Phylogeny and Molecular Evolution

Maximum Likelihood Approach
Parsimony approach... can we make the search faster?
Branch and bound

>wanted to improve their score and have only one solution for each solution set.

A C G T G A C
B c t......
C c a g....
D c a a c.
E c a a t

A C
B D
E A
B C
D E

4
6 7 7
9 9 8 8
7
Parsimony approach... can we employ a less naïve (stochastic) model?
Parsimony approach... can we employ a less naïve (stochastic) model?

Given additional (trained) data, for example:

- The probability of a specific letter in one species turning into another specific letter in its offspring (given the elapsed time between them).
- The probability of a certain letter to appear at the root of a rooted phylogenetic tree
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 topology with 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 taking into account all possible reconstructions of $T$. 
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