CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu
https://users.cs.fiu.edu/~giri/teach/BioinfF18.html
Evolution and Phylogeny
Charles Darwin’s 1859 book (On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life) introduced the Theory of Evolution.

Struggle for existence induces a natural selection. Offspring are dissimilar from their parents (that is, variability exists), and individuals that are more fit for a given environment are selected for. In this way, over long periods of time, species evolve. Groups of organisms change over time so that descendants differ structurally and functionally from their ancestors.
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**.
Five kingdom system (Haeckel, 1879)

Slide by Pevsner

mammals
vertebrates
invertebrates
protozoa

animals
plants
fungi
protists
monera
At the molecular level, evolution is a process of mutation with selection.

Molecular evolution is the study of changes in genes and proteins throughout different branches of the tree of life.

Phylogeny is the inference of evolutionary relationships. Traditionally, phylogeny relied on the comparison of morphological features between organisms. Today, molecular sequence data are also used for phylogenetic analyses.
Questions for Phylogenetic Analysis

- How many genes are related to my favorite gene?
- How related are whales, dolphins & porpoises to cows?
- Where and when did HIV or other viruses originate?
- What is the history of life on earth?
- Was the extinct quagga more like a zebra or a horse?

Slide by Pevsner
Molecular phylogeny uses trees to depict evolutionary relationships among organisms. These trees are based upon DNA and protein sequence data.
Tree Roots

- The root of a phylogenetic tree represents the common ancestor of the sequences. Some trees are unrooted, and thus do not specify the common ancestor.

- A tree can be rooted using an outgroup (that is, a taxon known to be distantly related from all other OTUs).
Tree nomenclature: roots

Rooted tree (specifies evolutionary path)

Unrooted tree
Numbers of rooted and unrooted trees: 3 OTUs

For three operational taxonomic units (OTUs) there is one possible unrooted tree.

Any of the three edges can be selected to form a root.

Three rooted trees are possible.
Numbers of rooted and unrooted trees: 4 OTUs

For 4 OTUs there are three possible unrooted trees.

For 4 OTUs there are 15 possible rooted trees.

There is only one of these 15 trees that accurately describes the evolutionary process by which these four sequences evolved.
Tree Nomenclature

Fig. 7.8

Slide by Pevsner
Tree nomenclature

operational taxonomic unit (OTU) such as a protein sequence

taxon

time

one unit
Tree nomenclature

Node (intersection or terminating point of two or more branches)

branch (edge)
Tree nomenclature

Branches are unscaled...

Branches are scaled...

...OTUs are neatly aligned, and nodes reflect time

...branch lengths are proportional to number of amino acid changes

Fig. 7.8

Slide by Pevsner
Tree nomenclature

bifurcating internal node

multifurcating internal node

Slide by Pevsner
Examples of multifurcation: failure to resolve the branching order of some metazoans and protostomes

Rokas A. et al., Animal Evolution and the Molecular Signature of Radiations Compressed in Time, Science 310:1933 (2005), Fig. 1.
Clade ABF (monophyletic group)

Clade

group of organisms believed to have evolved from a common ancestor

Monophyletic

group of organisms that consists of all the descendants of a common ancestor
Tree nomenclature: clades
Tree nomenclature

Clade CDH
Tree nomenclature

Clade ABF/CDH/G

Fig. 7.8
Page 232
Examples of clades

Lindblad-Toh et al., *Nature* 438: 803 (2005), fig. 10
Tree nomenclature: roots

Rooted tree
(specifies evolutionary path)

Unrooted tree

Fig. 7.10
Page 234
Tree nomenclature: outgroup rooting

Rooted tree

Outgroup (used to place the root)
Molecular Clock Hypothesis

- The molecular clock is a figurative term for a technique that uses the **mutation rate** of biomolecules to deduce the time in prehistory when two or more life forms diverged.
- In the 1960s, sequence data were accumulated for small, abundant proteins such as globins, cytochromes c, and fibrinopeptides.
- Some proteins appeared to evolve slowly, while others evolved rapidly.
- Linus Pauling, Emanuel Margoliash and others proposed the hypothesis of a molecular clock:
  - For every given protein, the rate of molecular evolution is approximately constant in all evolutionary lineages.
Molecular Clock Hypothesis

![Graph showing corrected amino acid changes per 100 residues (n) against millions of years since divergence. The y-axis represents amino acid changes, and the x-axis represents millions of years. Different species and their divergence times are indicated, such as birds/reptiles (240 MY), mammals/reptiles (300 MY), higher vertebrates/fish (400 MY), carp/limpet (600 MY), vertebrates/insects (600 MY), and plants/animals (1200 MY).]
Molecular Clock Hypothesis — Implications

- If protein sequences evolve at constant rates, they can be used to estimate the times that sequences diverged. This is analogous to dating geological specimens by radioactive decay.
Molecular evolutionary studies can be complicated by the fact that both species and genes evolve. Speciation usually occurs when a species becomes reproductively isolated. In a species tree, each internal node represents a speciation event.

Genes (and proteins) may duplicate or otherwise evolve before or after any given speciation event. The topology of a gene (or protein) based tree may differ from the topology of a species tree.
Species trees versus gene/protein trees

![Diagram showing the relationship between species trees and gene/protein trees. The diagram illustrates the past and present relationships between species 1 (e.g., human) and species 2 (e.g., rat), highlighting a speciation event.](https://example.com/diagram.png)
A gene (e.g. a globin) may duplicate before or after two species diverge!
Stage 1: Use of DNA, RNA, or protein

For phylogeny, DNA can be more informative.

Some substitutions in a DNA sequence alignment can be directly observed: single nucleotide substitutions, sequential substitutions, coincidental substitutions. Additional mutational events can be inferred by analysis of ancestral sequences.
Two sequences (human and mouse) and their common ancestor: we can infer which DNA changes occurred over time.
Two sequences (human and mouse) and their common ancestor: we can infer which DNA changes occurred over time.

- Parallel substitutions
- Single substitution
- Sequential substitution
- Coincidental substitutions
- Convergent substitutions
- Back substitution
Step matrices: number of steps required to change a character

For amino acids, between 1 and 3 nucleotide changes are required to change one residue to another.
The fundamental basis of a phylogenetic tree is a multiple sequence alignment.

(If there is a misalignment, or if a nonhomologous sequence is included in the alignment, it will still be possible to generate a tree.)

Consider the following alignment of 13 homologous globin proteins (see Fig. 3.2)
Multiple alignment of myglobins, alpha globins, beta globins

myoglobin_kanga
---

myoglobin_harbo
---

myoglobin_gray
---

alpha_globin_ho
---

alpha_globin_ka
---

alpha_globin_do
---

beta_globin_dog
---

beta_globin_rab
---

beta_globin_kan
---

globin_riveram
---

globin_sealamp
---

globin_soysign
---

globin_insect
---

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---

myoglobin_kanga

myoglobin_harbo

myoglobin_gray

alpha_globin_ho

alpha_globin_ka

alpha_globin_do

beta_globin_dog

beta_globin_rab

beta_globin_kan

globin_riveram

globin_sealamp

globin_soysign

globin_insect

---
Open circles: positions that distinguish myoglobins, alpha globins, beta globins

<table>
<thead>
<tr>
<th></th>
<th>myoglobin_kanga</th>
<th>myoglobin_harbo</th>
<th>myoglobin_gray_</th>
<th>alpha_globin_ho</th>
<th>alpha_globin_ka</th>
<th>alpha_globin_do</th>
<th>beta_globin_dog</th>
<th>beta_globin_rab</th>
<th>beta_globin_kan</th>
<th>globin_riverlam</th>
<th>globin_sealamp</th>
<th>globin_soybean</th>
<th>globin_insect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MGLSDGEWQLVLNIGKVTDEGGGHGKDVLIRLFKGHPETLEKFDFK</td>
<td>MGLSegEwqlvlnvWgkveadlaghqdvlIRLFKgHPETLEKFDFK</td>
<td>MGLSDGEWHLVLNVWgKvetdlagHqevlIRLFKgHPETLEKFDFK</td>
<td>MV-LSAADKTNVKAASKGESHGAGEBVGAEALERMFGLGFPTTKYFPHF</td>
<td>V-LSAADKGHVKAIWGKVGGHAGEYAAEGERTFSFPTTKYFPHF</td>
<td>V-LSADKTNKSTWDK1GGHADYGGEALDRTQFSFPTTKYFPHF</td>
<td>MVHLTAEKSLVSGLWGKV--NVEVGGEALGRLLIVYPWTQRFDFSF</td>
<td>MVHLSSAEKSVLALWGKV--NVEEVGGEALGRLLIVYPWTQRFDFSF</td>
<td>VHLTAEEKNAITSLWGKV--AEQTGGEALGRLLIVYPWTSRFDFSF</td>
<td>-PIVDS--GSPAVLSAEEKTKIRSAWAPVYSYETSVDLVKFFTSTPAAQEBFPFK</td>
<td>MPIVDT--GSVAPLSAEEKTIKIRSAWAPVYSTYETSVDLVKFFTSTPAAQEBFPFK</td>
<td>VAFTKQDALVSFEAKIPQYSVVYFTISLEKAPAADKDLFSFL</td>
<td>MKFLILALCFAASALSADQISTVQASFDKVKG...--PVGILYAVFKADSIPMAKFTQF</td>
</tr>
</tbody>
</table>

\[ ▼ \text{gaps} \]

\[ □ 100\% \text{ conserved} \]
Stage 2: Multiple sequence alignment

[1] Confirm that all sequences are homologous

[2] Adjust gap creation and extension penalties as needed to optimize the alignment

[3] Restrict phylogenetic analysis to regions of the multiple sequence alignment for which data are available for all taxa (delete columns having incomplete data).
2 broad categories:

- **Distance-based methods**
  - Ultrametric
  - Additive:
    - UPGMA
    - Transformed Distance
    - Neighbor-Joining

- **Character-based**
  - Maximum Parsimony
  - Maximum Likelihood
  - Bayesian Methods
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 $D(i,j)$, $D(j,k)$, $D(i,k)$
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
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>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>

### Diagram

- **Root**: A
- **Nodes**: B, D, F, H
- **Leaves**: C, G
## 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>
<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></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 diagram on the right visualizes the ultrametric relationships:

- **5**
  - **4**
    - **3**
      - **2**
        - **1**
          - **A**
          - **C,G**
          - **B**
          - **F**
    - **E**
    - **H**

The ultrametric relationship is indicated by the distances in the table and the hierarchical structure in the diagram.
# 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>2</td>
<td></td>
<td></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>

- **Diagram:**
  - The distances between the elements are represented as edges in a network.
  - Elements A, C, G, B, and F are connected in a way that reflects their distances.
  - The distances are indicated by the numbers on the edges.

- **Legend:**
  - A, C, G, B, F are connected in a network that represents the distances computed.
  - The network is a tree structure indicating the ultrametric property, where each node is connected to another node with a distance that is greater than or equal to the sum of the distances to its children.

- **Notes:**
  - The distances computed are consistent with the ultrametric property, where the distance between any two elements is less than or equal to the sum of the distances to any other element.
### 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>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Tree Representation

```
      5
     /  \
    4    3
   /  \
  2   2
 / \  /\
1  1 1 2
```

- A
- C
- G
- B
- F
- D
- H
- E

4/5/15 CAP5510 / CGS5166
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
Lab-based methods: **hybridization**
- Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.

Sequence-based methods: **distance**
Additive-Distance Trees

Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.

<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>7</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

The table above shows the edge weights between nodes A, B, C, and D.
Unrooted Trees on 4 Taxa
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)

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

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

\[
D_{ij}' = \frac{D_{ij} - D_{i0} - D_{j0}}{2} + \left( \frac{\sum_{k=1}^{n} D_{k0}}{n} \right)
\]

- Now apply UPGMA
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

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

http://en.wikipedia.org/wiki/Neighbor_joining
Neighbor-joining method

[Diagram of evolutionary relationships showing nodes and branches labeled with letters from A to E, with arrows indicating the sequence of branch additions.]
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 ⇔ changed traits

### 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>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>

### Diagram

```
A -- 3 -- 2
   | 4
   v
 D <--- B
   |
   v
 E<-- A
   |
   v
 C
```

4/5/15  CAP5510 / CGS5166  59
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>

Characters/Sites

<table>
<thead>
<tr>
<th></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><strong>Sequences</strong></td>
<td>1</td>
<td></td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td></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>
</tr>
<tr>
<td><strong>3</strong></td>
<td></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>
</tr>
<tr>
<td><strong>4</strong></td>
<td></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>
</tr>
</tbody>
</table>

4/5/15

CAP5510 / CGS5166
### Maximum Parsimony Method: Example

<table>
<thead>
<tr>
<th>Sequences</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
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
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
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).

\[
L(5) = \text{Prob}\left(\begin{array}{c}
\text{C} \\
\text{C}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{A} - \text{A} \\
\text{C}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{C} - \text{A} \\
\text{G}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{A} - \text{C} \\
\text{G}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{C} - \text{C} \\
\text{C}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{C} - \text{C} \\
\text{C}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{C} - \text{T} \\
\text{C}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{T} - \text{C} \\
\text{C}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{G} - \text{A} \\
\text{G}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{G} - \text{C} \\
\text{G}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{T} - \text{G} \\
\text{C}
\end{array}\right) + \text{Prob}\left(\begin{array}{c}
\text{G} - \text{G} \\
\text{G}
\end{array}\right)
\]

(d) \( L = L(1) \times L(2) \times L(3) \times \ldots \times L(n) = \prod_{i=1}^{n} L(i) \)

(e) \( \ln L = \ln L(1) + \ln L(2) + \ln L(3) + \ldots + L(n) = \sum_{i=1}^{n} \ln L(i) \)