Introduction to Bioinformatics for Computer Scientists

Lecture 7

Last Time

- Distinction between distance-based and character-based tree inference algorithms
- Distance-based approaches
  - Heuristics
    - Neighbor Joining $O(n^3)$
    - UPGMA $O(n^2)$
  - Criteria
    - Least Squares
    - Minum Evolution

Last Time

- Search Heuristic Components
  - Methods to infer comprehensive starting trees
  - NNI, SPR, and TBR moves for improviong the score of trees - searching tree space
- Character-based Criteria
  - Criteria
    - Parsimony
  - Example of an heuristic parsimony search strategy
    - The `Parsimonator` algorithm

Today

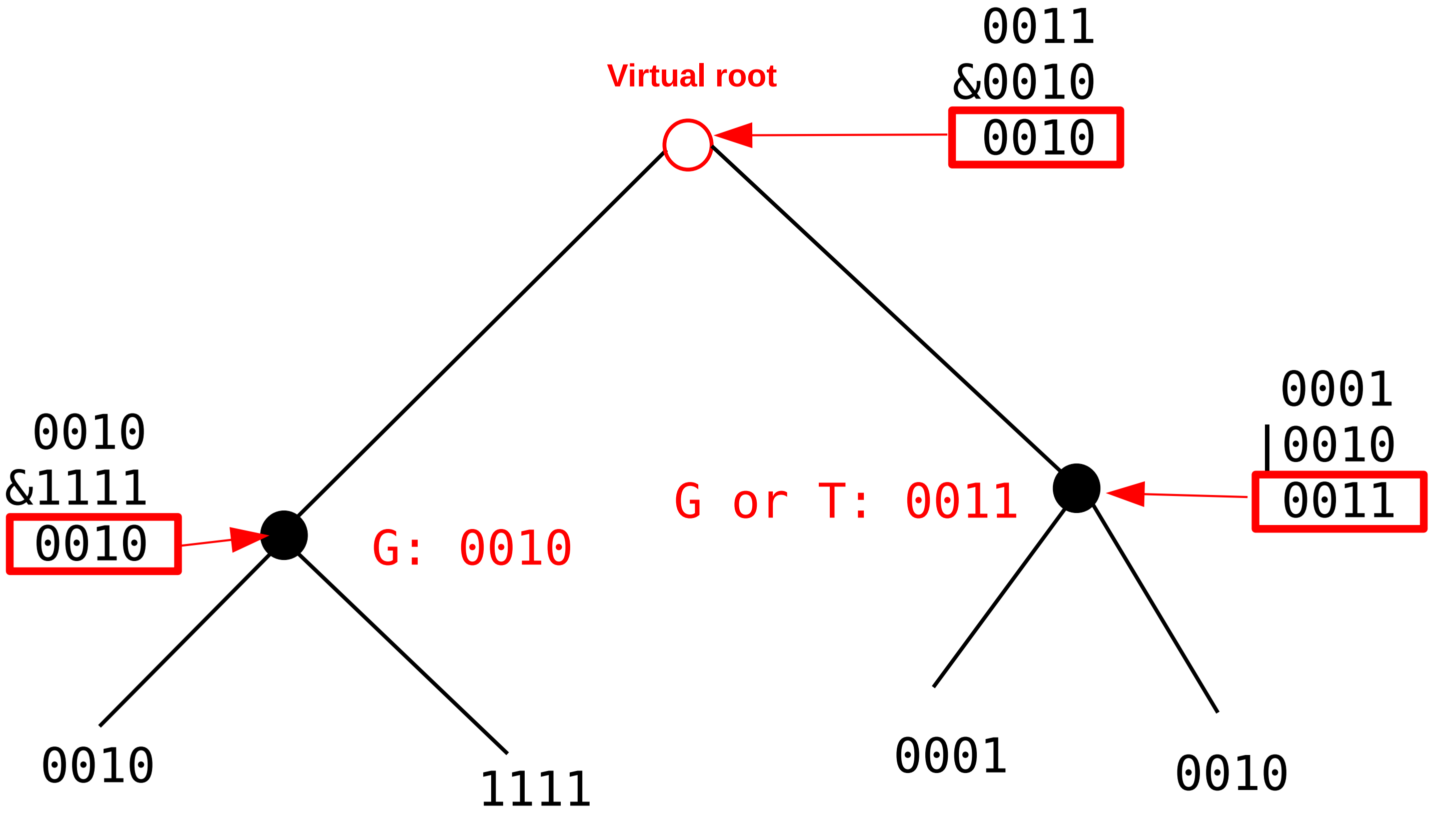
- **Data structures for unrooted trees**
- Why do we need statistical/probabilistic models of evolution?
- Introduction to the phylogenetic likelihood function
    - A detour on Markov Chains
    - The phylogenetic likelihood function

Data Structures for unrooted Trees

- Unrooted trees with dynamically changing virtual roots need a dedicated tree data structure

Remember! - Parsimony

1+2+1+1= 5



Virtual root

0011
&0010
0010

0010
&1111
0010

G: 0010

G or T: 0011

0001
|0010
0011

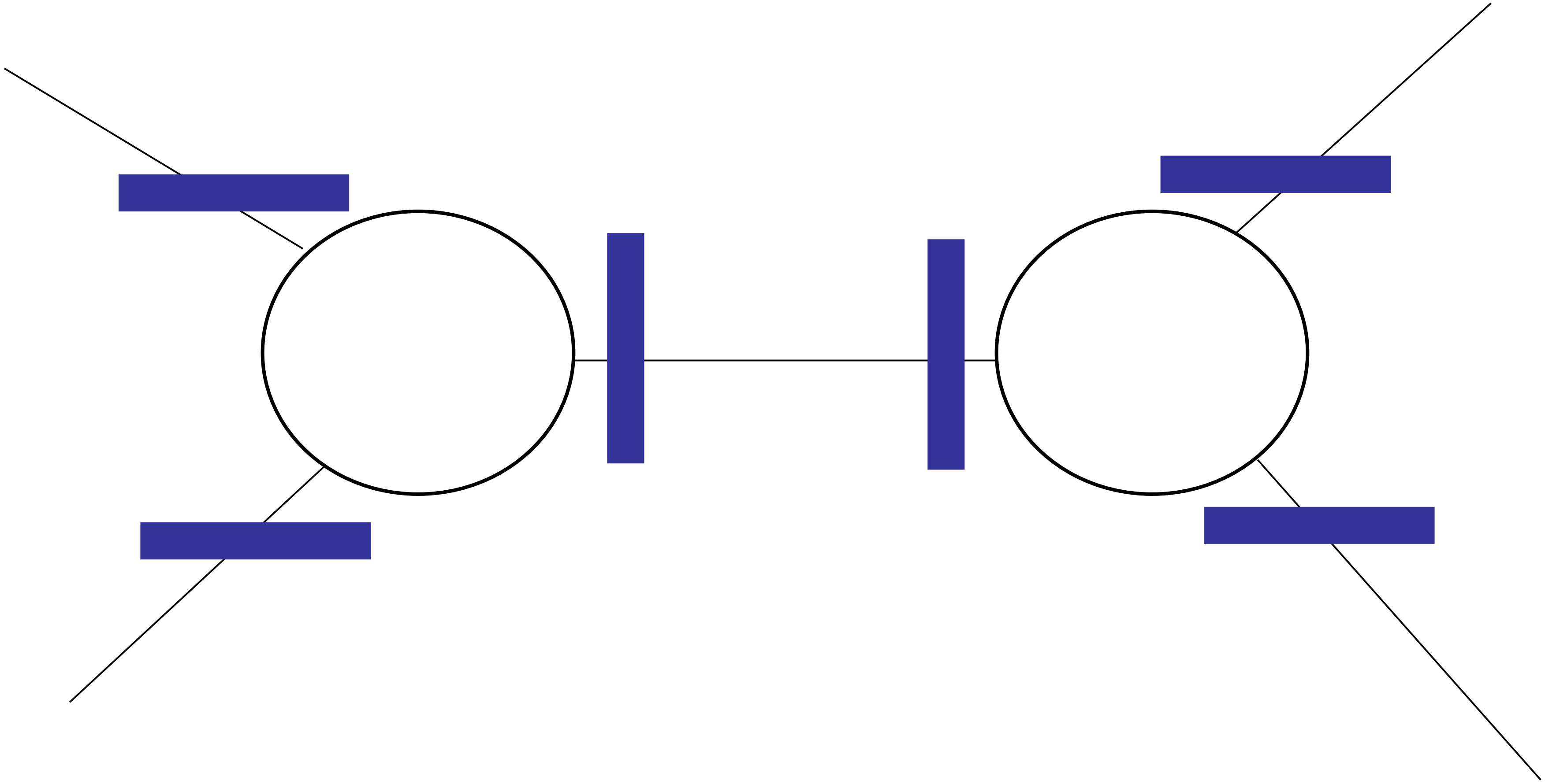
0010

1111

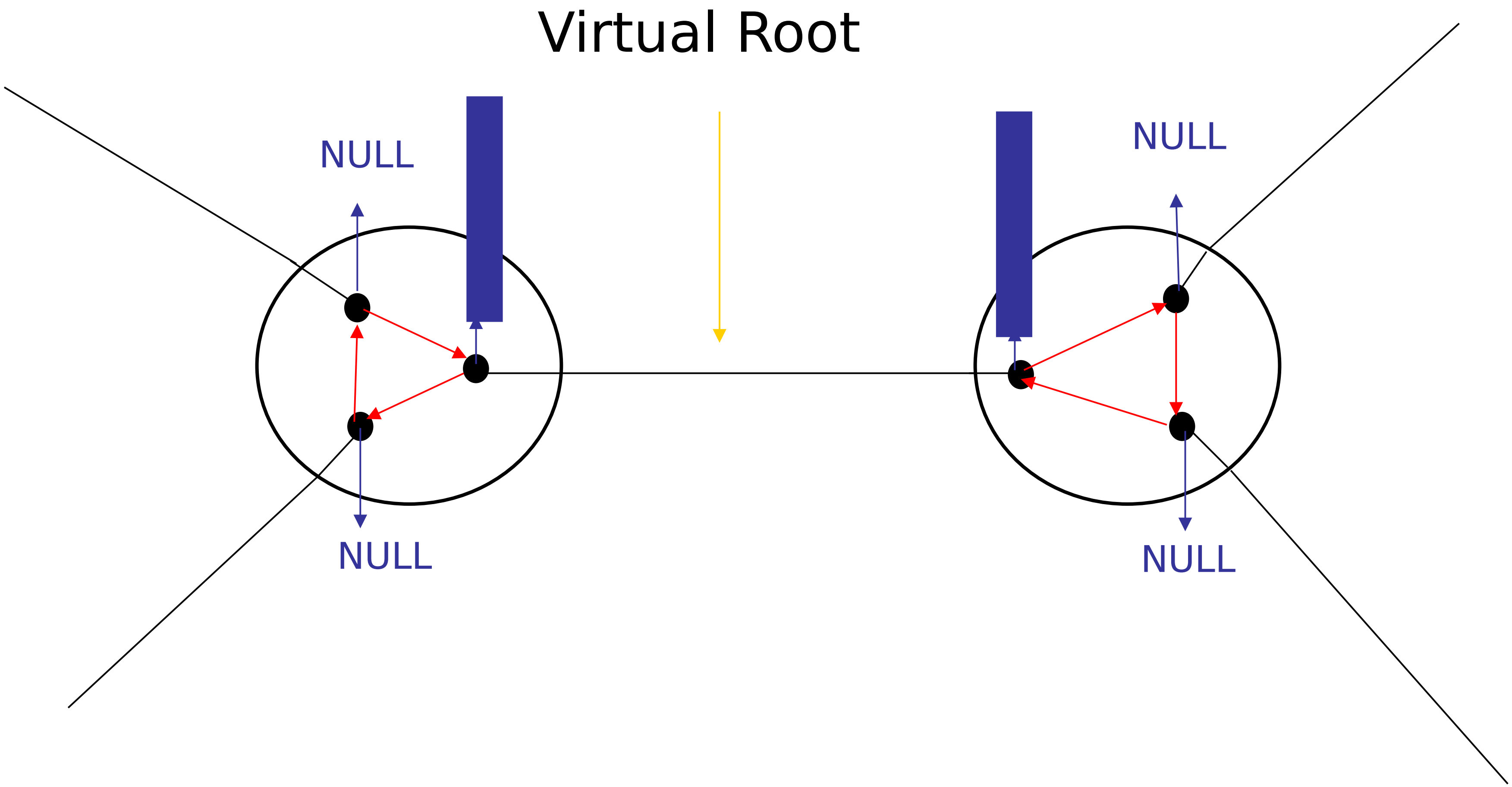
0001

0010

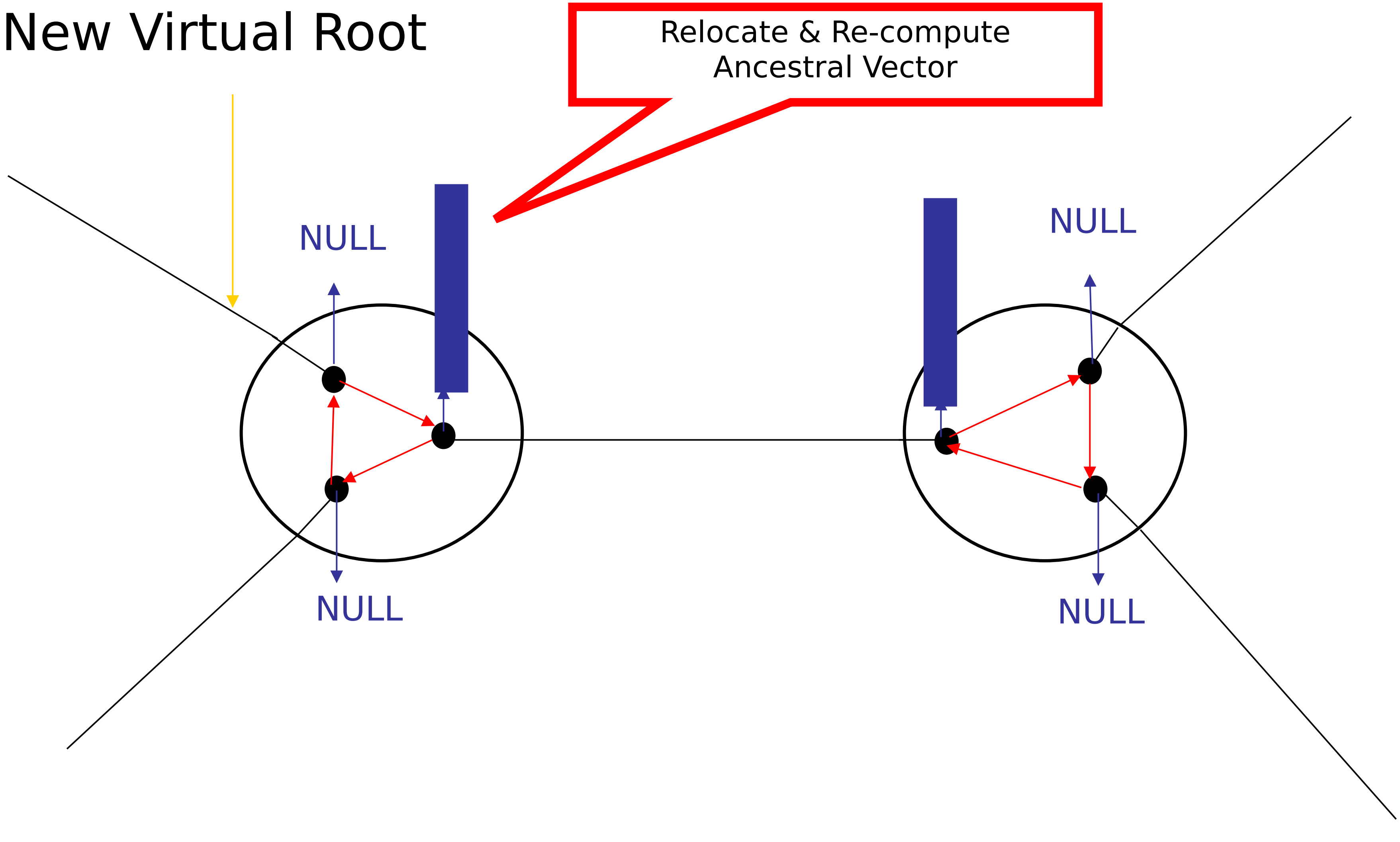
Memory Organization: Ancestral Vectors with an Unrooted View



Memory Organization: Ancestral Vectors with a Rooted View

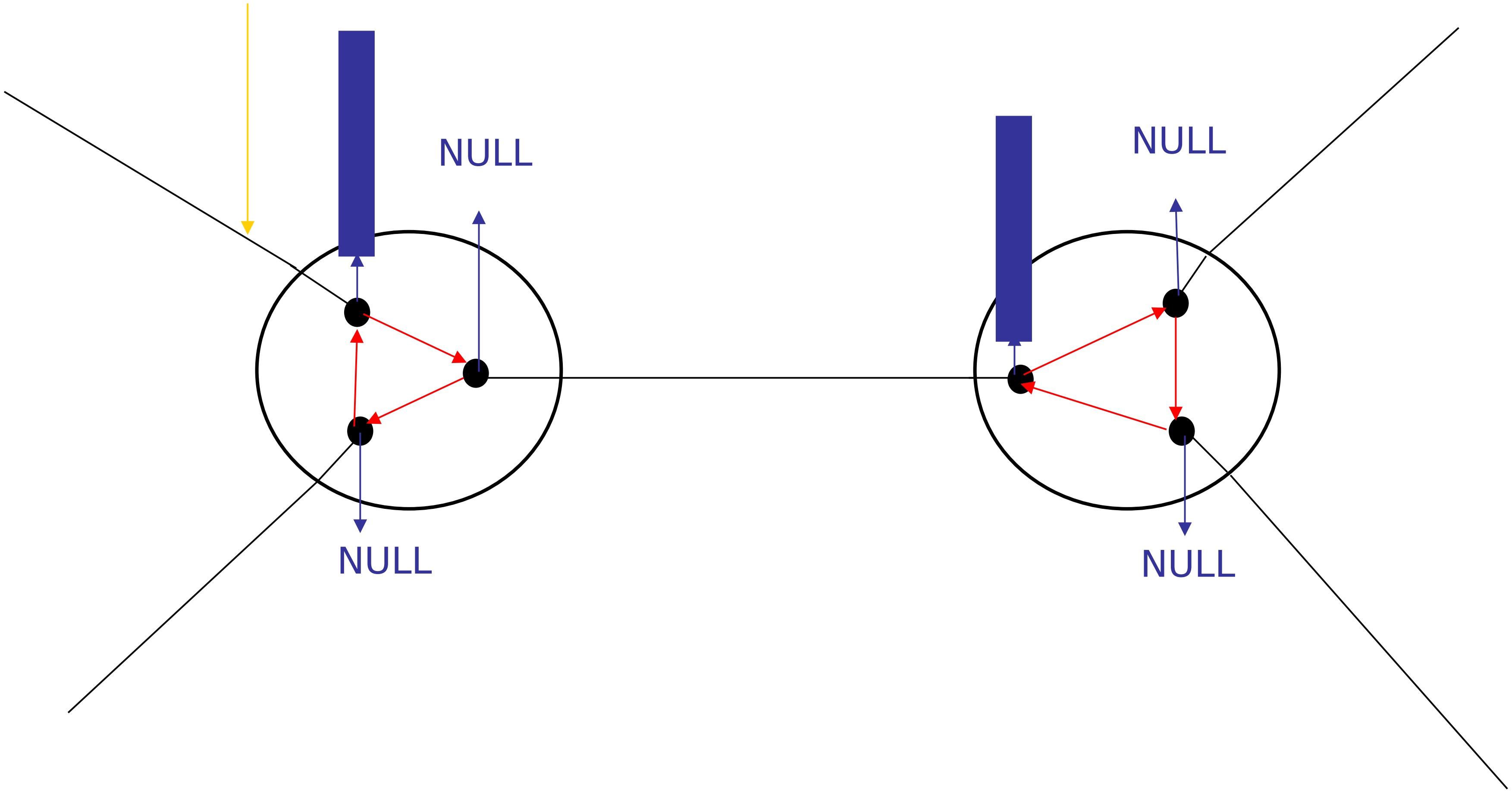


Memory Organization: Ancestral Vectors with a Rooted View



Memory Organization: Ancestral Vectors with a Rooted View

New Virtual Root

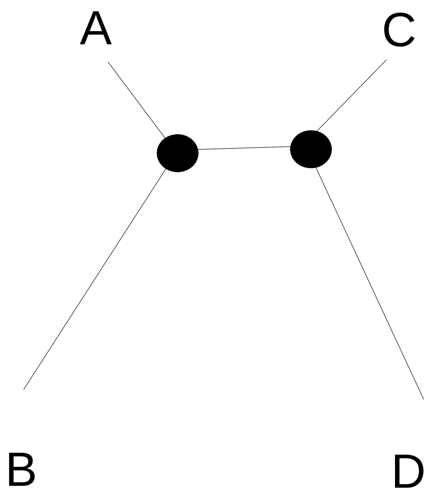


Today

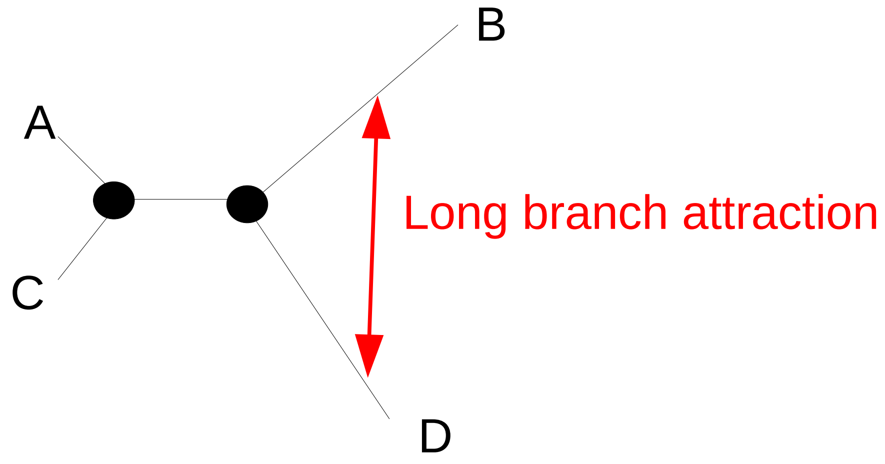
- Data structures for unrooted trees
- **Why do we need statistical/probabilistic models of evolution?**
- Introduction to the phylogenetic likelihood function
  - A detour on Markov Chains
  - The phylogenetic likelihood function

Parsimony & Long Branch Attraction

- Because parsimony tries to minimize the number of mutations it faces some problems on trees with long branches



Correct tree



Wrong tree inferred by parsimony

Parsimony & Long Branch Attraction

- Settings under which parsimony recovers the wrong tree are also called “the Felsenstein Zone” after *Joe Felsenstein* who has made numerous very important contributions to the field, e.g.
  - The Maximum Likelihood model
  - The Bootstrapping procedure
- If you are interested in statistics, there are some on-line courses by Joe at http://evolution.gs.washington.edu/courses.html

Today

- Data structures for unrooted trees
- Why do we need statistical/probabilistic models of evolution?
- **Introduction to the phylogenetic likelihood function**
  - **A detour on Markov Chains**
  - The phylogenetic likelihood function

A Detour to Markov Chains

- Before we start looking at likelihood models for phylogenetics
- We will review the concept of Markov Chains
- This will be useful to
  - Better understand likelihood models
  - Better understand Markov Chain Monte Carlo Sampling for Bayesian statistics

Markov Chains - Outline

- We will mostly talk about discrete Markov chains as this is conceptually easier
- Then, we will talk how to get from discrete Markov chains to continuous Markov chains … which are used in phylogenetics

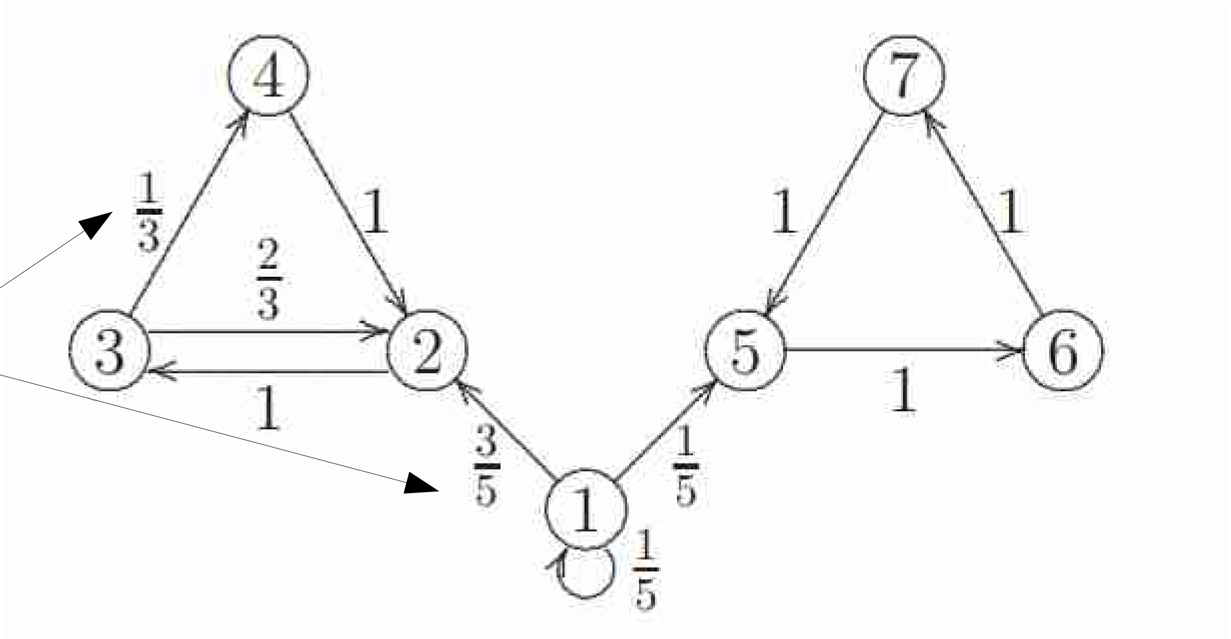
Markov Chains

- Stochastic processes with transition diagrams
- Process, is written as $\{X_0, X_1, X_2, \ldots\}$

  where X_t is the state at discrete time t
- Markov property: X_{t+1} **ONLY** depends on X_t
- Such processes are called **Markov Chains**

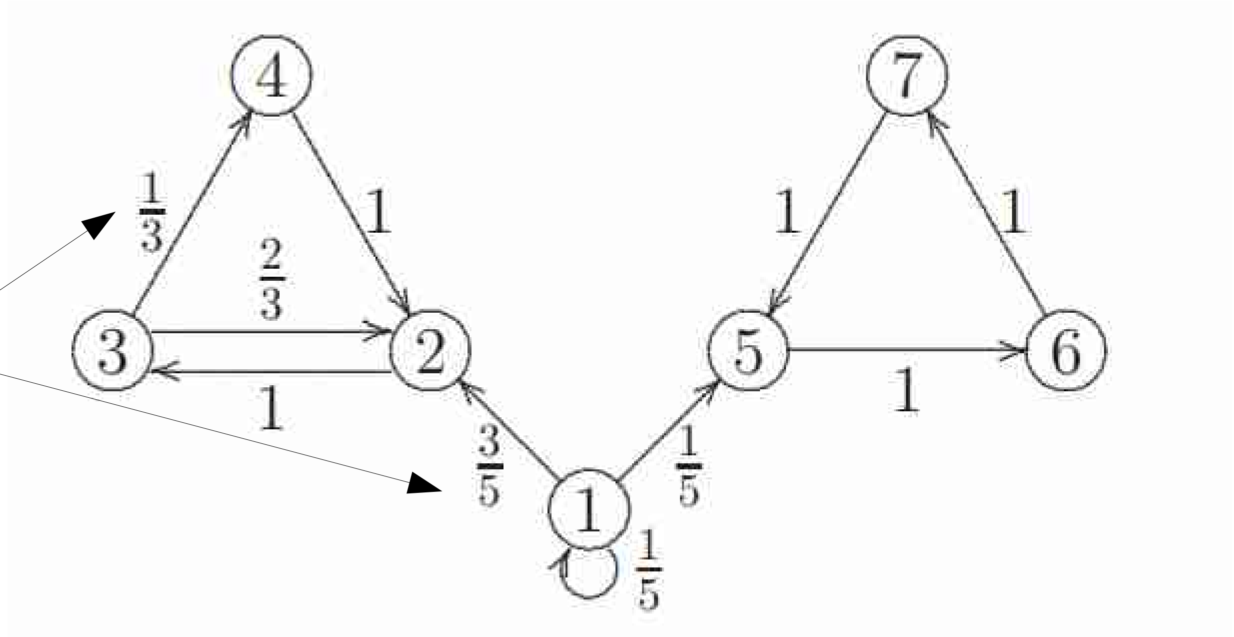
An Example

State transition probabilities



The Markov flea example: flea hopping around **at random** on this diagram **according to the probabilities** shown

An Example

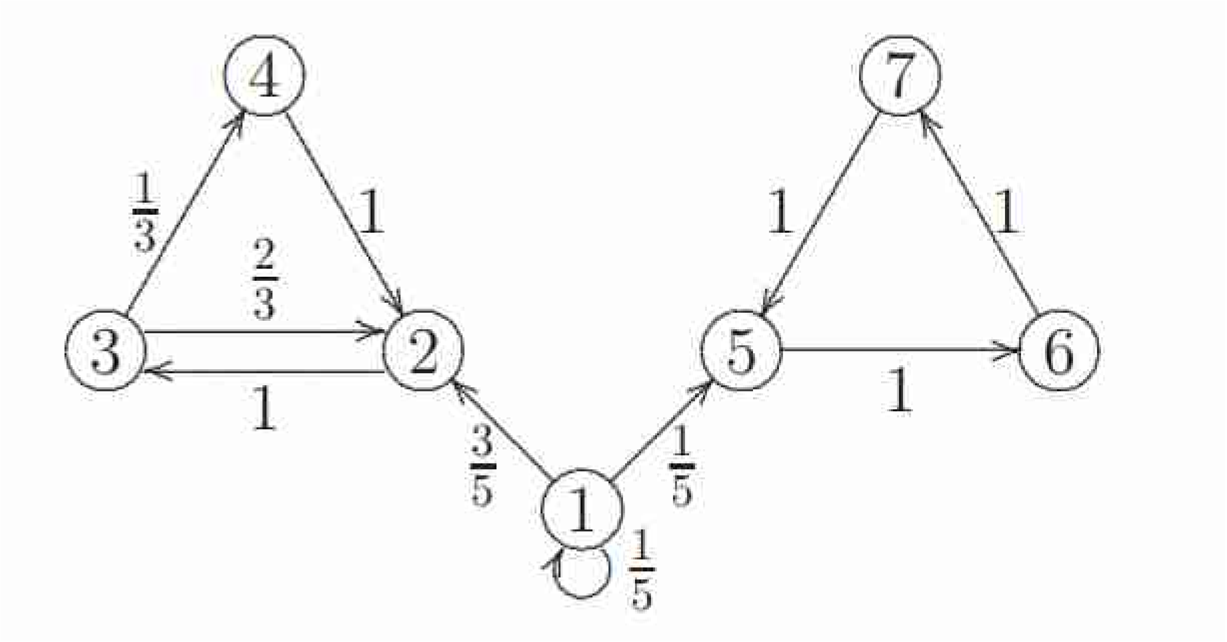


State transition probabilities

The Markov flea example: flea hopping around **at random** on this diagram **according to the probabilities** shown

State space *S* = *{1,2,3,4,5,6,7}*

An Example



- What is the probability of ever reaching state *7* from state *1*?
- Starting from state *2*, what is the expected time taken to reach state *4*?
- Starting from state *2*, what is the long-run proportion of time spent in state 3?
- Starting from state 1, what is the probability of being in state 2 at time *t*? Does the probability converge as t → ∞, and if so, to what?

Definitions

- The Markov chain is the process $X_0, X_1, X_2, \ldots$.
- **Definition:** The state of a Markov chain at time t is the value of X_t

  For example, if $X_t = 6$, we say the process is in state 6 at time t.
- **Definition**: The state space of a Markov chain, S, is the set of values that each X_t can take.

  For example, $S = \{1, 2, 3, 4, 5, 6, 7\}$.

  Let S have size N (possibly infinite).
- **Definition**: A trajectory of a Markov chain is a particular set of values for $X_0, X_1, X_2, \ldots$

  For example, if $X_0 = 1, X_1 = 5$, and $X_2 = 6$, then the trajectory up to time $t = 2$ is $1, 5, 6$.

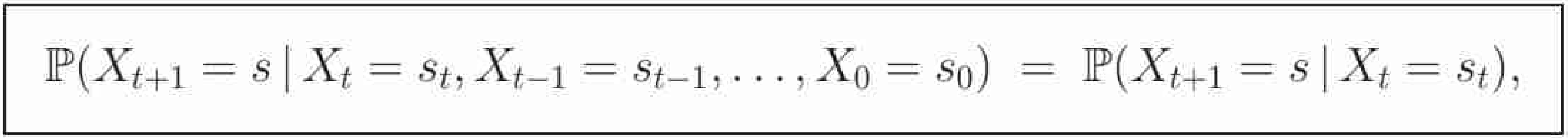
  More generally, if we refer to the trajectory $s_0, s_1, s_2, s_3, \ldots$ we mean that

  $X_0 = s_0, X_1 = s_1, X_2 = s_2, X_3 = s_3, \ldots$

  ‘Trajectory’ is just a word meaning ‘path'

Markov Property

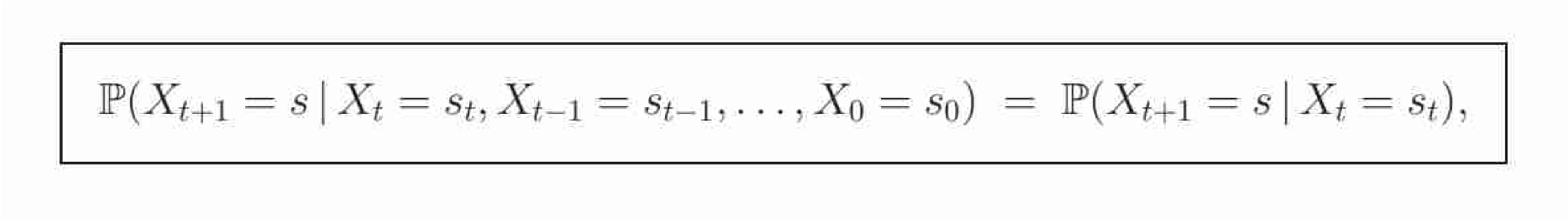
- Only the most recent point X_t affects what happens next, that is, X_{t+1} only depends on X_t, but not on X_{t-1}, X_{t-2}, . . .
- More formally:



$$\mathbb{P}(X_{t+1} = s \mid X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) \;=\; \mathbb{P}(X_{t+1} = s \mid X_t = s_t),$$

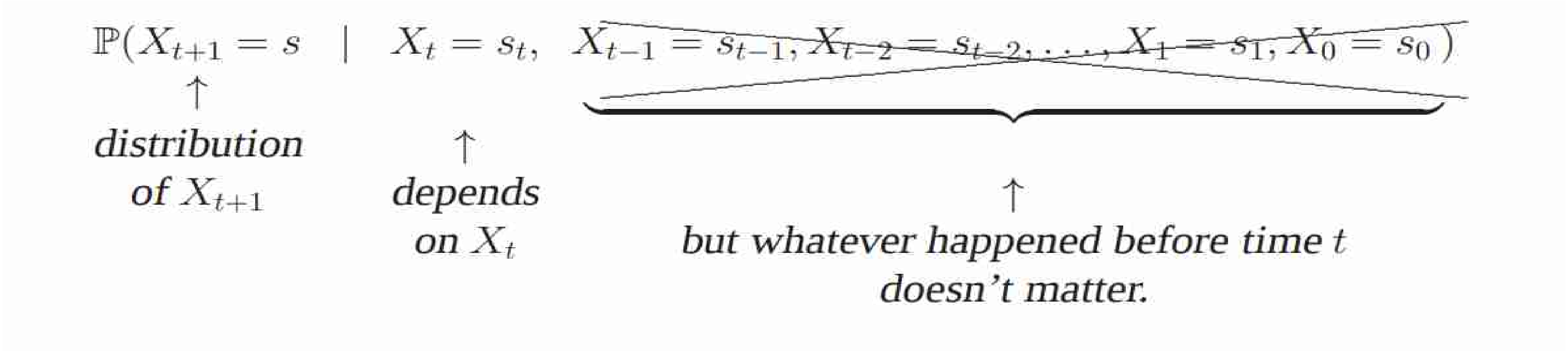
Markov Property

- Only the most recent point X_t affects what happens next, that is, X_{t+1} only depends on X_t, but not on X_{t-1}, X_{t-2}, . . .
- More formally:



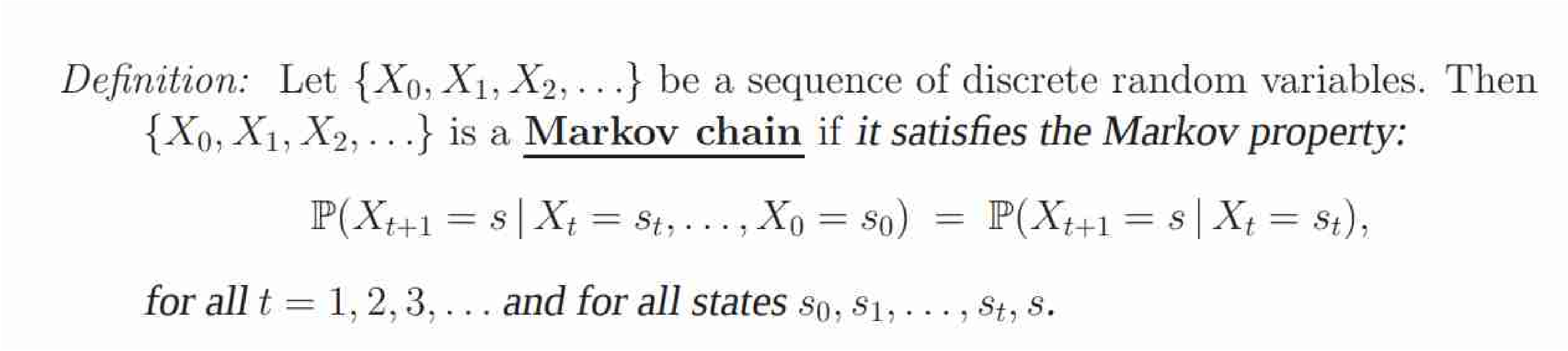
$$\boxed{\mathbb{P}(X_{t+1} = s \,|\, X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) \;=\; \mathbb{P}(X_{t+1} = s \,|\, X_t = s_t),}$$

- Explanation



$$\underset{\substack{\uparrow\\ \textit{distribution}\\ \textit{of } X_{t+1}}}{\mathbb{P}(X_{t+1} = s} \quad \Big| \quad \underset{\substack{\uparrow\\ \textit{depends}\\ \textit{on } X_t}}{X_t = s_t,} \quad \underbrace{\xcancel{X_{t-1} = s_{t-1}, X_{t-2} = s_{t-2}, \ldots, X_1 = s_1, X_0 = s_0\,)}}_{\substack{\uparrow\\ \textit{but whatever happened before time } t\\ \textit{doesn't matter.}}}$$

Definition



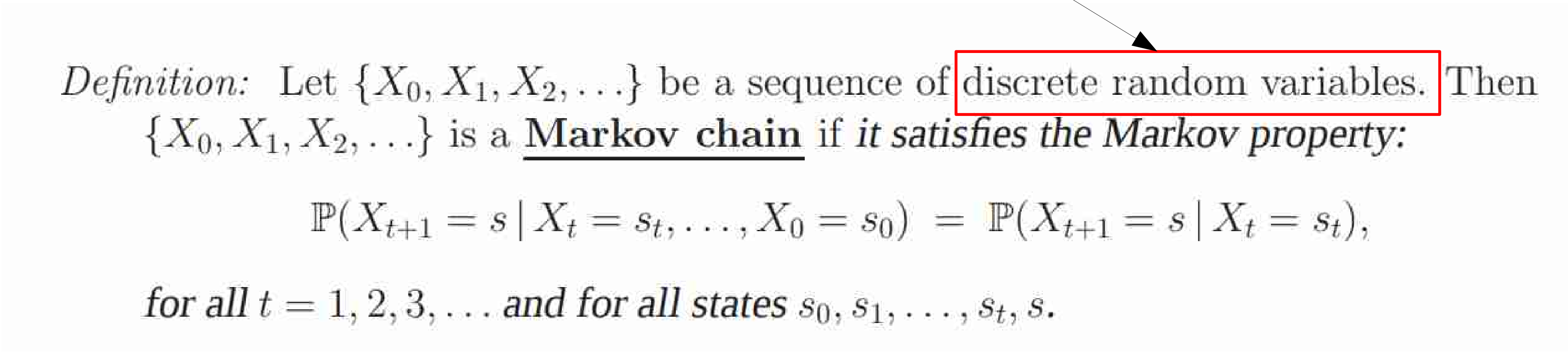
Definition: Let $\{X_0, X_1, X_2, \ldots\}$ be a sequence of discrete random variables. Then $\{X_0, X_1, X_2, \ldots\}$ is a **Markov chain** if *it satisfies the Markov property:*

$$\mathbb{P}(X_{t+1} = s \,|\, X_t = s_t, \ldots, X_0 = s_0) \;=\; \mathbb{P}(X_{t+1} = s \,|\, X_t = s_t),$$

for all $t = 1, 2, 3, \ldots$ *and for all states* $s_0, s_1, \ldots, s_t, s$.

Definition

Discrete states, e.g., *A, C, G, T*

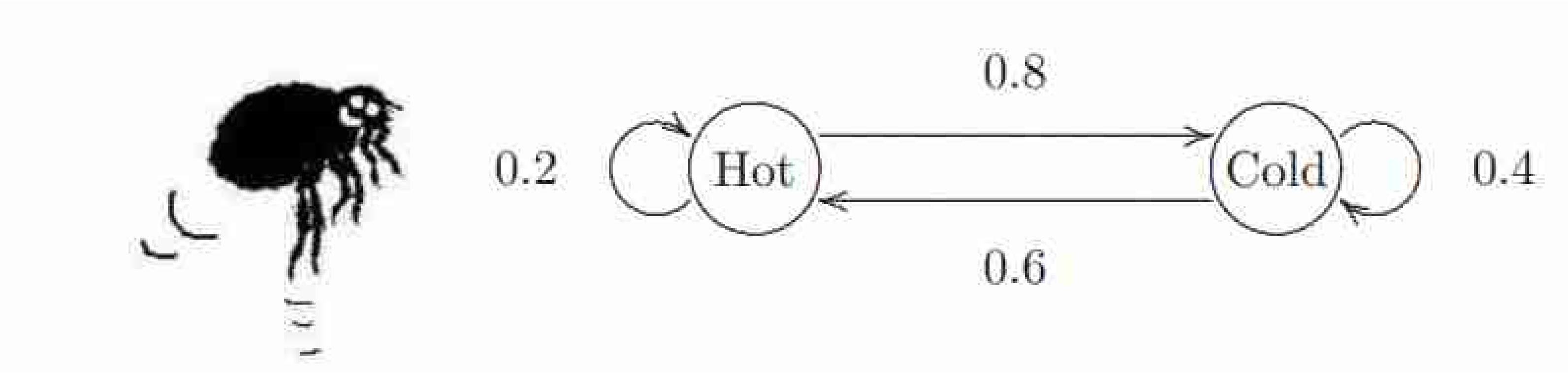


Definition: Let $\{X_0, X_1, X_2, \ldots\}$ be a sequence of discrete random variables. Then $\{X_0, X_1, X_2, \ldots\}$ is a **Markov chain** if *it satisfies the Markov property:*

$$\mathbb{P}(X_{t+1} = s \,|\, X_t = s_t, \ldots, X_0 = s_0) \;=\; \mathbb{P}(X_{t+1} = s \,|\, X_t = s_t),$$

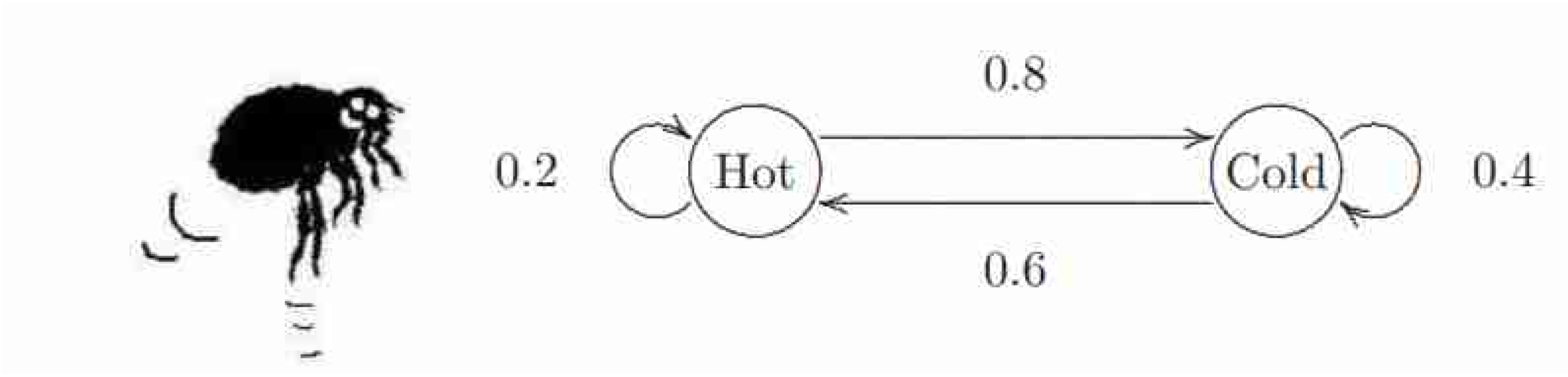
for all $t = 1, 2, 3, \ldots$ *and for all states* $s_0, s_1, \ldots, s_t, s$.

The Transition Matrix



Let us transform this into an equivalent transition matrix which is just another equivalent way of describing this diagram.

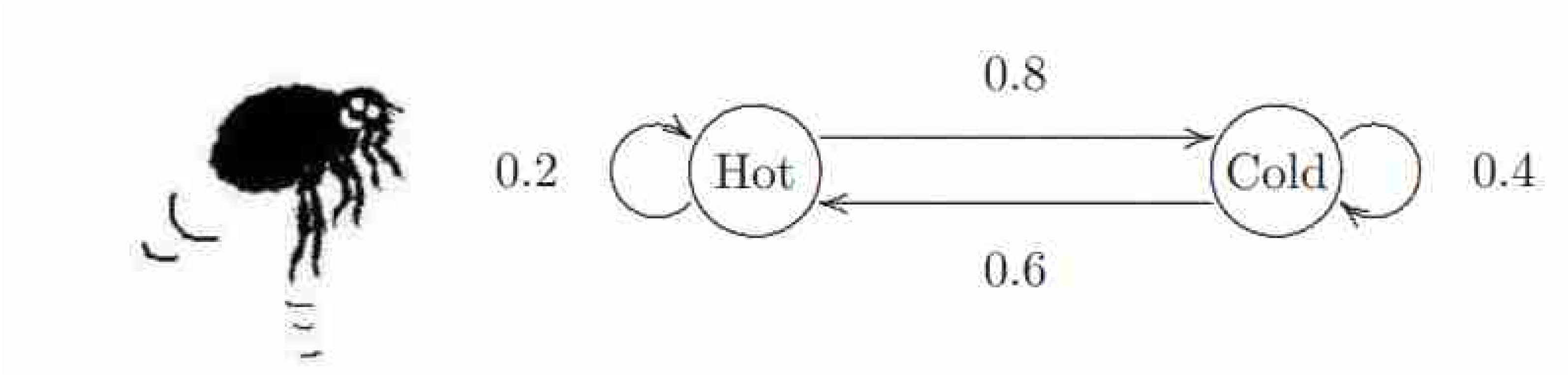
The Transition Matrix



Let us transform this into an equivalent transition matrix which is just another equivalent way of describing this diagram.

$$\begin{array}{ll} & \overbrace{\begin{array}{cc}\text{Hot} & \text{Cold}\end{array}}^{X_{t+1}} \\ X_t \left\{ \begin{array}{c}\text{Hot} \\ \text{Cold}\end{array} \right. & \begin{pmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{pmatrix} \end{array}$$

The Transition Matrix



Let us transform this into an equivalent transition matrix which is just another Equivalent way of describing this diagram.

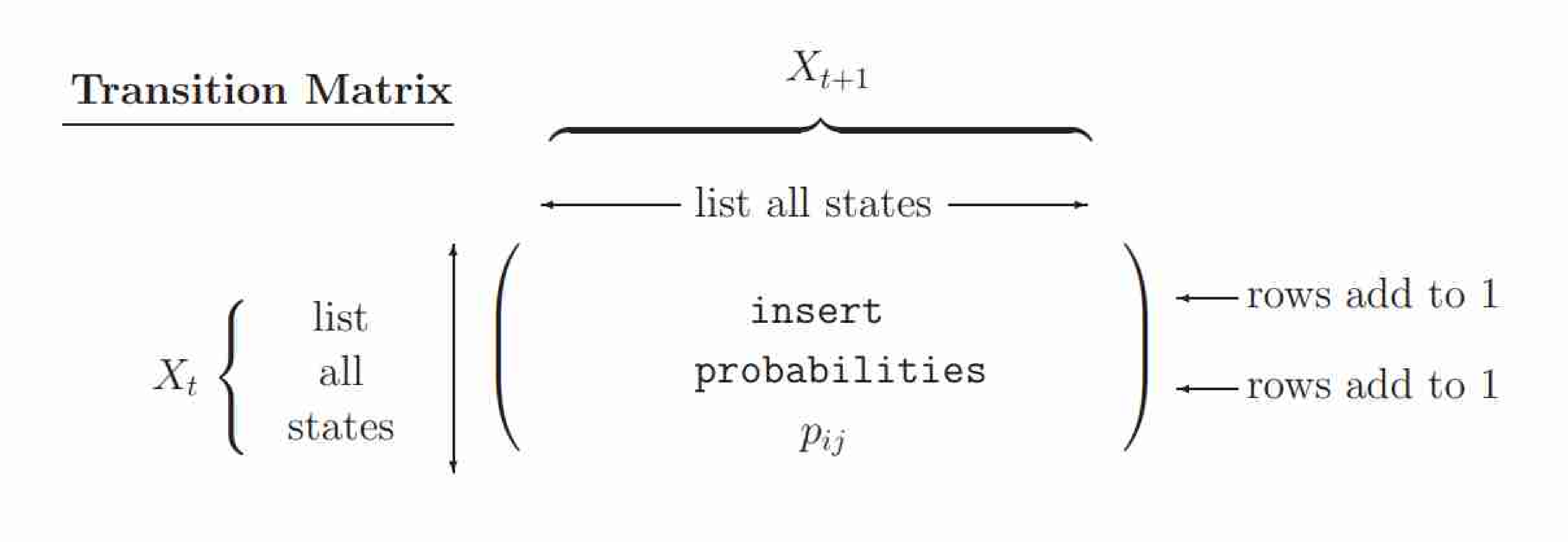
$$\begin{array}{cc} & \overbrace{\begin{array}{cc}\text{Hot} & \text{Cold}\end{array}}^{X_{t+1}} \\ X_t \left\{\begin{array}{c}\text{Hot}\\ \text{Cold}\end{array}\right. & \begin{pmatrix}0.2 & 0.8\\ 0.6 & 0.4\end{pmatrix}\end{array}$$

Values of states

States

Transition probabilities

More formally



Transition Matrix

X_{t+1}

list all states

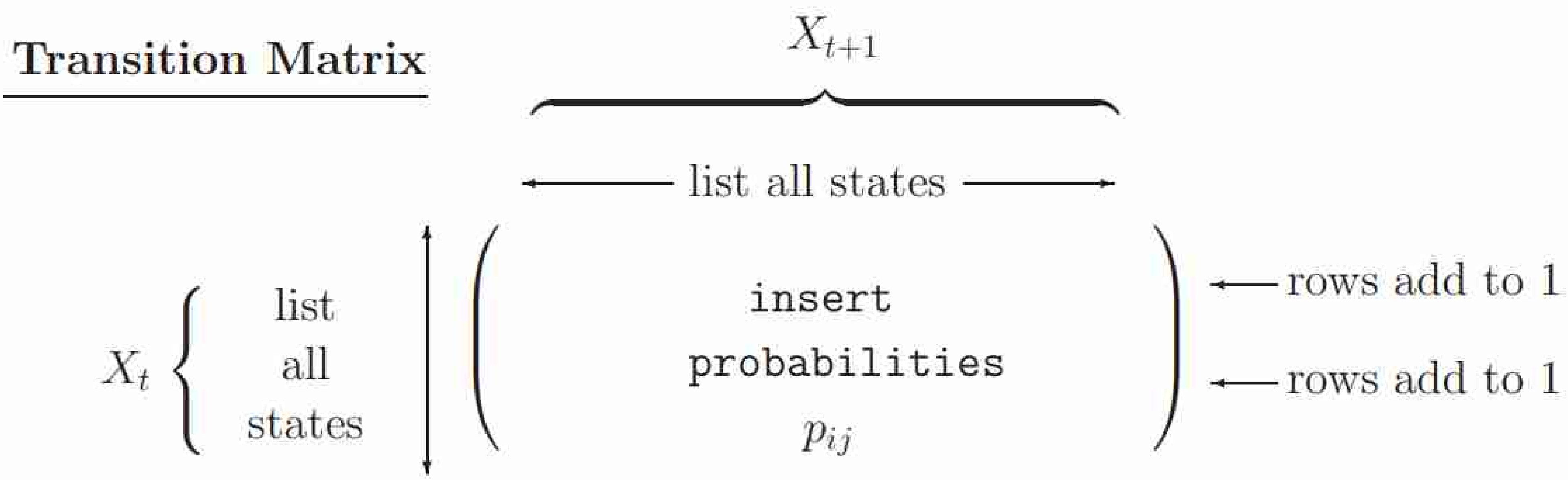
X_t list all states

insert probabilities p_{ij}

rows add to 1

rows add to 1

More formally



Transition Matrix

X_{t+1}

list all states

X_t

list all states

insert probabilities p_{ij}

rows add to 1

rows add to 1

The transition matrix is usually given the symbol *P* = (p_{ij})

In the transition matrix *P*:

the **ROWS** represent **NOW**, or **FROM X_t**

the **COLUMNS** represent **NEXT**, or **TO X_{t+1}**

Matrix entry *i,j* is the **CONDITIONAL** probability that **NEXT = *j***, given that **NOW = *i***: the probability of going **FROM** state ***i*** **TO** state ***j***.
$p_{ij} = P(X_{t+1} = j \mid X_t = i)$.

A Review of Probabilities

This is not a transition matrix!

		Hair color		
		brown	blonde	Σ
Eye color	light	5/40	15/40	20/40
	dark	15/40	5/40	20/40
	Σ	20/40	20/40	**40/40**

A Review of Probabilities

Hair color

		brown	blonde	Σ
Eye color	light	5/40	15/40	20/40
	dark	15/40	5/40	20/40
	Σ	20/40	20/40	**40/40**

Joint probability: probability of observing both A and B: *Pr(A,B)*
For instance, *Pr(brown, light) = 5/40 = 0.125*

A Review of Probabilities

Hair color

Eye color

	brown	blonde	Σ
light	5/40	15/40	20/40
dark	15/40	5/40	20/40
Σ	20/40	20/40	**40/40**

Marginalize over hair color

Marginal Probability: *unconditional* probability of an observation *Pr(A)*
For instance, *Pr(dark) = Pr(dark,brown) + Pr(dark,blonde) = 15/40 + 5/40 = 20/40 = 0.5*

A Review of Probabilities

		Hair color		
		brown	blonde	Σ
Eye color	light	5/40	15/40	20/40
	dark	15/40	5/40	20/40
	Σ	20/40	20/40	**40/40**

Conditional Probability: The probability of observing A given that B has occurred:
Pr(A|B) is the fraction of cases *Pr(B)* in which *B* occurs where *A* also occurs with *Pr(AB)*
Pr(A|B) = Pr(AB) / Pr(B)

For instance, *Pr(blonde|light) = Pr(blonde,light) / Pr(light) = (15/40) / (20/40) = 0.75*

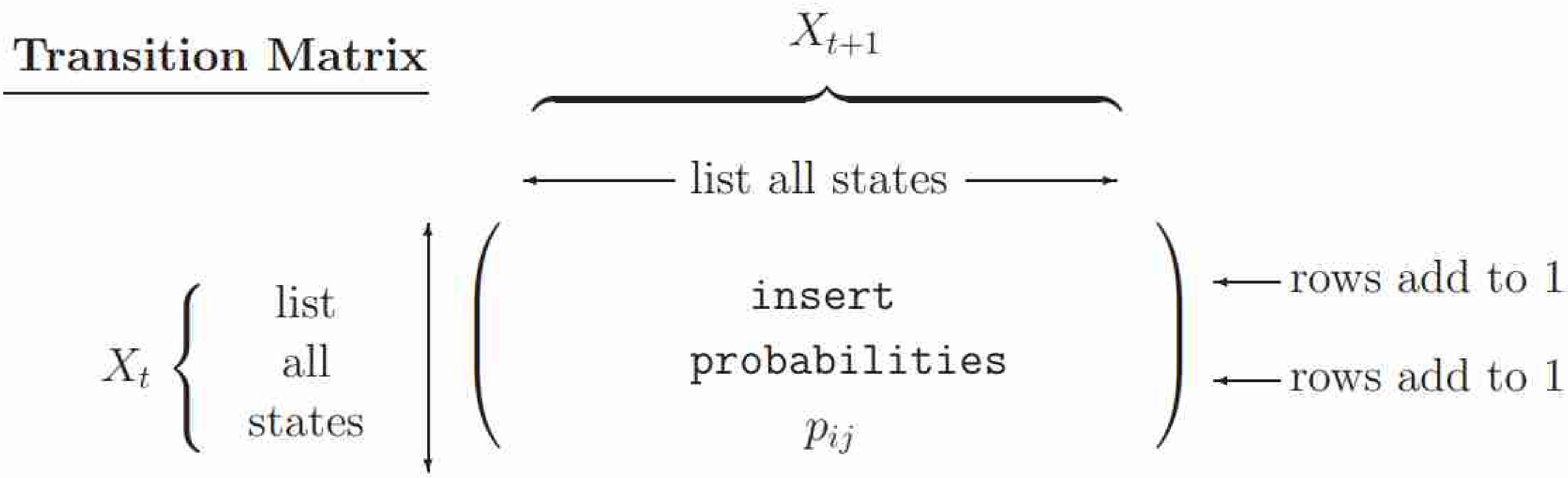
A Review of Probabilities

		Hair color		
		brown	blonde	Σ
Eye color	light	5/40	15/40	20/40
	dark	15/40	5/40	20/40
	Σ	20/40	20/40	**40/40**

Statistical Independence: Two events A and B are independent
If their joint probability *Pr(A,B)* equals the product of their marginal probability *Pr(A) Pr(B)*

For instance, *Pr(light,brown) ≠ Pr(light) Pr(brown)*, that is, the events are not independent!

More formally



Transition Matrix

X_{t+1}

list all states

X_t

list all states

insert probabilities p_{ij}

rows add to 1

rows add to 1

The transition matrix is usually given the symbol $P = (p_{ij})$

In the transition matrix P:

the **ROWS** represent **NOW**, or **FROM** $\boldsymbol{X_t}$

the **COLUMNS** represent **NEXT**, or **TO** $\boldsymbol{X_{t+1}}$

Matrix entry *i,j* is the **CONDITIONAL** probability that **NEXT =** ***j***, given that **NOW =** ***i***: the probability of going **FROM** state ***i*** **TO** state ***j***.
$p_{ij} = P(X_{t+1} = j \mid X_t = i)$.

Notes

1. The transition matrix *P* must list all possible states in the state space *S*.

2. *P* is a square *N* × *N* matrix, because X_{t+1} and X_t both take values in the same state space *S* of size *N*.

3. The **rows** of *P* should each sum to *1*:

$$\sum_{j=1}^{N} p_{ij} = \sum_{j=1}^{N} \mathbb{P}(X_{t+1} = j \mid X_t = i) = \sum_{j=1}^{N} \mathbb{P}_{\{X_t = i\}}(X_{t+1} = j) = 1.$$

The above simply states that X_{t+1} must take one of the listed values.

4. The columns of *P* do in general NOT sum to 1.

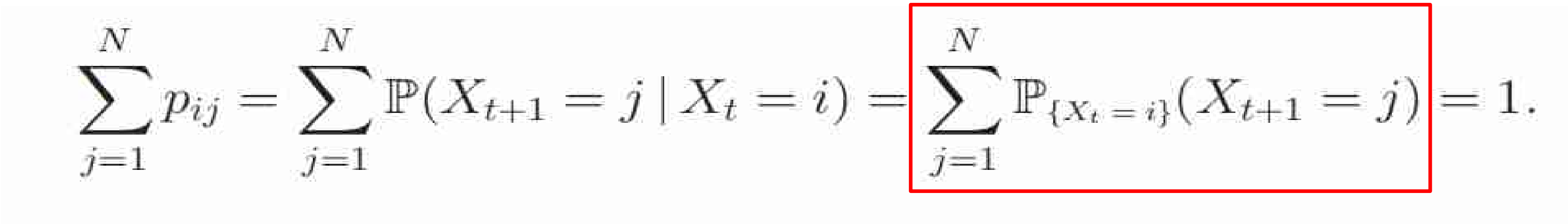
Notes

This is just another way of writing this conditional probability.

1. The transition matrix *P* must list all possible states in the state space *S*.

2. *P* is a square *N* × *N* matrix, because X_{t+1} and X_t both take values in the same state space *S* of size *N*.

3. The **rows** of *P* should each sum to *1*:



$$\sum_{j=1}^{N} p_{ij} = \sum_{j=1}^{N} \mathbb{P}(X_{t+1} = j \mid X_t = i) = \boxed{\sum_{j=1}^{N} \mathbb{P}_{\{X_t = i\}}(X_{t+1} = j)} = 1.$$

The above simply states that X_{t+1} must take one of the listed values.

4. The columns of *P* do in general NOT sum to 1.

t-step Transition Probabilites

- Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain with state space $S = \{1, 2, \ldots, N\}$
- Recall that the elements of the transition matrix P are defined as

  $(P)_{ij} = p_{ij} = P(X_1 = j \mid X_0 = i) = P(X_{n+1} = j \mid X_n = i)$ for any n.

- p_{ij} is the probability of making a transition **FROM** state *i* **TO** state **j** in a **SINGLE** step
- **Question:** what is the probability of making a transition from state *i* to state *j* over two steps? i.e. what is

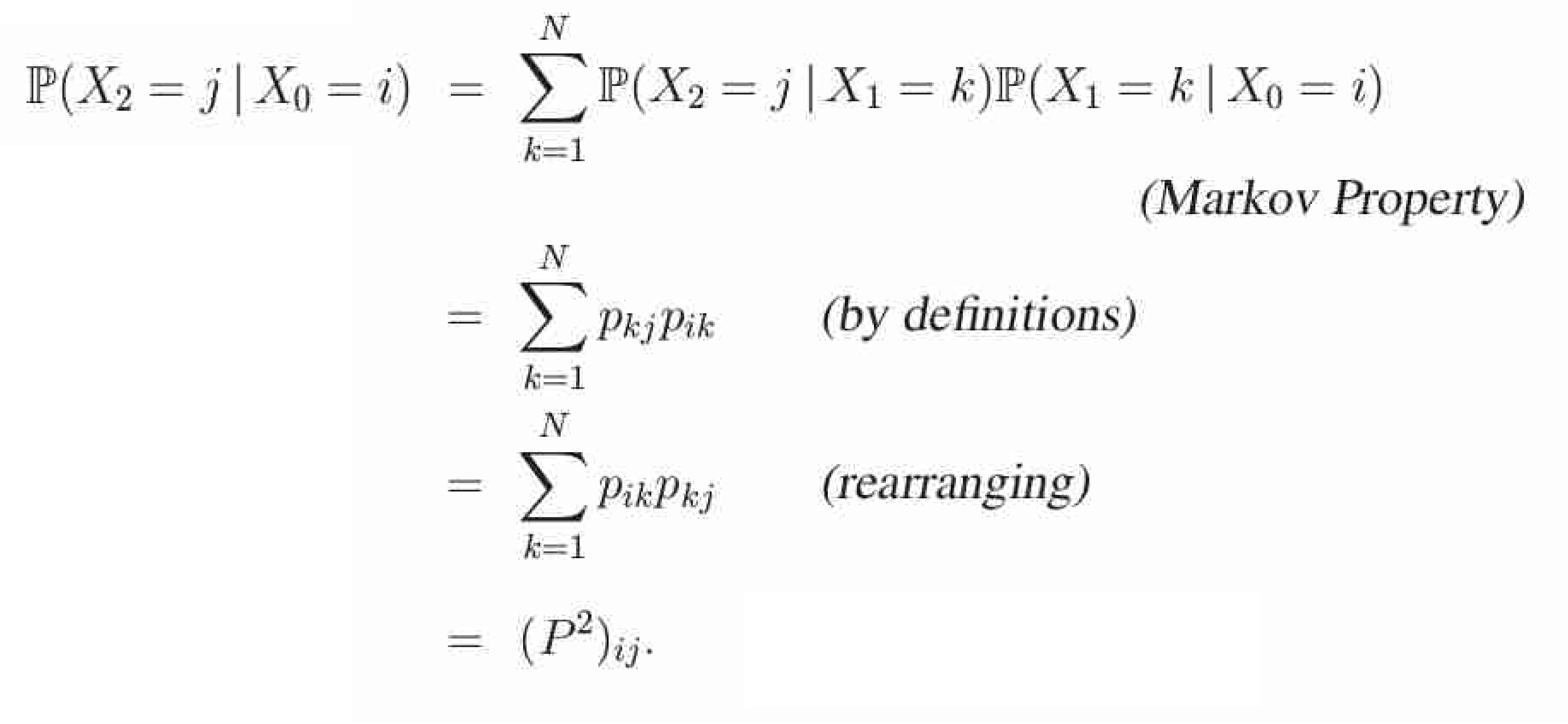
  $P(X_2 = j \mid X_0 = i)$?

t-step transition probs

$$\mathbb{P}(X_2 = j \mid X_0 = i) \;\; =$$

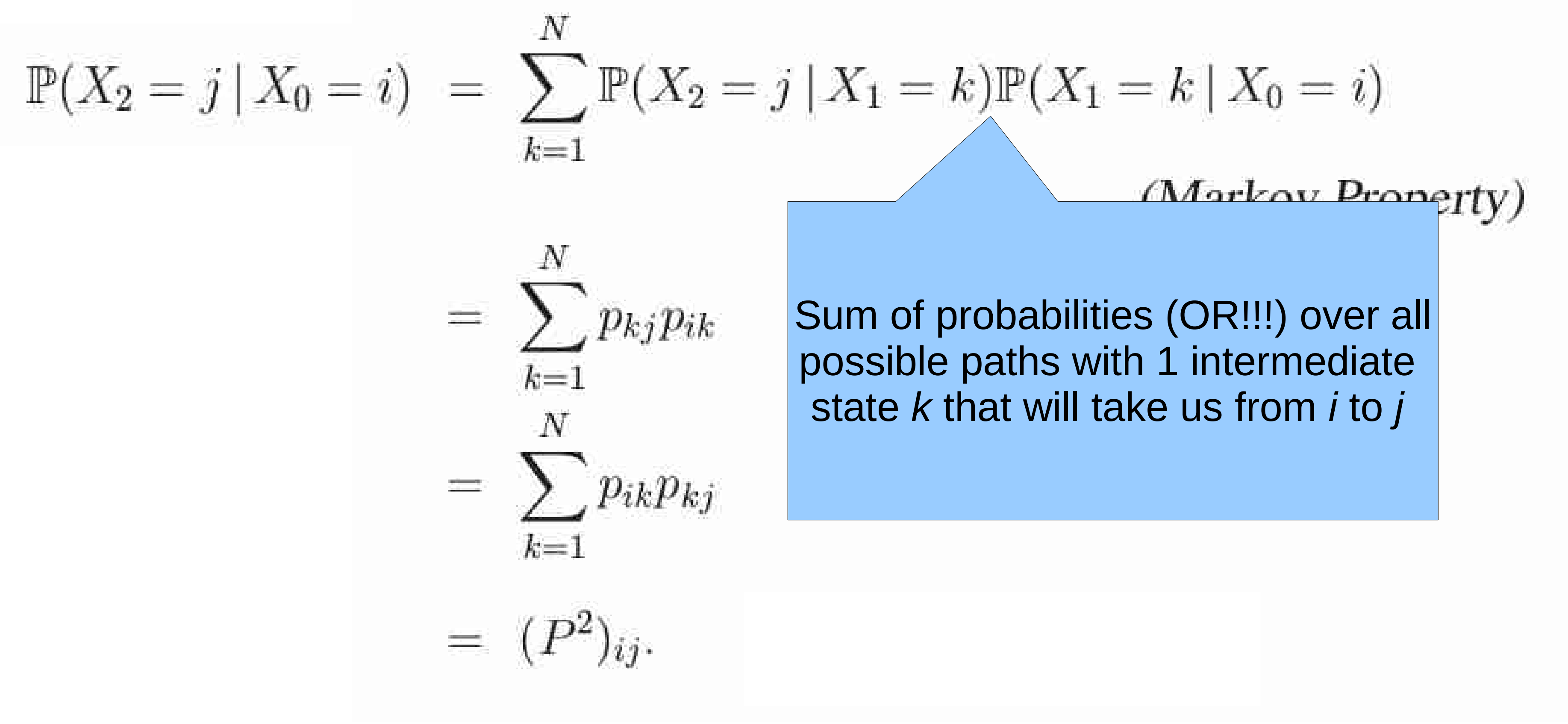
Any ideas?

t-step transition probs



$$\begin{aligned}
\mathbb{P}(X_2 = j \mid X_0 = i) &= \sum_{k=1}^{N} \mathbb{P}(X_2 = j \mid X_1 = k)\mathbb{P}(X_1 = k \mid X_0 = i) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{(Markov Property)} \\
&= \sum_{k=1}^{N} p_{kj} p_{ik} \qquad \textit{(by definitions)} \\
&= \sum_{k=1}^{N} p_{ik} p_{kj} \qquad \textit{(rearranging)} \\
&= (P^2)_{ij}.
\end{aligned}$$

t-step transition probs



$$\begin{aligned}
\mathbb{P}(X_2 = j \mid X_0 = i) &= \sum_{k=1}^{N} \mathbb{P}(X_2 = j \mid X_1 = k)\mathbb{P}(X_1 = k \mid X_0 = i) \\
&\qquad\qquad \textit{(Markov Property)} \\
&= \sum_{k=1}^{N} p_{kj} p_{ik} \\
&= \sum_{k=1}^{N} p_{ik} p_{kj} \\
&= (P^2)_{ij}.
\end{aligned}$$

Sum of probabilities (OR!!!) over all possible paths with 1 intermediate state *k* that will take us from *i* to *j*

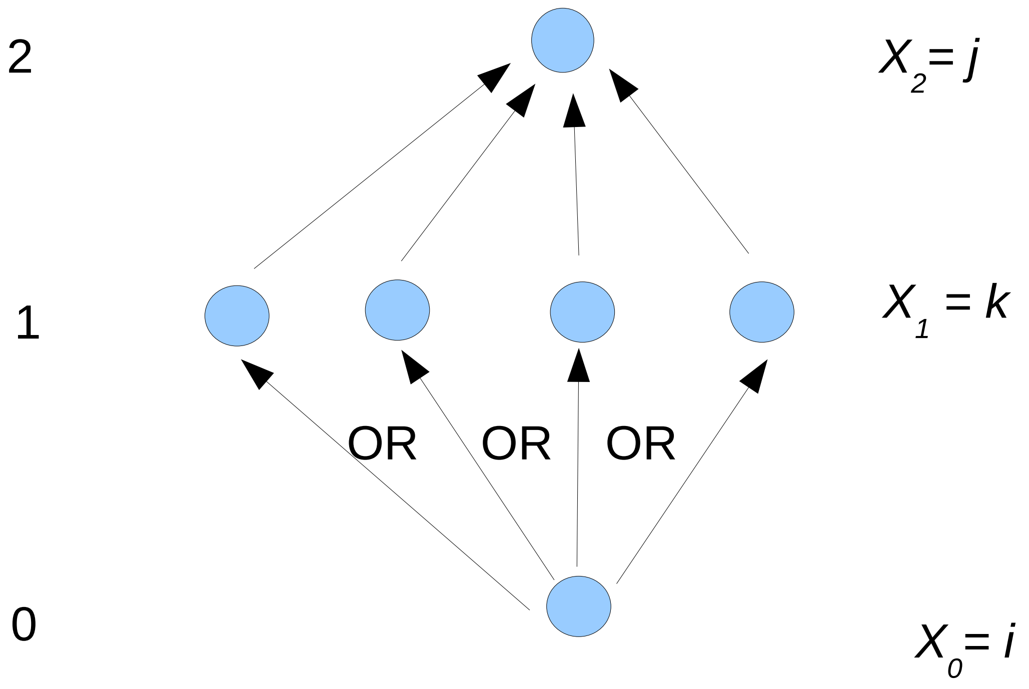
t-step transition probs

$$\begin{aligned}
\mathbb{P}(X_2 = j \mid X_0 = i) &= \sum_{k=1}^{N} \mathbb{P}(X_2 = j \mid X_1 = k)\mathbb{P}(X_1 = k \mid X_0 = i) \\
& \qquad\qquad\qquad\qquad\qquad\qquad \textit{(Markov Property)} \\
&= \sum_{k=1}^{N} p_{kj} p_{ik} \qquad \textit{(by definitions)} \\
&= \sum_{k=1}^{N} p_{ik} p_{kj} \qquad \textit{(rearranging)} \\
&= (P^2)_{ij}.
\end{aligned}$$

The two step-transition probabilities, in fact, for any *n* are thus:

$$\mathbb{P}(X_2 = j \mid X_0 = i) = \mathbb{P}(X_{n+2} = j \mid X_n = i) = \left(P^2\right)_{ij}$$

All possible paths



2

$X_2 = j$

1

$X_1 = k$

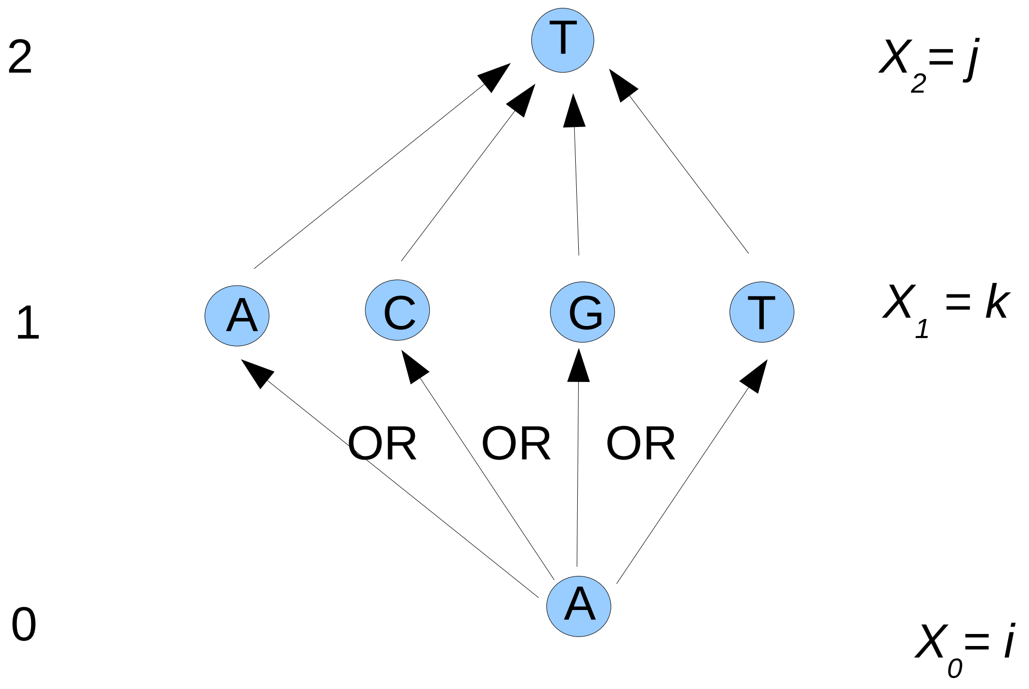
OR OR OR

0

$X_0 = i$

Sum over k

All possible paths



$X_2 = j$

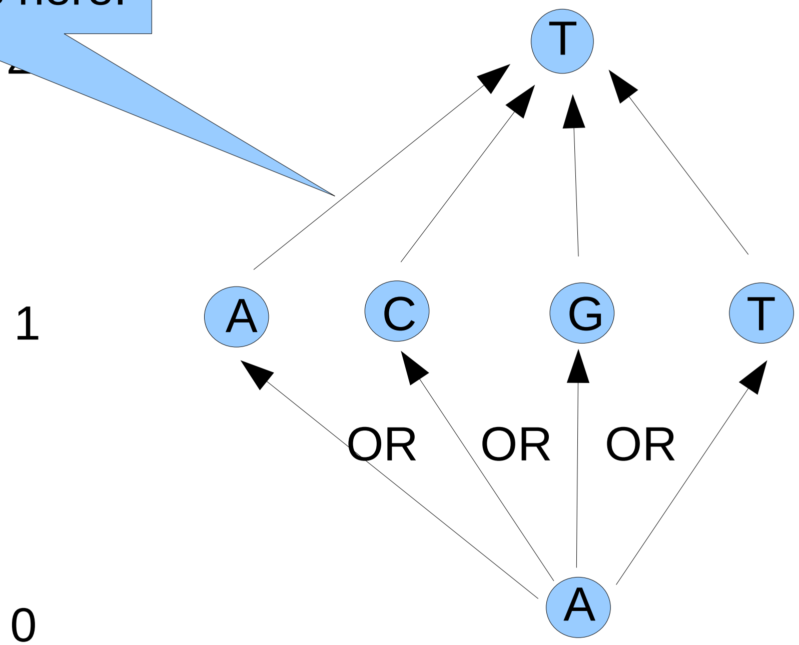
$X_1 = k$

$X_0 = i$

Sum over k

All possible paths

We are still thinking
In discrete steps here!



$X_2 = j$

$X_1 = k$

Sum over k

$X_0 = i$

3-step transitions

- What is: $P(X_3 = j \mid X_0 = i)$?

3-step and *t*-step transitions

- What is: $P(X_3 = j \mid X_0 = i)$?

  → $(P^3)_{ij}$

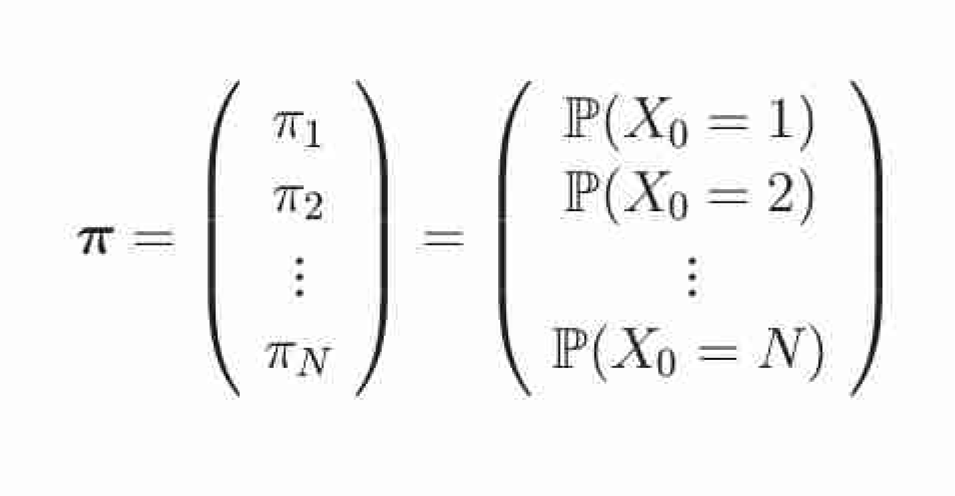
- General case with *t* steps for any *t* **and** any *n*

$$\mathbb{P}(X_t = j \mid X_0 = i) = \mathbb{P}(X_{n+t} = j \mid X_n = i) = \left(P^t\right)_{ij}$$

Distribution of X_t

- Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain with state space $S = \{1, 2, \ldots, N\}$.
- Now each X_t is a random variable → it has a **probability distribution**.
- We can write down the probability distribution of X_t as vector with N elements.
- For example, consider X_0 . Let π be a vector with N elements denoting the probability distribution of $X_{0.}$

The π vector

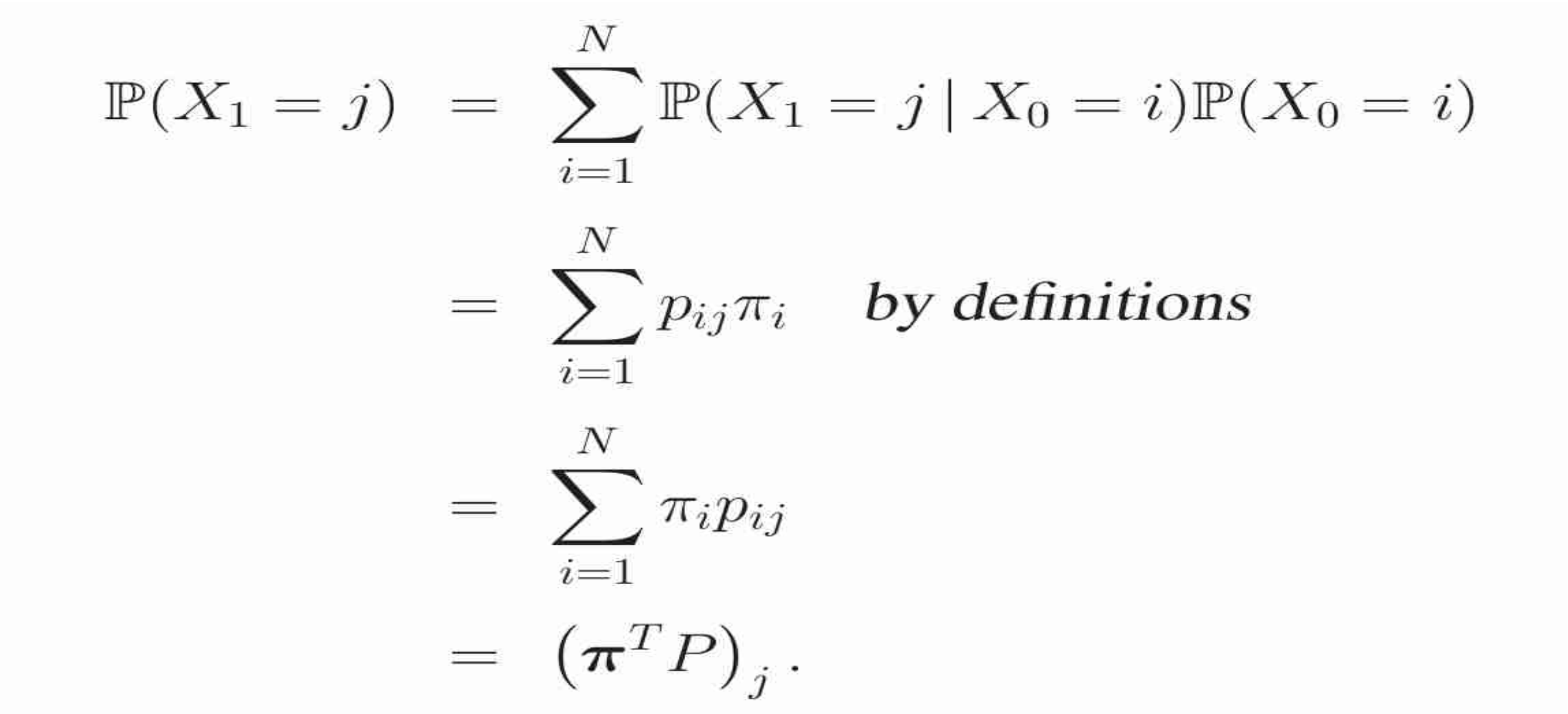


$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{pmatrix} = \begin{pmatrix} \mathbb{P}(X_0 = 1) \\ \mathbb{P}(X_0 = 2) \\ \vdots \\ \mathbb{P}(X_0 = N) \end{pmatrix}$$

This means that our Markov process choses at random in which state (e.g., A, C, G, or T) it **starts** with probability: P(start in state A) = π_A

This is why those vectors are also called prior probabilities.

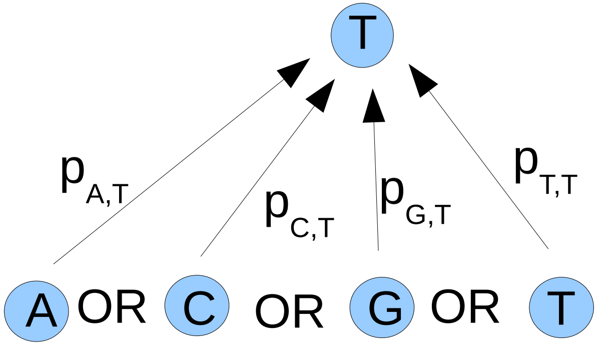
Probability of X_1



$$\begin{aligned}
\mathbb{P}(X_1 = j) &= \sum_{i=1}^{N} \mathbb{P}(X_1 = j \mid X_0 = i)\mathbb{P}(X_0 = i) \\
&= \sum_{i=1}^{N} p_{ij}\pi_i \quad \textit{by definitions} \\
&= \sum_{i=1}^{N} \pi_i p_{ij} \\
&= \left(\boldsymbol{\pi}^T P\right)_j .
\end{aligned}$$

So, here we are asking what the probability of ending up in state *j* at X_1 is, for starting in all possible states *N* at X_0

All possible paths



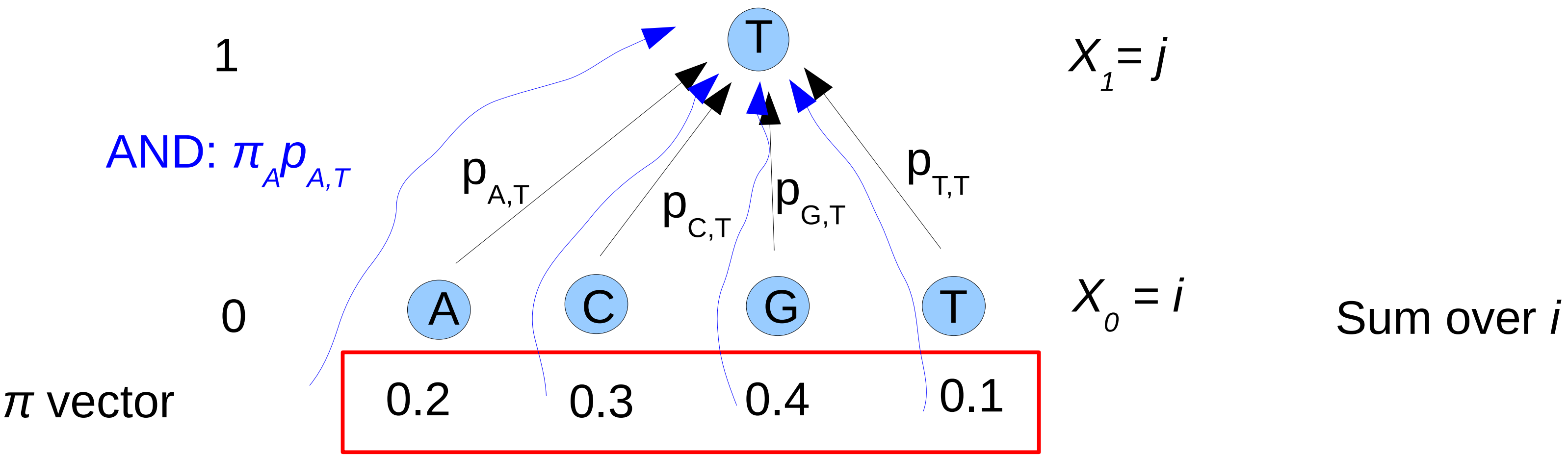
1 T $X_1 = j$

$p_{A,T}$ $p_{C,T}$ $p_{G,T}$ $p_{T,T}$

0 A OR C OR G OR T $X_0 = i$ Sum over *i*

π vector	0.2	0.3	0.4	0.1

All possible paths



1

AND: $\pi_A p_{A,T}$

0

π vector

T

$p_{A,T}$

$p_{C,T}$

$p_{G,T}$

$p_{T,T}$

A C G T

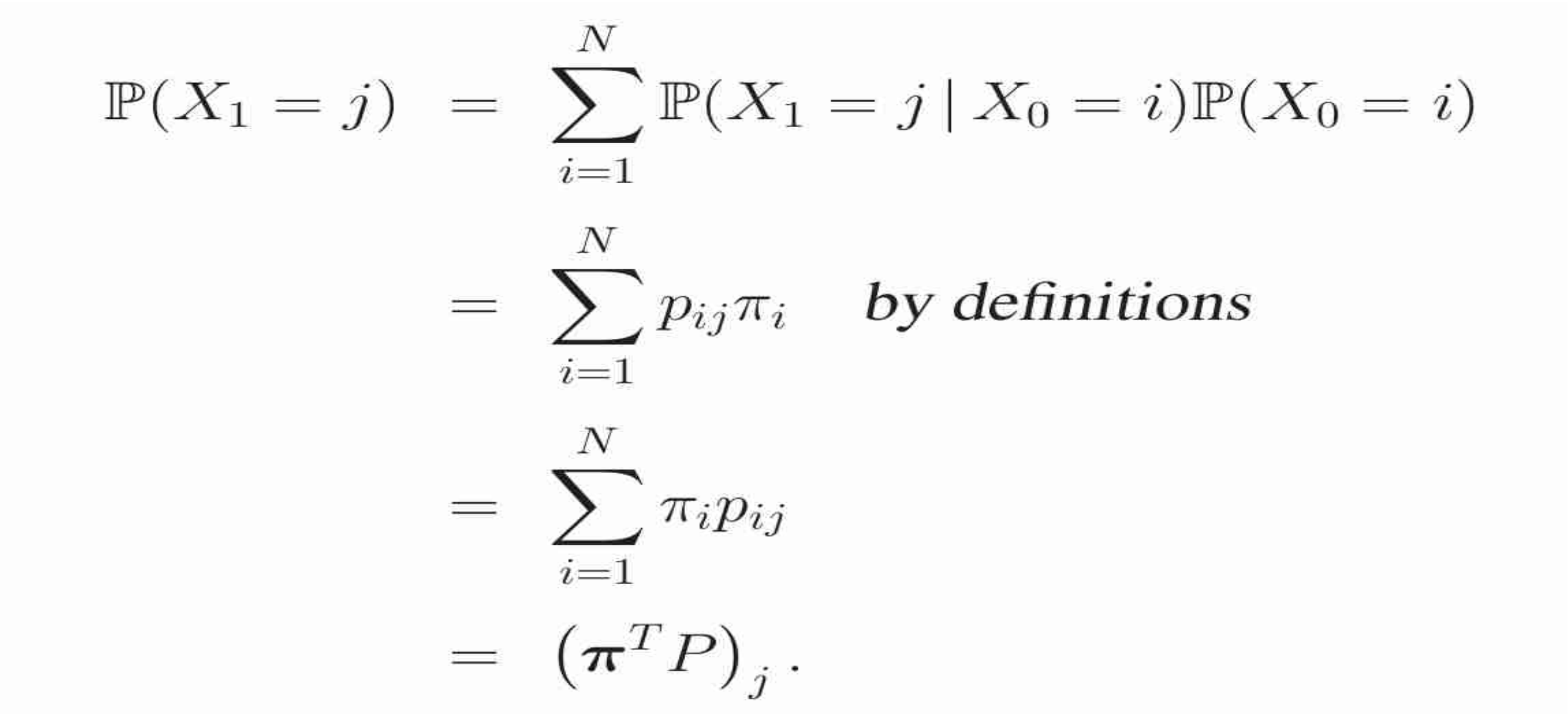
0.2 0.3 0.4 0.1

$X_1 = j$

$X_0 = i$

Sum over *i*

Probability Distribution of X_1



$$\begin{aligned}\mathbb{P}(X_1 = j) &= \sum_{i=1}^{N} \mathbb{P}(X_1 = j \mid X_0 = i)\mathbb{P}(X_0 = i) \\ &= \sum_{i=1}^{N} p_{ij}\pi_i \quad \textbf{\textit{by definitions}} \\ &= \sum_{i=1}^{N} \pi_i p_{ij} \\ &= \left(\boldsymbol{\pi}^T P\right)_j .\end{aligned}$$

This shows that $P(X_1 = j) = \pi^T P_j$ for all j .

The row vector $\pi^T P$ is therefore the probability distribution over all possible states for X_1, more formally:

$X_0 \sim \pi^T$

$X_1 \sim \pi^T P$

Distribution of X_2

- What do you think?

Distribution of X_2

- What do you think?

$$\mathbb{P}(X_2 = j) = \sum_{i=1}^{N} \mathbb{P}(X_2 = j \mid X_0 = i)\mathbb{P}(X_0 = i) = \sum_{i=1}^{N} \left(P^2\right)_{ij} \pi_i = \left(\boldsymbol{\pi}^T P^2\right)_j .$$

and in general:

$$\boxed{\begin{aligned} X_0 &\sim \boldsymbol{\pi}^T \\ X_1 &\sim \boldsymbol{\pi}^T P \\ X_2 &\sim \boldsymbol{\pi}^T P^2 \\ &\vdots \\ X_t &\sim \boldsymbol{\pi}^T P^t. \end{aligned}}$$

Theorem

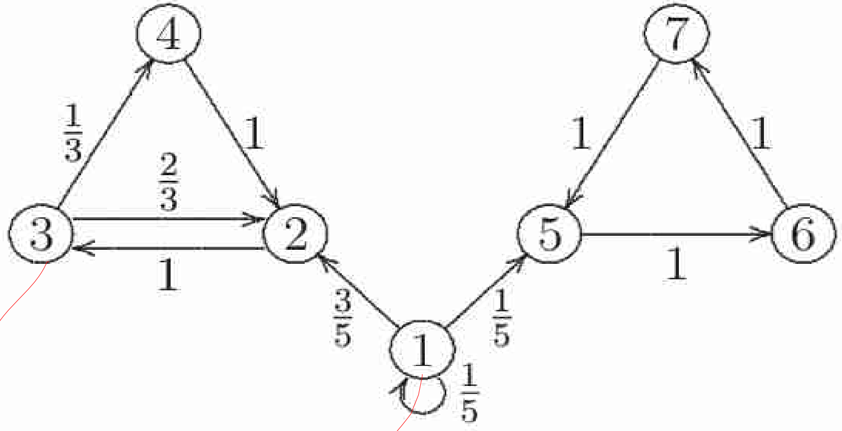
- Let $\{X_0 , X_1 , X_2 , . . .\}$ be a Markov chain with a $N \times N$ transition matrix P.
- If the probability distribution of X_0 is given by the $1 \times N$ row vector π^T, then the probability distribution of X_t is given by the $1 \times N$ row vector $\pi^T P^t$. That is,

  $X_0 \sim \pi^T \Rightarrow X_t \sim \pi^T P^t$.

Example – Trajectory probability

Recall that a trajectory is a sequence of values for $X_0, X_1, \ldots, X_t$.

Because of the Markov Property, we can find the probability of any trajectory by multiplying together the starting probability and all subsequent single-step probabilities.



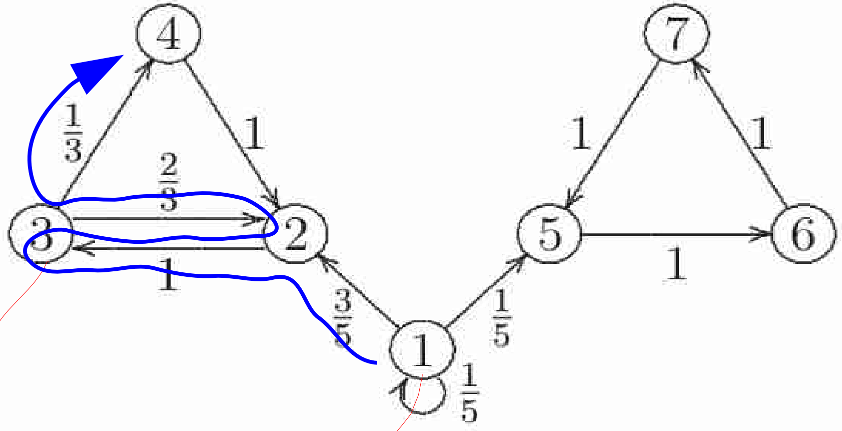
Example: Let $X_0 \sim (\frac{3}{4}, 0, \frac{1}{4}, 0, 0, 0, 0)$. What is the probability of the trajectory 1, 2, 3, 2, 3, 4?

$$\begin{aligned}
\mathbb{P}(1,2,3,2,3,4) &= \mathbb{P}(X_0 = 1) \times p_{12} \times p_{23} \times p_{32} \times p_{23} \times p_{34} \\
&= \frac{3}{4} \times \frac{3}{5} \times 1 \times \frac{2}{3} \times 1 \times \frac{1}{3} \\
&= \frac{1}{10}.
\end{aligned}$$

Example – Trajectory probability

Recall that a trajectory is a sequence of values for $X_0, X_1, \ldots, X_t$.

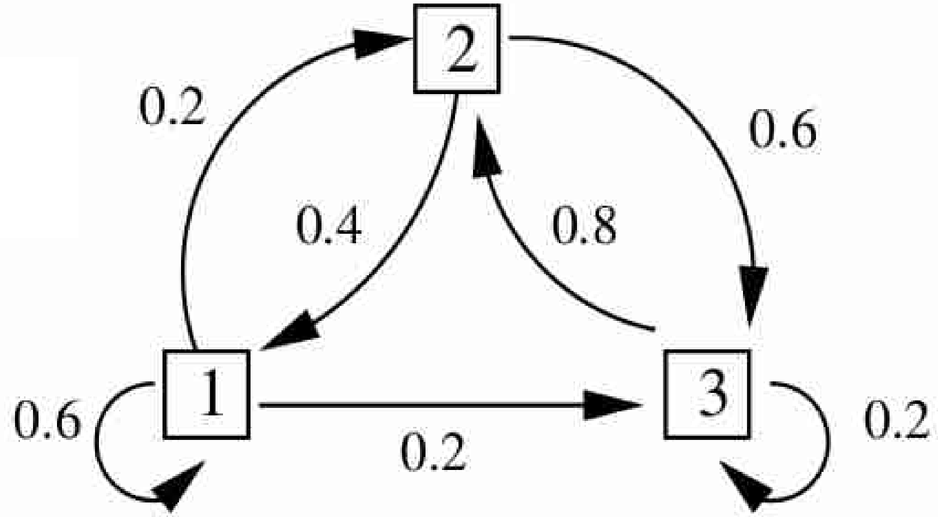
Because of the Markov Property, we can find the probability of any trajectory by multiplying together the starting probability and all subsequent single-step probabilities.



Example: Let $X_0 \sim (\frac{3}{4}, 0, \frac{1}{4}, 0, 0, 0, 0)$. What is the probability of the trajectory 1, 2, 3, 2, 3, 4?

$$\begin{aligned}
\mathbb{P}(1,2,3,2,3,4) &= \mathbb{P}(X_0 = 1) \times p_{12} \times p_{23} \times p_{32} \times p_{23} \times p_{34} \\
&= \frac{3}{4} \times \frac{3}{5} \times 1 \times \frac{2}{3} \times 1 \times \frac{1}{3} \\
&= \frac{1}{10}.
\end{aligned}$$

Exercise



- Find the transition matrix *P*
- Find *$P(X_2=3 \mid X_0=1)$*
- Suppose that the process is equally likely to start in any state at time 0
  → Find the probability distribution of *X_1*
- *Suppose that the process begins in state 1 at time 0*
  → *Find the probability distribution of X_2*
- *Suppose that the process is equally likely to start in any state at time 0*
  → *Find the probability of obtaining the trajectory (3, 2, 1, 1, 3).*

Class Structure

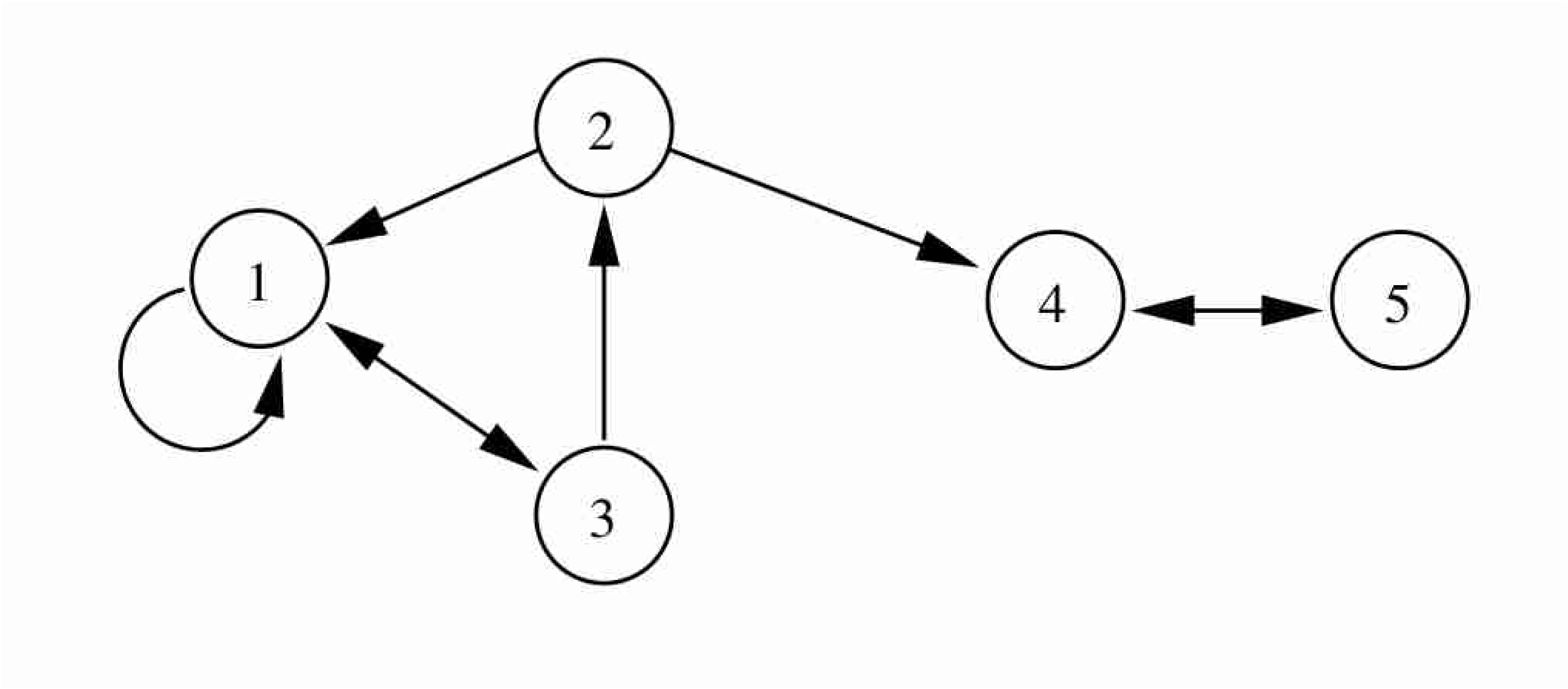
- The state space of a Markov chain can be partitioned into a set of non-overlapping *communicating classes*.
- States *i* and *j* are in the same communicating class if there is some way of getting from state $i \rightarrow j$, **AND** there is some way of getting from state $j \rightarrow i$.
- It needn't be possible to get from $i \rightarrow j$ in a single step, but it must be possible over some number of steps to travel between them both ways.
- We write: $i \leftrightarrow j$

Definition

- Consider a Markov chain with state space *S* and transition matrix *P*, and consider states *i, j in S.* Then state *i* communicates with state *j* if:
  - there exists some *t* such that $(P^t)_{ij} > 0$, **AND**
  - there exists some *u* such that $(P^u)_{ji} > 0$.
- Mathematically, it is easy to show that the communicating relation ↔ is an equivalence relation, which means that it *partitions* the state space *S* into *non-overlapping* equivalence classes.
- **Definition:** States *i* and *j* are in the same communicating class if $i \leftrightarrow j$: i.e., if each state is accessible from the other.
- Every state is a member of *exactly one* communicating class.

Example

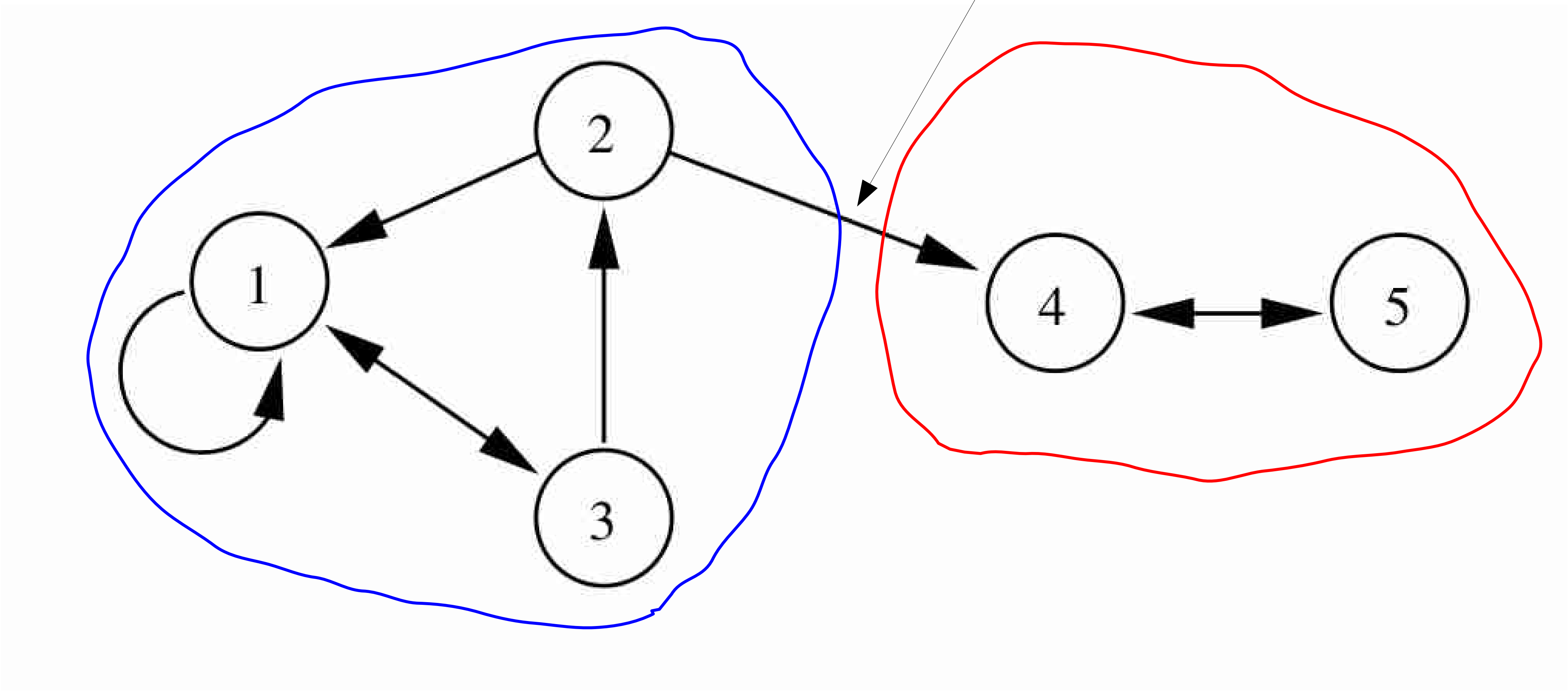
- Find the communicating classes!



Example

- Find the communicating classes!

No way back!



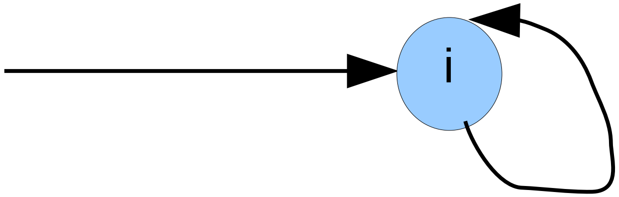
{1, 2, 3} and {4, 5}

Properties of Communicating Classes

- **Definition:** A communicating class of states is closed if it is not possible to leave that class.

  That is, the communicating class *C* is closed if $p_{ij} = 0$ whenever *i in C* and *j not in C*

- **Example:** In the transition diagram from the last slide:
  - Class *{1, 2, 3}* is not closed: it is possible to escape to class *{4, 5}*
  - Class *{4, 5}* is closed: it is not possible to escape.
- **Definition:** A state *i* is said to be absorbing if the set *{i}* is a closed class.

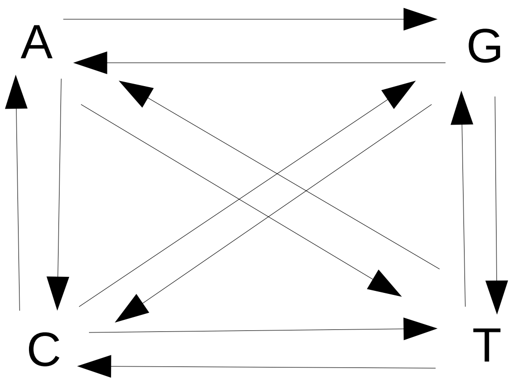


Irreducibility

- **Definition:** A Markov chain or transition matrix P is said to be **irreducible** if $i \leftrightarrow j$ (i communicates with j) for all $i, j \in S$. That is, the chain is irreducible if the state space S is a single communicating class.
- Do you know an example for an irreducible transition matrix P?

Irreducibility

- **Definition:** A Markov chain or transition matrix P is said to be **irreducible** if $i \leftrightarrow j$ for all $i, j \in S$. That is, the chain is irreducible if the state space S is a single communicating class.
- Do you know an example for an irreducible transition matrix P?



Equilibrium

- We saw that if $\{X_0, X_1, X_2, \ldots\}$ is a Markov chain with transition matrix P, then $X_t \sim \pi^T \Rightarrow X_{t+1} \sim \pi^T P$
- **Question:** is there any distribution π at some time t such that $\pi^T P = \pi^T$?
- If $\pi^T P = \pi^T$, *then*

$$
\begin{aligned}
X_t \sim \pi^T \quad & \Rightarrow X_{t+1} \sim \pi^T P = \pi^T \\
& \Rightarrow X_{t+2} \sim \pi^T P = \pi^T \\
& \Rightarrow X_{t+3} \sim \pi^T P = \pi^T \\
& \Rightarrow \ldots
\end{aligned}
$$

Equilibrium

- We saw that if $\{X_0, X_1, X_2, . . .\}$ is a Markov chain with transition matrix P, then $X_t \sim \pi^T \Rightarrow X_{t+1} \sim \pi^T P$
- **Question:** is there any distribution π at some time t such that $\pi^T P = \pi^T$?
- If $\pi^T P = \pi^T$, *then*

$$
\begin{aligned}
X_t \sim \pi^T \quad & \Rightarrow X_{t+1} \sim \pi^T P = \pi^T \\
& \Rightarrow X_{t+2} \sim \pi^T P = \pi^T \\
& \Rightarrow X_{t+3} \sim \pi^T P = \pi^T \\
& \Rightarrow . . .
\end{aligned}
$$

- In other words, if $\pi^T P = \pi^T$ AND $X_t \sim \pi^T$, then

$$X_t \sim X_{t+1} \sim X_{t+2} \sim X_{t+3} \sim . . .$$

- Thus, once a Markov chain has reached a distribution π^T such that $\pi^T P = \pi^T$, **it will stay there**

Equilibrium

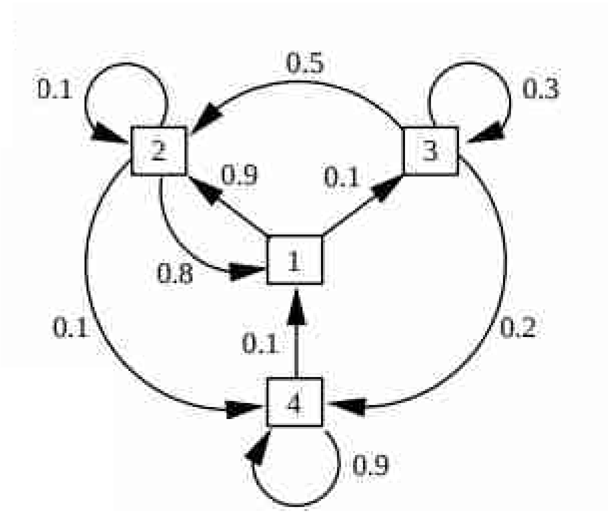
- If $\pi^T P = \pi^T$, we say that the distribution π^T is an equilibrium distribution.
- Equilibrium means there will be no further change in the distribution of X_t as we wander through the Markov chain.
- **Note:** Equilibrium **does not mean** that the actual **value** of X_{t+1} equals the value of X_t
- It means that the distribution of X_{t+1} is the same as the distribution of X_t, e.g.

  $P(X_{t+1} = 1) = P(X_t = 1) = \pi_1$;

  $P(X_{t+1} = 2) = P(X_t = 2) = \pi_2$, etc.

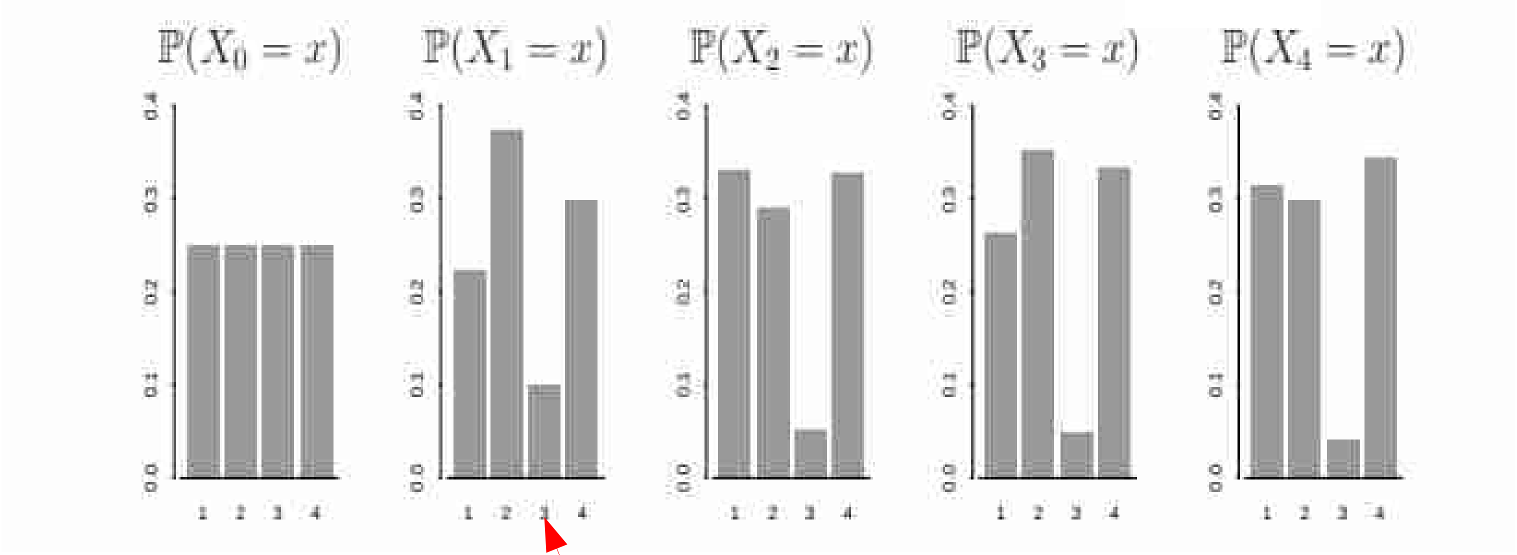
Example

$$P = \begin{pmatrix} 0.0 & 0.9 & 0.1 & 0.0 \\ 0.8 & 0.1 & 0.0 & 0.1 \\ 0.0 & 0.5 & 0.3 & 0.2 \\ 0.1 & 0.0 & 0.0 & 0.9 \end{pmatrix}$$



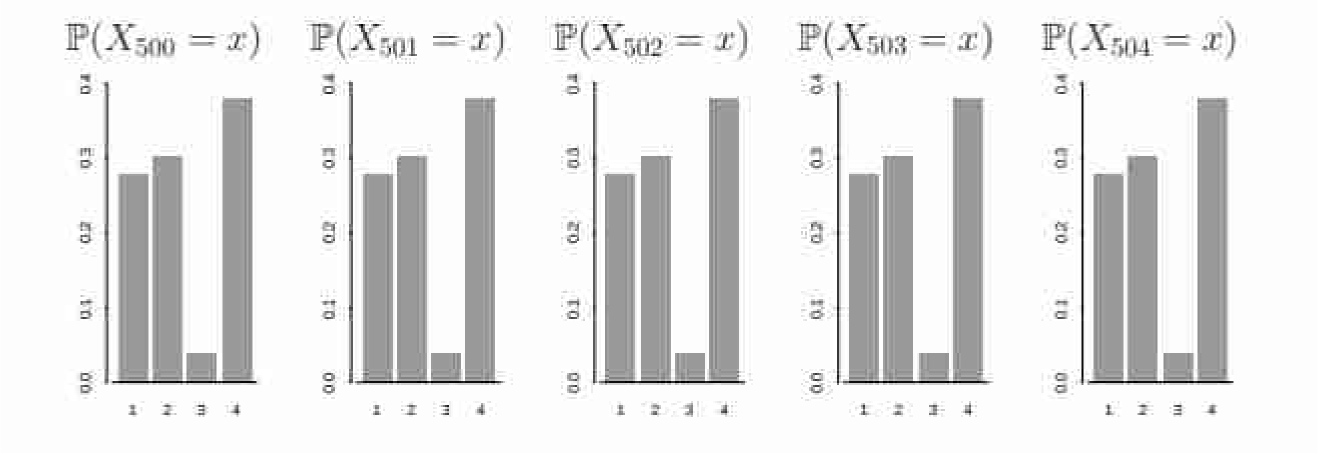
Suppose we start at time *t:=0* with X_0 ~ (¼, ¼, ¼, ¼) : so the chain is equally likely to start in any of the four states.

First Steps



Probability of being in state 1, 2, 3, or 4

Later Steps



We have reached equilibrium, the chain has forgotten about the initial Probability distribution of *(¼, ¼, ¼, ¼).*

Note: There are several other names for an equilibrium distribution. If π^T is an equilibrium distribution, it is also called:

- **invariant:** it doesn’t change π^T
- **stationary:** the chain ‘stops’ here

Calculating the Equilibrium Distribution

- For the example, we can explicitly calculate the equilibrium distribution by solving $\pi^T P = \pi^T$, under the restriction that:

1. The sum over all entries π_i in vector π^T is *1*
2. All π_i are larger or equal to *0*

- I will spare you the details, the equilibrium frequencies for our example are: *(0.28, 0.30, 0.04, 0.38)*

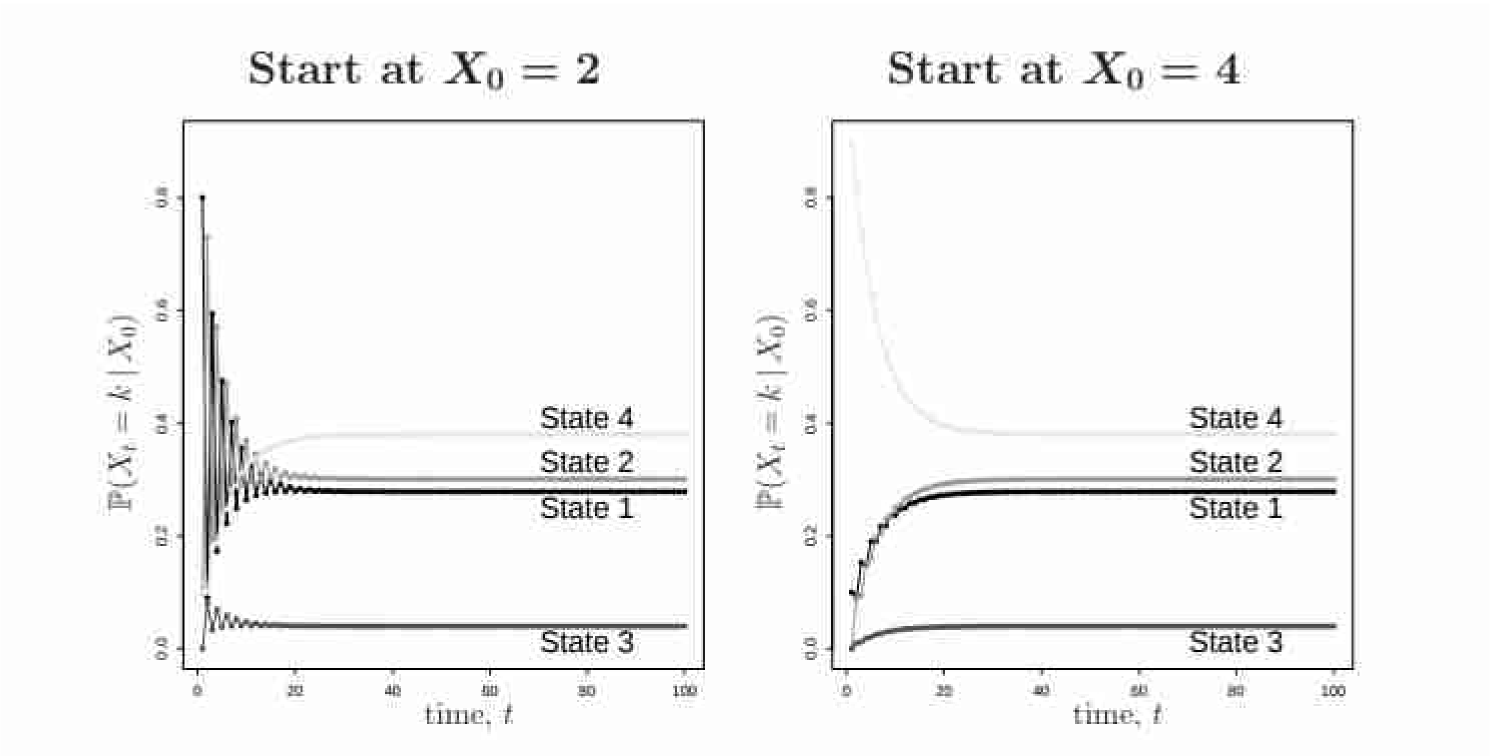
Convergence to Equilibrium

- What is happening here is that each row of the transition matrix Pt converges to the equilibrium distribution (0.28, 0.30, 0.04, 0.38) as t → ∞

$$P = \begin{pmatrix} 0.0 & 0.9 & 0.1 & 0.0 \\ 0.8 & 0.1 & 0.0 & 0.1 \\ 0.0 & 0.5 & 0.3 & 0.2 \\ 0.1 & 0.0 & 0.0 & 0.9 \end{pmatrix} \quad \Rightarrow \quad P^t \to \boxed{\begin{pmatrix} 0.28 & 0.30 & 0.04 & 0.38 \\ 0.28 & 0.30 & 0.04 & 0.38 \\ 0.28 & 0.30 & 0.04 & 0.38 \\ 0.28 & 0.30 & 0.04 & 0.38 \end{pmatrix}} \text{ as } t \to \infty.$$

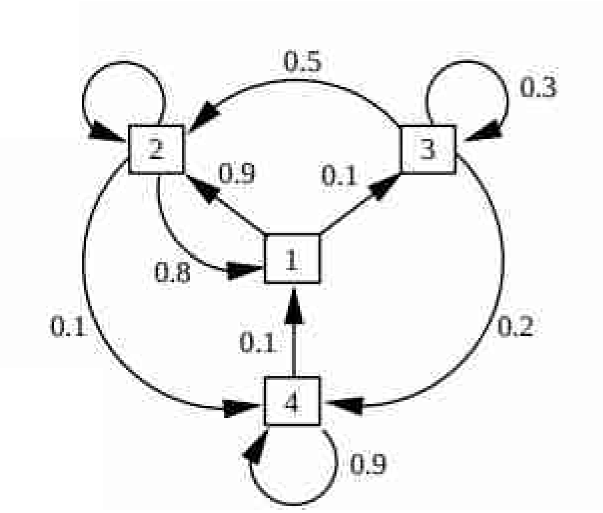
All rows become identical.

Impact of Starting Points

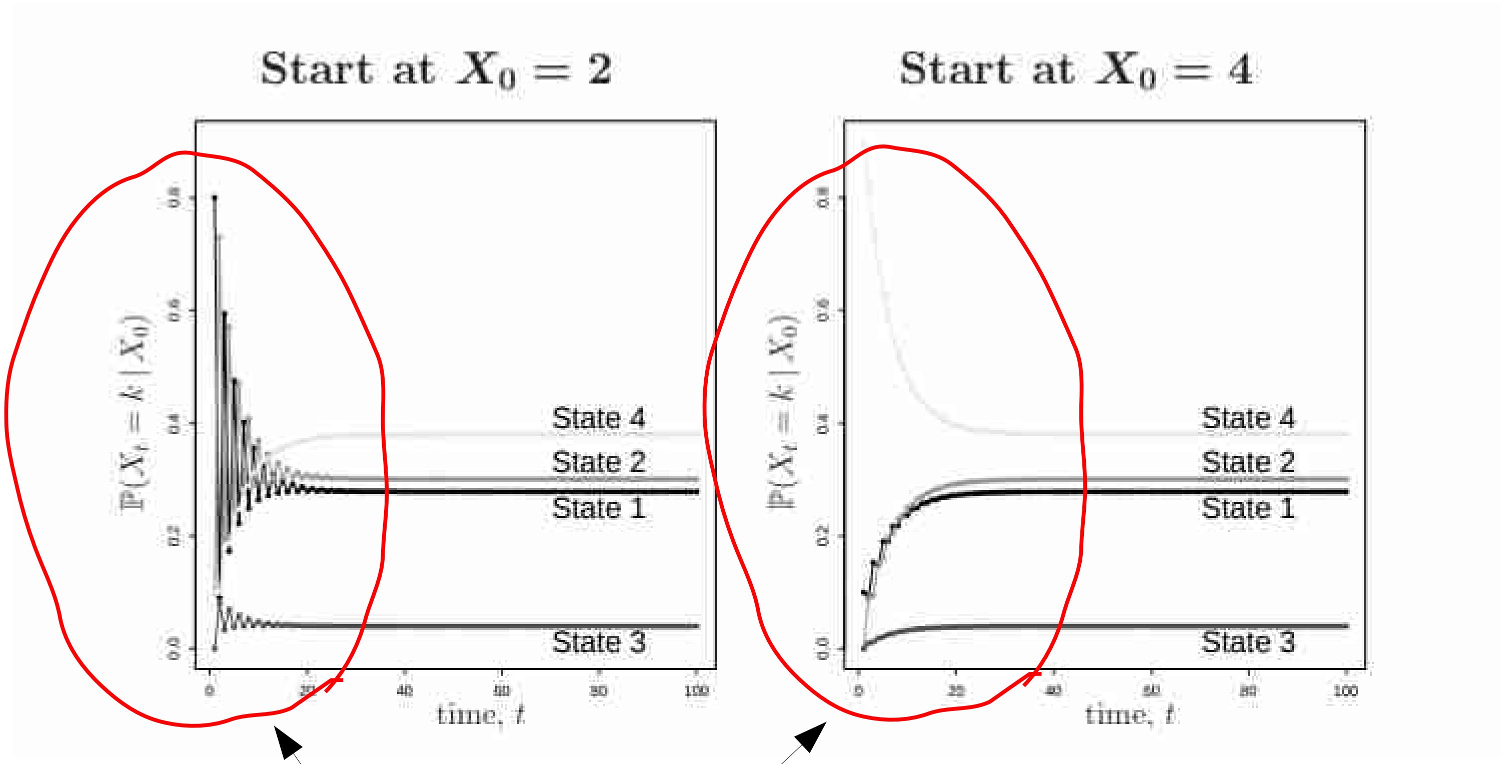


Start at $X_0 = 2$

Start at $X_0 = 4$

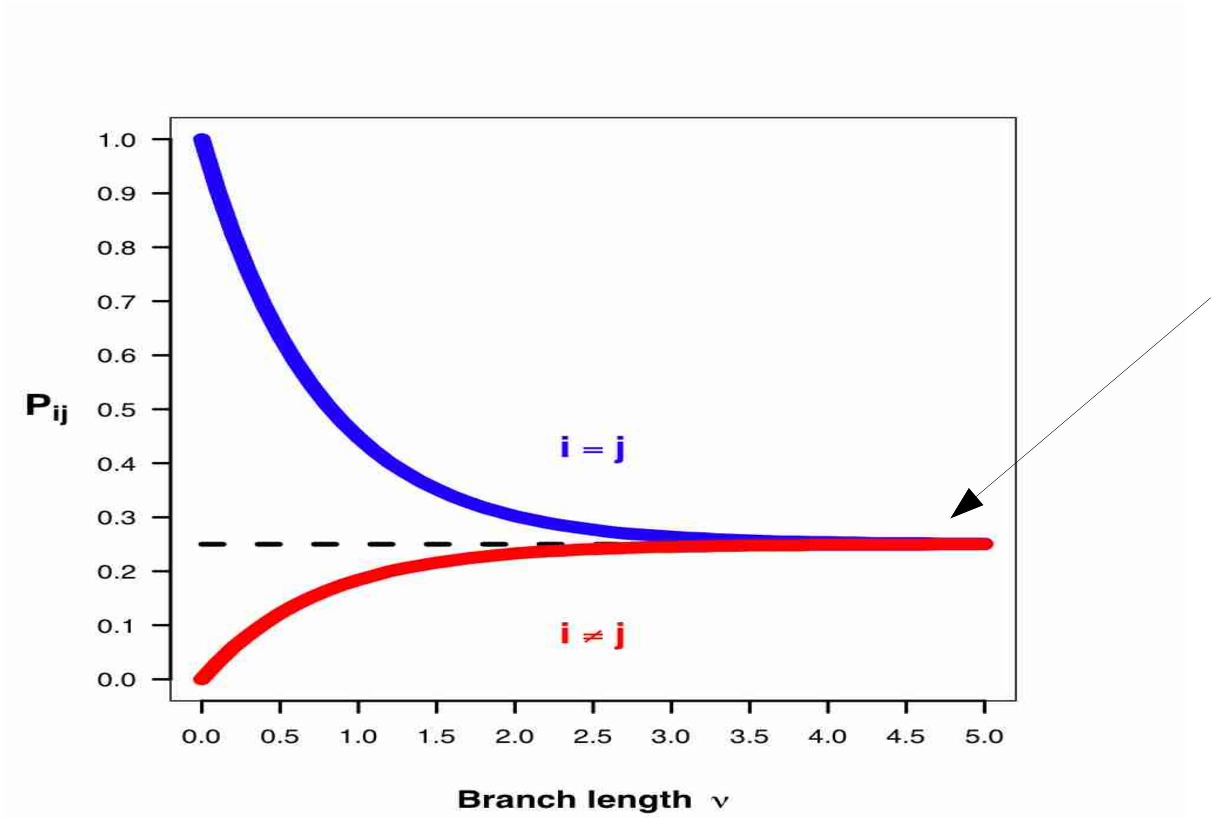


Impact of Starting Points



Initial behavior is different!

Continuous Time Models

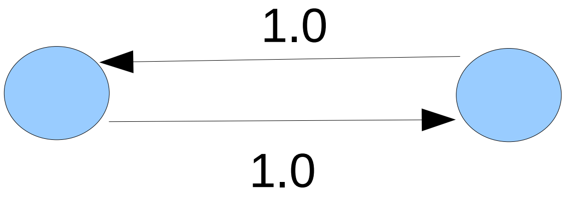


Convergence to stationary distribution of the *Jukes Cantor* Model: *(0.25,0.25,0.25, 0.25)*

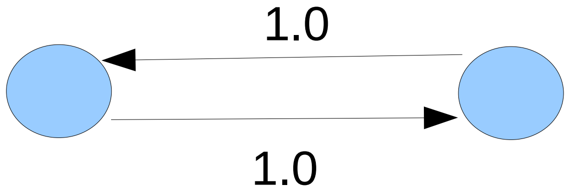
Time steps *t*

Probability of ending in state *j* when starting in state *i* over time (branch length) *v* where *i* = *j* for the blue curve and i ≠ j for the red one.

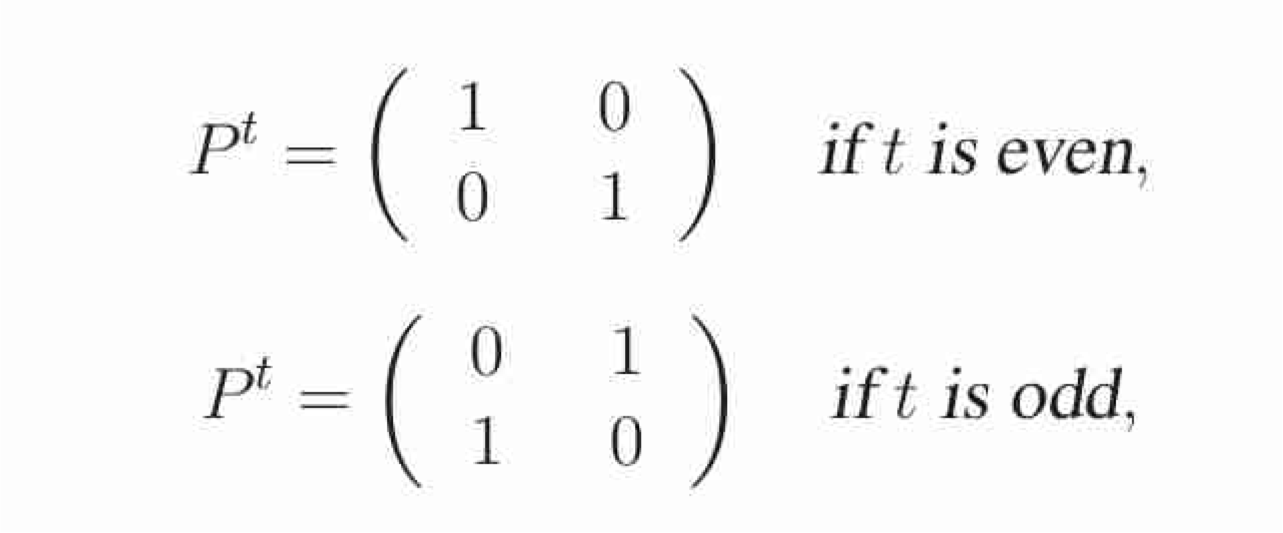
Is there always convergence to an equilibrium distribution?



Is there always convergence to an equilibrium distribution?



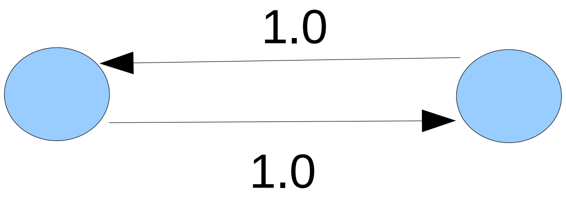
In this example, P^t never converges to a matrix with both rows identical as *t* becomes large. The chain never 'forgets' its starting conditions as $t \rightarrow \infty$.



$$P^t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \textit{if } t \textit{ is even},$$

$$P^t = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \textit{if } t \textit{ is odd},$$

Is there always convergence to an equilibrium distribution?



In this example, P^t never converges to a matrix with both rows identical as t becomes large. The chain never 'forgets' its starting conditions as $t \rightarrow \infty$.

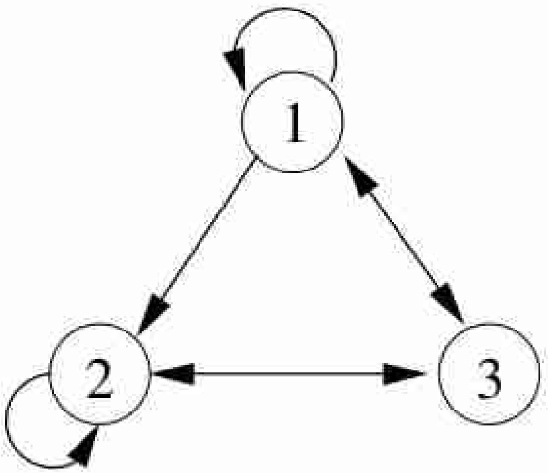
$$P^t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \textit{if } t \textit{ is even,}$$

$$P^t = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \textit{if } t \textit{ is odd,}$$

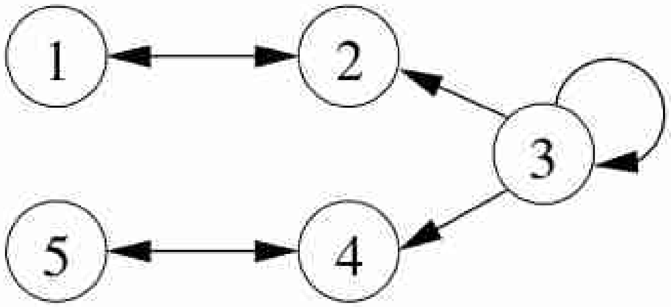
The chain does have an equilibrium distribution $\pi^T = (½, ½)$. However, the chain does not converge to this distribution as $t \rightarrow \infty$.

Convergence

- If a Markov chain is irreducible and aperiodic, and if an equilibrium distribution π^T exists, then the chain converges to this distribution as $t \to \infty$, regardless of the initial starting states.
- Remember: irreducible means that the state space is a single communicating class!



irreducible



non-irreducible

Periodicity

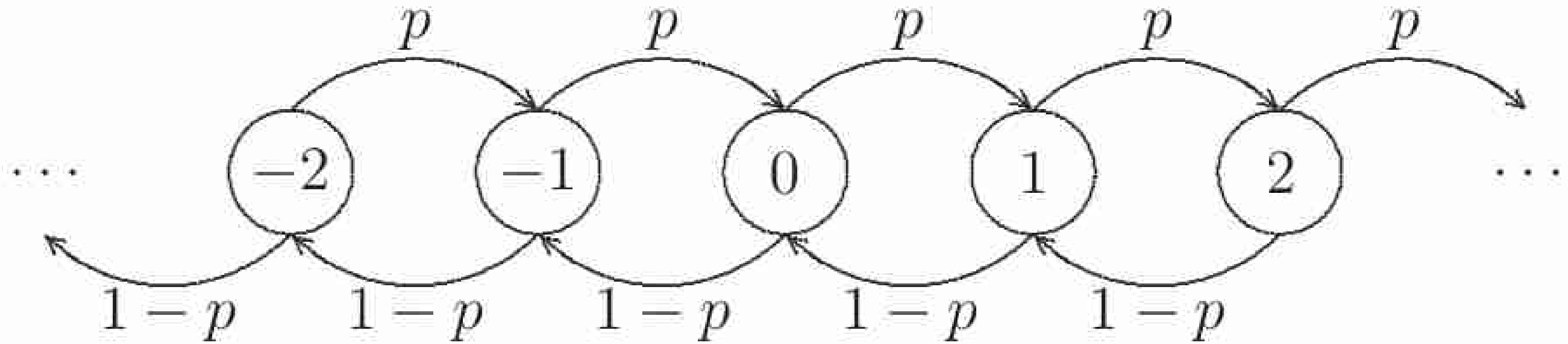
- In general, the chain can return from state *i* back to state *i* again in *t* steps if $(P^t)_{ii} > 0$. This leads to the following definition:
- **Definition:** The period $d(i)$ of a state *i* is

  $d(i) = gcd\{t : (P^t)_{ii} > 0\},$

  the greatest common divisor of the times at which return is possible.
- **Definition:** The state *i* is said to be periodic if $d(i) > 1$

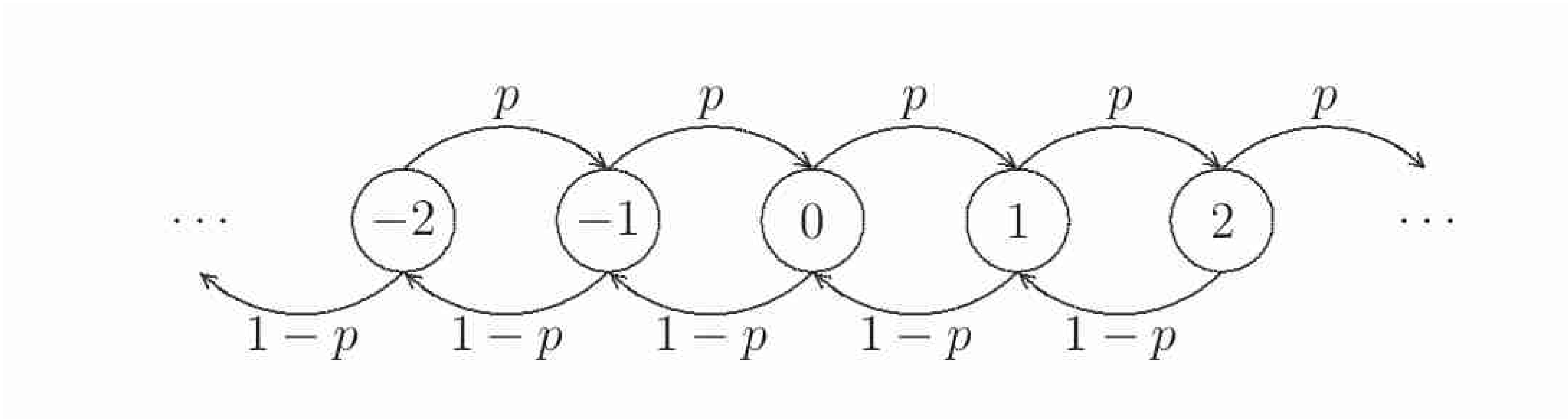
  For a periodic state *i*, $(P^t)_{ii} = 0$ if *t* is **not** a multiple of $d(i)$
- **Definition:** The state *i* is said to be aperiodic if $d(i) = 1$

Example



d(0) = ?

Example



d(0) = gcd{2, 4, 6, …} = 2

The chain is irreducible!

Result

- If a Markov chain is **irreducible** and has one **aperiodic** state, then all states are aperiodic.
- Theorem: Let $\{X_0, X_1, \ldots\}$ be an **irreducible** and **aperiodic** Markov chain with transition matrix P . Suppose that there exists an equilibrium distribution π^T . Then, from any starting state i, and for any end state j,

  $P(X_t = j \mid X_0 = i) \to \pi_j$ as $t \to \infty$.

  In particular,

  $(P^t)_{ij} \to \pi_j$ as $t \to \infty$, for all i and j,

  so P^t converges to a matrix with all rows identical and equal to π^T

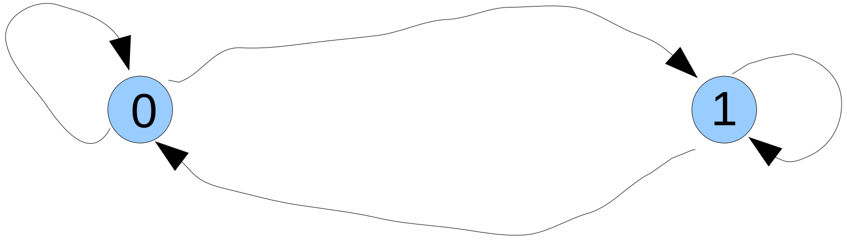
Why?

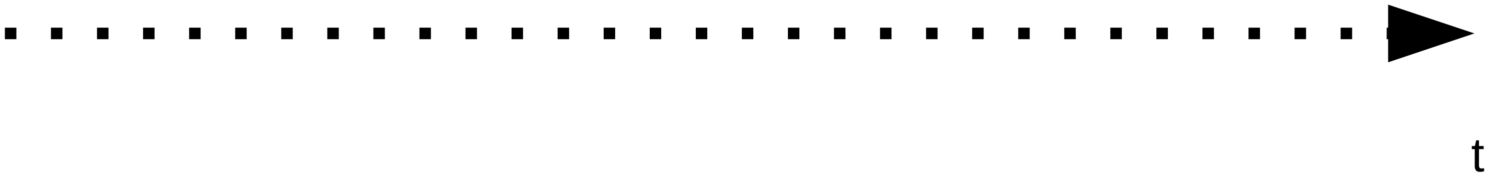
- The stationary distribution gives information about the stability of a random process.

Continuous Time Markov Chains (CTMC)

- Tranistions/switching between states at random times and not at clock ticks like in a CPU, for example!

→ no periodic oscillation, concept of waiting times!

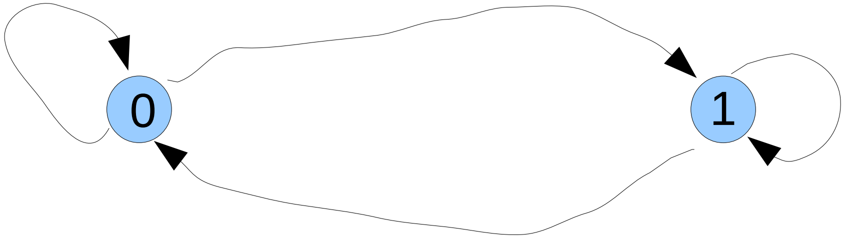


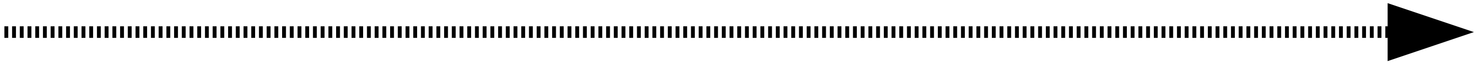


Continuous Time Markov Chains

- Tranistions/switching between states at random times and not at clock ticks like in a CPU, for example!

→ no periodic oscillation, concept of waiting times!

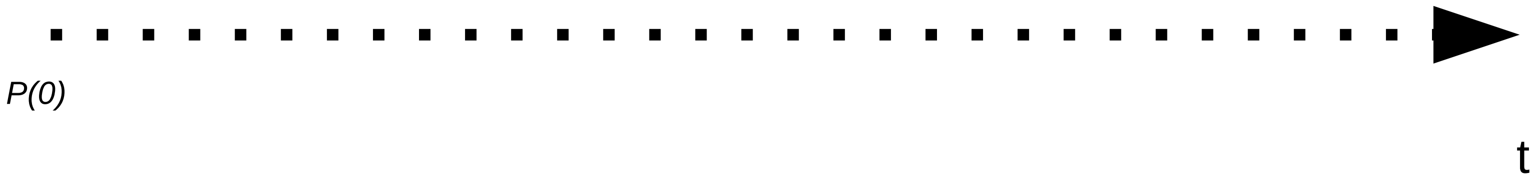




Understand what happens as we go toward *dt*

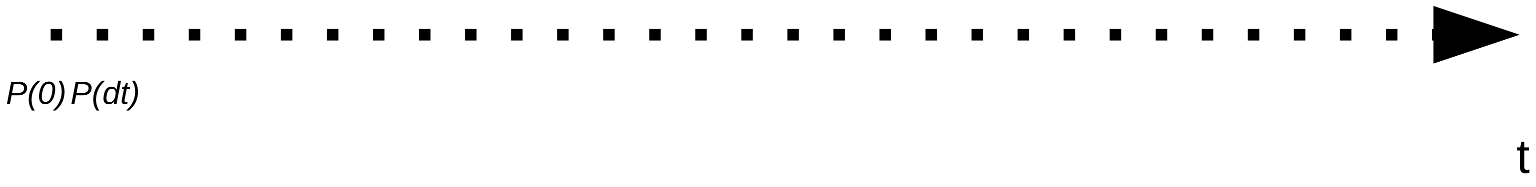
Use Calculus

- Now write the transition probability matrix *P* as a function of time *P(t)*



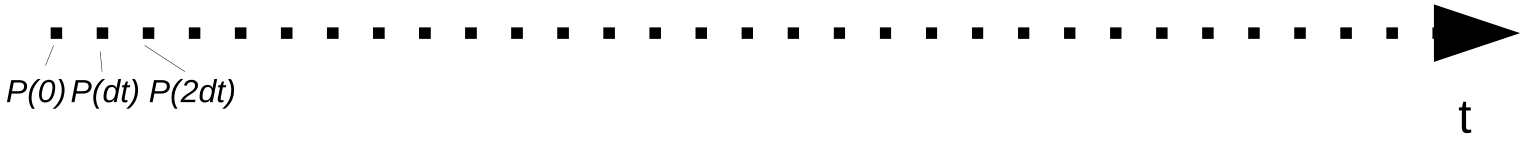
Use Calculus

- Now write the transition probability matrix *P* as a function of time *P(t)*



Use Calculus

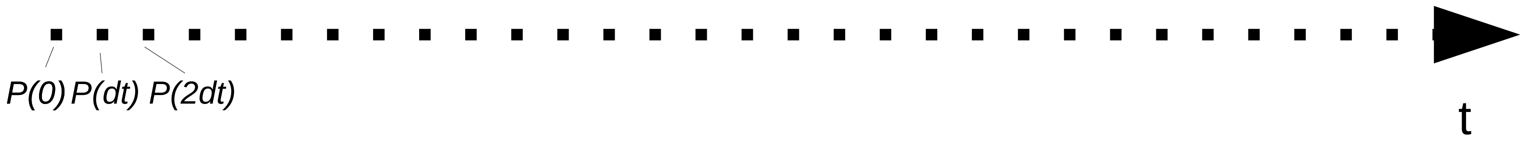
- Now write the transition probability matrix *P* as a function of time *P(t)*



P(t) is a function that returns a matrix! However, most standard maths on scalar functions can be applied.

Use Calculus

- Now write the transition probability matrix *P* as a function of time *P(t)*



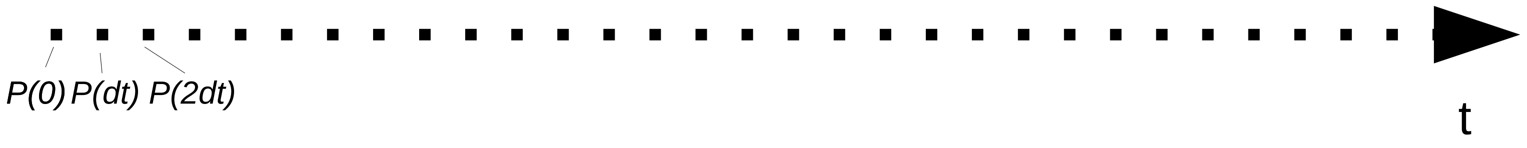
P(t) is a function that returns a matrix! However, most standard maths on scalar functions can be applied.

Derivative: $dP(t) / dt = lim_{\delta t \rightarrow 0} [P(t + \delta t) - P(t)] / \delta t$

Here only *dt* is a scalar value, everything else is a matrix!

Use Calculus

- Now write the transition probability matrix *P* as a function of time *P(t)*



P(t) is a function that returns a matrix! However, most standard maths on scalar functions can be applied.

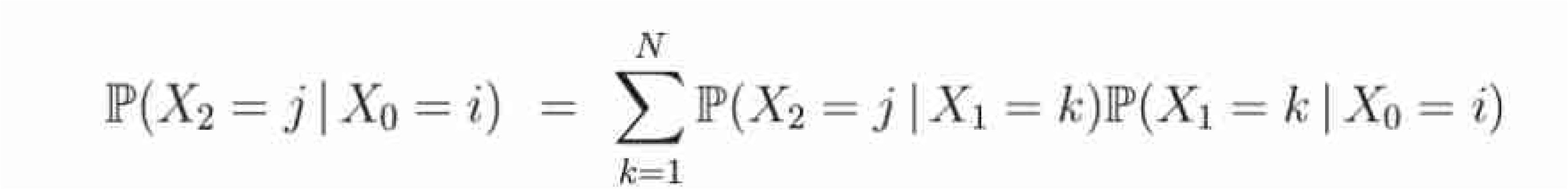
Derivative: $dP(t) / dt = \lim_{\delta t \rightarrow 0} [P(t + \delta t) - P(t)] / \delta t$

Here only *dt* is a scalar value, everything else is a matrix!

The derivative of a matrix is obtained by individually differentiating all of its entries, the same holds for the limit.

Calculating the limit

- Calculating $lim_{\delta t \rightarrow 0}$ *[P(t + δt) – P(t)] / δt r*equires solving a differential equation.
- If we can solve this, then we can calculate *P(t)*
- *Remember, for discrete chains:*



$$\mathbb{P}(X_2 = j \mid X_0 = i) \;=\; \sum_{k=1}^{N} \mathbb{P}(X_2 = j \mid X_1 = k)\mathbb{P}(X_1 = k \mid X_0 = i)$$

This is also known as the **Chapman-Kolmogorov relationship** and can be written differently as

$P^{n+m} = P^{n}P^{m}$

for any discrete number of steps *n* and *m.*

Calculating the limit

- *Calculating* $\lim_{\delta t \to 0} [P(t + \delta t) - P(t)] / \delta t$ requires solving a differential equation.
- If we can solve this, then we can calculate *P(t)*
- *Remember, for discrete chains:*

$$\mathbb{P}(X_2 = j \mid X_0 = i) \;=\; \sum_{k=1}^{N} \mathbb{P}(X_2 = j \mid X_1 = k)\mathbb{P}(X_1 = k \mid X_0 = i)$$

This is also known as the **Chapman-Kolmogorov relationship** and can be written differently as

$P^{n+m} = P^n P^m$

for any discrete number of steps *n* and *m.* Thus for continuous time we want: $P(t+h) = P(t)P(h)$

Calculating the limit

$$\lim_{\delta t \to 0} [P(t + \delta t) - P(t)] / \delta t$$

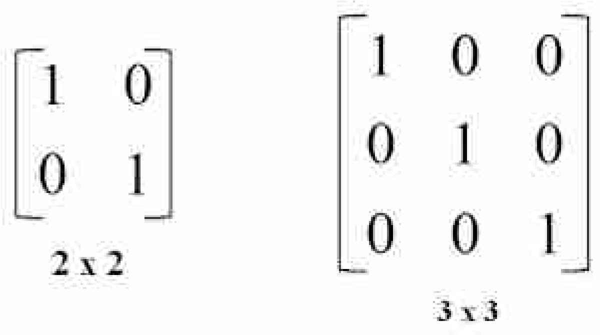
$$\downarrow$$

$$\lim_{\delta t \to 0} [P(t)P(\delta t) - P(t)] / \delta t$$

$$\downarrow$$

$$\lim_{\delta t \to 0} [P(t)(P(\delta t) - I)] / \delta t$$

Identity matrix, analogous to *1* in the scalar case



$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

2 x 2

3 x 3

Calculating the limit

$$\begin{tikzcd}
\lim_{\delta t \to 0} [P(t + \delta t) - P(t)] / \delta t \arrow[d] \\
\lim_{\delta t \to 0} [P(t)P(\delta t) - P(t)] / \delta t \arrow[d] \\
\lim_{\delta t \to 0} [P(t)(P(\delta t) - I)] / \delta t \arrow[d, "\text{The limit doesn't depend on } P(t)\text{!}"] \\
P(t) \lim_{\delta t \to 0} (P(\delta t) - I) / \delta t
\end{tikzcd}$$

Calculating the limit

$$\begin{tikzcd}
\lim_{\delta t \to 0} [P(t + \delta t) - P(t)] / \delta t \arrow[d] \\
\lim_{\delta t \to 0} [P(t)P(\delta t) - P(t)] / \delta t \arrow[d] \\
\lim_{\delta t \to 0} [P(t)(P(\delta t) - I)] / \delta t \arrow[d] \\
P(t) \lim_{\delta t \to 0} \boxed{(P(\delta t) - I) / \delta t}
\end{tikzcd}$$

The limit doesn't depend on *P(t)*!

This is the famous *Q* matrix

Calculating the limit

$$\lim_{\delta t \to 0} [P(t + \delta t) - P(t)] / \delta t$$

$$\Big\downarrow$$

$$\lim_{\delta t \to 0} [P(t)P(\delta t) - P(t)] / \delta t$$

$$\Big\downarrow$$

$$\lim_{\delta t \to 0} [P(t)(P(\delta t) - I)] / \delta t$$

$$\Big\downarrow$$

The limit doesn't depend on *P(t)*!

$$P(t) \boxed{\lim_{\delta t \to 0} (P(\delta t) - I) / \delta t}$$

This is the famous *Q* matrix

The values of *Q* can be anything, but rows must sum to *0*. Remember that rows of *P* must sum to *1*.

What we have so far

dP(t)/dt = P(t)Q

Q is also called the **jump rate matrix**, or **instantaneous transition matrix**

Now, imagine that *P(t)* is a scalar function and *Q* just some scalar constant:

P(t) = exp(Qt)

the same holds for matrices.

What we have so far

$dP(t)/dt = P(t)Q$

Q is also called the **jump rate matrix**, or **instantaneous transition matrix**

Now, imagine that *P(t)* is a scalar function and *Q* just some scalar constant:

$P(t) = exp(Qt)$

the same holds for matrices.

However calculating a matrix exponential is not trivial, it's not just taking the exponential of each of its elements!

$exp(Qt) = I + Qt + 1/2!\ Q^2t^2 + 1/3!\ Q^3t^3 + \dots$

$$P(t)=e^{Qt}$$

- There is no general solution to analytically calculate this matrix exponential, it depends on *Q*.
- In some cases we can come up with an analytical equation, like for the aforementioned *Jukes Cantor* model
- For the GTR model we already need to use creepy numerical methods (Eigenvector/Eigenvalue) decomposition, we might see that later
- For non-reversible models it gets even more nasty

Equilibrium Distribution

- Assume there exists a row vector π^T such that $\pi^T Q = 0$

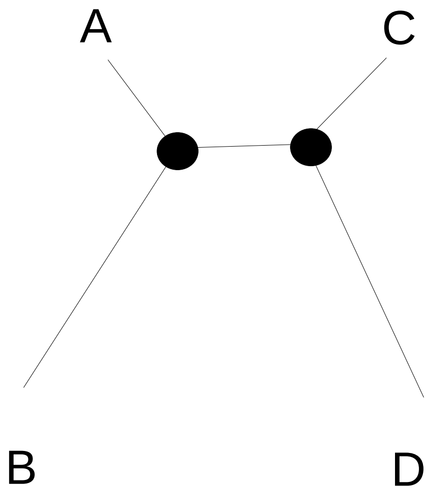
→ π^T *is the equilibrium distribution*

Today

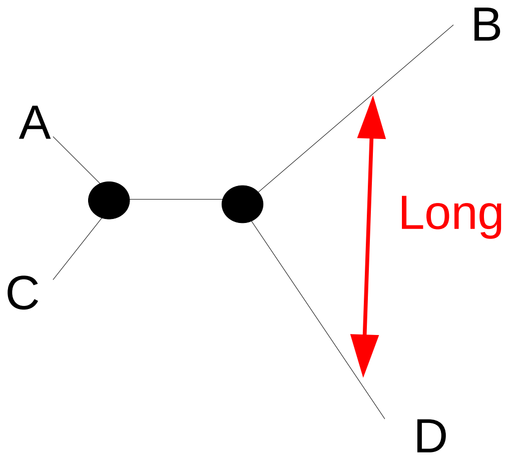
- Data structures for unrooted trees
- Why do we need statistical/probabilistic models of evolution?
- **Introduction to the phylogenetic likelihood function**
  - A detour on Markov Chains
  - **The phylogenetic likelihood function**

Parsimony & Long Branch Attraction

- Because parsimony tries to minimize the number of mutations it faces some problems on trees with long branches



Correct tree



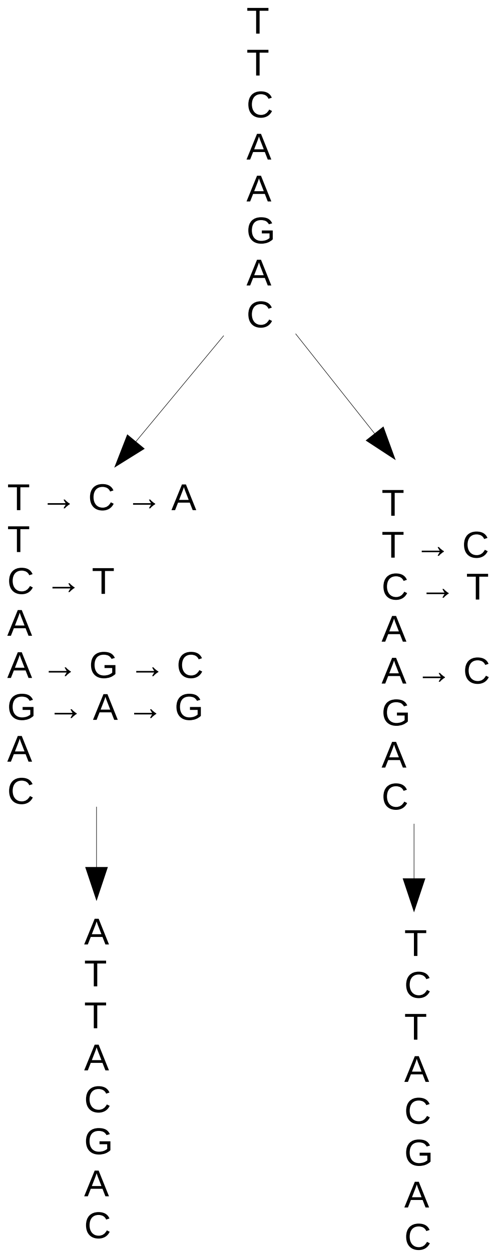
Long branch attraction

Wrong tree inferred by parsimony

Parsimony & Long Branch Attraction

- Settings under which parsimony recovers the wrong tree are also called **“the Felsenstein Zone”** after *Joe Felsenstein* who has made numerous very important contributions to the field, e.g.
  - The Maximum Likelihood model
  - The Bootstrapping procedure
- If you are interested in statistics, there are some on-line courses by Joe at http://evolution.gs.washington.edu/courses.html

Likelihood tries to fix this Problem



T
T
C
A
A
G
A
C

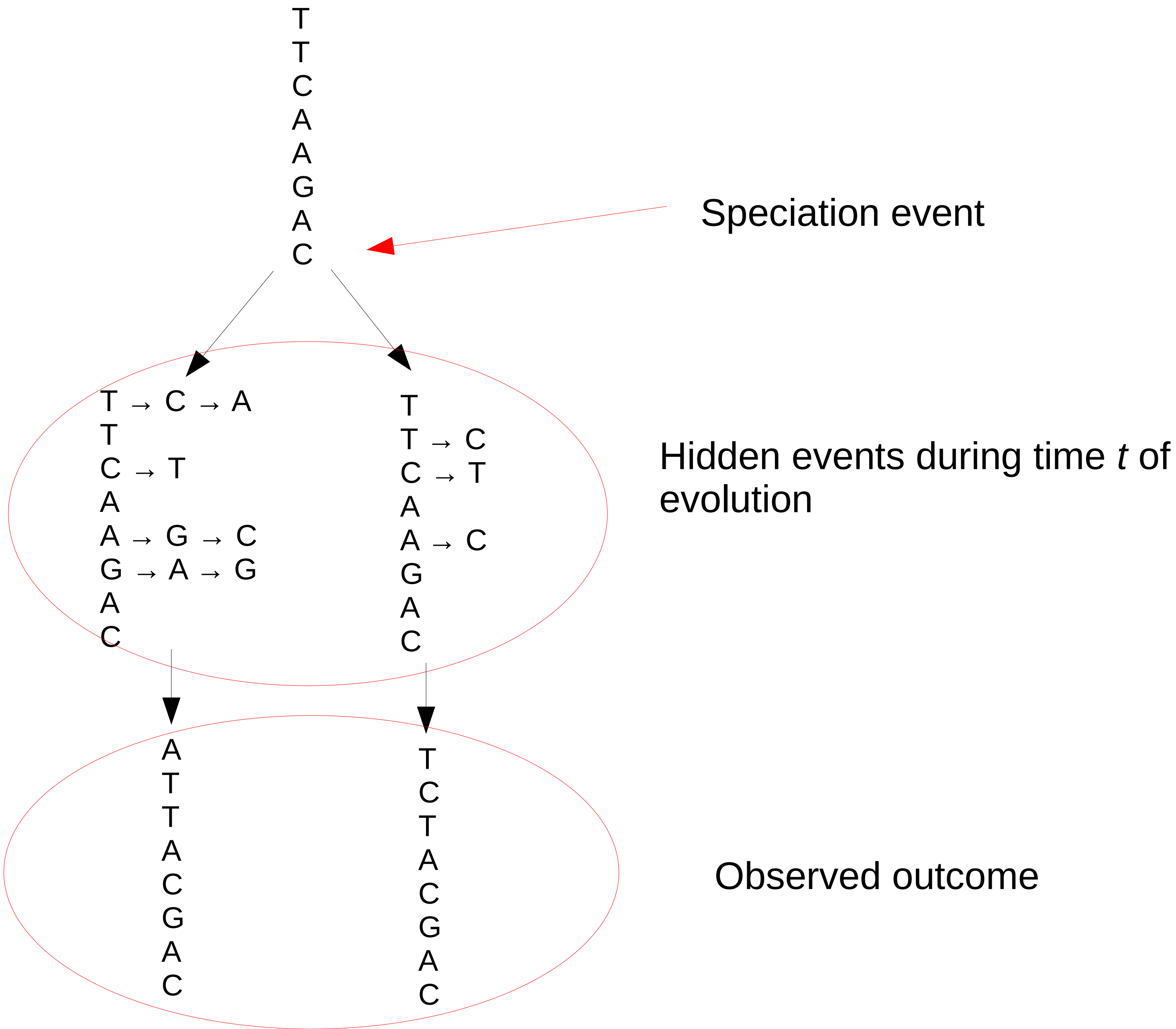
T → C → A
T
C → T
A
A → G → C
G → A → G
A
C

T
T → C
C → T
A
A → C
G
A
C

A
T
T
A
C
G
A
C

T
C
T
A
C
G
A
C

Likelihood tries to fix this Problem



T
T
C
A
A
G
A
C

Speciation event

T → C → A
T
C → T
A
A → G → C
G → A → G
A
C

T
T → C
C → T
A
A → C
G
A
C

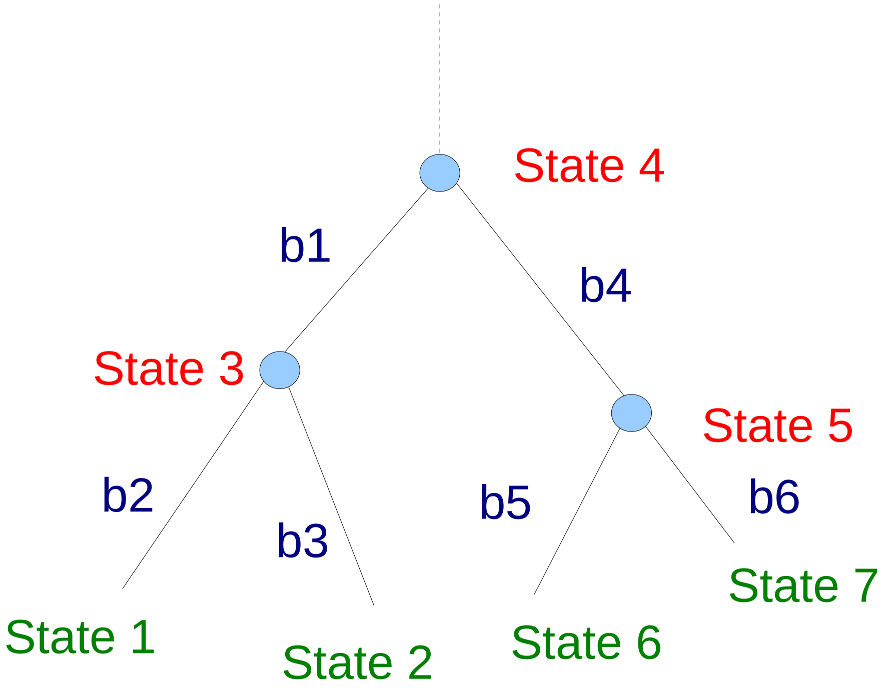
Hidden events during time *t* of evolution

A
T
T
A
C
G
A
C

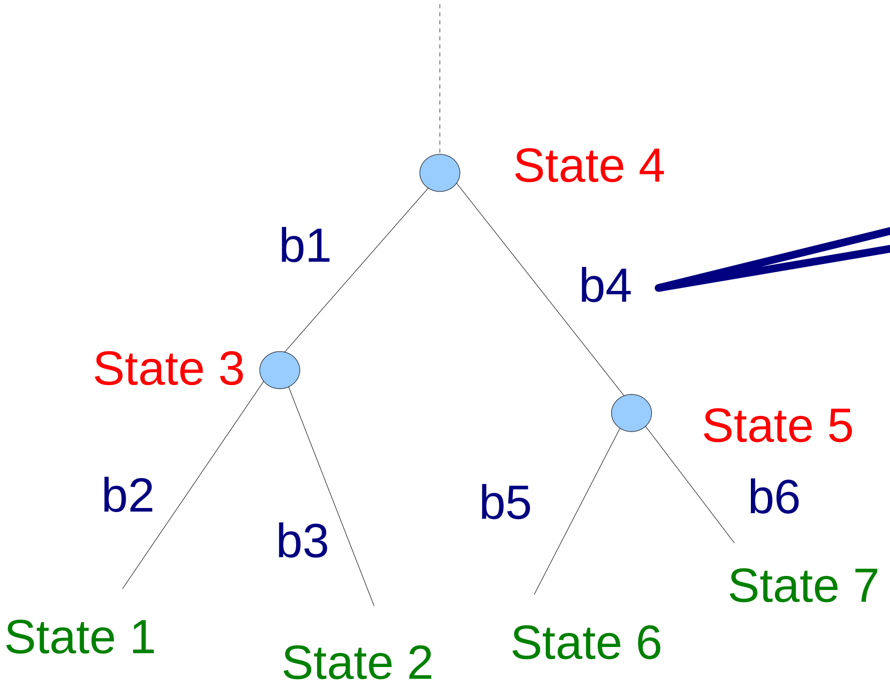
T
C
T
A
C
G
A
C

Observed outcome

Today and next time → How do we compute the likelihood on a tree?



Next Time → How do we maximize the likelihood on a tree?



We need, for instance, to optimize/estimate the branch lengths

Likelihood

- Given:
  - MSA
  - Tree topology with branch lengths
  - Model
  - We can calculate $P_{x\text{->}z}(b)$ for a branch length (or time) *b*
  - $P_{x\text{->}z}(b)$ is our continuous time Markov Model of sequence evolution!
  - We obtain $P_{x\text{->}z}(b)$ by exponentiating the instantaneous rate matrix Q

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

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

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 $\boldsymbol{\Theta}$: $L(\boldsymbol{\Theta}|D)$

Likelihood

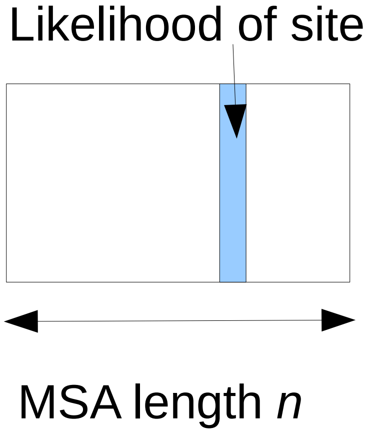
- $L(T|D) = P(D|T)$
- $L(T|D) = \Pi\ P(s_i|T)$
- $\log(L(T|D)) = \Sigma \log(P(s_i|T))$

How do we compute this?

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

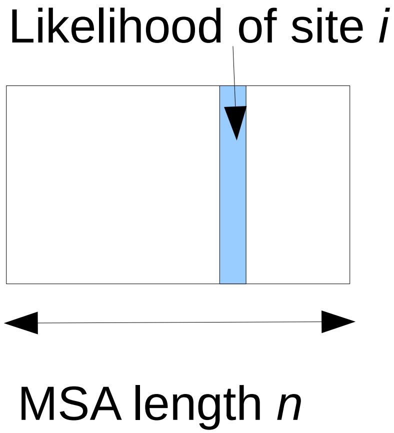
Likelihood of a Tree

- We assume that sites evolve independently



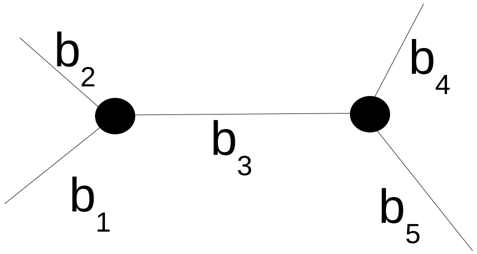
Likelihood of a Tree

- We assume that sites evolve independently



Likelihood of site *i*

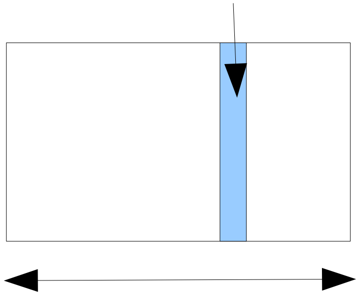
MSA length *n*



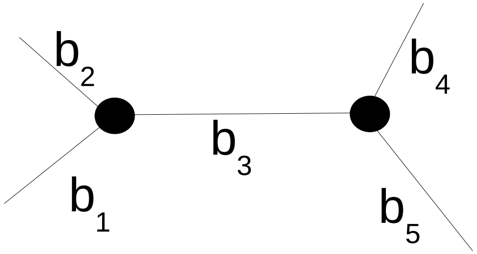
Likelihood of a Tree

- We assume that sites evolve independently

Likelihood of site *i*



MSA length *n*



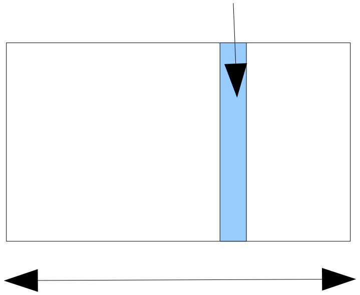
Model ***M***

$P_{ij}(t)$

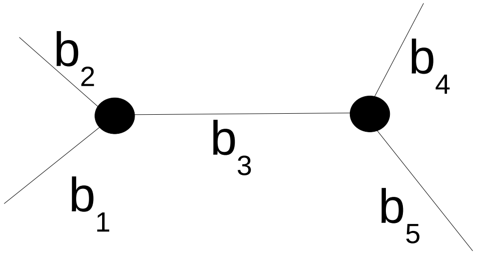
Likelihood of a Tree

- We assume that sites evolve independently

Likelihood of site *i*



MSA length *n*



Model ***M***

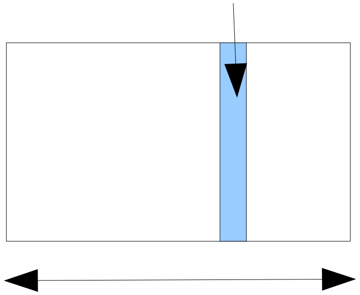
$P_{ij}(t)$

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

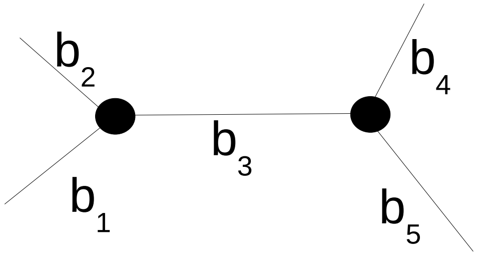
Likelihood of a Tree

- We assume that sites evolve independently

Likelihood of site *i*



MSA length *n*



Model ***M***

$P_{ij}(t)$

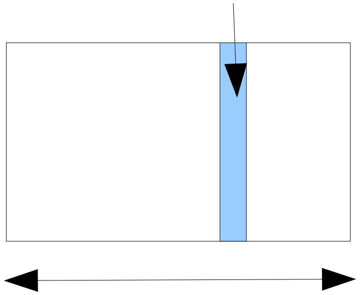
- Overall likelihood: $L := \Pi L_i$
- $P_{ij}(t)$ *i,j in {A, C, G, T}*

Branch length/relative time

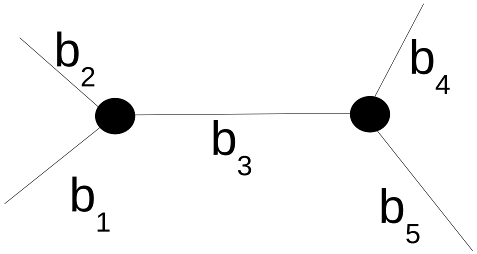
Likelihood of a Tree

- We assume that sites evolve independently

Likelihood of site *i*



MSA length *n*



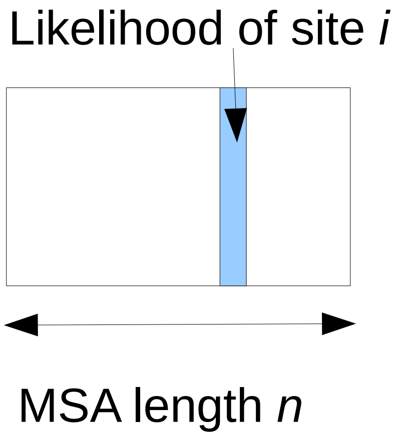
Model ***M***

$P_{ij}(t)$

- 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 continuous time Markov Process

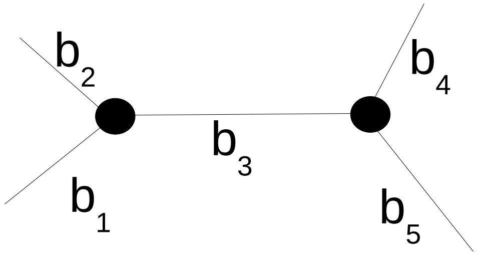
Likelihood of a Tree

- 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}*

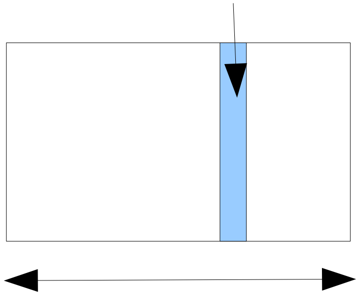
  → Probability of being in state *j* after time *t*

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

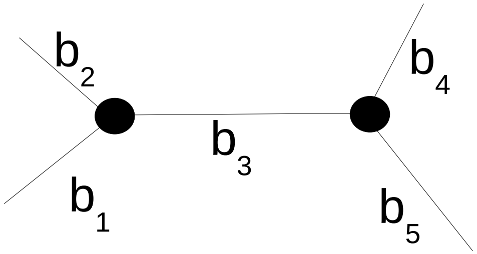
Likelihood of a Tree

- We assume that sites evolve independently

Likelihood of site *i*



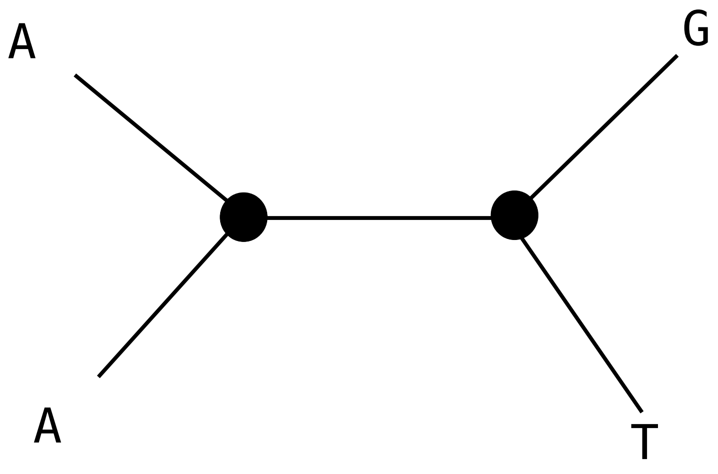
MSA length *n*



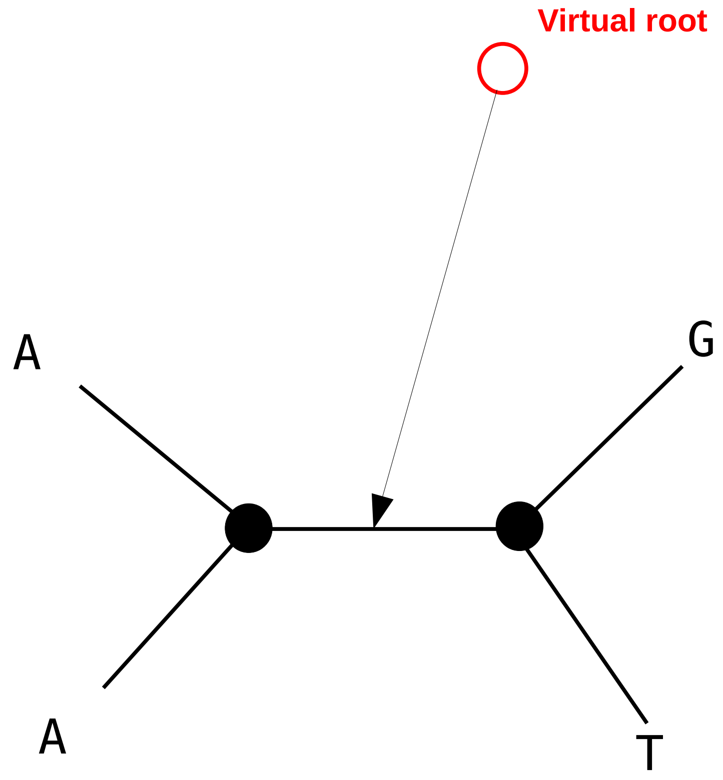
Model ***M***
$P_{ij}(t)$

- 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
- Equilibrium frequency vector $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$
- **Time reversibility:** $\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$

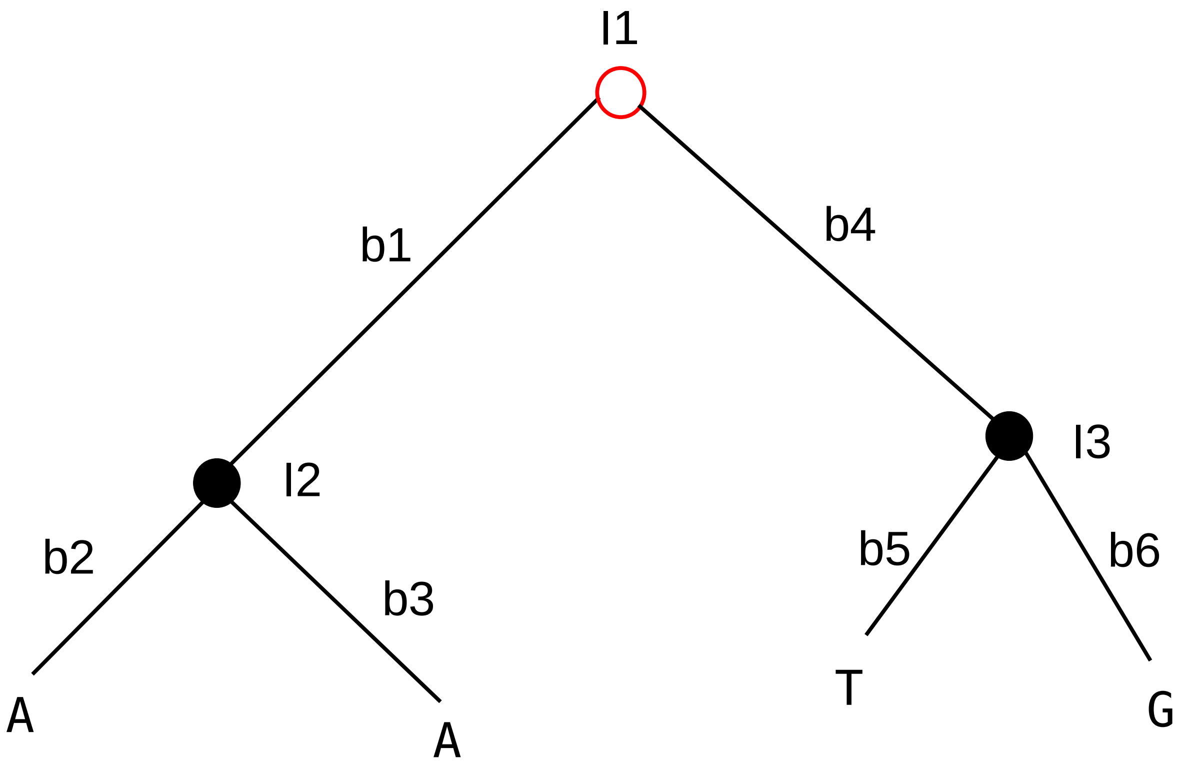
What's the likelihood of this tree?



What's the likelihood of this tree?

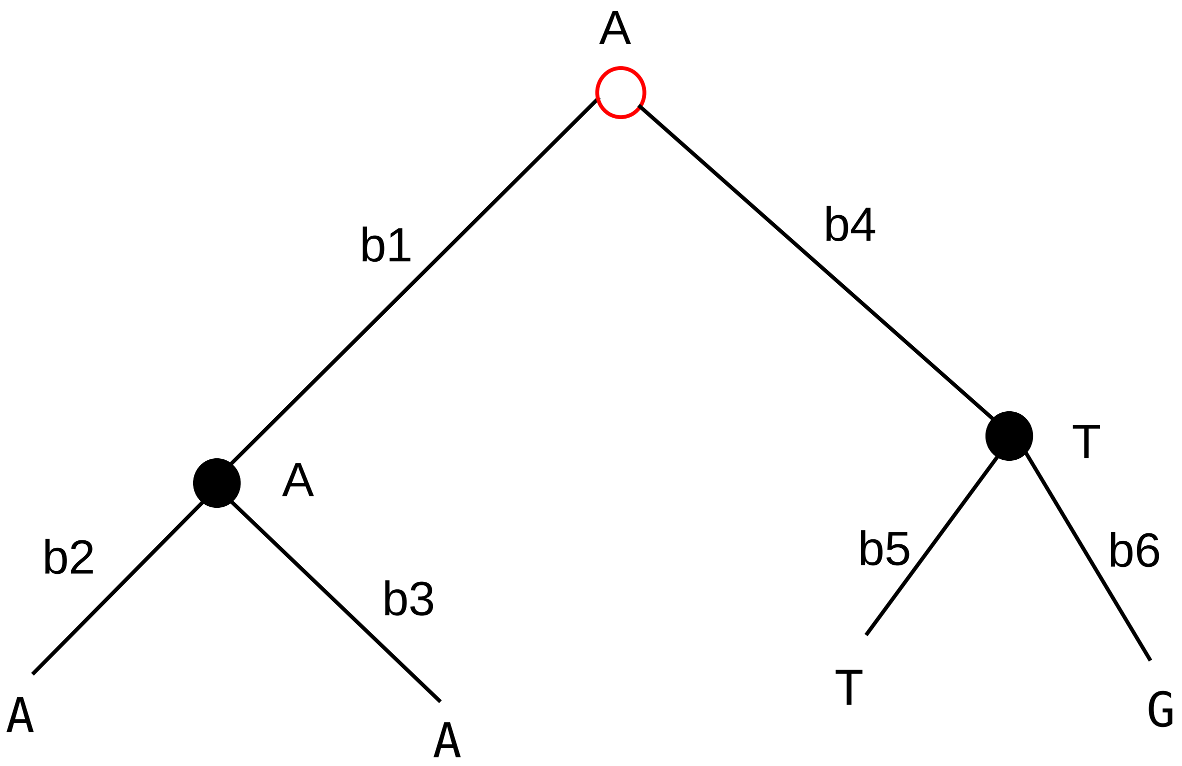


What's the likelihood of this tree?



What's the likelihood of this tree?

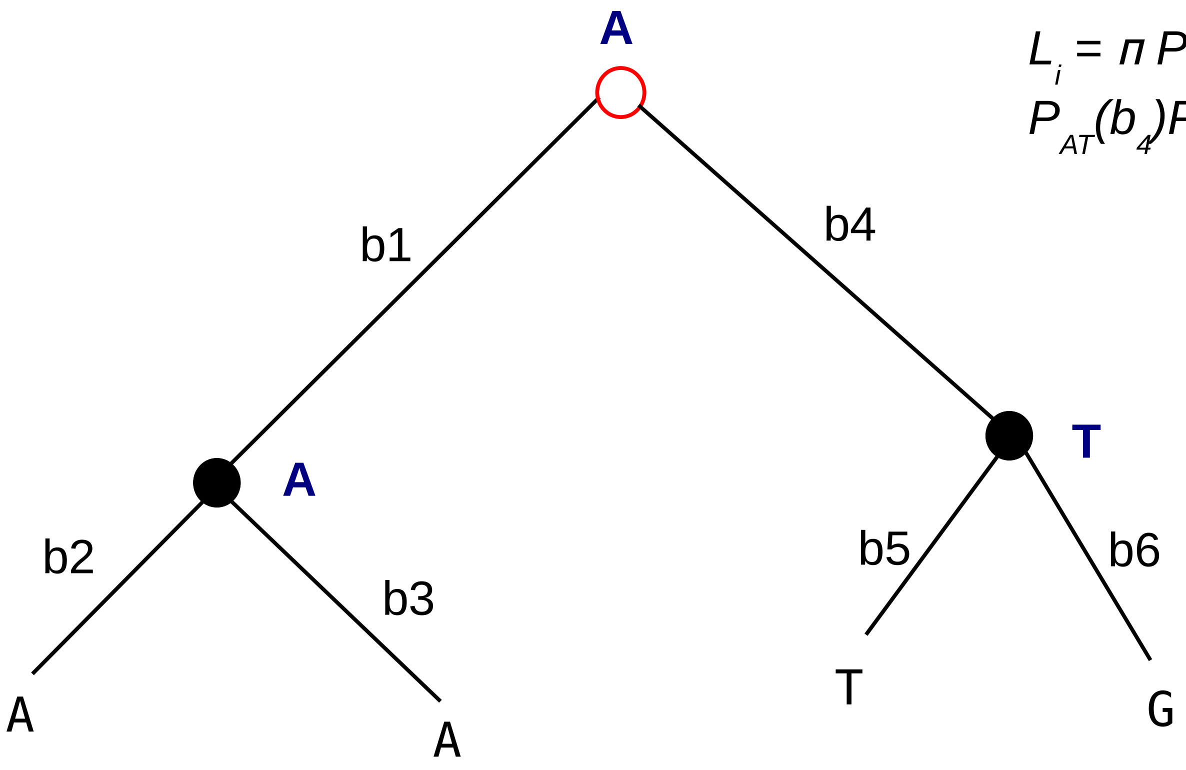
Assume the inner states are given!
What is the likelihood of the tree if we interpret it as **Markov** diagram?



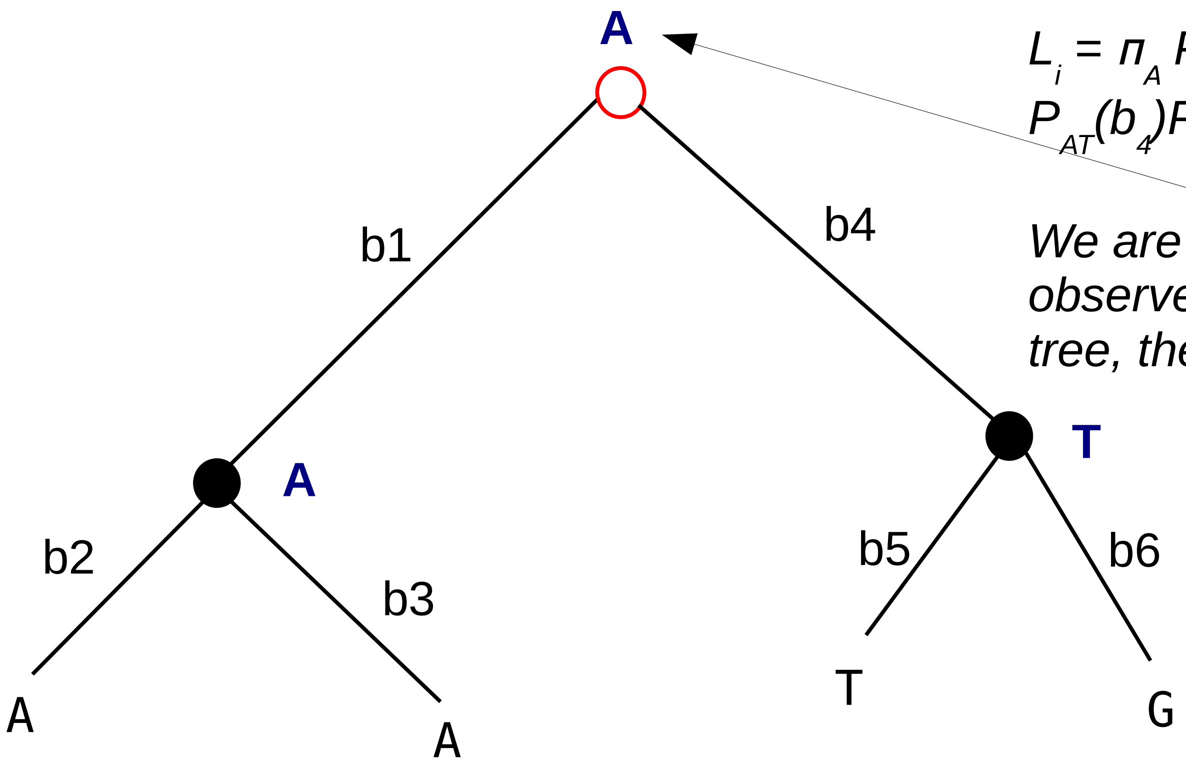
What's the likelihood of this tree?

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?

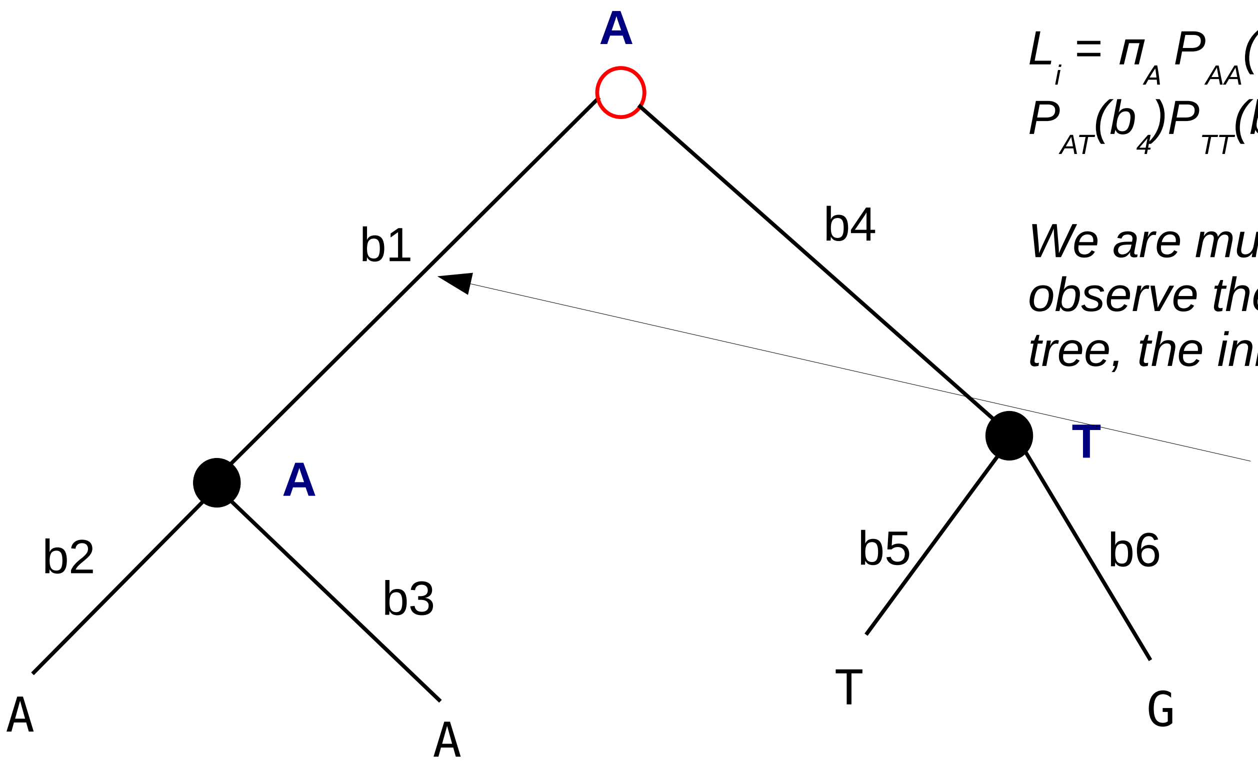


Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

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

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

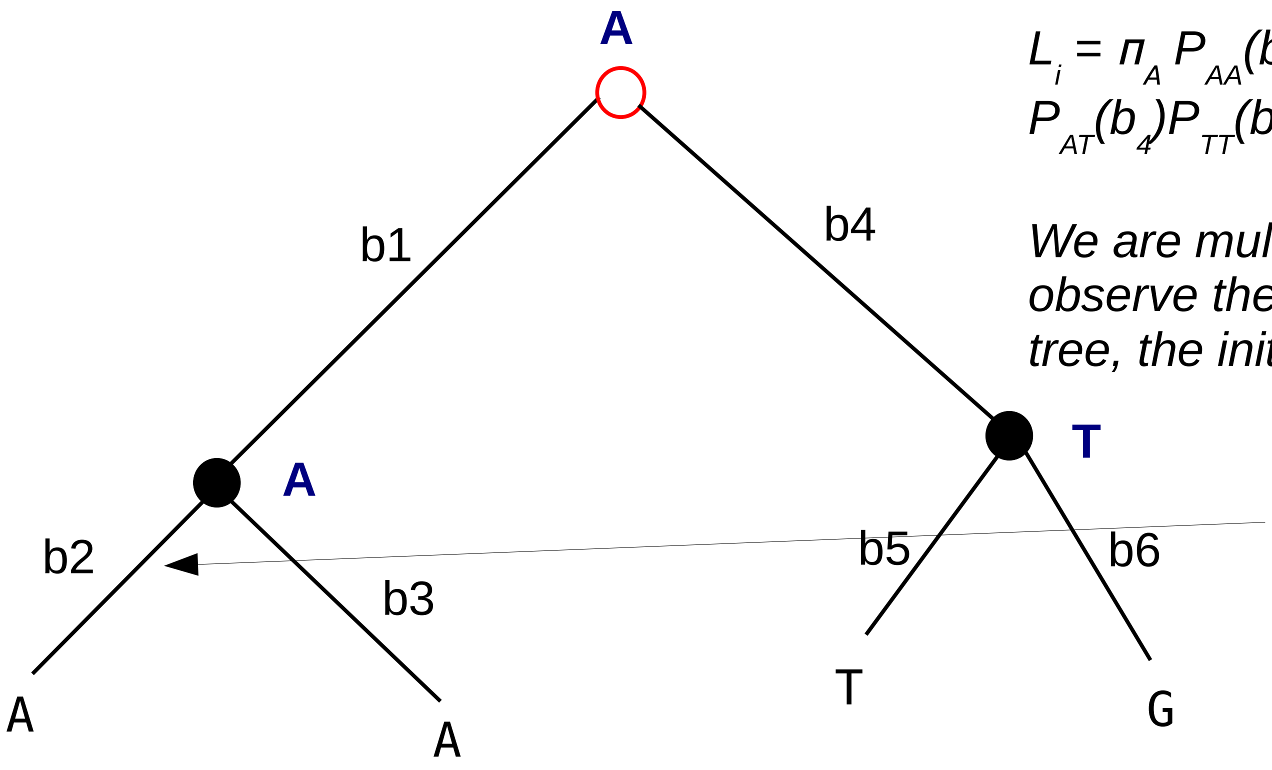
AND then this happened

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?

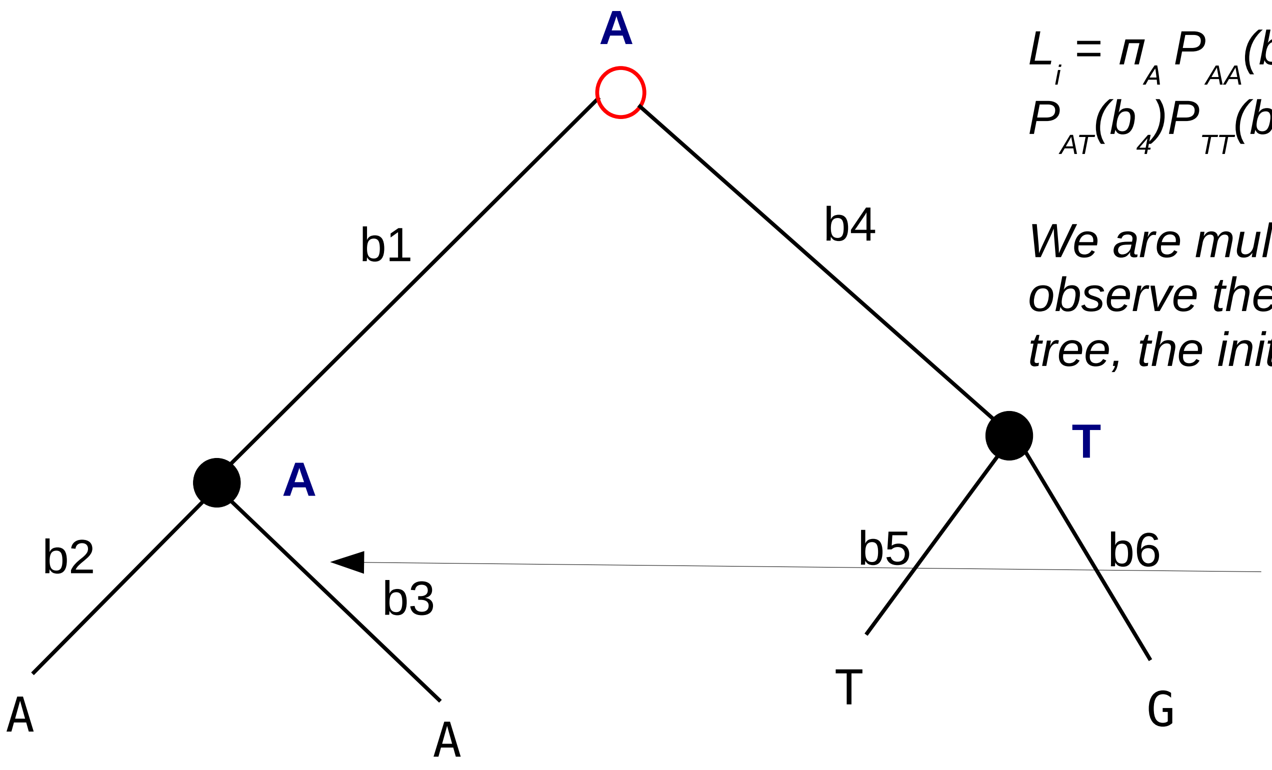
$L_i = \pi_A\, P_{AA}(b_1)\, P_{AA}(b_2)\, P_{AA}(b_3)$
$P_{AT}(b_4) P_{TT}(b_5)\, P_{TG}(b_6)$

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?



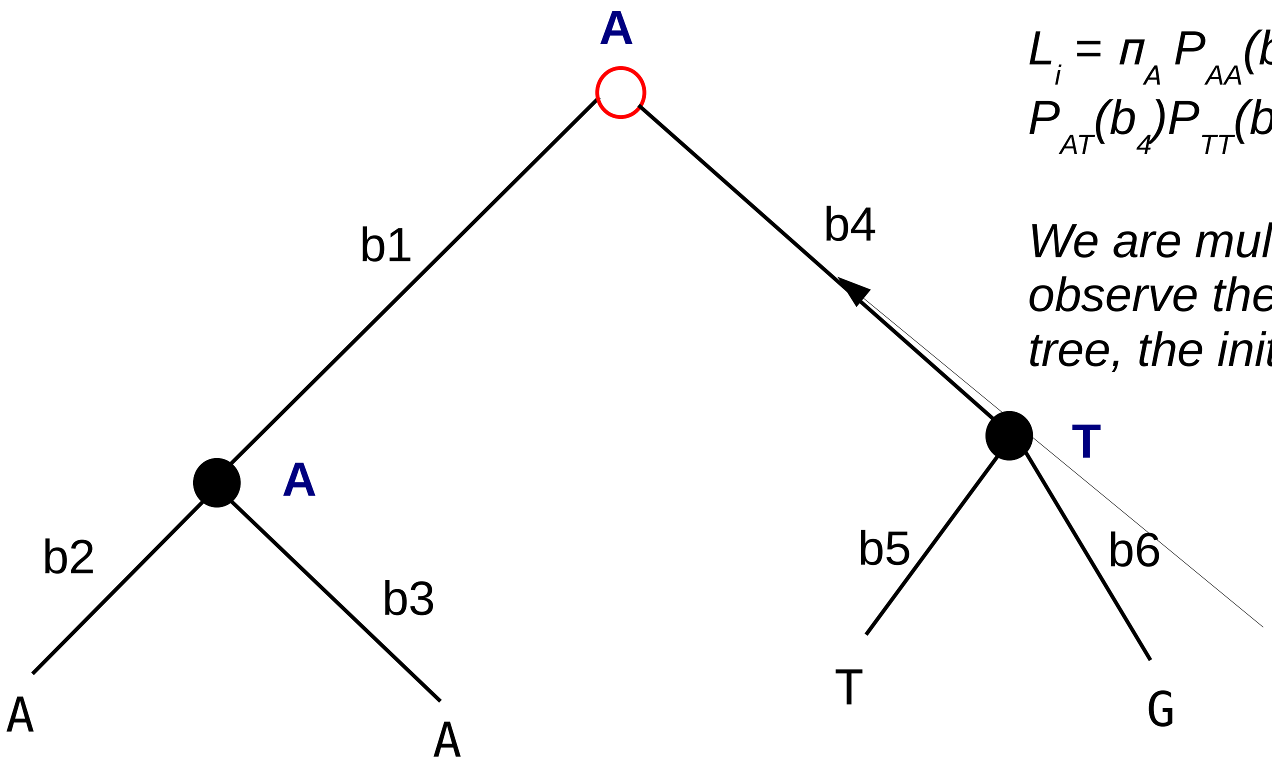
Assume the inner states are given!
What is the likelihood of the tree if we
Interpret it as **Markov** diagram?

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

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

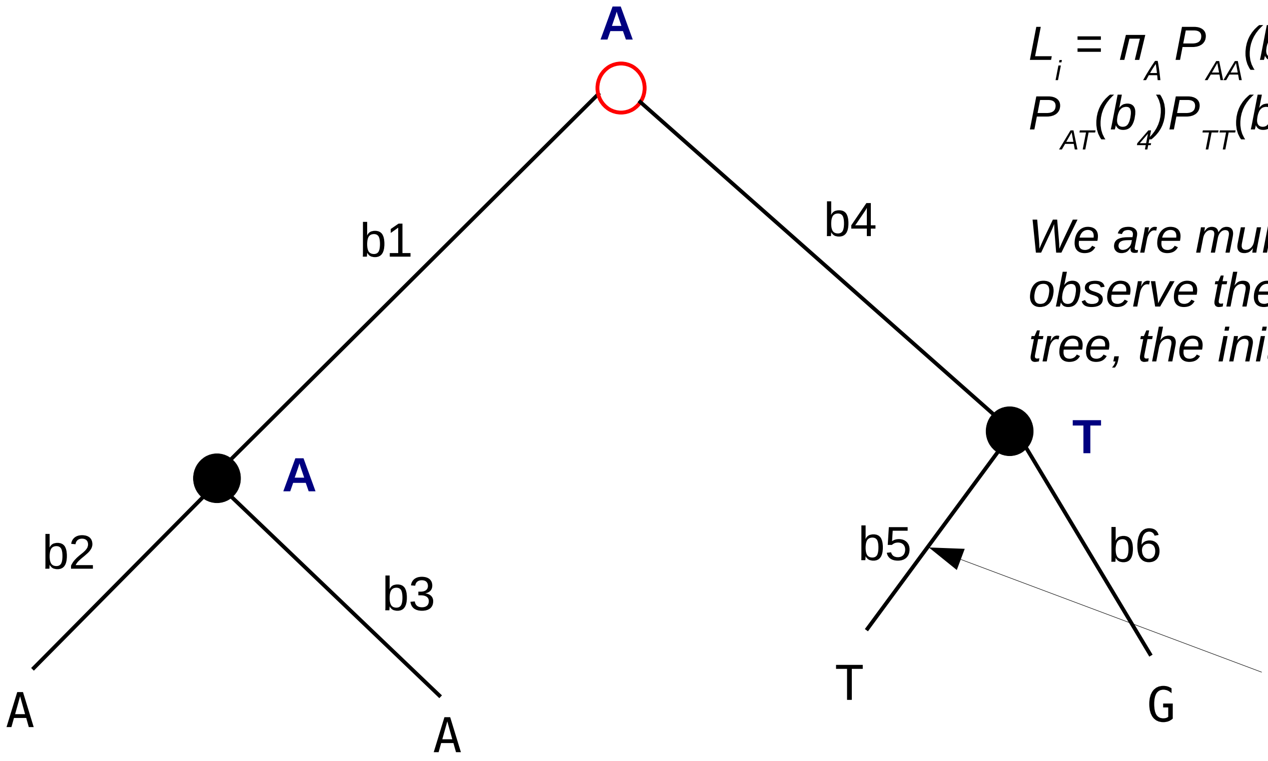
AND then this happened
AND this
AND this
AND this

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?

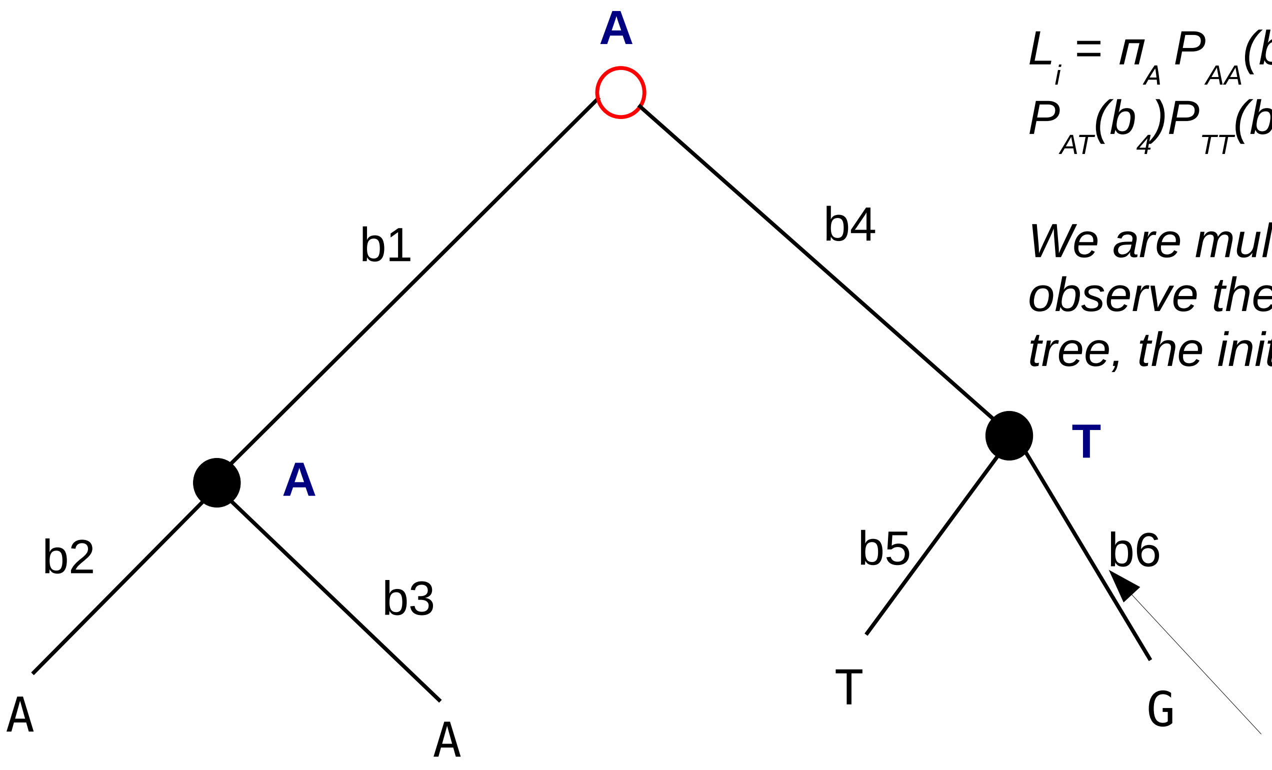
$$L_i = \pi_A P_{AA}(b_1) P_{AA}(b_2) P_{AA}(b_3) P_{AT}(b_4) P_{TT}(b_5) P_{TG}(b_6)$$

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?

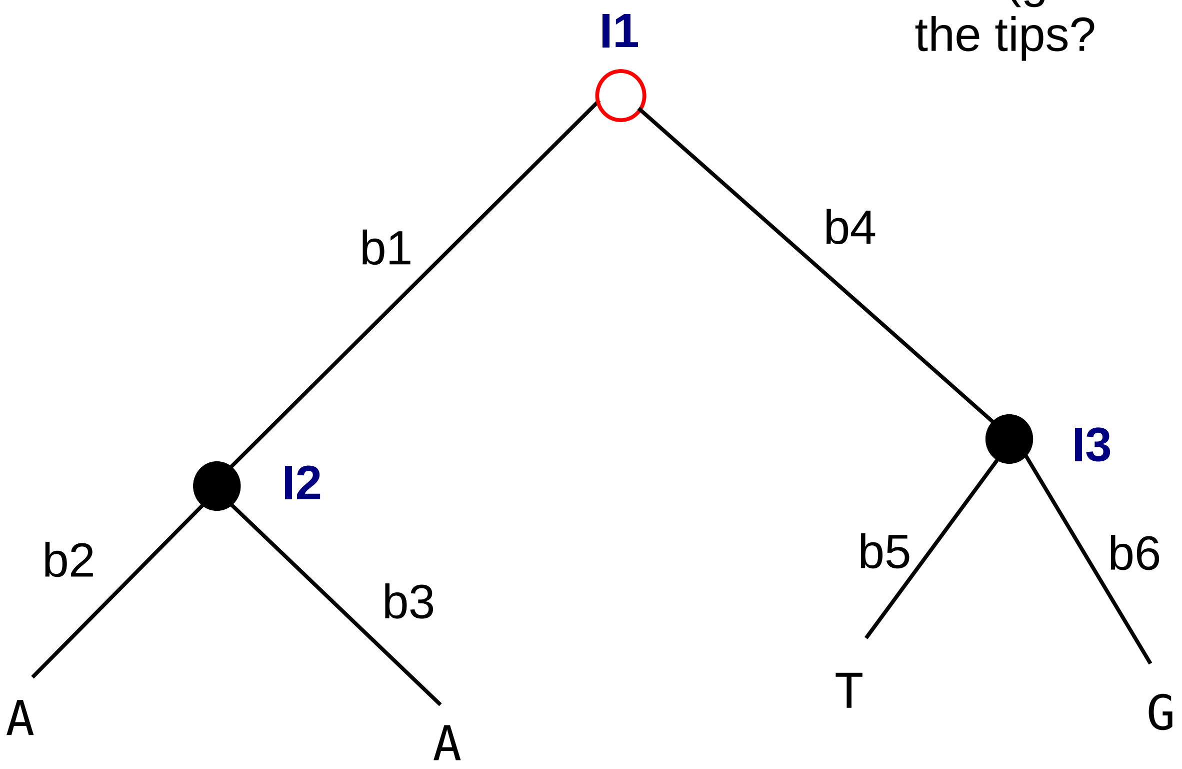
$L_i = \pi_A\, P_{AA}(b_1)\, P_{AA}(b_2)\, P_{AA}(b_3)$
$P_{AT}(b_4) P_{TT}(b_5)\, P_{TG}(b_6)$

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

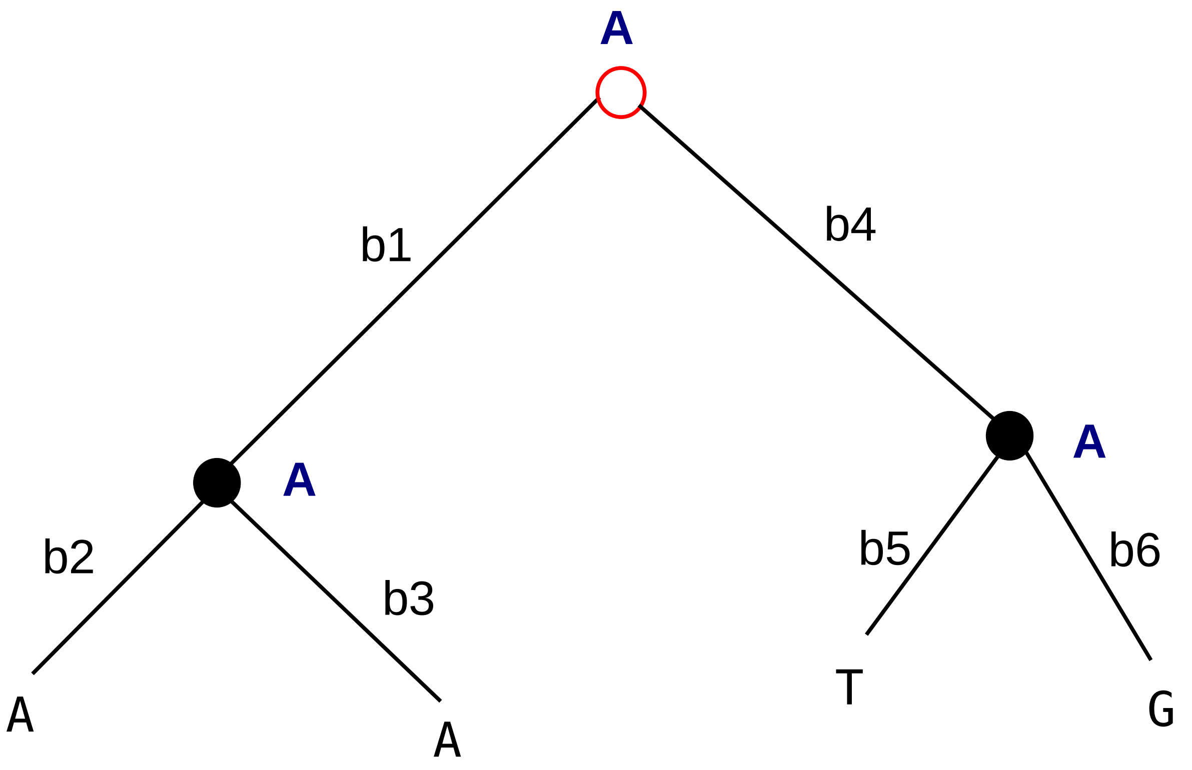
What's the likelihood of this tree?

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?



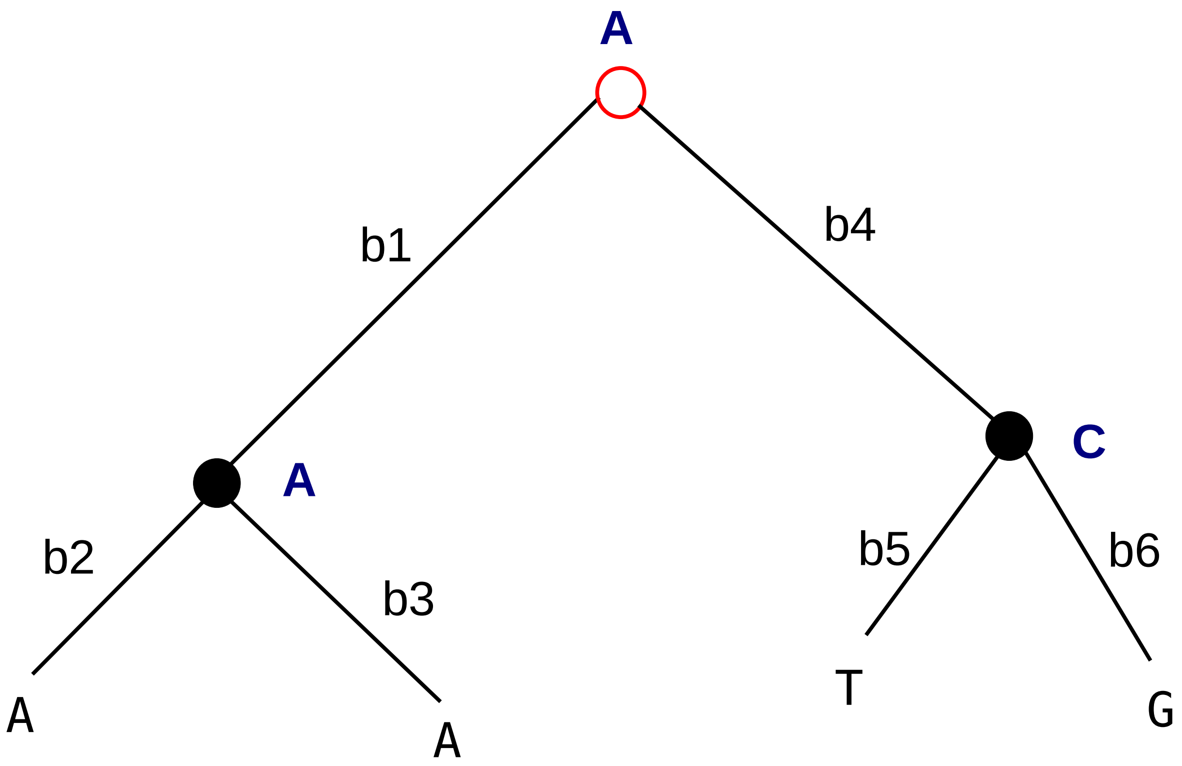
What's the likelihood of this tree?

It could be this



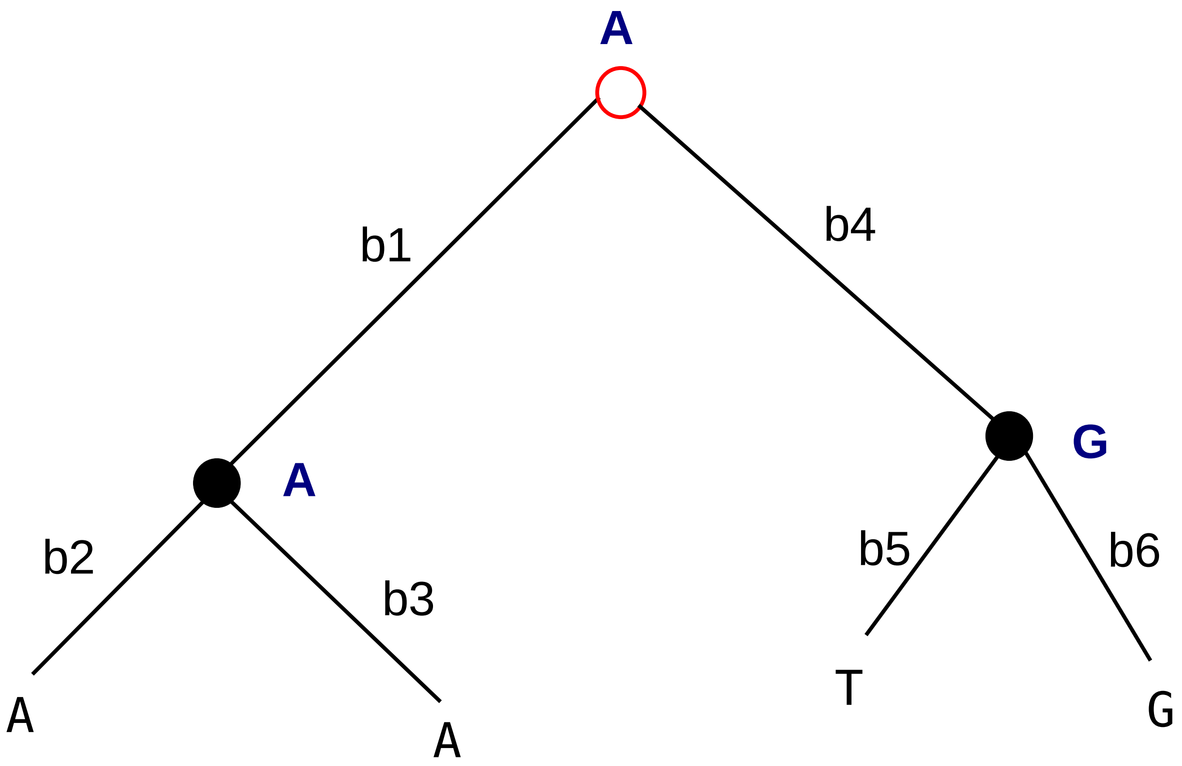
What's the likelihood of this tree?

It could be this
OR this

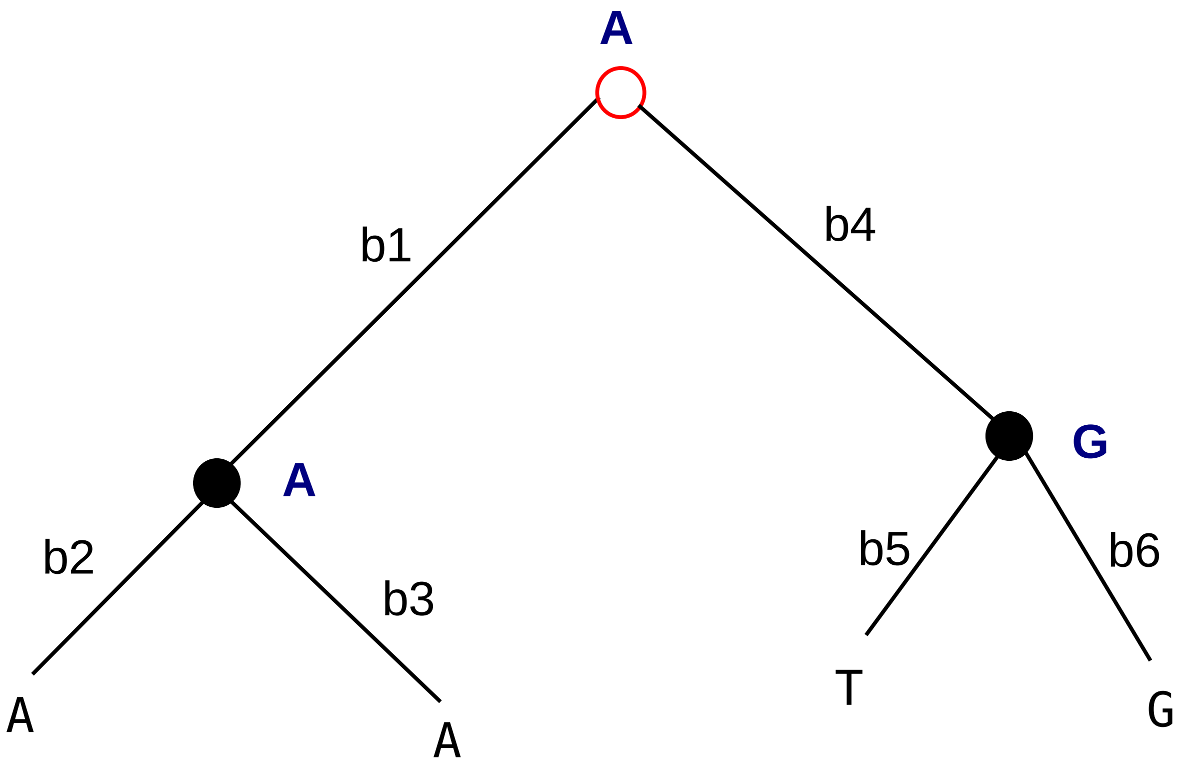


What's the likelihood of this tree?

It could be this
OR this
OR this

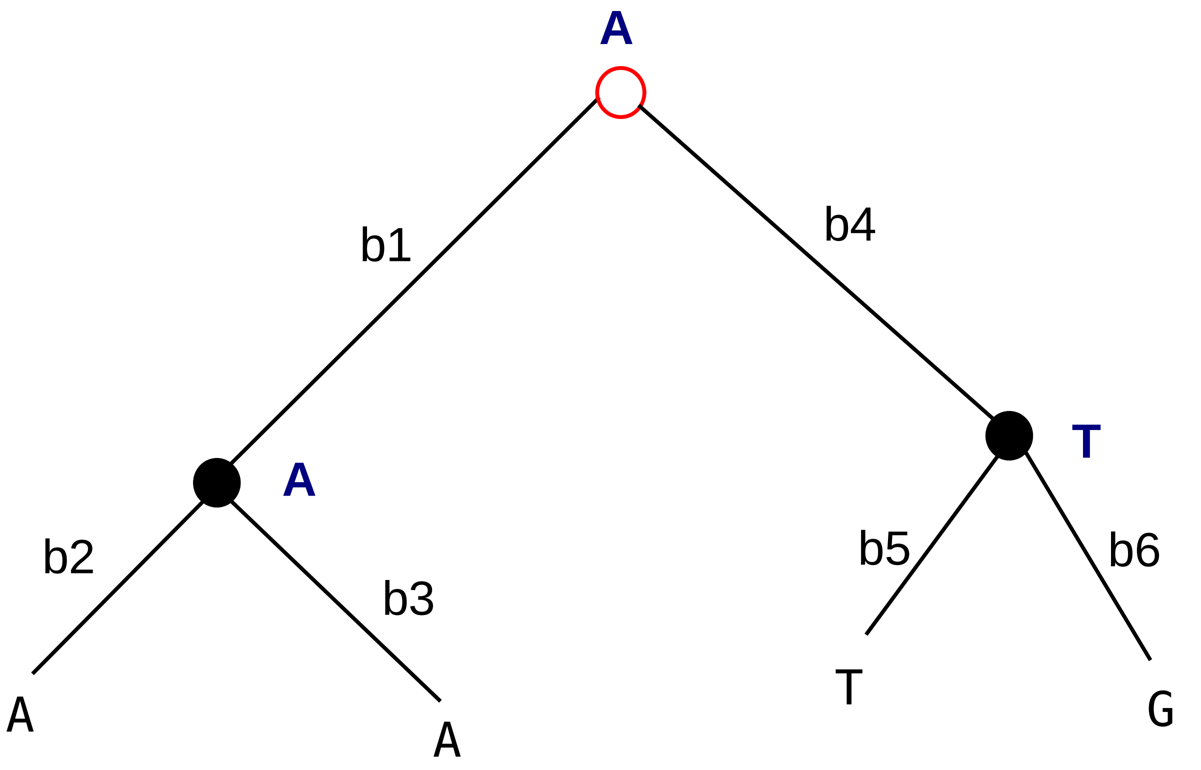


What's the likelihood of this tree?



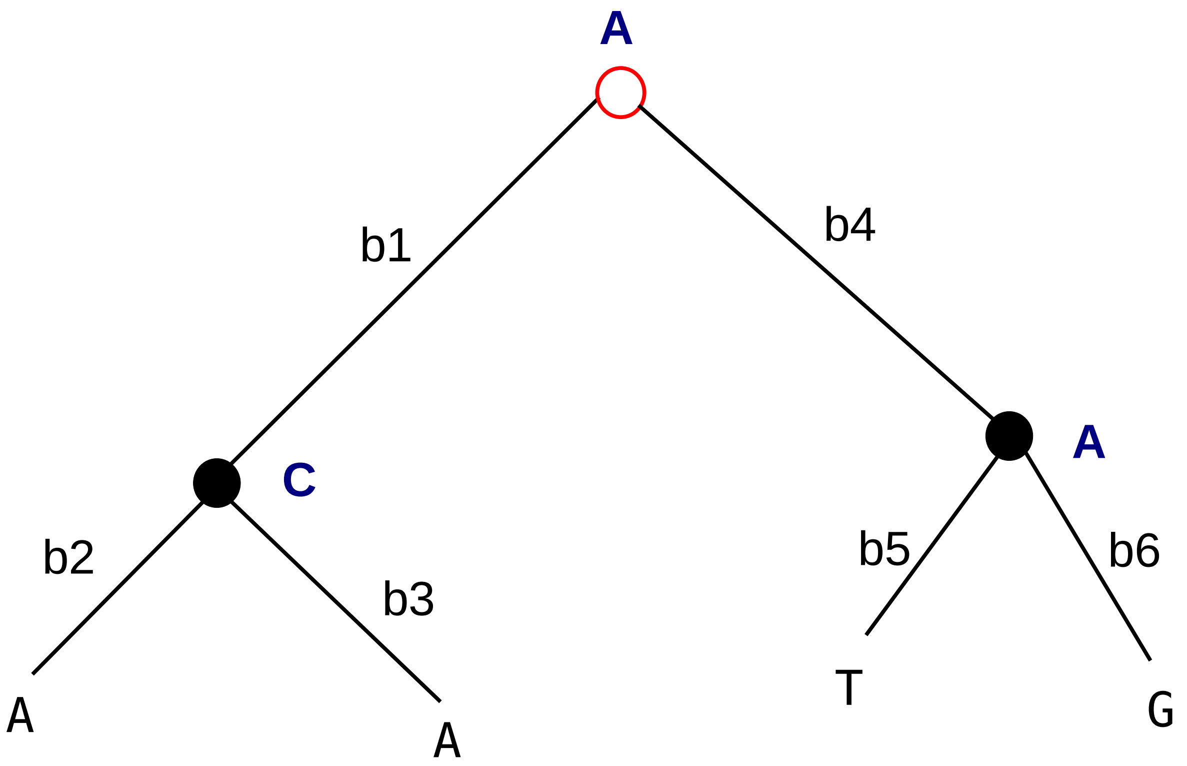
It could be this
OR this
OR this
OR this

What's the likelihood of this tree?



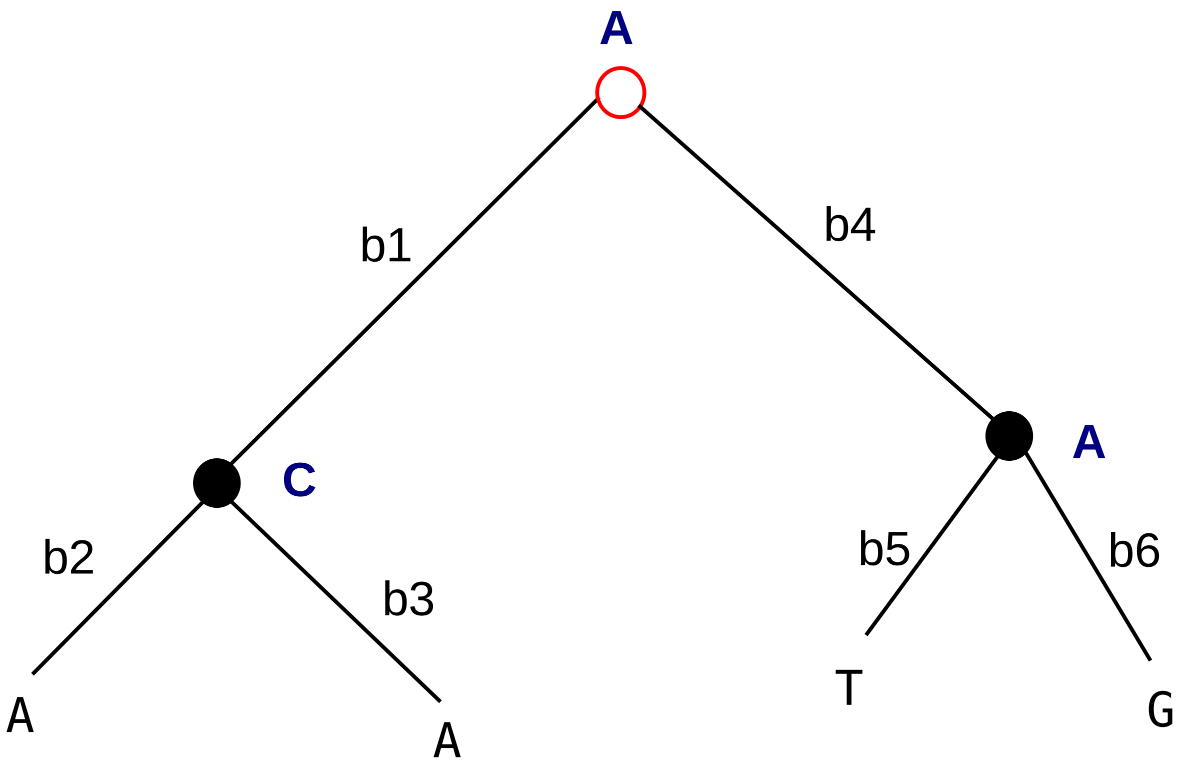
It could be this
OR this
OR this
OR this
OR this

What's the likelihood of this tree?



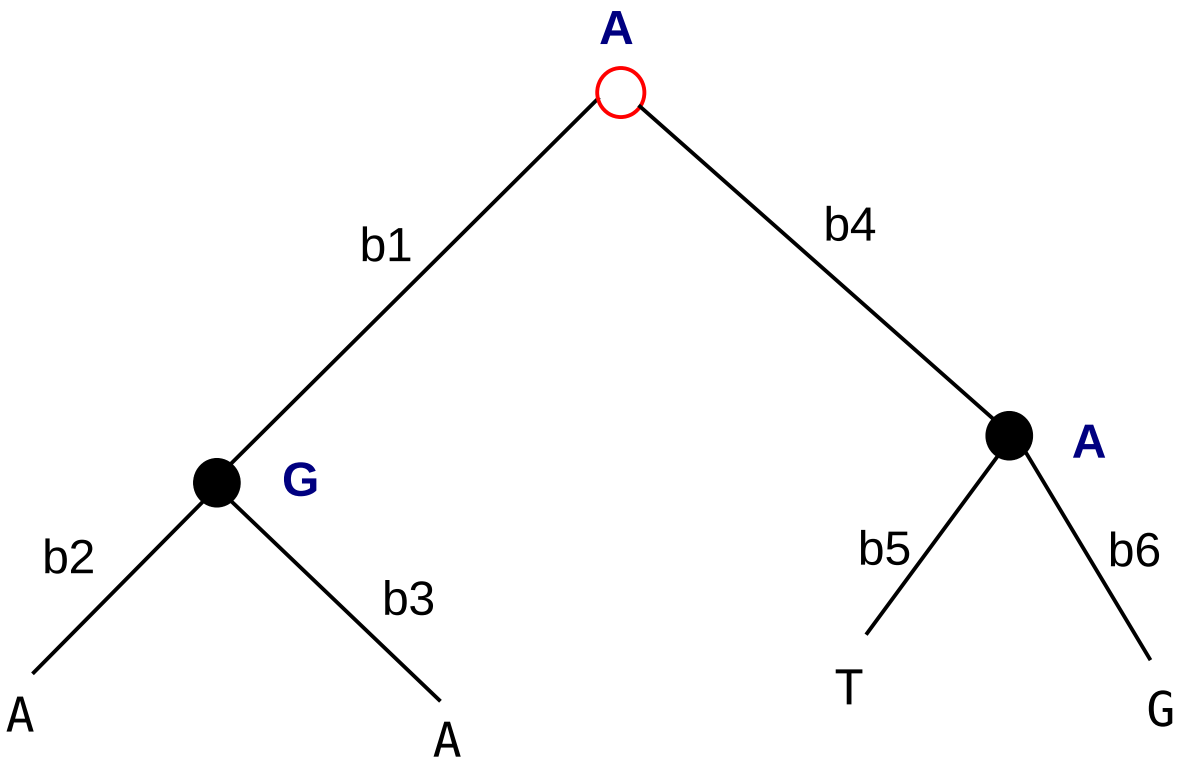
It could be this
OR this
OR this
OR this
OR this
OR this

What's the likelihood of this tree?



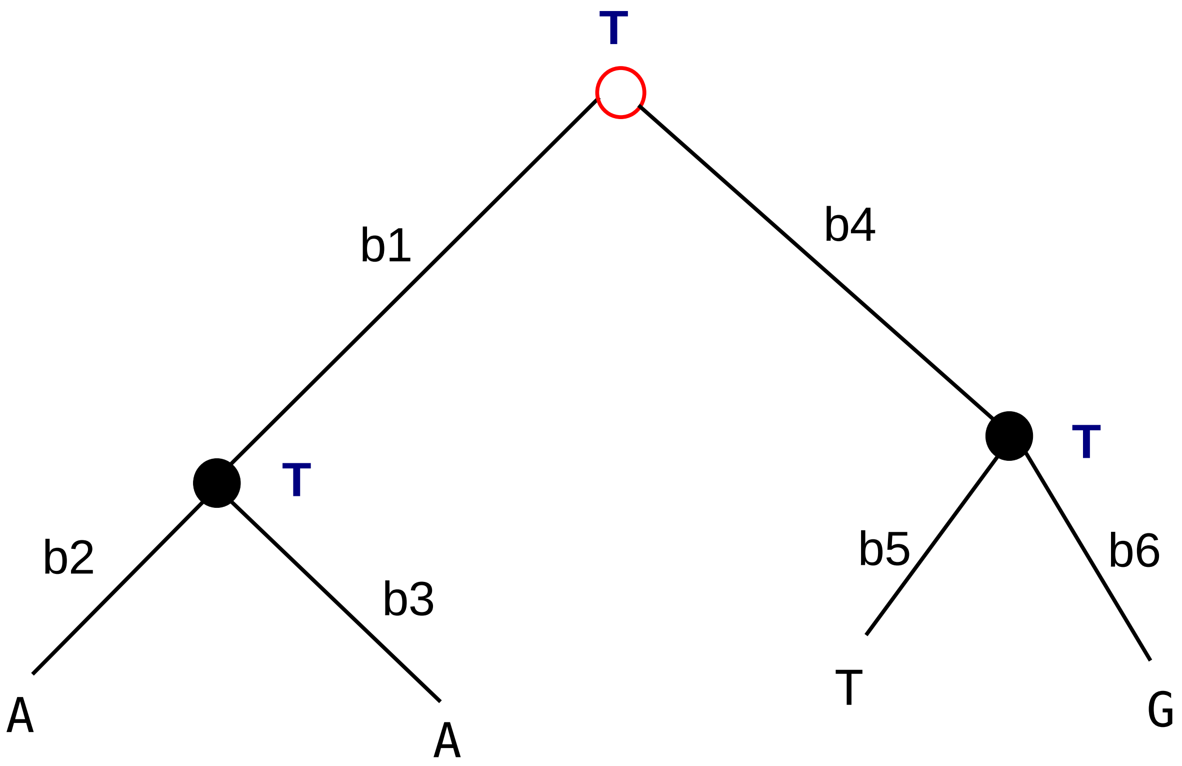
It could be this
OR this
OR this
OR this
OR this
OR this
OR this

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

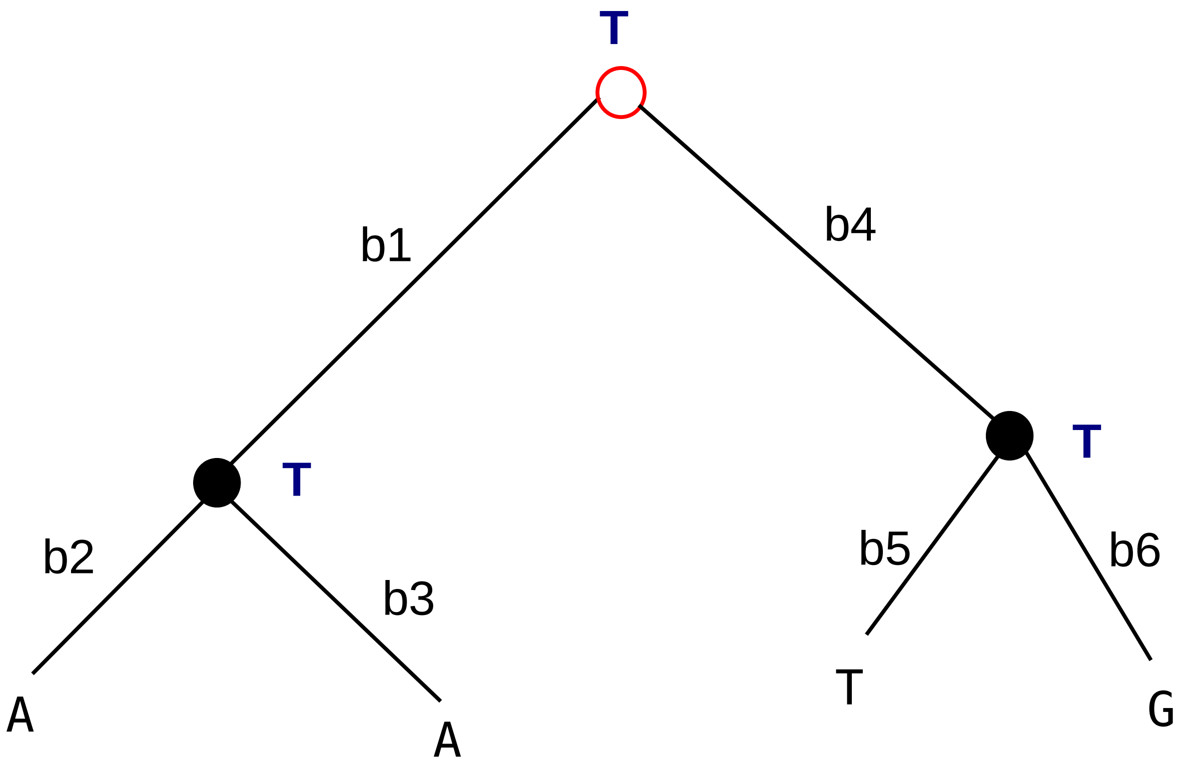
What's the likelihood of this tree?



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?

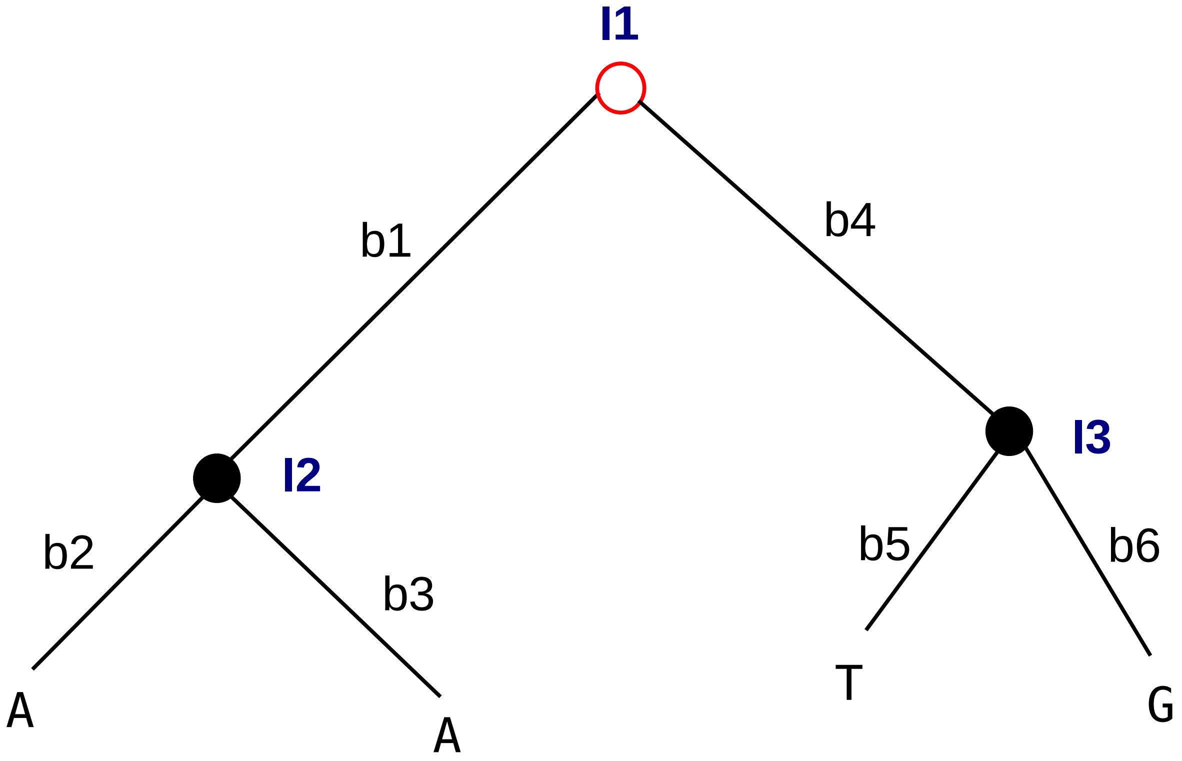
So the likelihood of the tree is the sum (**OR!**) over the likelihoods of all possible assignments of A, C, G, and T (all possible evolutionary histories) to the inner nodes *I1, I2, I3* of the tree.



What's the likelihood of this tree?

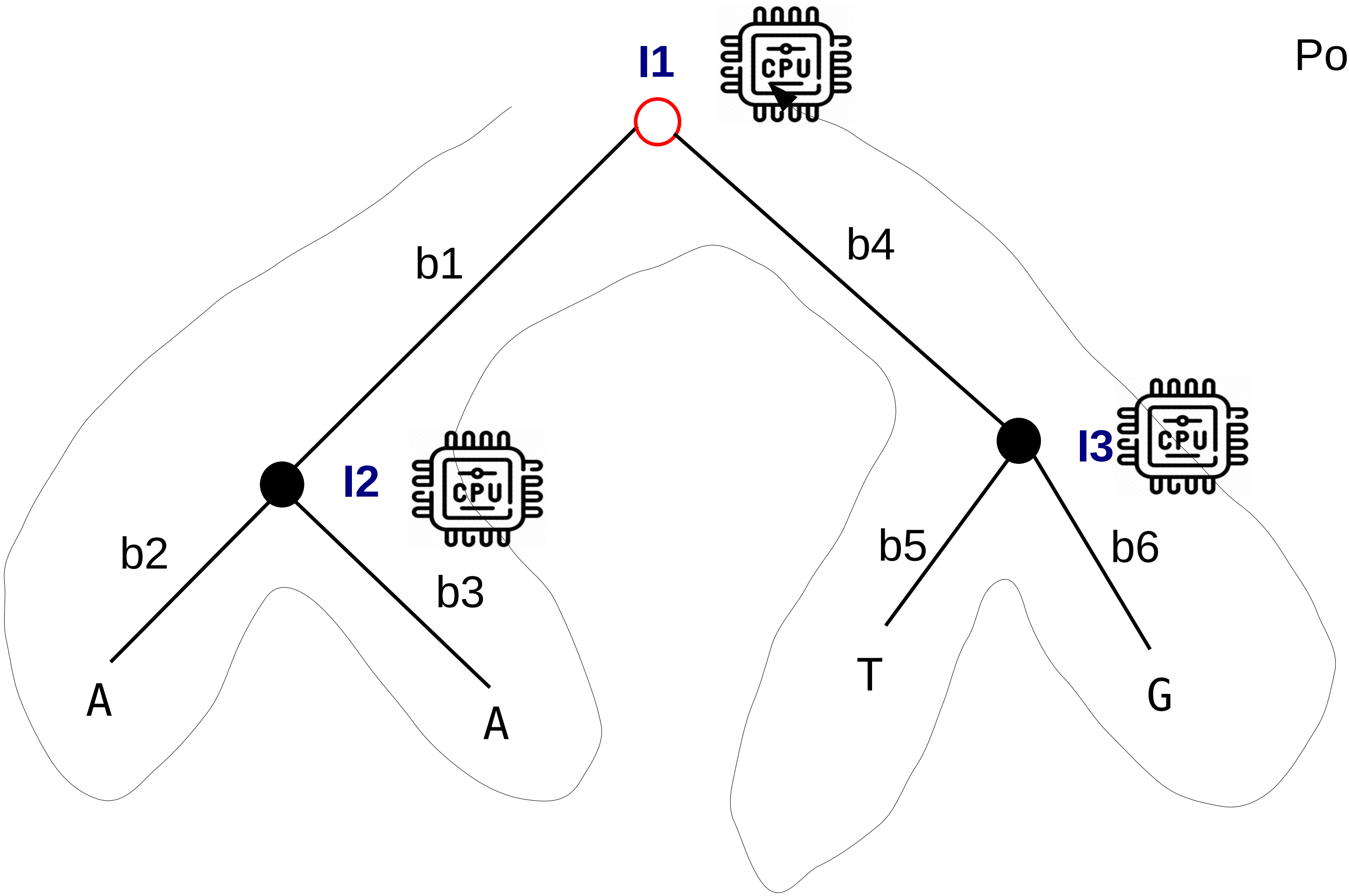
So the likelihood of the tree is the sum (**OR!**) over the likelihoods of all possible assignments of A, C, G, and T (all possible evolutionary histories) to the inner nodes *I1, I2, I3* of the tree.

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

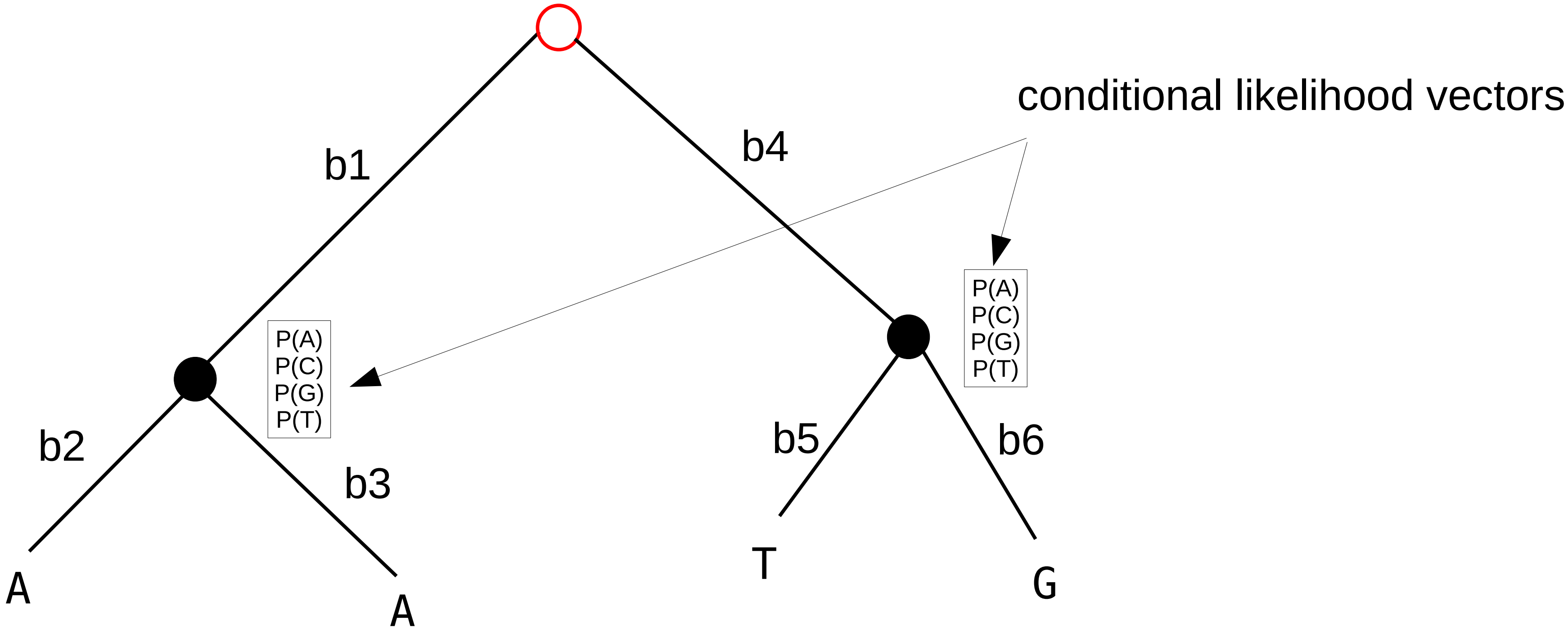


The Felsenstein Pruning Algorithm

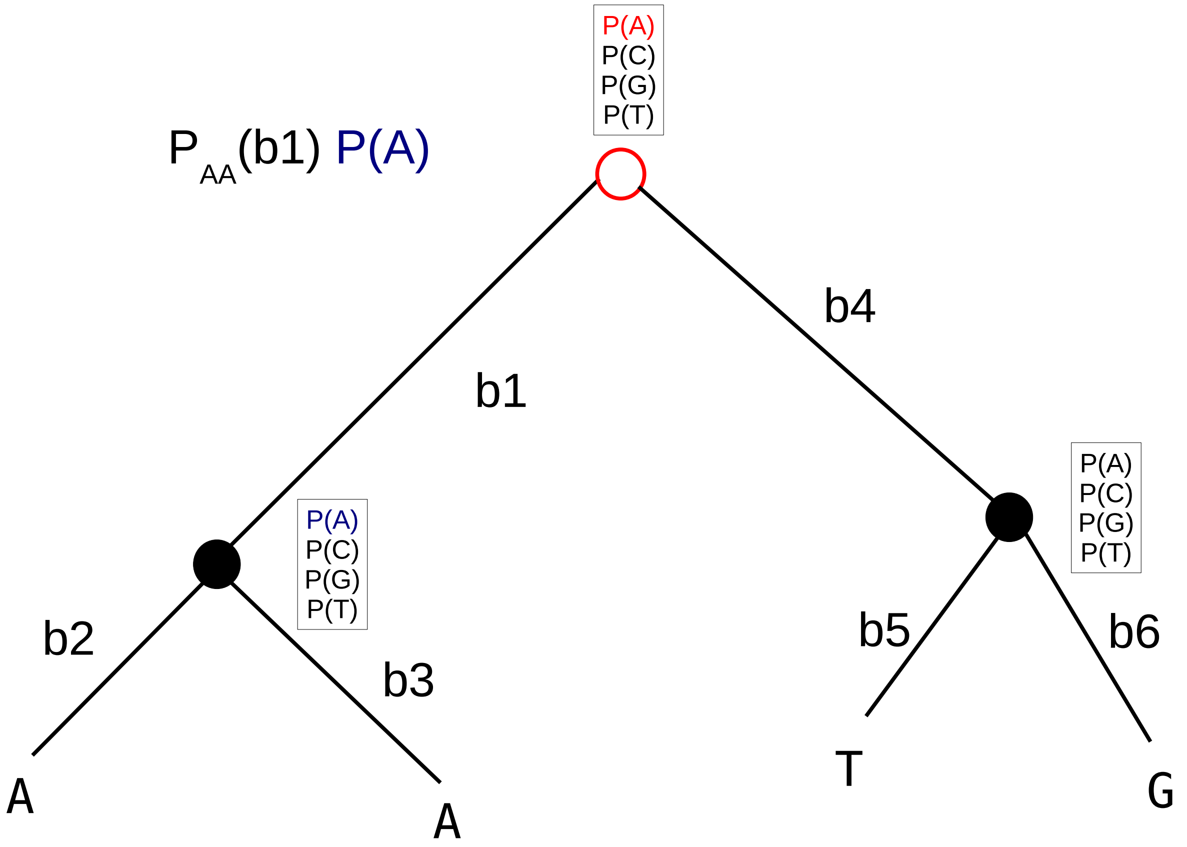
Post order traversal



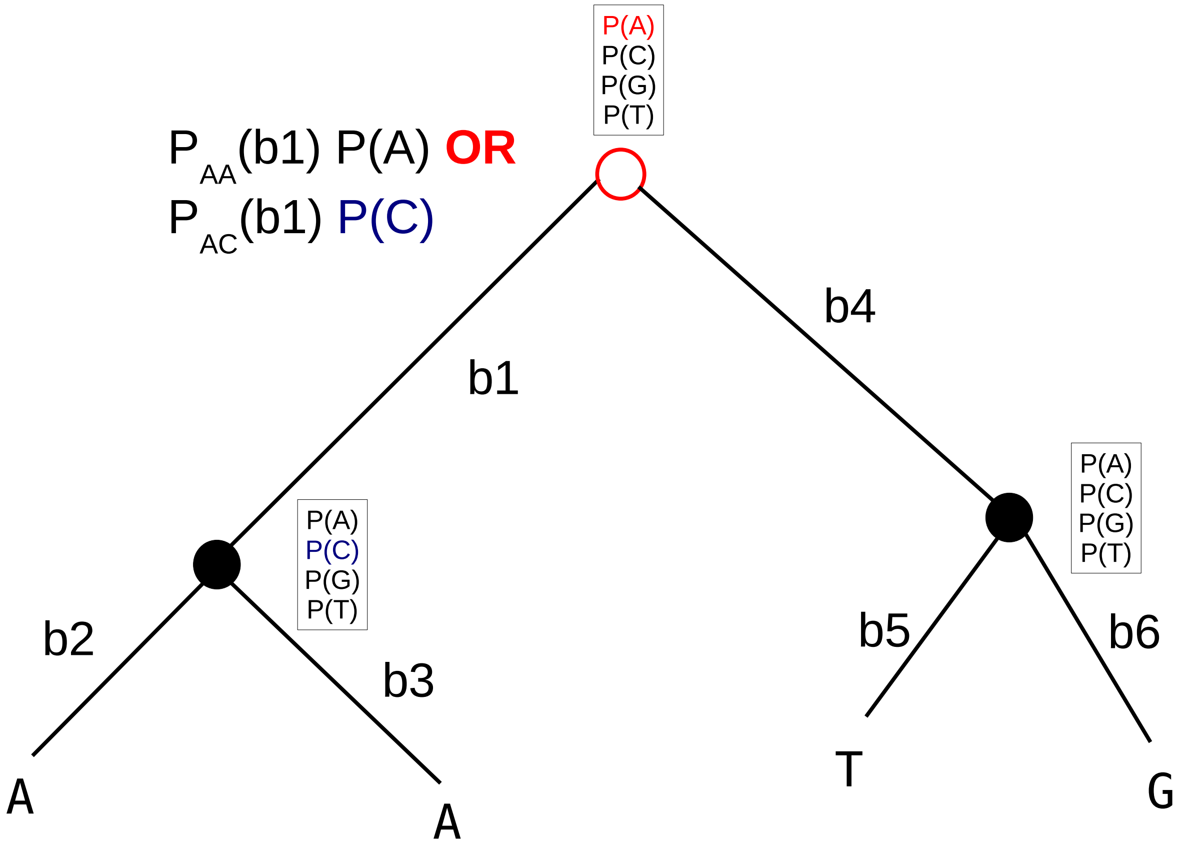
Felsenstein Pruning



Felsenstein Pruning



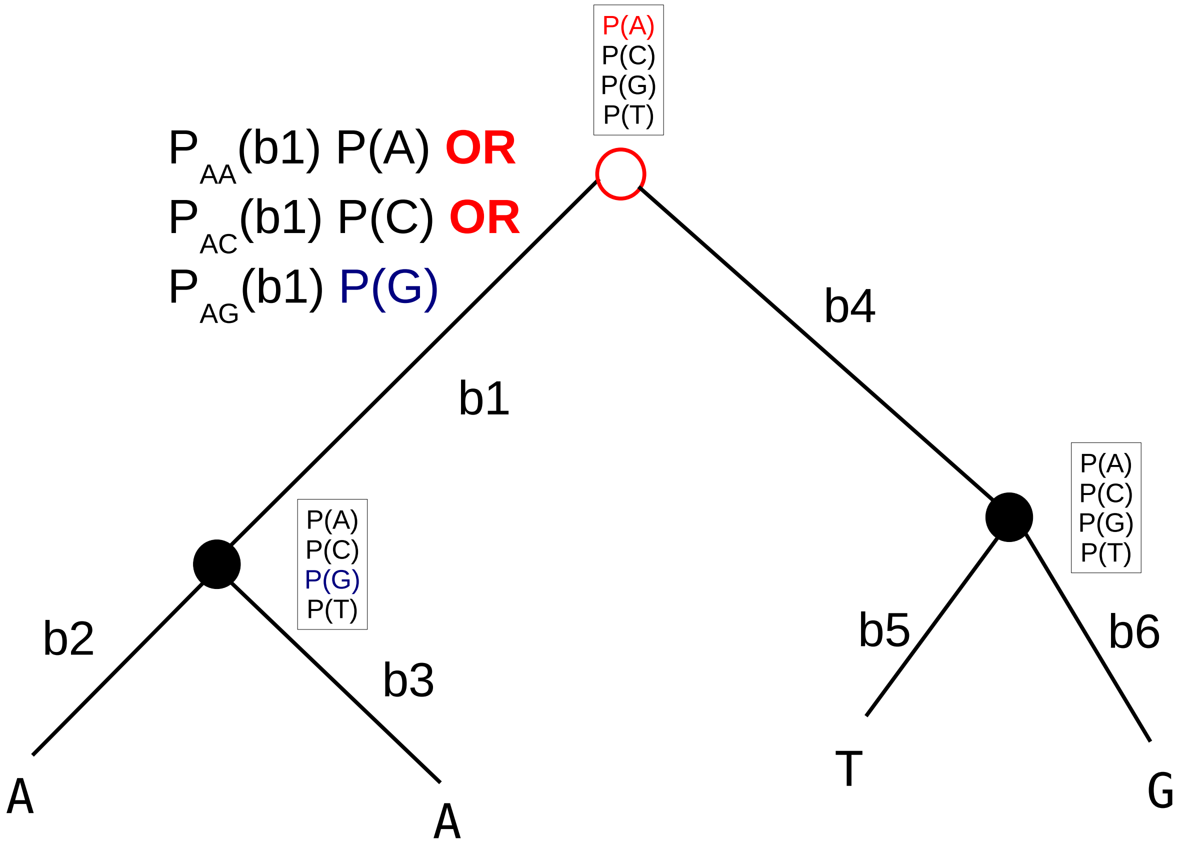
Felsenstein Pruning



$P_{AA}(b1)\,P(A)$ **OR**

$P_{AC}(b1)\,P(C)$

Felsenstein Pruning

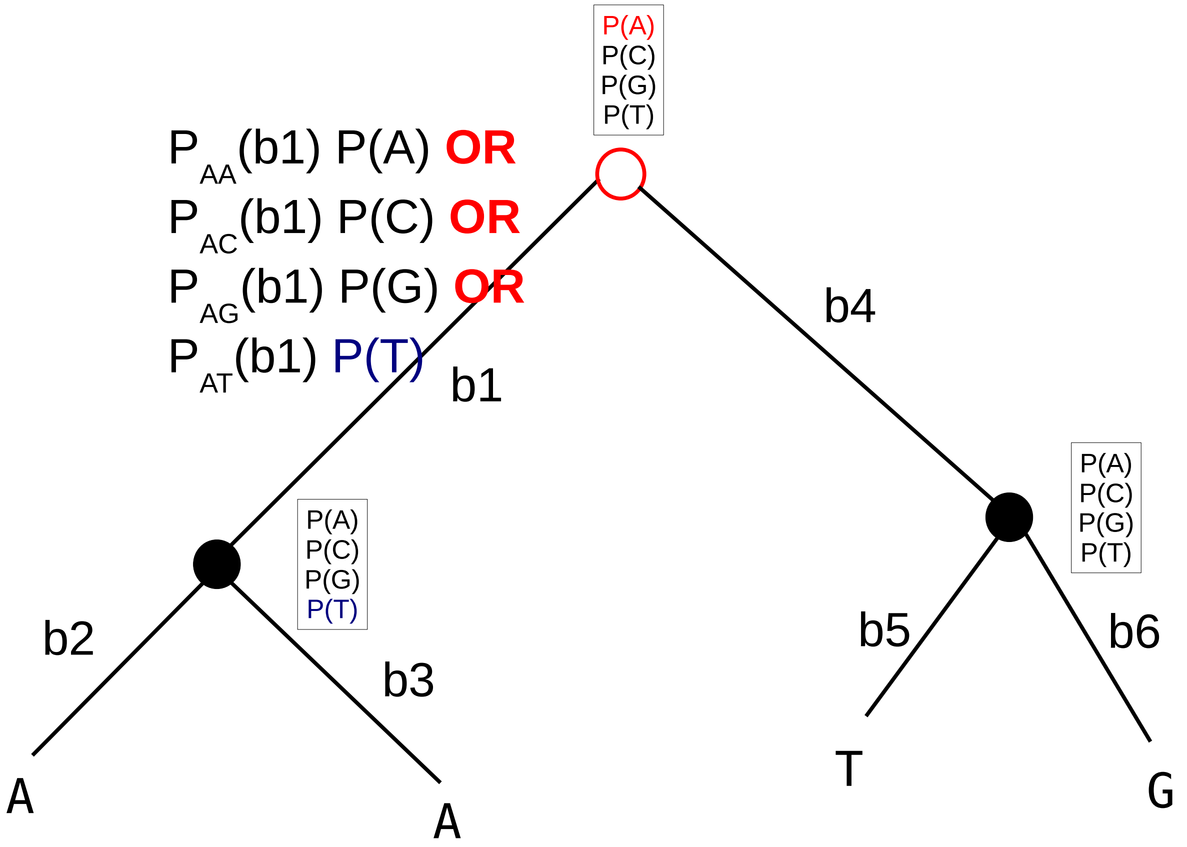


$P_{AA}(b1)\ P(A)$ **OR**

$P_{AC}(b1)\ P(C)$ **OR**

$P_{AG}(b1)\ P(G)$

Felsenstein Pruning



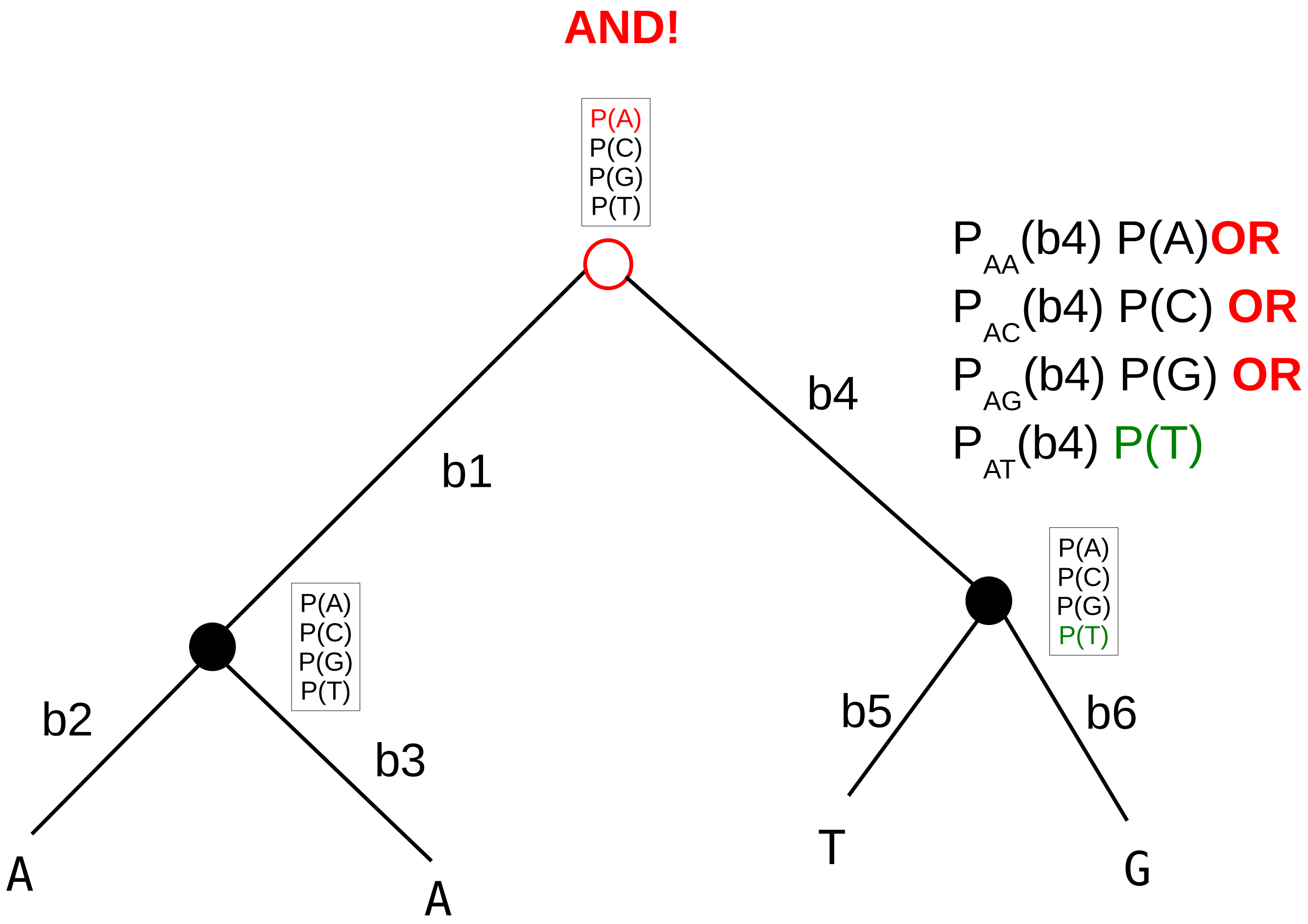
P_{AA}(b1) P(A) **OR**

P_{AC}(b1) P(C) **OR**

P_{AG}(b1) P(G) **OR**

P_{AT}(b1) P(T)

Felsenstein Pruning



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

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

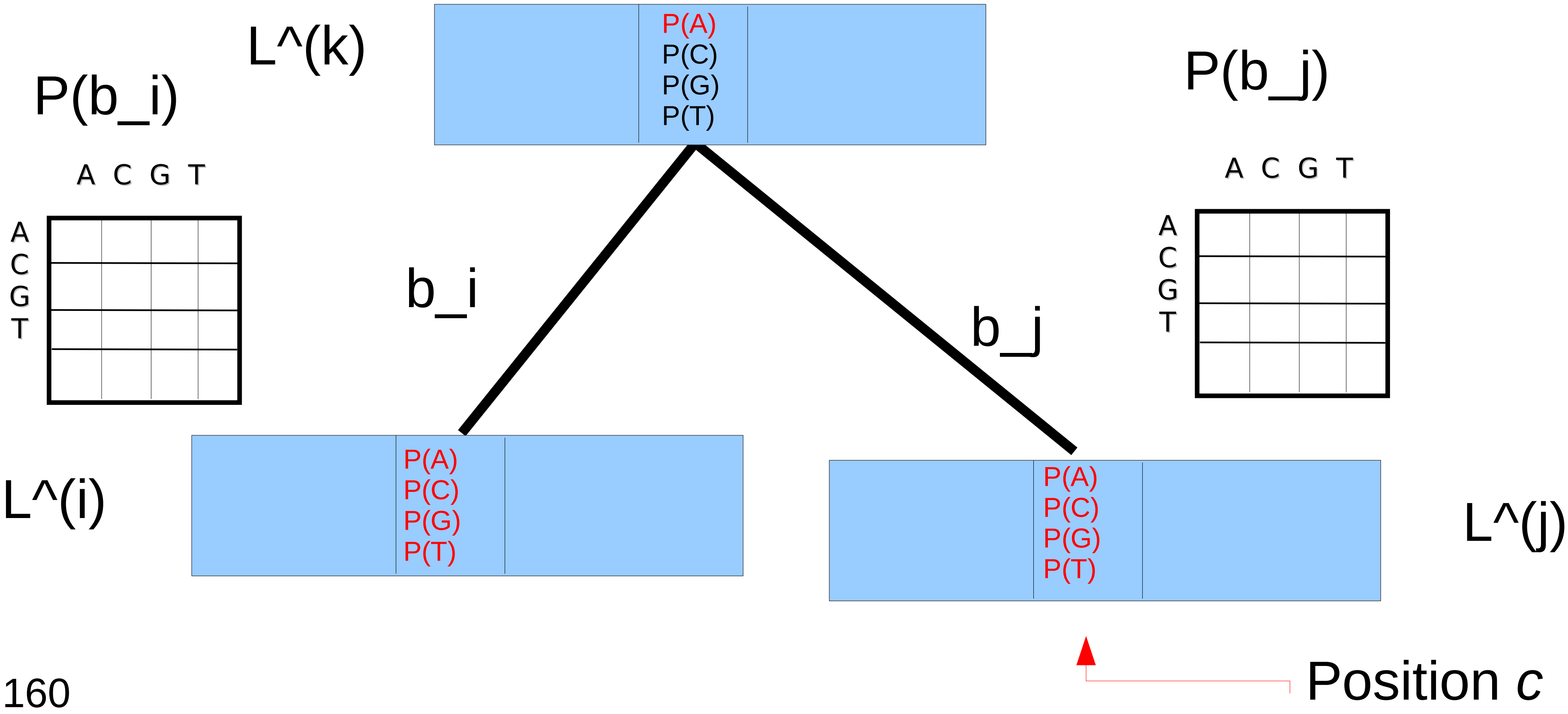
P_{AG}(b4) P(G) **OR**

P_{AT}(b4) P(T)

Felsenstein Pruning

AND (left branch/right branch)

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

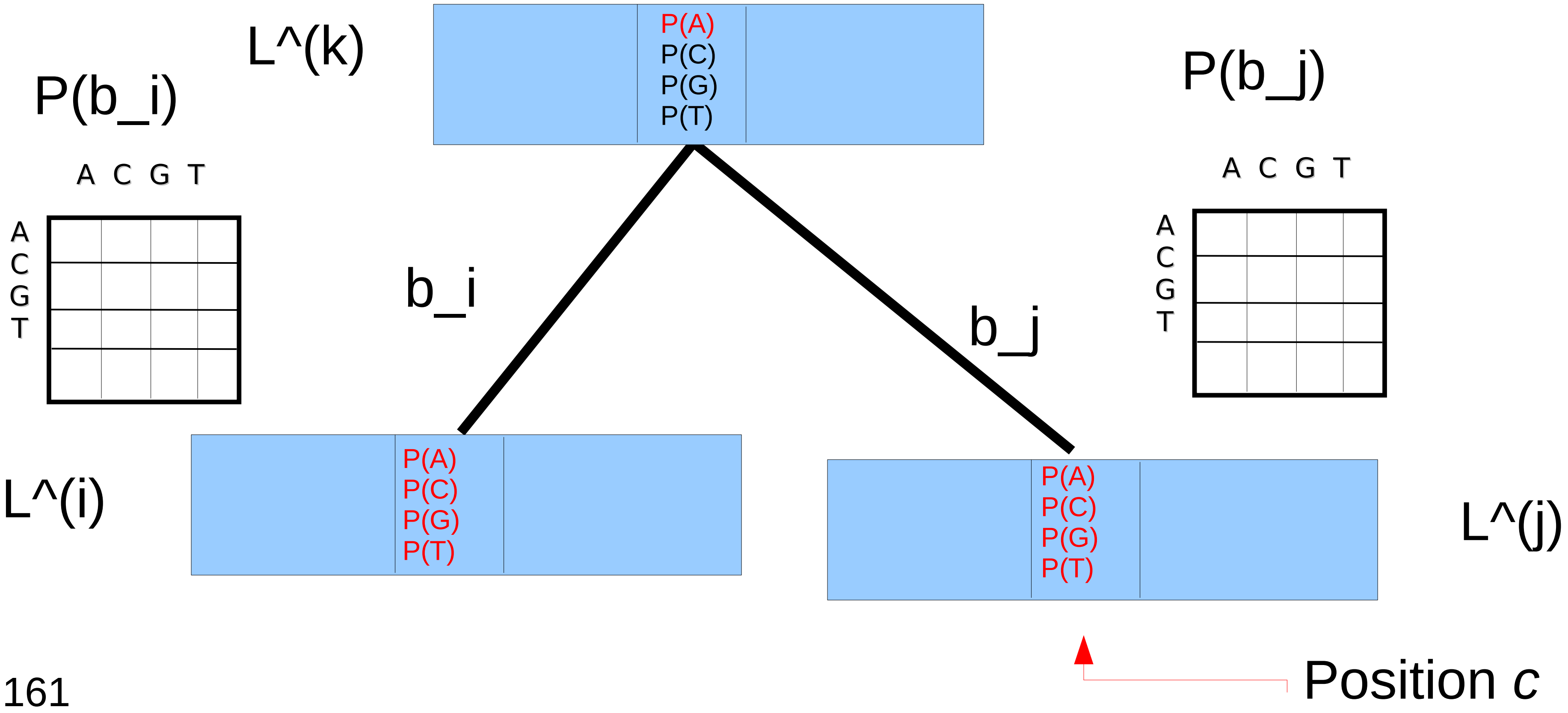


Position *c*

Felsenstein Pruning

OR (along left branch)

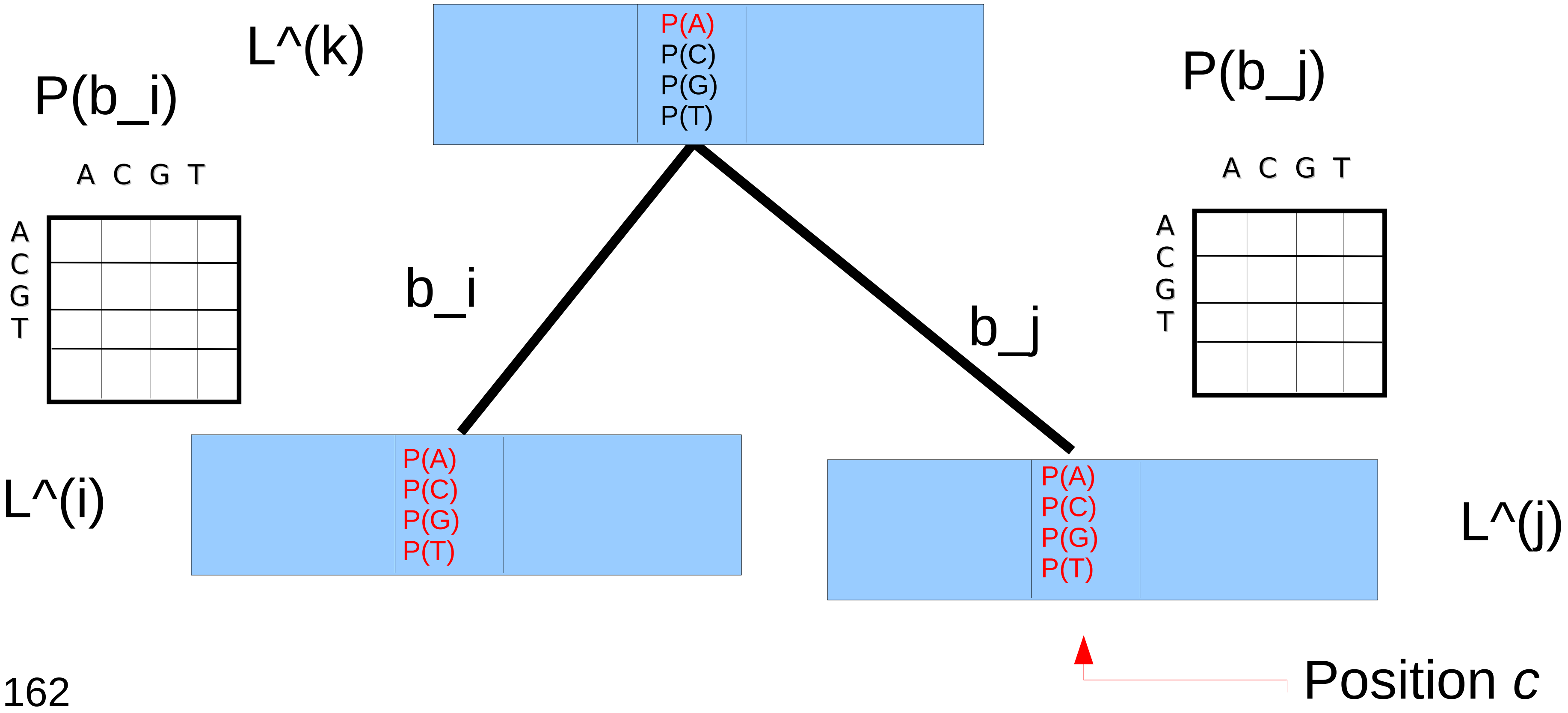
$$\vec{L}_A^{(k)}(c) = (\sum_{S=A}^{T} P_{AS}(b_i)\vec{L}_S^{(i)}(c))(\sum_{S=A}^{T} P_{AS}(b_j)\vec{L}_S^{(j)}(c))$$



Felsenstein Pruning

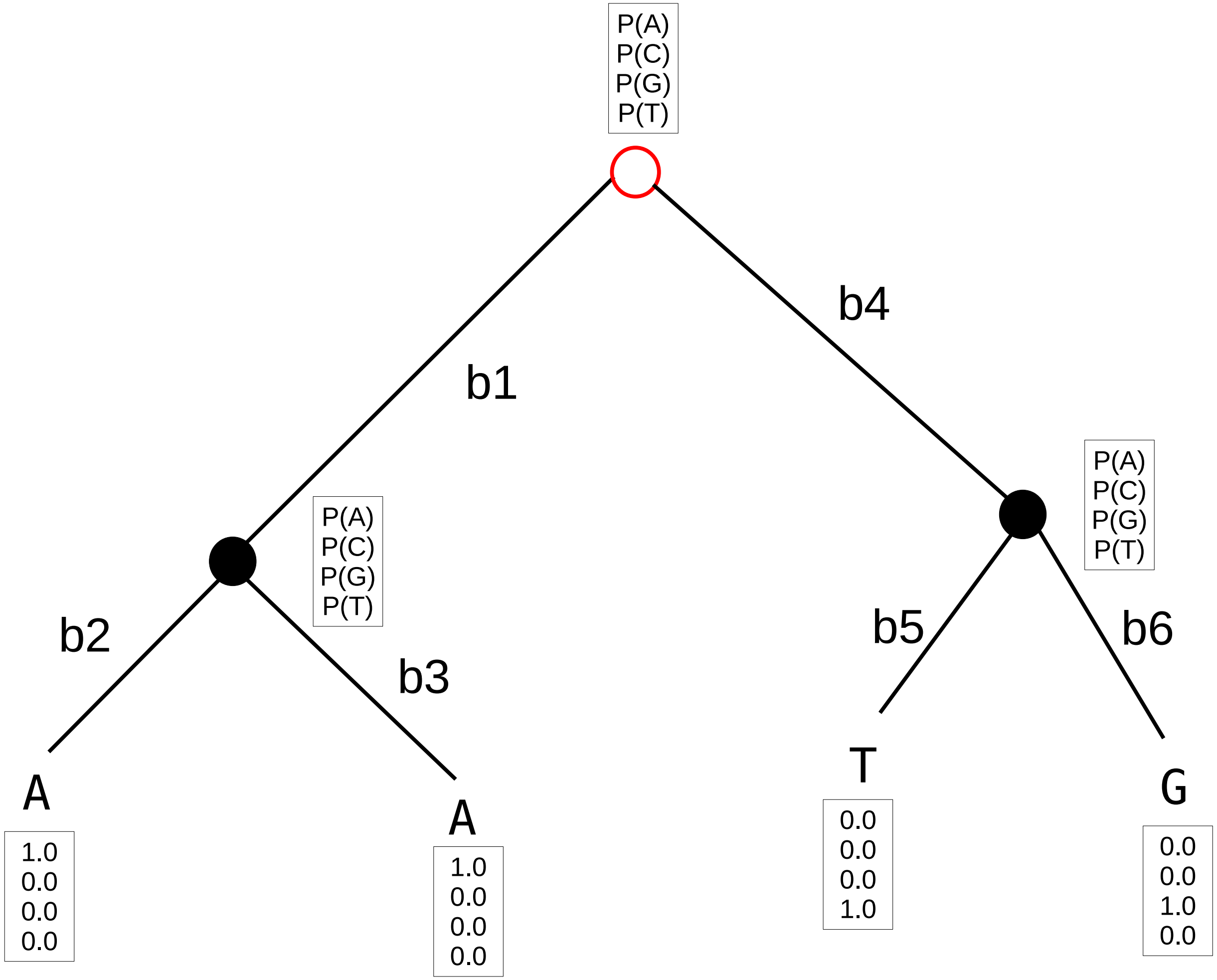
OR (along right branch)

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



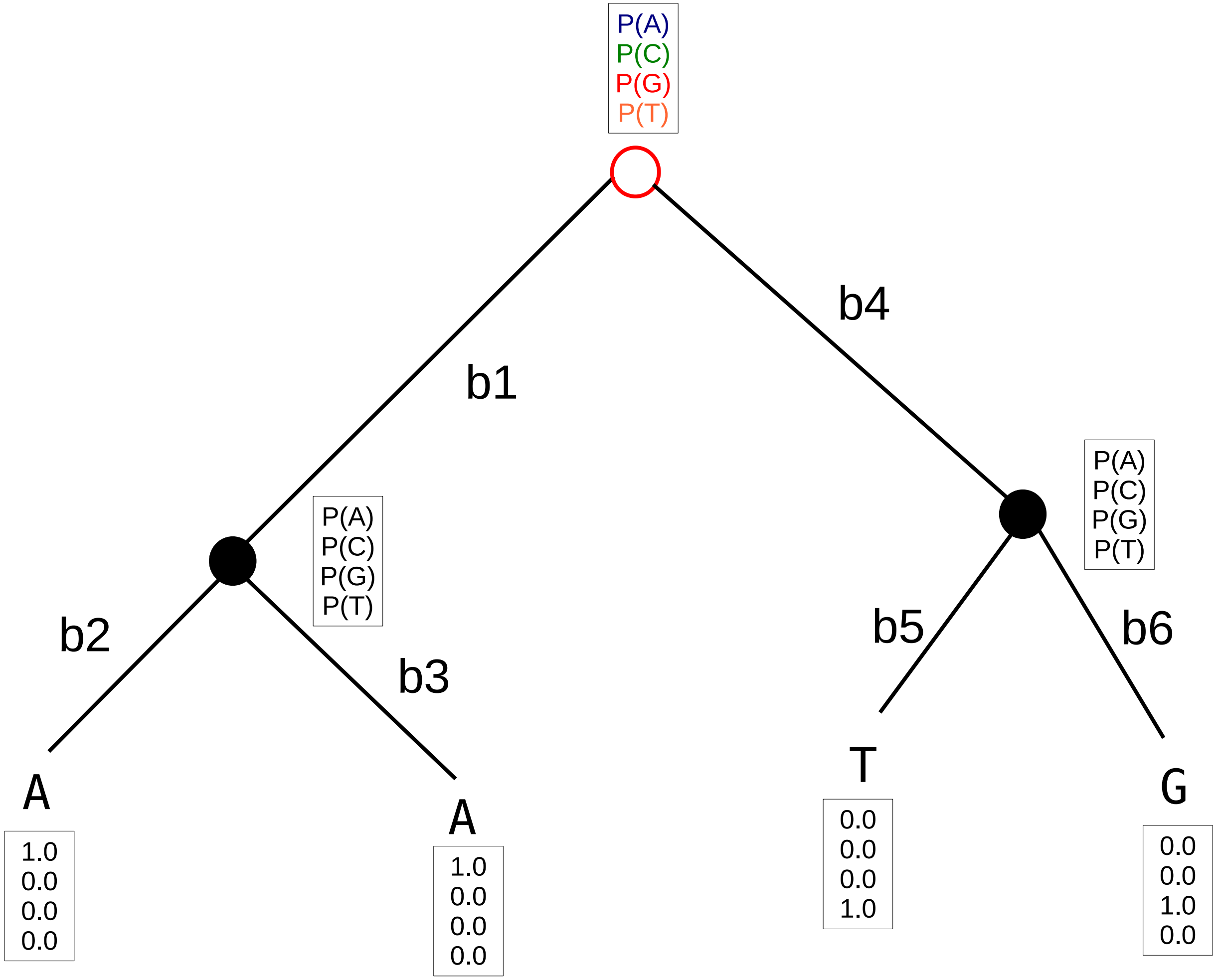
Position *c*

Felsenstein Pruning



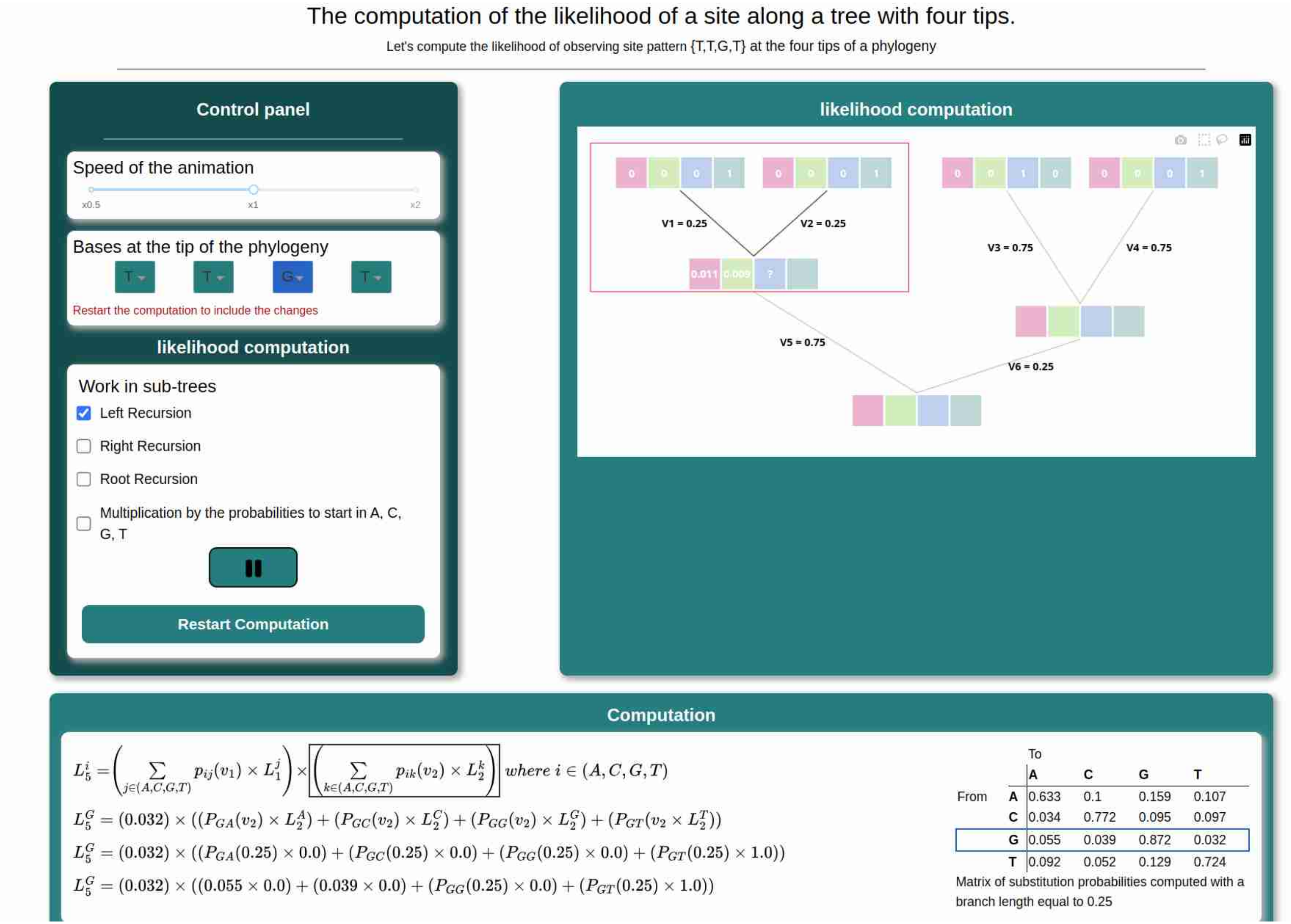
Felsenstein Pruning

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



An Excellent Tool to viualize and revise this concept

- https://phylanim.univ-lyon1.fr/LikelihoodTreeComputation



The computation of the likelihood of a site along a tree with four tips.

Let's compute the likelihood of observing site pattern {T,T,G,T} at the four tips of a phylogeny

Control panel

Speed of the animation

Bases at the tip of the phylogeny

Restart the computation to include the changes

likelihood computation

Work in sub-trees

☑ Left Recursion

□ Right Recursion

□ Root Recursion

□ Multiplication by the probabilities to start in A, C, G, T

likelihood computation

Computation

$$L_5^i = \left(\sum_{j \in (A,C,G,T)} p_{ij}(v_1) \times L_1^j \right) \times \boxed{\left(\sum_{k \in (A,C,G,T)} p_{ik}(v_2) \times L_2^k \right)} \; where \; i \in (A,C,G,T)$$

$$L_5^G = (0.032) \times ((P_{GA}(v_2) \times L_2^A) + (P_{GC}(v_2) \times L_2^C) + (P_{GG}(v_2) \times L_2^G) + (P_{GT}(v_2 \times L_2^T))$$

$$L_5^G = (0.032) \times ((P_{GA}(0.25) \times 0.0) + (P_{GC}(0.25) \times 0.0) + (P_{GG}(0.25) \times 0.0) + (P_{GT}(0.25) \times 1.0))$$

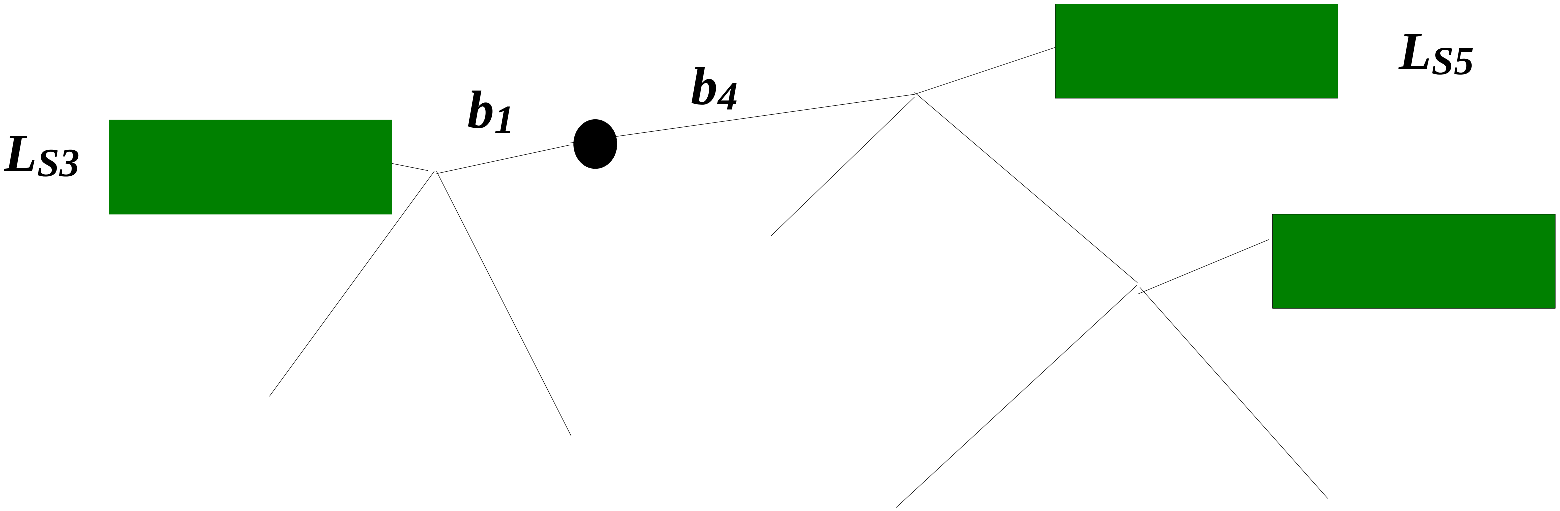
$$L_5^G = (0.032) \times ((0.055 \times 0.0) + (0.039 \times 0.0) + (P_{GG}(0.25) \times 0.0) + (P_{GT}(0.25) \times 1.0))$$

		To			
		A	**C**	**G**	**T**
From	**A**	0.633	0.1	0.159	0.107
	C	0.034	0.772	0.095	0.097
	G	0.055	0.039	0.872	0.032
	T	0.092	0.052	0.129	0.724

Matrix of substitution probabilities computed with a branch length equal to 0.25

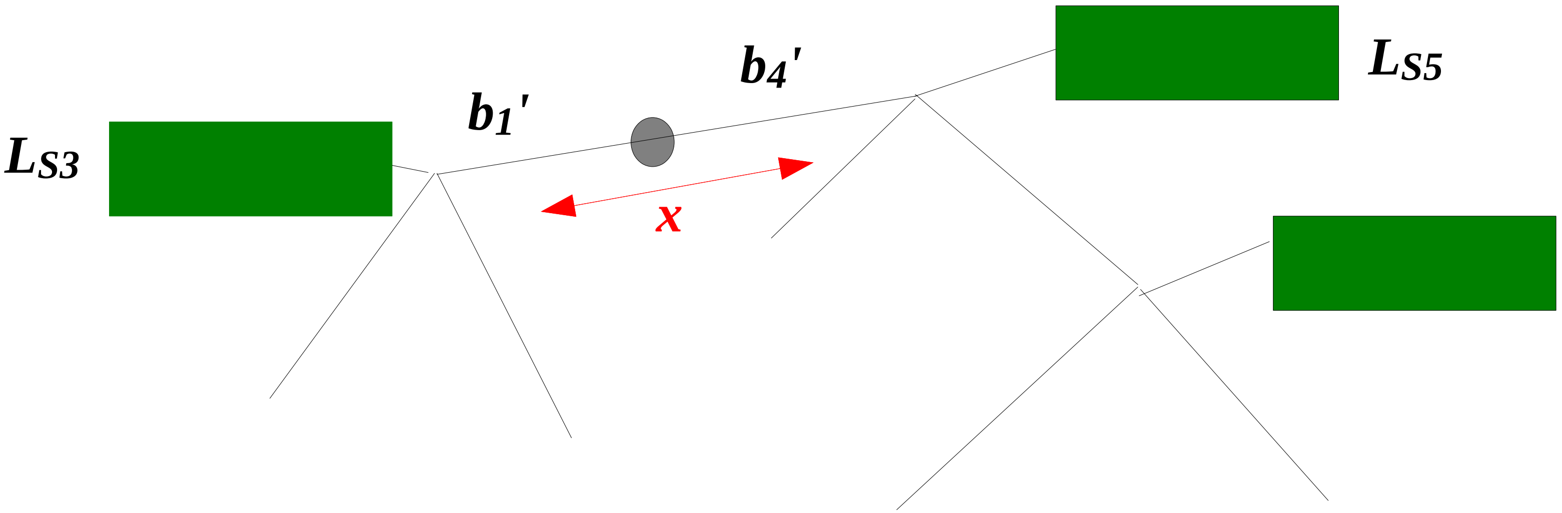
Why is time-reversibility important?

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



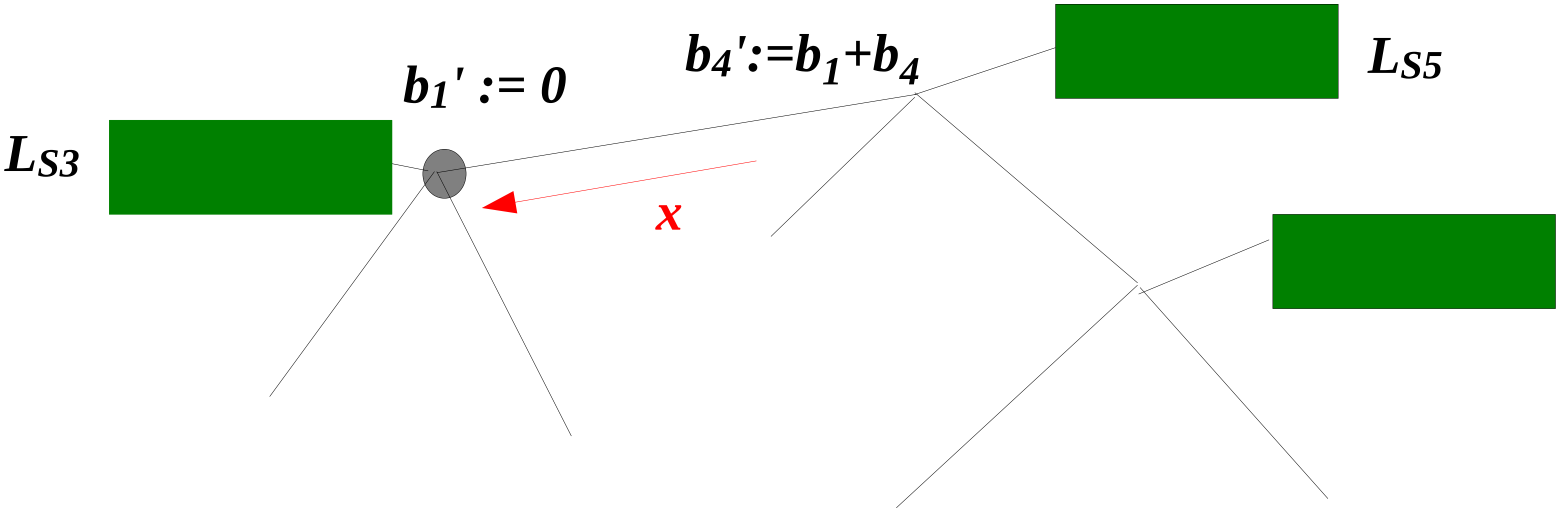
Why is time-reversibility important?

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



Why is time-reversibility important?

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



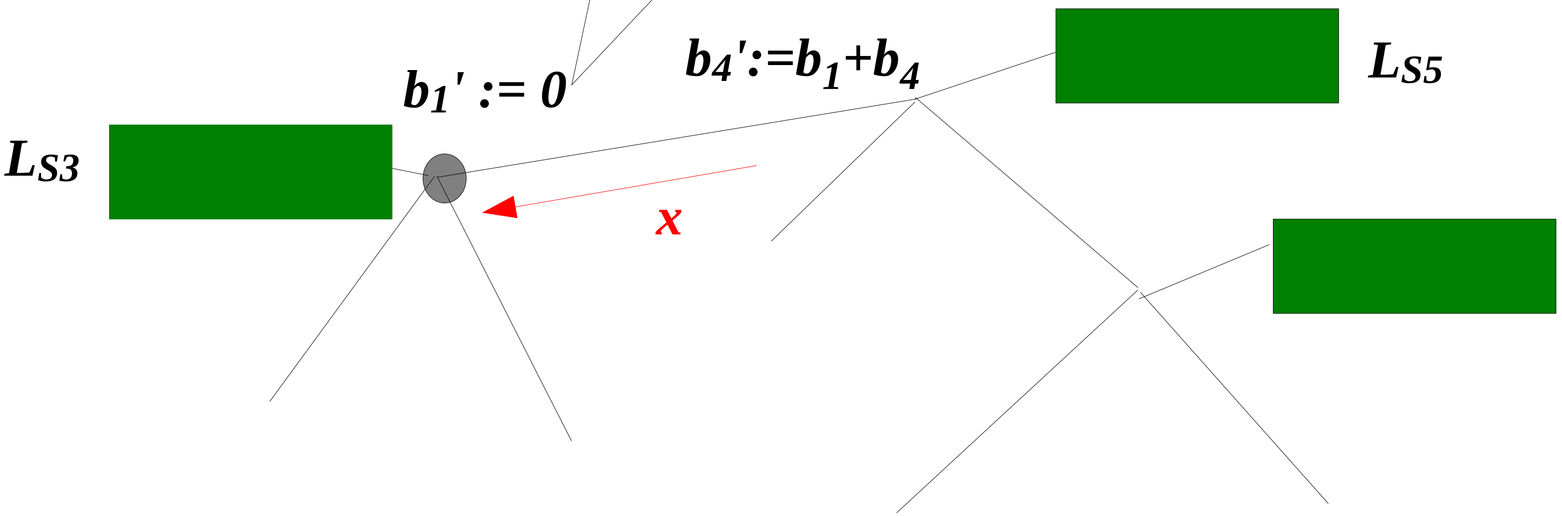
Why is time-reversibility important?

This observation can be applied recursively to the tree

→

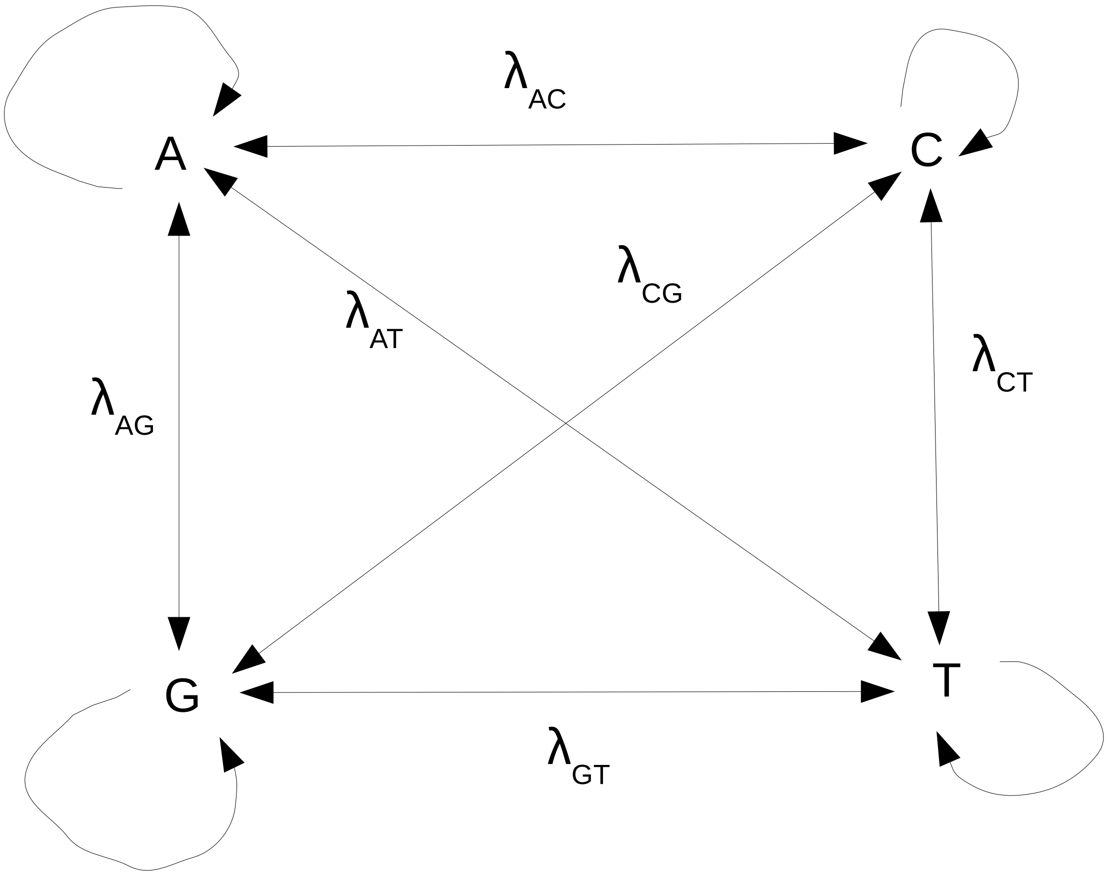
It does not matter at all where we place the root!

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



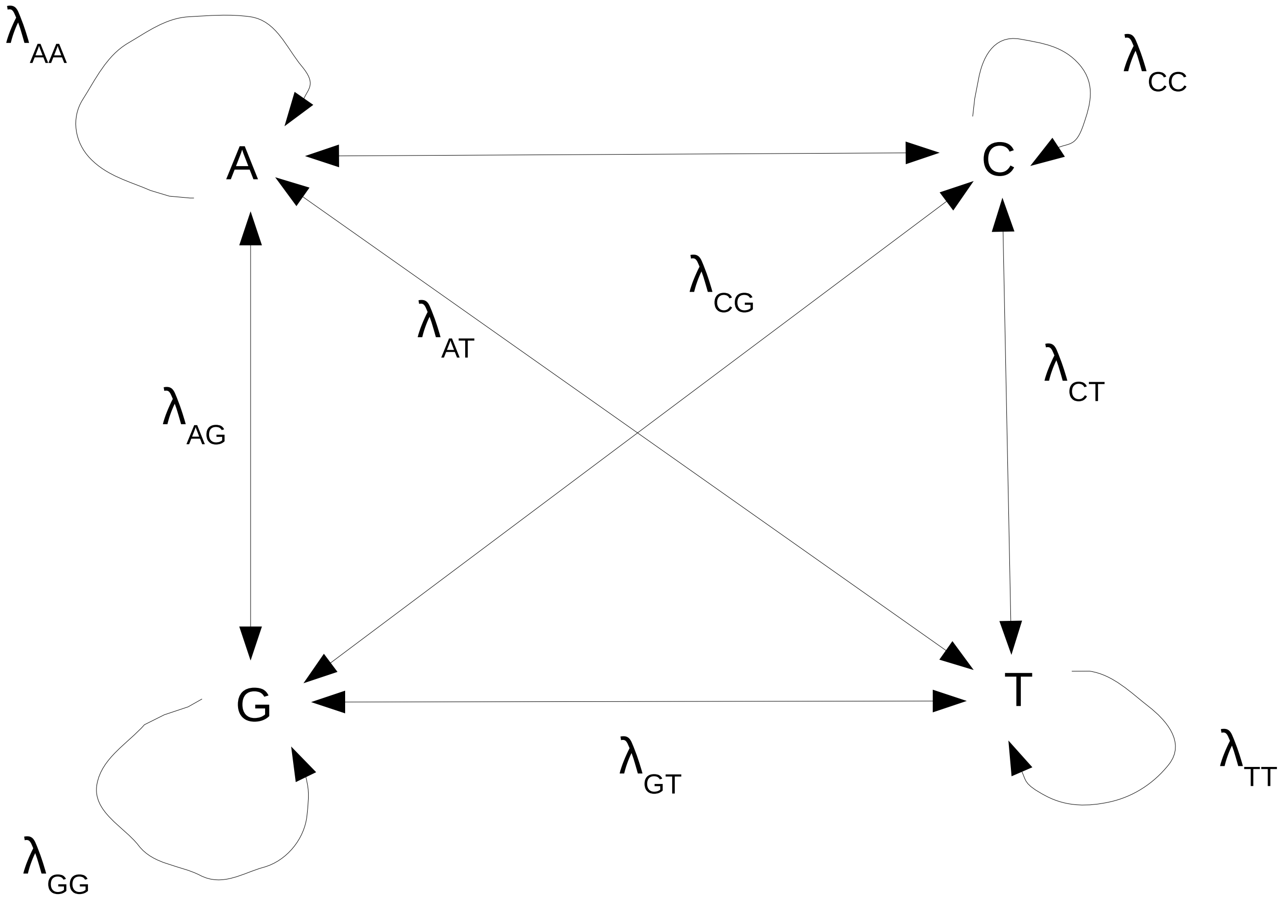
What's in the black box $P_{ij}(t)$?

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., $\lambda_{AA} = -(\lambda_{AC} + \lambda_{AG} + \lambda_{AT})$ → remember from lecture on Markov models:
rows in the *R* matrix need to sum to **0**



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

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

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

Diagonal values are given by the off-diagonal values (R matrix property)
$\lambda_{AA} = -(\lambda_{AC} + \lambda_{AG} + \lambda_{AT})$

$$\begin{array}{cc}
 & \begin{array}{cccc} A & C & G & T \end{array} \\
\begin{array}{c} A \\ C \\ G \\ T \end{array} & \left(\begin{array}{cccc}
* & \lambda_{AC} & \lambda_{AG} & \lambda_{AT} \\
 & * & \lambda_{CG} & \lambda_{CT} \\
 & & * & \lambda_{GT} \\
 & \text{Symmetric} & & *
\end{array}\right)
\end{array}$$

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 simple Jukes-Cantor model

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

Π = (1/4, 1/4, 1/4, 1/4)

The Felsenstein 81 model

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

$\Pi_i \neq \Pi_j$

Kimura 2-parameter model 1980

$$\begin{array}{ccccc} & A & C & G & T \\ A & * & \lambda & \zeta & \lambda \\ C & & * & \zeta & \lambda \\ G & & & * & \zeta \\ T & & & & * \end{array}$$

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

HKY85

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

$\Pi_i \neq \Pi_j$

GTR 1986

$$\begin{array}{cc} & \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{array}{c} A \\ C \\ G \\ T \end{array} & \begin{pmatrix} * & \alpha & \beta & \gamma \\ & * & \delta & \varepsilon \\ & & * & \zeta \\ & & & * \end{pmatrix} \end{array}$$

$\Pi_i \neq \Pi_j$

GTR 1986

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & * & \alpha & \beta & \gamma \\ C & & * & \delta & \varepsilon \\ G & & & * & \zeta \\ T & & & & * \end{array}$$

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

$\Pi_i \neq \Pi_j$

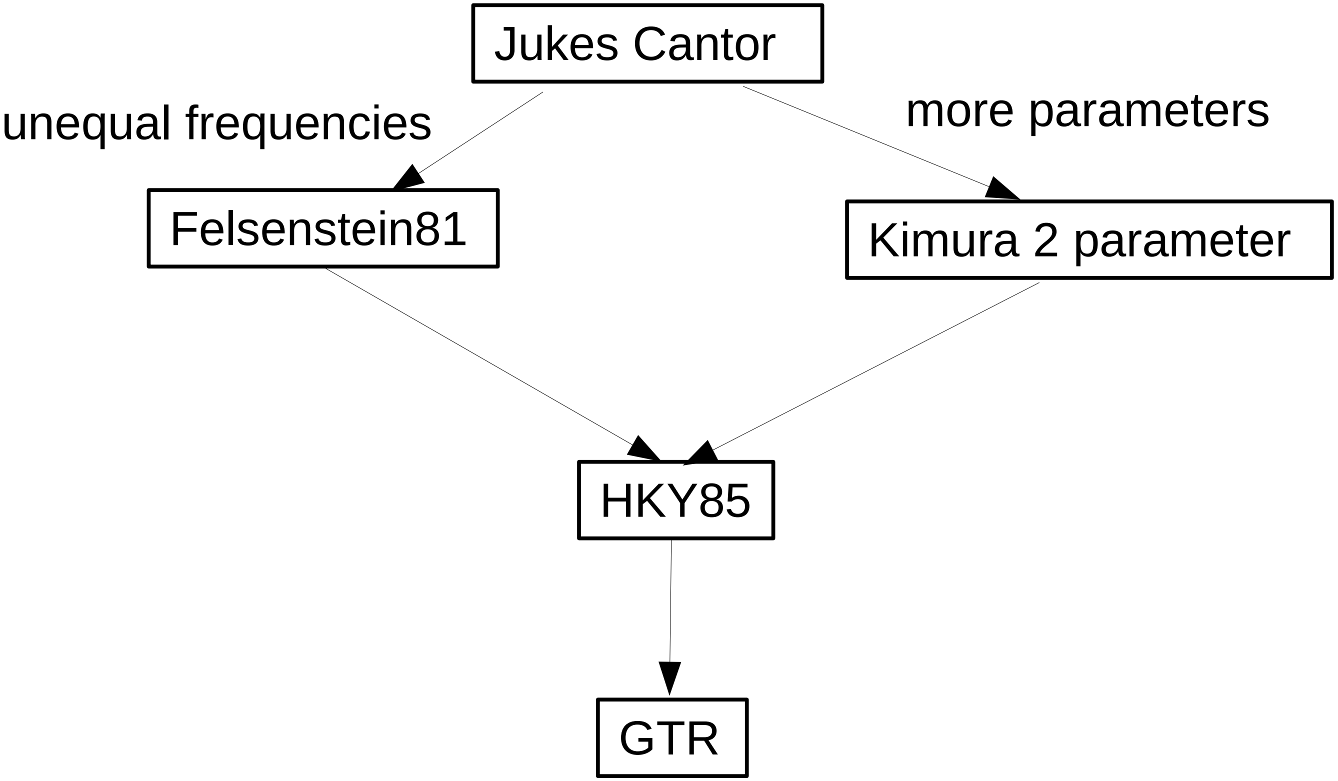
GTR 1986

$$\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & * & \alpha & \beta & \gamma \\
C & & * & \delta & \varepsilon \\
G & & & * & 1.0 \\
T & & & & *
\end{array}$$

Note that these are **relative** rates, their values only matter relative to each other, so we can set $\zeta := 1.0$ 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 $\boldsymbol{\pi_i r_{ij} = \pi_j r_{ji}}$

$$
\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & * & \alpha & \beta & \gamma \\
C & & * & \delta & \varepsilon \\
G & & & * & 1.0 \\
T & & & & *
\end{array}
$$

$\Pi_i \neq \Pi_j$

GTR 1986

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & * & \alpha & \beta & \gamma \\ C & & * & \delta & \varepsilon \\ G & & & * & 1.0 \\ T & & & & * \end{array}$$

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(π) R*

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

$\Pi_i \neq \Pi_j$

GTR 1986

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & * & \alpha & \beta & \gamma \\ C & & * & \delta & \varepsilon \\ G & & & * & 1.0 \\ T & & & & * \end{array}$$

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(π) 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:

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

- 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 Eigenvector/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:

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

- 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 Eigenvector/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!

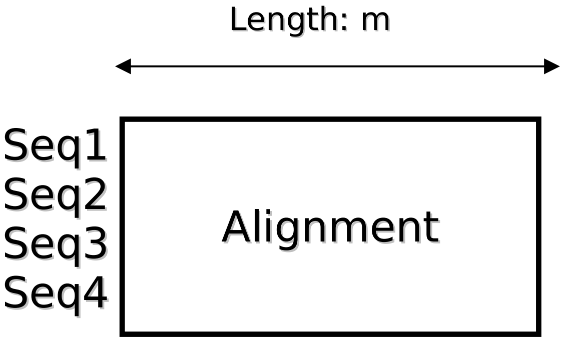
Likelihood Calculations

- 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
  1. functions for optimizing continuous parameters
  2. functions for searching the discrete space of trees

Outline – Lecture 10

- Maximum Likelihood – motivation
- Computing the Likelihood on a tree
- **Computing the Maximum Likelihood on a tree**

Maximum Likelihood

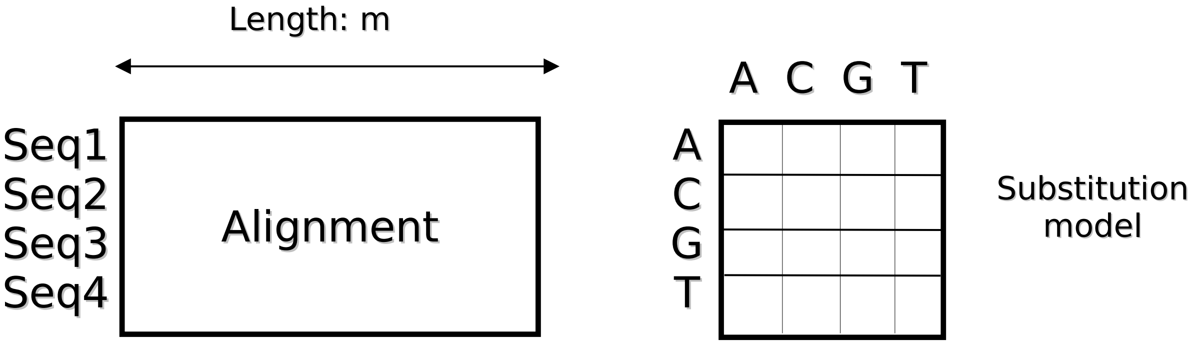


Length: m

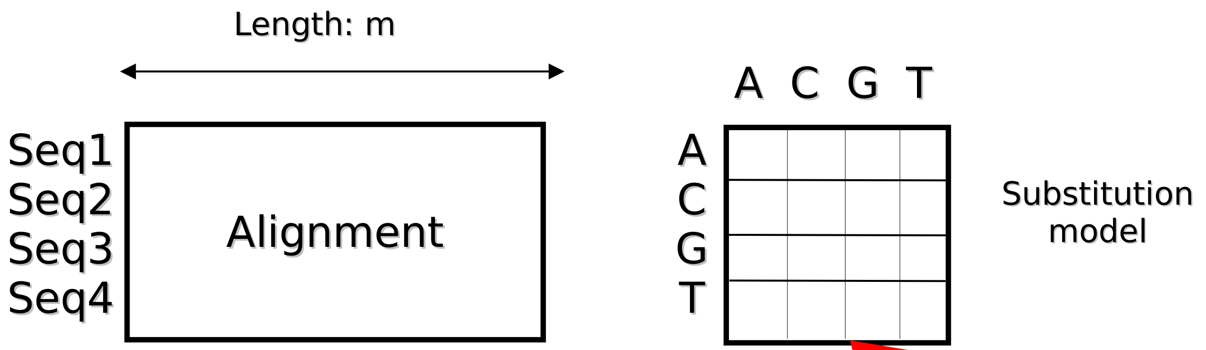
Seq1
Seq2
Seq3
Seq4

Alignment

Maximum Likelihood



Maximum Likelihood



Length: m

Seq1
Seq2
Seq3
Seq4

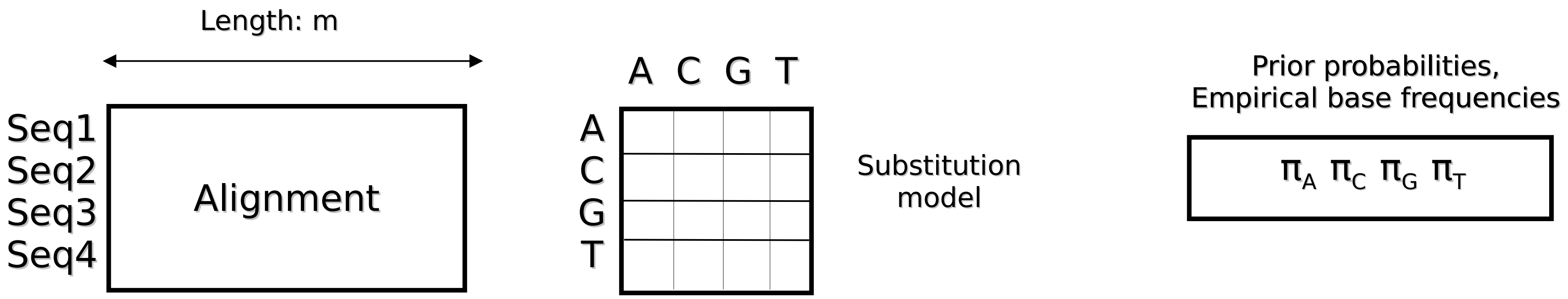
Alignment

	A	C	G	T
A				
C				
G				
T				

Substitution model

Commonly denoted as *Q* matrix: transition probs for time *dt*, for time *t*: $P(t)=e^{Qt}$

Maximum Likelihood



Length: m

Seq1
Seq2
Seq3
Seq4

Alignment

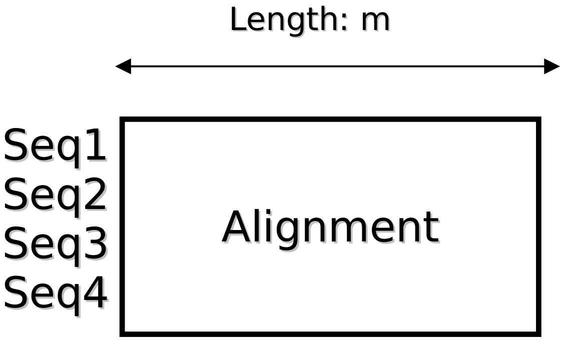
	A	C	G	T
A				
C				
G				
T				

Substitution model

Prior probabilities,
Empirical base frequencies

π_A π_C π_G π_T

Maximum Likelihood

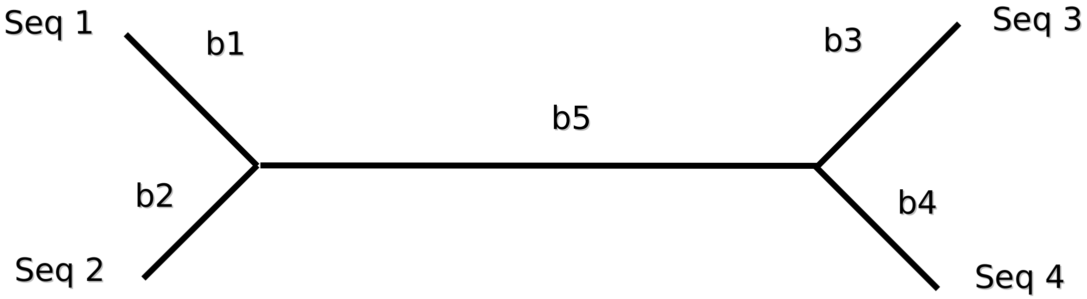


	A	C	G	T
A				
C				
G				
T				

Substitution model

Prior probabilities, Empirical base frequencies

π_A π_C π_G π_T



Maximum Likelihood

Length: m

Seq1
Seq2
Seq3
Seq4

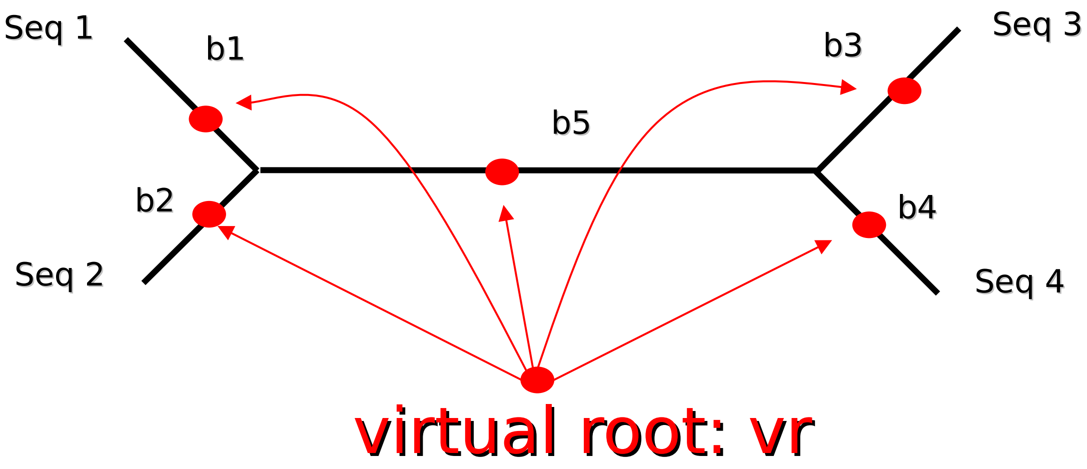
Alignment

	A	C	G	T
A				
C				
G				
T				

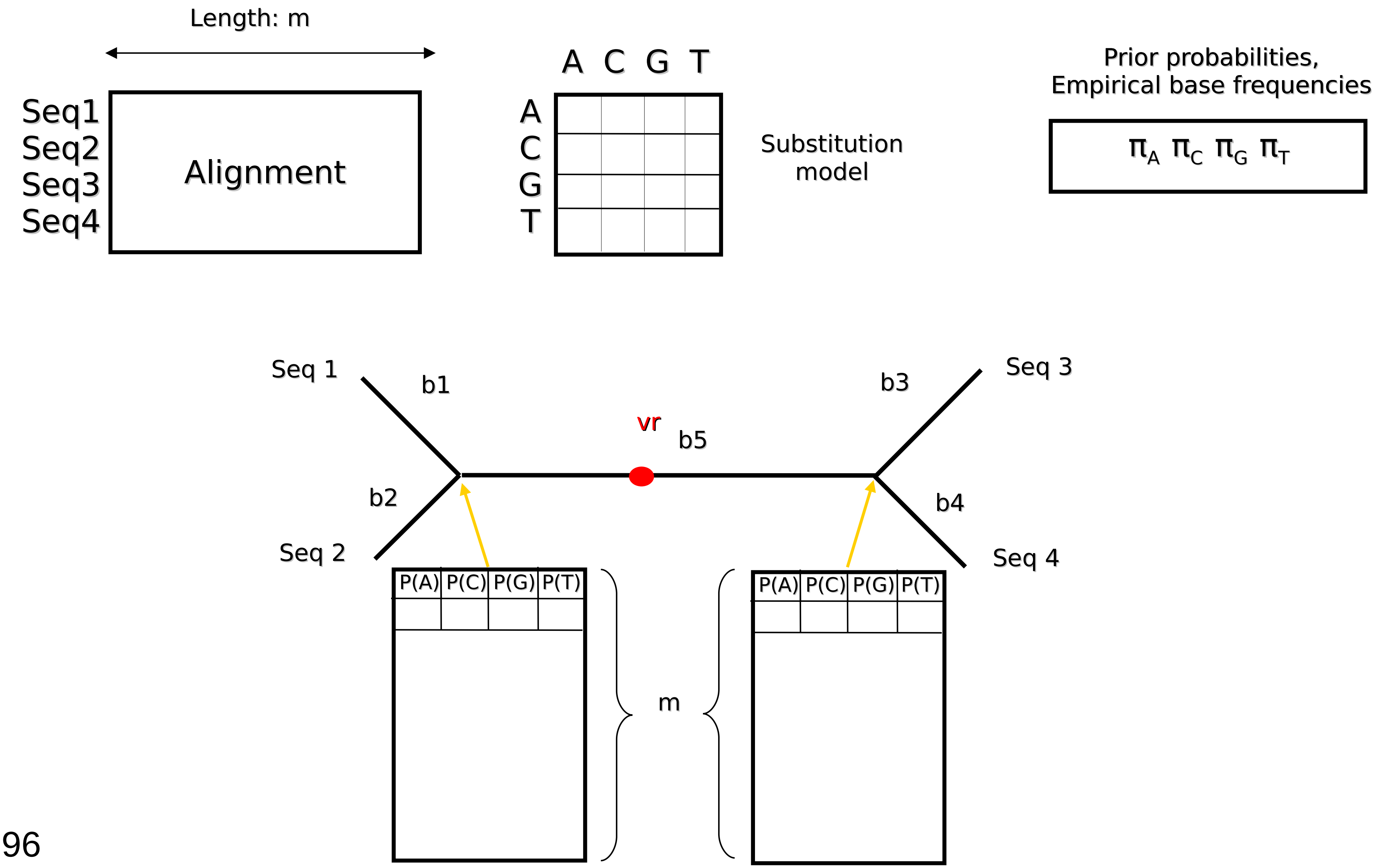
Substitution model

Prior probabilities, Empirical base frequencies

$\pi_A \; \pi_C \; \pi_G \; \pi_T$



Maximum Likelihood



Length: m

Seq1
Seq2
Seq3
Seq4

Alignment

	A	C	G	T
A				
C				
G				
T				

Substitution model

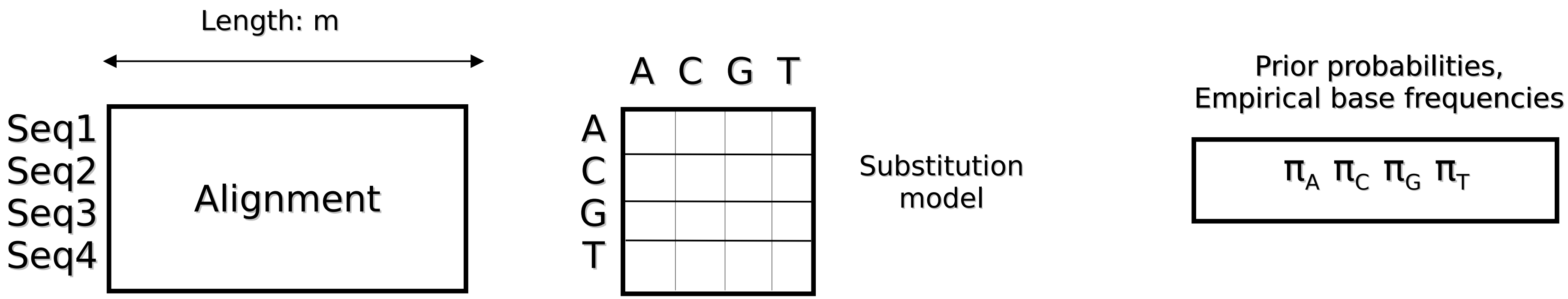
Prior probabilities,
Empirical base frequencies

$\pi_A\ \pi_C\ \pi_G\ \pi_T$

P(A)	P(C)	P(G)	P(T)

P(A)	P(C)	P(G)	P(T)

Maximum Likelihood



Length: m

Seq1
Seq2
Seq3
Seq4

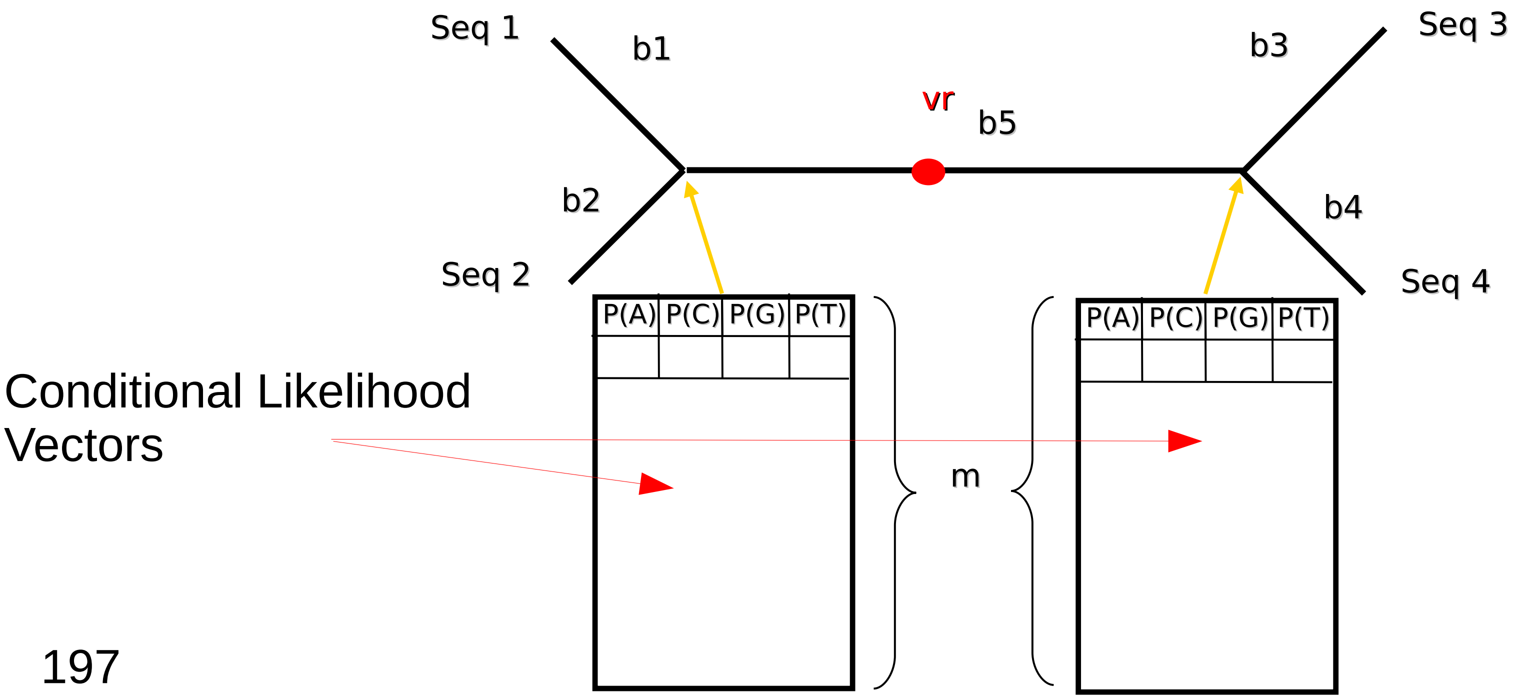
Alignment

	A	C	G	T
A				
C				
G				
T				

Substitution model

Prior probabilities, Empirical base frequencies

π_A π_C π_G π_T



Seq 1
b1
Seq 3
b3
vr
b5
b2
b4
Seq 2
Seq 4

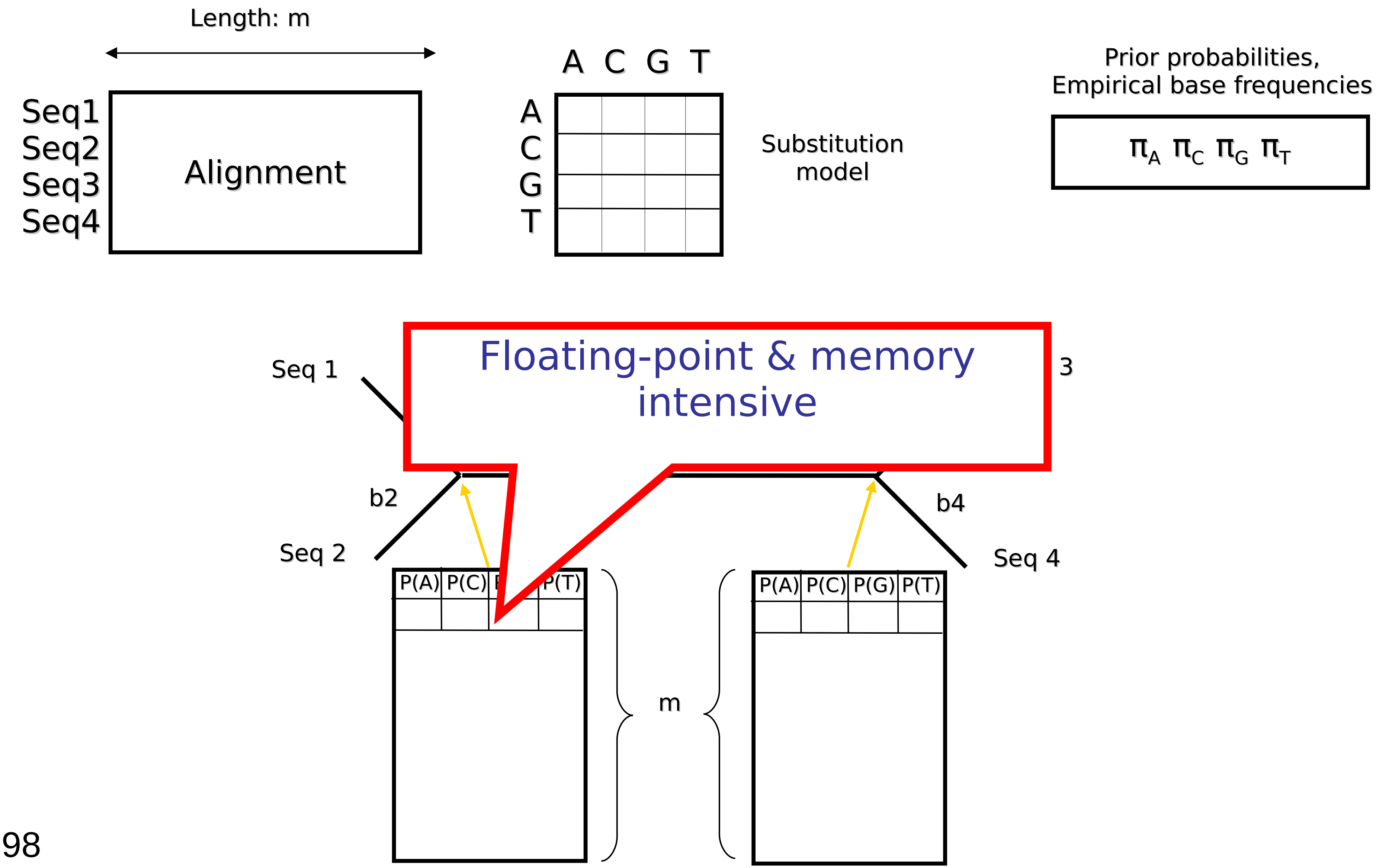
P(A)	P(C)	P(G)	P(T)

m

P(A)	P(C)	P(G)	P(T)

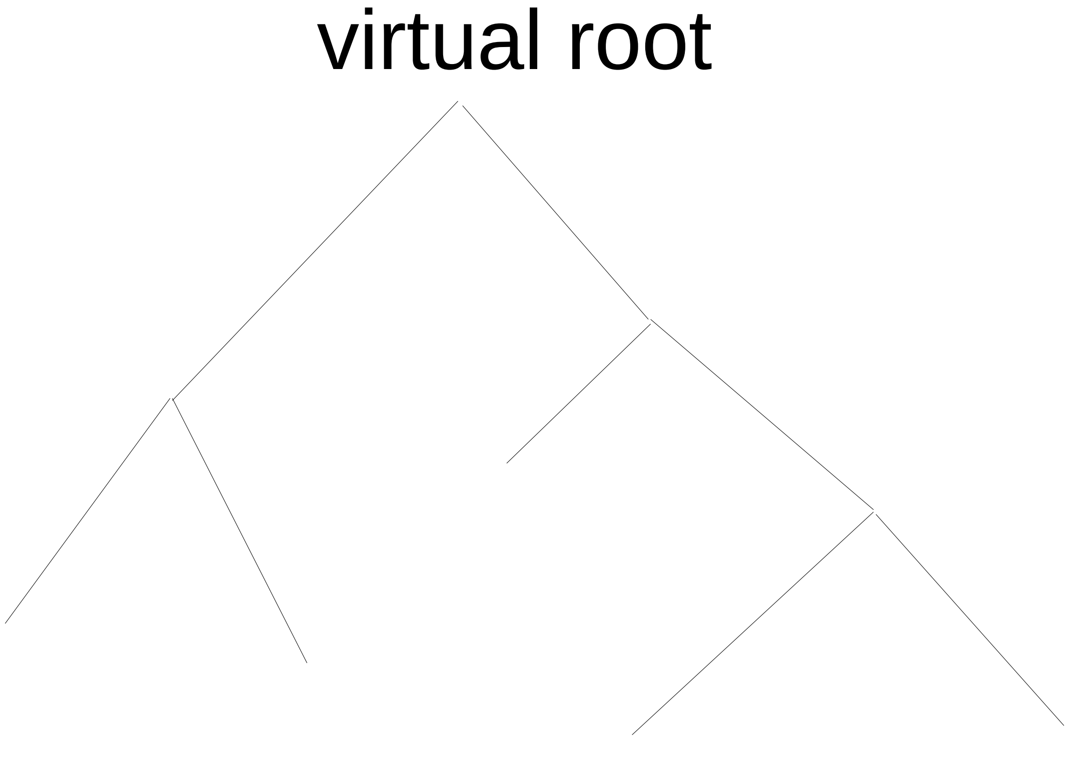
Conditional Likelihood Vectors

Maximum Likelihood



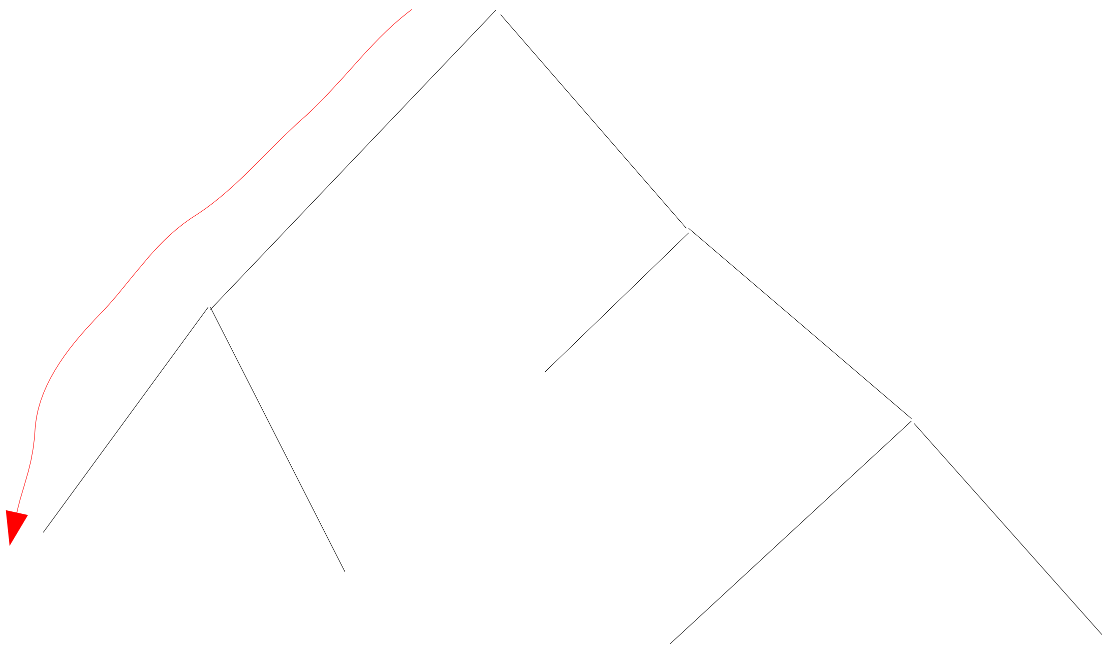
Floating-point & memory intensive

Post-order Traversal

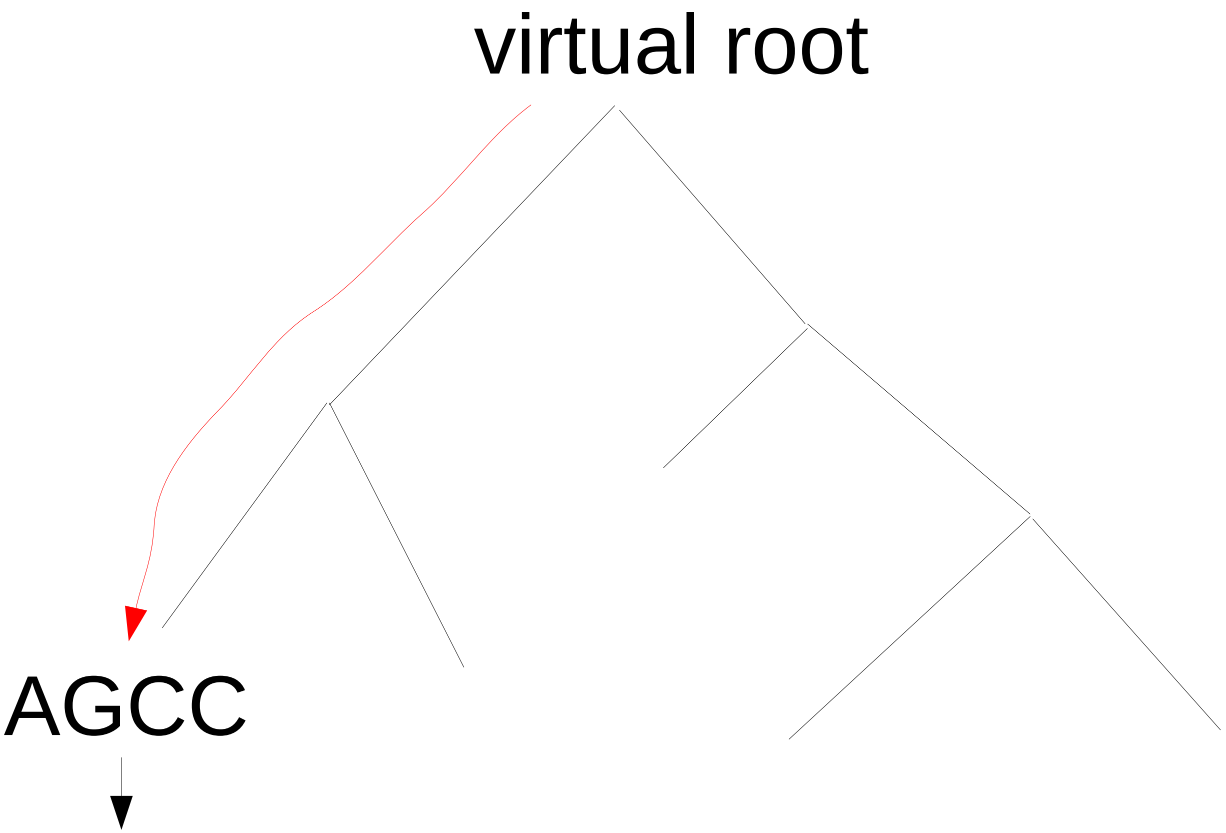


Post-order Traversal

virtual root



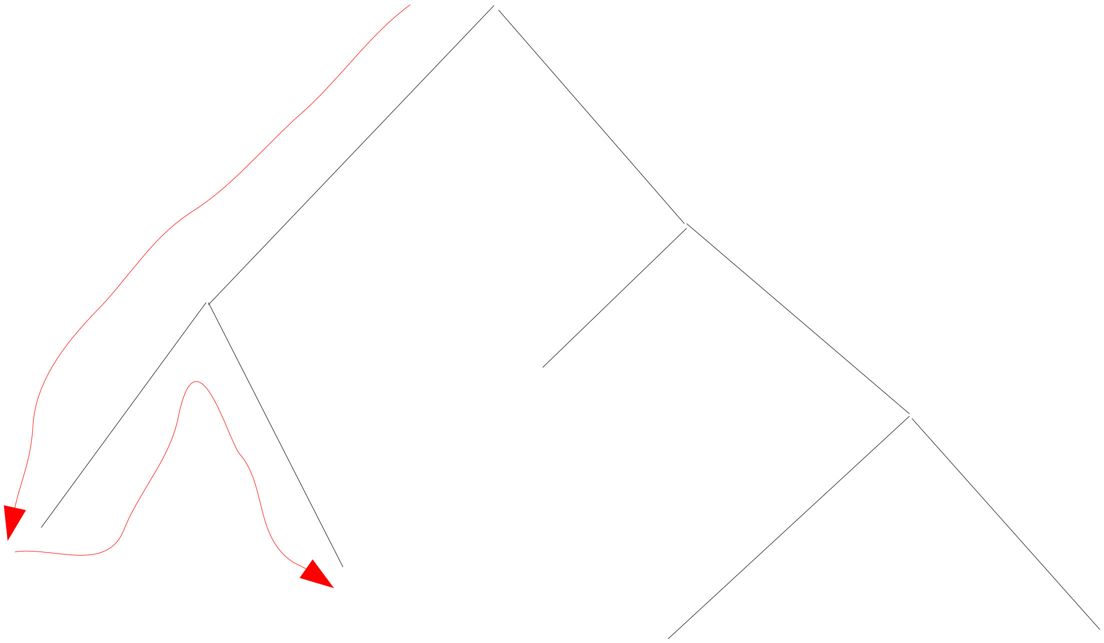
Post-order Traversal



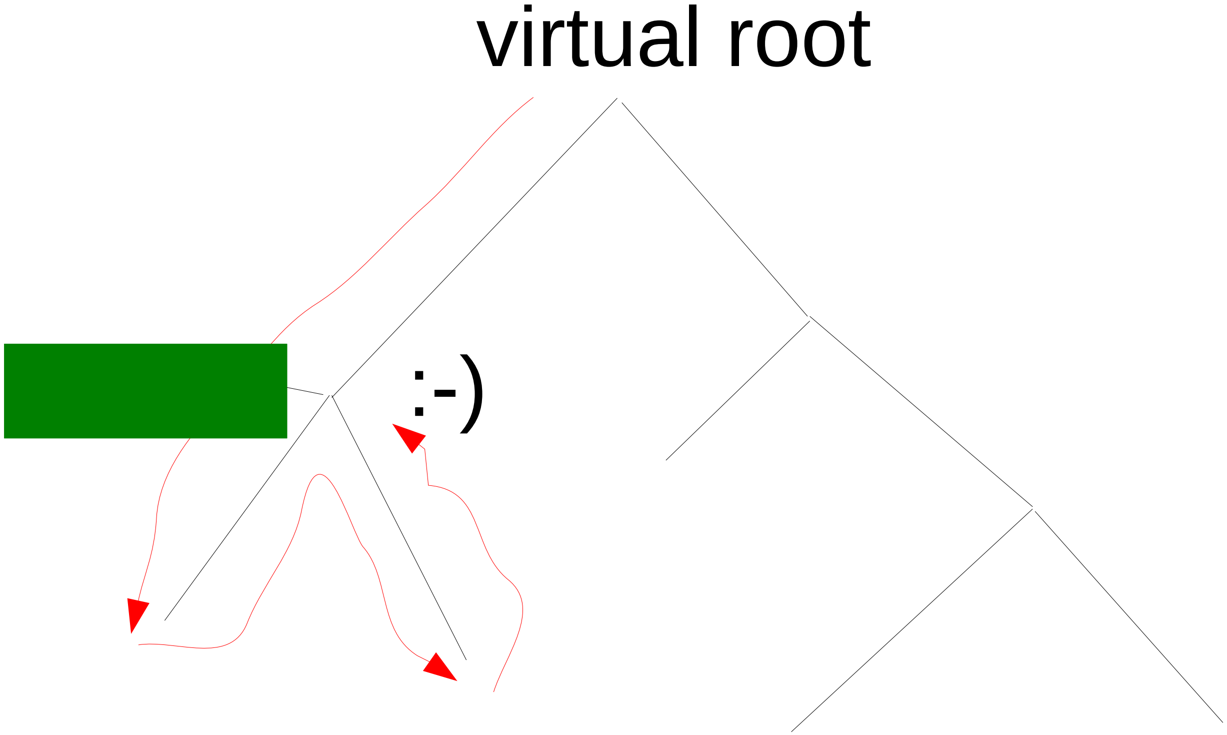
A	G	C	C
1.0	0.0	0.0	0.0
0.0	0.0	**1.0**	**1.0**
0.0	**1.0**	0.0	0.0
0.0	0.0	0.0	0.0

Post-order Traversal

virtual root

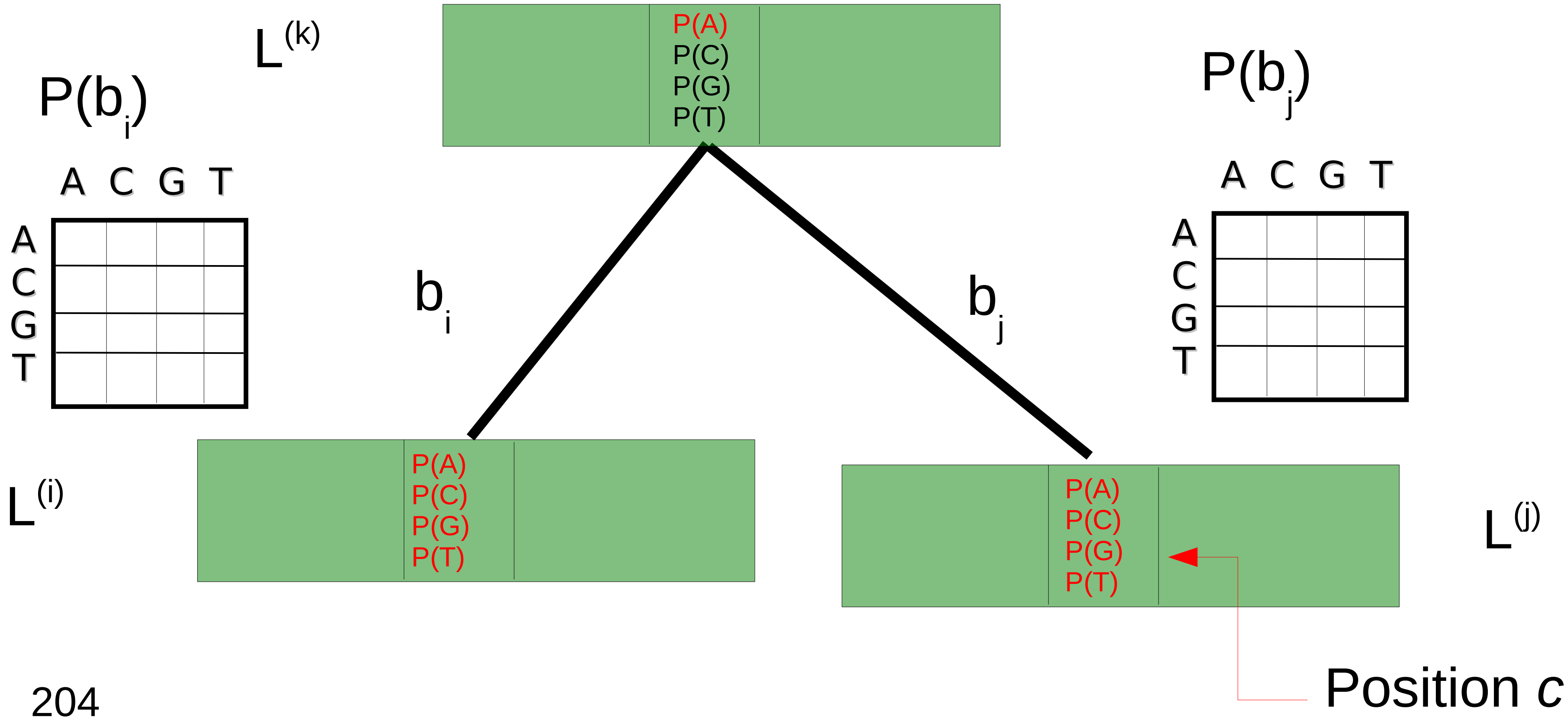


Post-order Traversal

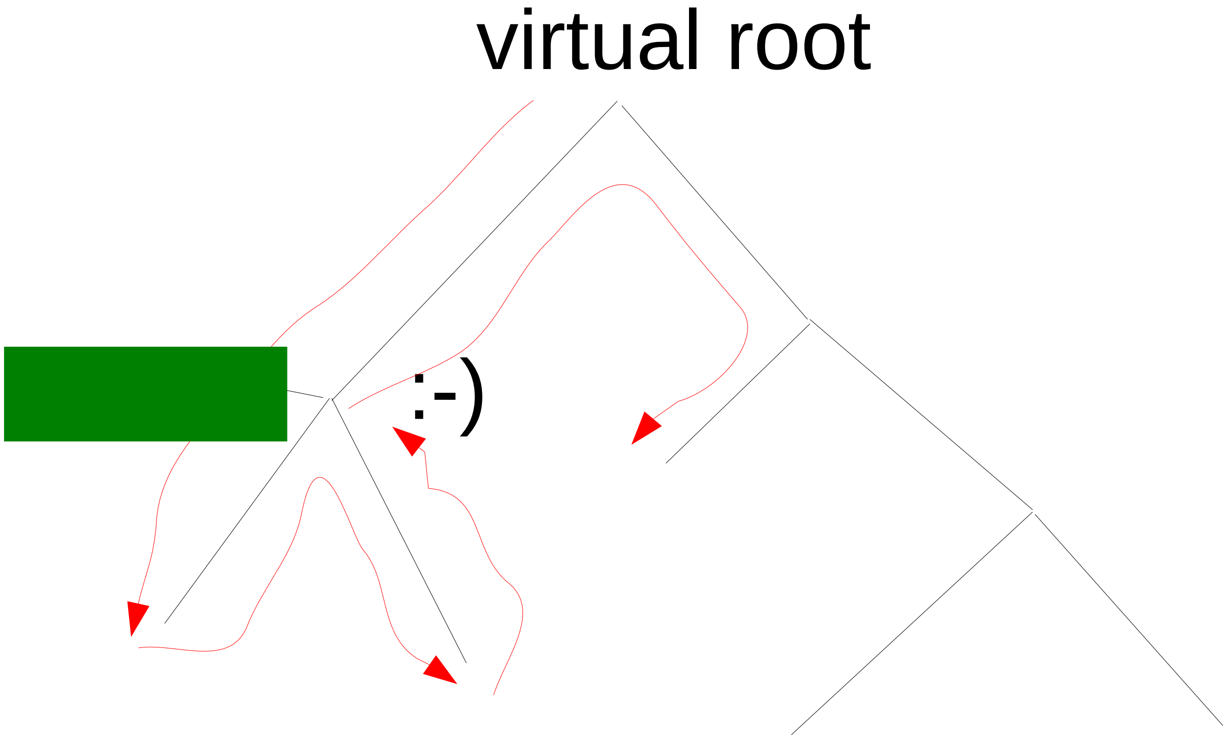


What happens when we compute this inner vector?

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



Post-order Traversal

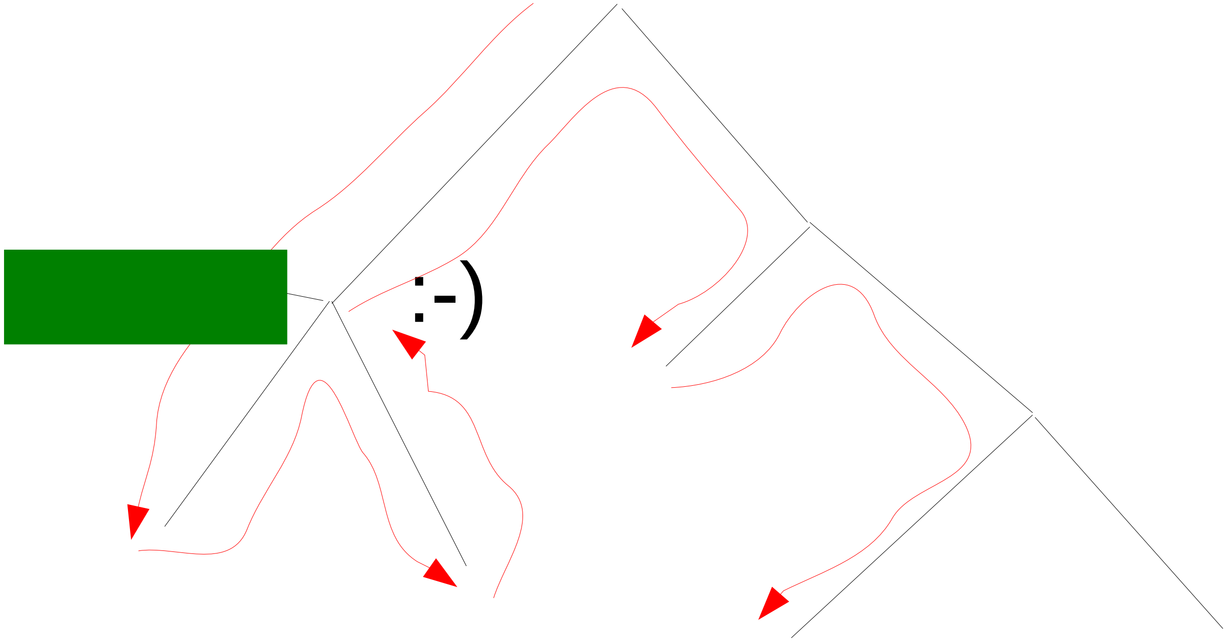


virtual root

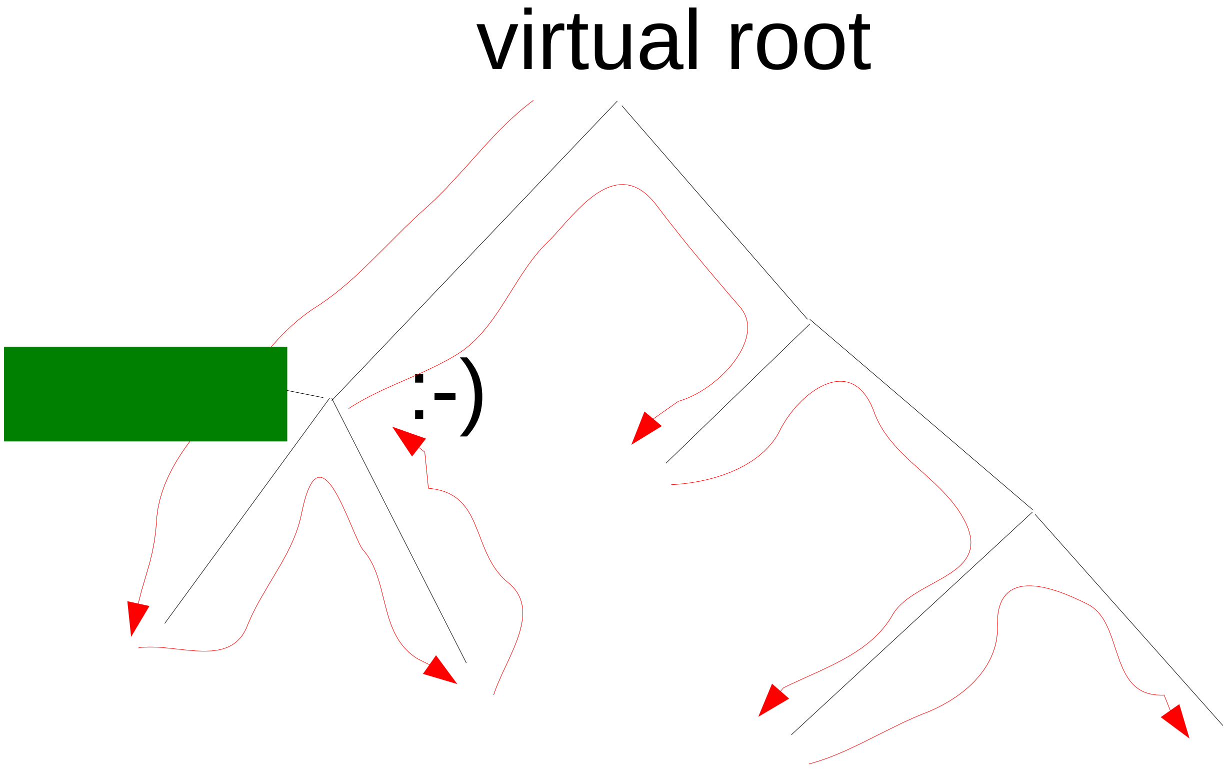
:-)

Post-order Traversal

virtual root



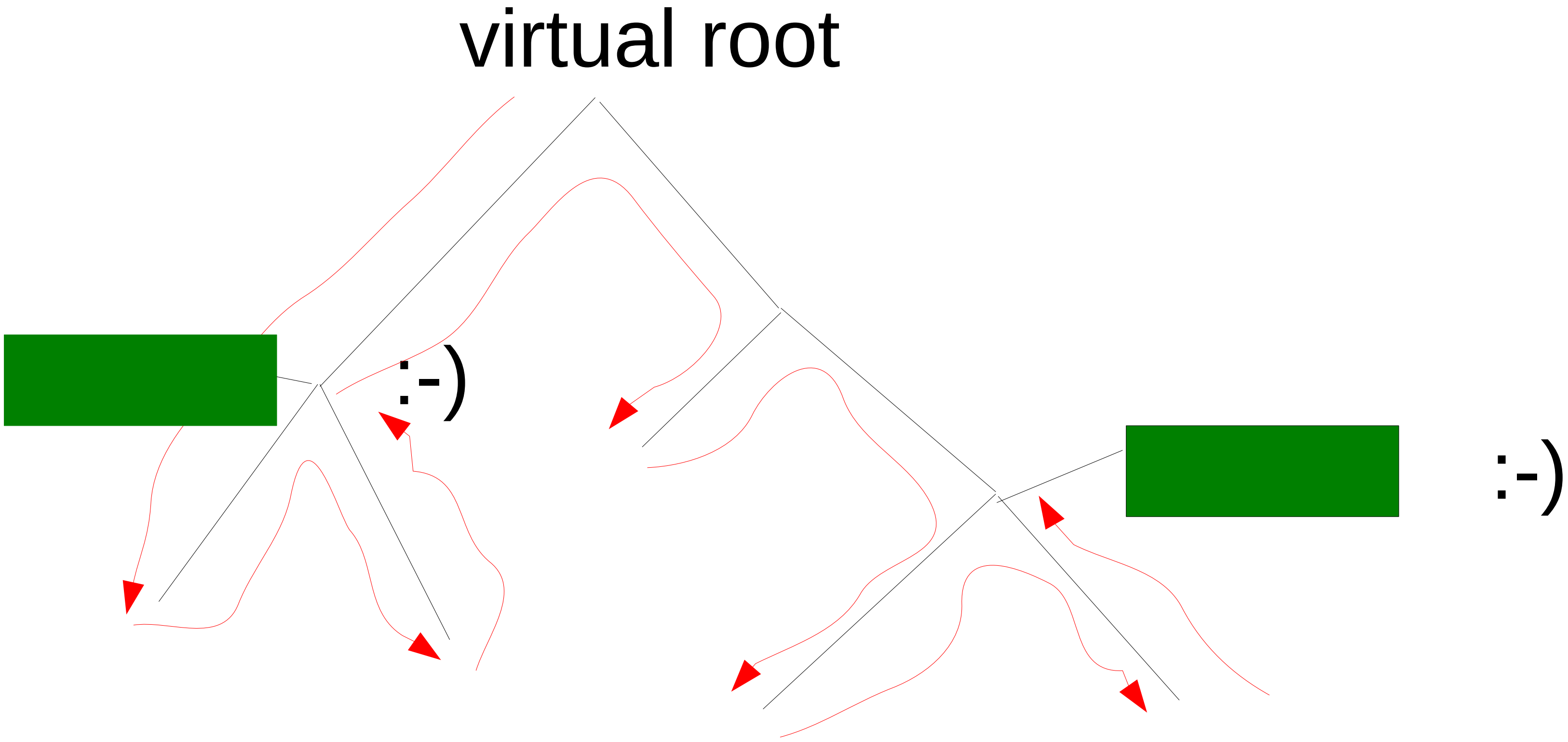
Post-order Traversal



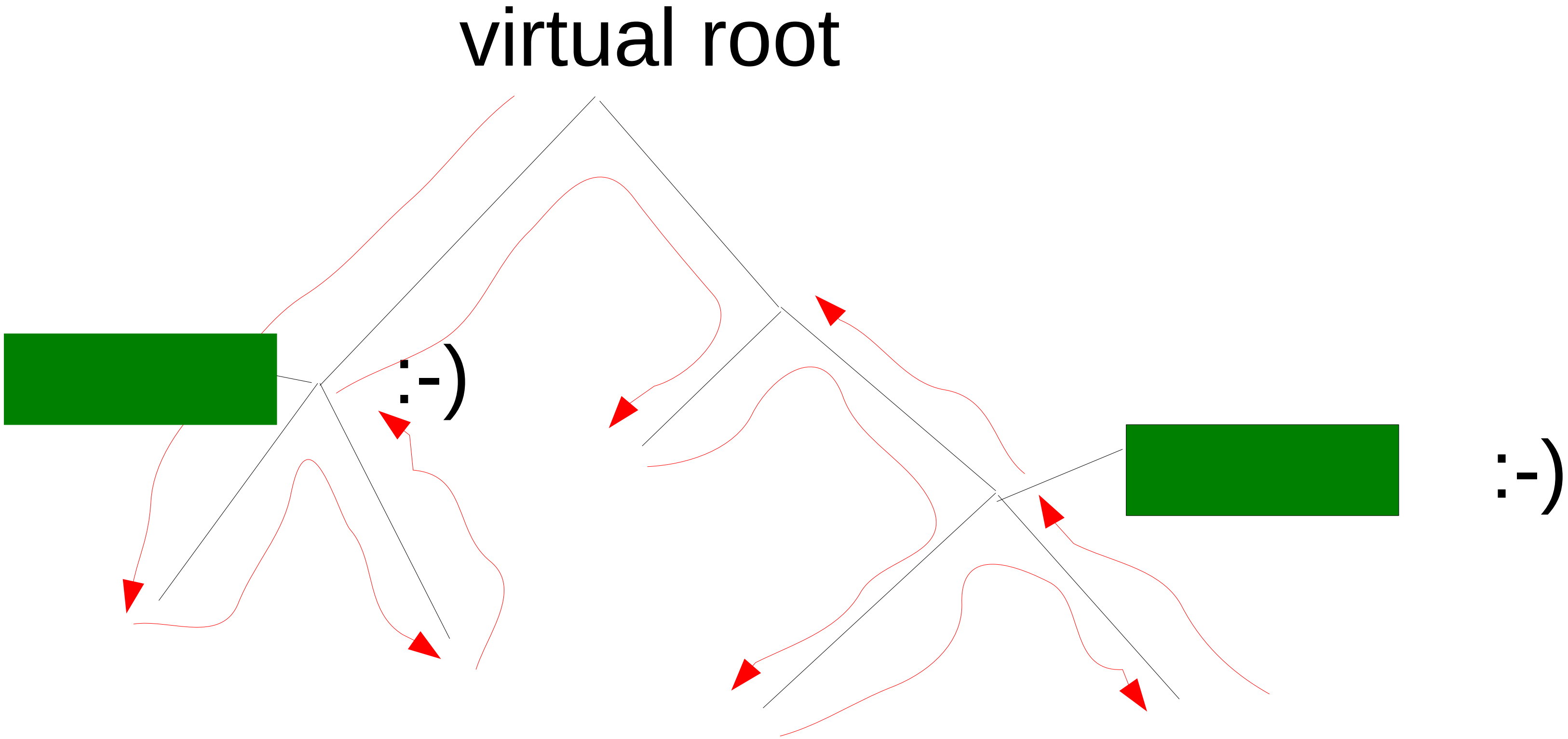
virtual root

:-)

Post-order Traversal

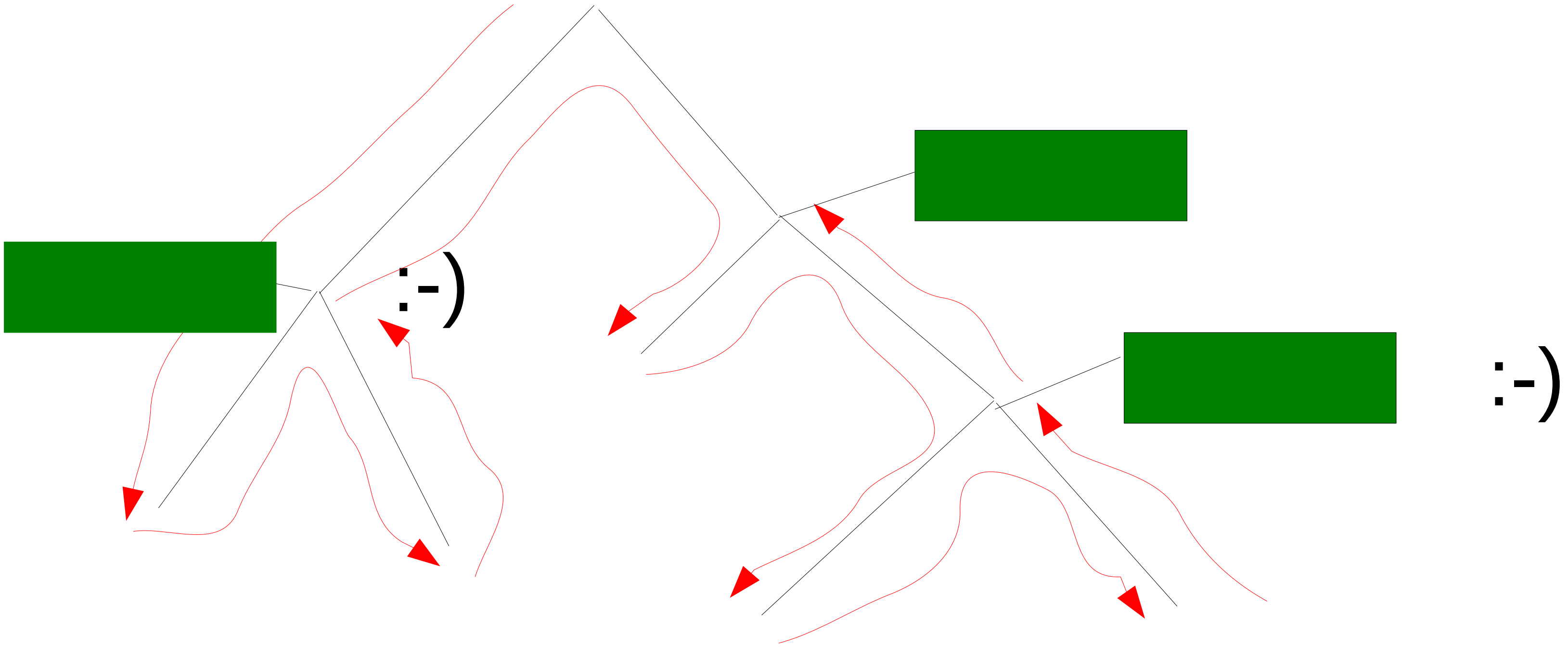


Post-order Traversal

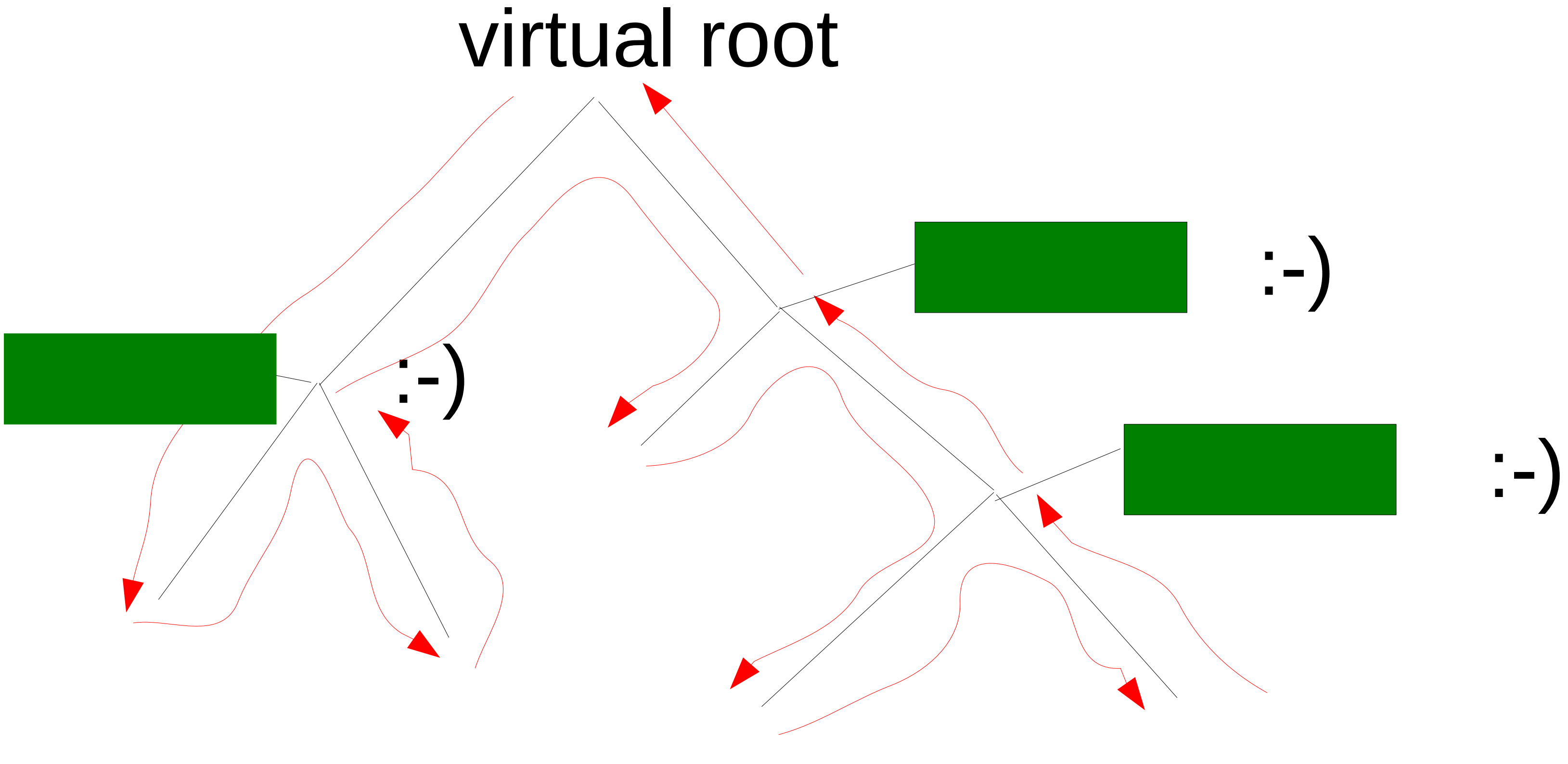


Post-order Traversal

virtual root

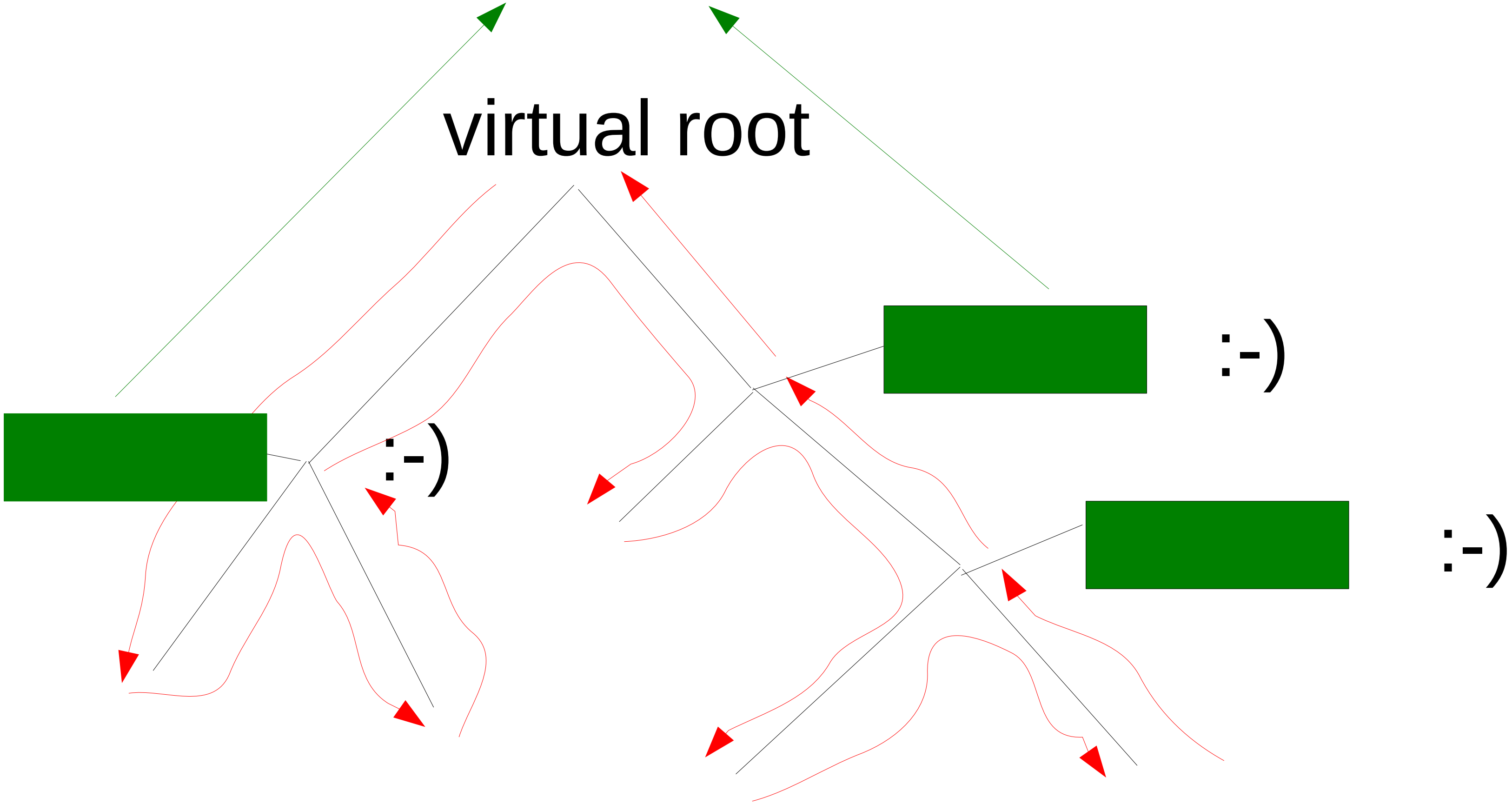


Post-order Traversal



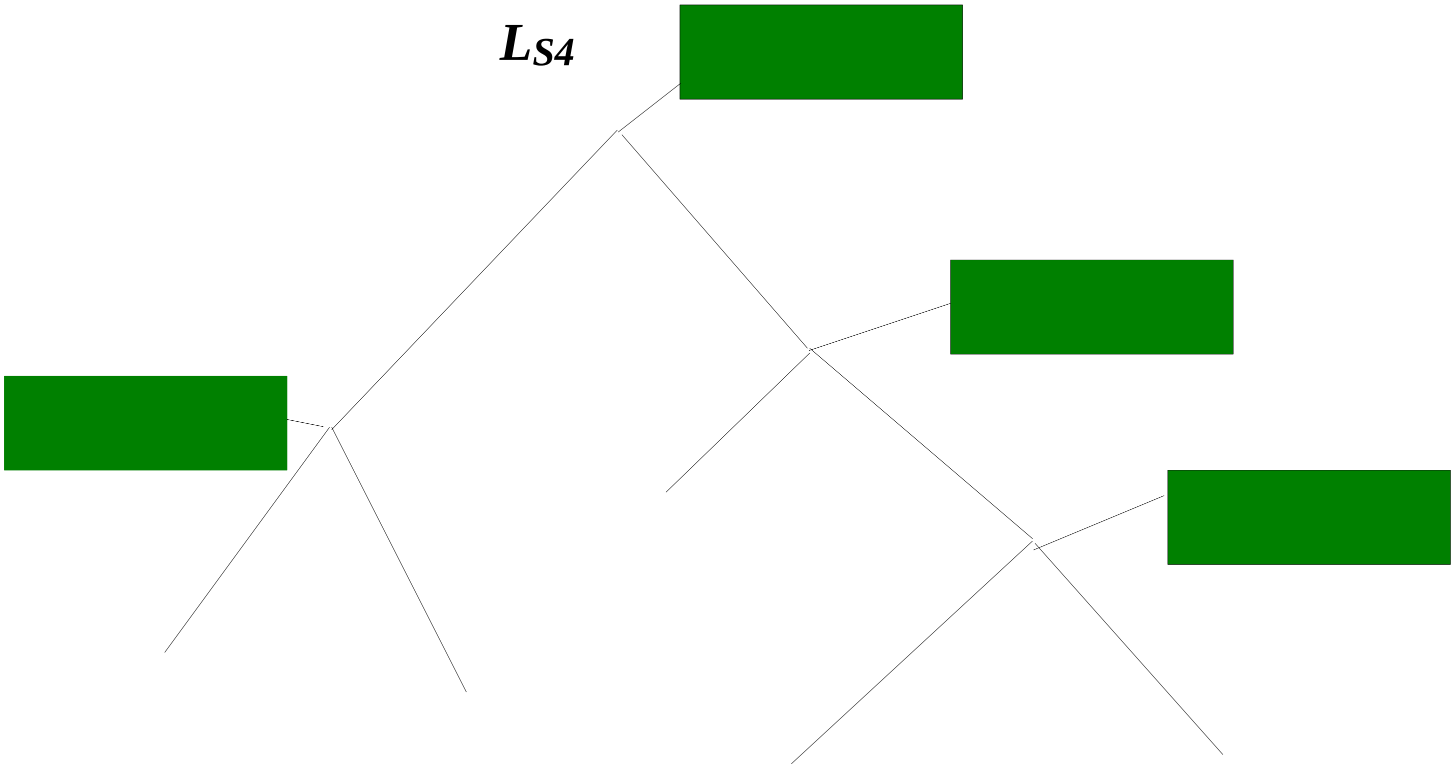
Post-order Traversal

Overall likelihood: sum over logarithms of per-site likelihoods

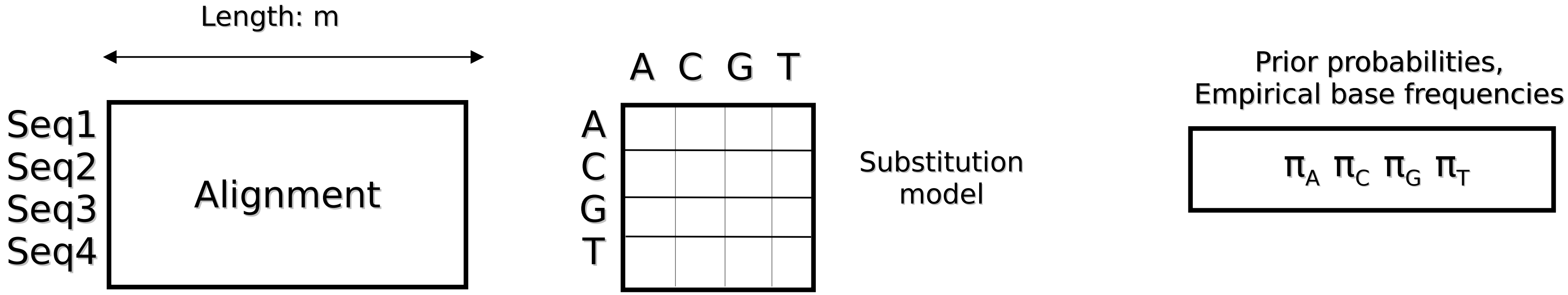


Post-order Traversal

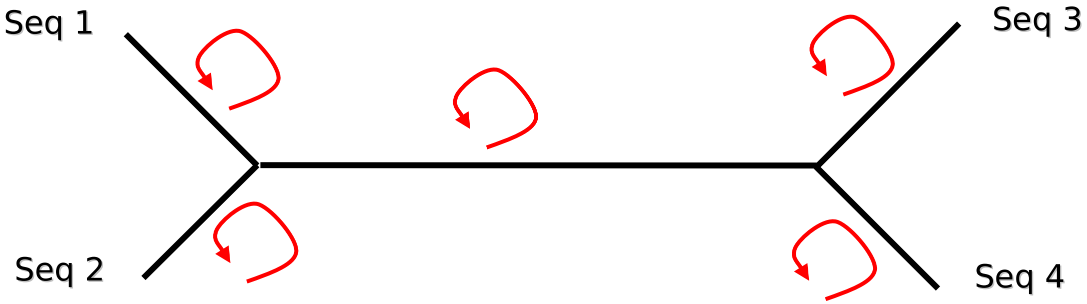
$$L = \sum_{S_4=A}^{T} \pi_{S_4} L_{S_4}$$



Maximum Likelihood

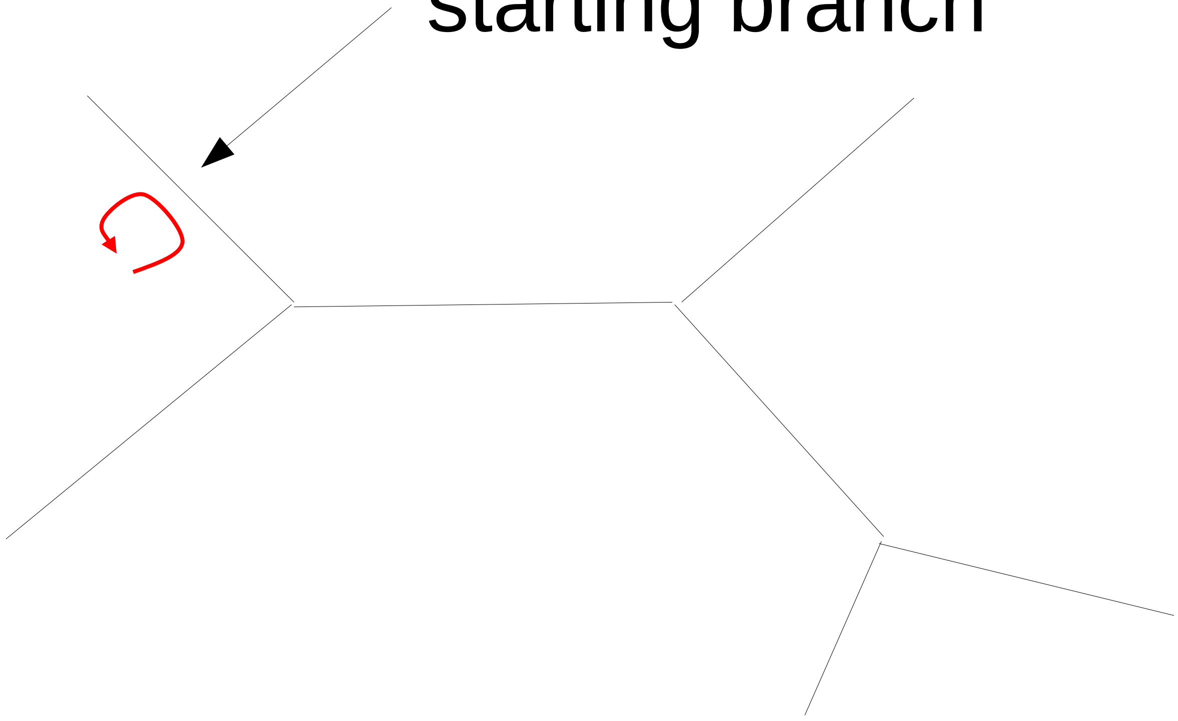


π_A π_C π_G π_T



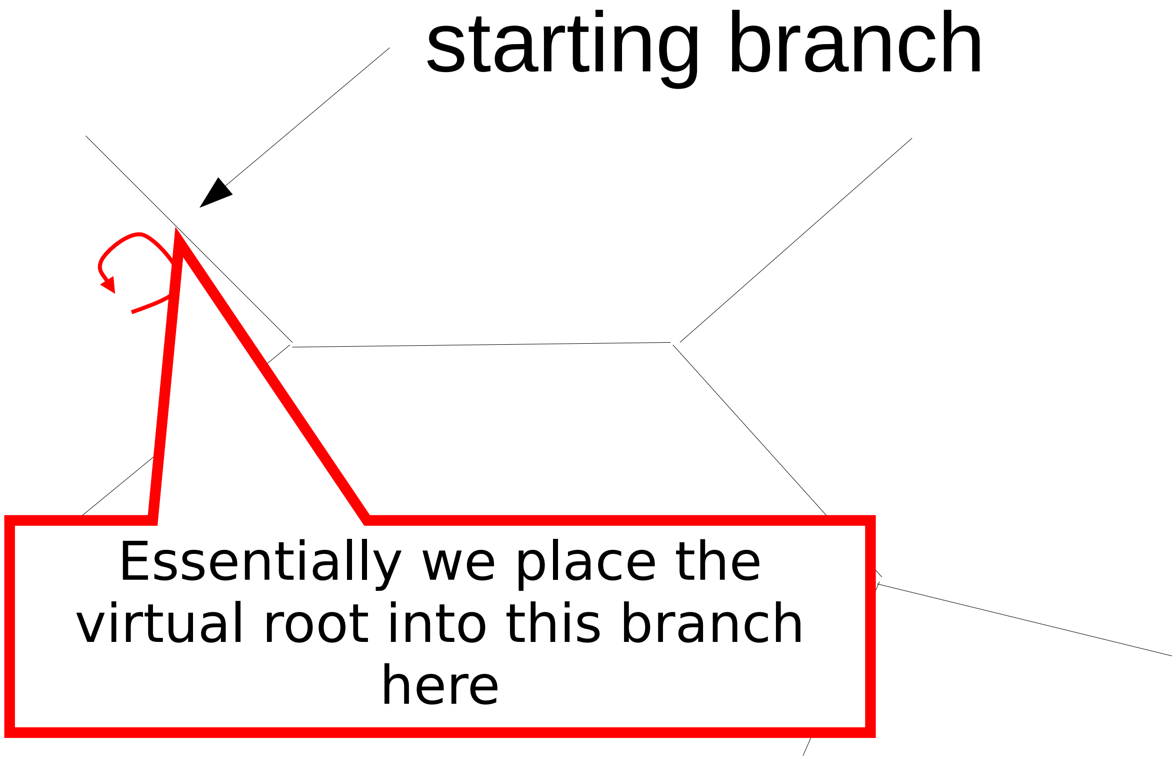
optimize branch lengths

Branch Length Optimization



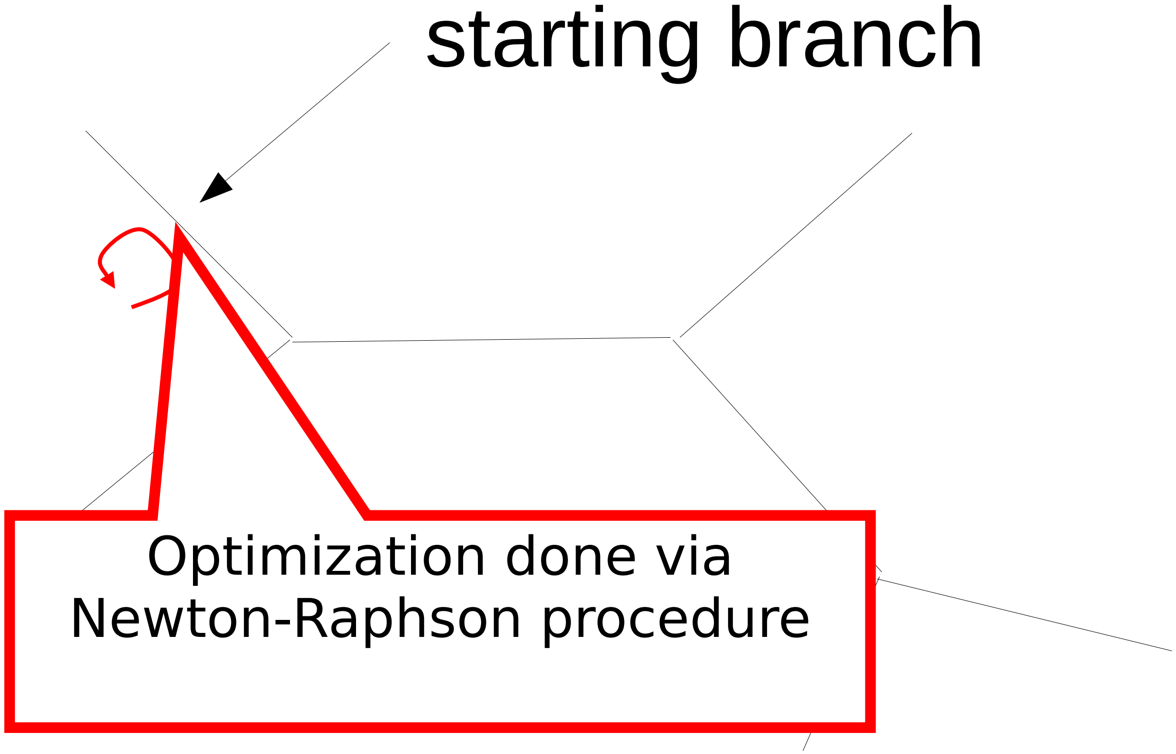
starting branch

Branch Length Optimization



Essentially we place the virtual root into this branch here

Branch Length Optimization



starting branch

Optimization done via
Newton-Raphson procedure

Newton Rapshon

- We want to find the branch length *b* that maximizes the likelihood *L(b)* of the tree
- For this, we want to know where the *first* derivative of *L(b)* is *0*
- To achieve this numerically we use the Newton-Raphson procedure for root finding deploying the first and second derivative of the likelihood *L'(b)* and *L''(b)*
- Note that, the likelihood only depends on branch *b*, all other model parameters (*Q* matrix, base frequencies, tree topology) remain fixed

Derivatives of *L(b)*

- To compute the derivatives of *L(b)*, we essentially need to be able to compute the derivatives of *P(b)* since the rest is just sums and does not depend on *b*
- Recall

$P(b) = e^{Qb} = Ue^{\Lambda b}U^{-1}$

- thus

$(P(b))' = U\Lambda e^{\Lambda b}U^{-1}$

- and

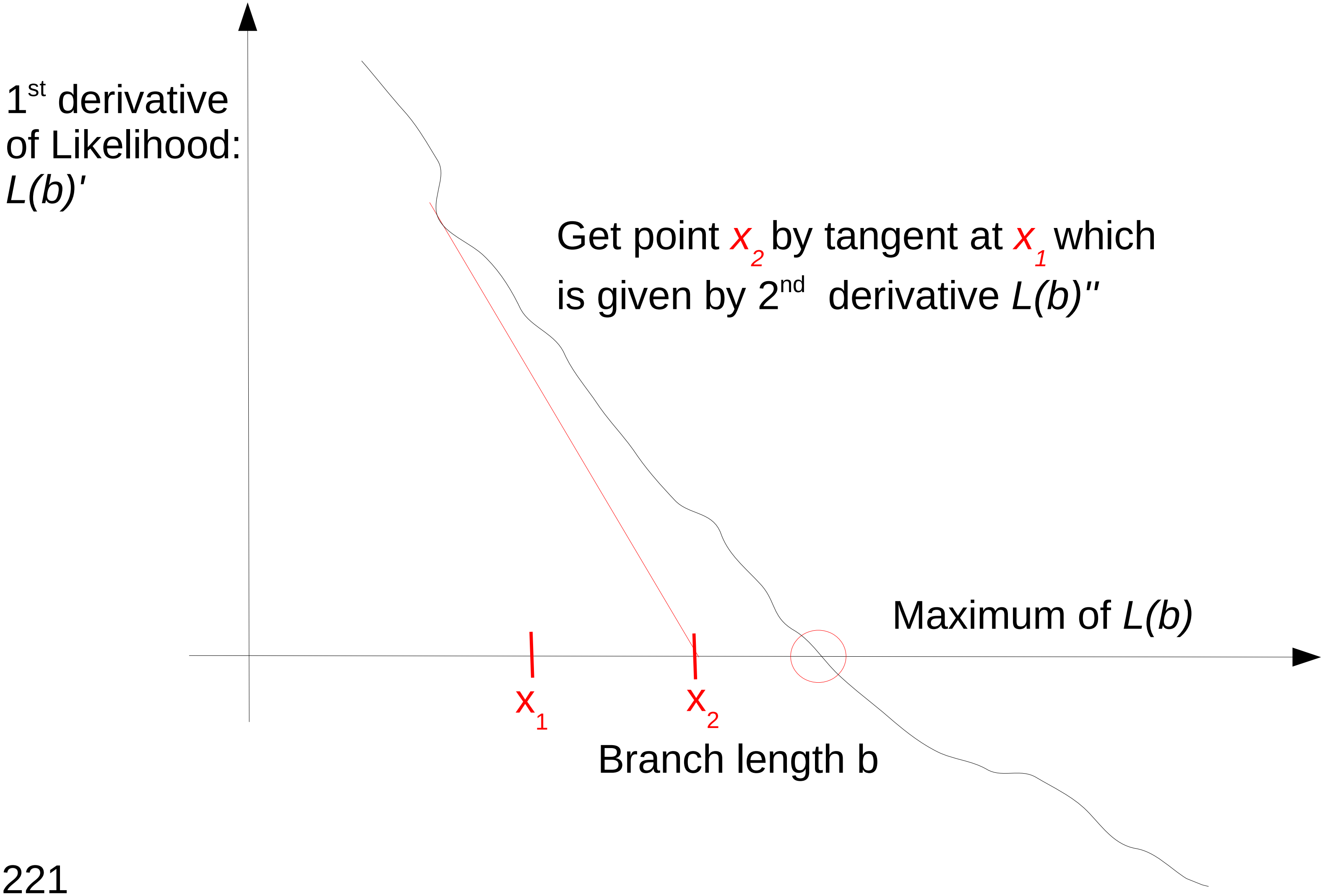
$(P(b))'' = U\Lambda^{2} e^{\Lambda b}U^{-1}$

- In practice we compute the derivatives of the log likelihood *log(L(b))*, but it is essentially the same (see next slide)

Derivatives of *log(L(b))*

- 1st derivative: $L(b)'/L(b)$
- *2nd derivative:* $(L(b)\ L(b)'' - (L(b)'')^2\ )\ /\ L(b)^2$

Newton Raphson



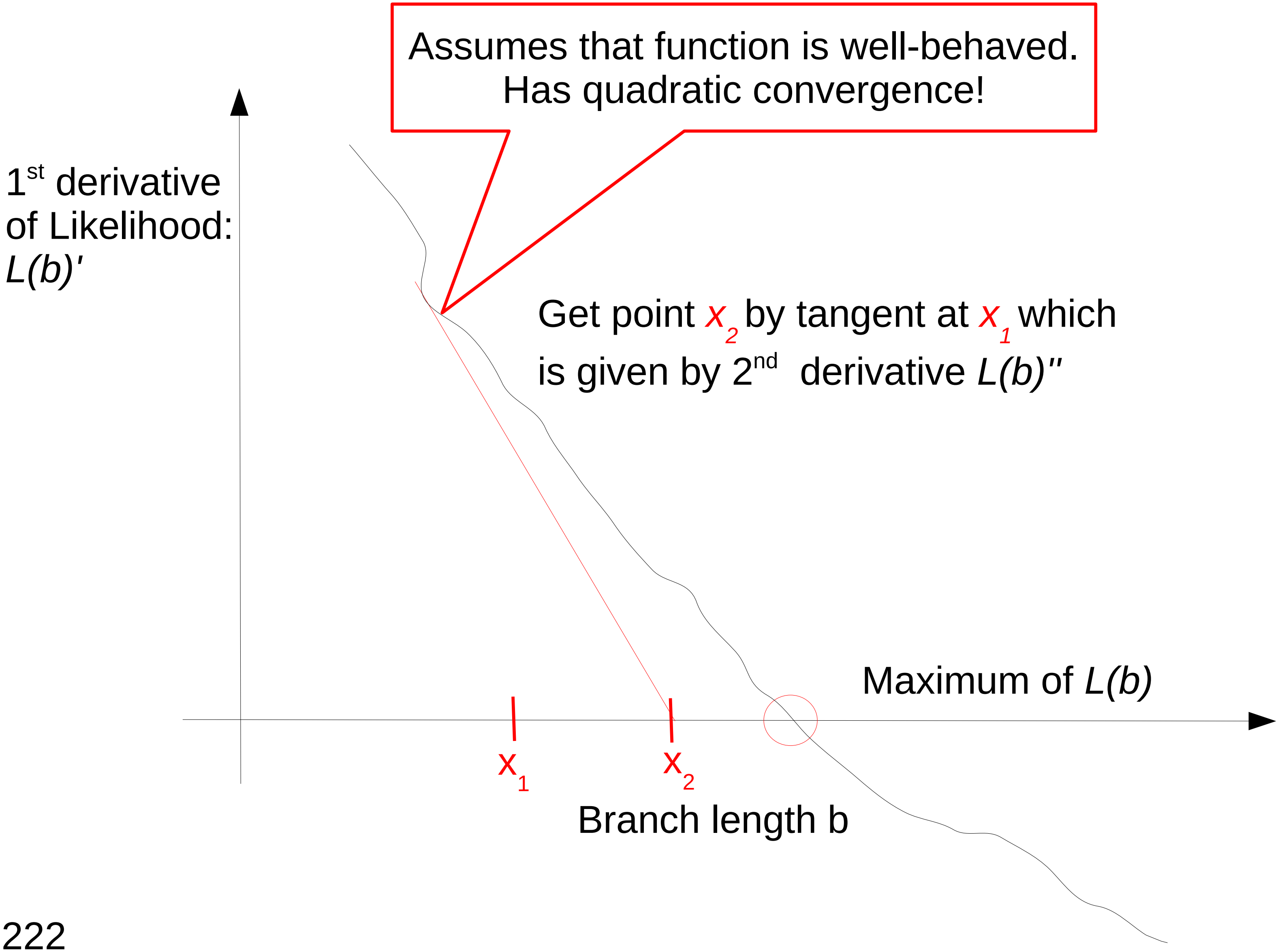
1st derivative of Likelihood: *L(b)'*

Get point *x_2* by tangent at *x_1* which is given by 2nd derivative *L(b)''*

Maximum of *L(b)*

Branch length b

Newton Raphson



Assumes that function is well-behaved.
Has quadratic convergence!

1^{st} derivative of Likelihood: *L(b)'*

Get point x_2 by tangent at x_1 which is given by 2^{nd} derivative *L(b)''*

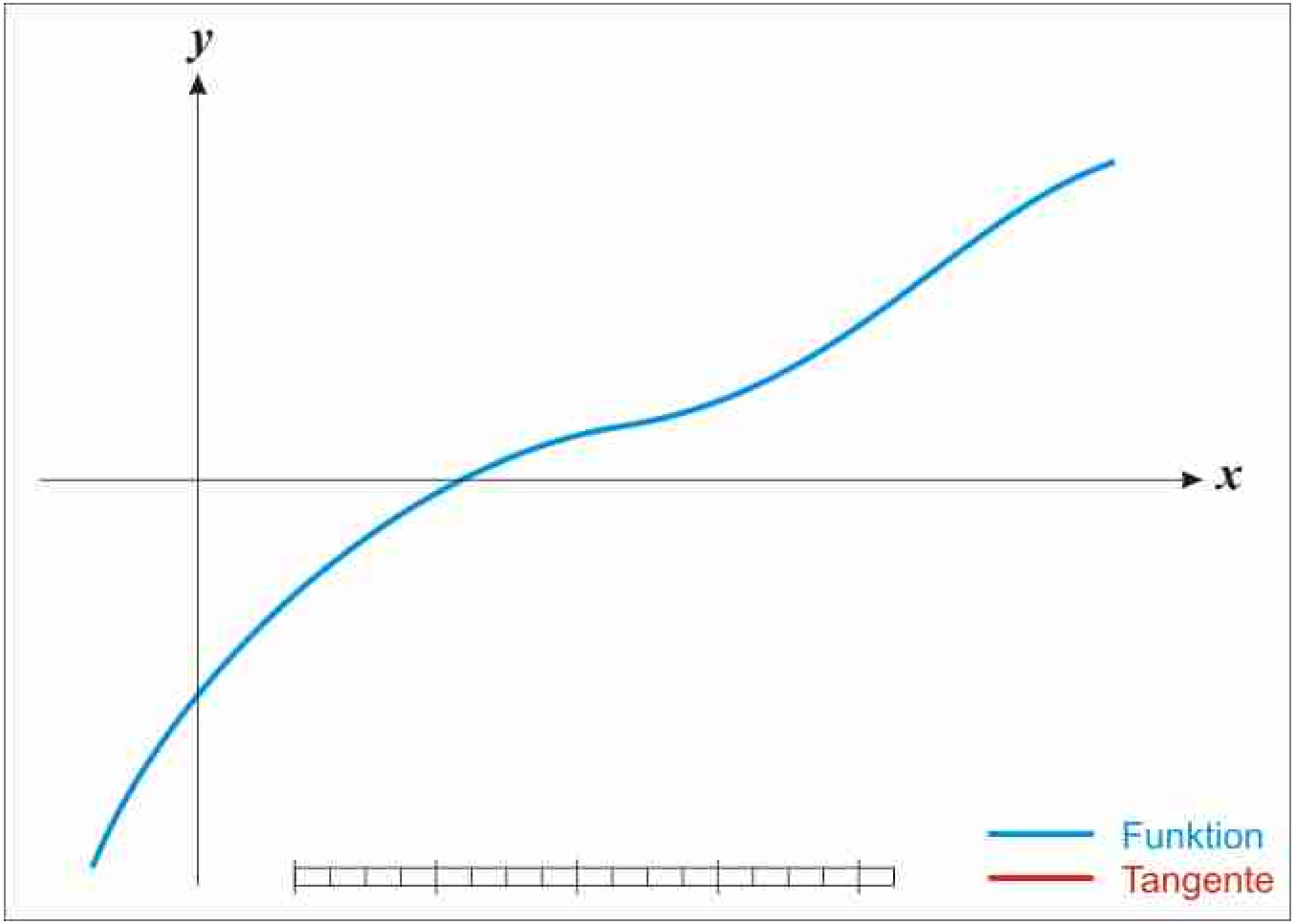
Maximum of *L(b)*

x_1

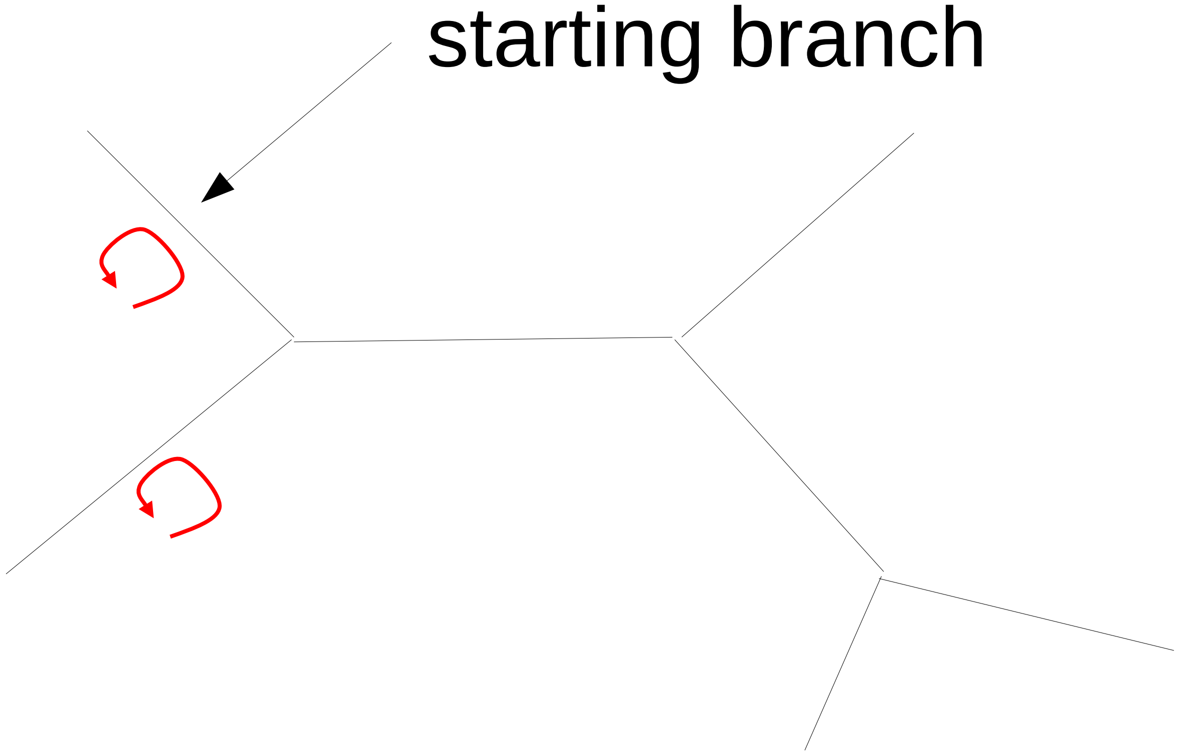
x_2

Branch length b

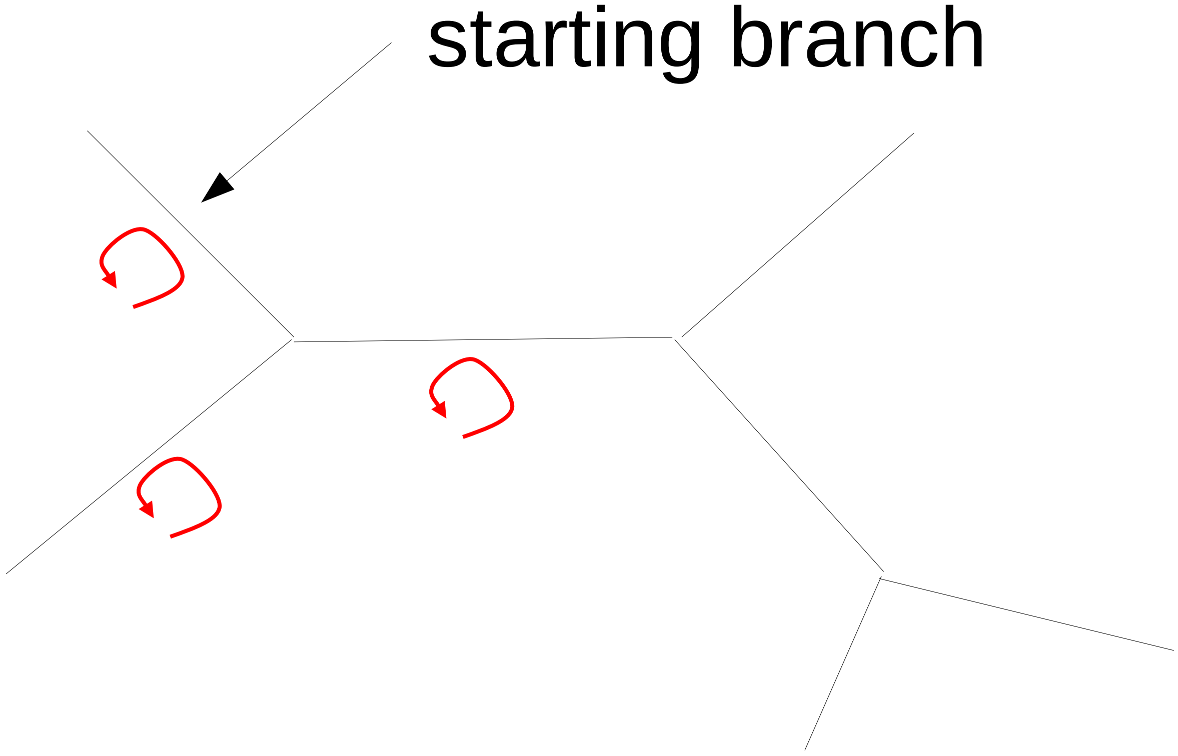
An animation



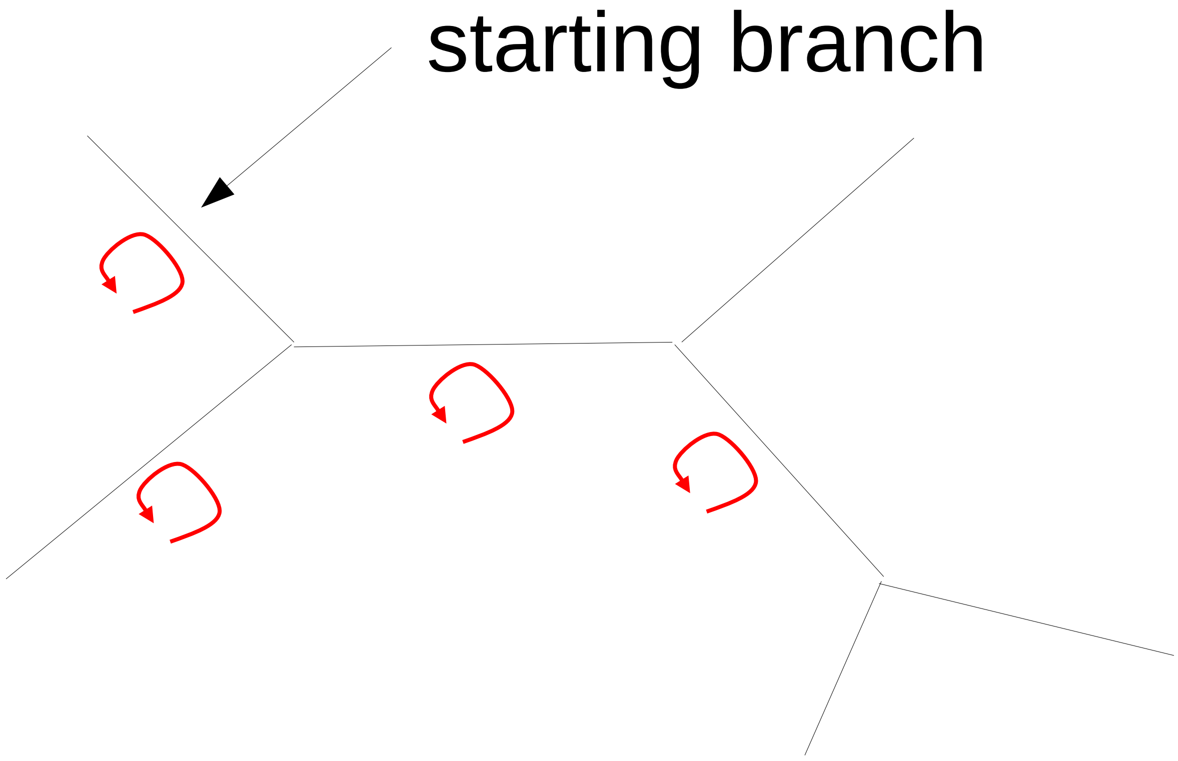
Branch Length Optimization



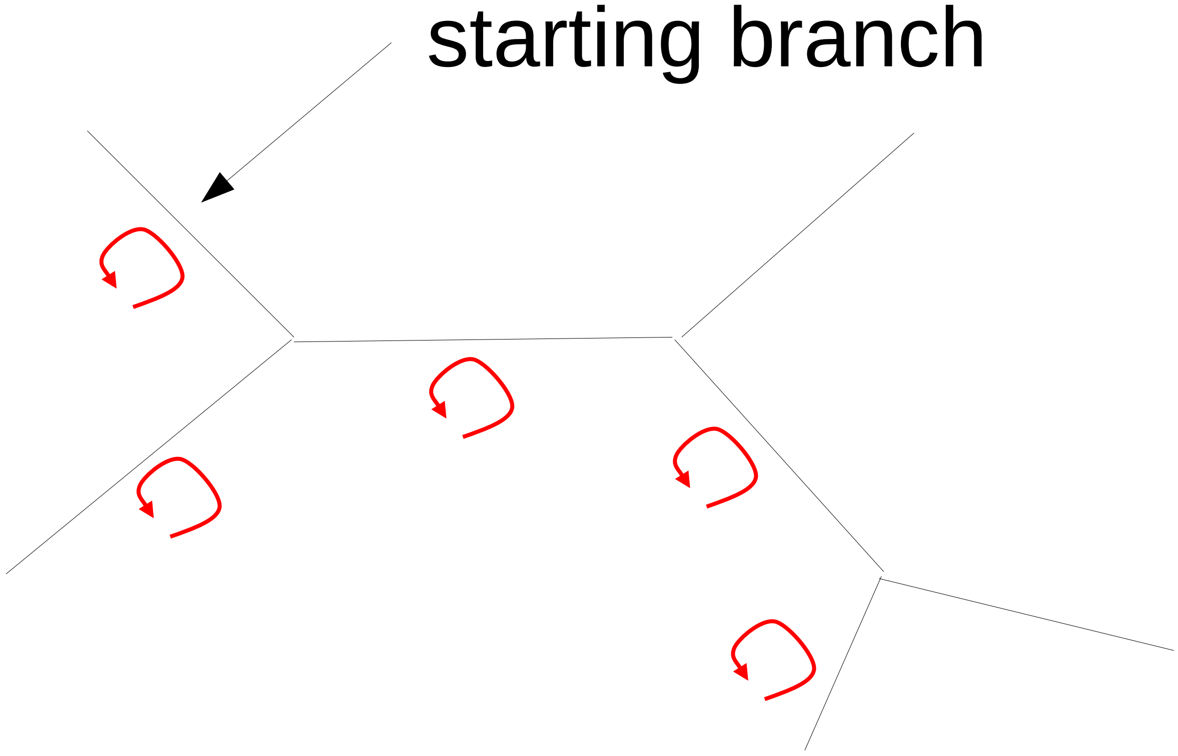
Branch Length Optimization



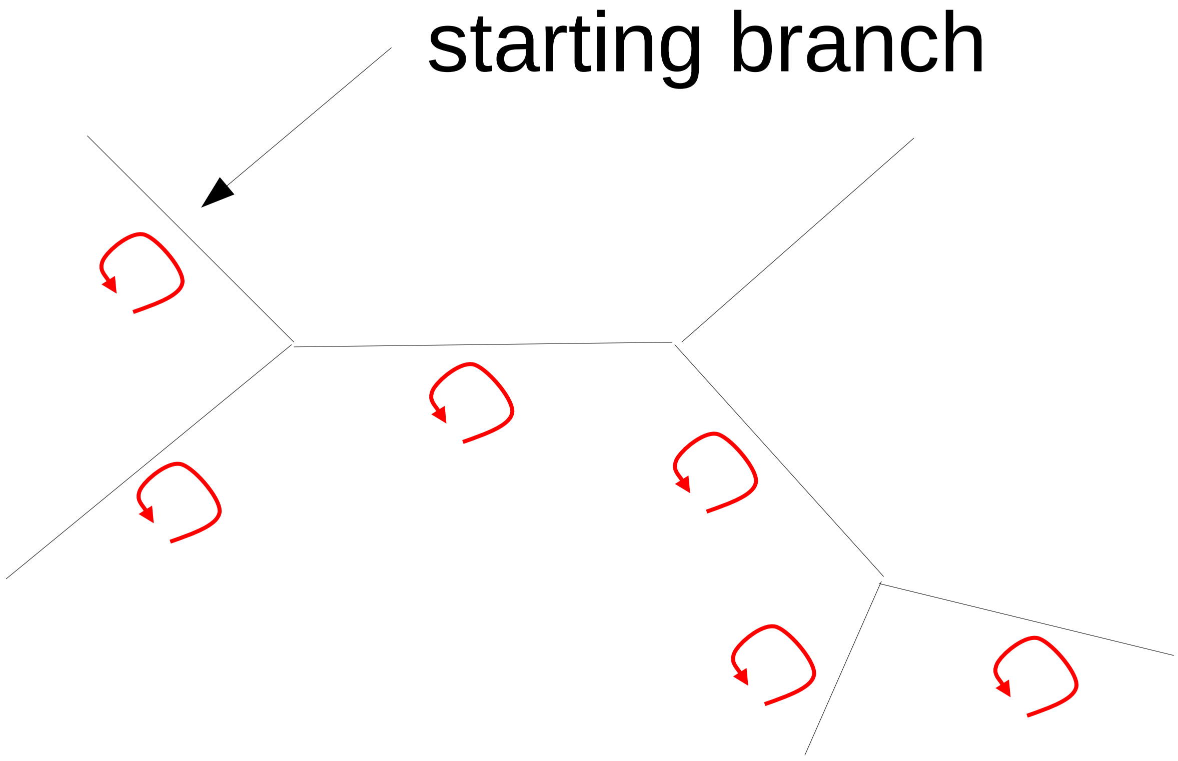
Branch Length Optimization



Branch Length Optimization

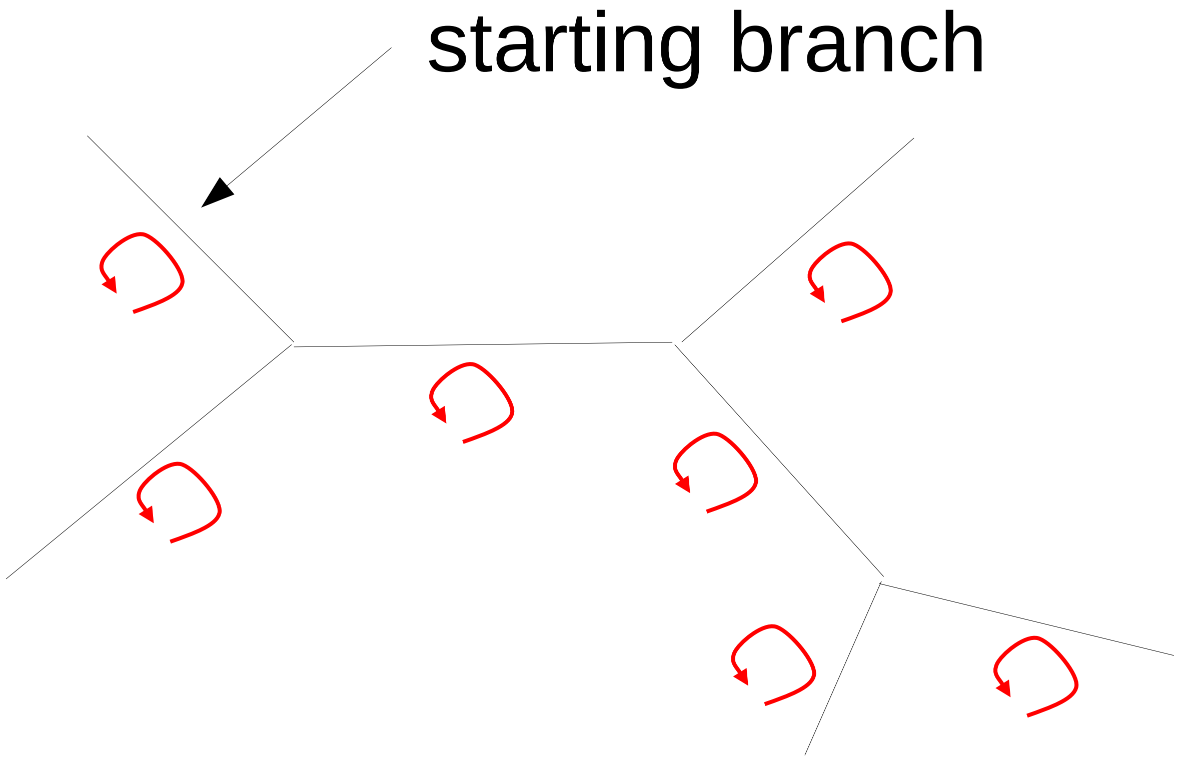


Branch Length Optimization

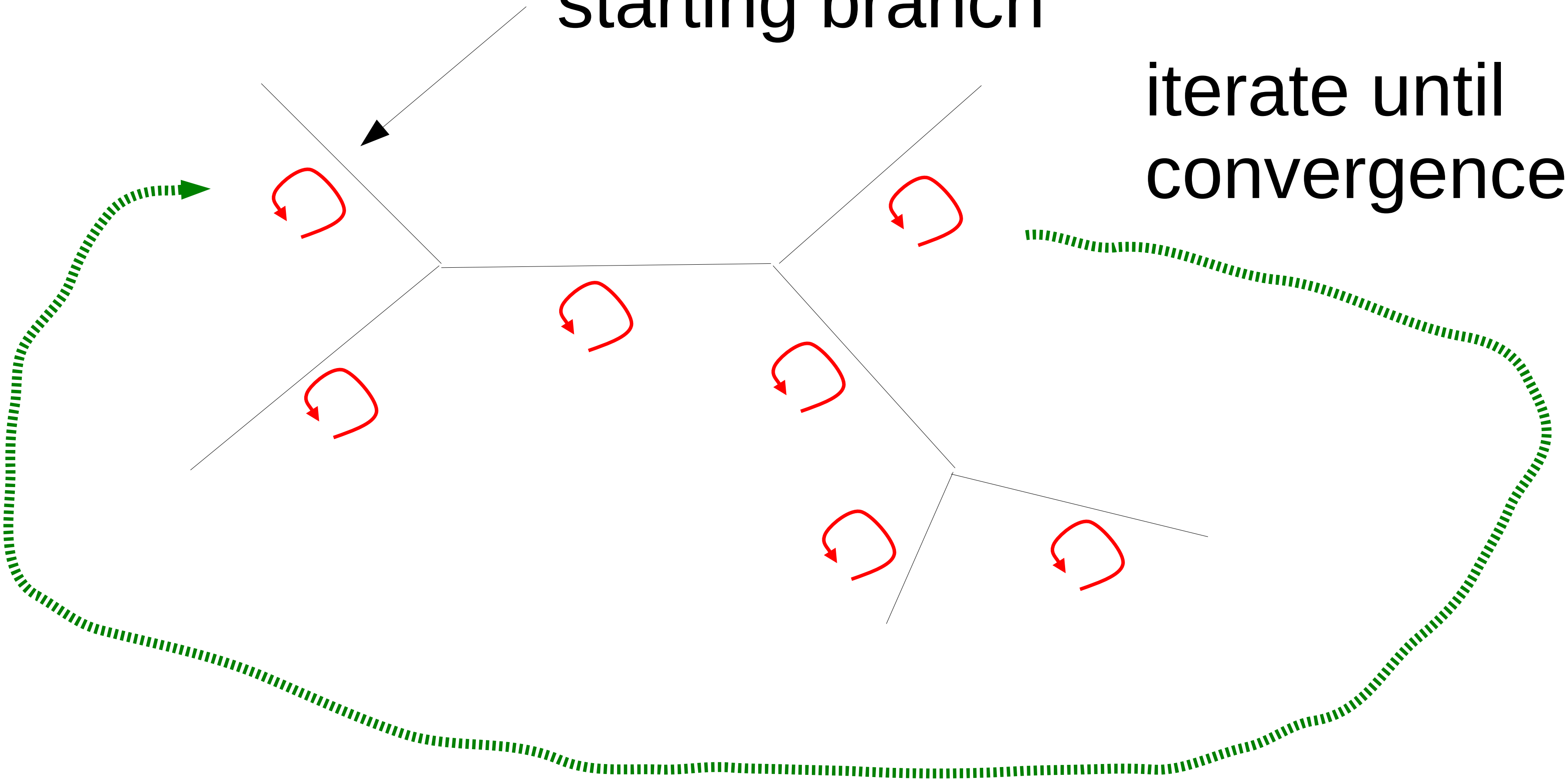


starting branch

Branch Length Optimization



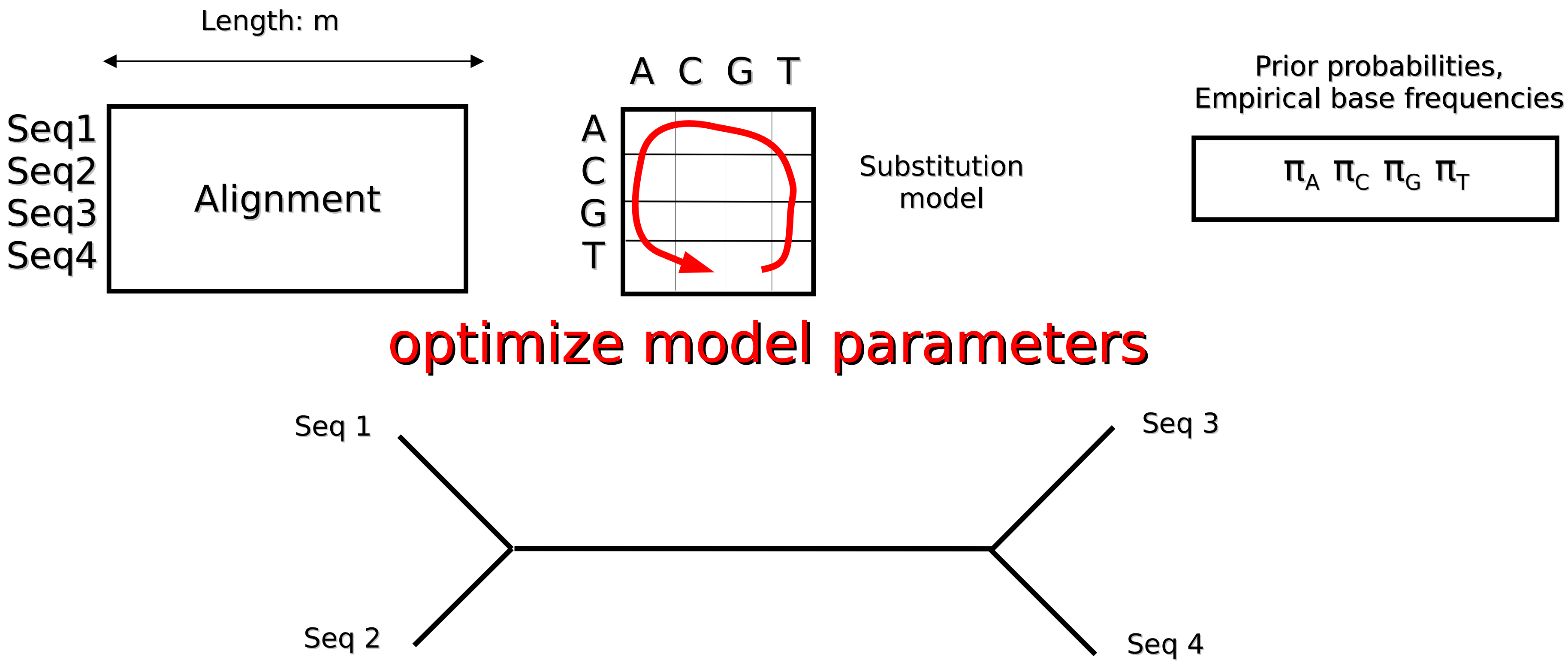
Branch Length Optimization



starting branch

iterate until convergence

Maximum Likelihood



Length: m

Seq1
Seq2
Seq3
Seq4

Alignment

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

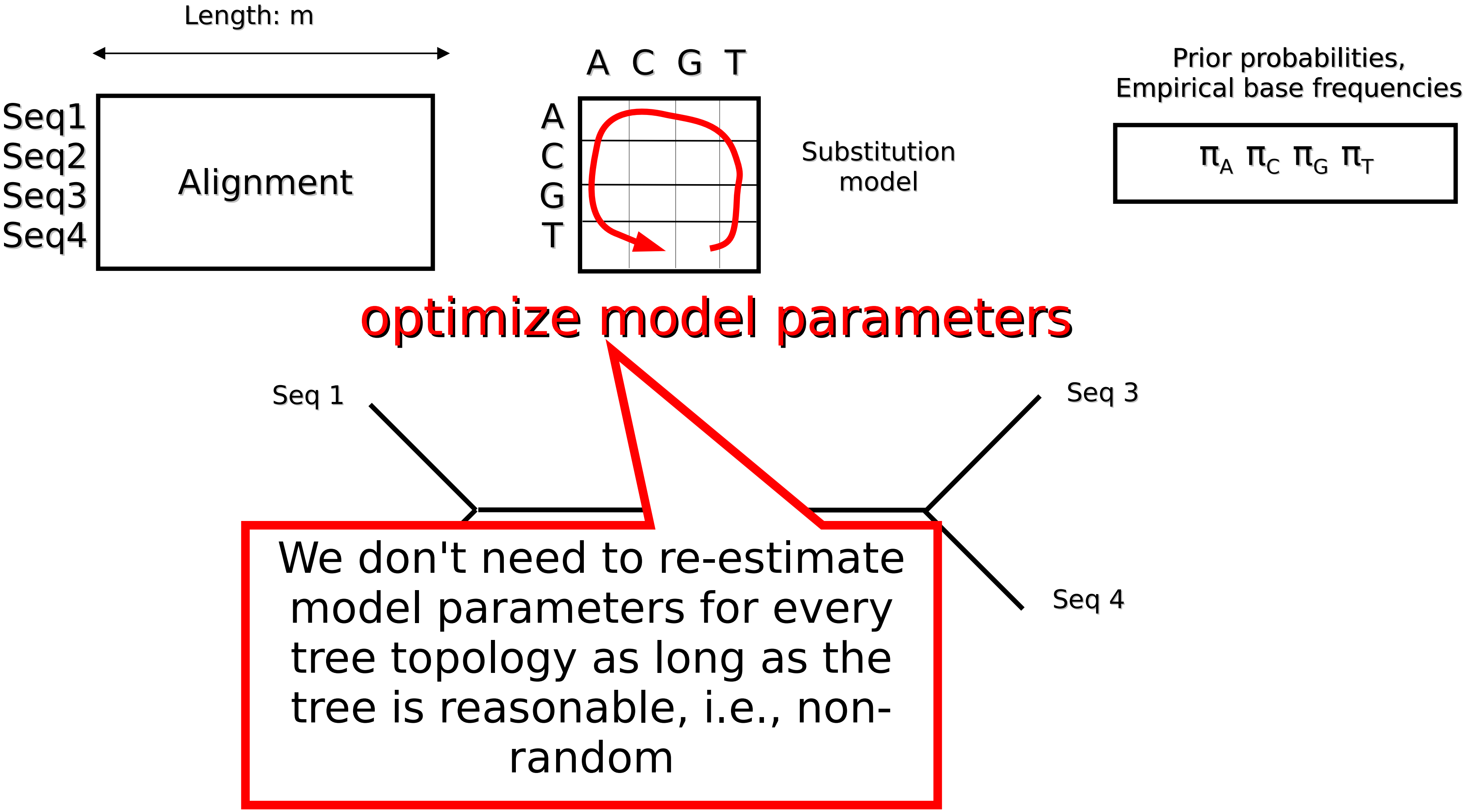
Substitution model

Prior probabilities,
Empirical base frequencies

π_A π_C π_G π_T

optimize model parameters

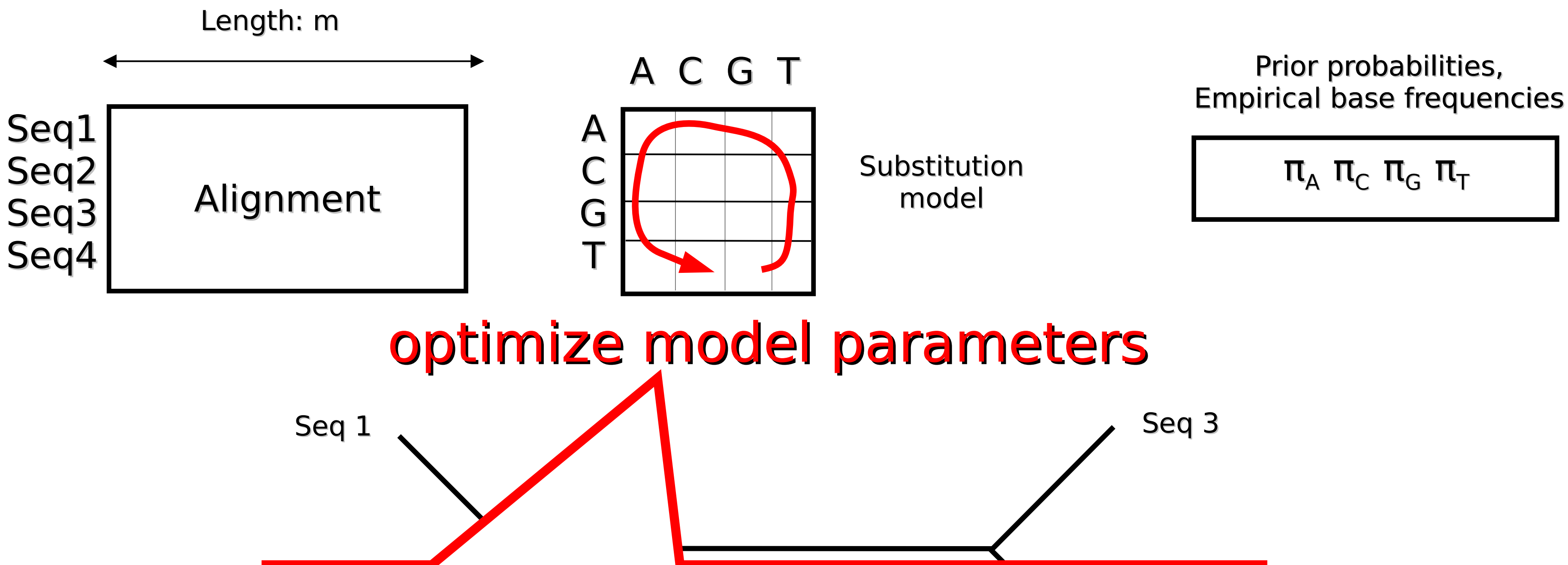
Maximum Likelihood



optimize model parameters

We don't need to re-estimate model parameters for every tree topology as long as the tree is reasonable, i.e., non-random

Maximum Likelihood



optimize model parameters

Methods used for model parameter optimization (other than branch lengths)

1. BFGS
2. Brent's method
3. Expectation maximization approaches

Numerical Optimization Procedures

- See chapters 9 & 10 of: *Numerical Recipes in C – The Art of Scientific Computing*

Basic Operations
Maximum Likelihood

- Compute Conditional Likelihood Vector at an inner node
- Compute Likelihood at Virtual Root
- Optimize a Branch Length for a given Branch
- Optimize all Branch Lengths
- Optimize other Model Parameters

Basic Operations
Maximum Likelihood

- Compute Conditional Likelihood Vector at an inner node
- Compute Likelihood at Virtual Root
- Optimize a Branch Length for a given Branch
- Optimize all Branch Lengths
- Optimize other Model Parameters

The optimizers are the tricky routines!