C(D(A,B)))**
- **r(D(C(A,B)))**

---

Adapted from N. Prokert & D. Guttman
More Terminology

Ancestral State

Homology - Similarity due to common ancestry

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

Ancestral State

Derived Character

Ancestral Character

Homoplasy

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
Phylogenetic Tree Construction Methods
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**

- **Muscle**

- **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.

---

### 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**
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.

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)
### Neighbor Joining

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

---

#### UPGMA vs. Neighbor-Joining

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>17</td>
<td>21</td>
<td>31</td>
<td>23</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>30</td>
<td>34</td>
<td>21</td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td>-</td>
<td>28</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>E</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**UPGMA**

**NJ**

---

Adapted from N. Provart & D. Guttman
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)

Models of Sequence Evolution

- Jukes-Cantor (JC)
  - Equal base freq \( p_A = p_C = p_G = p_T \)
  - All subst equally likely \( (a = b) \)
  - Allow for ts / tv bias
  - Allow base freq to vary

- Kimura 2 Parameter (K2P)
  - Equal base freq \( p_A = p_C = p_G = p_T \)
  - Ts and Tv diff subst rates \( (a \neq b) \)
  - Allow base freq to vary

- Felsenstein (F81)
  - Unequal base freq \( p_A \neq p_C \neq p_G \neq p_T \)
  - All subst equally likely \( (a = b) \)
  - All subst equally likely \( (a = b) \)
  - Allow for ts / tv bias

- Hasegawa et al. (HKA85)
  - Unequal base freq \( p_A \neq p_C \neq p_G \neq p_T \)
  - Ts and Tv diff subst rates \( (a \neq b) \)
  - Allow all six pairs of subst to have diff rates

- General Time-Reversible (GTR)
  - Unequal base freq \( p_A \neq p_C \neq p_G \neq p_T \)
  - All six pairs of subst have diff rates

Adapted from N. Provart & D. Guttman
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.

Phylogenetic Trees

Maximum Parsimony

Works under the principle of “Occam’s razor”

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
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.

Likelihood example:

\[
L(Data|Model1) = \text{Prob}(H|Model1) \times \text{Prob}(H|Model1) \times \text{Prob}(T|Model1) \times \text{Prob}(H|Model1) \times \text{Prob}(T|Model1) \times \text{Prob}(H|Model1) = 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 = 0.0156
\]

\[
L(Data|Model2) = 1.0 \times 1.0 \times 0.0 \times 1.0 \times 0.0 \times 1.0 = 0.0
\]

\[
L(Data|Model3) = 0.0 \times 0.0 \times 1.0 \times 0.0 \times 1.0 \times 0.0 = 0.0
\]
Likelihood of the tree = product of the likelihoods for each site.

\[ L = L_1 \times L_2 \times \ldots \times L_N = \prod_{j=1}^{N} L_j \]

Usually evaluated as the sum of the log likelihoods.

\[ \ln L = \ln L_1 + \ln L_2 + \ldots + \ln L_N = \sum_{j=1}^{N} \ln L_j \]

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.
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.

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.

Adapted from N. Provart & D. Guttman

ML-bootstrap

Bayesian MC³


---


---

© 2023·Nature Publishing Group

---

© 2023·Nature Publishing Group
**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 analysis/graphics
- **ggdendro** - ggplot2 hclust graphics

(Originally, CMGM has a site-license)

**geneIOUS**

NJ, UPGMA, PAUP*, PhyML, MrBayes
(including “cloud” MrBayes)

---

**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

---


---

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, 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);


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

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


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.
Example Tree Plots: “How to Read a Tree”

Example 1: Determine species names of unlabeled \textit{Lactobacillus} species in the GreenGenes database

Consensus Sequence logo

1. \textit{Atopobium vaginae} DSM 15829 AF325325
2. 324730
3. 3217202
4. 3127261
5. \textit{Lactobacillus} neuteri strain DSM 20016 NR_119609
6. 46166859
7. 7252845
8. 316787
9. \textit{Lactobacillus} vaginalis strain NCTC 12197 NR_118977
10. 137043
11. \textit{Lactobacillus} iners ATCC 55195 NZ_G1622635
12. 12974
13. \textit{Lactobacillus} gasseri strain CIP 102991 NR_117573
14. 4428811
15. \textit{Lactobacillus} gasseri strain Gasser 626 NR_217073
16. 11171
17. \textit{Lactobacillus} crispatus strain DSM 25854 NR_119274
18. \textit{Lactobacillus} acidophilus strain VPI 6032 NR_117606
19. 444742
20. 3851592
21. 4441684
22. 444169
23. 4441819
24. 5812141
25. 1141968
26. 4460180

Full Length 16S database and type strains
Example 1: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database

Does the sequenced region of 16S rRNA actually discriminate *Lactobacillus* species?
Example 1: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database

**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), ...)"
Tree Method: UniFrac

(Unweighted) UniFrac Distance
A proposal for using the phylogenetic tree and OTU table

“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
(Unweighted) UniFrac Distance
A proposal for using the phylogenetic tree and OTU table

Lozupone & Knight (2005) Applied and Environmental Microbiology

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
UniFrac Comparison
(Fraction of branch lengths not shared)

Unweighted

Weighted

gray branches have no weight


Tree Method: UniFrac

Ordination Examples:
- PCoA/MDS (very common)
- NMDS

Exploratory Analyses often
rarefy - UniFrac - PCoA - Write Paper
(Not that we recommend this approach)
Tree Method: DPCoA
“Double Principle Coordinates Analysis”, DPCoA, implemented in ade4

Suppose we have n species in p locations and a (euclidean) matrix $\Delta$ 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 $\Delta$.
- 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


Comparison of UniFrac and DPCoA

Original description
square root of Rao’s distance based on the square root of the patristic distances

New formula
$\left[\sum_i b_i (A_i/A_T - B_i/B_T)^2\right]^{1/2}$

Properties
Most sensitive to outliers, least sensitive to noise, upweights deep differences, gives OTU locations

Less sensitive to outliers/more sensitive to noise than DPCoA

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)
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.

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.
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.