Phylogeny
and
Molecular Evolution

Distance Based Phylogeny
Credit

- Serafim Batzoglou (UPGMA slides)
  http://www.stanford.edu/class/cs262/Slides
- Notes by Nir Friedman, Dan Geiger, Shlomo Moran, Ron Shamir, Sagi Snir, Michal Ziv-Ukelson
- Durbin et al.
- Jones and Pevzner’s lecture notes
- Bioinformatics Algorithms book by Phillip Compeau and Pavel Pevzner – all book photos shown in this lecture are from there.
Phylogenetic Trees
Phylogenetic Trees are Unordered
Phylogenetic Trees could be Rooted or Unrooted
Type of Tree Reconstruction

- **Character-based**
  - Input is a multiple alignment of the sequences at the leaves. (find the topology that best explains the evolution of leaf sequences via mutations)

- **Distance-based**
  - Input is a matrix of distances between species.
Type of Tree Reconstruction

- **Character-based**
  - Input is a multiple alignment of the sequences at the leaves. (find the topology that best explains the evolution of leaf sequences via mutations)

- **Distance-based**
  - Input is a matrix of distances between species.
Distance Based Tree Reconstruction

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>ALIGNMENT</th>
<th>DISTANCE MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Chimp</td>
</tr>
<tr>
<td>Chimp</td>
<td>ACGTAGGCCT</td>
<td>0</td>
</tr>
<tr>
<td>Human</td>
<td>ATGTAAGACT</td>
<td>3</td>
</tr>
<tr>
<td>Seal</td>
<td>TCGAGAGCAC</td>
<td>6</td>
</tr>
<tr>
<td>Whale</td>
<td>TCGAAAGGCAT</td>
<td>4</td>
</tr>
</tbody>
</table>
## Distance Based Tree Reconstruction

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>ALIGNMENT</th>
<th>DISTANCE MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chimp</td>
<td>ACGTAGGCT</td>
<td>Chimp: 0</td>
</tr>
<tr>
<td>Human</td>
<td>ATGTAAGACT</td>
<td>3</td>
</tr>
<tr>
<td>Seal</td>
<td>TCGAGAGCAC</td>
<td>6</td>
</tr>
<tr>
<td>Whale</td>
<td>TCGAAAGCAT</td>
<td>4</td>
</tr>
</tbody>
</table>
Distances:

Are symmetric: for all $i$ and $j$, $D_{i,j} = D_{j,i}$

Are non-negative (for all $i$ and $j$, $D_{i,j} \geq 0$)
Distances in Trees

• Edges may have weights, which reflect:
  • Number of mutations on evolutionary path from one species to another
  • Or, time estimate for evolution of one species into another
• In a tree $T$ with $n$ leaves, we often compute the length of a path between leaves $i$ and $j$, $d_{ij}(T)$
  • $d_{ij}$ refers the the distance between $i$ and $j$ and is the sum of the weight of the edges between $i$ and $j$
Distance in Trees (cont’d)

For \( i = 1, j = 4 \), \( d_{ij} \) is:

\[
d(1,4) = 12 + 13 + 14 + 17 + 13 = 69
\]
Additive Distance Matrices

A distance matrix is called ADDITIVE if there exists a tree $T$ with $d_{ij}(T) = D_{ij}$
Additive Distance Matrices

Is this matrix additive???
Additive Distance Matrices

A distance matrix is called ADDITIVE if there exists a tree $T$ with $d_{ij}(T) = D_{ij}$
Additive Distance Matrices

Is this matrix additive???
Additive Distance Matrices

A distance matrix is called **ADDITIVE** if there exists a tree $T$ with $d_{ij}(T) = D_{ij}$

**NONADDITIVE** otherwise
Additive Matrices have a Simple Tree Fitting
Distance Based Phylogeny Problem

- **Goal**: Reconstruct an evolutionary tree from a distance matrix
- **Input**: $n \times n$ distance matrix $D_{ij}$
- **Output**: weighted unrooted (or rooted) tree $T$ with $n$ leaves fitting $D$

- If $D$ is additive, this problem has a solution and there are simple algorithms to solve it (we will not learn them in class)
- **However usually** $D$ is not additive
Rooted Ultrametric Trees
UPGMA
Unweighted Pair Group Method with Arithmetic Mean

• UPGMA is a clustering algorithm that:
  • Computes the distance between clusters using average pairwise distance
  • Assigns a height to every vertex in the tree, effectively assuming the presence of a molecular clock and dating every vertex
  • Assumes the matrix D is additive, so the generated tree fits D.
  • If D is not additive, UPGMA will generate a heuristic solution that does not fit D
  • Produces ultrametric trees – all leaves are equi-distant from the root.
UPGMA Algorithm (cont’d)
UPGMA Algorithm (cont’d)
UPGMA Algorithm (cont’d)
UPGMA Algorithm (cont’d)
UPGMA Algorithm (cont’d)
UPGMA’s Weakness

• The algorithm produces an ultrametric tree: the distance from the root to any leaf is the same

• UPGMA assumes a constant molecular clock: all species represented by the leaves in the tree are assumed to accumulate mutations (and thus evolve) at the same rate. This is one of the major pitfalls of UPGMA.
UPGMA’s Weakness: Example

Correct tree

UPGMA
Example for the next slide
Clustering in UPGMA

Given two disjoint clusters \( C_i, C_j \) of sequences,

\[
1 \quad \text{d}_{ij} = \frac{\sum_{p \in C_i, q \in C_j} d_{pq}}{|C_i| \times |C_j|}
\]

Note that if \( C_k = C_i \cup C_j \), then distance to another cluster \( C_l \) is:

\[
\text{d}_{il} |C_i| + \text{d}_{jl} |C_j| \quad \text{d}_{kl} = \frac{|C_i| + |C_j|}{|C_i| + |C_j|}
\]
UPGMA Algorithm

**Initialization:**
Assign each $x_i$ into its own cluster $C_i$ (clusters of size $1$)
Define one leaf per sequence, height $0$

**Iteration:**
Find two clusters $C_i, C_j$ s.t. $d_{ij}$ is min
Let $C_k = C_i \cup C_j$
Define node connecting $C_i, C_j$, & place it at height $d_{ij}/2$
Delete $C_i, C_j$

**Termination:**
When two clusters $i, j$ remain, place root at height $d_{ij}/2$
Example
Fig. 8.1. A: Rooted phylogenetic tree of seven globin sequences calculated using the UPGMA algorithm that assumes a molecular clock. The thin vertical line indicates the present point in time. B: Midpoint rooted phylogenetic tree of the same globin sequences calculated using the neighbor-joining algorithm that does not assume a molecular clock. C: Unrooted version of B. D: Outgroup rooted version of B. M: position of midpoint root; O: position of outgroup root.