Distance based phylogeny reconstruction

Usman Roshan
Phylogenetics

- Study of how species relate to each other
- Rich in computational problems
- Fundamental tool in comparative bioinformatics
Why phylogenetics?

• Study of evolution
  – Origin and migration of humans
  – Origin and spread of disease

• Many applications in comparative bioinformatics
  – Sequence alignment
  – Motif detection (phylogenetic motifs, evolutionary trace, phylogenetic footprinting)
  – Correlated mutation (useful for structural contact prediction)
  – Protein interaction
  – Gene networks
  – Vaccine development
  – And many more…
Phylogeny Problem

AGGGGCAT  TAGCCCCA  TAGACTTT  TGCACAAA  TGCACGCTT
Phylogeny Problem

• Two main methodologies:
  – Alignment first and phylogeny second
    • Construct alignment using one of the MANY alignment programs in the literature
    • Do manual (eye) adjustments if necessary
    • Apply a phylogeny reconstruction method
    • Fast but biologically not realistic
    • Phylogeny is highly dependent on accuracy of alignment (but so is the alignment on the phylogeny!)
  – Simultaneously alignment and phylogeny reconstruction
    • Output both an alignment and phylogeny
    • Computationally much harder
    • Biologically more realistic as insertions, deletions, and mutations occur during the evolutionary process
First methodology

• Compute alignment (for now we assume we are given an alignment)
• Construct a phylogeny (two approaches)
• Distance-based methods
  – Input: Distance matrix containing pairwise statistical estimation of aligned sequences
  – Output: Phylogenetic tree
  – Fast but less accurate
• Character-based methods
  – Input: Sequence alignment
  – Output: Phylogenetic tree
  – Accurate but computationally very hard
Definitions

• Tree:
  – Set of nodes and edges
  – Undirected graph
  – No cycles
  – Connected

• Examples

• Degree of node = number of edges connected to the node

• Binary tree: every node has at most two children

• Phylogeny: unrooted binary tree
Distance-based methods

**TRUE TREE**

**DNA SEQUENCES**
- $S_1$: ACAATTAGAAC
- $S_2$: ACCCTTAGAAC
- $S_3$: ACCATTCCAAAC
- $S_4$: ACCAGACCAAC

**STATISTICAL ESTIMATION OF PAIRWISE DISTANCES**

**METHODS SUCH AS NEIGHBOR JOINING**

**DISTANCE MATRIX**

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>$S_2$</td>
<td></td>
<td>0</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>$S_3$</td>
<td></td>
<td></td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>$S_4$</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Distance methods

TRUE TREE

DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

METHODS SUCH AS NEIGHBOR JOINING

DISTANCE MATRIX
Distance methods

• UPGMA: similar to hierarchical clustering but not additive
• Neighbor-joining: more sophisticated and additive
• What is additivity?
Additivity

A distance matrix $D$ is additive if there exists a tree, $T = (V, E)$, and $w : E \rightarrow \mathbb{R}^+$ such that $D_{ij} = \sum_{e \in P_{ij}} w(e)$.

Waterman et al, 1977, showed that:

\begin{tabular}{c|cccc}
  & S_1 & S_2 & S_3 & S_4 \\
\hline
S_1 & 0 & 3 & 6 & 5 \\
S_2 & 0 & 5 & 4 & \\
S_3 & 0 & 5 & & \\
S_4 & 0 & & &
\end{tabular}

POLYTIME INVERTIBLE
UPGMA

UPGMA is not additive but works for ultrametric trees. Takes $O(n^3)$ time.
**UPGMA**

**Input**: distance matrix $D$, **Output**: Phylogeny $T$

1. Set $d=D$
2. Initialize $n$ clusters where each cluster $i$ contains the sequence $I$
3. Find closest pair of clusters $i, j$, using distances in matrix $d$
4. Make them neighbors in the tree by adding new node $(ij)$, and set distance from $(ij)$ to $i$ and $j$ as $d_{ij}/2$
5. Update distance matrix $d$ with weighted average. For all clusters $k$ do the following ($ni$ and $nj$ are size of clusters $i$ and $j$ respectively)

$$d(i, j) = \frac{1}{n_i n_j} \sum_{i' \in L(i)} \sum_{j' \in L(j)} D(i', j')$$
UPGMA

6. Delete columns and rows for $i$ and $j$ in $d$ and add new ones corresponding to cluster $(ij)$ with distances as computed above.

7. Goto step 2 until only one cluster is left.
UPGMA

<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>6</td>
<td>32</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>32</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A → B

13

3

3

A

B

C

D
UPGMA

Doesn’t work (in general) for non ultrametric trees
UPGMA

UPGMA constructs incorrect tree here

<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>13</td>
<td>19</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>12</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
UPGMA

Bipartition \((BC, AD)\) is not in true tree