Sequence Analysis '17--lecture 10

Trees

- types of trees
- Newick notation
- UPGMA
- Fitch Margoliash
- Distance vs Parsimony
Phylogenetic trees

What is a phylogenetic tree?
A model of evolutionary relationships -- common ancestors and speciation events.

Why build phylogenetic trees?
To trace the branch order of "taxa" (taxon = a gene, a species, a population, etc.)
To understand the evolution of traits
As part of a multiple sequence alignment algorithm

Trees can be "rooted" or "unrooted"
Tree Terminology

Lineages

Common ancestors (hypothetical)

outgroup

taxa

time (rooted trees only)

root

Taxa are observed species or genes.
Inferring evolutionary *relationships* between the taxa requires rooting the tree:

To root a tree mentally, imagine that the tree is made of string. Grab the string at the root \(\bigcirc\) and tug on it until the ends of the string (the taxa) fall opposite the root:
Where the tree is rooted changes its meaning.

Each of these trees is possible by choosing a different root.

This one says C and D branched late.

This one says C and D branched early.
Taxon order doesn't matter.

Rotating an ancestor node does not change the relationships.

UGENE: “swap siblings”
Two strategies for rooting a tree: "outgroup" and "midpoint"

1. Choose the **midpoint** between the two most distant branches.

2. Choose one taxon as the "**out group**." (it branches first.)

A good outgroup is not too distant from the rest of the tree.
Newick notation

\[(D, (C, (A, B)))\]

Trees can be represented in plain text Newick notation. Each set of parentheses represents a branch-point (split), the comma separates left and right lineages. Implies a rooted tree.

Newick notation can contain sequence labels or not.
Here is a Newick Tree for 50 taxa

```
((((((('Phoca caspica':0.07021230607681982,'Halichoerus grypus':0.07021230607681982),
      'Phoca sibirica':0.07611418753416775),
     'Phoca largha':0.04443967284675786),
    'Phoca vitulina':0.038343188853643255),
   'Cystophora cristata':0.01378136689767208),
  'Hydrurga leptonyx':0.15128664129846624),
 ((('Arctocephalus australis':0.057399989415323844,'Arctocephalus forsteri':0.057399989415323844),
   'Neophoca cinerea':0.18323401004037046),
  'Callorhinus ursinus':0.5246989884225943),
 (((('Enhydra lutris':0.59169678085082259,'Lontra canadensis':0.7507168117788424),
   'Martes americana':0.1104991428194073),
   'Alopex lagopus':0.326843130446780029364),
  'Ailuropoda melanoleuca':1.116821080259208),
   'Manis tetradactyla':4.385450608816927);
```
Did the *Florida Dentist* infect his patients with HIV?

Phylogenetic tree of HIV sequences from the DENTIST, his Patients, & Local HIV-infected People:

Yes:
The HIV sequences from these patients fall within the clade of HIV sequences found in the dentist.

No:
The HIV sequences from these patients do not fall within the clade of HIV sequences found in the dentist.

From Ou et al. (1992) and Page & Holmes (1998)
Evolutionary time

Cladogram

Phylogram

Ultrametric tree


Newick format with distances
Neighbor joining: cladogram

Choose the closest neighbors. Add a node between them. Choose the next closest, ad so on.

<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>Species A</td>
<td>----</td>
<td>0.20</td>
<td>0.50</td>
<td>0.45</td>
<td>0.40</td>
</tr>
<tr>
<td>Species B</td>
<td>0.23</td>
<td>----</td>
<td>0.40</td>
<td>0.55</td>
<td>0.50</td>
</tr>
<tr>
<td>Species C</td>
<td>0.87</td>
<td>0.59</td>
<td>----</td>
<td>0.15</td>
<td>0.40</td>
</tr>
<tr>
<td>Species D</td>
<td>0.73</td>
<td>1.12</td>
<td>0.17</td>
<td>----</td>
<td>0.25</td>
</tr>
<tr>
<td>Species E</td>
<td>0.59</td>
<td>0.89</td>
<td>0.61</td>
<td>0.31</td>
<td>----</td>
</tr>
</tbody>
</table>
**UPGMA**

Assumes constant evolutionary clock. All sequences have the same distance to the common ancestor. *Ultrametric* tree.

### Raw p-distances

<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>Species A</td>
<td>---</td>
<td>0.20</td>
<td>0.50</td>
<td>0.45</td>
<td>0.40</td>
</tr>
<tr>
<td>Species B</td>
<td>0.23</td>
<td>---</td>
<td>0.40</td>
<td>0.55</td>
<td>0.50</td>
</tr>
<tr>
<td>Species C</td>
<td>0.87</td>
<td>0.59</td>
<td>---</td>
<td>0.15</td>
<td>0.40</td>
</tr>
<tr>
<td>Species D</td>
<td>0.73</td>
<td>1.12</td>
<td>0.17</td>
<td>---</td>
<td>0.25</td>
</tr>
<tr>
<td>Species E</td>
<td>0.59</td>
<td>0.89</td>
<td>0.61</td>
<td>0.31</td>
<td>---</td>
</tr>
</tbody>
</table>

### J-C corrected distances

1) Generate neighbor-joining tree. (NJ)
2) For first neighbors, distance to ancestor is $d_{ij}/2$
3) For next neighbors, distance to ancestor is average pairwise distance between taxa in two clades, divided by two.
4) Subtract to get lineage distances.

Distance to common ancestor =
Average distance between taxa in two clades, divided by two.

\[
\frac{0.115 + 0.085}{2} = ____.\]

[Diagram and calculations shown]
Problem statement: Given a tree and a set of sequence distances, derive lineages such that the tree distances maximally match the sequence distances.
Fitch-Margoliash algorithm for calculating the branch lengths

1. Find the most closely-related pair of sequences, A and B

2. Calculate the average distance from A to all other sequences, then from B to all other sequences.

3. Adjust the position of the common ancestor node for A and B so that the difference between the averages is equal to the difference between the A and B branch lengths, while the sum of the branch lengths is still equal to $D_{AB}$. ($D=$sequence distance, $d=$lineage distance)

$$d_A - d_B = (D_{AC} + D_{AD})/2 - (D_{BC} + D_{BD})/2$$

NOTE: the difference between the averages may be greater than $D(A,B)$, making step 3 impossible.
Exercise 8: create a rooted phylogram with 4 taxa

Directions:
1. Make a distance matrix. (p-distance, then convert to J-C distance)
2. Use Neighbor-joining to make a tree.
3. Adjust branch lengths using Fitch-Margoliash.
4. Choose the root using the Midpoint method.
5. Write tree in Newick format.
Orthologs: homologs originating from a speciation event.
Paralogs: homologs originating from a gene duplication event.

clam A and fish A are orthologs. clam A and crab B are paralogs.
How do I know it’s a paralog?

- If it’s a **paralog**, then at some point in evolutionary history, a species existed with two identical genes in it.
  - One may have been lost since then. (Descendants are still paralogs!)
  - Paralogs can be from **different species**.
- Paralogous genes have **more than the expected sequence divergence**.
  - Because they are more likely to have different functions
  - Because they diverged earlier than the speciation event.
- Without **species information** or **functional information**, it’s impossible to tell orthologs from paralogs.
Maximum parsimony -- it's "character-building"

Optimality criterion: The 'most-parsimonious' tree is the one that requires the \textit{fewest number of evolutionary events} (e.g., nucleotide substitutions, amino acid replacements) to explain the sequences.

For this column, and this tree, one mutation event is required.
character-based tree-building

For this other column, the same tree requires **two** mutation events. A different tree would require only one.

A: ATGGCTATTCTTAAGTACG
B: ATCGCTAGTCTTATTACA
C: TTCACTAGACCTGTGGTCCA
D: TTGACCAAGACCTGTGGTCCG
E: TTGACCAAGTTCTCTAGTTCCG
Finding the minimum number of mutations

Given a tree and a set of taxa, one-letter each (1) choose optional characters for each ancestor. (2) Select the root character that minimizes the number of mutations by selecting each and propagating it through the tree.

minimum 2 mutations

minimum 1 mutation
Parsimony tips:

Ignore non-informative sites

- **No** mismatches --> *noninformative*!
  Adds 0 mutations in all trees
- **One** mismatch --> *noninformative*!
  Adds 1 mutation in all trees.
- **All** different --> *noninformative*!
  Adds same number of mutation in all trees. Only possible if number of sequences ≤ alphabet
Find noninformative columns

<p>| | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ATGGCGATTTCTTTATAGTACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATGGCCTAGTGTATTACAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATCAACAGATCTGTGGTCCAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATGACCAGAACTGTGGTCCG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sum the Max Unweighted Parsimony for 4 5-taxa trees

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ATGGCTATTCTTAGTGTCACG</td>
<td>ATGGCTAGTCTTATATTTACA</td>
<td>TTCACTAGGCTCCTGTGTTCCA</td>
<td>TTGACCAAGATCCTGTGTTCCG</td>
<td>TTGACCAAGTTCTCTAGTTTCG</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence similarity</td>
<td>Method to use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strong</td>
<td>distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>weak</td>
<td>parsimony</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very weak</td>
<td>maximum likelihood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Review

- What do nodes and lineages in a tree represent?
- What are the two strategies for rooting a tree? Why root a tree?
- What is maximum parsimony? How do I find the most parsimonious tree?
- What is maximum likelihood? How does it relate to maximum parsimony?
- What kind of MSA positions are not informative of the tree?
- What problem does Fitch-Margoliash solve? How?
- What is an ortholog? Paralog? How can I tell them apart?
- Unroot and re-root this tree: ((A,B),((C,D),E))