Using Trees in Microbiome Analysis

OTUS
- samples
- DPOCFGS

OTU table
Taxonomy table
Phylogenetic Tree
Reference Sequences

> OTU1
ACCGTTT

> OTU2
GCCGAT

Evolutionary tree of reptiles:
- Cenozoic era
  - Turtles
  - Lizards
  - Snakes
  - Tuataras
  - Crocodiles and Alligators
- Jurassic period
  - Ichthyosaurs
  - Pleiosaurus
  - Pterosaurs
  - Saurischians
  - Ornithischians
- Triassic period
  - Cotyllosaurus
  - Dinosaurus
- Permian period
  - Pelycosaurus
  - Therapsids
  - Thecodons
  - Stem reptiles
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
Phylogenetic Trees

Evolutionary Tree, Known Bacteria

1987

Archaea

Eukarya

0.1

Bacteria, 1987

1997

Archaea

Eukarya

Bacteria, 1997

0.1 changes per site

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

**Bacteria**

- Candidate Phyla Radiation
  - (mostly uncultivated)

**Archaea**

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

2016
Phylogenetic Trees

Motivations:

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

- The study of evolutionary relationships.
- Conversion of DNA or protein sequence data into a branching diagram ("tree") that shows the relationships between the sequences.

**the anatomy of a tree**

- Branches
- Edges
- Nodes
- Clades
- Leaves
- Tips
- 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. Provart & D. Guttman
Phylogenetic Trees

example

Adapted from N. Provart & D. Guttman
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

Ancestral Character

Derived Character

Ancestral Character

Homoplaspy

Homology - Similarity due to common ancestry

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

Adapted from N. Provart & D. Guttman
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. Provar & D. Guttman
Ancestral sequence

ACTGAACGTAACCGC

Single substitution

A

C

T

G

Multiple substitution

A

C → A

T

G

Coincidental substitutions

A

C → A

G

Parallel substitutions

A

T → A

A

Convergent substitution

A

A → T

C

G

Back substitution

C

T
Phylogenetic Tree Construction Methods
All tree-building begins with multiple-alignment

- Naïve multiple sequence alignment is NP-complete.
- Students typically don’t want to spend time multiple alignment details.
- 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)*

  **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 usually a worthwhile thing to do.
Phylogenetic Tree Construction Methods

Distance-based tree 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) tree methods

Maximum Parsimony

Maximum Likelihood

Bayesian Methods
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.
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)
Phylogenetic Tree Construction Methods

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

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
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

**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 \( a = b \)
- 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 base freq to vary

**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
- Allow all six pairs of subst to have diff rates
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

Implementation:

- In parsimony, the score is simply the minimum number of mutations that could possibly produce the data.
- Pro: There are fast algorithms that guarantee that any tree can be scored correctly
- Con: 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
Character-Based Phylogenetic Methods

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
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.

Phylogenetic Trees

Unrooted tree for the 4 taxa
Arbitrarily rooted tree for site $j$

Adapted from N. Provart & D. Guttman
**Phylogenetic Trees**

\[ L_j = P(A) \times P(G) \times P(C) \times P(T) \]

The likelihood of the tree is the product of the likelihoods for each site. Usually evaluated as the sum of the log likelihoods.

**Maximum Likelihood**

ML evaluates:
- all possible ancestral states
- at all variable sites
- 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
Phylogenetic Trees

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.
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
Typically 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.

to the context of an

any downloading other

study to root the

lues about the early

evolution.

gle-sequence

ides (or amino

a common ancestral

es, or that a base was

sequence evolution

logenetic methods

terplay between
REVIEWS

Previous experiments. Based on general knowledge or derived from first principles, or available data. Priors can value (The 'prior'). The probability of a value without reference to the (The 'prior'). The probability of a value without reference to the value). The probability of a value without reference to the value).

PRIOR PROBABILITY

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 and graphics
- **ggtree** - ggplot2-specific for trees

---

**Software**

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

---

**References**

- **RAxML**

- **MrBayes**

- **BEAUti / BEAST 1.7**

---

Phylogenetic Tree Construction Methods

**But** we’re not going to build trees in this workshop…

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)1.000:0.0,
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);

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
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, \ldots, n, and internal nodes are coded n+ 1, \ldots, n+m (n+ 1 is the root). Both series are numbered without gaps.

dedge\.length, ndnode\.label, rroot\.edge are optional annotation slots in “phylo” list

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.
Exercise:
Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database

Research Motivation:
Does the region of 16S rRNA gene in my data actually discriminate *Lactobacillus* species?
Example 1: Determine species names of unlabeled \textit{Lactobacillus} species in the GreenGenes database
Example 1: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database

![Phylogenetic tree showing species relationships](image)
Example 1: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database

Consensus

Sequence Logo

Lactobacillus reuteri strain DSM 20016 NR_119069
Lactobacillus iners ATCC 55195 NZ_GL622335
Lactobacillus gasseri strain CIP 102991 NR_117573
Lactobacillus vaginalis strain NCTC 12197 NR_118977
Lactobacillus jensenii strain Gasser 62G NR_117072
Lactobacillus crispatus strain DSM 20584 NR_119274
Lactobacillus acidophilus strain VPI 6032 NR_117062

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

GGOTUID - Species
129798 - L. iners
4428313 - L. gasseri
31171 - L. jensenii
4447432 - L. crispatus / acidophilus
137043 - L. reuteri / vaginalis
338757 - L. mucosae
4441804 - L. brevis
4463108 - L. ruminis
4480189 - L. zeae
586141 - ??
577716 - ??
3851582 - ??
1757845 - ??
4416659 - ??
137043 - ??
Manipulating Trees in phyloseq

• Trees are automatically pruned to match data operations on other parts of phyloseq object

• Use standard taxa functions
  • prune_taxa(), filter_taxa(), subset_taxa()

• Agglomeration
  • tip_glom()
  • tax_glom()

• ape functions after accession:
  • plot.tree(phy_tree(physeq))
  • root(phy_tree(physeq), …)
(Tree-based) Distances between microbiomes
Community Distance

Communities are a vector of abundances:
\[ \mathbf{x} = \{x_1, x_2, x_3, \ldots\} \]

\[ \begin{align*}
E. \text{ coli:} & \quad \bullet \bullet \bullet \\
P. \text{ fluorescens:} & \quad \bullet \\
B. \text{ subtilis:} & \quad \bullet \\
P. \text{ acnes:} & \\
D. \text{ radiodurans:} & \\
H. \text{ pylori:} & \quad \bullet \bullet \bullet \bullet \bullet \bullet \bullet \\
L. \text{ crispatus:} & \\
\end{align*} \]

\[ \mathbf{x} = \{3, 1, 1, 0, 0, 7, 0\} \]
Community Distance Properties

- Range from 0 to 1
- Distance to self is 0
- If no shared taxa, distance is 1
- Triangle inequality (metric)
- Joint absences do not affect distance (biology)
- Independent of absolute counts (metagenomics)
The Distance Spectrum

<table>
<thead>
<tr>
<th>Presence/Absence</th>
<th>Categorical</th>
<th>Phylogenetic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jaccard</td>
<td>Unifrac</td>
</tr>
<tr>
<td>Quantitative Abundance</td>
<td>Bray-Curtis</td>
<td>Weighted Unifrac</td>
</tr>
</tbody>
</table>

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)
Unifrac

\[ \text{Dist}(x, y) = \frac{+}{+ + +} \]

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone and Knight (2008)
Unifrac

\[ \text{Dist}(x, y) = \frac{\text{red} + \text{blue}}{\text{red} + \text{blue} + \text{purple}} \]

**Intuition:** Fraction of shared **tree** unique to one of the communities

---

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)
Weighted Unifrac

Weighted UniFrac
branchlengths weighted by difference in red and blue

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)
**Intuition**: The cost of turning one distribution into the other; where the cost is the amount of “dirt” moved times the distance by which it is moved.
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
Jaccard:
Bray:
Unifrac:
W-Unifrac:

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)
Jaccard: d=0
Bray: Distant
Unifrac:
W-Unifrac:

Jaccard: Distant
Bray: Similar
Unifrac:
W-Unifrac:

Jaccard: Distant
Bray: Distant
Unifrac:
W-Unifrac:

Slide graciously provided by Benjamin Callahan, not necessarily with permission 😃)
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)
Jaccard: d=0  
Bray: Distant  
Unifrac: d=0  
W-Unifrac: Distant

Jaccard: Distant  
Bray: Similar  
Unifrac: Similar  
W-Unifrac: Similar

Jaccard: Distant  
Bray: Distant  
Unifrac: Distant  
W-Unifrac: Similar

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)
The Distance Spectrum

<table>
<thead>
<tr>
<th>Presence/Absence</th>
<th>Categorical</th>
<th>Phylogenetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>Unifrac</td>
<td></td>
</tr>
<tr>
<td>Bray-Curtis</td>
<td>Weighted Unifrac</td>
<td></td>
</tr>
</tbody>
</table>

**phyloseq distances**
- manhattan
- euclidean
- canberra
- bray
- kulczynski
- jaccard
- gower
- altGower
- morisita-horn
- mountford
- raup
- binomial
- chao
- cao
- jensen-shannon
- unifrac
- weighted-unifrac
...
That’s great, Joey… What do we do with these distances???

Alex is going to go over ordination methods for interpreting the distance matrix derived from comparing all the samples in your data…

What we learned here was…
• A survey about how to think about trees
• How trees are represented and interact with phyloseq
• An introduction about different definitions for a distance between two microbiomes
End