Phylogeny and Molecular Evolution

Maximum Likelihood Approach
Large Parsimony

A character-based method

Input:

\(n\) sequences (one per species), all of length \(m\).

• **Goal:**
  • Find a tree with the input sequences at its leaves, and an assignment of sequences to internal nodes, such that the total number of changes is minimized.
Parsimony Based Reconstruction

Two separate components:

1. A procedure to find the minimum number of changes needed to explain the data (for a given tree topology)

2. A search through the space of trees.

3. We already say that (1) can be done in polynomical time, we will now show that (2) is hard.
Exploring the Space of Rooted Trees

<table>
<thead>
<tr>
<th>taxa (n)</th>
<th># trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>30,405,375</td>
</tr>
</tbody>
</table>
How many distinct rooted trees with $n$ leaves?
Exploring the Space of Trees

<table>
<thead>
<tr>
<th>taxa (n)</th>
<th># trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>30,405,375</td>
</tr>
</tbody>
</table>
Probabilistic Approaches

Consider the phylogenetic tree to be a stochastic process. When the data consists only of the leaves sequences (but the topology is fixed):

The likelihood of transition from character \( x \) to character \( y \) during the time \( t_{vu} \) is given by parameters \( P_{x \rightarrow y}(t_{vu}) \). The likelihood of a letter \( x \) in the root is \( q_x \).

Given the complete tree, its probability is defined by the values of the \( P_{x \rightarrow y}(t_{vu}) \)'s and the \( q_x \)'s.
Probabilistic Approaches

Definition

- *Labels* are the vectors of $m$ character values associated with each species, or node in the tree.

- A *reconstruction* is a full labeling of the tree’s internal nodes.

- A *branch length* $t_{vu}$ is the length of the edge between nodes $v$ and $u$, and it measures the biological time, or genetic distance, between the species associated with these nodes.
Probabilistic Approaches

**Problem:** Calculating the Likelihood of a tree.

**INPUT:**
- A matrix $M$ describing a set of $m$ characters for each one of $n$ given species.
- A tree $T$ with the above species at its leaves and with known branch lengths $t_{vu}$.

**GOAL:** Maximize probability of $P(M|T)$ by finding an optimal reconstruction $T$, its internal nodes labeling and branch lengths.
Probabilistic Approaches

Assumptions:

• Characters are independent of each other.
• Markov model - probability of a label depends only on its parent and the branch length, $t$, between them.
Likelihood of a Tree

Simple case:
There is only one character identifying each species.
Need to sum over all possible reconstructions.
Likelihood of a Tree

Example:

\[ L = P(M \mid T) = \sum_{r} \sum_{v} P(r) \cdot P_{r \rightarrow s}(t_{rs}) \cdot P_{r \rightarrow v}(t_{rv}) \cdot P_{v \rightarrow u}(t_{vu}) \cdot P_{v \rightarrow w}(t_{vw}) \]
Likelihood of a Tree

General case (sequences longer than one character at the leaves):

• Repeat the above calculation for each character separately, and then multiply the results (assumption: characters are pairwise independent).
Likelihood of a Tree

• The general equation is

\[
L = P(M \mid T) = \prod_{\text{character } j} P(M_j \mid T) \\
= \prod_{\text{character } j} \left\{ \sum_{\text{reconstruction } R} P(M_j, R \mid T) \right\} \\
= \prod_{\text{character } j} \left\{ \sum_{\text{reconstruction } R} \left[ P(\text{root}) \cdot \prod_{\text{edge } u \rightarrow v} P_{u \rightarrow v}(t_{uv}) \right] \right\}
\]
Example Maximal Likelihood
Exploring the Space of Trees

- we’ve considered how to find the minimum number of changes for a given tree topology

- need some search procedure for exploring the space of tree topologies

- Given $n$ sequences there are $(2n - 3)!!$ possible rooted trees, where

$$(2n - 3)!! = \prod_{i=3}^{n}(2i - 3) = 3 \cdot 5 \cdot \ldots \cdot (2n - 3)$$
Exploring the Space of Trees

<table>
<thead>
<tr>
<th>taxa (n)</th>
<th># trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>30,405,375</td>
</tr>
</tbody>
</table>