Phylogenetics: Parsimony

COMP 571
Luay Nakhleh, Rice University
The Problem

- Input: Multiple alignment of a set $S$ of sequences
- Output: Tree $T$ leaf-labeled with $S$
Assumptions

* Characters are mutually independent
* Following a speciation event, characters continue to evolve independently
In parsimony-based methods, the inferred tree is fully labeled.
A Simple Solution: Try All Trees

Problem:

* (2n-3)!! rooted trees
* (2m-5)!! unrooted trees
A Simple Solution: Try All Trees

<table>
<thead>
<tr>
<th>Number of Taxa</th>
<th>Number of unrooted trees</th>
<th>Number of rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
<td>10395</td>
</tr>
<tr>
<td>8</td>
<td>10395</td>
<td>135135</td>
</tr>
<tr>
<td>9</td>
<td>135135</td>
<td>2027025</td>
</tr>
<tr>
<td>10</td>
<td>2027025</td>
<td>34459425</td>
</tr>
<tr>
<td>20</td>
<td>2.22E+20</td>
<td>8.20E+21</td>
</tr>
<tr>
<td>30</td>
<td>8.69E+36</td>
<td>4.95E+38</td>
</tr>
<tr>
<td>40</td>
<td>1.31E+55</td>
<td>1.01E+57</td>
</tr>
<tr>
<td>50</td>
<td>2.84E+74</td>
<td>2.75E+76</td>
</tr>
<tr>
<td>60</td>
<td>5.01E+94</td>
<td>5.86E+96</td>
</tr>
<tr>
<td>70</td>
<td>5.00E+115</td>
<td>6.85E+117</td>
</tr>
<tr>
<td>80</td>
<td>2.18E+137</td>
<td>3.43E+139</td>
</tr>
</tbody>
</table>
Solution

- Define an optimization criterion
- Find the tree (or, set of trees) that optimizes the criterion
- Two common criteria: parsimony and likelihood
Parsimony
The parsimony of a fully-labeled unrooted tree $T$, is the sum of lengths of all the edges in $T$

Length of an edge is the Hamming distance between the sequences at its two endpoints

$PS(T)$
Parsimony score = 5
Maximum Parsimony (MP)

* Input: a multiple alignment $S$ of $n$ sequences

* Output: tree $T$ with $n$ leaves, each leaf labeled by a unique sequence from $S$, internal nodes labeled by sequences, and $PS(T)$ is minimized
The three trees are equally good MP trees.
ACT
GTT
GTA
ACA
Weighted Parsimony

* Each transition from one character state to another is given a weight
* Each character is given a weight
* See a tree that minimizes the weighted parsimony
Both the MP and weighted MP problems are NP-hard
A Heuristic For Solving the MP Problem

- Starting with a random tree $T$, move through the tree space while computing the parsimony of trees, and keeping those with optimal score (among the ones encountered)
- Usually, the search time is the stopping factor
Two Issues

* How do we move through the tree search space?
* Can we compute the parsimony of a given leaf-labeled tree efficiently?
Searching Through the Tree Space

* Use tree transformation operations (NNI, TBR, and SPR)
Searching Through the Tree Space

* Use tree transformation operations (NNI, TBR, and SPR)

Local maximum

Global maximum
Computing the Parsimony Length of a Given Tree

- Fitch's algorithm
- Computes the parsimony score of a given leaf-labeled rooted tree
- Polynomial time
Fitch's Algorithm

* Alphabet $\Sigma$
* Character $c$ takes states from $\Sigma$
* $v_c$ denotes the state of character $c$ at node $v$
Fitch's Algorithm

- **Bottom-up phase:**

  - For each node $v$ and each character $c$, compute the set $S_{c,v}$ as follows:
    - If $v$ is a leaf, then $S_{c,v} = \{v_c\}$
    - If $v$ is an internal node whose two children are $x$ and $y$, then
      
      $$S_{c,v} = \begin{cases} 
      S_{c,x} \cap S_{c,y} & \text{if } S_{c,x} \cap S_{c,y} \neq \emptyset \\
      S_{c,x} \cup S_{c,y} & \text{otherwise}
      \end{cases}$$
Fitch's Algorithm

* Top-down phase:
  * For the root $r$, let $r_c = a$ for some arbitrary $a$ in the set $S_{c,r}$
  * For internal node $v$ whose parent is $u$,
    
    $$v_c = \begin{cases} 
    u_c & u_c \in S_{c,v} \\
    \text{arbitrary } \alpha \in S_{c,v} & \text{otherwise}
    \end{cases}$$
3 mutations
Fitch’s Algorithm

- Takes time $O(nkm)$, where $n$ is the number of leaves in the tree, $m$ is the number of sites, and $k$ is the maximum number of states per site (for DNA, $k=4$)
Informative Sites and Homoplasy

* Invariable sites: In the search for MP trees, sites that exhibit exactly one state for all taxa are eliminated from the analysis.

* Only variable sites are used.
Informative Sites and Homoplasy

* However, not all variable sites are useful for finding an MP tree topology

* Singleton sites: any nucleotide site at which only unique nucleotides (singletons) exist is not informative, because the nucleotide variation at the site can always be explained by the same number of substitutions in all topologies
C, T, G are three singleton substitutions \( \Rightarrow \) non-informative site

All trees have parsimony score 3
Informative Sites and Homoplasy

* For a site to be informative for constructing an MP tree, it must exhibit at least two different states, each represented in at least two taxa.
* These sites are called informative sites.
* For constructing MP trees, it is sufficient to consider only informative sites.
Informative Sites and Homoplasy

* Because only informative sites contribute to finding MP trees, it is important to have many informative sites to obtain reliable MP trees.

* However, when the extent of homoplasy (backward and parallel substitutions) is high, MP trees would not be reliable even if there are many informative sites available.
Measuring the Extent of Homoplasy

+ The consistency index (Kluge and Farris, 1969) for a single nucleotide site (i-th site) is given by $c_i = m_i / s_i$, where

+ $m_i$ is the minimum possible number of substitutions at the site for any conceivable topology (= one fewer than the number of different kinds of nucleotides at that site, assuming that one of the observed nucleotides is ancestral)

+ $s_i$ is the minimum number of substitutions required for the topology under consideration
Measuring the Extent of Homoplasy

- The lower bound of the consistency index is not 0
- The consistency index varies with the topology
- Therefore, Farris (1989) proposed two more quantities: the retention index and the rescaled consistency index
The Retention Index

The retention index, $r_i$, is given by $(g_i-s_i)/(g_i-m_i)$, where $g_i$ is the maximum possible number of substitutions at the $i$-th site for any conceivable tree under the parsimony criterion and is equal to the number of substitutions required for a star topology when the most frequent nucleotide is placed at the central node.
The Retention Index

* The retention index becomes 0 when the site is least informative for MP tree construction, that is, $s_i = g_i$
The Rescaled Consistency Index

\[ rC_i = \frac{g_i - s_i}{g_i - m_i} \cdot \frac{m_i}{s_i} \]
Ensemble Indices

* The three values are often computed for all informative sites, and the ensemble or overall consistency index (CI), overall retention index (RI), and overall rescaled index (RC) for all sites are considered.
Ensemble Indices

\[ CI = \frac{\sum_i m_i}{\sum_i s_i} \]

\[ RI = \frac{\sum_i g_i - \sum_i s_i}{\sum_i g_i - \sum_i m_i} \]

\[ RC = CI \times RI \]

These indices should be computed only for informative sites, because for uninformative sites they are undefined.
The homoplasy index is  \( HI = 1 - CI \).

When there are no backward or parallel substitutions, we have  \( HI = 0 \). In this case, the topology is uniquely determined.
A Major Caveat

* Maximum parsimony is not statistically consistent!
Questions?