Introduction to Bioinformatics for Computer Scientists

Lecture 9b

Likelihood

- Given:
  - MSA
  - Tree topology with branch lengths
  - Model
  - We can calculate $P_{x->z}(b)$ for a branch length (or time) *b*

Likelihood

- $L(T|D) = P(D|T)$

Probability that the tree generated the data (generating process)

Likelihood

- $L(T|D) = P(D|T)$

Likelihood of the tree, given the data

Likelihood

- L(T|D) = P(D|T)

Likelihood: 10 coin flips → 10 heads
What's the likelihood that the coin is fair?

Probability: Probability of landing heads up 10 times

Likelihood

- $L(T|D) = P(D|T)$
- $L(T|D) = \Pi\ P(s_i|T)$

Alignment site *i*

Likelihood

- $L(T|D) = P(D|T)$
- $L(T|D) = \Pi \, P(s_i|T)$

Alignment site *i*

What is problematic about this term?

Likelihood

- $L(T|D) = P(D|T)$
- $L(T|D) = \Pi\ P(s_i|T)$
- $\log(L(T|D)) = \Sigma \log(P(s_i|T))$

Likelihood

- $L(T|D) = P(D|T)$
- $L(T|D) = \Pi P(s_i|T)$
- $\log(L(T|D)) = \Sigma \log(P(s_i|T))$

This is the model
1. Tree topology
2. Branch lengths
3. Model of nucleotide substitution
   → generally lumped into parameter vector **Θ**: $L(\Theta|D)$

Likelihood

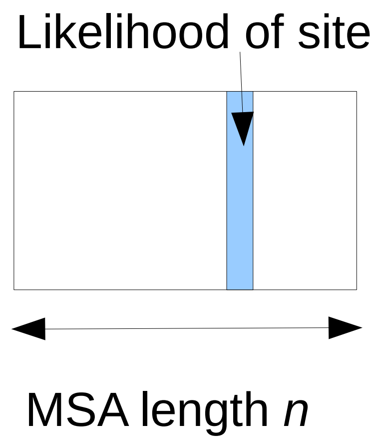
- $L(T|D) = P(D|T)$
- $L(T|D) = \Pi\ P(s_i|T)$
- $\log(L(T|D)) = \Sigma \log(P(s_i|T))$

This is the model
1. Tree topology
2. Branch lengths
3. Model of nucleotide substitution
   → generally lumped into parameter vector **Θ**: $L(\Theta|D)$

How do we compute this?

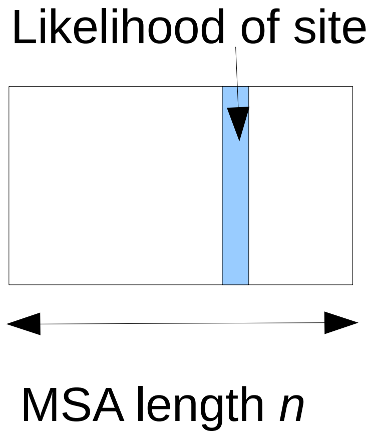
Likelihood of a Tree

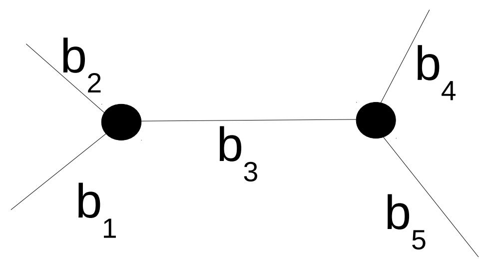
- We assume that sites evolve independently



Likelihood of a Tree

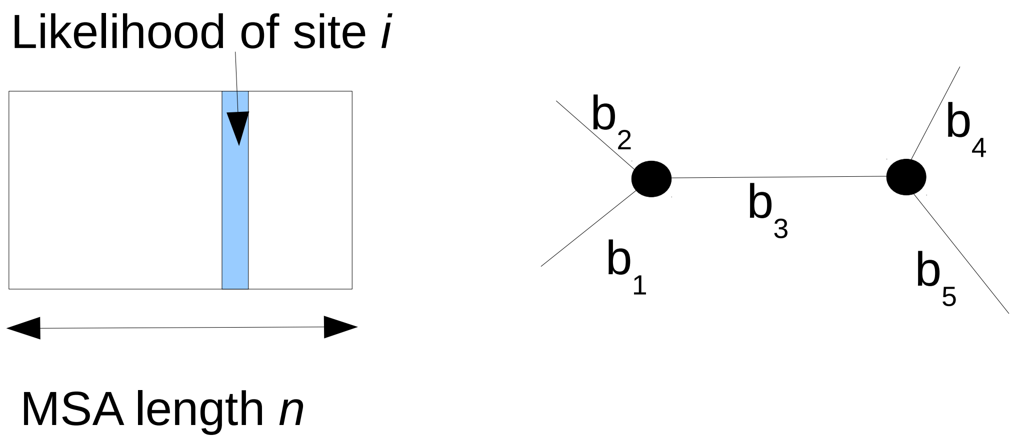
- We assume that sites evolve independently





Likelihood of a Tree

- We assume that sites evolve independently

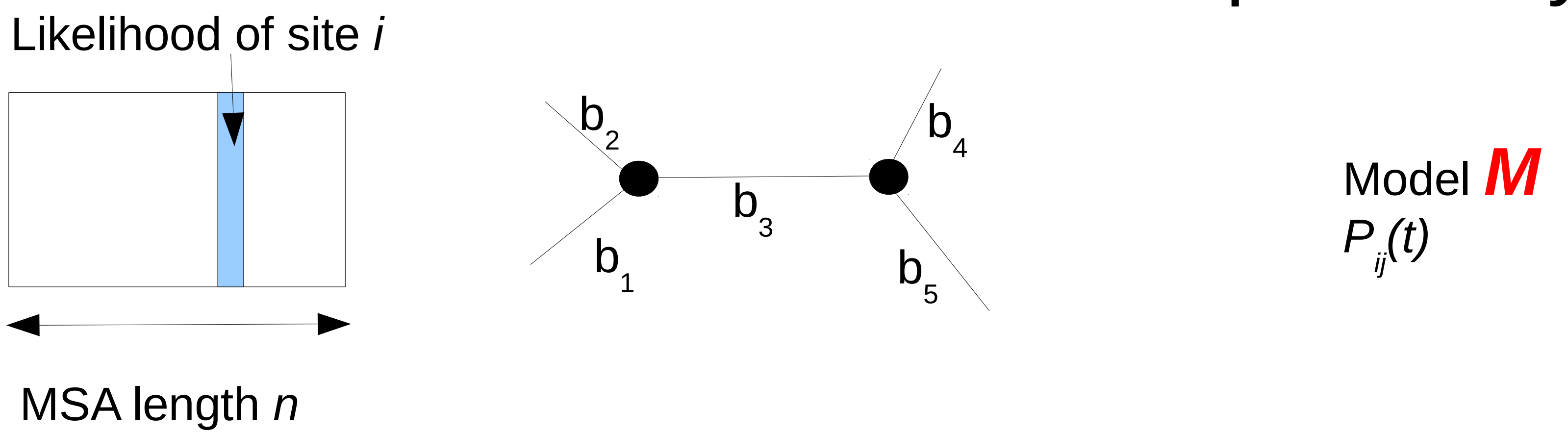


Model ***M***

$P_{ij}(t)$

Likelihood of a Tree

- We assume that sites evolve independently

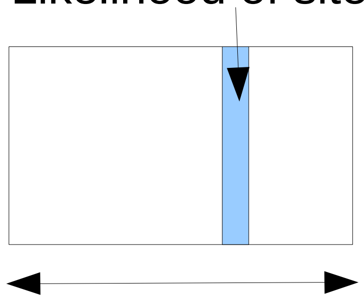


- Overall likelihood: $L := \Pi L_i$

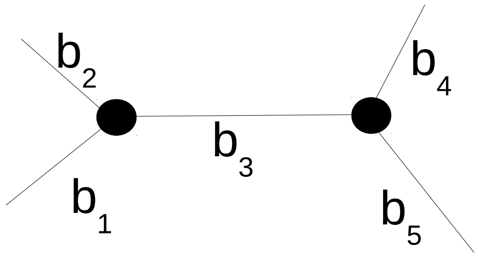
Likelihood of a Tree

- We assume that sites evolve independently

Likelihood of site *i*



MSA length *n*



Model ***M***

$P_{ij}(t)$

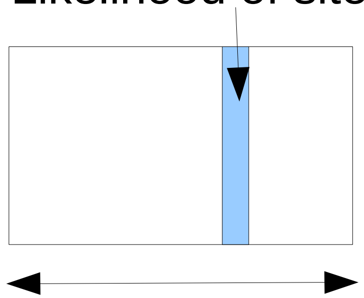
- Overall likelihood: $L := \Pi L_i$
- $P_{ij}(t)$ i,j *in* $\{A, C, G, T\}$

Branch length/time

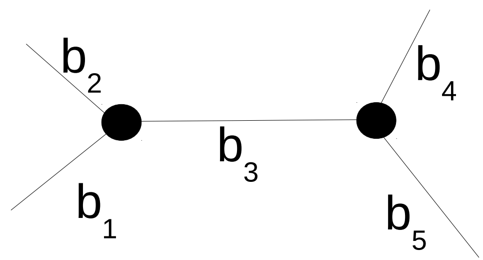
Likelihood of a Tree

- We assume that sites evolve independently

Likelihood of site *i*



MSA length *n*



Model ***M***

$P_{ij}(t)$

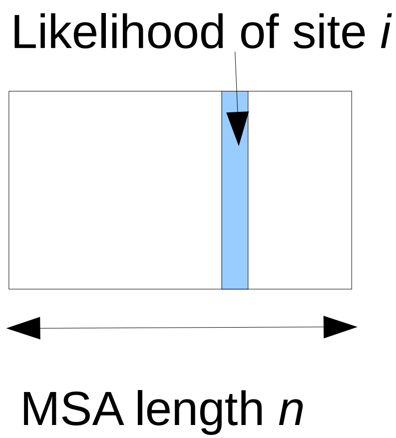
- Overall likelihood: $L := \Pi L_i$
- $P_{ij}(t)$ *i,j in {A, C, G, T}*

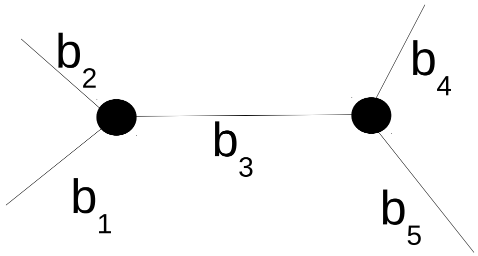
→ Probability of being in state *j* after time *t*

→ We assume that $P_{ij}(t)$ is a Markov Process

Likelihood of a Tree

- We assume that sites evolve independently





Model ***M***
$P_{ij}(t)$

- Overall likelihood: $L := \Pi L_i$
- $P_{ij}(t)$ *i,j in {A, C, G, T}*

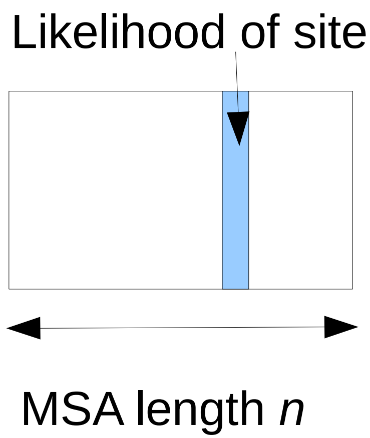
→ Probability of being in state *j* after time *t*

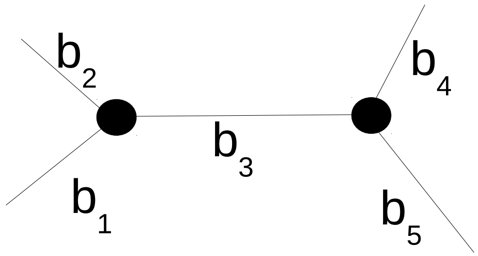
→ We assume that $P_{ij}(t)$ is a Markov Process

- Equilibrium frequency vector $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$

Likelihood of a Tree

- We assume that sites evolve independently





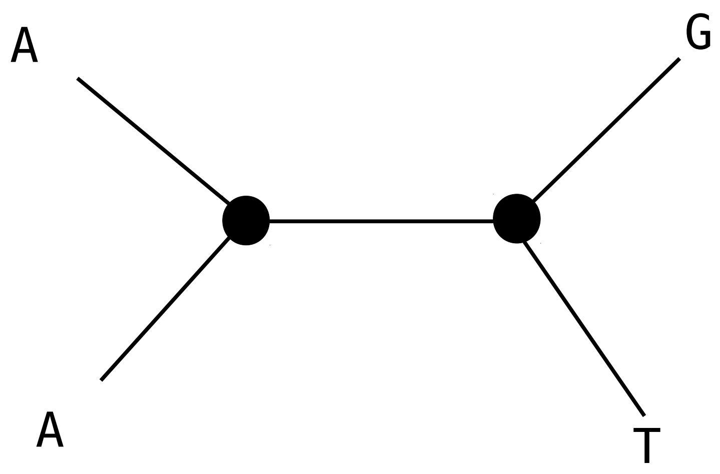
Model ***M***
$P_{ij}(t)$

- Overall likelihood: $L := \Pi L_i$
- $P_{ij}(t)$ *i,j in {A, C, G, T}*

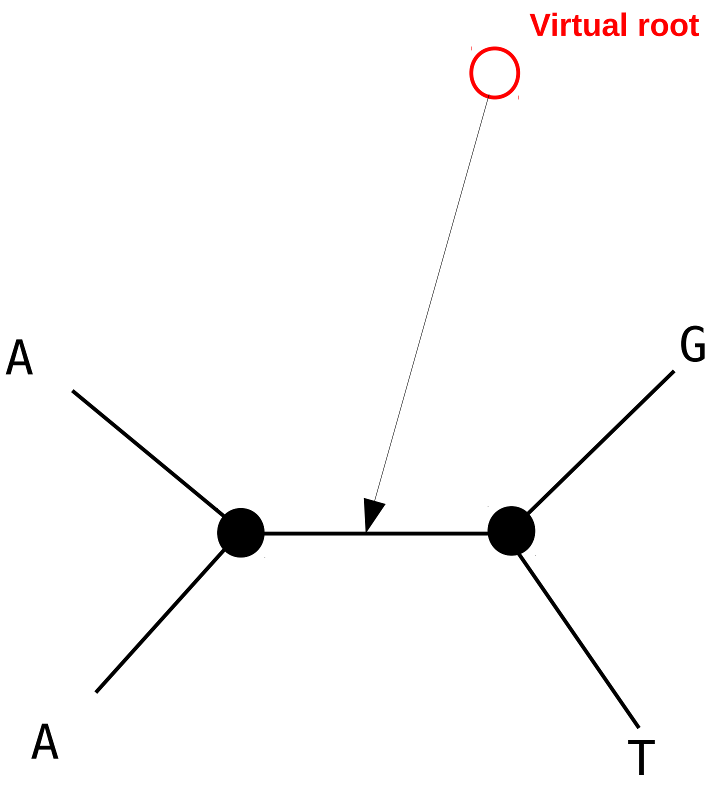
  → Probability of being in state *j* after time *t*

  → We assume that $P_{ij}(t)$ is a Markov Process
- Equilibrium frequency vector $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$
- **Time reversibility:** $\pi_i P_{ij}(t) = \pi_j P_{ij}(t)$

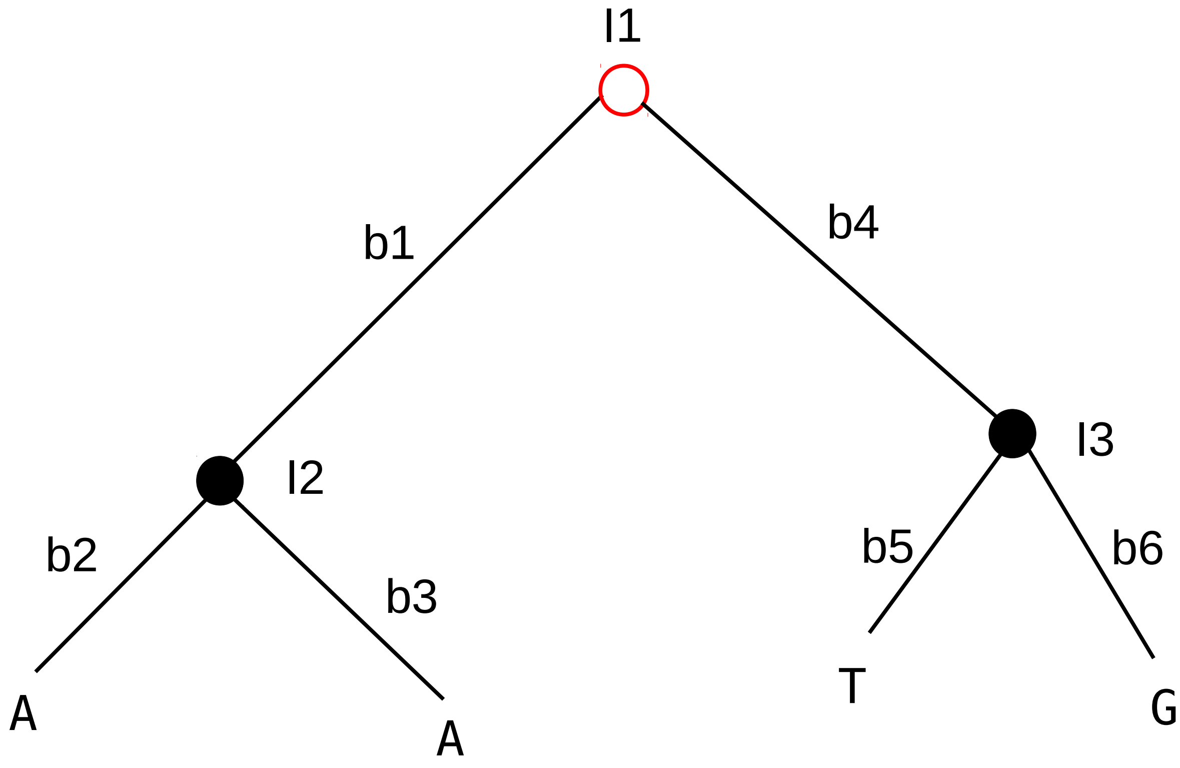
What's the likelihood of this tree?



What's the likelihood of this tree?

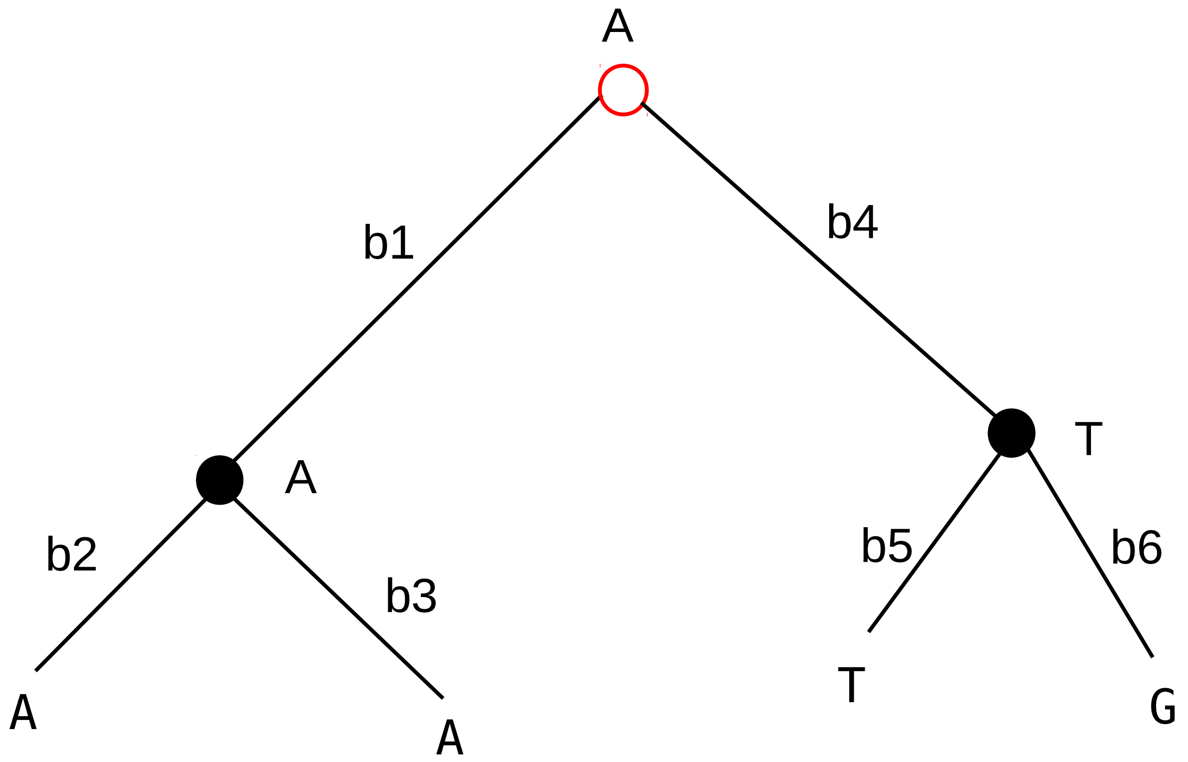


What's the likelihood of this tree?

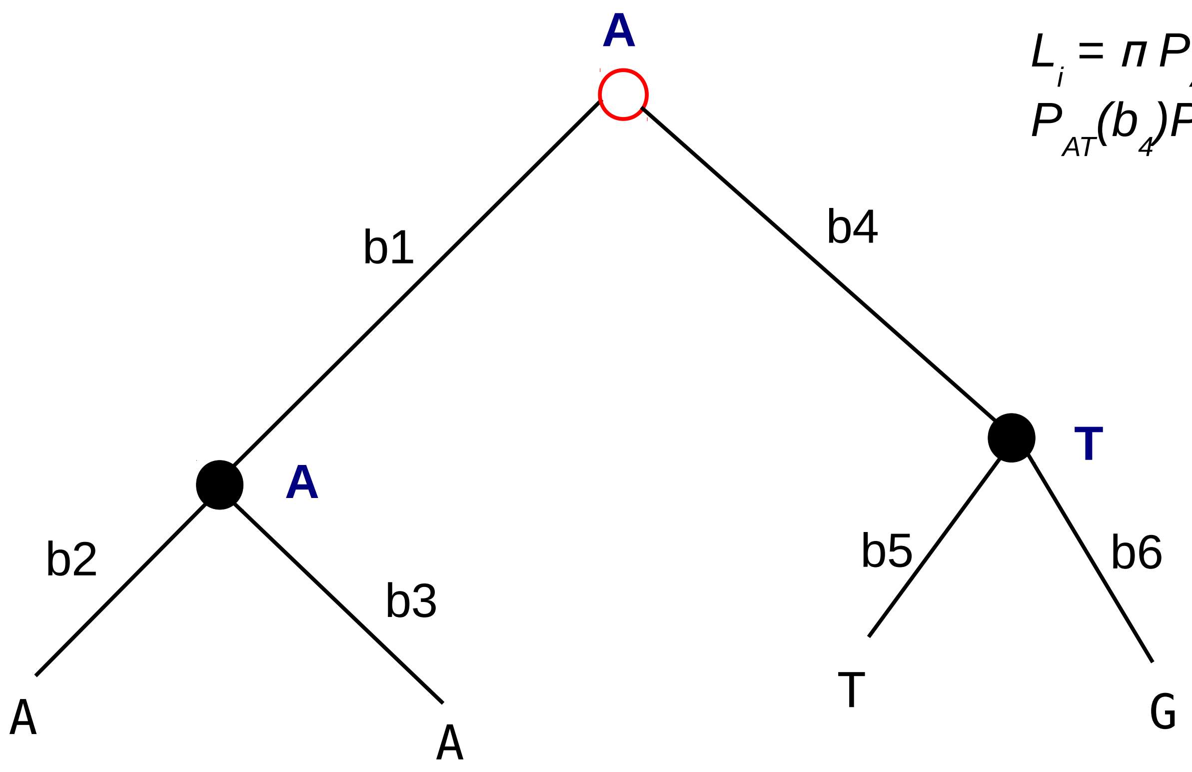


What's the likelihood of this tree?

Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?



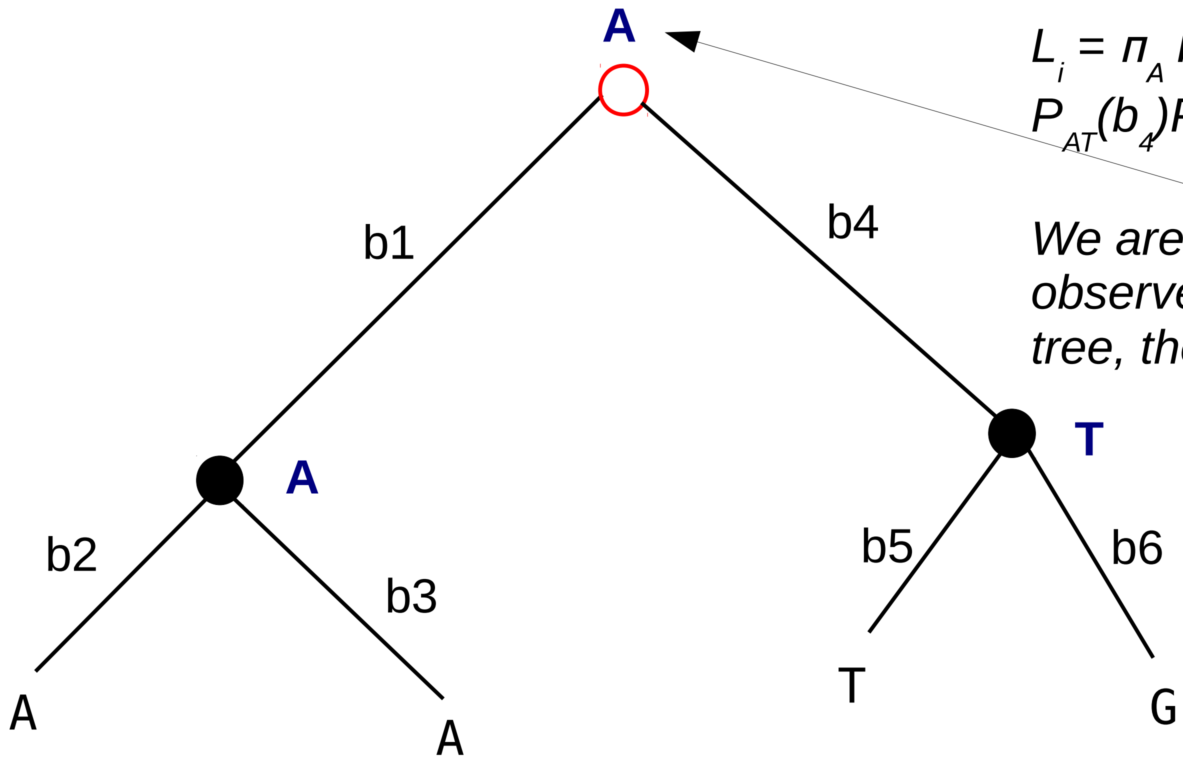
What's the likelihood of this tree?



Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

$L_i = \pi P_{AA}(b_1)\, P_{AA}(b_2)\, P_{AA}(b_3)$
$P_{AT}(b_4) P_{TT}(b_5)\, P_{TG}(b_6)$

What's the likelihood of this tree?

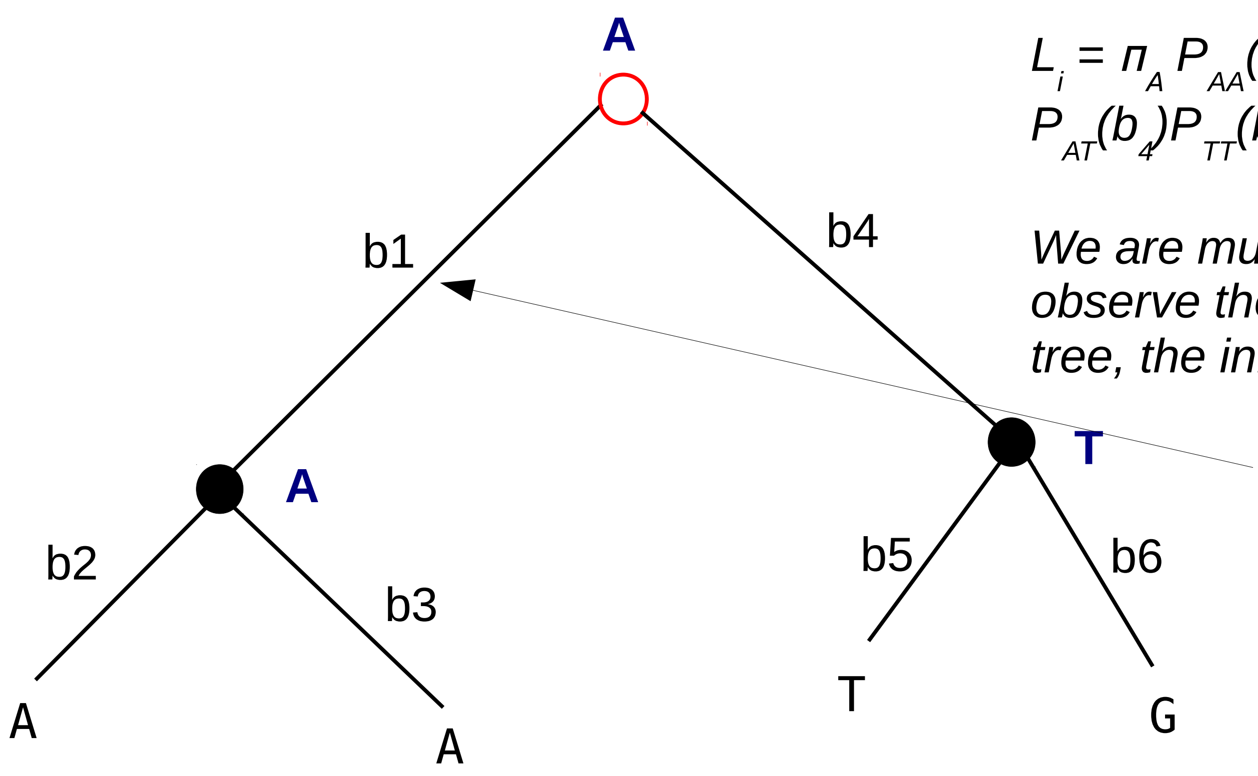


Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

$L_i = \pi_A \, P_{AA}(b_1) \, P_{AA}(b_2) \, P_{AA}(b_3) \, P_{AT}(b_4) P_{TT}(b_5) \, P_{TG}(b_6)$

We are multiplying here, because to observe the data at the tips, given the tree, the initial state must be **A** π_A

What's the likelihood of this tree?



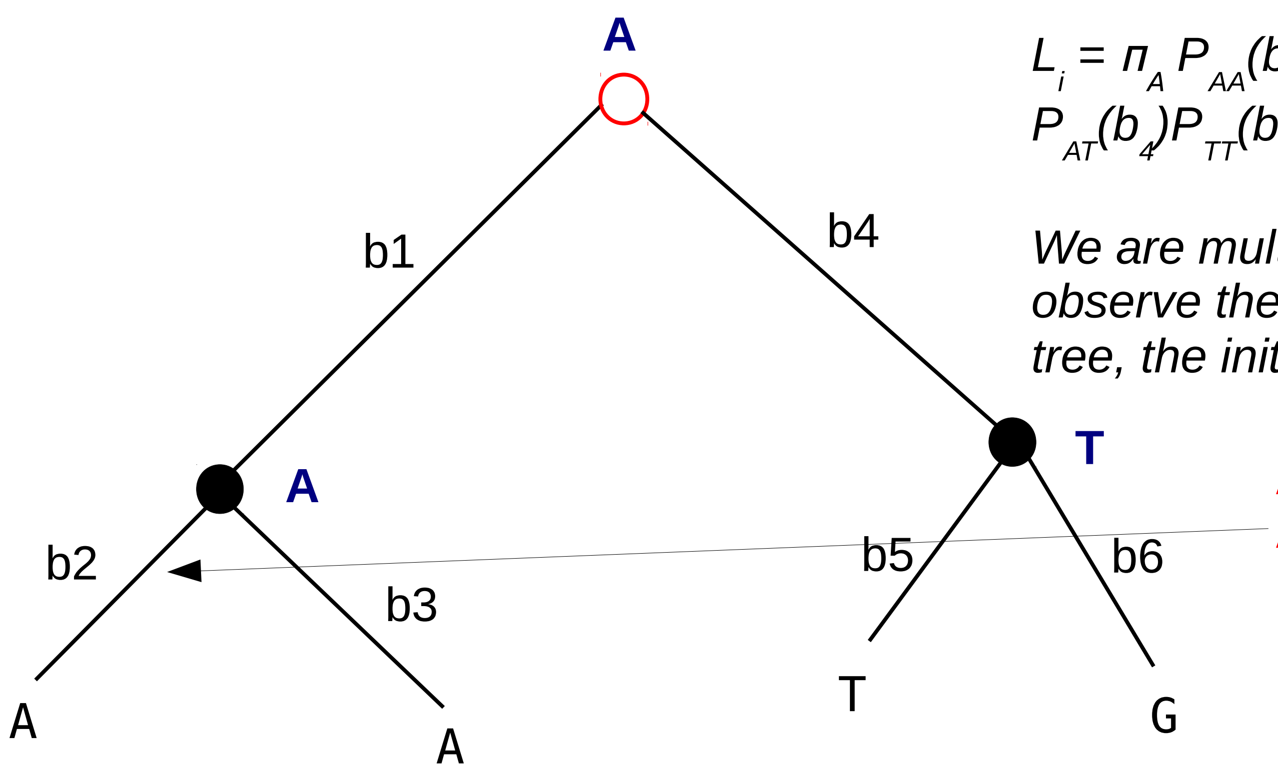
Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

$L_i = \pi_A \, P_{AA}(b_1) \, P_{AA}(b_2) \, P_{AA}(b_3)$
$P_{AT}(b_4) P_{TT}(b_5) \, P_{TG}(b_6)$

We are multiplying here, because to observe the data at the tips, given the tree, the initial state must be **A** π_A

AND then this happened

What's the likelihood of this tree?



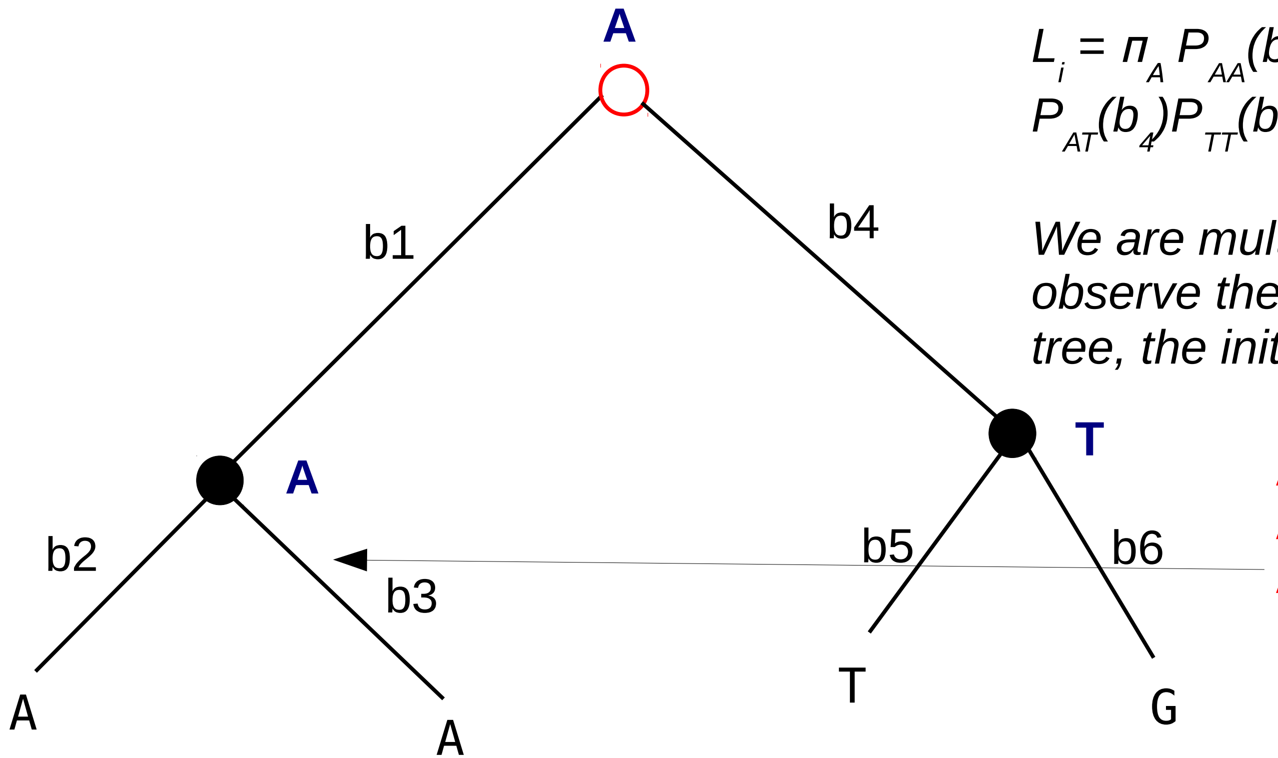
Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

$L_i = \pi_A \, P_{AA}(b_1) \, P_{AA}(b_2) \, P_{AA}(b_3)$
$P_{AT}(b_4) P_{TT}(b_5) \, P_{TG}(b_6)$

We are multiplying here, because to observe the data at the tips, given the tree, the initial state must be **A** π_A

AND then this happened
AND this

What's the likelihood of this tree?



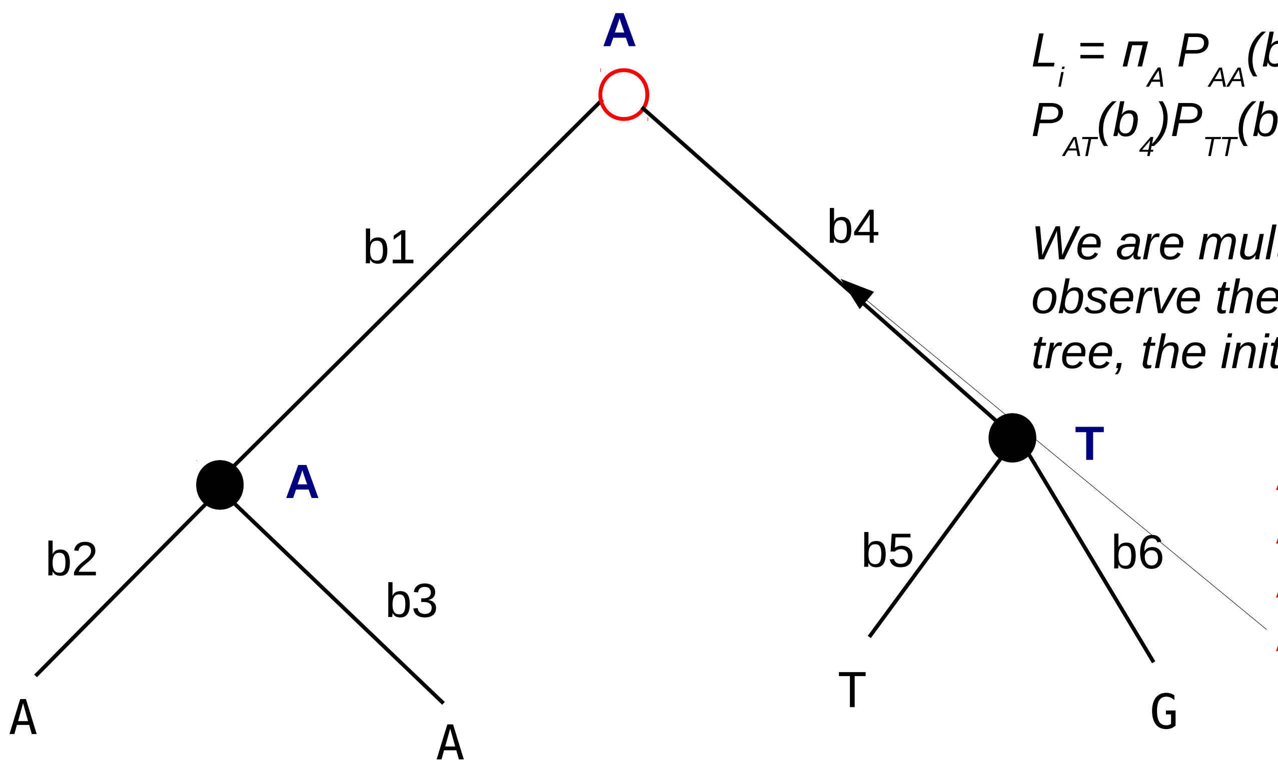
Assume the inner states are given!
What is the likelihood of the tree if we Interpret it as **Markov** diagram?

$L_i = \pi_A P_{AA}(b_1) P_{AA}(b_2) P_{AA}(b_3) P_{AT}(b_4) P_{TT}(b_5) P_{TG}(b_6)$

We are multiplying here, because to observe the data at the tips, given the tree, the initial state must be **A** π_A

AND then this happened
AND this
AND this

What's the likelihood of this tree?



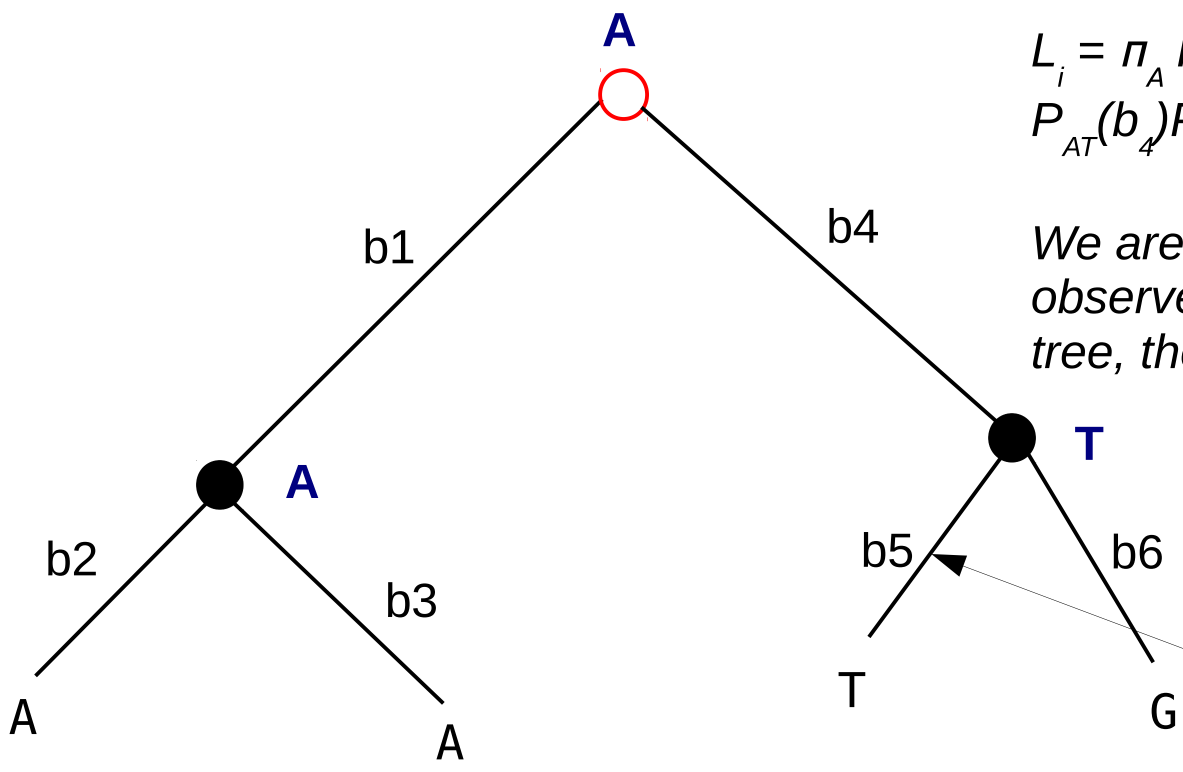
Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

$L_i = \pi_A\, P_{AA}(b_1)\, P_{AA}(b_2)\, P_{AA}(b_3)$
$P_{AT}(b_4) P_{TT}(b_5)\, P_{TG}(b_6)$

We are multiplying here, because to observe the data at the tips, given the tree, the initial state must be **A** π_A

AND then this happened
AND this
AND this
AND this

What's the likelihood of this tree?



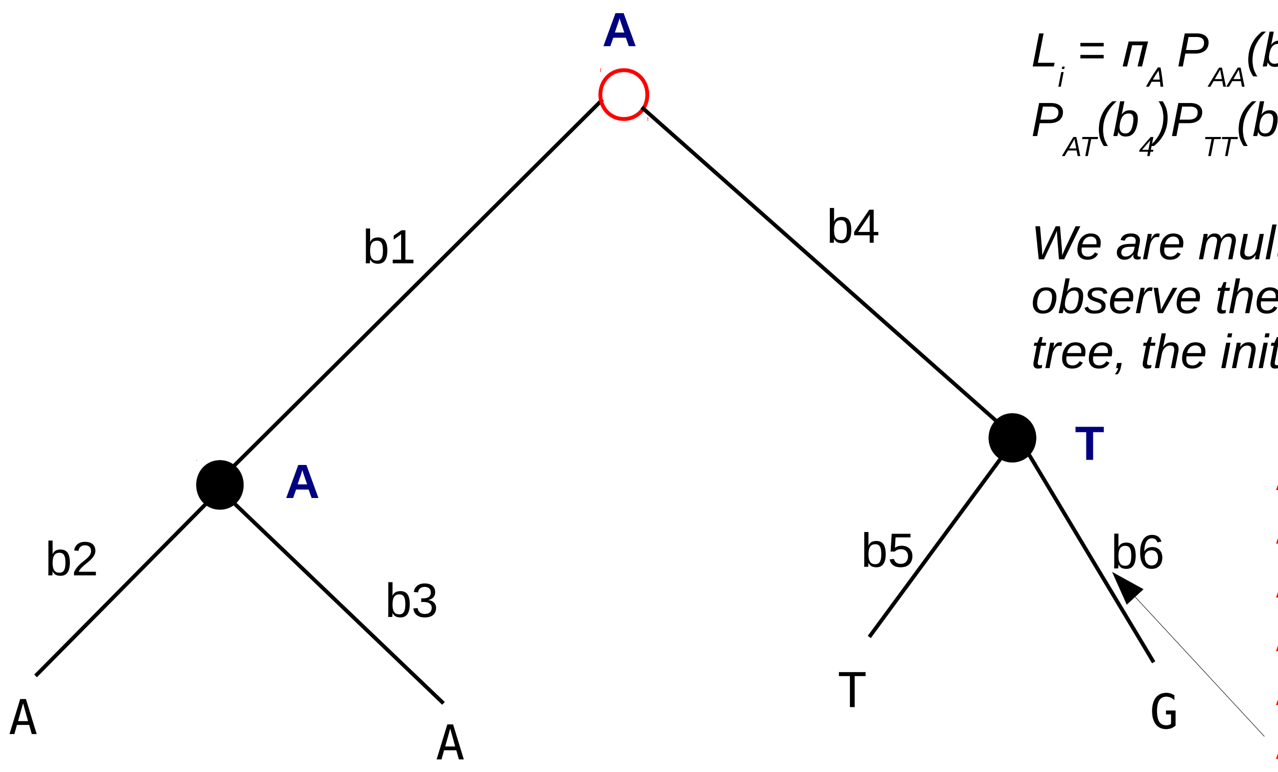
Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

$L_i = \pi_A P_{AA}(b_1) P_{AA}(b_2) P_{AA}(b_3)$
$P_{AT}(b_4) P_{TT}(b_5) P_{TG}(b_6)$

We are multiplying here, because to observe the data at the tips, given the tree, the initial state must be **A** π_A

AND then this happened
AND this
AND this
AND this
AND this

What's the likelihood of this tree?



Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

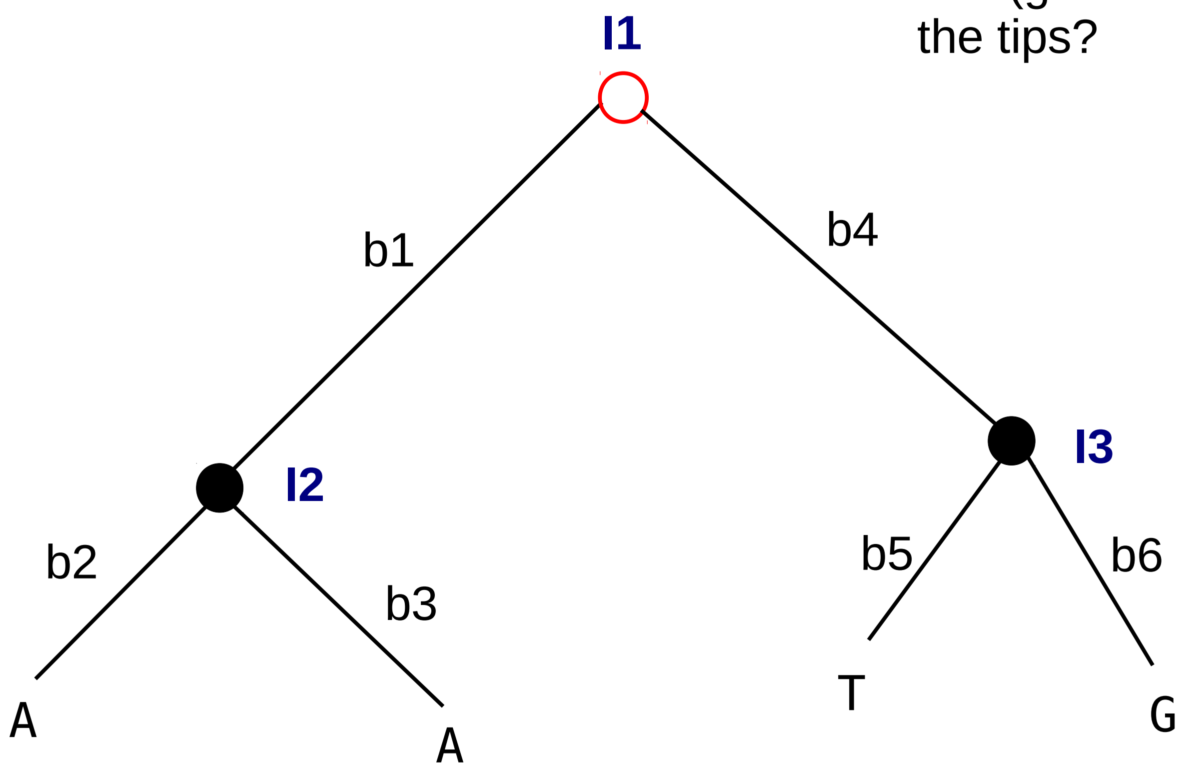
$L_i = \pi_A \, P_{AA}(b_1) \, P_{AA}(b_2) \, P_{AA}(b_3) \, P_{AT}(b_4) P_{TT}(b_5) \, P_{TG}(b_6)$

We are multiplying here, because to observe the data at the tips, given the tree, the initial state must be **A** π_A

AND then this happened
AND this
AND this
AND this
AND this
AND this

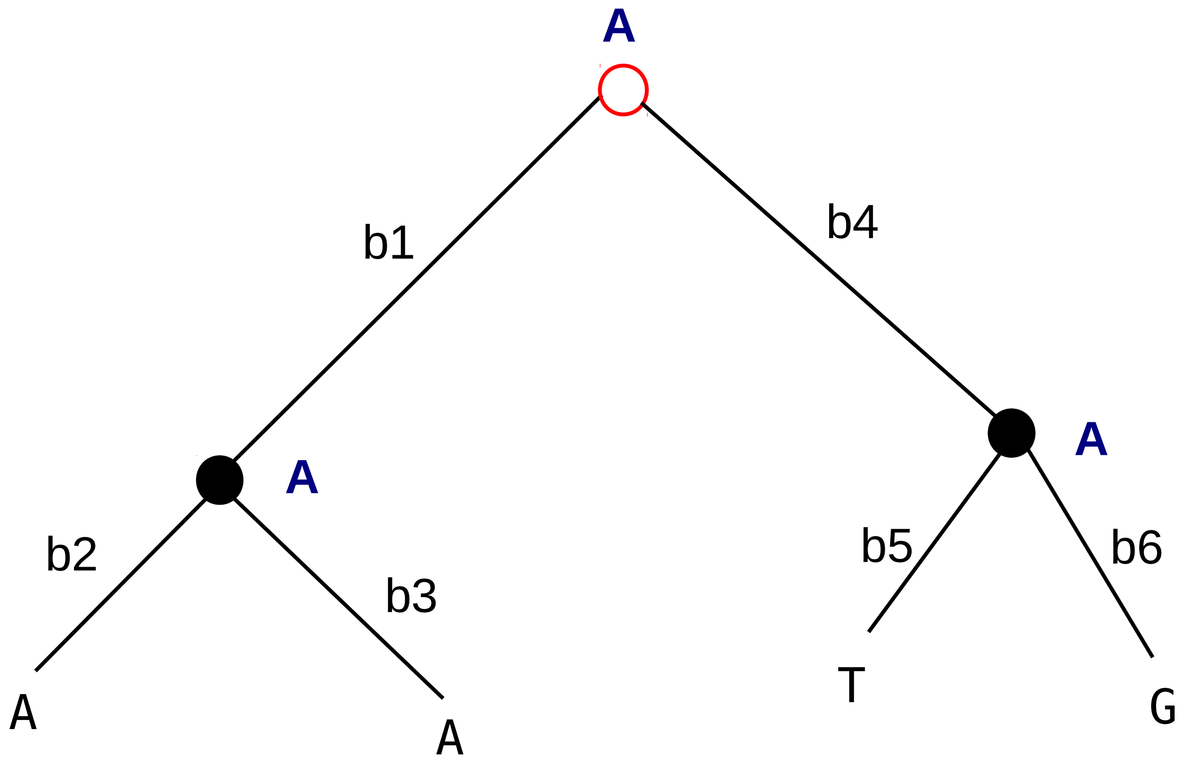
What's the likelihood of this tree?

However, we don't know the inner states :-(
So the question is: What are the possible evolutionary histories that could have given rise (generated) to the data we observe at the tips?



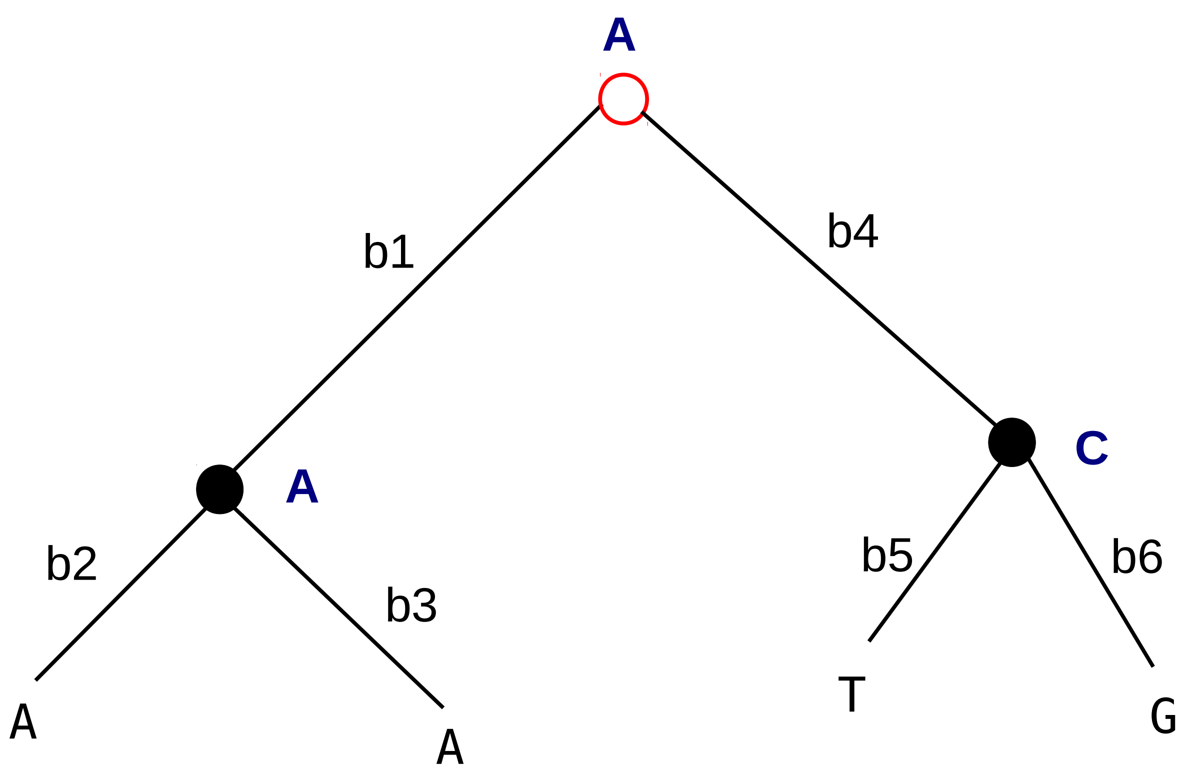
What's the likelihood of this tree?

It could be this



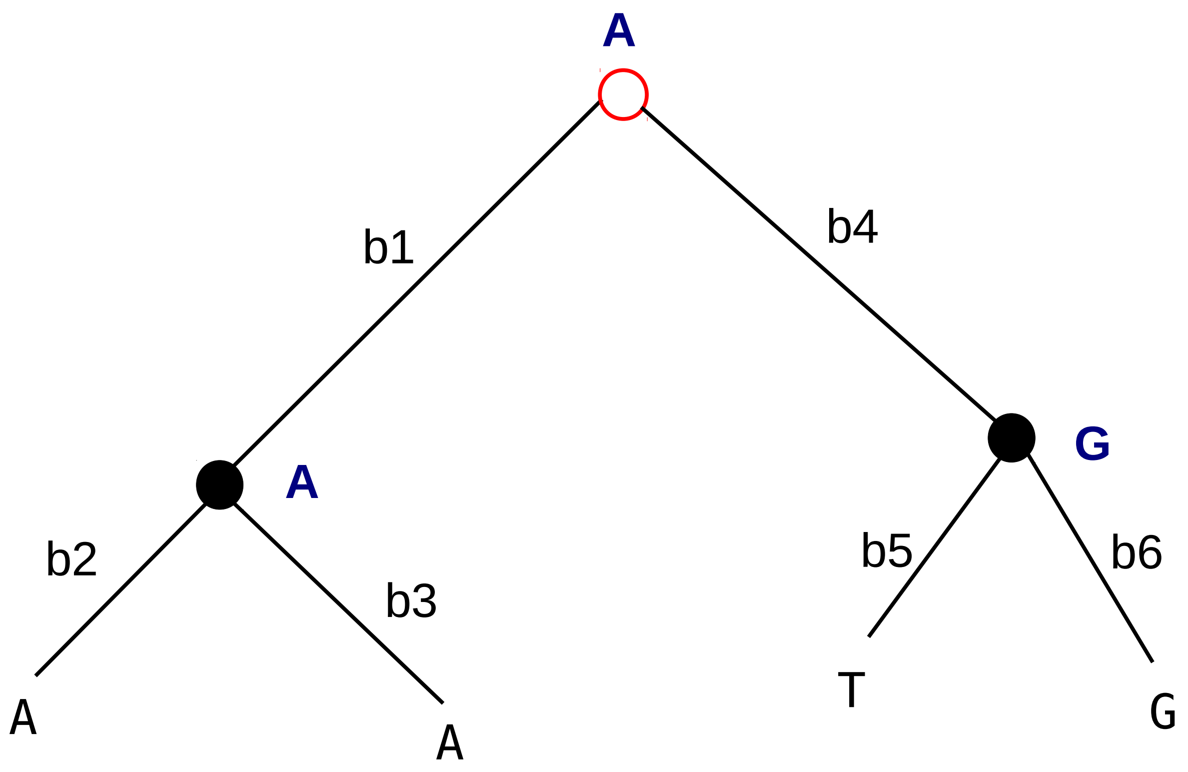
What's the likelihood of this tree?

It could be this
OR this



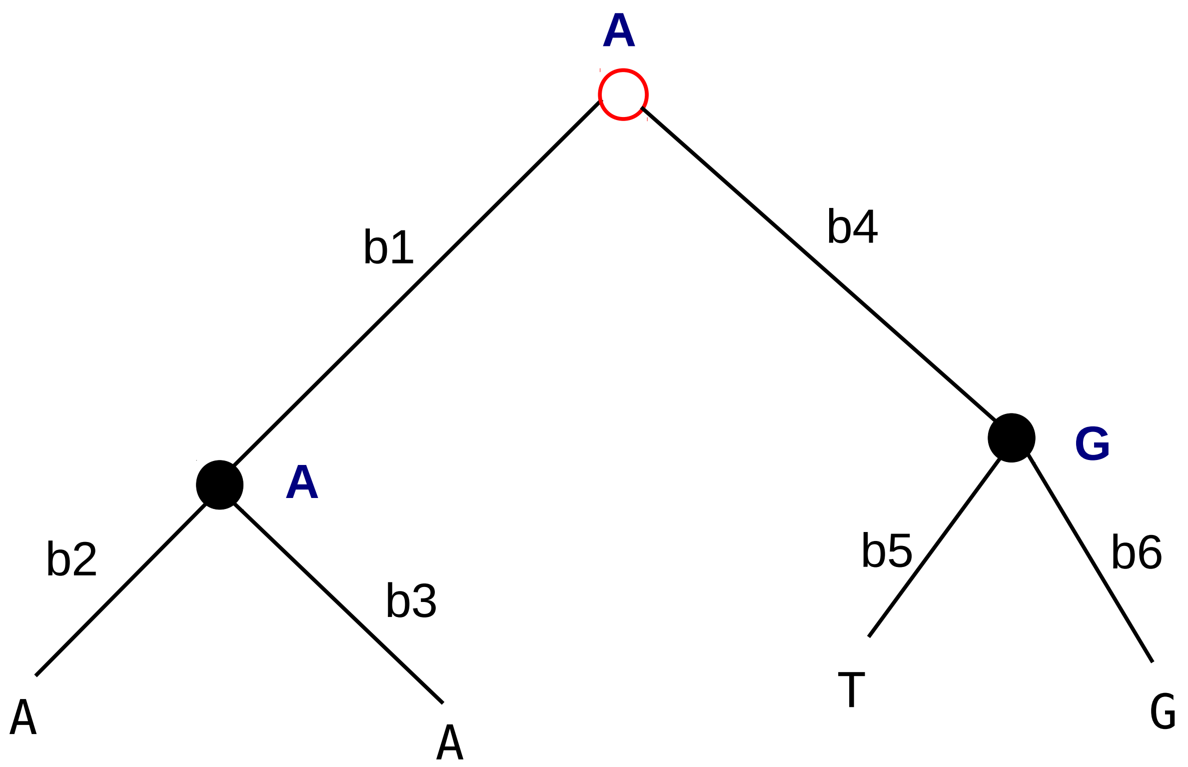
What's the likelihood of this tree?

It could be this
OR this
OR this

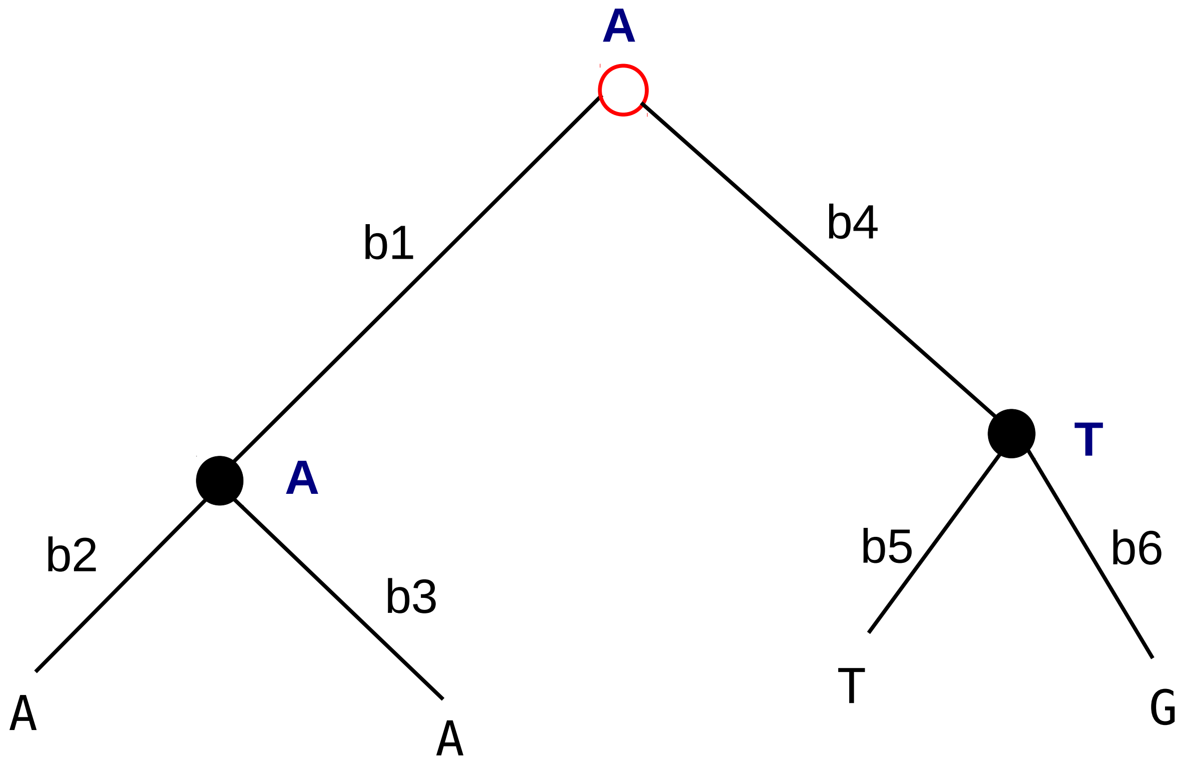


What's the likelihood of this tree?

It could be this
OR this
OR this
OR this

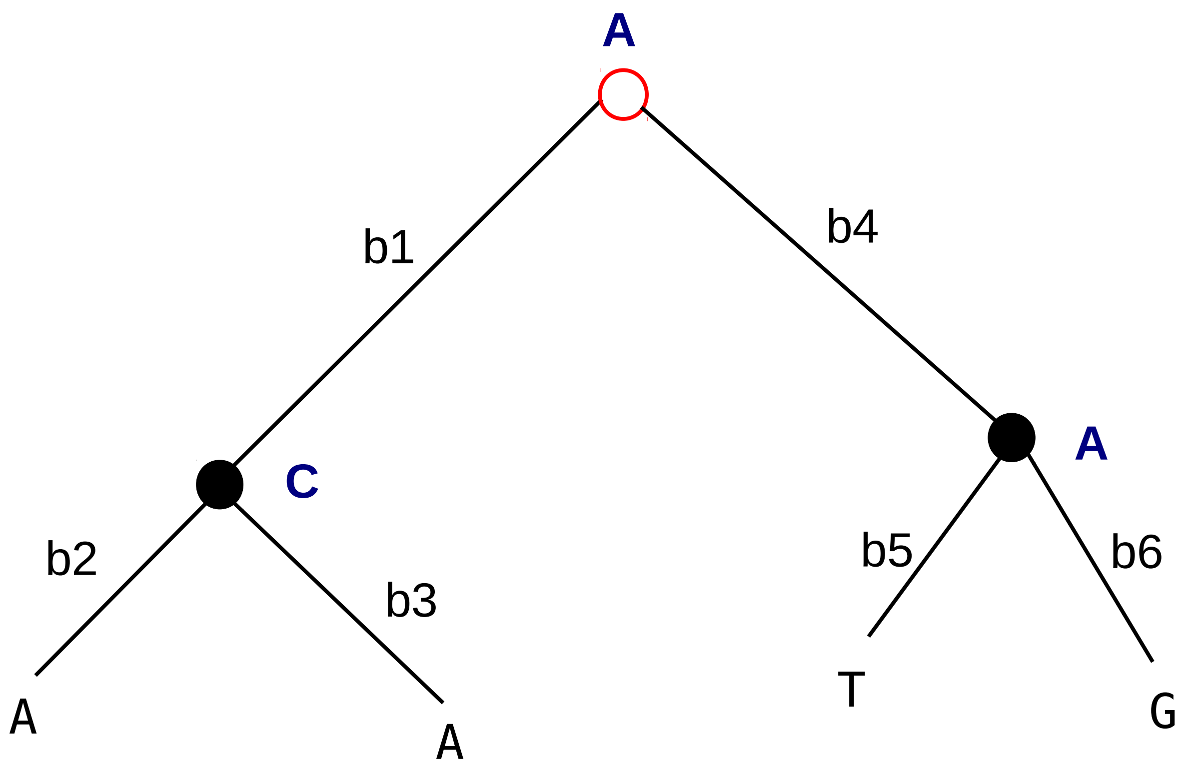


What's the likelihood of this tree?



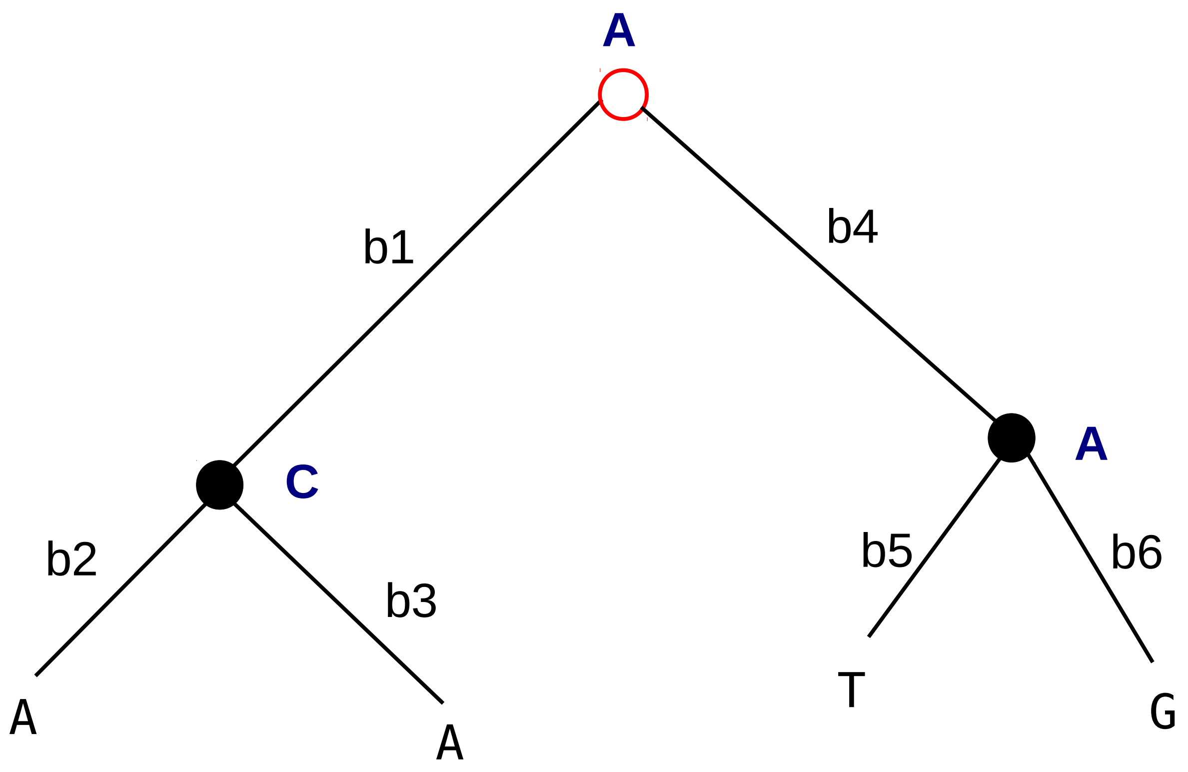
It could be this
OR this
OR this
OR this
OR this

What's the likelihood of this tree?



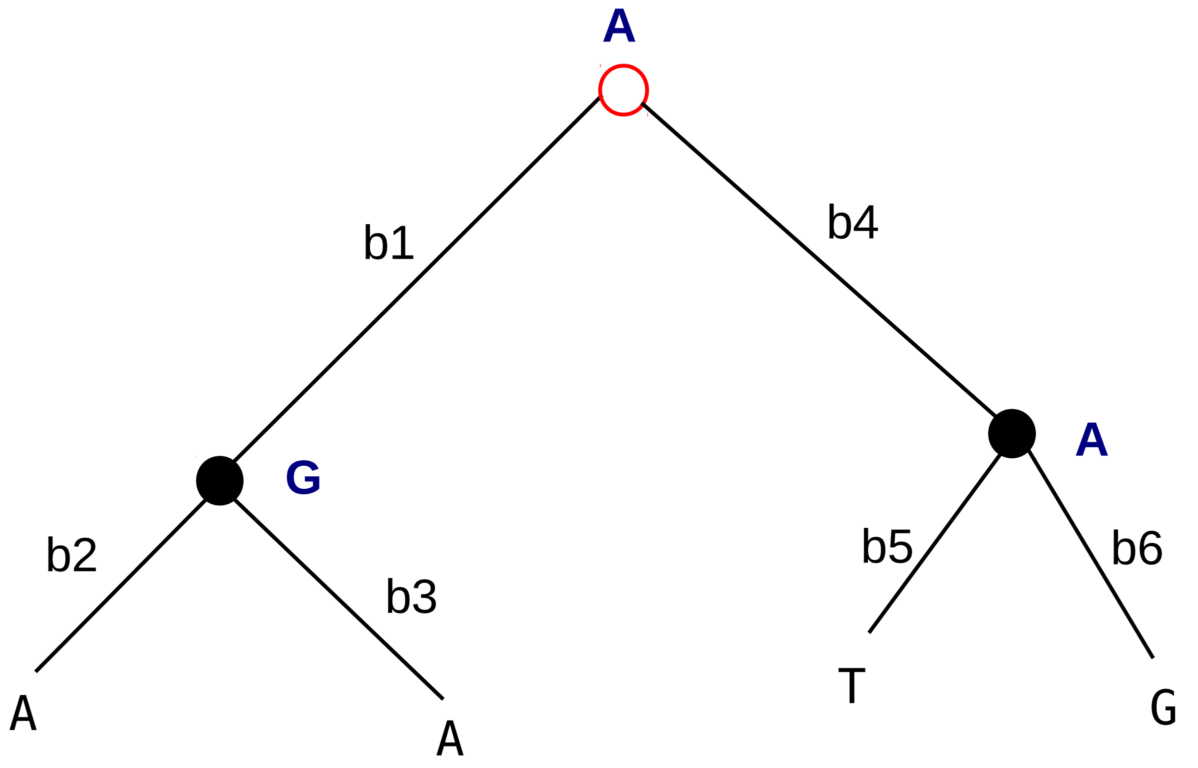
It could be this
OR this
OR this
OR this
OR this
OR this

What's the likelihood of this tree?



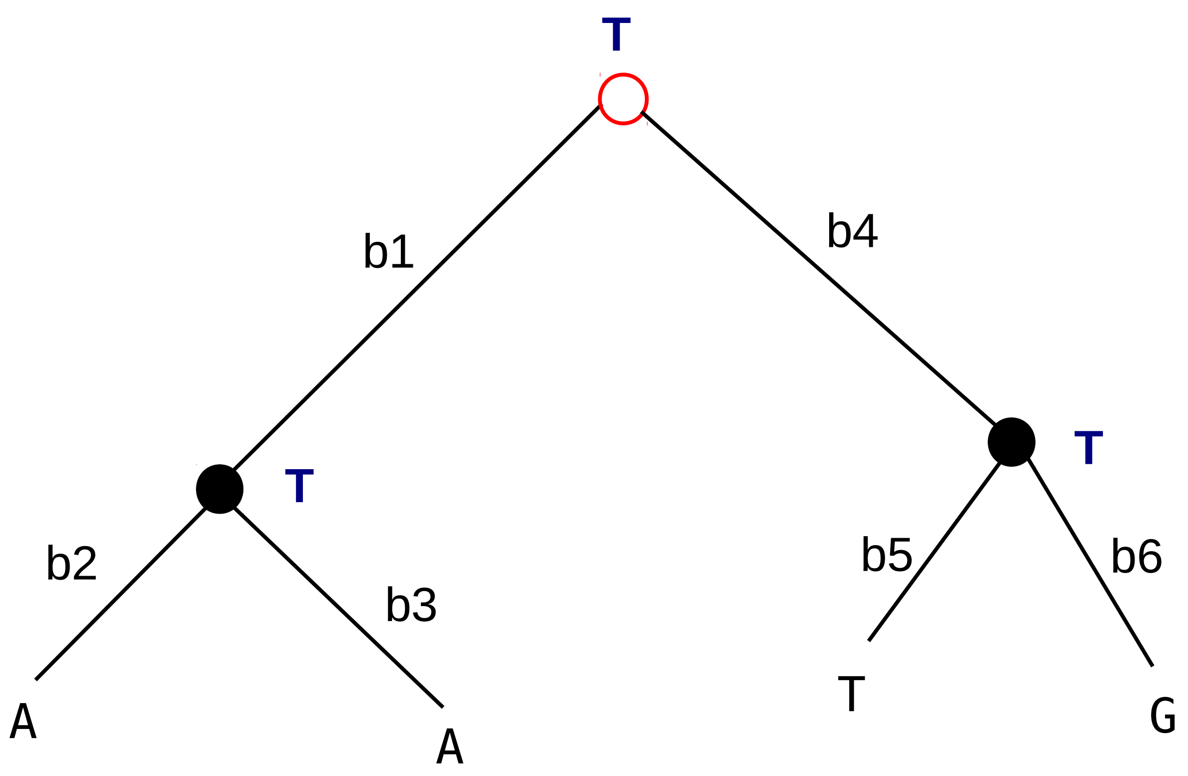
It could be this
OR this
OR this
OR this
OR this
OR this
OR this

What's the likelihood of this tree?



It could be this
OR this
OR this
OR this
OR this
OR this
OR this
OR this

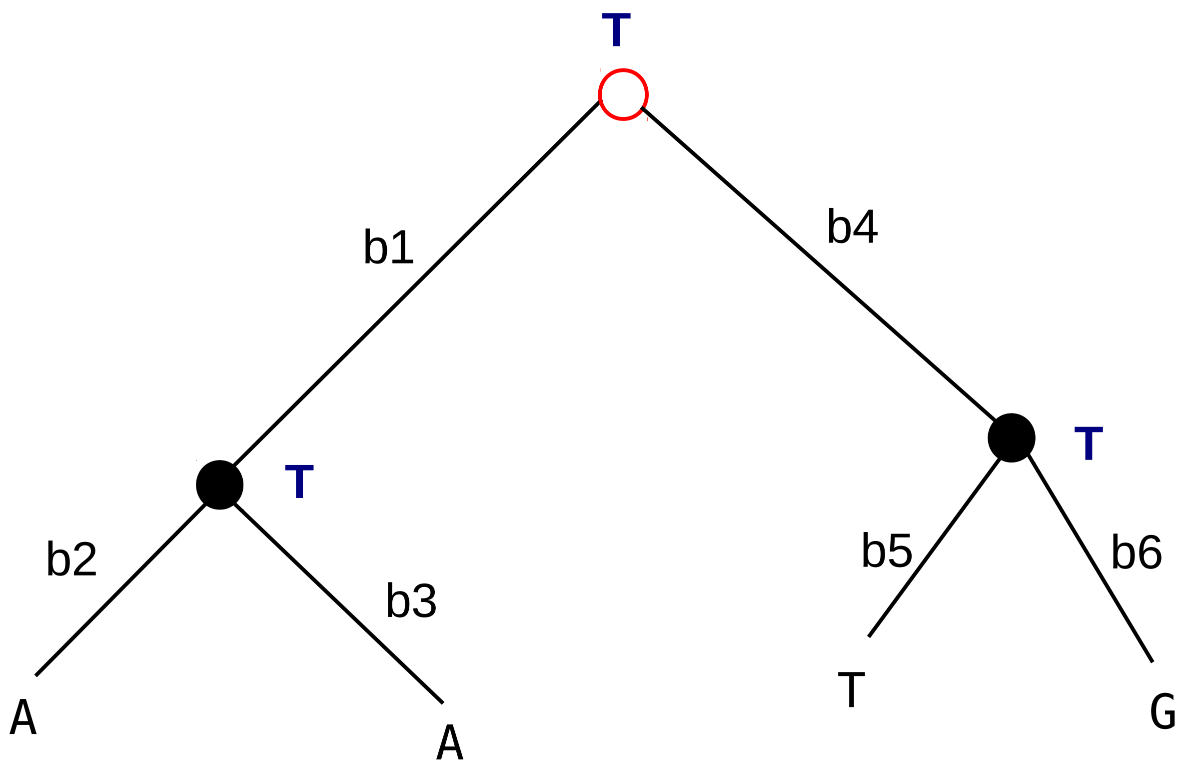
What's the likelihood of this tree?



It could be this
OR this
OR this
OR this
OR this
OR this
OR this
OR this
…
OR this

What's the likelihood of this tree?

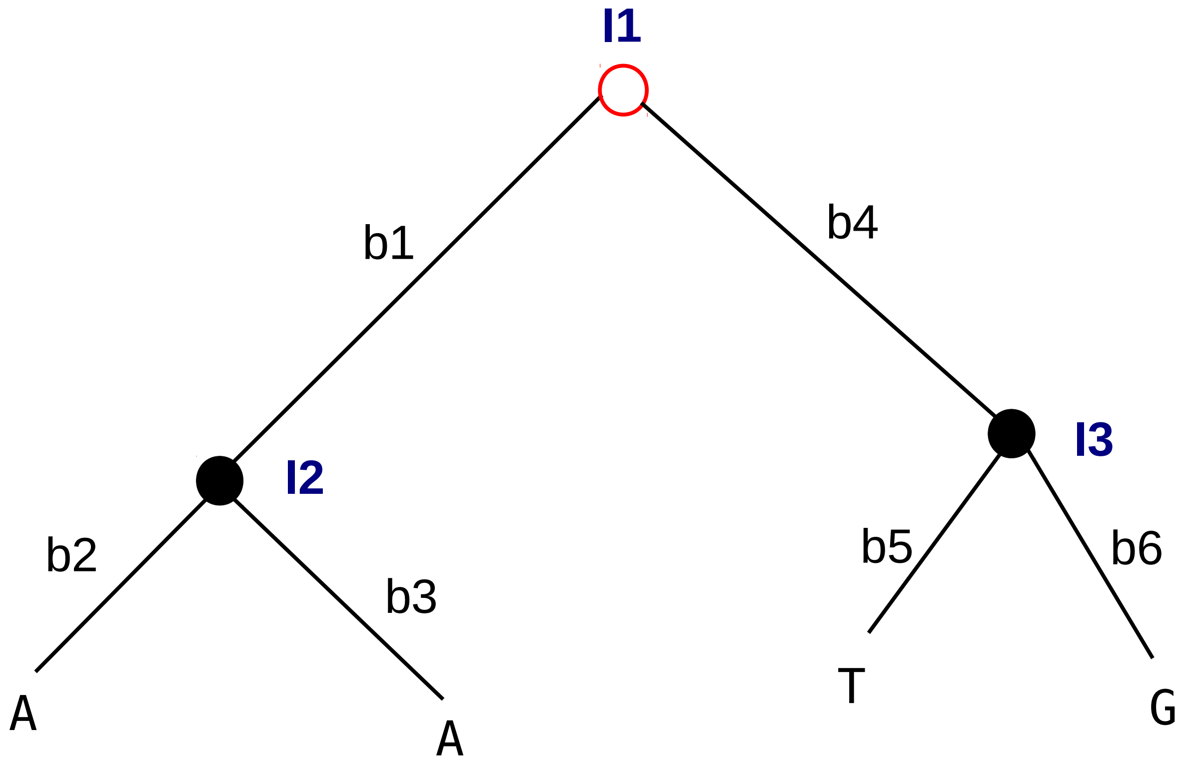
So the likelihood of the tree is the sum (**OR!**) over the likelihoods of all possible assignments of A, C, G, and T (all possible evolutionary histories) to the inner nodes *I1, I2, I3* of the tree.



What's the likelihood of this tree?

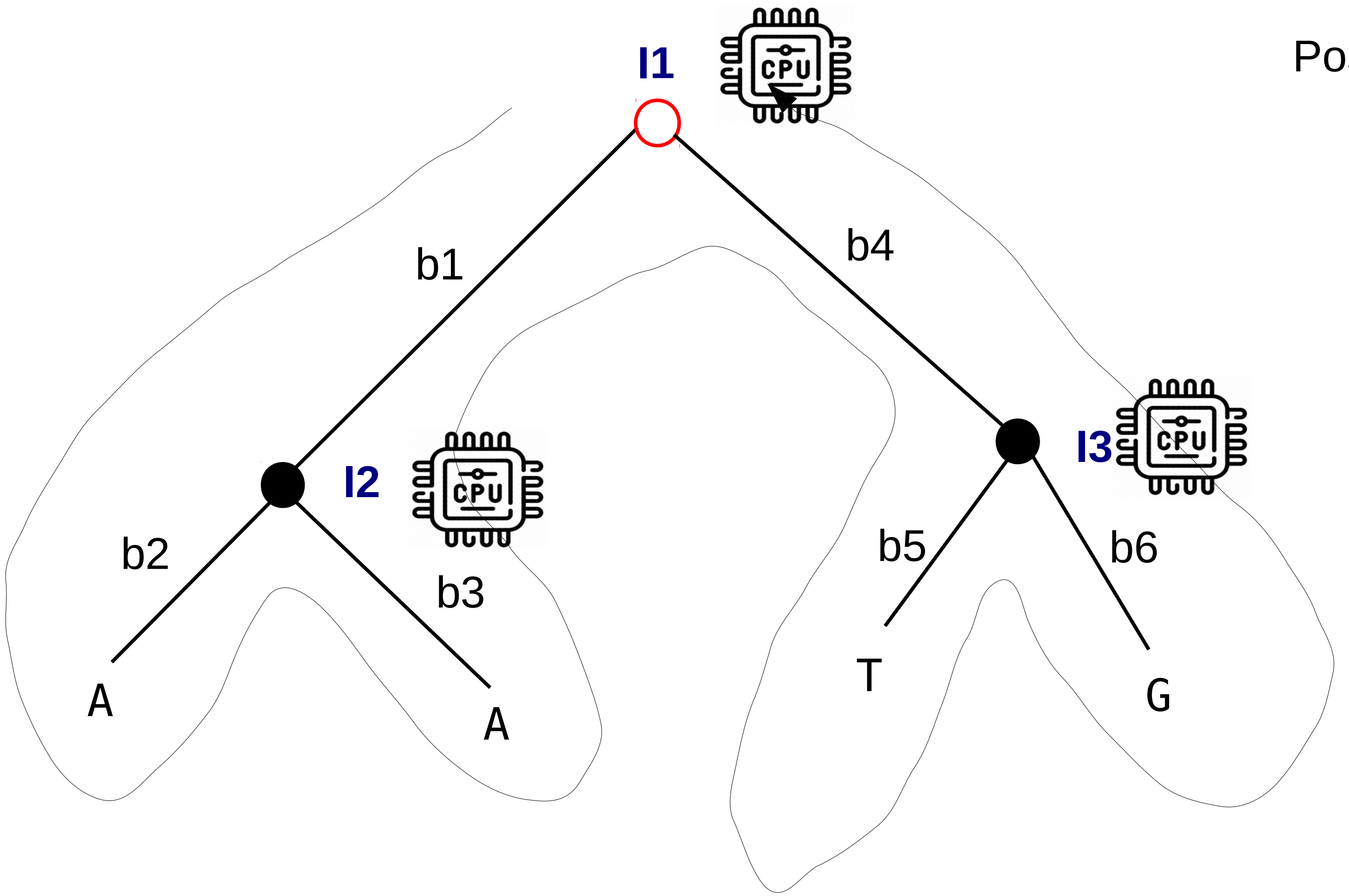
So the likelihood of the tree is the sum (**OR!**) over the likelihoods of all possible assignments of A, C, G, and T (all possible evolutionary histories) to the inner nodes *I1, I2, I3* of the tree.

There are 4 x 4 x 4 possible assignments in our example
→ this sounds very compute-intensive :-(

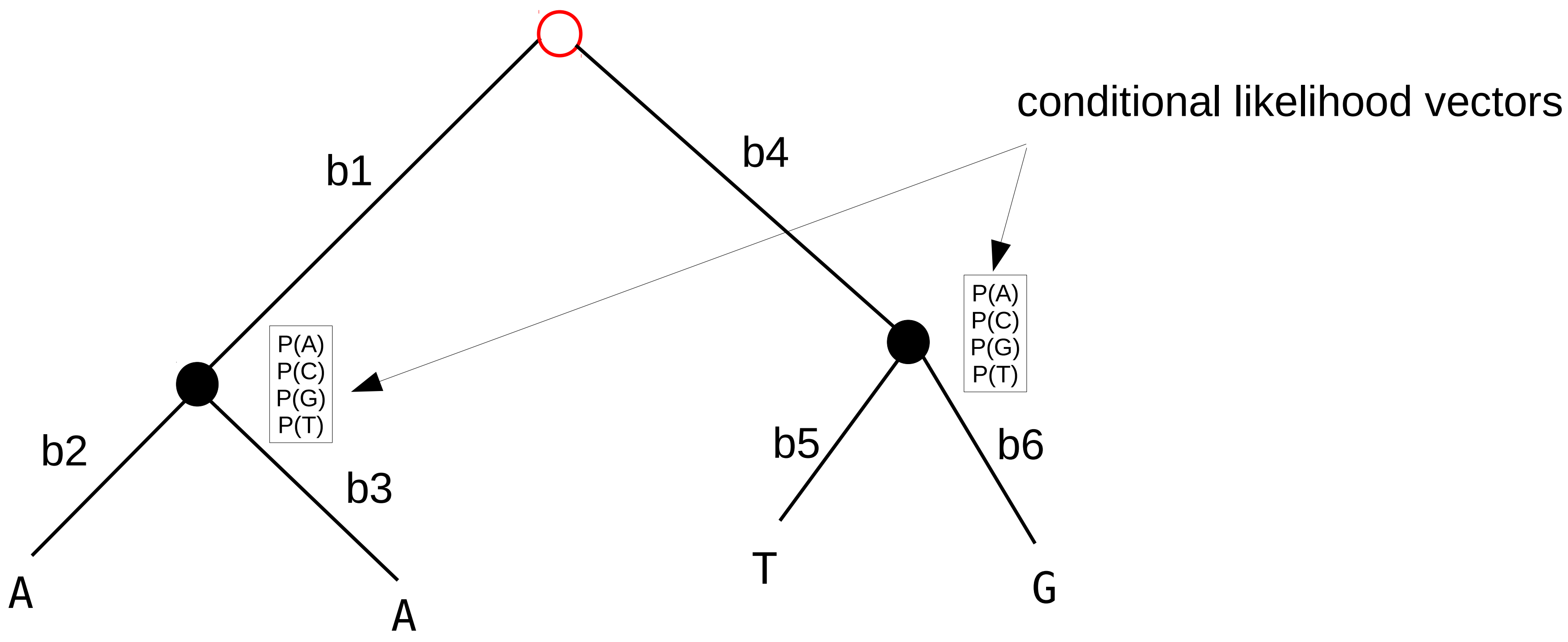


The Felsenstein Pruning Algorithm

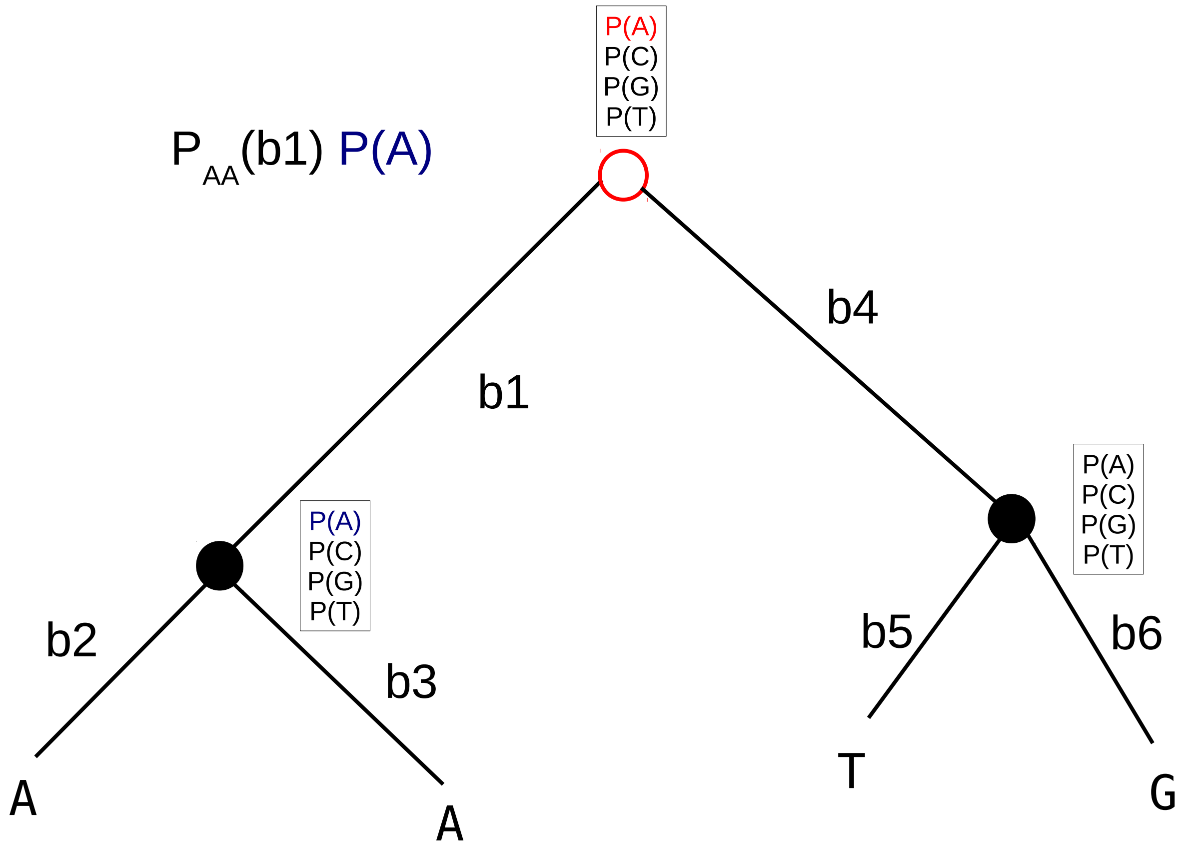
Post order traversal



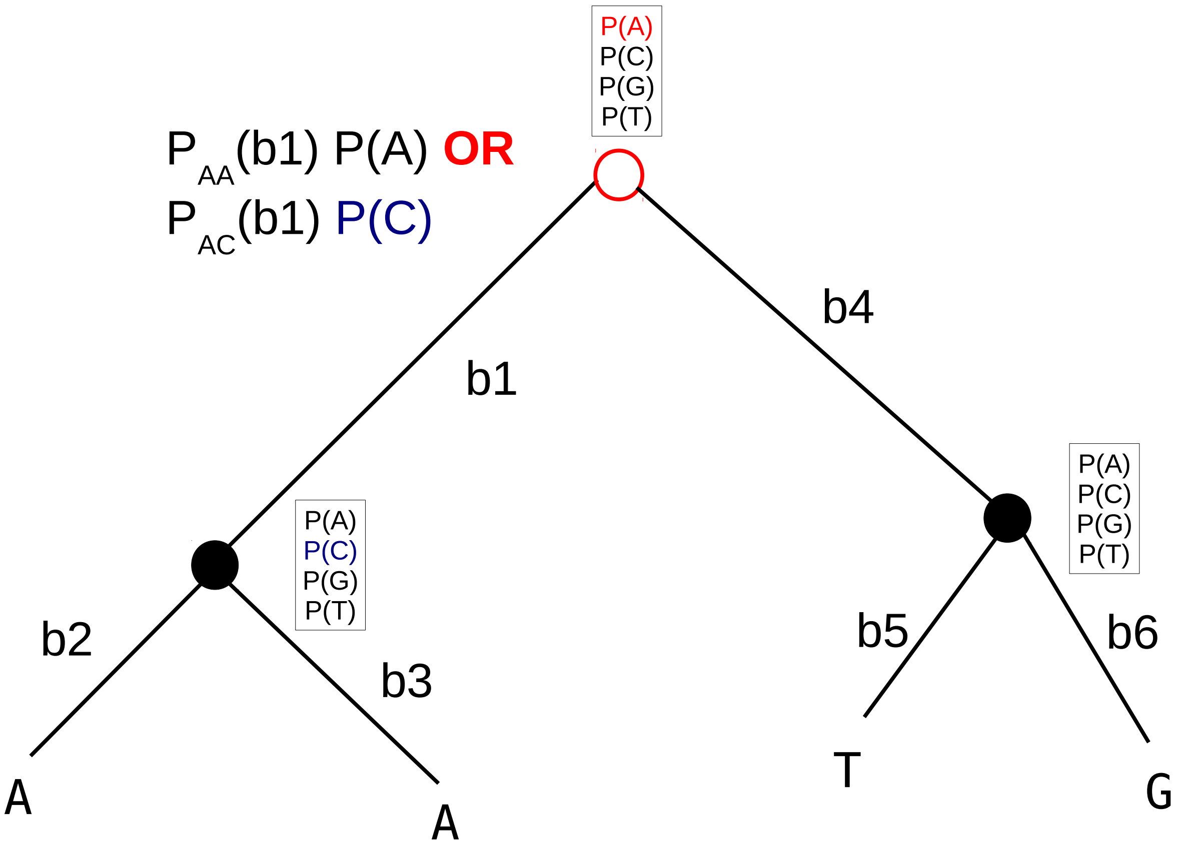
Felsenstein Pruning



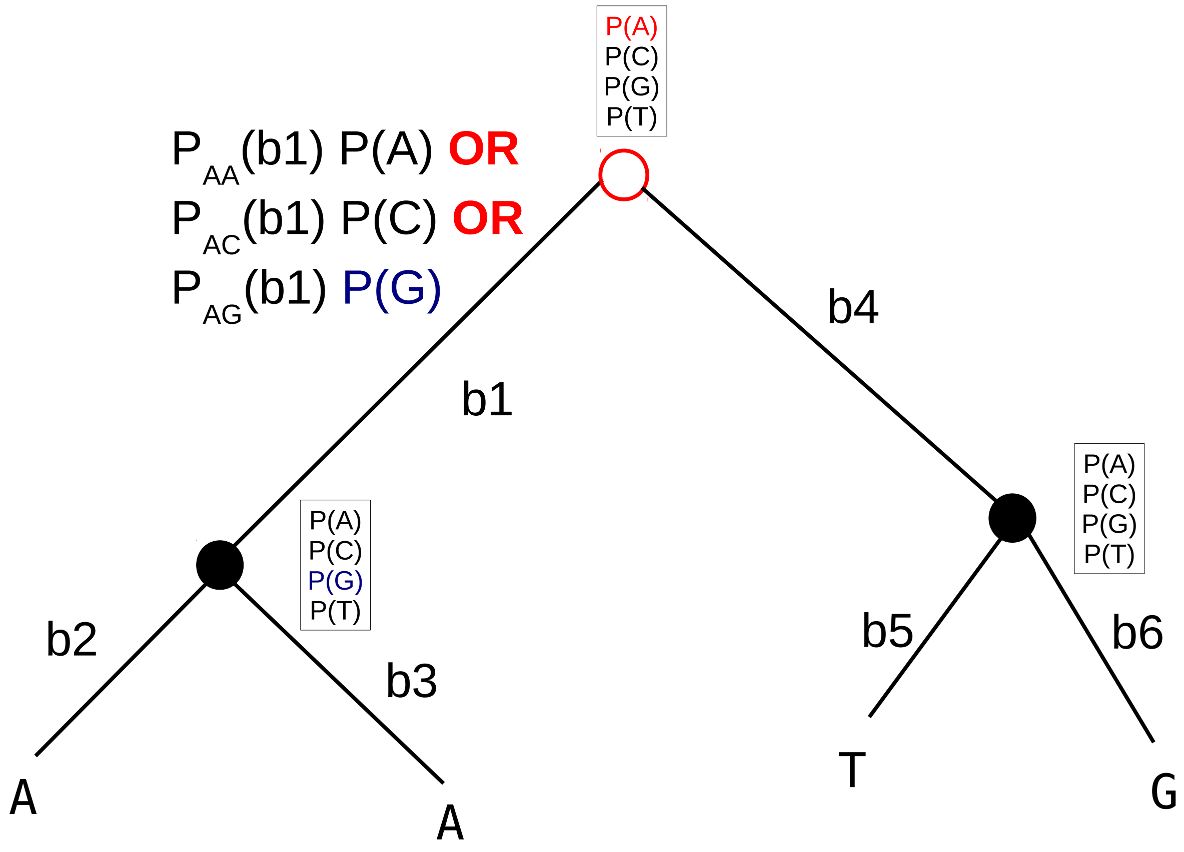
Felsenstein Pruning



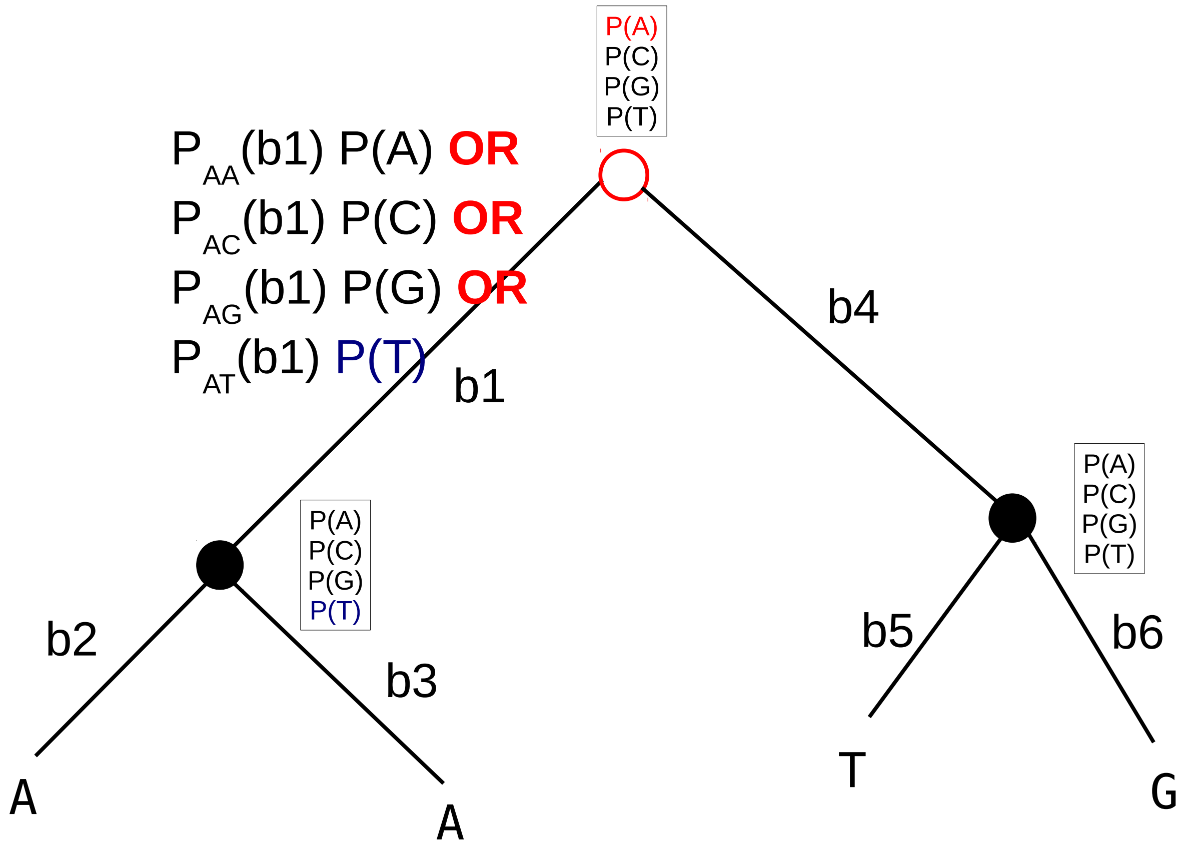
Felsenstein Pruning



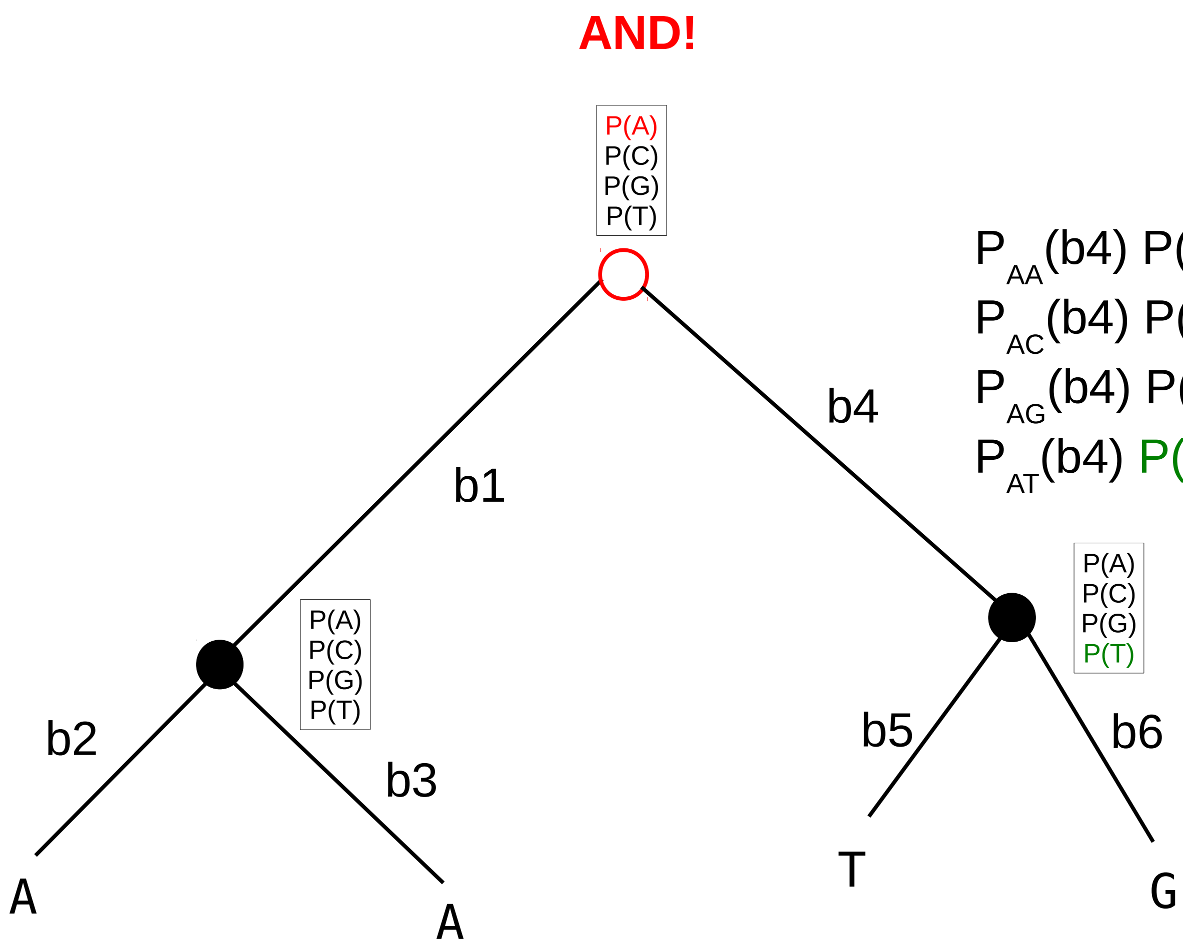
Felsenstein Pruning



Felsenstein Pruning



Felsenstein Pruning



$P_{AA}(b4)\ P(A)$ **OR**

$P_{AC}(b4)\ P(C)$ **OR**

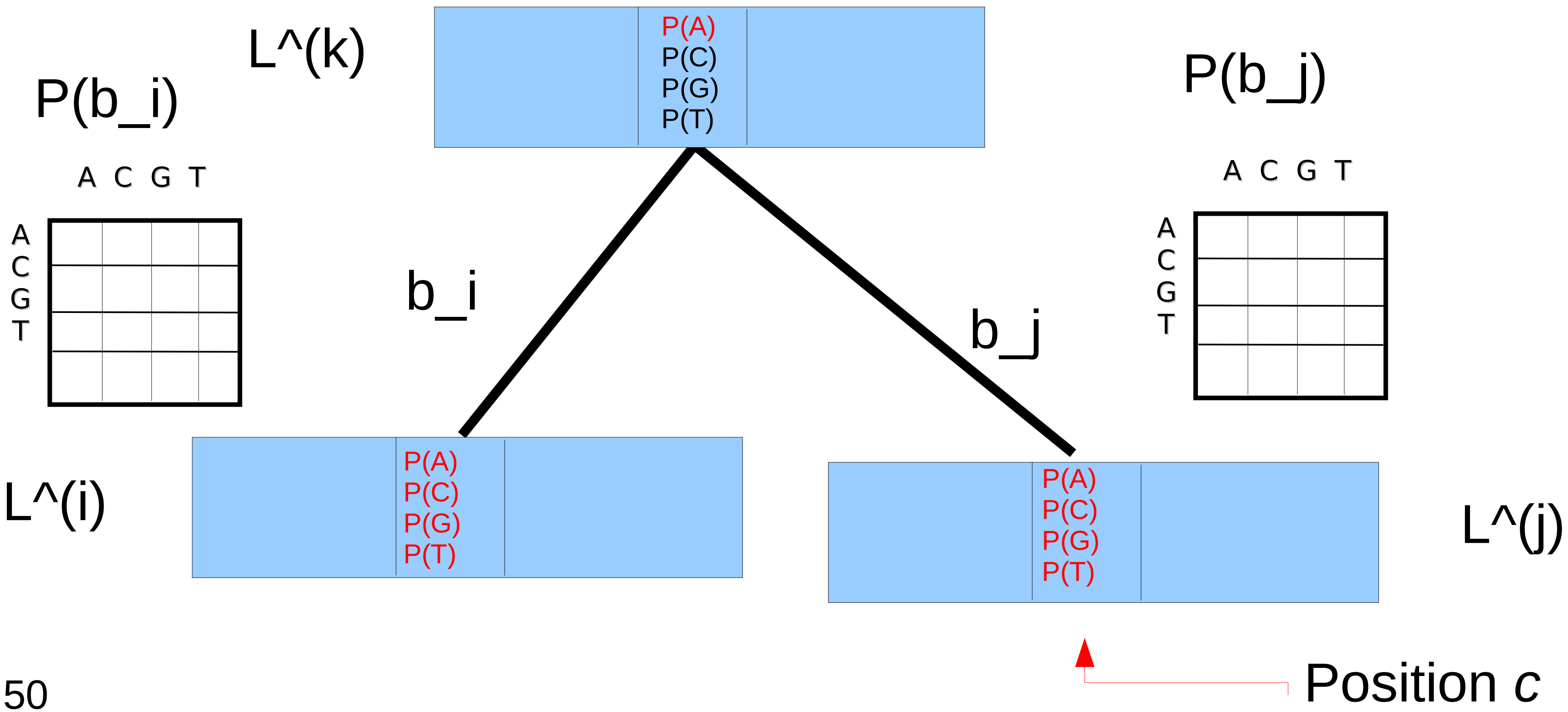
$P_{AG}(b4)\ P(G)$ **OR**

$P_{AT}(b4)\ P(T)$

Felsenstein Pruning

AND (left branch/right branch)

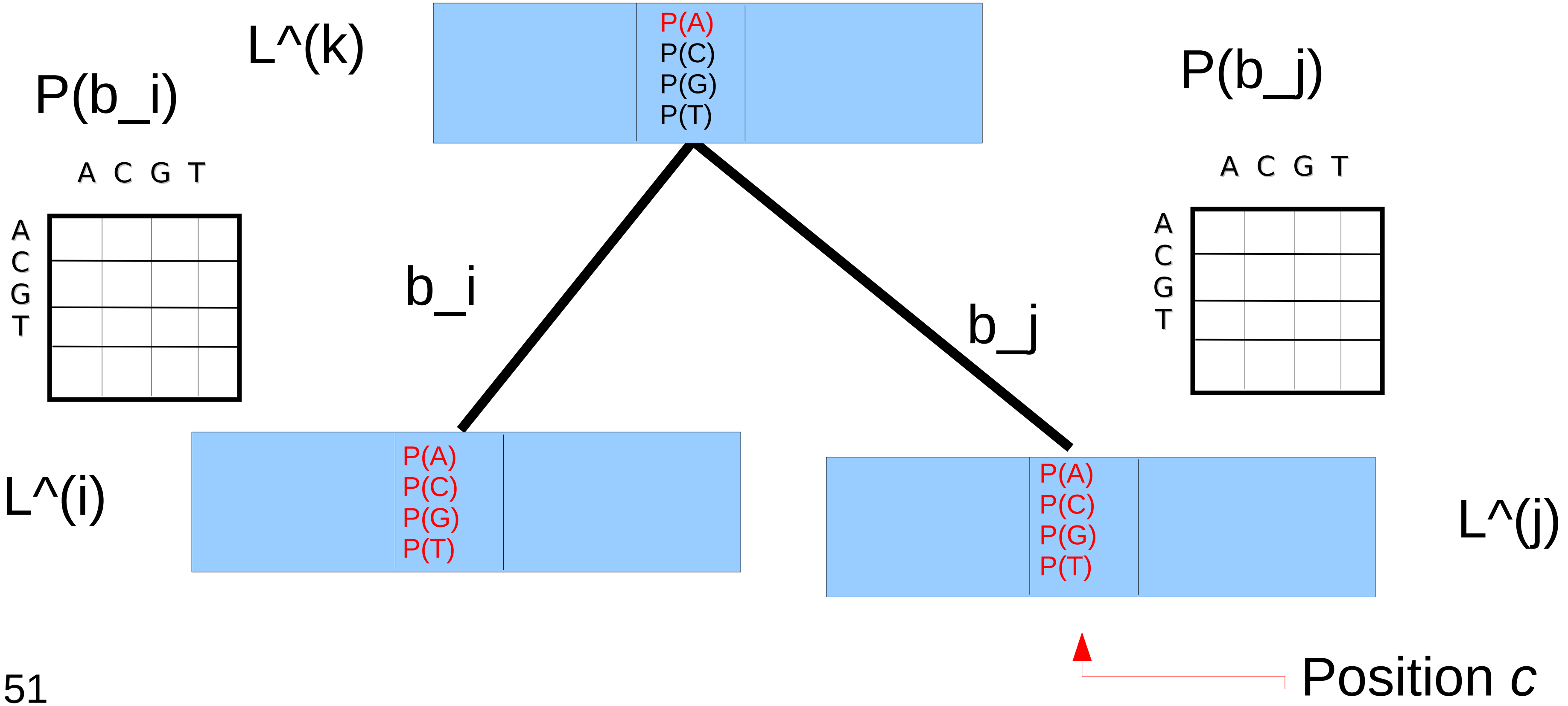
$$\vec{L}_A^{(k)}(c) = \Big(\sum_{S=A}^{T} P_{AS}(b_i)\vec{L}_S^{(i)}(c)\Big)\Big(\sum_{S=A}^{T} P_{AS}(b_j)\vec{L}_S^{(j)}(c)\Big)$$



Felsenstein Pruning

OR (along left branch)

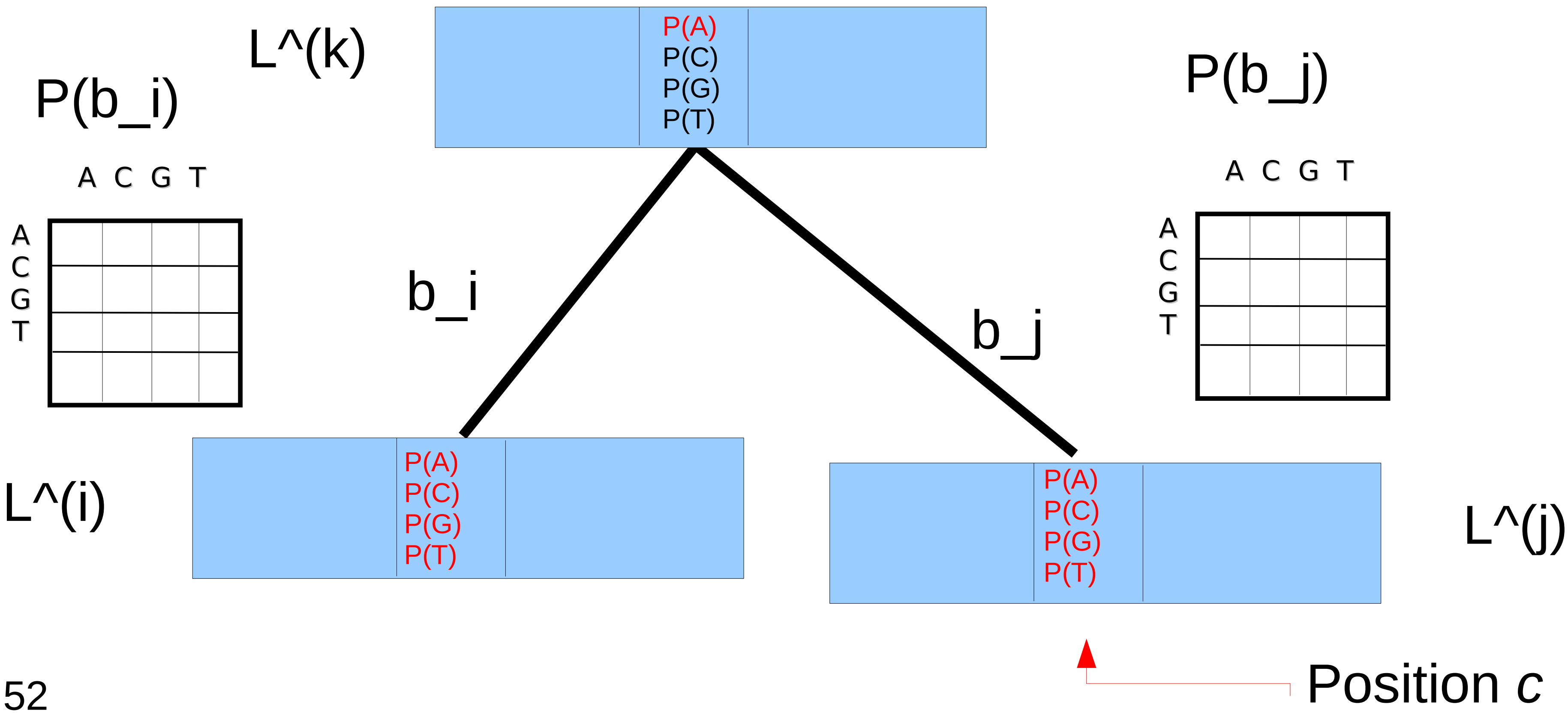
$$\vec{L}_A^{(k)}(c) = \left(\sum_{S=A}^{T} P_{AS}(b_i)\vec{L}_S^{(i)}(c)\right)\left(\sum_{S=A}^{T} P_{AS}(b_j)\vec{L}_S^{(j)}(c)\right)$$



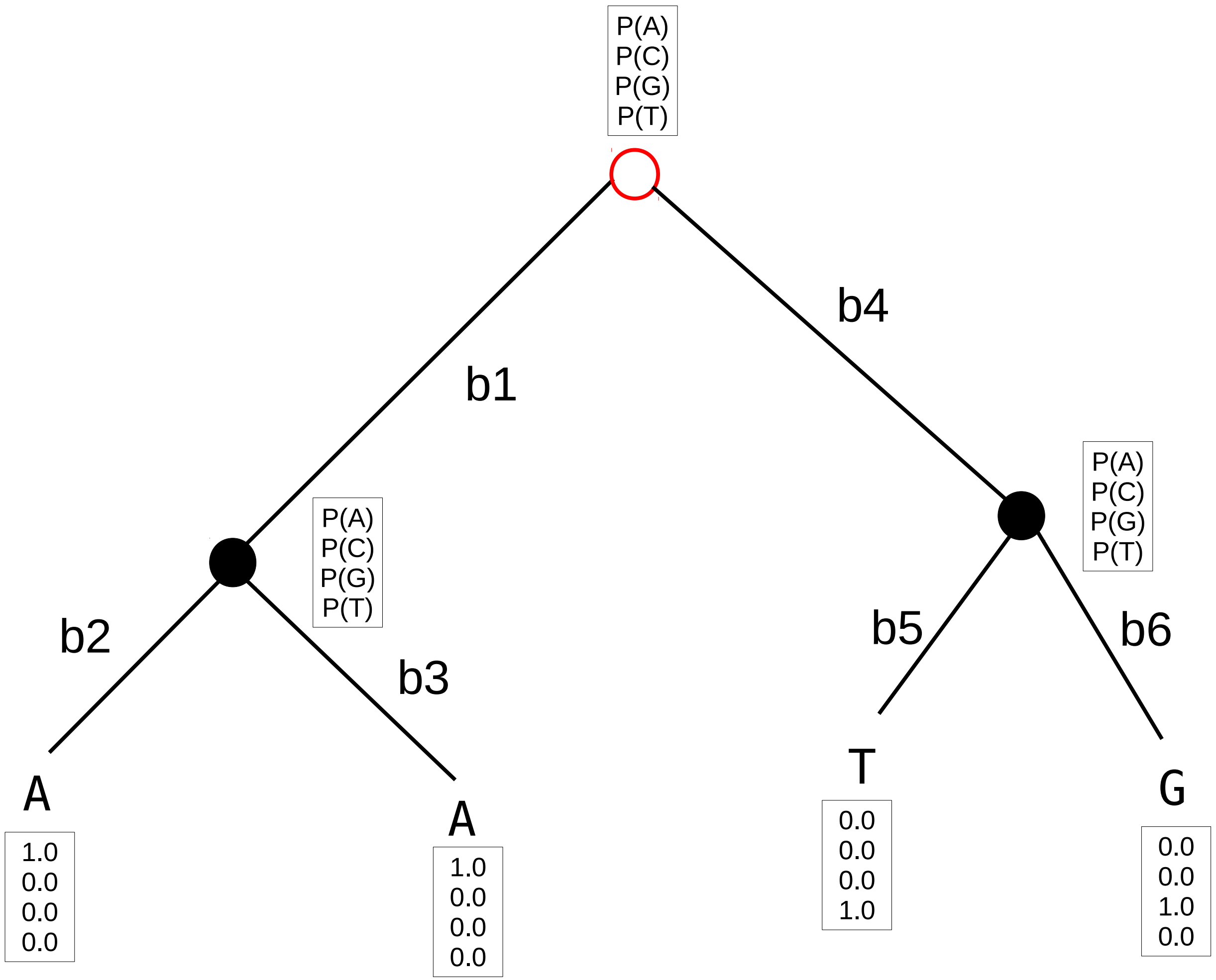
Felsenstein Pruning

OR (along right branch)

$$\vec{L}_A^{(k)}(c) = \left(\sum_{S=A}^{T} P_{AS}(b_i)\vec{L}_S^{(i)}(c)\right)\left(\sum_{S=A}^{T} P_{AS}(b_j)\vec{L}_S^{(j)}(c)\right)$$

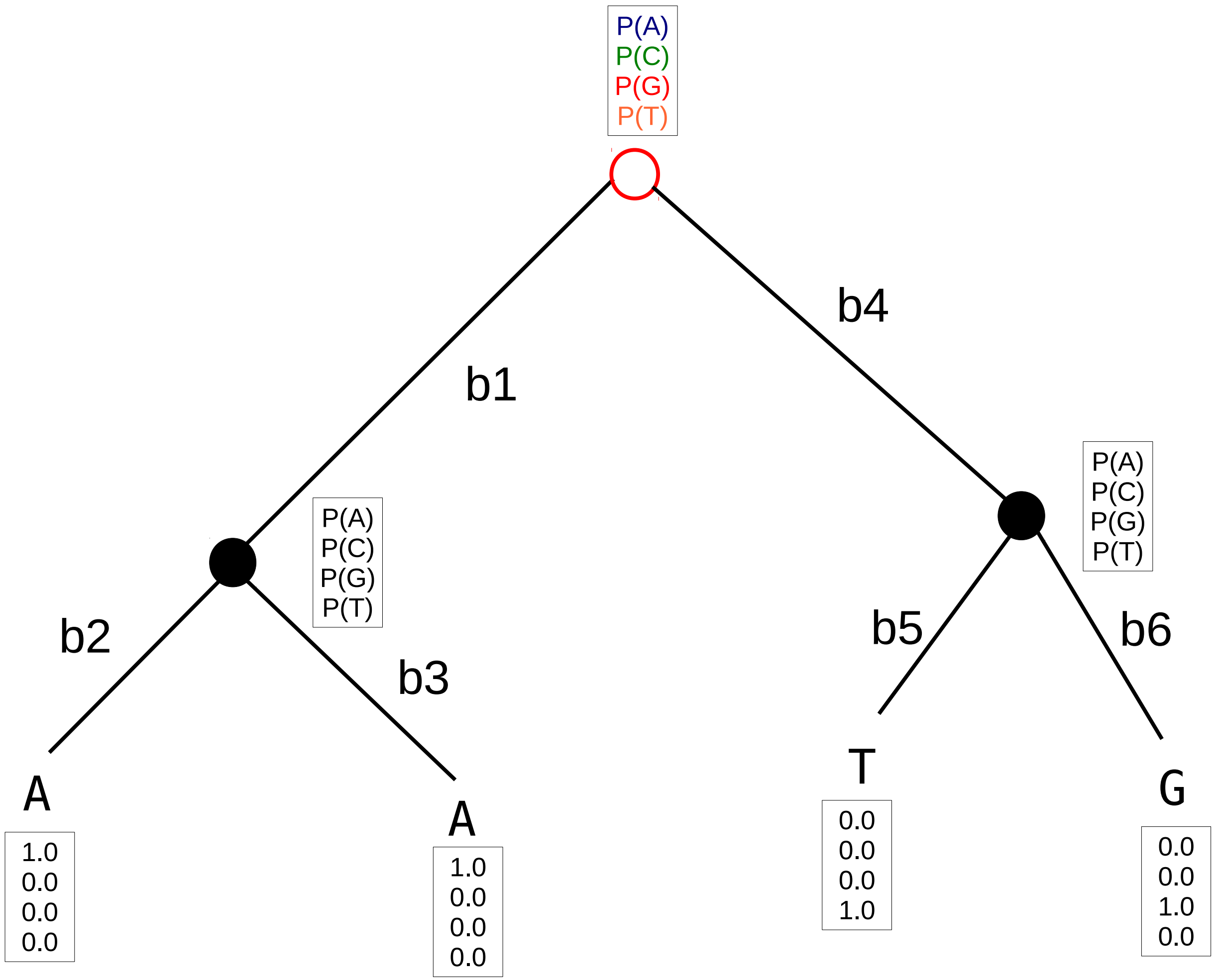


Felsenstein Pruning



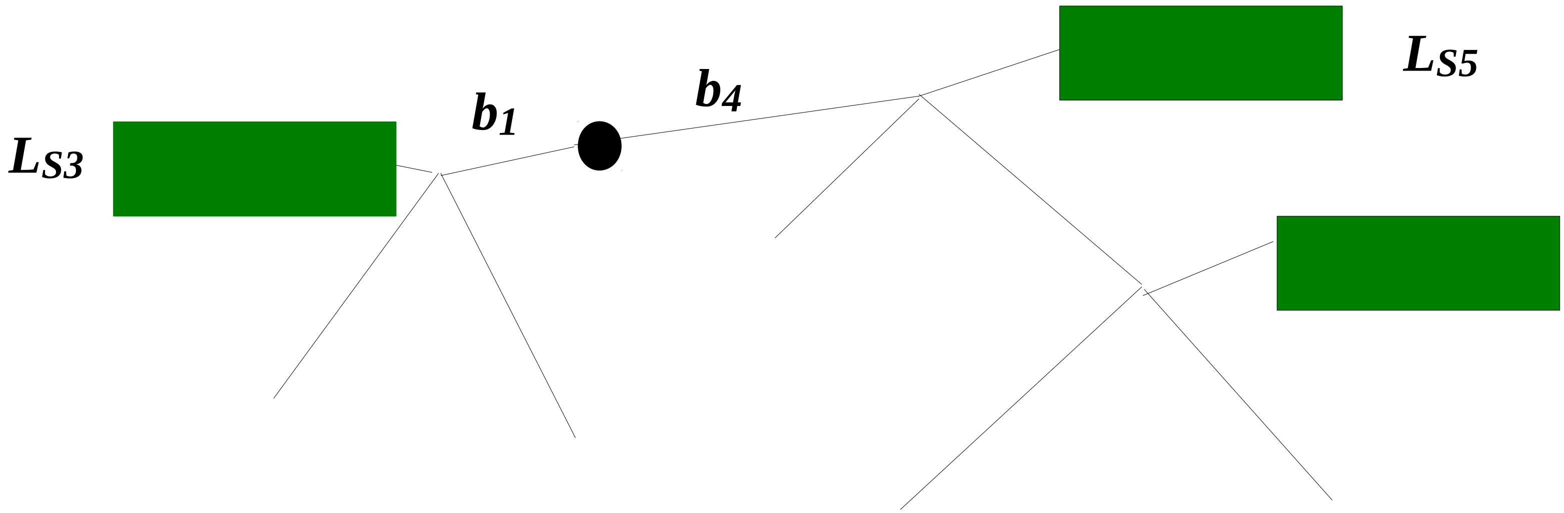
Felsenstein Pruning

Likelihood at the root: $L_i = \pi_A\ P(A) + \pi_C\ P(C) + \pi_G\ P(G) + \pi_T\ P(T)$



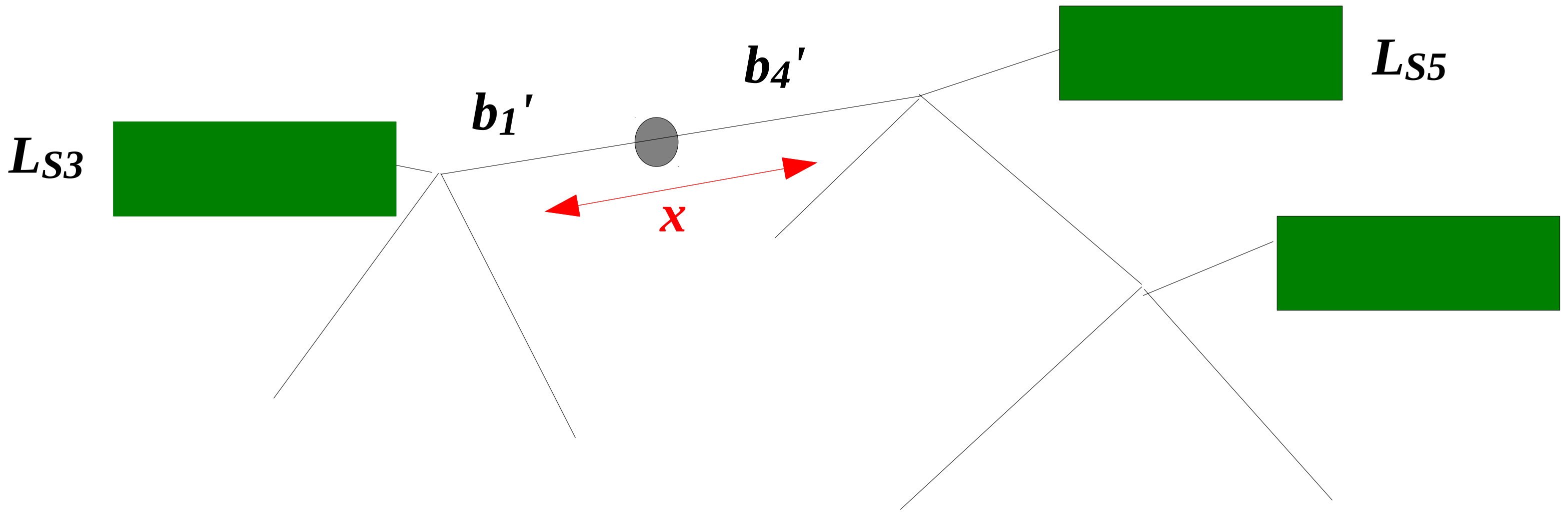
Why is time-reversibility important?

$$L = \sum_{S_4=A}^{T} \pi_{S_4} \sum_{S_3=A}^{T} P_{S_4 S_3}(b_1) L_{S_3}^{(3)} \sum_{S_5=A}^{T} P_{S_4 S_5}(b_4) L_{S_5}^{(5)}$$



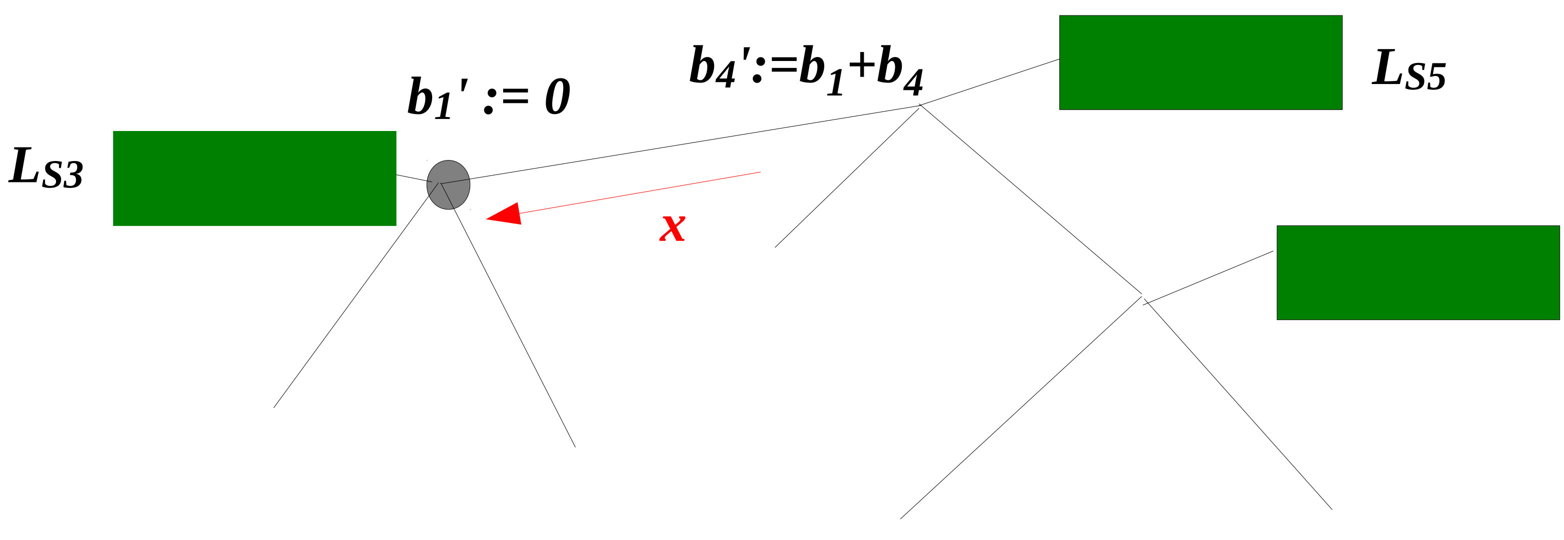
Why is time-reversibility important?

$$L = L' = \sum_{S_4=A}^{T} \pi_{S_4} \sum_{S_3=A}^{T} P_{S_4 S_3}(b_1 + x) L_{S_3}^{(3)} \sum_{S_5=A}^{T} P_{S_4 S_5}(b_4 - x) L_{S_5}^{(5)}$$



Why is time-reversibility important?

$$L = L' = \sum_{S_4=A}^{T} \pi_{S_4} \sum_{S_3=A}^{T} P_{S_4 S_3}(b_1 + x) L_{S_3}^{(3)} \sum_{S_5=A}^{T} P_{S_4 S_5}(b_4 - x) L_{S_5}^{(5)}$$



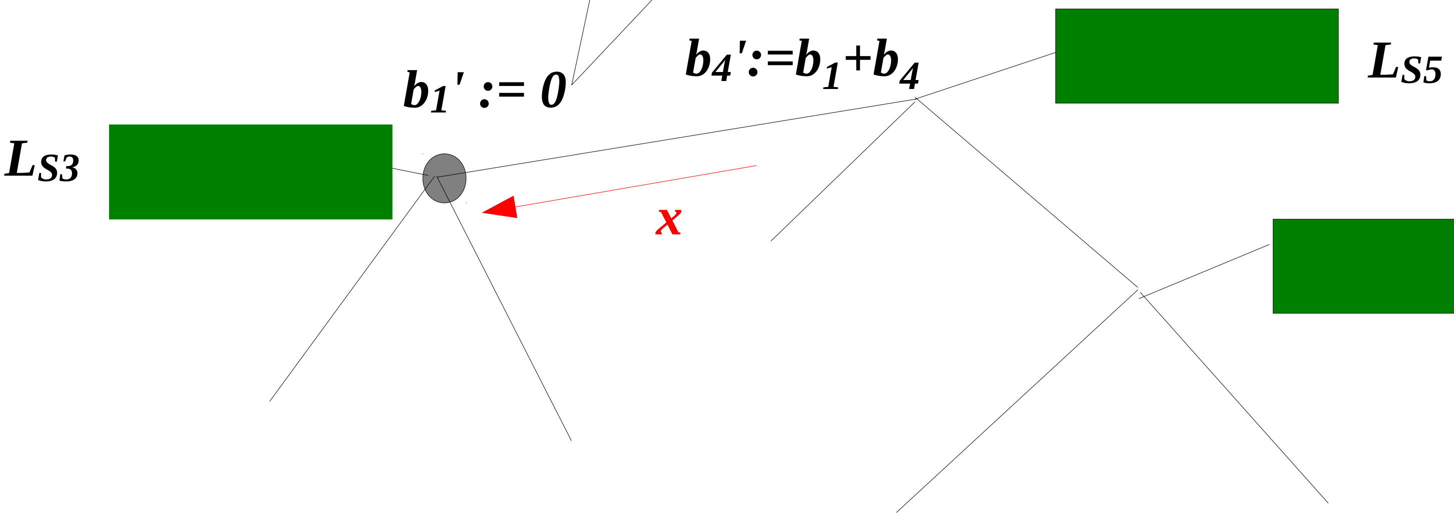
Why is time-reversibility important?

$$L = L' = \sum_{S_4=A}^{T} \cdots \sum_{S_5=A}^{T} P_{S_4 S_5}(b_4 - x) L_{S_5}^{(5)}$$

This observation can be applied recursively to the tree

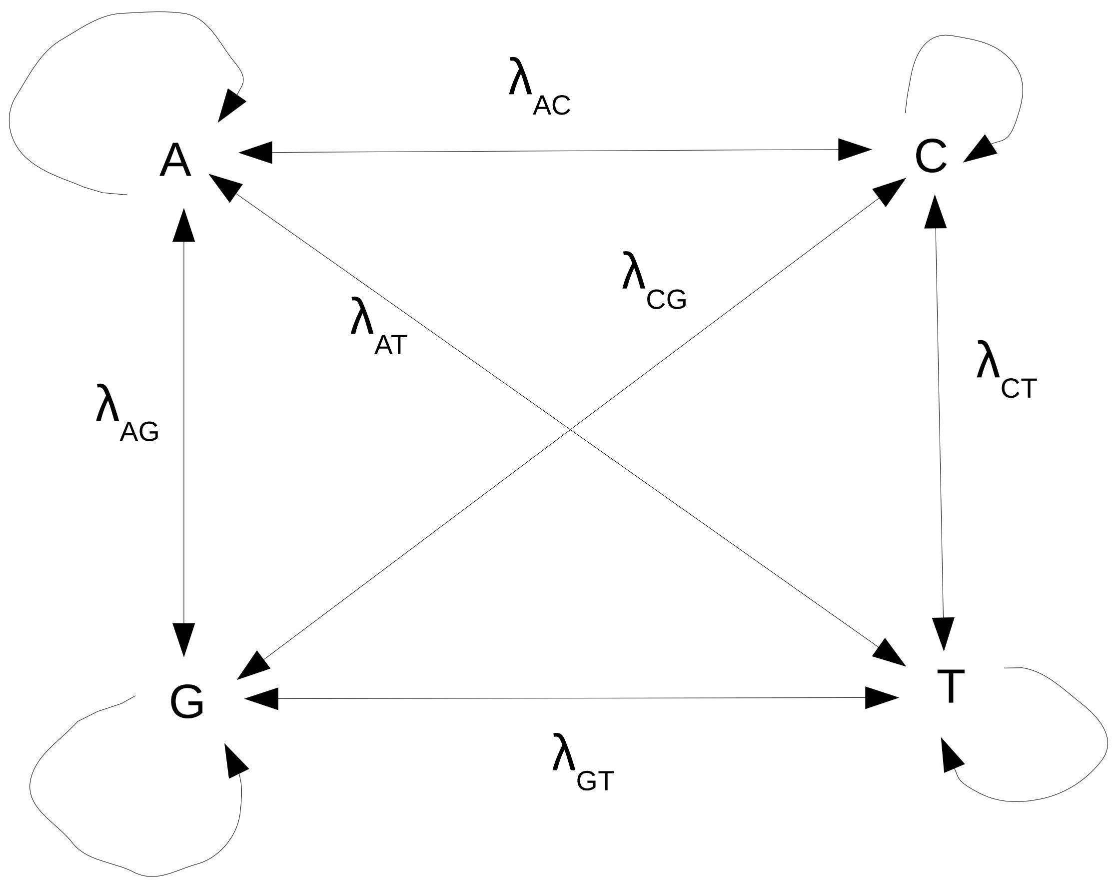
→

It does not matter at all where we place the root!



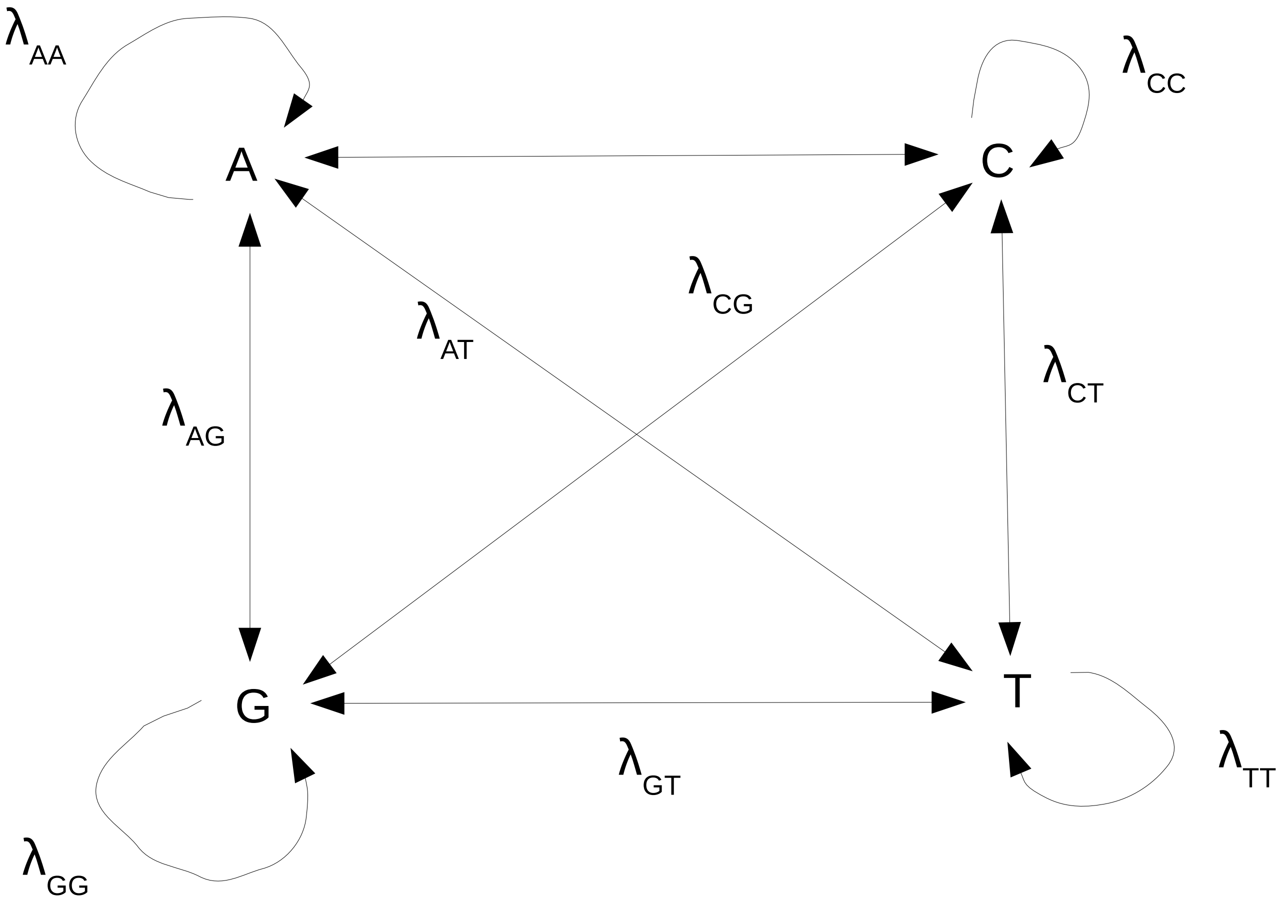
What's in the black box $P_{ij}(t)$?

Instantaneous rate matrix *R*!



What's in the black box $P_{ij}(t)$?

What about the probabilities of staying in the current state?
→ they are given by the properties of continuous Markov chains!
e.g., $\lambda_{AA} = -(\lambda_{AC} + \lambda_{AG} + \lambda_{AT})$ rows in the *R* matrix need to sum to **0**



What's in the black box $P_{ij}(t)$?

	A	C	G	T
A	*	λ_{AC}	λ_{AG}	λ_{AT}
C		*	λ_{CG}	λ_{CT}
G			*	λ_{GT}
T	Symmetric			*

What's in the black box $P_{ij}(t)$?

Diagonal values are given by the off-diagonal values (R matrix property)

$\lambda_{AA} = -(\lambda_{AC} + \lambda_{AG} + \lambda_{AT})$

	A	C	G	T
A	*	λ_{AC}	λ_{AG}	λ_{AT}
C		*	λ_{CG}	λ_{CT}
G			*	λ_{GT}
T	Symmetric			*

What's in the black box $P_{ij}(t)$?

$$
\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & * & \lambda_{AC} & \lambda_{AG} & \lambda_{AT} \\
C & & * & \lambda_{CG} & \lambda_{CT} \\
G & & & * & \lambda_{GT} \\
T & \text{Symmetric} & & & *
\end{array}
$$

Equilibrium frequency vector $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ *where* $\pi_A + \pi_C + \pi_G + \pi_T = 1$

The Jukes-Cantor model

$$
\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & * & \lambda & \lambda & \lambda \\
C & & * & \lambda & \lambda \\
G & & & * & \lambda \\
T & & & & *
\end{array}
$$

$\Pi = (1/4,\ 1/4,\ 1/4,\ 1/4)$

Felsenstein 81

	A	C	G	T
A	*	λ	λ	λ
C		*	λ	λ
G			*	λ
T				*

$\Pi_i \neq \Pi_j$

Kimura 2-parameter model 1980

	A	C	G	T
A	*	λ	ζ	λ
C		*	ζ	λ
G			*	ζ
T				*

$\Pi = (1/4,\ 1/4,\ 1/4,\ 1/4)$

HKY85

	A	C	G	T
A	*	λ	ζ	λ
C		*	ζ	λ
G			*	ζ
T				*

$\Pi_i \neq \Pi_j$

GTR 1986

	A	C	G	T
A	*	α	β	γ
C		*	δ	ε
G			*	ζ
T				*

$\Pi_i \neq \Pi_j$

GTR 1986

	A	C	G	T
A	*	α	β	γ
C		*	δ	ε
G			*	ζ
T				*

Note that these are **relative** rates, their Values only matter relative to each other, so we can set *ζ := 1.0* by default

$\Pi_i \neq \Pi_j$

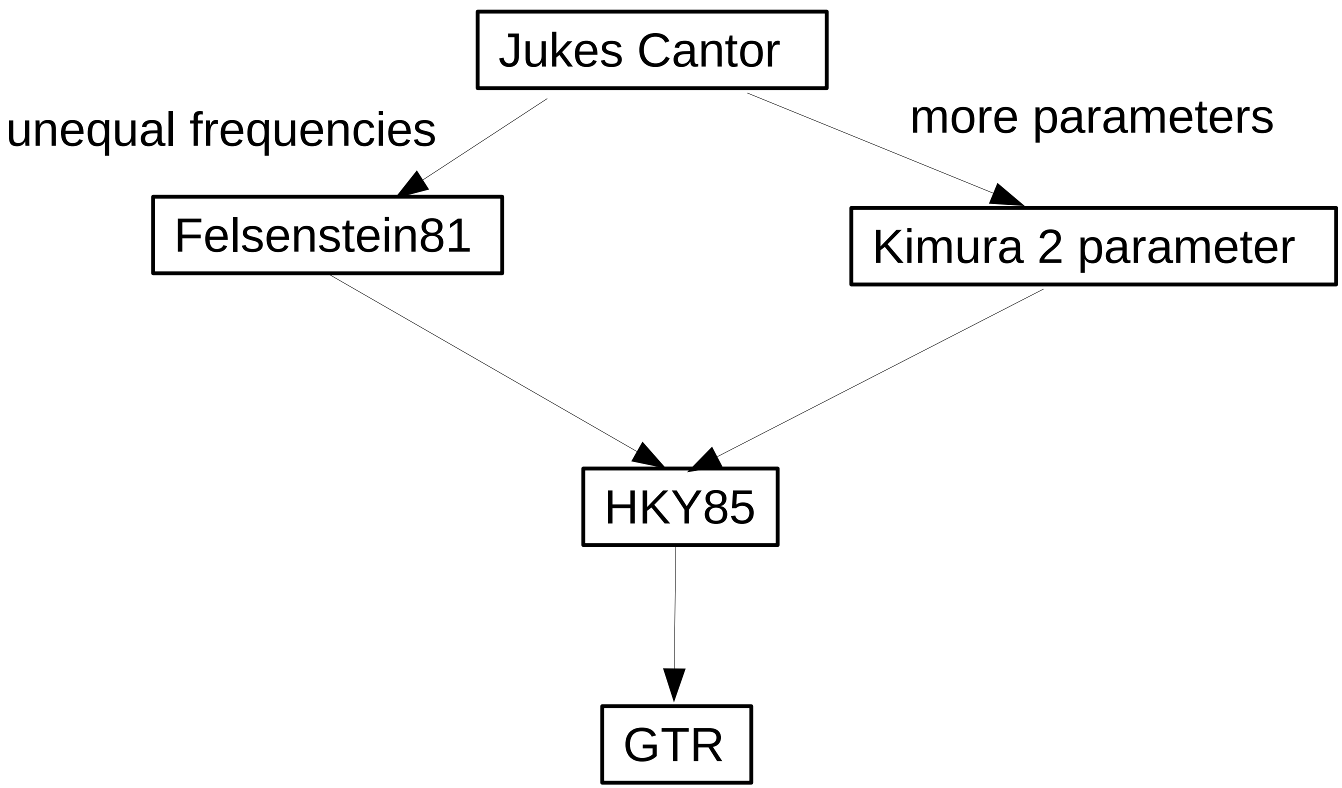
GTR 1986

	A	C	G	T
A	*	α	β	γ
C		*	δ	ε
G			*	1.0
T				*

Note that these are **relative** rates, their values only matter relative to each other, so we can set *ζ := 1.0* by default. Although the GTR model has 6 rates, it only has 5 free parameters!

$\Pi_i \neq \Pi_j$

Model Hierarchy



GTR 1986

This is a rate matrix, time reversibility would Require $\boldsymbol{\pi_i r_{ij} = \pi_j r_{ji}}$

	A	C	G	T
A	*	α	β	γ
C		*	δ	ε
G			*	1.0
T				*

$\Pi_i \neq \Pi_j$

GTR 1986

This is a rate matrix,
time reversibility would
Require $\boldsymbol{\pi_i r_{ij} = \pi_j r_{ji}}$
Solution: introduce a
Q matrix $Q := diag(\pi)\ R$

	A	C	G	T
A	*	α	β	γ
C		*	δ	ε
G			*	1.0
T				*

$$\begin{pmatrix} \pi_A & & & \\ & \pi_C & & \\ & & \pi_G & \\ & & & \pi_T \end{pmatrix}$$

$\Pi_i \neq \Pi_j$

GTR 1986

	A	C	G	T
A	*	α	β	γ
C		*	δ	ε
G			*	1.0
T				*

This is a rate matrix, time reversibility would Require $\boldsymbol{\pi_i r_{ij} = \pi_j r_{ji}}$

Solution: introduce a *Q* matrix $Q := diag(\pi)\ R$

$$\begin{pmatrix} \pi_A & & & \\ & \pi_C & & \\ & & \pi_G & \\ & & & \pi_T \end{pmatrix}$$

$\Pi_i \neq \Pi_j$

Then, $\pi_i r_{ij} = \pi_j r_{ji}$ holds

So how do we compute P(t) from Q?

- As we have seen in the lecture on Markov chains:

  $P(t) = e^{Qt}$ = I + Qt + 1/2! (Qt)^2 + 1/3! (Qt)^3 + …

- but this is unfortunately a matrix eponential :-(
- I will spare you the details, but in general, e.g., for GTR we need to apply an egienvector/eigenvalue decomposition of Q to calculate:

  $P(t) = U \, exp(diag(\lambda_i)t) \, U^{-1}$

Matrix and inverse matrix of eigenvectors of *Q*

So how do we compute P(t) from Q?

- As we have seen in the lecture on Markov chains:

  $P(t) = e^{Qt} = I + Qt + 1/2!\,(Qt)^2 + 1/3!\,(Qt)^3 + \ldots$

- but this is unfortunately a matrix exponential :-(
- I will spare you the details, but in general, e.g., for GTR we need to apply an egienvector/eigenvalue decomposition of Q to calculate:

  $P(t) = U\, exp(diag(\lambda_i)t)\, U^{-1}$

Diagonal matrix of eigenvalues of *Q*, here the exponential function *exp()* is invoked on scalar values!

Likelihood Calculations

- So far, we have only seen how to calculate **a** likelihood on a
    - given, fixed tree topology
    - with given fixed branch lengths
    - and given, fixed remaining model parameters
- Computing the **maximum** likelihood score, is much more complicated as it requires functions for optimizing continuous parameters and functions for searching the discrete space of trees !