Categories of tree reconstruction methods

	Parsimony	Distance	Maximum likelihood estimation	Bayesian methods
Character data	x		x	x
Pairwise distances		x		

Finding the optimal tree

Given *k* taxa,

- Consider all trees with *k* leaves
- ➢ Score each tree with respect to chosen evolutionary model.
- Select highest scoring tree(s)

Criteria for evaluating which tree best fits the data:

- Maximum parsimony (character data)
- Minimum evolution (distance data)
- ➢ Maximum Likelihood (character data)

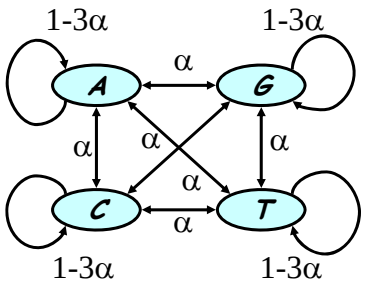
Finding the optimal tree

Given an MSA for *k* taxa:

- For each tree, *t*, with *k* leaves
- Determine *P(MSA| t),* the likelihood of the data given the tree.
- Select the tree(s) that maximize(s) the likelihood.

We will use a Markov model of sequence substitution to calculate *P(MSA|t),* so we will briefly review substitution models first….

Recap: Models of DNA substitution



Jukes Cantor model

- Total substitution rate: 3α
- Stationary distribution:

  π=(1/4, 1/4, 1/4, ¼)
- Assumptions:

  All substitutions are equally likely

  Site independence

Correcting distances

Given an ungapped alignment of sequences *s* and *t* of length *n* with *m* mismatches, what is the expected number of substitutions at site *i?*

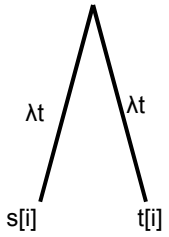
$2\lambda t = 6\alpha t$

Goal: obtain an estimate for αt in terms of *m* and *n* using the Jukes Cantor model

From the JC rate matrix, we derived the probability of observing a match or mismatch given α and *t:*

$P_{aa}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t,}$

$P_{ab}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t,}$ *if* $a \neq b$



Correcting distances continued…

We further obtained and expression for the expected number of substitutions, given the observed number of mismatches:

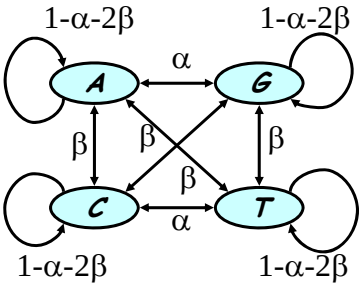
$6\alpha t = -3/4 \ln(1 - 4/3\ m/n)$

For estimating the likelihood of an MSA given a tree, we will make use of these equations:

$P_{aa}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t,}$

$P_{ab}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t,}$ *if* $a \neq b$

More realistic models



- Separate rates for transitions and transversions
- Stationary distribution:
  π=(1/4, 1/4, 1/4, ¼)

Kimura 2 parameter

More more realistic models

Model	Description	N° parameters
Jukes-Cantor (JC)	Equal base frequencies, all substitutions equally likely	1
Kimura (K2P)	Equal base frequencies, allow for transition/transversion bias	2
Felsenstein (F81)	actual base frequencies, all substitutions equally likely	4
Hasegawa et al (HKY85)	actual base frequencies, allow for transition/transversion bias	5
General Time Reversible (GTR)	actual base frequencies, all six pairs of substitutions have different rates	9

Maximum likelihood estimation

- Likelihood: Probability of the data given the hypothesis (or model) = P(D|H)
- Maximum likelihood estimation: The hypothesis that maximizes the likelihood is the best explanation for the data

Maximum Likelihood Estimation for Phylogeny Reconstruction

Consider all topologies, T_i,

Select T_i such that $P(D|T_i)$ is maximum, where D = *MSA.*

Note:

Character based method

Assumes neutral evolution

Correction for multiple substitutions is built into the method.

Maximum Likelihood Estimation for Phylogeny Reconstruction

Data: Multiple sequence alignment, *n* sites, *k* taxa

Model: sequence evolution, e.g. Jukes Cantor

Parameters to be estimated:

Branch lengths, $\underline{\boldsymbol{x}} = (x_1, x_2 \ldots x_j)$

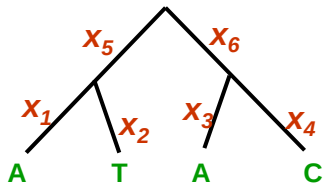
Model parameters?

Select T_i, $\underline{x}$ such that $P(MSA \mid T_i, \underline{\boldsymbol{x}})$ is maximum

Calculating the likelihood of MSA given a tree, T_i, with branch lengths, $\underline{x} = (x_1, x_2, \ldots)$

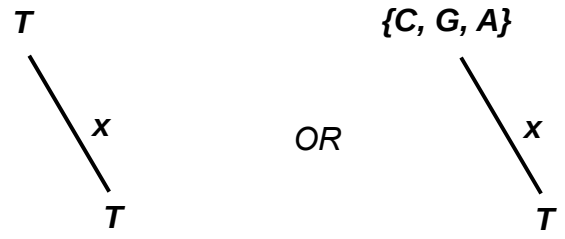
Calculate likelihood of each site, and take the product to get the likelihood of the entire MSA:

```
...TCAGG...
...TGTCG...
...TGACG...
...TCCGA...
```



$$P(MSA \mid T_i, \underline{x}) = \prod_{a=1}^{k} P(site_a \mid T_i, \underline{x})$$

Calculating the likelihood for a single branch



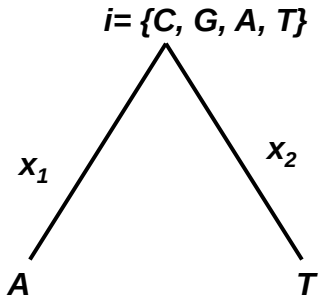
$p(T)P_{TT}(x)$ + $p(A)\,P_{AT}(x)$ + $p(C)\,P_{CT}(x)$ + $p(G)\,P_{GT}(x)$

Probabilities given by, e.g., Jukes Cantor model:

$P_{TT}(x) = (1/4 + 3/4\, e^{-4x_i})$, $P_{TA}(x) = (1/4 - 1/4\, e^{-4x_i})$, etc.

where p(k) is the background frequency of *k.*

Calculating the likelihood for 2 branches taking common ancestry into account:



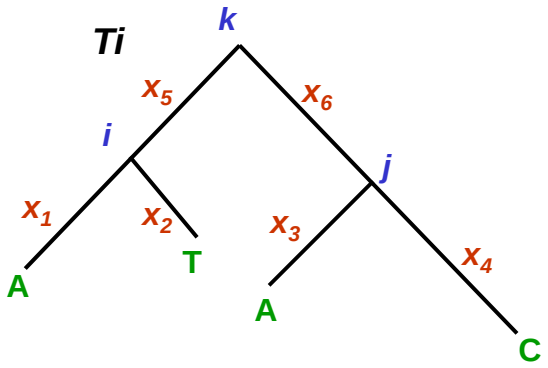
$p(i \rightarrow A|x_1)\; p(i \rightarrow T|x_2) =$

$p(A)P(x_1)_{AA}P(x_2)_{AT} + p(T)P(x_1)_{TA}P(x_2)_{TT}$

$+ p(C)P(x_1)_{CA}P(x_2)_{CT} + p(G)P(x_1)_{GA}P(x_2)_{GT}$

Calculating the likelihood the site for the entire tree:

...TCAGG...
...TGTCG...
...TGACG...
...TCCGA...



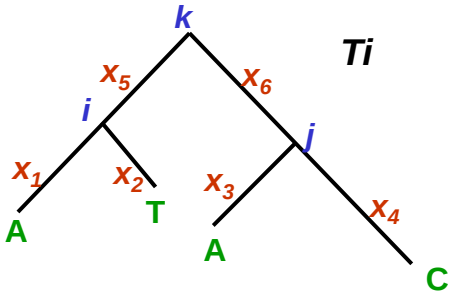
$$P(\{A,T,A,C\}^T \mid T_i, \underline{l}, \underline{x}) =$$

$$\sum_{i\in\{A,C,G,T\}}^{4} \sum_{j\in\{A,C,G,T\}}^{4} \sum_{k\in\{A,C,G,T\}}^{4} p(k)\,p(k\to i \mid x_5)\,p(k\to j \mid x_6)$$

$$\times\, p(i\to A \mid x_1)\,p(i\to T \mid x_2)\,p(j\to A \mid x_3)\,p(j\to C \mid x_4)$$

Likelihood of the entire MSA for this tree:

...TCAGG...
...TGTCG...
...TGACG...
...TCCGA...



$$P(MSA \mid T_i, \underline{x}) = \prod_{a=1}^{k} P(site_a \mid T_i, \underline{x})$$

Assumptions:

Sites are independent: score each site separately

Lineages are independent (Markov property): compute each branch separately

Maximum Likelihood Estimation for Phylogeny Reconstruction

Note we need to consider

- All sites: O(n)
- All trees: O($\mathcal{T}_{\text{rooted}}$(k))
- All combinations of internal labels: O($|\Sigma|^{k)}$
- A branch lengths: O(k) branches

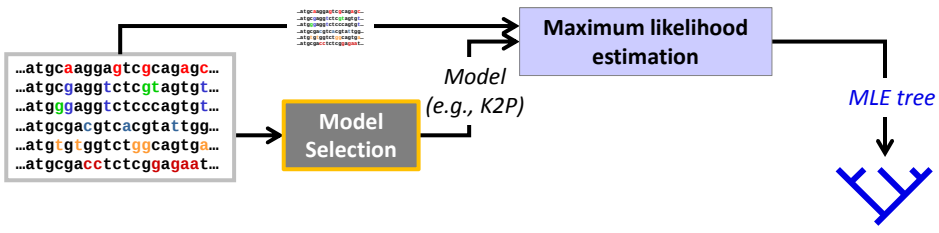
  Branch lengths are estimated numerically

In theory, we select the substitution model, the substitution rate parameters, and the tree in a single unified maximum likelihood calculation.

In practise, model selection and tree inference are calculated in separate steps to reduce the computational cost:

1. Model selection: estimate a rough tree from the MSA and, for each model, calculate the likelihood of the MSA given the model.
2. Tree inference: given the model and rate parameters selected in step 1, infer the tree that maximizes the probability of the MSA.

Several model selection prgrams are available including ModelGenerator, Jmodeltest and Prottest.



Maximum Likelihood Estimation for Phylogeny Reconstruction

- Computationally intensive
- Consistent (more data, better estimation)
- If evolutionary model is a reversible Markov chain (e.g., JC), then the MLE distance matrix converges to additive.

  ➔ Neighbor Joining is a consistent method

  Farach and Kannan, 96
- Note that parsimony is not consistent.

Selecting data for tree reconstruction

- For reconstructing recent events, use DNA sequences
- For reconstructing distant events, use amino acid sequences
- Select sequences that
  - Are present in all taxa
  - Contain a conserved region
  - Exhibit variation within that region
  - e.g., Ribosomal (16sRNA) genes were used to reconstruct the tree of life. These genes encode products use in all organisms from bacteria to mammals.
- Pitfalls: duplicated genes, horizontal gene transfer, mosaic genes.

Comparison of Phylogeny Reconstruction Methods

- Parsimony
  - Selection dominates, e.g., ribosomal genes
  - Exhaustive or heuristic search, branch and bound
- Distance
  - Neutral mutation dominates, e.g., immunoglobulin sequences
  - Exhaustive or heuristic search, greedy methods.
  - Neighbor Joining finds correct tree in quadratic time if data is additive.
  - UPGMA finds correct tree in quadratic time if data is ultrametric.
- Maximum Likelihood
  - Neutral mutation dominates, e.g., immunoglobulin sequences
  - Exhaustive or heuristic search

	Parsimony	**Distance**	**MLE**
Data	Character	Distance	Character
NP-complete	Yes	Yes	Yes
Topology	Yes	Yes	Yes
Branch lengths	Yes	Yes	Prob
Ancestral states	Yes	No	Prob
Consistent	No	Yes	Yes
Selective pressure	Yes	No	No
Model of mutational change	No	Yes	Yes