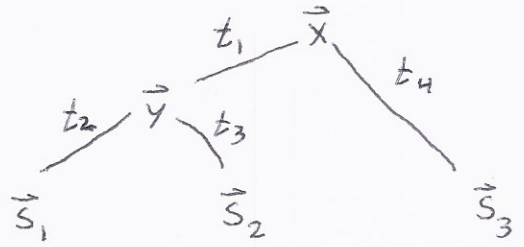
Let's imagine we have some gene sequences $\vec{s}_1$, $\vec{s}_2$, and $\vec{s}_3$. If they evolved without recombination, then there is some true evolutionary relationship among them that can be represented by a phylogenetic tree:



Here is one possible tree topology. The number of possible topologies increases rapidly with the number of sequences

The tree, T, is fully specified by the topology and the branch lengths (t_1, t_2, t_3, t_4).

Let's say we have some evolutionary model that tells us ~~the probability that~~ ~~how likely~~ each sequence $\vec{s}_1$ is likely to evolve to some other sequence $\vec{s}_2$ after some duration of time t:

$$Pr(\vec{s}_2 \mid \vec{s}_1, t)$$ ← probability of being $\vec{s}_2$ after duration t given that sequence was $\vec{s}_1$ at start of time duration

Note that this transition probability does not depend on the absolute date, only the duration of time t. So we are <u>assuming</u> evolution is time homogenous. When might this be untrue?

Let's return to the tree above. We would like to write the <u>likelihood</u> of the sequence data given the tree:

$$Pr(\vec{s}_1, \vec{s}_2, \vec{s}_3 \mid T) = \sum_{\vec{x}} \sum_{\vec{y}} Pr(\vec{x}) \cdot Pr(\vec{s}_3 \mid \vec{x}, t_4) \cdot Pr(\vec{y} \mid \vec{x}, t_1) \cdot Pr(\vec{s}_1 \mid \vec{y}, t_2) \cdot Pr(\vec{s}_2 \mid \vec{y}, t_3)$$

Felsenstein pruning algorithm, a form of dynamic programming →

$$= \sum_{\vec{x}} Pr(\vec{x}) \cdot Pr(\vec{s}_3 \mid \vec{x}, t_4) \left[\sum_{\vec{y}} Pr(\vec{y} \mid \vec{x}, t_1) \cdot Pr(\vec{s}_1 \mid \vec{y}, t_2) \cdot Pr(\vec{s}_2 \mid \vec{y}, t_3) \right]$$

The $Pr(\vec{s} \mid \vec{x}, t_4)$ terms are the transition probabilities given by our evolutionary model.

What is $Pr(\vec{x})$? It is the root frequencies or stationary state. This is the principal eigenvector of the transition matrix.

We now have an equation to compute:

$$Pr(\text{sequences} \mid \text{tree} = T, \text{model} = Pr(\vec{s}_1 \mid \vec{s}_2, t))$$

Maximum likelihood:

Find tree that maximizes

$Pr(\text{sequences} \mid \text{tree, model})$

Bayesian:

Place prior over trees: $Pr(\text{tree})$

Then compute posterior over trees:

$$Pr(\text{tree} \mid \text{sequences}) \propto Pr(\text{tree}) \cdot Pr(\text{sequences} \mid \text{tree})$$

Computation uses MCMC

Choice of method depends on taste, computational difficulty, and question at hand.

Sometimes a researcher just wants a tree to use as figure.

Other times they want to compute posterior distribution over a statistic calculated from tree.