Inferencia Filogenética

Types of Data

Tree-building method	Distances	Nucleotide sites
Clustering algorithm	UPGMA neighbour-joining	
Optimality criterion	Minimum evolution	Maximum parsimony Maximum likelihood

Maximum likelihood

The best explanation for the observed outcome

$$\text{Prob}\left(D \mid H\right)$$

Likelihood de H

A maximum likelihood:

Model of sequence evolution
Tree - topology and branch lengths
Observed data

- For a tree topology, what set of branch lengths makes the oberserved data most likely

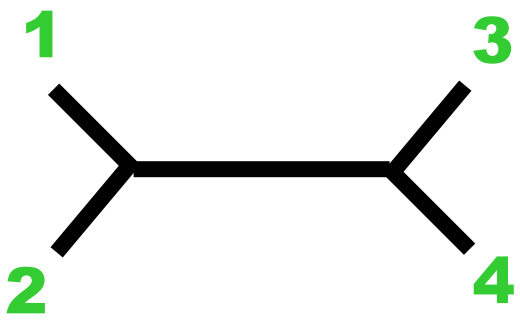
- Which tree of all possible trees has the greatest likelihood

Dois pressupostos:

1 – Evolution in different sites (on the given tree) is independent

2 – Evolution in different lineages is independent

Exemplo simples: Dados, Modelo e Árvore



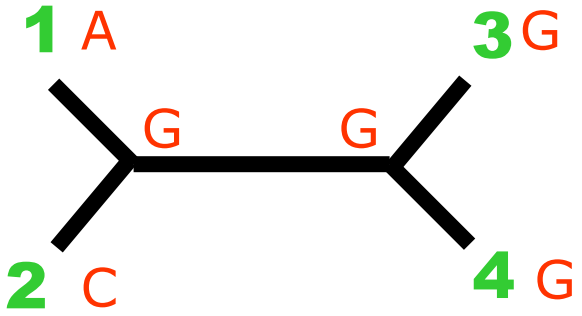
site 1 2 3 4 5

Species 1- A T A T A
Species 2- A T C G C
Species 3- G C A G G
Species 4- G C C G G

$$\begin{bmatrix} p_{AA} & \alpha & \alpha & \alpha \\ \alpha & p_{CC} & \alpha & \alpha \\ \alpha & \alpha & p_{GG} & \alpha \\ \alpha & \alpha & \alpha & p_{TT} \end{bmatrix}$$

$p_{AA} = 1-3\alpha$

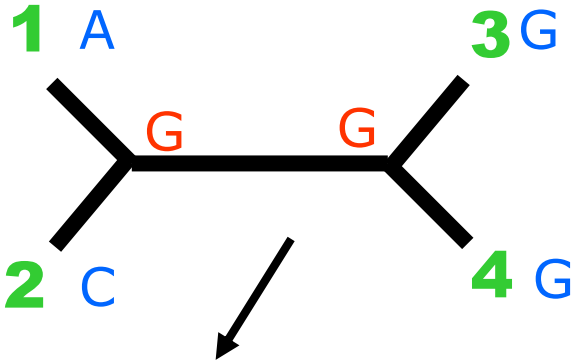
Maximum likelihood



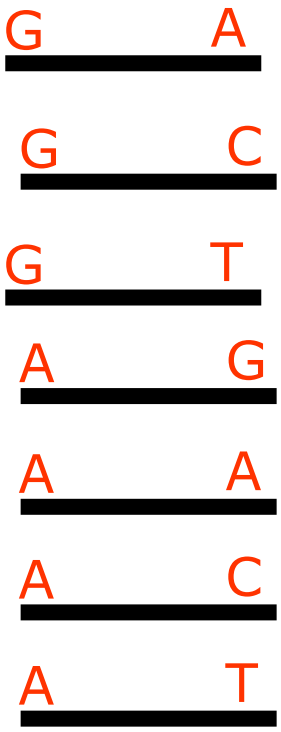
site	1	2	3	4	5
Species 1-	A	T	A	T	A
Species 2-	A	T	C	G	C
Species 3-	G	C	A	G	G
Species 4-	G	C	C	G	G

$$\alpha^2 \times (1-3\alpha)^3$$

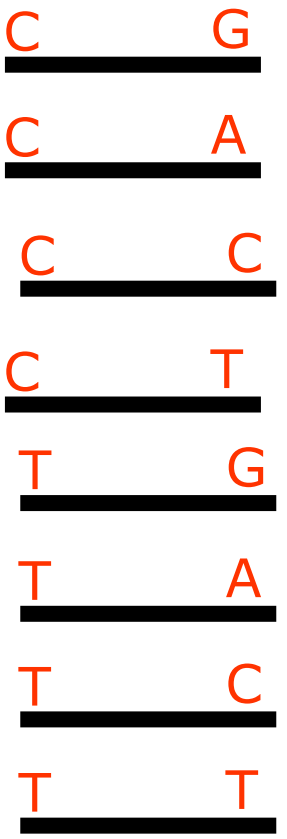
Calcula-se a probabilidade de cada combinação para um dado site e somam-se todas



$$\alpha^2 \times (1-3\alpha)^3$$



$$\alpha^3 \times (1-3\alpha)^2$$



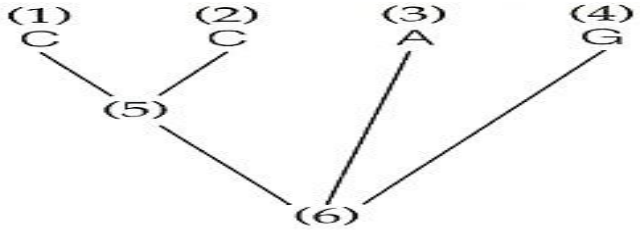
Maximum likelihood

(A)

	1							j							N
(1)	C	…	G	G	A	C	A	C	G	T	T	T	A	…	C
(2)	C	…	A	G	A	C	A	C	C	T	C	T	A	…	C
(3)	C	…	G	G	A	T	A	A	G	T	C	T	A	…	C
(4)	C	…	G	G	A	T	A	G	C	C	T	A	G	…	C

(B)

(C)



Somam-se as combinações para cada site

(D)

$$L_{(j)} = \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ \backslash & / & / & / \\ & A & & \\ & \backslash & / & \\ & & A & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & \backslash & / & / \\ & & C & \\ & & \backslash & / \\ & & & A \end{array}\right)$$

$$+ \quad \ldots \quad + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & \backslash & / & / \\ & G & & \\ & \backslash & / & \\ & & C & \end{array}\right)$$

$$+ \quad \ldots \quad + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & \backslash & / & / \\ & T & & \\ & \backslash & / & \\ & & T & \end{array}\right)$$

Repete-se para todos os sites

(E)

$$L = L_{(1)} \cdot L_{(2)} \cdot \ldots \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

(F)

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}$$

A soma dos log likelihoods dá-nos o log likelihood de uma dada árvore

Maximum likelihood

Repete-se o cáculo para *todas* as árvores e escolhe-se aquela

com maior probabilidade de ocorrer.

Searching methods para escolher as árvores que apresentam valores elevados de probabilidade