CS5263 Bioinformatics

Guest Lecture
Part II
Phylogenetics
Up to now we have focused on finding similarities, now we start focusing on differences (dissimilarities leading to distance measures). Identifying sequences has been the goal so far. Now we are to arrange the sequences according to their ancestry.

Phylogenetic trees are a graphical representation of the distance between sequences or species. Here we have the tree of the 3 major groups of living organisms (excluding viruses). Until recently, most phylogenies were based on rRNA sequences for prokaryotes and mitochondrial sequences for eukaryotes.
Terminology

- **Phylogeny**
  - The evolutionary relationships among organisms, based on a common ancestor

- **Phylogenetics**
  - Area of research concerned with finding the genetic relationships between species

(Greek: *phylon* = race and *genetic* = birth)
Terminology

- Phylogenetic tree: Visual representation of evolutionary distances between species
Introduction to terminology for phylogeny lecture

• Speciation
• Gene Duplication
• Homologous
  – Orthologs
  – Paralogs
**Allopatric speciation:** populations are separated by a barrier. After some time, even if the barrier is removed the two populations can no longer form hybrids (now different species).
Sympatric speciation: the population shares the environment, mate selection effectively separates gene pools.
Gene duplication diagram:
The three bands are duplicated.
Consider gene A is the ancestor species. Following duplication and modification, A1 and A2 variants of gene A was fixated in the ancestor. The ancestor species diverged into species X and Y. The two variants A1 and A2 evolves independently in the two lineages into A1X - A2x, and A1Z - A2Z in species X and Z, respectively.

Paralogous genes are derived from duplication, such as A1 and A2. Orthologous genes are derived from speciation, such as A1x - A1z, and A2x - A2z.

Genetic similarity among taxa should be estimated by comparing orthologous sequences. A phylogeny should be computed to determine which similar sequences are orthologs.
homologs

orthologs

paralogs

orthologs

frog α  chick α  mouse α  mouse β  chick β  frog β

α-chain gene

gene duplication

β-chain gene

to

early globin gene
Orthologues / Paralogues

Figure 7.1 Above: a tree of orthologues based on a set of alpha haemoglobins. Below: a tree of paralogues, the alpha, beta, gamma, delta, epsilon, zeta and theta chains of human haemoglobins, and human myoglobin. The orthologues are the alpha haemoglobins with SWISS-PROT identifiers HBA_ACCGE, HBA_AEGMO, HBA_AILFU, HBA_AILME, HBA_ALCAA, HBA_ALLMI, HBA_AMBME, and HBA_ANAPL, chosen because they were alphabetically the first eight alpha globins in PFAM [Sonnhammer, Eddy & Durbin 1997] (http://genome.wustl.edu/Pfam/). The paralogues are globins with SWISS-PROT identifiers HBAT_HUMAN, HBAZ_HUMAN, HBA_HUMAN, HBB_HUMAN, HBD_HUMAN, HBE_HUMAN, HBG_HUMAN, and MYG_HUMAN. The trees were made by neighbour-joining, Section 7.3, using J. Felsenstein's package PHYLIP (http://evolution.genetics.washington.edu/phylip.html). The distances used for neighbour-joining were the PAM-based ML distances (see p. 228) determined by the program PROTDIST in PHYLIP.
Definitions

• Classic phylogenetic analysis uses *morphological* features
  • Anatomy, size, number of legs, beak shape…

• Modern phylogenetic analysis uses *molecular* information
  • Genetic material (DNA and protein sequences)

   Molecular phylogenetic analysis
Phylogenetic reconstruction

• Goal: given a set of species (genes), reconstruct the tree which best explains their evolutionary history
Phylogenetic reconstruction

• “Nothing in evolution makes sense except in the light of phylogeny.” -- Joe Felsenstein
A brief history of molecular phylogeny

- phylogenetic inference is old (for Biology)

Charles Darwin – Origin of Species (1859)
Illustration of ‘descent with modification’

Ernst Haeckel “Tree of life” (1891)
Tracing evolutionary history

**phylogeny** – pattern and timing of evolutionary branching events ("evolutionary tree")

- Branching happened in the past
  - Common ancestors cannot be observed
  - Must infer from data

- Internal node - common ancestor (CA)
- External node - operational taxonomic unit (OTU)
- Order of branches define the relationships (topology)
- Branch length defines the number of changes

A common ancestor of A & B
A common ancestor of C & D
A common ancestor of A, B, C, D
Most reconstruction methods produce unrooted trees. To root a tree we need “external information” (e.g. outgroup).
What are phylogenies good for?
(1) classification

- Systematics: a scientific field devoted to classification of organisms
  - Phenetics: a classification scheme based on grouping populations according to similarities
  - **Cladistics**: a classification scheme based on evolutionary relationships (phylogenies)
Monophyletic vs paraphyletic

- **Monophyletic group**: including all descendents of a common ancestor
- **Paraphyletic group**: a set of species that includes a common ancestor and some, but not all, of its descendants.
Paraphyletic groups
What are phylogenies good for? 

(2) detecting coevolution

• Aphid-bacteria
• Mutualistic
• cospeciation
What are phylogenies good for?
(3) origin of pathogens

- Black plague
- Pathogen: *Yersinia pestis*
- 36 strains
What are phylogenies good for?

(4) Tree of life

Animal Kingdom
Rrooting the tree:

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

Note that in this rooted tree, taxon A is no more closely related to taxon B than it is to C or D.
**Number of OTUs (tips) vs. number of possible trees**

<table>
<thead>
<tr>
<th># OTUs (n)</th>
<th># unrooted trees</th>
<th># rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>954</td>
</tr>
<tr>
<td>7</td>
<td>954</td>
<td>10,395</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
<td>135,135</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
<td>2,027,025</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
<td>34,459,425</td>
</tr>
</tbody>
</table>

**true tree** - true evolutionary history is one of many possibilities. Difficult to infer true tree when # OTUs is large

**inferred tree** - obtained using data and reconstruction method. Not necessarily the same as the true tree - a hypothesis
Counting Trees

\[(2N - 5)!! = \# \text{ unrooted trees for } N \text{ taxa}\]
\[(2N - 3)!! = \# \text{ rooted trees for } N \text{ taxa}\]

<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>\cdot</td>
<td>\cdot</td>
</tr>
<tr>
<td>\cdot</td>
<td>\cdot</td>
</tr>
<tr>
<td>\cdot</td>
<td>\cdot</td>
</tr>
<tr>
<td>\cdot</td>
<td>\cdot</td>
</tr>
<tr>
<td>30</td>
<td>\approx3.58 \times 10^{36}</td>
</tr>
</tbody>
</table>

(2N - 5)!! = \# \text{ unrooted trees for } N \text{ taxa}
(2N - 3)!! = \# \text{ rooted trees for } N \text{ taxa}
Reconstruct phylogenetic trees

Input → Algorithm → Tree
- **Morphology Based Input:** n-by-m table, with rows = species, columns = properties.
- **Sequence Based Input:** n aligned sequences, one per species.
Methods of phylogenetic reconstruction

• **Distance based**
  – pairwise evolutionary distances computed for all taxa
  – tree constructed using algorithm based on relationships between distances

• **Maximum parsimony**
  – nucleotides or amino acids are considered as character states
  – best phylogeny is chosen as the one that minimizes the number of changes between character states

• **Maximum likelihood**
  – statistical method of phylogeny reconstruction
  – explicit model for how data set generated - nucleotide or amino acid substitution
  – find topology that maximizes the probability of the data given the model and the parameter values (estimated from data)
2. Determine the evolutionary distances and build distance matrix

- For molecular data, evolutionary distances can be the observed number of *nucleotide differences* between the pairs of species.
- **Distance matrix**: simply a table showing the evolutionary distances between all pairs of sequences in the dataset.
2. Determine the evolutionary distances and build distance matrix - A simple example using DNA sequences

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>0.2</td>
<td>0.05</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.25</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In this example the evolutionary distance is expressed as the number of nucleotide differences for each sequence pair. For example, sequences 1 and 2 are 20 nucleotides in length and have four differences, corresponding to an evolutionary difference of 4/20 = 0.2.
3. Phylogenetic Tree Construction example (UPGMA algorithm)

<table>
<thead>
<tr>
<th>(D_{ij})</th>
<th>Bear</th>
<th>Raccoon</th>
<th>Weasel</th>
<th>Seal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bear</td>
<td>-</td>
<td>0.26</td>
<td>0.34</td>
<td>0.29</td>
</tr>
<tr>
<td>Raccoon</td>
<td>-</td>
<td>0.42</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Weasel</td>
<td>-</td>
<td>0.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seal</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Pick smallest entry \(D_{ij}\)

2. Join the two intersecting species and assign branch lengths \(D_{ij}/2\) to each of the nodes
3. Phylogenetic Tree Construction example (UPGMA algorithm)

<table>
<thead>
<tr>
<th>$D_{ij}$</th>
<th>Bear</th>
<th>Raccoon</th>
<th>Weasel</th>
<th>Seal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bear</td>
<td>-</td>
<td>0.26</td>
<td>0.34</td>
<td>0.29</td>
</tr>
<tr>
<td>Raccoon</td>
<td>-</td>
<td>0.42</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Weasel</td>
<td>-</td>
<td>-</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Seal</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

3. Compute new distances to the other species using arithmetic means

$$D_{W(BR)} = \frac{D_{WB} + D_{WR}}{2} = \frac{0.34 + 0.42}{2} = 0.38$$

$$D_{S(BR)} = \frac{D_{SB} + D_{SR}}{2} = \frac{0.29 + 0.44}{2} = 0.365$$
3. Phylogenetic Tree Construction example (UPGMA algorithm)

<table>
<thead>
<tr>
<th>D_{ij}</th>
<th>BR</th>
<th>Weasel</th>
<th>Seal</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>-</td>
<td>0.38</td>
<td>0.365</td>
</tr>
<tr>
<td>Weasel</td>
<td>-</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Seal</td>
<td></td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

1. Pick smallest entry \( D_{ij} \)

2. Join the two intersecting species and assign branch lengths \( D_{ij}/2 \) to each of the nodes
3. Phylogenetic Tree Construction example (UPGMA algorithm)

<table>
<thead>
<tr>
<th>$D_{ij}$</th>
<th>BR</th>
<th>Weasel</th>
<th>Seal</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>-</td>
<td>0.38</td>
<td>0.365</td>
</tr>
<tr>
<td>Weasel</td>
<td>-</td>
<td>-</td>
<td>0.44</td>
</tr>
<tr>
<td>Seal</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

3. Compute new distances to the other species using arithmetic means

$$D_{W(BRS)} = \frac{D_{WB} + D_{WR} + D_{WS}}{3} = \frac{0.34 + 0.42 + 0.44}{3} = 0.4$$
3. Phylogenetic Tree Construction example (UPGMA algorithm)

<table>
<thead>
<tr>
<th>$D_{ij}$</th>
<th>BRS</th>
<th>Weasel</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRS</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>Weasel</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

1. Pick smallest entry $D_{ij}$.

2. Join the two intersecting species and assign branch lengths $D_{ij}/2$ to each of the nodes.

3. Done!
UPGMA clustering can be done using protein sequences

Calculation of a phylogeny from molecular comparisons.

Cytochrome c comparisons (from Fitch and Margoliash, Science Vol. 155, 20 Jan. 1967). The selected comparisons have been arranged randomly (no particular order), as this makes no difference in the application of UPGMA (unweighted pair-group method using arithmetic averages) clustering.

The numbers in the cells show differences between the cytochrome c molecules of various species: for example, there is only 1 difference in the amino acid sequences between man and monkey, but there are 19 differences between man and turtle.

### UPGMA CLUSTERING INPUT DATA

<table>
<thead>
<tr>
<th></th>
<th>Turtle A</th>
<th>Man B</th>
<th>Tuna C</th>
<th>Chicken D</th>
<th>Moth E</th>
<th>Monkey F</th>
<th>Dog G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turtle A</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man B</td>
<td>19</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuna C</td>
<td>27</td>
<td>12</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken D</td>
<td>8</td>
<td>18</td>
<td>10</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moth E</td>
<td>33</td>
<td>36</td>
<td>26</td>
<td>13</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monkey F</td>
<td>18</td>
<td>1</td>
<td>32</td>
<td>14</td>
<td>12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Dog G</td>
<td>13</td>
<td>13</td>
<td>29</td>
<td>14</td>
<td>28</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

The UPGMA method

• The UPGMA method is applied to the cytochrome c data sample. At each cycle of the method, the smallest entry is located, and the entries intersecting at that cell are "joined." The height of the branch for this junction is one-half the value of the smallest entry. Thus, since the smallest entry at the beginning is 1 (between B=man and F=monkey), B and F are joined with branch heights of 0.5 (=1.0/2).

Then, the comparison matrix is reduced by combining cells. These combinations are indicated with colors in the next slide. For example, the comparisons of A to B (19.0) and A to F (18.0) are consolidated as 18.5 = (19.0+18.0)/2 (red cells), while the comparisons of E to B (36.0) and E to F (35.0.0) are consolidated as 35.5 = (36.0+35.0)/2 (blue cells).

• The process is repeated on the reduced comparison matrix, resulting in a smaller matrix with each cycle. When the matrix is completely reduced, the calculation is finished.
APPLICATION OF UPGMA CLUSTERING METHOD ON SELECTED CYTOCHROME C DATA TO CALCULATE PHYLOGENETIC RELATIONS

Key: the boldface number on the left side indicates the smallest entry (closest match), and directs which entries are to be joined. The height of the new branch is 1/2 times this smallest value. The matrix is reduced as the entries are joined; cells of one color on the left are combined (averaged) to form the new entries (same color) on the right.
What makes such calculations of phylogenies interesting is the fact that the results so often agree with evolutionary trees developed from other methods (anatomy, fossils, or other proteins or genes). Indeed, molecular comparisons provide ample "repeat experiments" of the hypothesis of evolution.

The final phylogeny calculated from tables. It is in perfect accord with the fossil record, showing fish ancestral to reptiles, reptiles ancestral to mammals, birds splitting from reptiles after the reptile/mammal split, and so forth. The lengths of branches indicate time since last common ancestry; for example, moths and tuna (18.2 branch length) separated long before turtles and chickens (4.0 branch length).
Weakness of UPGMA

• UPGMA assumes a constant molecular clock (i.e. accumulate mutations at the same rate)
  – All leaves in the same level
• Only constructs rooted trees
Example: morphology-based input

Morphology Based Input & Distance Based Algorithm

• Input: 10-by-13 table (10 species, 13 properties).

• Algorithm - 2 stages:

  1. Build a simple distance matrix: Distance between a pair of species is the % of properties where they differ.

  2. Construct a tree by iteratively clustering species with small distances (“neighbors”).
<table>
<thead>
<tr>
<th>Taxon</th>
<th>amnion</th>
<th>legs</th>
<th>scales</th>
<th>blood</th>
<th>nostrils</th>
<th>septum</th>
<th>fenestrations</th>
<th>hemipenes</th>
<th>gizzard</th>
<th>teeth</th>
<th>feathers</th>
<th>wings</th>
<th>vertebrae</th>
<th>pedicillate</th>
</tr>
</thead>
<tbody>
<tr>
<td>perch</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>cold</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>coelocanth</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>cold</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>salamander</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>cold</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>frog</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>cold</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>turtle</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>cold</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>warm</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>gecko</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>cold</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>snake</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>cold</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>alligator</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>cold</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>budgy</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>warm</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

The preceding matrix can be represented numerically (for convenience) as:

```
   1 2 3 4 5 6 7 8 9 10 11 12 13
perch     0 0 0 0 0 0 0 0 0 0 0 0 0
coelocanth 0 0 0 1 1 0 0 0 0 0 0 0 0
salamander 0 1 1 0 0 1 1 0 0 1 0 0 0
frog       0 1 1 0 1 1 0 0 0 1 0 0 0
turtle     1 1 0 0 1 1 0 0 0 0 0 0 0
human      1 1 1 1 1 0 0 0 0 0 0 0 0
gecko      1 1 0 0 1 1 1 1 0 0 0 0 0
snake      1 0 0 0 1 1 1 1 0 0 0 0 0
alligator  1 1 0 0 1 1 1 0 1 0 0 0 0
budgy      1 1 1 1 1 1 1 0 1 0 1 1 0
```
Algorithm – Stage 1: construct the distance matrix

- **Distance** between any species is defined by the % of properties where they disagree (out of total number of properties)

<table>
<thead>
<tr>
<th></th>
<th>perch</th>
<th>coelocanth</th>
<th>salamander</th>
<th>frog</th>
<th>turtle</th>
<th>human</th>
<th>gecko</th>
<th>snake</th>
<th>alligator</th>
<th>budgy</th>
</tr>
</thead>
<tbody>
<tr>
<td>perch</td>
<td>85</td>
<td>62</td>
<td>62</td>
<td>69</td>
<td>54</td>
<td>54</td>
<td>62</td>
<td>54</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>coelocanth</td>
<td>77</td>
<td>77</td>
<td>85</td>
<td>69</td>
<td>69</td>
<td>77</td>
<td>69</td>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>salamander</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frog</td>
<td>77</td>
<td>77</td>
<td>54</td>
<td>46</td>
<td>62</td>
<td>62</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>turtle</td>
<td>85</td>
<td>85</td>
<td>76</td>
<td>85</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>69</td>
<td>62</td>
<td>69</td>
<td>69</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gecko</td>
<td>92</td>
<td>85</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>snake</td>
<td>76</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alligator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>budgy</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Similarity** **100**
Algorithm – Stage 2: cluster close neighbors

- Iteration #1: identify the two closest taxa from the distance matrix
- In our case, one pair has zero distance: salamander & frog
- We join them together, and update the distance matrix
- Updated matrix has only 9 species (8 “old”, 1 new).

The closest pair here are the gecko+snake (distance=8)
Algorithm – Stage 2: cluster close neighbors

- Iteration #2: join the gecko and snake.
- Add the new pair to the forest, equally distributing their distance (8).
- Update the distance matrix.
- One must calculate the similarity of each as-yet ungrouped taxa to the two groups already formed above.
- Also calculate the similarity of the two groups to each other.
Tie problem

<table>
<thead>
<tr>
<th></th>
<th>perch</th>
<th>coelocanth</th>
<th>turtle</th>
<th>human</th>
<th>alligator</th>
<th>budgy</th>
<th>SaFr</th>
<th>SnGe</th>
</tr>
</thead>
<tbody>
<tr>
<td>perch</td>
<td>85</td>
<td>69</td>
<td>54</td>
<td>54</td>
<td>23</td>
<td>75</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>coelocanth</td>
<td>85</td>
<td>69</td>
<td>69</td>
<td>38</td>
<td>75</td>
<td>69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>turtle</td>
<td>85</td>
<td>85</td>
<td>54</td>
<td>85</td>
<td>84</td>
<td>74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>69</td>
<td>69</td>
<td>69</td>
<td>85</td>
<td>75</td>
<td>84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alligator</td>
<td>69</td>
<td>75</td>
<td>84</td>
<td>64</td>
<td>64</td>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>budgy</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SaFr</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SnGe</td>
<td>65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The last entry in this table is calculated as the average similarity across two pre-existing groups (salamander:frog) and (snake:gecko). Therefore, it is the average of the 6 pairwise similarity values: \(\frac{100+92+54+54+46+46}{6} = 65\).
- The ties problem: two possible things we could do with the resulting similarities: either connect turtle to Salamander + Frog or connect Human to Salamander + Frog, or create two separate equally optimal solutions and continue for each.
If you break ties “systematically”, according to the order of appearance in the matrix, you will get the tree 1; if you break ties randomly, you make get the tree 2.
Distance method
(2) Neighbor-joining (NJ) method

uses ‘star decomposition’ – identification of neighbors that sequentially minimize the total length of the tree

1 - start with star tree - no topology
S = total branch length of tree

2 - separate pair of OTUs from all others
S_{12} = total branch length of tree

3 - choose pair of OTUs that minimizes total branch lengths in the tree

4 - this pair collapsed as single OTU and distance matrix recalculated

5 - next pair of OTUs that gives smallest branch length is chosen

6 - iterate until complete
The “1-star” Sum of the Branch Lengths

\[ S_0 = \sum_{i=1}^{N} L_{ix} = \frac{1}{N-1} \sum_{i<j}^{N} D_{ij} \]

\[ = \frac{T}{(N-1)} \]

- \( D \) and \( L \) as the distance between OTUs and the branch length between nodes
- Each branch is counted \( N-1 \) times when all distances are added
The “paired-2-star” Tree Size

\[ L_{XY} = \frac{1}{2(N - 2)} \left[ \sum_{k=3}^{N} (D_{1k} + D_{2k}) - (N - 2)(L_{1X} + L_{2X}) - 2 \sum_{i=3}^{N} L_{iY} \right] \]

\[ L_{1X} + L_{2X} = D_{12} \]

\[ \sum_{i=3}^{N} L_{iY} = \frac{1}{N - 3} \sum_{3 \leq i < j} D_{ij} \]

\[ S_{12} = L_{XY} + (L_{1X} + L_{2X}) + \sum_{i=3}^{N} L_{iY} \]

\[ = \frac{1}{2(N - 2)} \sum_{k=3}^{N} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N - 2} \sum_{3 \leq i < j} D_{ij} \]
Neighbor-joining example

A. Cycle 1: Neighbors = [1, 2]

<table>
<thead>
<tr>
<th>OTU</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>36.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>38.33</td>
<td>38.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>39.00</td>
<td>39.00</td>
<td>38.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>40.33</td>
<td>40.33</td>
<td>40.00</td>
<td>39.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>40.33</td>
<td>40.33</td>
<td>40.00</td>
<td>39.67</td>
<td>37.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>40.17</td>
<td>40.17</td>
<td>39.83</td>
<td>39.50</td>
<td>38.83</td>
<td>38.83</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>40.17</td>
<td>40.17</td>
<td>39.83</td>
<td>39.50</td>
<td>38.83</td>
<td>38.83</td>
<td>37.67</td>
</tr>
</tbody>
</table>

B. Cycle 2: Neighbors = [5, 6]

<table>
<thead>
<tr>
<th>OTU</th>
<th>1-2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>31.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>32.30</td>
<td>32.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>33.90</td>
<td>33.90</td>
<td>33.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>33.90</td>
<td>33.90</td>
<td>33.70</td>
<td>31.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>33.70</td>
<td>33.70</td>
<td>33.50</td>
<td>33.10</td>
<td>33.10</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>33.70</td>
<td>33.70</td>
<td>33.50</td>
<td>33.10</td>
<td>33.10</td>
<td>31.90</td>
</tr>
</tbody>
</table>
Neighbor-Joining: Complexity

• The method performs a search using time $O(n^2)$ and using time $O(n^2)$ to update distance matrix.
• Giving a total time complexity of $O(n^3)$, and a space complexity of $O(n^2)$. 
Neighbor-joining method …

- Extremely fast and efficient method, widely used
- Tends to perform fairly well in simulation studies
- May produce tie trees from data set but this appears to be rare
- Algorithm is ‘greedy’ and so can get stuck in local optima
- Main criticism is that it produces only one tree and does not give any idea of how many other trees are equally well or almost as supported by the data
Maximum Parsimony (MP)

What is parsimony?

• A criterion for selecting among alternative patterns based on minimizing the total amount of evolutionary change

• Ancestral characters are inferred for each site and the total number of changes between nodes for a given topology are determined

• Best topology is the one that requires the fewest number of residue changes between nodes across all sites
Counting substitutions on a tree

• For an alignment site and a topology, **ancestral residues** are inferred so that the minimum number of residue changes between nodes is required.

Unambiguous (1)

Ambiguous (2)
Choosing the shortest tree with parsimony

<table>
<thead>
<tr>
<th>Sites</th>
<th>3</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree I</td>
<td>4 steps</td>
<td>best tree</td>
<td></td>
</tr>
<tr>
<td>1A 3C 2A 4C</td>
<td>1T 3T 2A 4A</td>
<td>1T 3G 2T 4G</td>
<td></td>
</tr>
<tr>
<td>Tree II</td>
<td>5 steps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A 2A 3C 4C</td>
<td>1T 2A 3T 4A</td>
<td>1T 2T 3G 4G</td>
<td></td>
</tr>
<tr>
<td>Tree III</td>
<td>6 steps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A 2A 3C 4C</td>
<td>1T 2A 3T 4A</td>
<td>1T 2T 3G 4G</td>
<td></td>
</tr>
</tbody>
</table>

OTU  1 2 3 4 5 6 7 8 9 10 Site
1 T C A G A T C T A G
2 T T A G A A C T A G
3 T T C G A T C G A G
4 T T C T A A G G A C

Lecture #7
Advantages and disadvantages of parsimony

- Advantages:
  - based on a logically coherent and biologically plausible model of evolution
  - free from assumptions used in distance estimations
  - better than distance methods when extent of sequence divergence is low (10%), rate of substitution is constant, number of residues is large
  - very useful for certain types of molecular data e.g. indels

- Disadvantages:
  - gives incorrect topologies when backward substitutions are present (common with nucleotides) and when the number of sites is fairly small
  - gives incorrect topologies when rate of substitution varies substantially across lineages
  - **long branch attraction** – long branches (and short branches) tend to group together on reconstructed tree
  - difficult to treat the results in a statistical framework
Maximum Likelihood

• Statistical (probabilistic) method for inferring phylogenies
  – substitution model is chosen for sequence data (alignment)
  – likelihood of observing the sequence data given the substitution model is obtained for each topology evaluated (parameter fitting on branch lengths)
    • Probability of each tree is product of mutation rates in each branch
    • Likelihoods given by each column multiplied to give the likelihood of the tree
  – topology that gives the highest likelihood is chosen as the best tree
Maximum Likelihood

- Extremely slow method so heuristic methods almost always have to be employed to search for best tree
- Method very dependent on model of substitution used
- Method estimates branch lengths not topology, so may give wrong topology
Assessing significance of Tree

- need some way to assess the support for the topologies (evolutionary relationships) of reconstructed phylogenies

- **bootstrapping**: re-sample alignment and construct trees from re-sampled data
Bootstrap test
(Felsenstein 1985)

- assess the support for individual interior branches
- re-sample alignment columns with replacement
- testing the signal : noise ratio in the data (homoplasies)
- repeat many times (100 - 1000) and get consensus tree

<table>
<thead>
<tr>
<th>OTU</th>
<th>Site</th>
<th>OTU</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T C A G A T C T A G</td>
<td>1</td>
<td>C A T A G C A T T A</td>
</tr>
<tr>
<td>2</td>
<td>T T A G A A C T A G</td>
<td>2</td>
<td>T A A A G C A A T A</td>
</tr>
<tr>
<td>3</td>
<td>T T C G A T C G A G</td>
<td>3</td>
<td>T C T A G C A T T A</td>
</tr>
<tr>
<td>4</td>
<td>T T C T A A G G A C</td>
<td>4</td>
<td>T C A A C G A A T A</td>
</tr>
</tbody>
</table>

The original alignment is transformed into a sampled alignment by re-sampling columns with replacement and assessing the support for individual interior branches.

![Alignment Diagram]

The alignment tree is shown side by side with the alignment, illustrating the support for each branch.
Interpreting bootstrap results

- 71 = the percentage of trees built from re-sampled alignments that included the interior branch in question
- bootstrap values are said to “support”
  - that interior branch
  - the interior nodes adjacent (terminal) to the branch

- Rule of the thumb: >70% considered good evidence for proper placement of branches
- <50%, uncertain (unresolved) branching pattern (polytomy)
## Comparison of Methods

<table>
<thead>
<tr>
<th>Neighbor-joining</th>
<th>Maximum parsimony</th>
<th>Maximum likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uses only pairwise distances</td>
<td>Uses only shared derived characters</td>
<td>Uses all data</td>
</tr>
<tr>
<td>Minimizes distance between nearest neighbors</td>
<td>Minimizes total distance</td>
<td>Maximizes tree likelihood given specific parameter values</td>
</tr>
<tr>
<td>Very fast</td>
<td>Slow</td>
<td><strong>Very</strong> slow</td>
</tr>
<tr>
<td>Easily trapped in local optima</td>
<td>Assumptions fail when evolution is rapid (Long branch attraction)</td>
<td>Highly dependent on assumed evolution model</td>
</tr>
<tr>
<td>Good for generating tentative tree, or choosing among multiple trees, or working on large-scale data sets</td>
<td>Best option when tractable (&lt;30 taxa)</td>
<td>Good for very small data sets and for testing trees built using other methods</td>
</tr>
</tbody>
</table>
Which Method to Choose?

• depends upon the sequences that are being compared
  – strong sequence similarity:
    • maximum parsimony
  – clearly recognizable sequence similarity
    • distance methods
  – All others:
    • maximum likelihood

• Best to choose at least two approaches

• Compare the results – if they are similar, you can have more confidence
Which programs to use?

• Distance method:
  – MEGA

• Maximum Parsimony method
  – PAUP
  – MacClade

• Maximum Likelihood method
  – PHYLIP
  – PAML
MEGA 4: Molecular Evolutionary Genetics Analysis


MEGA is an integrated tool for conducting automatic and manual sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, and testing evolutionary hypotheses.

MEGA 4 has been tested on the following Microsoft Windows® operating systems:

- Windows 95/98, NT, 2000, XP, and Vista.
Phylogenetics and forensic evidence

Victim & patient strains more closely related to each other than controls (monophyletic)

Victims’ HIV sequences were a subset of the doctor’s patient’s sequences

Doctor guilty of attempted murder

Louisiana doctor accused of injecting victim with HIV

Baylor grad student compares sequences of victim’s HIV & Doctor’s patient HIV & local control strains


HIV strain analysis debuts in murder trial.

Vogel G.
Phylogenetics and forensic evidence ...

Molecular epidemiology of HIV transmission in a dental practice.


The Florida dentist

- In 1986 dentist tested positive for HIV; later AIDS.
- Continued general dentistry for 2 years.
- Patient with no risk factors discovered she had HIV
- Dentist publicly urges other patients to take test
- 10 of 1100 patients are HIV+

Did the dentist pass HIV to his patients?
Phylogenetics and forensic evidence

Forensics—Transmission of HIV by Florida dentist

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

- DENTIST
- Patient C
- Patient A
- Patient G
- Patient B
- Patient E
- Patient A
- DENTIST

Local control 2
Local control 3
Patient F
Local control 9
Local control 35
Local control 3
Patient D

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

No

From Ou et al. (1992) and Page & Holmes (1998), redrawn by Caro-Beth Stewart
Bayesian and GA software

- BEAST (Bayesian Evolutionary Analysis Sampling Trees): bayesian, MCMC
- MrBayes: bayesian, MCMC and MCMCMC
- Phycas: bayesian, for DNA seqs, python
- GARLI (Genetic Algorithm for Rapid Likelihood Inference): uses a stochastic genetic algorithm-like approach, Computational analogue of evolution by natural selection, not actually genetic algorithm
Software to evaluate trees

• Readseq is a program that edits sequences into 18 different formats
• AWTY (are we there yet?) is used to calculate whether MCMC has run long enough
• Tracer is similarly used to analyze MCMC based program runs
• FigTree is used to edit trees for publication

• And so much more
Probabilistic Methods

- The phylogenetic tree represents a generative probabilistic model (like HMMs) for the observed sequences.
- Background probabilities: $q(a)$
- Mutation probabilities: $P(a|b, t)$
- Models for evolutionary mutations
  - Jukes Cantor
  - Kimura 2-parameter model
- Such models are used to derive the probabilities
Jukes Cantor model

• A model for mutation rates

• Mutation occurs at a constant rate
• Each nucleotide is equally likely to mutate into any other nucleotide with rate $\alpha$. 
Kimura 2-parameter model

- Allows a different rate for transitions and transversions.
Optimal Tree Search

- Perform search over possible topologies

$T_1$ $T_2$ $T_3$ $T_4$ $\ldots$ $T_n$

Parameter space

Parametric optimization (EM)

Local Maxima
Computational Problem

- Such procedures are computationally expensive!
- Computation of optimal parameters, per candidate, requires non-trivial optimization step.
- Spend non-negligible computation on a candidate, even if it is a low scoring one.
- In practice, such learning procedures can only consider small sets of candidate structures
Current status of phylogenetic analysis

- Bayesian approaches widely implemented
- Maximum likelihood remains gold-standard
- Novel genetic algorithms also currently implemented, but not yet widely tested
- Large datasets still very computationally expensive
- Very few reiterative methods where phylogeny directs alignments and vice-versa
- Still difficult for biologists to evaluate results of different algorithms
Useful links

• IUPAC codes
  http://www.bioinformatics.org/sms/iupac.html

• Molecular Evolution Course website
  http://www.molecularevolution.org/

• Tree of Life web project
  http://tolweb.org/tree/


• Introduction to evolution
  http://evolution.berkeley.edu/
References

UPGMA protein example:
http://www.nmsr.org/upgma.htm

• Joe Felsenstein, *Phylogeny methods*,