

Maximum likelihood estimation

Alice tosses a coin eight times and obtains six heads and two tails. *What is the probability that the coin is biased given the observation of six heads and two tails?*

- Observation: Data, $D = 6 H, 2 T$
- What process generated this data?
 - Alternative hypothesis: $H_a (q \neq 0.5)$
 - Null hypothesis: $H_o (q = 0.5)$
- *What is $P(H_a | D)$?*

Bayes Rule

$$P(H_a | D) = \frac{P(D | H_a)P(H_a)}{P(D)}$$

posterior probability likelihood prior probability
probability of data independent of hypothesis

Bayes Rule

$$P(H_a | D) = \frac{P(D | H_a)P(H_a)}{P(D)}$$

$$\frac{P(H_a | D)}{P(H_o | D)} = \frac{P(D | H_a)P(H_a)}{P(D | H_o)P(H_o)}$$

prior probability is hard to estimate

Bayes Rule

$$P(Ha | D) = \frac{P(D | Ha)P(Ha)}{P(D)}$$

$$\frac{P(Ha | D)}{P(H_0 | D)} = \frac{P(D | Ha)P(Ha)}{P(D | H_0)P(H_0)}$$

$$\frac{P(Ha | D)}{P(H_0 | D)} = \frac{P(D_1 | Ha)P(D_2 | Ha)P(D_3 | Ha) \dots P(Ha)}{P(D_1 | H_0)P(D_2 | H_0)P(D_3 | H_0) \dots P(H_0)}$$

Bayes Rule

$$P(Ha | D) = \frac{P(D | Ha)P(Ha)}{P(D)}$$

$$\frac{P(Ha | D)}{P(H_0 | D)} = \frac{P(D | Ha)P(Ha)}{P(D | H_0)P(H_0)}$$

$$\frac{P(Ha | D)}{P(H_0 | D)} = \frac{P(D_1 | Ha)P(D_2 | Ha)P(D_3 | Ha) \dots P(Ha)}{P(D_1 | H_0)P(D_2 | H_0)P(D_3 | H_0) \dots P(H_0)}$$

$$\frac{P(Ha | D)}{P(H_0 | D)} \approx \frac{\prod_i P(D_i | Ha)}{\prod_i P(D_i | H_0)}$$

Hypothesis testing using a likelihood ratio

How likely is the data under the alternate hypothesis compared with the likelihood under the null hypothesis?

$$\text{Likelihood ratio: } \frac{P(D | H_a)}{P(D | H_0)} = \frac{P(6 \text{ heads in } 8 \text{ tosses} | q)}{P(6 \text{ heads in } 8 \text{ tosses} | 0.5)}$$

$P(\text{toss yields heads}): H_a: q \neq 0.5, H_0: 0.5$

Need to estimate q

Note: There are other ways to test a hypothesis; e.g., a p -value.

Maximum Likelihood Estimation

What process generated this data?

- Model with parameters: e.g., binomial with parameter q

$$P(n, k, q) = \binom{n}{k} q^k (1 - q)^{n-k}$$

- The best estimate of q is the value that maximizes the likelihood of the data. To obtain q , solve:

$$\frac{dP(D | H)}{dq} = 0 \quad \frac{d\left(\binom{6}{2} q^6 (1 - q)^2\right)}{dq} = 0 \quad \boxed{q = 0.75}$$

Hypothesis testing using a (log) odds ratio

How likely is the data under the alternate hypothesis compared with the likelihood under the null hypothesis?

$$\text{Likelihood ratio: } \frac{P(D | H_a)}{P(D | H_0)} = \frac{P(6 \text{ heads in 8 tosses} | 0.75)}{P(6 \text{ heads in 8 tosses} | 0.5)}$$
$$\frac{(0.75)^6(0.25)^2}{(0.5)^8} = 2.85$$

Observing 6 heads in 8 coin tosses is 2.85 times as likely if $q = 0.75$ than if the coin is fair.

Note: the sample size is very small!

Note:

- The estimate improves as the sample size increases. A method is *consistent* if $\lim_{n \rightarrow \infty} \hat{q} = q$
- For mathematical convenience we may use the log likelihood ratio: $\log \frac{P(D | H_a)}{P(D | H_0)}$
- In general, the probability distribution is unknown. Select a model and maximize the likelihood with respect to that model. Results can vary with the choice of model
- We estimated a parameter and determined the likelihood in a single, unified process.

Maximum Likelihood Estimation for Phylogeny Reconstruction

Consider all topologies, T_i

Select T_i such that $P(D|T_i)$ is maximum, where $D = \text{MSA}$.

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:

Internal labels, $\underline{l} = (l_1, l_2, \dots, l_j)$

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

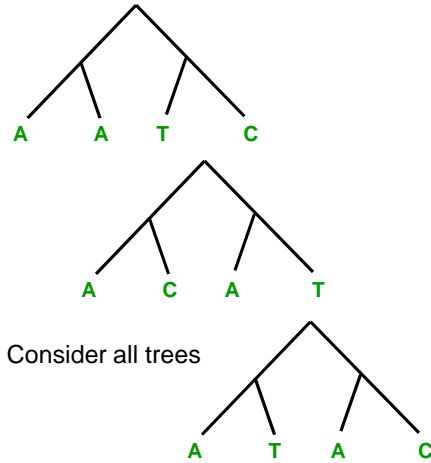
Model parameters?

Select $T_i, \underline{l}, \underline{x}$ such that $P(\text{MSA} | T_i, \underline{l}, \underline{x})$ is maximum

Likelihood of MSA

Consider all sites

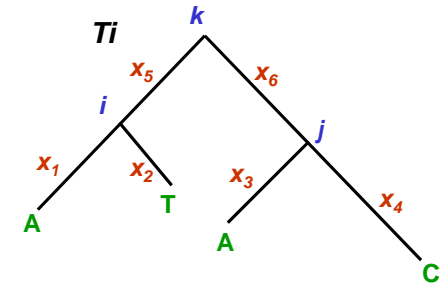
...TCAGG...
 ...TGTCG...
 ...TGACG...
 ...TCCGA...



Consider all trees

For each site and each tree

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



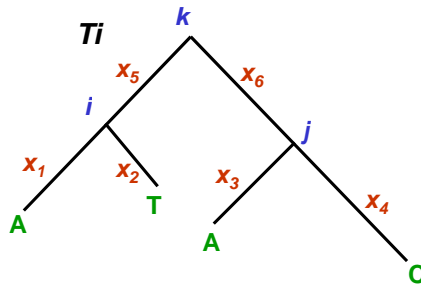
Consider all

Internal labels, $l = (i, j, k)$

Branch lengths, $\underline{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$

Likelihood of one site:

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



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

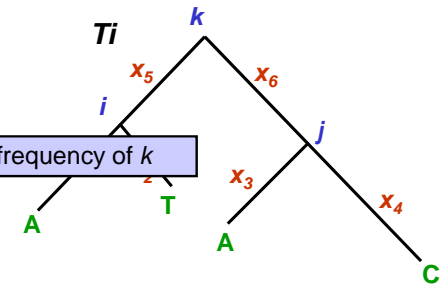
$$\sum_{i \in \{A, C, G, T\}} \sum_{j \in \{A, C, G, T\}} \sum_{k \in \{A, C, G, T\}} p(k) p(k \rightarrow i | x_5) p(k \rightarrow j | x_6)$$

$$\times p(i \rightarrow A | x_1) p(i \rightarrow T | x_2) p(j \rightarrow A | x_3) p(j \rightarrow C | x_4)$$

Likelihood of one site:

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

$p(k)$ is the background frequency of k



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

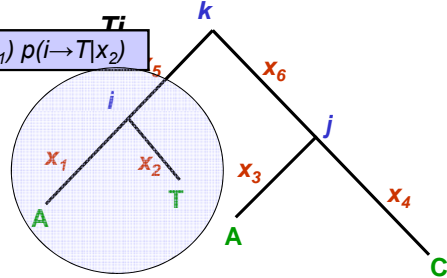
$$\sum_{i \in \{A, C, G, T\}} \sum_{j \in \{A, C, G, T\}} \sum_{k \in \{A, C, G, T\}} p(k) p(k \rightarrow i | x_5) p(k \rightarrow j | x_6)$$

$$\times p(i \rightarrow A | x_1) p(i \rightarrow T | x_2) p(j \rightarrow A | x_3) p(j \rightarrow C | x_4)$$

Likelihood of one site:

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

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

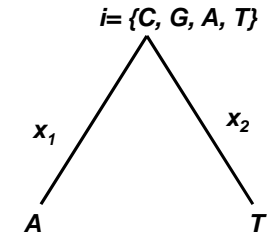


$$P(\{A, T, A, C\}^T | T_i, \underline{l}, \underline{x}) =$$

$$\sum_{i \in \{A, C, G, T\}} \sum_{j \in \{A, C, G, T\}} \sum_{k \in \{A, C, G, T\}} p(k) p(k \rightarrow i | x_1) p(k \rightarrow j | x_2)$$

$$\times p(i \rightarrow A | x_1) p(i \rightarrow T | x_2) p(j \rightarrow A | x_3) p(j \rightarrow C | x_4)$$

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



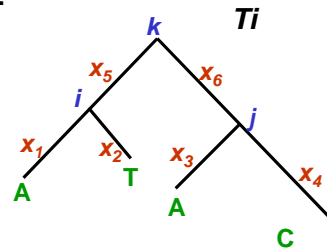
$$P(i=A)P(x_1)_{AA}P(x_2)_{AT} + P(i=T)P(x_1)_{TA}P(x_2)_{TT} \\ + P(i=C)P(x_1)_{CA}P(x_2)_{CT} + P(i=G)P(x_1)_{GA}P(x_2)_{GT}$$

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

$$P(x)_{CC} = (1/4 + 3/4 e^{-4x_i}), P(x)_{CG} = (1/4 - 1/4 e^{-4x_i}), \text{ etc.}$$

Likelihood of MSA:

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



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

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

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.

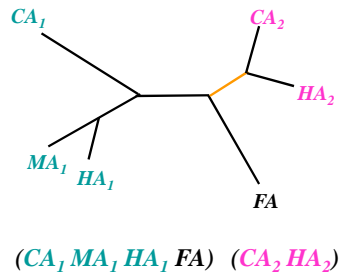
Comparison of Phylogeny Reconstruction Methods

- 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	Max Likelihood
Data	Character	Distance	Character
NP-complete	Yes	Yes	Yes
Topology	Yes	Yes	Yes
Branch lengths	Yes	Yes	Prob
Ancestral states	Yes	No	Prob
DNA	Yes	Yes	Yes
Amino acids	Yes	Yes	Very slow
Consistent	No	Yes	Yes
Selective pressure	Yes	No	No
Model of mutational change	No	Yes	Yes

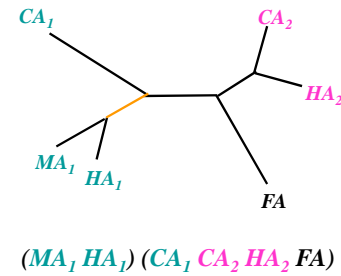
Bootstrapping, Branches and Partitions

- Every edge partitions a tree into two groups of taxa



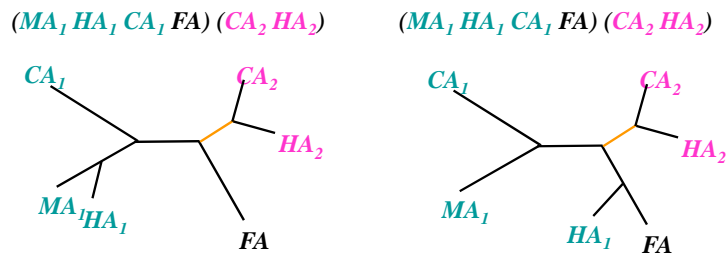
Bootstrapping, Branches and Partitions

- Every edge partitions a tree into two groups of taxa



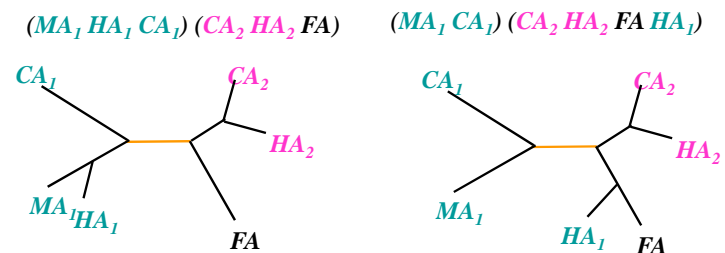
Bootstrapping, Branches and Partitions

- These two trees are different, but they share a partition



Bootstrapping, Branches and Partitions

- Neither of these partitions exist in the other tree



Bootstrapping a gene tree

- For $i = 1$ to N
 - Construct MSA' by sampling columns from the original MSA *with replacement*
 - Construct a new tree, t' , from MSA'
 - Tabulate the partitions in t' .
- For every partition, p , in the original tree, $score(p) = (the\ number\ observations\ of\ p)/N$

