Theory of Evolution

• Charles Darwin
  - 1858-59: *Origin of Species*
  - 5 year voyage of H.M.S. Beagle (1831-36)
  - Populations have variations.
  - Natural Selection & Survival of the fittest: *nature selects best adapted varieties to survive and to reproduce.*
  - Speciation arises by splitting of one population into subpopulations.
  - Gregor Mendel and his work (1856-63) on inheritance.
Dominant View of Evolution

• All existing organisms are derived from a common ancestor and that new species arise by splitting of a population into subpopulations that do not cross-breed.

• Organization: **Directed Rooted Tree**; Existing species: **Leaves**; Common ancestor species (divergence event): **Internal node**; Length of an edge: **Time**.
Phylogeny
Constructing Evolutionary/Phylogenetic Trees

• 2 broad categories:
  - **Distance-based methods**
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - **Character-based**
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods
Ultrametric

- An ultrametric tree:
  - decreasing internal node labels
  - distance between two nodes is label of least common ancestor.

- An ultrametric distance matrix:
  - Symmetric matrix such that for every i, j, k, there is tie for maximum of \( D(i,j) \), \( D(j,k) \), \( D(i,k) \)
Ultrametric: Assumptions

- **Molecular Clock Hypothesis, Zuckerkandl & Pauling, 1962**: Accepted point mutations in amino acid sequence of a protein occurs at a **constant rate**.
  - Varies from protein to protein
  - Varies from one part of a protein to another
Ultrametric Data Sources

• Lab-based methods: *hybridization*
  - Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.

• Sequence-based methods: *distance*
## Ultrametric: Example

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

**Diagram:**
- Node A is at a distance of 5 from the root.
- Node C is at a distance of 3 from the root.
- Nodes B, D, F, H are at a distance of 4 from the root.
- Nodes E, B, D, F, H are at a distance of 3 from the root.

**Key:**
- A, C, G are leaf nodes.
- E is an internal node.
- B, D, F, H are an internal node with distance 4 from the root.
- A is a leaf node with distance 5 from the root.
Ultrametric: Example

A  B  C  D  E  F  G  H
A  0  4  3  4  5  4  3  4
B  0  4  2  5  1  4  4  4
C  D  E  F  G
H

Bioinformatics (Lec 20)  3/31/05
### Ultrametric: Distances Computed

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The tree diagram on the right corresponds to the distances computed in the table. The distances between the points are shown on the branches of the tree.
Additive-Distance Trees

Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.
Unrooted Trees on 4 Taxa
Four-Point Condition

- If the true tree is as shown below, then
  1. $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, and
  2. $d_{AB} + d_{CD} < d_{AD} + d_{BC}$
Unweighted pair-group method with arithmetic means (UPGMA)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td>$d_{AB}$</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>$d_{AC}$</td>
<td></td>
<td>$d_{BC}$</td>
</tr>
<tr>
<td>D</td>
<td>$d_{AD}$</td>
<td>$d_{BD}$</td>
<td>$d_{CD}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AB</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td></td>
<td>$d_{(AB)C}$</td>
</tr>
<tr>
<td>D</td>
<td>$d_{(AB)D}$</td>
<td></td>
</tr>
</tbody>
</table>

\[
d_{(AB)C} = \frac{(d_{AC} + d_{BC})}{2}
\]
Transformed Distance Method

• UPGMA makes errors when rate constancy among lineages does not hold.
• Remedy: introduce an outgroup & make corrections

\[
D_{ij}' = \frac{D_{ij} - D_{iO} - D_{jO}}{2} + \frac{1}{n} \left( \sum_{k=1}^{n} D_{kO} \right)
\]

• Now apply UPGMA
Saitou & Nei: Neighbor-Joining Method

- Start with a **star topology**.
- Find the pair to separate such that the total length of the tree is minimized. The pair is then replaced by its arithmetic mean, and the process is repeated.

\[
S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \leq i \leq j \leq n} D_{ij}
\]
Neighbor-Joining

\[ S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \leq i \leq j \leq n} D_{ij} \]
Constructing Evolutionary/Phylogenetic Trees

- **2 broad categories:**
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - **Character-based**
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods
Character-based Methods

- **Input:** characters, morphological features, sequences, etc.
- **Output:** phylogenetic tree that provides the history of what features changed. [*Perfect Phylogeny Problem*]
- one leaf/object, 1 edge per character, path $\Leftrightarrow$ changed traits

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Example

- **Perfect phylogeny does not always exist.**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Maximum Parsimony

• Minimize the total number of mutations implied by the evolutionary history
Examples of Character Data

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Characters/Sites</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
</tbody>
</table>
### Maximum Parsimony Method: Example

<table>
<thead>
<tr>
<th>Sequence</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
</tbody>
</table>
Unrooted Trees on 4 Taxa

A
  B

A
  B

C
  D

A
  B

C
  D

B

D
FIGURE 5.14  Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number and the nucleotide type at homologous positions in the extant species. Each dot on a branch means a substitution is inferred on that branch. Note that the nucleotides at the two internal nodes of each tree represent one possible reconstruction from among several alternatives. For example, the nucleotides at both the internal nodes of tree III(d) (bottom right) can be A instead of T. In this case, the two substitutions will be positioned on the branches leading to species 2 and 4. Alternatively, other combinations of nucleotides can be placed at the internal nodes. However, these alternatives will require three substitutions or more. The minimum number of substitutions required for site 9 is two.
Inferring nucleotides on internal nodes

Figure 5.15  Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an internal node is dependent on the tree. Modified from Fitch (1971).
Searching for the Maximum Parsimony Tree: Exhaustive Search

Figure 5.16 Exhaustive stepwise construction of all 15 possible trees for five OTUs. In step 1, we form the only possible unrooted tree for the first three OTUs (A, B, and C). In step 2, we add OTU D to each of the three branches of the tree in step 1, thereby generating three unrooted trees for four OTUs. In step 3, we add OTU E to each of the five branches of the three trees in step 2, thereby generating 15 unrooted trees. Additions of OTUs are shown as heavier lines. Modified from Swofford et al. (1996).
Searching for the Maximum Parsimony Tree: Branch-&-Bound
Probabilistic Models of Evolution

- Assuming a model of substitution,
  - \( \Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\} \),
- Using this formula it is possible to compute the likelihood that data \( D \) is generated by a given phylogenetic tree \( T \) under a model of substitution. Now find the tree with the maximum likelihood.

- Time elapsed? \( \Delta \)
- Prob of change along edge? \( \Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\} \)
- Prob of data? Product of prob for all edges
Computing Maximum Likelihood Tree

**Figure 5.19** Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length \( n \). (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood of a particular site, in this case site 5, equals the sums of the 16 probabilities of every possible reconstruction of ancestral states at nodes 5 and 6 in (b). (d) The likelihood of the tree in (b) is the product of the individual likelihoods for all \( n \) sites. (e) The likelihood is usually evaluated by summing the logarithms of the likelihoods at each site, and reported as the log likelihood of the tree. Modified from Swofford et al. (1996).

\[
L(5) = \text{Prob}(C\rightarrow A\rightarrow A) + \text{Prob}(C\rightarrow A\rightarrow C) + \text{Prob}(C\rightarrow A\rightarrow T) + \text{Prob}(C\rightarrow A\rightarrow G)
\]

\[
+ \text{Prob}(C\rightarrow C\rightarrow A) + \text{Prob}(C\rightarrow C\rightarrow C) + \text{Prob}(C\rightarrow C\rightarrow T) + \text{Prob}(C\rightarrow C\rightarrow G)
\]

\[
+ \text{Prob}(C\rightarrow T\rightarrow A) + \text{Prob}(C\rightarrow T\rightarrow C) + \text{Prob}(C\rightarrow T\rightarrow T) + \text{Prob}(C\rightarrow T\rightarrow G)
\]

\[
+ \text{Prob}(C\rightarrow G\rightarrow A) + \text{Prob}(C\rightarrow G\rightarrow C) + \text{Prob}(C\rightarrow G\rightarrow T) + \text{Prob}(C\rightarrow G\rightarrow G)
\]

\[
L = L(1) \times L(2) \times L(3) \times \cdots \times L(n) = \prod_{i=1}^{n} L(i)
\]

\[
\ln L = \ln L(1) + \ln L(2) + \ln L(3) + \cdots + L(n) = \sum_{i=1}^{n} \ln L(i)
\]