Sequence Analysis '16
lecture 19

- Tree properties
- Tree comparison
- Tree evaluation
- Tree significance (bootstrapping)
METRIC DISTANCES between any two or three taxa (a, b, and c) have the following properties:

Property 1: \( d(a, b) \geq 0 \)  \quad \text{Non-negativity}

Property 2: \( d(a, b) = d(b, a) \)  \quad \text{Symmetry}

Property 3: \( d(a, b) = 0 \) if and only if \( a = b \)  \quad \text{Distinctness}

Property 4: \( d(a, c) \leq d(a, b) + d(b, c) \)  \quad \text{Triangle inequality}

triangle inequality
A special distance metric...

ULTRAMETRIC DISTANCES

...must satisfy the previous four conditions, plus:

Property 5  The distances from any branch point to the taxa in the clade defined by that branch point are equal.

If distances are ultrametric, then the sequences are evolving in a perfectly **clock-like manner**. So any two sequences always have the same distance to their common ancestor.
One more property of patristic distances

Additivity

Property 6: Example: if (a,b) are nearest neighbors,
\[ d(a, b) + d(c, d) \leq \text{maximum } [d(a, c) + d(b, d), d(a, d) + d(b, c)] \]

For distances to fit into an evolutionary tree, they must be additive. Estimated distances often fall short of these criteria, and thus can fail to produce correct evolutionary trees.

A lineage that violates additivity implicitly goes *backwards in time*. 
In class exercise:
What’s wrong with these distances?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>
In class exercise:
What’s wrong with this tree and these branch lengths?
In class exercise:
What’s wrong with these distances on this tree?
Tree comparison

- If two trees have the **same** topology, compare the branch lengths.
  \[ S = \sum_{i>j} w_{ij} (d_{ij} - \Delta_{ij})^m \]

- If two trees have **different** topologies, calculate the **symmetric distance**.
  \[ D = \text{number of “splits” that are different.} \]
  A “split” is a bisection of the tree at an **internal** (not leaf) lineage.
  A split is different if the taxa on either side of the split are different.

\[
\begin{align*}
\text{Splits} & \quad \text{Splits} \\
AB, CDE & \quad AB, CDE \\
ABC, DE & \quad ABD, CE
\end{align*}
\]

**symmetric distance = 1, out of 2**

A 4-taxa tree has 1 split. A 5-taxa tree has 2 splits. A N-taxa tree has N-3 splits.
Tree evaluation

• A tree topology may be evaluated relative to a set of distances. To select the tree that best fits the data we need...

(1) A way to score each tree.
   (a) Parsimony
   (b) Fitch-Margoliash
   (c) Least-squares
   (d) Maximum likelihood

(2) A way to generate trees.
   (a) Nearest neighbor interchange (NNI)
   (b) Subtree pruning and re-grafting (SPR)
   (c) Tree bisection and reconnection (TBR)
   (d) Exhaustive??
Why we can’t try all trees….

Explosive tree growth

<table>
<thead>
<tr>
<th># Taxa (N)</th>
<th># Unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10,935</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
</tr>
<tr>
<td>30</td>
<td>(\approx 3.58 \times 10^{36})</td>
</tr>
</tbody>
</table>

Exhaustive search is possible only for small numbers of taxa.

\[
\frac{(2N-5)!}{[2^{N-3}*(N-3)!]} = \text{# unrooted trees for N taxa}
\]
Heuristic method: find best tree

1. Start with a NJ tree.
2. Until convergence:
   2.1. Modify tree topology
   2.2. Score tree against sequences
   2.3. If score goes up, keep new tree, else revert.
Exploring similar tree topologies:
Nearest neighbor interchange

Choose internal branch (lineage)  Condense to 4 groups (clades)

interchange, 2 ways.

NNI makes small changes in the tree. “Local search”. 
Exploring similar tree topologies:

Subtree pruning and re-grafting

Choose internal branch (lineage)

Pick a branch. Split the tree. (part A, part B)

Pick a branch on Part B.
Graft the cut point of Part A to that branch.

SPR makes large changes in the tree.
Exploring similar tree topologies:

Tree bisection and reconnection

Choose internal branch (lineage)

Pick a branch.
Split the tree. (part A, part B)

Randomly select one lineage in A, one lineage in B.
Connect them.

TBR makes even larger changes in the tree.
Scoring trees based on parsimony

- Requires a MSA
- Considers the characters.
- May be done using non-character, non-metric information, such as phenotypes.
- Assumes minimum evolution.
Scoring trees based on distances

- Compare sequence distances (J-C corrected) to patristic distances.
- Does not consider the characters.
- May not be done using non-character, non-metric information, such as phenotypes.
- No minimum evolution assumption. Maximizes correlation.
- Ways to get patristic distances:
  1. UPGMA (rooted tree. ultrametric distances.)
  2. Fitch-Margoliash (unrooted or rooted)
  3. Least-squares (unrooted or rooted)

\[ S = \sum_{i > j} w_{ij} (d_{ij} - \Delta_{ij})^m \]

Weights. May be a function of \( d_{ij} \)

May be anything. Typically 2

Sequence distances

Tree (patristic) distances
Least squares solution for tree branch lengths

\[ S = \sum_{i>j} w_{ij} (d_{ij} - \Delta_{ij})^m \]

\[
\begin{pmatrix}
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
b_1 \\
b_2 \\
b_3 \\
b_4 \\
b_5
\end{pmatrix}
= 
\begin{pmatrix}
d_{AB} \\
d_{AC} \\
d_{AD} \\
d_{BC} \\
d_{BD} \\
d_{CD}
\end{pmatrix}

The \( T \) matrix expresses the branches needed to sum the tree distance. Branch distances \( b \) are found by squaring and inverting the \( T \) matrix and multiplying by the sequence distances \( d \). New tree distances \( \Delta \) are found by summing \( b \), using \( T \).

\[ b = (T^T T)^{-1} (T^T d) \]
Least squares: solve for $\mathbf{b}$

$\mathbf{Tb} = \mathbf{d}$

$\begin{bmatrix}
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
b_1 \\
b_2 \\
b_3 \\
b_4 \\
b_5
\end{bmatrix} =
\begin{bmatrix}
d_{AB} \\
d_{AC} \\
d_{AD} \\
d_{BC} \\
d_{BD} \\
d_{CD}
\end{bmatrix}
= \begin{bmatrix}3 \\ 5 \\ 7 \\ 4 \\ 6 \\ 3\end{bmatrix}$

$\mathbf{b} = (\mathbf{T}^\top \mathbf{T})^{-1}(\mathbf{T}^\top \mathbf{d})$

Invert this to get $(\mathbf{T}^\top \mathbf{T})^{-1}$
\[ b_1 \]
\[ b_2 \]
\[ b_{31} = (T^T T)^{-1} (T^T d) = \]
\[ b_3 \]
\[ b_4 \]
\[ b_5 \]

Sequence distances

\[
\begin{align*}
    d_{AB} &= 3 \\
    d_{AC} &= 5 \\
    d_{AD} &= 7 \\
    d_{BC} &= 4 \\
    d_{BD} &= 6 \\
    d_{CD} &= 3
\end{align*}
\]

Least squares patristic distances

\[
\begin{align*}
    \Delta_{AB} &= \begin{bmatrix}
        1 & 1 & 0 & 0 & 0 \\
        1 & 0 & 1 & 0 & 1 \\
        1 & 0 & 0 & 1 & 1 \\
        0 & 1 & 1 & 0 & 1 \\
        0 & 1 & 0 & 1 & 1 \\
        0 & 0 & 1 & 1 & 0
    \end{bmatrix}
\end{align*}
\]

\[
S = \sum_{i>j} (d_{ij} - \Delta_{ij})^2 = 0.0
\]
Least squares distances for the **wrong** tree

\[ b = (T^T T)^{-1} (T^T d) \]

\[
\begin{pmatrix}
\Delta_{AC} \\
\Delta_{AB} \\
\Delta_{AD} \\
\Delta_{BC} \\
\Delta_{CD} \\
\Delta_{BD}
\end{pmatrix} = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 0
\end{bmatrix} \begin{bmatrix}
1/2 & 0 & 0 & 0 & -1/4 \\
0 & 1/2 & 0 & 0 & -1/4 \\
0 & 0 & 1/2 & 0 & -1/4 \\
0 & 0 & 0 & 1/2 & -1/4 \\
-1/4 & -1/4 & -1/4 & -1/4 & -1/4 \\
-3/4 & 0 & 0 & 0 & 0
\end{bmatrix} \begin{bmatrix}
5 \\
3 \\
7 \\
4 \\
3 \\
6
\end{bmatrix}
\]

\[ S = \sum_{i>j} (d_{ij} - \Delta_{ij})^2 = 58.5 \]

Least-squares produces a negative distance for \( b_5 \)

Set negative distance to zero
Max Unweighted Parsimony: Trying all trees

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ATGGCTATTCTTTATAGTACG</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>ATGGCTAGTCTTTATATTACA</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>TTCACCTAGACCTGTGGTCCA</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>TTAGCCAGACCCTGTGGTCCG</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>TTAGCCAGTTCTCTAGTTTCG</td>
<td></td>
</tr>
</tbody>
</table>

**TOTALS**

```
1 0 1
1 0 1
2 0 1
1 0 1
```
Maximum likelihood

Given a tree with branch lengths and MSA, sum the probabilities of the branches over all possible ancestor bases (or amino acids).

Probabilities may be J-C or Kimura for nucleotides, or PAM or BLOSUM for proteins.

May be used in combination with NNI, SPR, TBR to evaluate tree topology, with or without branch lengths (if no branch lengths, assume all branch length are equal).

May be used to produce ancestral sequences.

Sum the total likelihood of all branches over all possible bases at two ? positions.
Scoring quartets using Parsimony

Given this sequence data

<table>
<thead>
<tr>
<th>taxon</th>
<th>sequence position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>J</td>
<td>T</td>
</tr>
<tr>
<td>H</td>
<td>G</td>
</tr>
<tr>
<td>P</td>
<td>G</td>
</tr>
</tbody>
</table>

which quartet is most likely?

Find quartet AHJP in this tree.

Score each tree for AHJP

*Maximum likelihood is parsimony taken to the next level. Instead of picking the most parsimonious ancestor sequence, we sum the probabilities of all options. We choose the tree that maximizes this sum.
Quartet puzzling. Tree based on ML quartets

1. Identify all choose-4 subsets of MSA => “quartets”
2. For each quartet, find the best 4-taxa tree (3 possibilities) using least-squares, Fitch-Margoliash, Parsimony, or Maximum Likelihood.
3. Start with a randomly selected quartet, ((A,B),(C,D))
4. Choose one new taxon, E.
5. For every branch, try adding the E to the branch.
   1. For all quartets containing E and three of the other taxa, ask whether the split is correct. If not, add 1 to the penalty.
6. Add E to the branch with the lowest penalty.
7. Continue at step 4 with a new taxon.

QP can be repeated, starting with different quartet, to generate any number of trees.
**Expert tree strategy**

- Get sequence **distances**.
- **QP with ML** -- Use ML to choose topology for each quartet. Use QP to build trees from quartets.
- **NNI, SPR, TBR with L-S, F-M** -- Modify the tree, calculate branch lengths, calculate patristic distances, calculate difference distance score $S$.
- **Consensus tree** -- if more than one tree has same best score, merge branches.
- **Confidence** — For each branch, assign a “bootstrap” value.
“Boot strap analysis”

- A method to validate the significance of a phylogenetic tree, branchpoint by branchpoint.
- Requires a means to generate independent trees. (For example using distances generated from different regions of the mitochondrial genome or from random subsets of sequences or random subsets of columns.)
- Choose the representative tree as the ‘parent’. Calculate the following:

For each branchpoint in the parent tree,
  For each tree, ask
    Is there a branchpoint having the same subclade contents (i.e. same taxa, any order)
Bootstrap value = number of trees having the branchpoint / total trees.
Comparing branchpoints

![Diagram showing branchpoints and bootstrap values.]

- **Bootstrap value** = \( P((A,B),C) = \frac{5}{8} \)

For each branchpoint in the parent tree,
   - For each tree, ask
     - Is there a branchpoint having the same subclade contents (i.e. same taxa, any order)

Bootstrap value = number of trees having the branchpoint / total trees.
Comparing branchpoints

bootstrap value = \( P((A,B,C),(D,E)) = \frac{6}{8} \)
Bootstrap values for this data
Compare lineages instead of branchpoints.

\[ \bullet = P(A,B) = \frac{6}{8} \]

For unrooted trees or rooted trees.
• Treat any lineage as the root.
• Ask how often the root branching is conserved.
Review

- What are patristic distances?
- How can you tell if a tree is non-additive? What does non-additivity imply?
- How do you compare two trees that have different topology?
- How do you score a set of distances against a tree?
- Assuming an initial tree, what ways exist for exploring other tree topologies?
- Describe nearest neighbor interchange.
- What kind of program is quartet puzzling?
- ... Fitch-Margoliash?
- Compare and contrast three ways to get branch lengths from distances: UPGMA, Fitch-Margoliash, Least-squares.