Phylogenetics

Pradipta Ray,
BIOL 6385 / BMEN 6389,
The University of Texas at Dallas
(some material based on content by PR in Eric Xing’s 10-810 Carnegie Mellon class)
Ernst Haeckel, Tree of life, 1879
Some nomenclature on trees

- **Binary** (bifurcating) or multiway (multifurcating)

- A multiway split can be envisioned as a series of binary splits
Recursively defining trees: subtrees

Trees can be defined in terms of themselves
Induced subtree

- For a subset of nodes, the induced subtree is the minimal set of edges and nodes to connect them together, taken from the original tree.
Parameterizing a tree

- Topology: **connectivity** of the tree
  - path between any two nodes is unique!
  - topology unchanged by “squishing” a drawing of the tree

- Length of edges: the **geometry** of the tree
  - notion of “distance” between neighboring nodes

- Labels of nodes (?): which nodes were not born equal: **identifying** the ones we are interested in
Rooted trees

• “Rooting”: a notion of direction of flow
  – in case of phylogenies, flow of time
• Root: a privileged node
• Direction flows outward from the root
• Rooted tree = Directed edges away from root

Figure 1. Topologies of phylogenetic trees: A. unrooted tree; B. rooted tree.

Perrero & Lopez, Gen Mol Res, Sep 2005
Directed edges

• Directed edges away from root = Rooting

This is a tree. (Pick it up by the shaded/root node)
Is this graph a rooted tree?

• Do a topological sorting on it, is there a unique root?

• Can we order all the nodes in the direction of the edges, and be left with a single topmost node
Rooted tree nomenclature

- **Parent / ancestor**
- **Child / descendant**
- **Sibling**
Ordered vs unordered branches

- Is the order in which we represent the siblings important? $2^n$ ways to draw for $n$ interior nodes (rooted)

- Which is different?
Proper vs improper trees

- Proper: each node has 0 or 2 children (rooted), each node has 1 or 3 neighbors (unrooted)
Phylogeny / phylogenetic tree

• Taxon/taxa/operational taxonomic unit (OTU)
  – unit of classification: species, subspecies, individual, etc

• Phylogeny = evolutionary tree
  – Hypothesis concerning evolutionary history of taxa
Molecular Phylogeny

- Molecular phylogeny = based on models (or distances based on models) of molecular evolution
Phylogenetic Trees

- Edges = branches
- Interior nodes: ancestral taxa
- Leaf / exterior nodes: contemporary taxa
- Topology: relation between species
- Branch length = edge weights: amount of change
- Labelled leaves, unlabelled ancestors
Phylogenies have unordered branches

Ordering has no evolutionary connotation
Phylogenetic trees are proper trees

Improper phylogenies transformed to proper ones

– Each node = Common ancestor of subset of species

– Unrooting
Most recent common ancestor
Rooting an unrooted tree

• Root at any branch

• Sometimes, we may not know (or may not care) where the root is
Counting edges

• For an unrooted tree on n leaves, we have
  – n-2 internal nodes
    • (proof by induction, tree on 3 leaves has 1 internal node, every additional leaf incorporated into a tree adds one leaf and one internal node)
  – 2n-3 edges [no of nodes minus one]
    • (think of shrinking the tree one node and edge at a time)
Counting labelled leaf, unlabelled ancestor phylogenetic trees

• Topologically equivalent = a tree changed to another by flipping neighbors, w/o breaking branches

• For a tree of n - 1 leaves, 2n – 5 branches: add nth leaf to a branch: T(n) = T(n-1) X (2n – 5)
  – T(3) = 1

• If rooted, rooting can happen at any branch: additional factor of (2n - 5)
No of unrooted trees : $T(k)$

$$T(k) = \frac{(2k - 5)!}{2^{k-3} (k - 3)!}$$
Grows fast

- Felsenstein, Counting trees

Table 1. The numbers of rooted trees with n labelled tips and with unlabelled interior nodes. The left column counts all trees, the right column only bifurcating trees.

<table>
<thead>
<tr>
<th>n</th>
<th>All trees</th>
<th>Bifurcating trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>236</td>
<td>105</td>
</tr>
<tr>
<td>5</td>
<td>2,725</td>
<td>945</td>
</tr>
<tr>
<td>6</td>
<td>39,205</td>
<td>10,395</td>
</tr>
<tr>
<td>7</td>
<td>650,032</td>
<td>135,135</td>
</tr>
<tr>
<td>8</td>
<td>12,818,912</td>
<td>2,027,025</td>
</tr>
<tr>
<td>9</td>
<td>282,137,924</td>
<td>34,459,425</td>
</tr>
<tr>
<td>10</td>
<td>6,930,807,856</td>
<td>654,729,075</td>
</tr>
<tr>
<td>11</td>
<td>188,666,182,784</td>
<td>13,749,310,575</td>
</tr>
<tr>
<td>12</td>
<td>5,617,340,020,544</td>
<td>316,334,143,225</td>
</tr>
<tr>
<td>13</td>
<td>181,750,701,299,729</td>
<td>7,905,853,580,625</td>
</tr>
<tr>
<td>14</td>
<td>6,383,726,042,466,112</td>
<td>213,438,646,678,875</td>
</tr>
<tr>
<td>15</td>
<td>239,513,970,965,250,048</td>
<td>6,190,333,353,629,375</td>
</tr>
<tr>
<td>16</td>
<td>9,571,020,586,418,569,216</td>
<td>191,998,730,962,510,625</td>
</tr>
<tr>
<td>17</td>
<td>408,837,905,660,430,516,224</td>
<td>6,332,059,870,792,850,625</td>
</tr>
<tr>
<td>18</td>
<td>18,523,305,410,364,568,764,416</td>
<td>221,643,095,476,699,771,875</td>
</tr>
<tr>
<td>19</td>
<td>887,094,711,304,694,383,095,296</td>
<td>8,200,794,332,637,891,559,375</td>
</tr>
<tr>
<td>20</td>
<td>44,782,216,857,751,551,081,214,592</td>
<td>319,530,986,772,877,770,815,625</td>
</tr>
<tr>
<td>21</td>
<td>2,576,615,641,929,796,906,245,518,104</td>
<td>13,113,079,437,987,698,803,440,625</td>
</tr>
</tbody>
</table>

- No of trees on 500 taxa ~ 1 X 10^1074
- No of atoms in observable universe ~ 10^80
Minimum evolutionary hypothesis

• In terms of topology: the star topology (not a binary tree)
  – only assumes a most recent common ancestor
Resolution

- Resolution = process of figuring out topology = generating more complicated evolutionary hypotheses
- Partial resolution = intermediate stage
Outgroups

• When studying a group, we may want a control which is outside that group

• One way to root a tree
2 schools of phylogeny reconstruction

- Distance based methods

<table>
<thead>
<tr>
<th></th>
<th>AFR</th>
<th>NEC</th>
<th>EUC</th>
<th>NEA</th>
<th>ANE</th>
<th>AME</th>
<th>SEA</th>
<th>PAI</th>
<th>NGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>African</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-European Caucasian</td>
<td>1340.0 ± 301</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European Caucasian</td>
<td>1655.6 ± 416</td>
<td>154.7 ± 29</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast Asian</td>
<td>1979.1 ± 452</td>
<td>640.4 ± 134</td>
<td>938.2 ± 217</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arctic Northeast Asian</td>
<td>2098.5 ± 387</td>
<td>708.2 ± 160</td>
<td>746.7 ± 210</td>
<td>459.7 ± 98</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amerindian</td>
<td>2261.4 ± 434</td>
<td>955.5 ± 204</td>
<td>1038.2 ± 276</td>
<td>746.5 ± 183</td>
<td>577.4 ± 89</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast Asian</td>
<td>2206.5 ± 529</td>
<td>939.6 ± 202</td>
<td>1240.4 ± 335</td>
<td>630.5 ± 299</td>
<td>1039.4 ± 326</td>
<td>3417 ± 416</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>2805.4 ± 648</td>
<td>953.7 ± 230</td>
<td>1344.7 ± 364</td>
<td>723.8 ± 262</td>
<td>1181.2 ± 331</td>
<td>1740.7 ± 544</td>
<td>436.7 ± 87</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>New Guinean and Australian</td>
<td>2472.0 ± 536</td>
<td>1179.1 ± 189</td>
<td>1345.7 ± 231</td>
<td>734.4 ± 118</td>
<td>1012.5 ± 257</td>
<td>1457.9 ± 283</td>
<td>1237.9 ± 277</td>
<td>808.7 ± 264</td>
<td>0.0</td>
</tr>
</tbody>
</table>
2 schools of phylogeny reconstruction

• Character based methods

A B C D E
A X C D E
A X Y D E

Figure: A Phylogenetic Tree

molgen.mpg.de

BIOL 6385, Computational Biology
What kind of data can we use?

- Both genetic and phenotype based models of evolution can be either character based or distance based.
Distance based methods

• How to infer evolutionary relationships on the basis of some similarity measure?

• Notion of a similarity measure
  – eg. Curvature and length
  – 2D Euclidean distance
Genetic distance

• Similar to alignment score
  – how far away are two orthologous sequences in sequence space?
  – alignment usually required (compare apple to apple)
  – one option: pairwise entropy measure
  – similar to $D_{OBS} =$ fraction of sites with substitutions in pairwise alignment
Model based genetic distances

• Typically such simple measures can't capture complicated evolutionary models (which model selection)

• However, in JC 69, all mutations are equally likely: can we use $D_{OBS}$ as distance measure?
Correction for multiple substitutions

- A → T → C (multiple mutations when we see one substitution)
- A → T → A (multiple mutations when we see no substitution)

\[ D_{JC} = -\frac{3}{4} \ln \left[ 1 - \frac{4 \text{Dobs}}{3} \right] \]

\[ P(2 \text{ or more mutation at each site}) = \text{infinite series [HW]} \]
Metric

- Mathematical notion of “distance”
- Intuitive properties

\[
\begin{align*}
    d(x, y) &> 0 \quad \text{for } x \neq y \\
    d(x, y) &= 0 \quad \text{for } x = y \\
    d(x, y) &= d(y, x) \quad \forall x, y \\
    d(x, y) &\leq d(x, z) + d(y, z) \quad \forall x, y, z \quad \text{(triangle inequality)}
\end{align*}
\]

- Space in which \(x, y, z\) lives + metric definition
  - Metric space
Additive metrics & ultrametrics

- Additive / tree metric
  - remember, route between 2 points on a tree is unique

- Ultrametric: $d(x,y) \leq \max(d(x,z), d(y,z))$

Where should we depict the set of metrics corr. to rooted trees?
Tree metrics : 4 point condition

• Consider every quartet of leaves in the tree

• 3 ways to split into two subsets of 2 nodes :
  – AB & CD, AC & BD, AD & BC

• \( d(AB) + d(CD) \leq \max ( d(AC) + d(BD), d(AD) + d(BC) ) \)
• \( d(AC) + d(BD) \leq \max ( d(AB) + d(CD), d(AD) + d(BC) ) \)
• \( d(AD) + d(BC) \leq \max ( d(AC) + d(BD), d(AB) + d(CD) ) \)
4PC iff additivity

\[ \binom{n}{2} \text{ equations} \]

\[ 2n - 3 \text{ variables} \]

\[ u + \omega = d \ (AB) \]
\[ u + \omega + x = d \ (AC) \]
\[ u + \omega + y = d \ (AD) \Rightarrow \text{have unique soln.} \]
\[ v + \omega + x = d \ (BC) \]
\[ v + \omega + y = d \ (BD) \]
\[ x + y = d \ (CD) \]
Ultrametric = 3 point condition

\[ d(x, y) \leq \max \left( d(x, z), d(y, z) \right) \]

- Consider every triplet of leaves A, B, C
- \( d(AB) \leq \max \left( d(AC), d(BC) \right) \)
- \( d(BC) \leq \max \left( d(AB), d(AC) \right) \)
- \( d(AC) \leq \max \left( d(AB), d(BC) \right) \)
3 PC satisfied (hence 4 PC satisfied)

**Diagram:***

- A, B, C, D

- Cladered tree

**Grid:**

```
  A  B  C  D
0   2   4  2
0   0   4  2
0   0   4  2
0   0   4  2
```

**Legend:**

- A
- B
- C
- D
A few changes to 3 PC matrix may still be 3 PC matrix!

Some changes may still keep the clock.
Some changes to 3 PC matrix may only satisfy 4 PC, not 3 PC

Is it docked?
Subtrees may still be docked
Some changes to 4 PC matrix can preserve 4 PC condition

Some changes can preserve additivity
Some changes to 4 PC: only triangle inequality satisfied

\[ AC + BD > \max( AB + CD, AD + BC) \]

Can we measure consistent branch lengths for any topology?
So, what changed?

We need to increase 1, x, or z. If other entries will become consistent, but we can plot them in a metric space!
Arbitrary changes to the distance matrix

\[ D_{BC} > D_{BA} + D_{AC} \]

We can no longer plot the points in any metric space.
## Comparison

### Distance metrics

<table>
<thead>
<tr>
<th>Distance Metric</th>
<th>Additive Metric</th>
<th>Ultrametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangle Inequality</td>
<td>4 Point Condition</td>
<td>3 Point Condition</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property</th>
<th>Polynomial time greedy algorithm to reconstruct tree for all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topology recovered, branch lengths may be inconsistent</td>
<td>Fits unique unrooted (or one of its many equivalent rooted) tree with unique branch lengths</td>
</tr>
<tr>
<td>Points may still be plotted in metric space</td>
<td>Tree may be drawn</td>
</tr>
<tr>
<td></td>
<td>Tree drawn, one point on tree (root) is equidistant from all leaves</td>
</tr>
<tr>
<td></td>
<td>Fits unique rooted, clocked tree with branch lengths</td>
</tr>
</tbody>
</table>

Stricter condition, set of matrices satisfying this shrinks.
Clustering

• Partitioning some data points (species or individuals in our case) based on some metric or distance measure
Clustering is hard

- k-means is NP hard
Hierarchical clustering

- We may want to find such clusters at different levels: all the way from the whole data set to individual data points.
- Notion of distance between clusters, between data points, and in between both categories.
Clustering ↔ Tree topology
Tree notation: the Newick format

Each comma corresponds to an internal node, hence unrooted tree has one ternary node.
Top down & bottom up

- Bottom up: Start with each element in its own cluster. Join clusters until one is left.

- Top down: Start with a single cluster of all elements. Split one cluster at each step until no clusters can be split further.
Simple bottom-up hierarchical clustering

• Start with every taxa in its own cluster
• Join the two nearest clusters at each step
• Recompute distance matrix
• Repeat until only one cluster is left

Distance between 2 clusters, cluster and an element based on distance between elements:

- Max – link
- Min – link
- Avg – link
Distance between clusters

- For single elements, consider them to be in their own cluster, then compute:
  - max link
  - min link
  - avg link
Input / output of hierarchical clustering

• **Input**: a metric, the data points / heatmap

• **Output**: tree relation of the clusters, if possible branch lengths of tree (if further possible a root)
Lets build some trees ...
UPGMA

• Unweighted pair-group method with arithmetic mean

• In our language: bottom up hierarchical clustering with average-link distance between clusters

• Gives a clocked (hence rooted) tree! (how?)
UPGMA

– See a beautiful handtrace here:

http://www.southampton.ac.uk/~re1u06/teaching/upgma/

– However, UPGMA is clocked: lineage specific evolutionary rates (if 3 PC is violated) cant be handled
Neighbor joining

Make each taxa a tree of one species

- At each step identify 2 trees which are the most similar to each other AND further apart from the rest
- Make a new ancestral node and connect these 2 trees to it with different branch lengths
  - Accounts for lineage specific evolutionary rates: makes an unclocked tree: so rooting is arbitrary if performed
- Update distance matrix

Repeat until two trees are left, then connect them
Neighbor joining
Neighbor joining

• Notion of how different from the rest of the leaves one element is:

\[ u_i = \sum_j d_{ij} / (n-2) \]

\[ m_{ij} = d_{ij} - u_i - u_j \]

Pick i & j based on lowest \( m_{ij} \)
Neighbor joining

• Joining the two trees:

$$w^o_c = \frac{d_{ij}}{2} + \frac{1}{2} (u_i - u_j)$$

$$w^o_j = \frac{d_{ij}}{2} + \frac{1}{2} (u_j - u_i)$$

$$\Rightarrow \text{VIOLATES MOLECULAR CLOCK}$$
Neighbor joining

- If the $i$th and $j$th trees were joined, then the distance from the new ancestor to any other remaining tree $k$ is updated, and the original 2 trees are removed from the distance matrix:

$$d_{n,k} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}$$

- Originally, when each tree has a single element, the distances matrix are just distances between taxa.
Neighbor joining

• Another beautiful handtrace:
Bottom up hierarchical clustering

• Decisions:
  – how to identify neighbors and merge them together into a single tree/cluster
  – how to assign branch lengths
  – how to update the distance matrix once merging two clusters is performed
In practice ...

• Most distance matrices don’t follow 3PC or 4PC
  — Converting an arbitrary distance matrix to 3PC or 4PC is difficult and may change the nature of relationships unfaithfully
  — Forcing a hierarchical clustering on a non-conforming matrix leads to wrong branch lengths and/or topology
  • Squared error = sum of squares of difference between distance matrix and distance on tree for each taxon pair
Then, what?

• Pick any tree (!!!) — you may start with NJ tree
• Calculate squared error between the original distance matrix and distances on the tree
• Change the tree slightly: if the squared error decreases, keep the changes
  — Keep doing this step till no more improvement

How to change a tree “slightly”? (coming up next class)
Notion of a “better” tree

• Distance methods do not provide a rational notion of whether one tree is better than the other
  – Most distance methods produce one tree
  – Comparing trees across methods is difficult

• Minimum evolution: minimal sum of branch lengths is better
  – Premise may not be true for distant species
Figuring out the ancestral state

• How to figure out the ancestral state?
  – going from phenotype to distance is relatively easy (define a metric)
  – going from distance to phenotype is difficult (may be ambiguous)
To the rescue ...

- Character based methods

Figure: A Phylogenetic Tree

molgen.mpg.de
Acknowledgements

• Eric Xing
• Dannie Durand
• R Ravi