

# Lesson 6

Maximum likelihood & Bayesian  
Theory

# So far in clustering algorithms for sequence and evolution...

Trees are the natural way of representing relationships among species.

The number of possible trees increases factorially with the number of OTUs.



# So far in clustering algorithms for sequence and evolution...

We need algorithms for retrieving the tree that generated the data without exploring ALL the space.

We have described movements that we can go from one tree to another in heuristic approaches for finding the best tree given a cost



# So far in clustering algorithms for sequence and evolution...

So far we have seen  
methods based on  
generating distances  
between OTUs

...methods that are  
based on using  
distance matrices...



# So far in clustering algorithms for sequence and evolution...

...and maximum parsimony, that is based in characters

Today we will look at maximum likelihood cost functions



# So far in clustering algorithms for sequence and evolution...

Wait a moment!  
LIKELIHOOD!!! That  
sounds like statistics!!!!!!

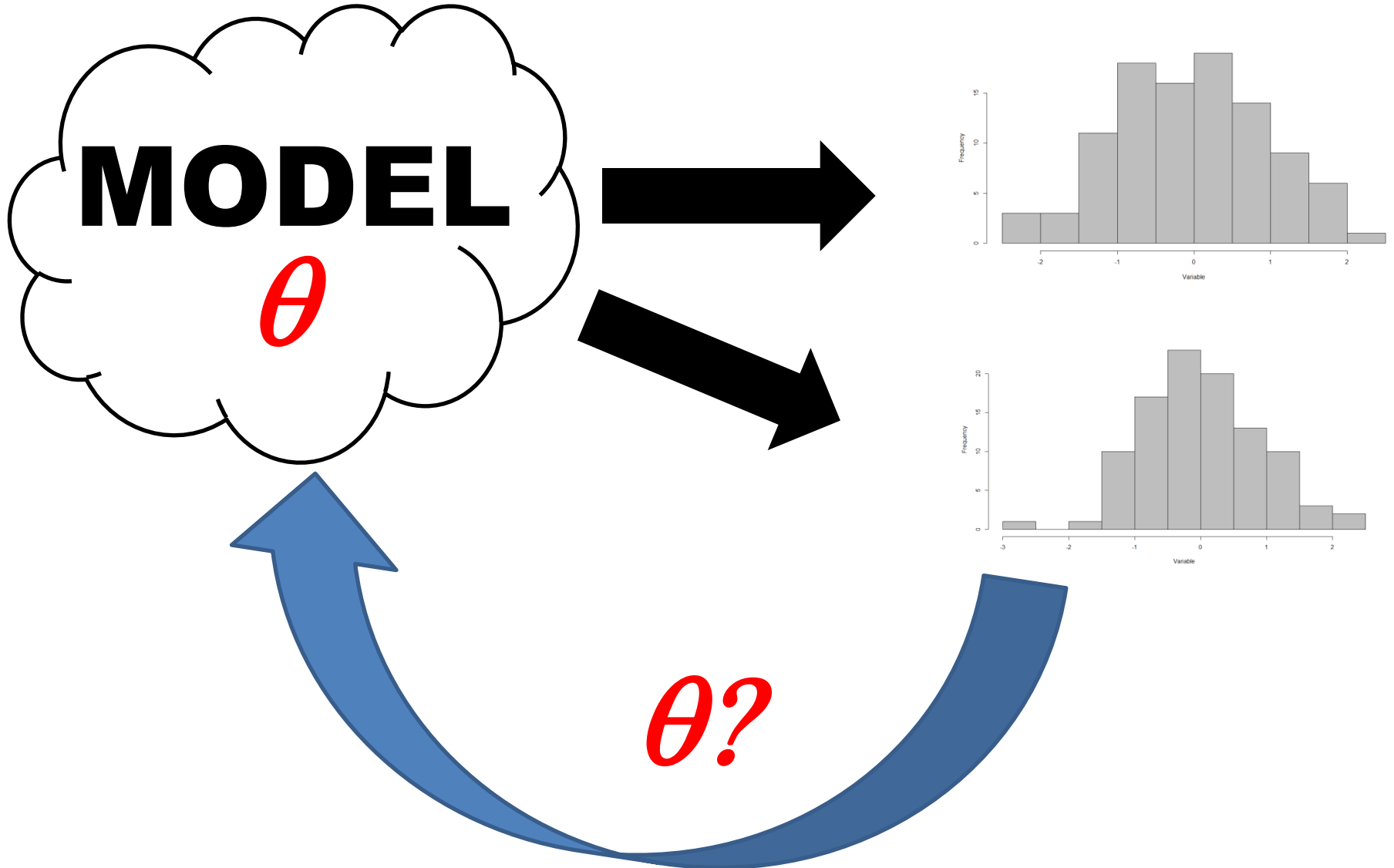
NOOOOOOOO!!!!!!  
**Statistics**  
NOOOOOOOOOOOO!!!!!!



So far in clustering algorithms for  
sequence and evolution...

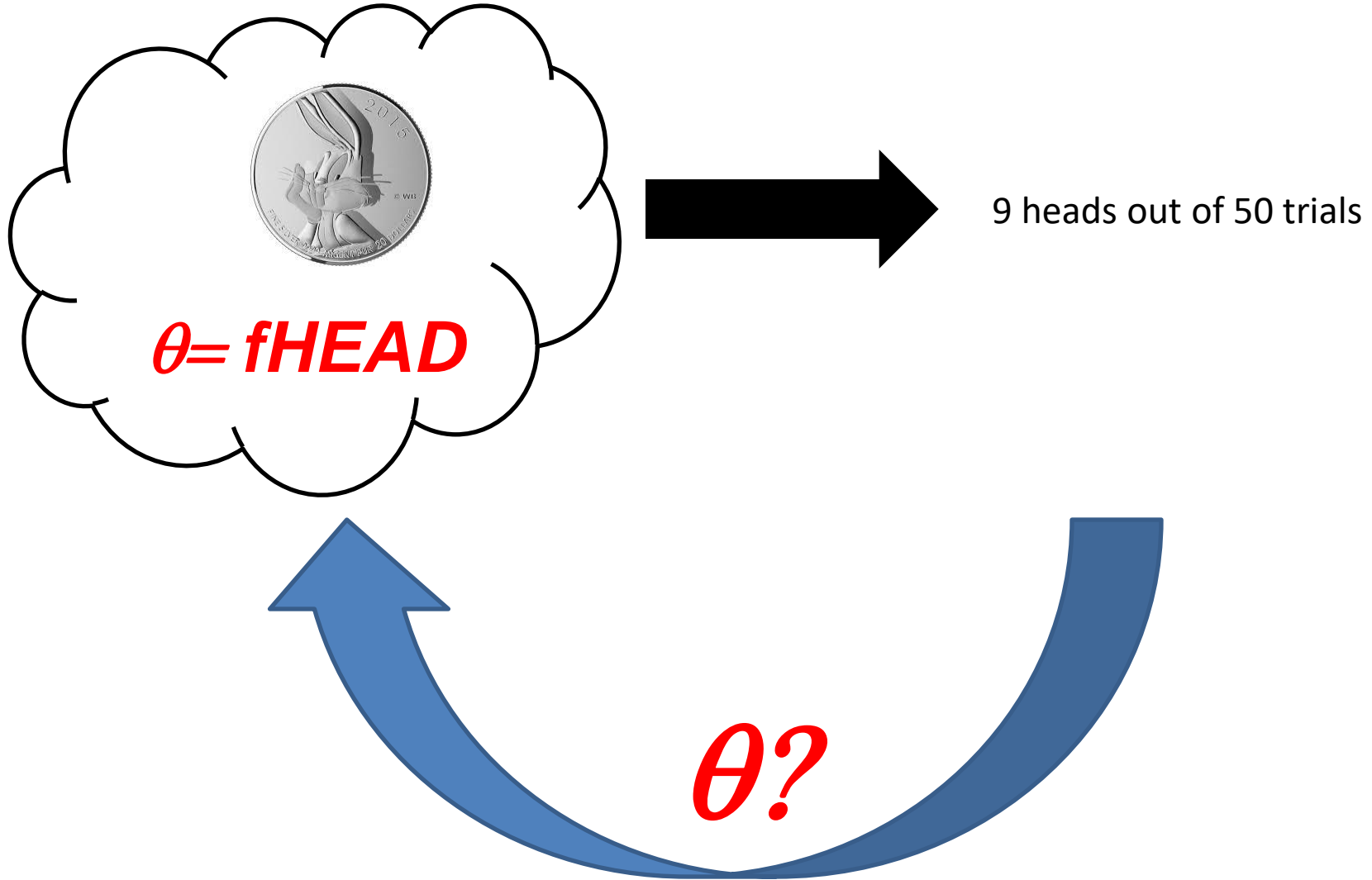


# Remember: statistics





# Remember: statistics



# Remember: statistics



**$\theta = f_{HEAD}$**

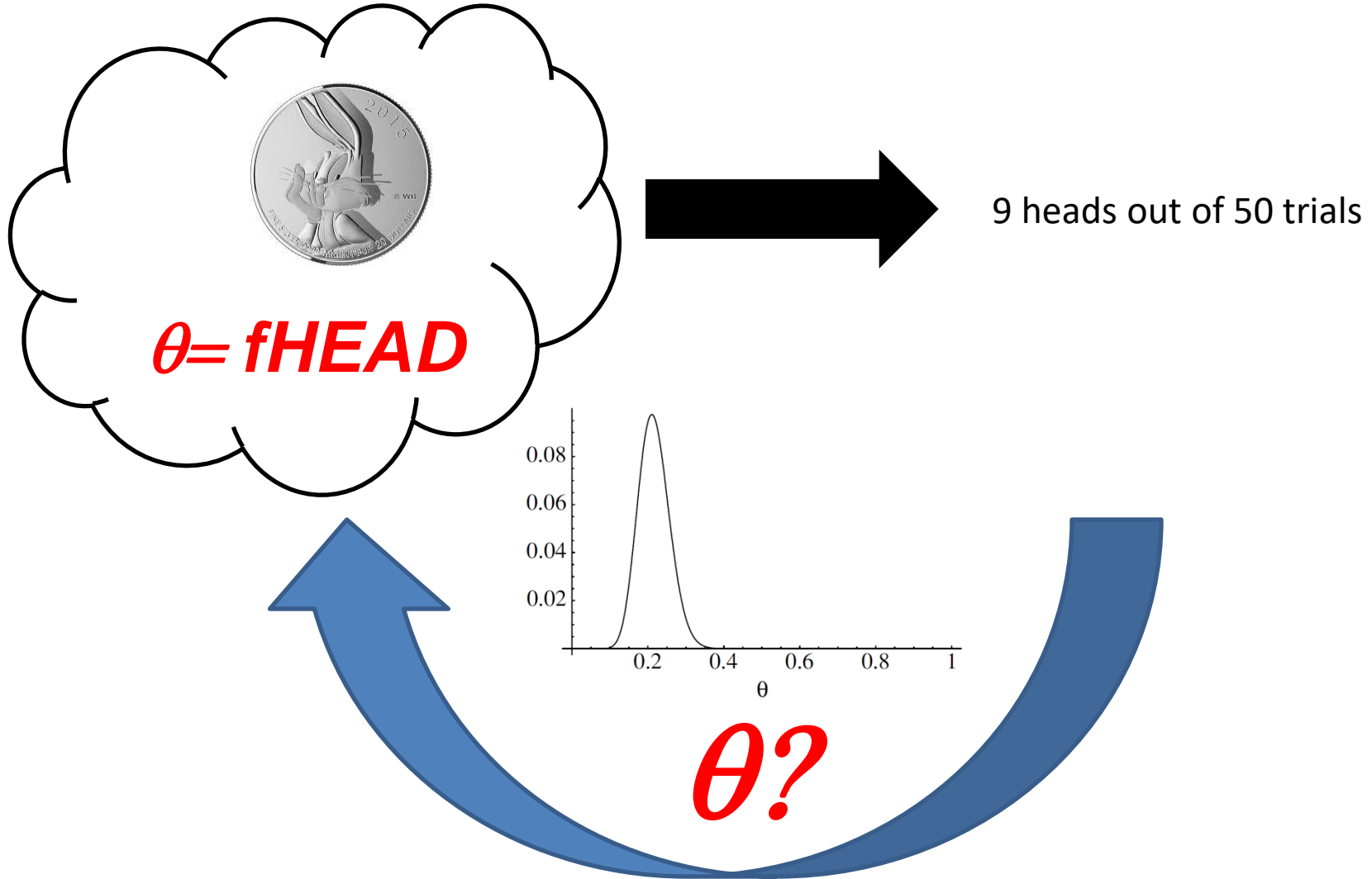


9 heads out of 50 trials

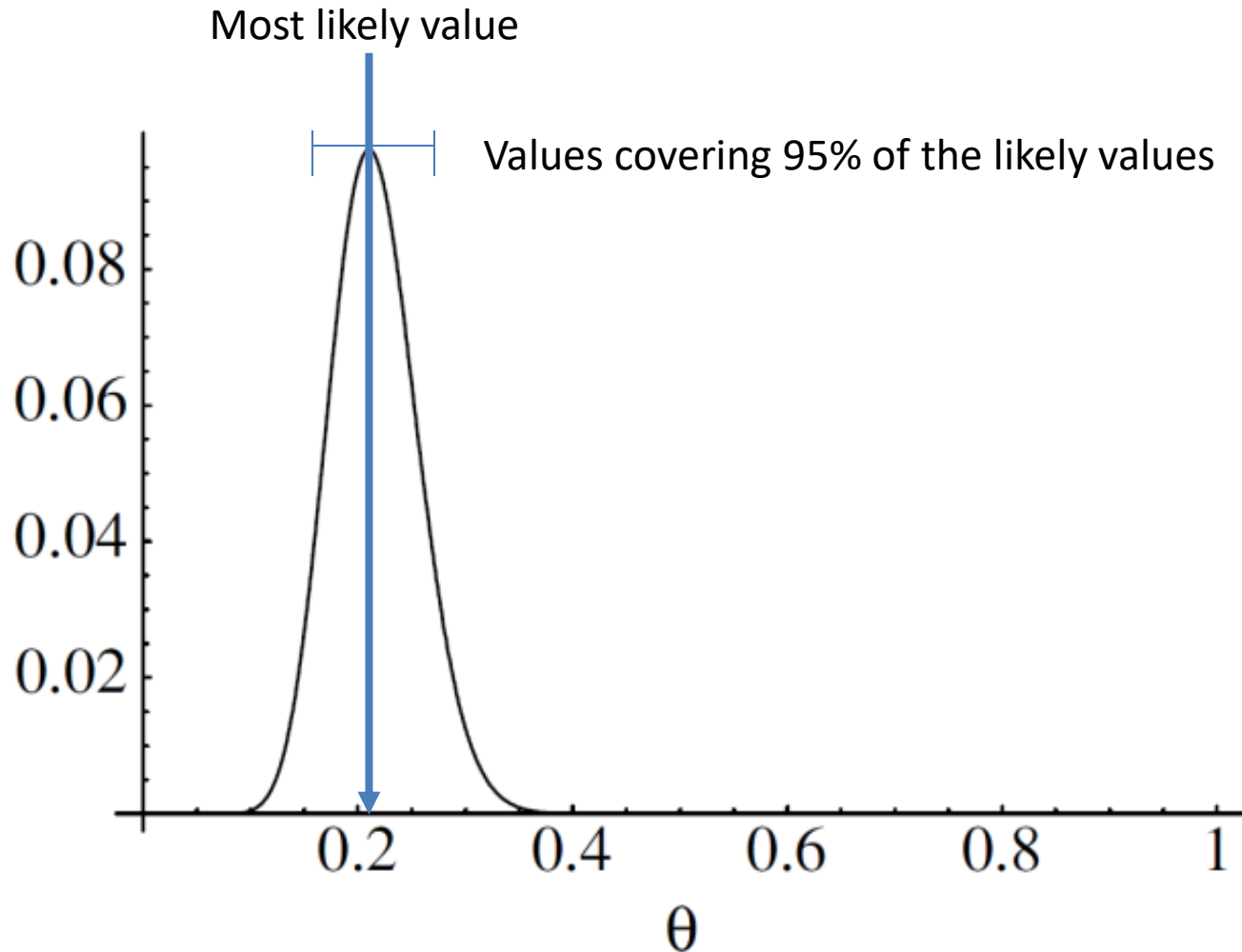
$$L(\theta) = \Pr[H = h] = \binom{n}{h} \theta^h (1 - \theta)^{n-h}$$

**$\theta?$**

# Remember: statistics



# Remember: statistics



# Remember: statistics

How do we recover the most likely value of the parameter  $\theta$  that produced the data?

$$L(\theta) = \Pr[H = h] = \binom{n}{h} \theta^h (1 - \theta)^{n-h}$$

$$h = 9$$

$$n = 50$$

$$\log(L(\theta)) = \log(C) + h * \log(\theta) + (n-h)*\log(1 - \theta)$$

$$L(\theta) = \log(C) + h * \log(\theta) + (n-h)*\log(1-\theta)$$

$$\frac{dL}{d\theta} = \frac{h}{\theta} - \frac{n-h}{1-\theta}$$

$$\frac{dL}{d\theta} = 0; \theta = \frac{h}{n}$$

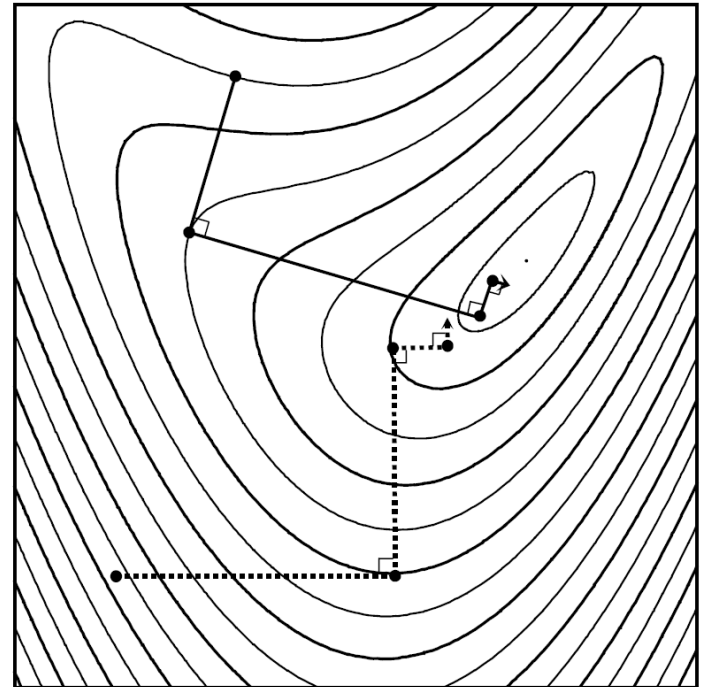
# Remember: statistics

How do we recover the most likely value of the parameter  $\theta$  that produced the data?

$$\frac{dL}{d\theta} = 0; ?$$

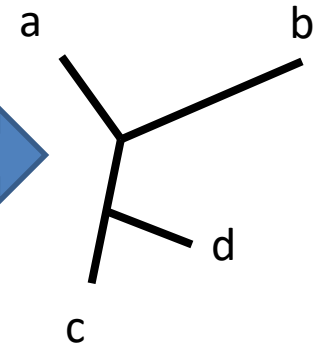
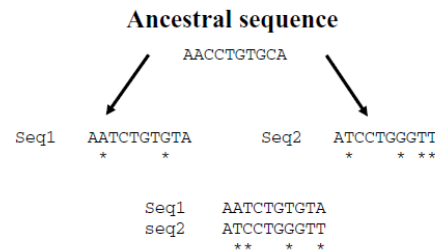
Newton-Raphson method for finding maximum

$$\theta_{k+1} = \theta_k - \frac{f'(\theta_k)}{f''(\theta_k)}.$$



# Remember: Character based methods

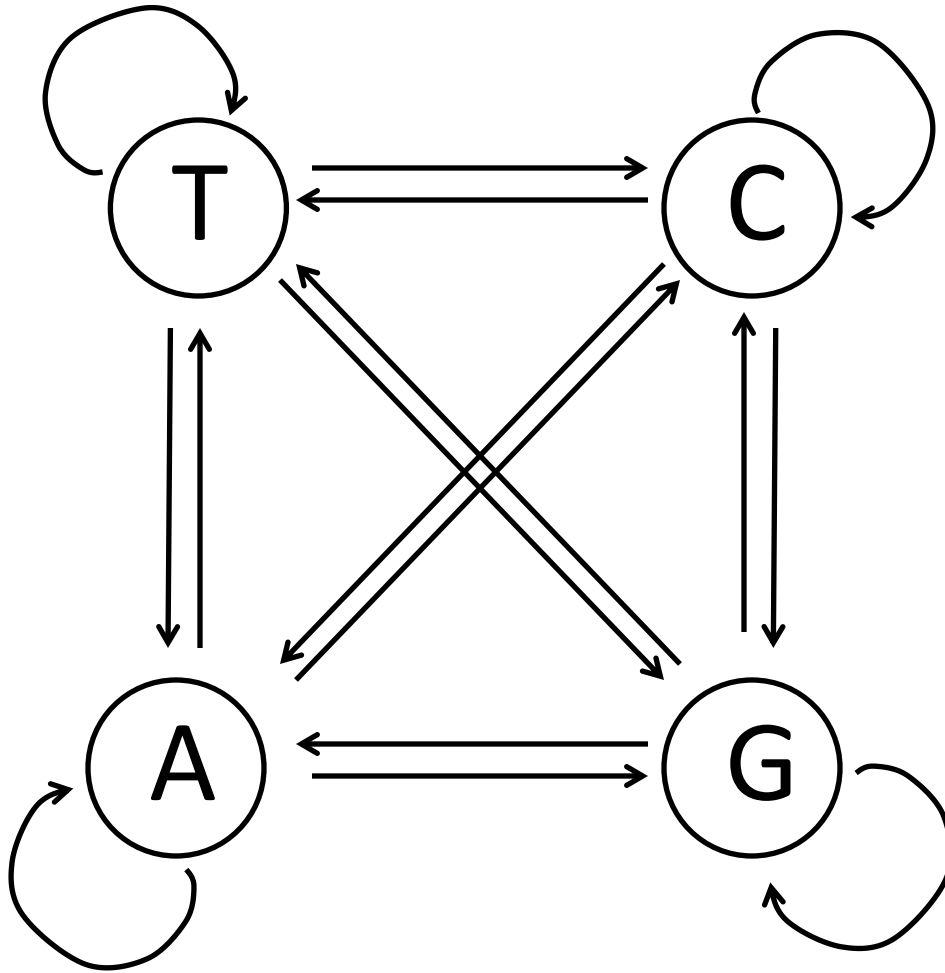
OTU	1	2	3	4	5	6
<b>a</b>	A	T	A	T	A	C
<b>b</b>	A	T	C	T	A	C
<b>c</b>	G	T	C	G	A	C
<b>d</b>	T	T	C	G	T	C



Model that explains  
how data is  
generated

How do you build the  
tree

# Remember: a model of evolution



Probability change matrix

$P =$

		Goes to			
		A	C	T	G
Comes from	A				
	C				
	T				
	G				

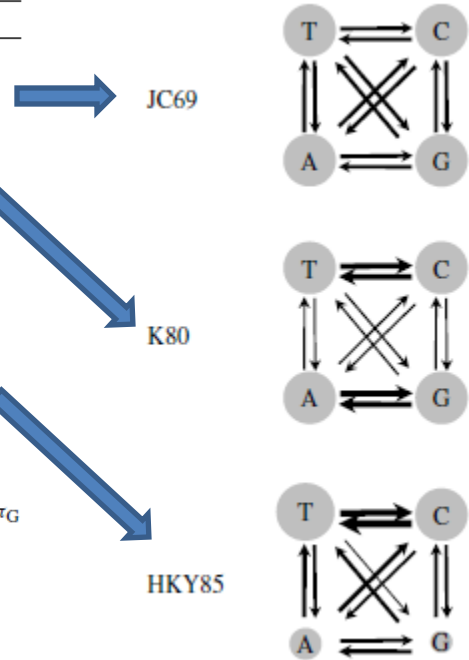


# Remember: a model of evolution

**Table 1.1** Substitution-rate matrices for commonly used Markov models of nucleotide substitution

	From	To			
		T	C	A	G
JC69 (Jukes and Cantor 1969)	T	.	$\lambda$	$\lambda$	$\lambda$
	C	$\lambda$	.	$\lambda$	$\lambda$
	A	$\lambda$	$\lambda$	.	$\lambda$
	G	$\lambda$	$\lambda$	$\lambda$	.
K80 (Kimura 1980)	T	.	$\alpha$	$\beta$	$\beta$
	C	$\alpha$	.	$\beta$	$\beta$
	A	$\beta$	$\beta$	.	$\alpha$
	G	$\beta$	$\beta$	$\alpha$	.
F81 (Felsenstein 1981)	T	.	$\pi_C$	$\pi_A$	$\pi_G$
	C	$\pi_T$	.	$\pi_A$	$\pi_G$
	A	$\pi_T$	$\pi_C$	.	$\pi_G$
	G	$\pi_T$	$\pi_C$	$\pi_A$	.
HKY85 (Hasegawa <i>et al.</i> 1984, 1985)	T	.	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$\alpha\pi_T$	.	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	.	$\alpha\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	.
F84 (Felsenstein, DNAML program since 1984)	T	.	$(1 + \kappa/\pi_Y)\beta\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$(1 + \kappa/\pi_Y)\beta\pi_T$	.	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	.	$(1 + \kappa/\pi_R)\beta\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$(1 + \kappa/\pi_R)\beta\pi_A$	.
TN93 (Tamura and Nei 1993)	T	.	$\alpha_1\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$\alpha_1\pi_T$	.	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	.	$\alpha_2\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$\alpha_2\pi_A$	.
GTR (REV) (Tavaré 1986; Yang 1994b; Zharkikh 1994)	T	.	$a\pi_C$	$b\pi_A$	$c\pi_G$
	C	$a\pi_T$	.	$d\pi_A$	$e\pi_G$
	A	$b\pi_T$	$d\pi_C$	.	$f\pi_G$
	G	$c\pi_T$	$e\pi_C$	$f\pi_A$	.
UNREST (Yang 1994b)	T	.	$q_{TC}$	$q_{TA}$	$q_{TG}$
	C	$q_{CT}$	.	$q_{CA}$	$q_{CG}$
	A	$q_{AT}$	$q_{AC}$	.	$q_{AG}$
	G	$q_{GT}$	$q_{GC}$	$q_{GA}$	.

The diagonals of the matrix are determined by the requirement that each row sums to 0. The equilibrium distribution is  $\pi = (1/4, 1/4, 1/4, 1/4)$  under JC69 and K80, and  $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$  under F81, F84, HKY85, TN93, and GTR. Under the general unrestricted (UNREST) model, it is given by the equations  $\pi Q = 0$  under the constraint  $\sum_i \pi_i = 1$ .



**Table 4.1** The discrete-rates model

Site class	1	2	3	...	$K$
Probability	$p_1$	$p_2$	$p_3$	...	$p_K$
Rate	$r_1$	$r_2$	$r_3$	...	$r_K$

# Some basic notation

X1

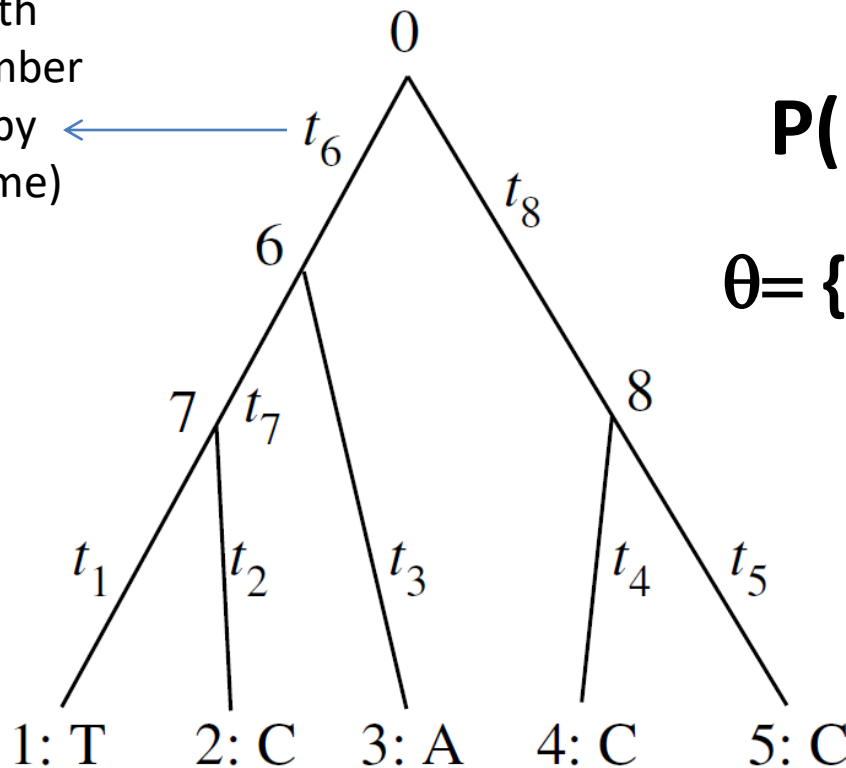


$X =$

<b>OTU</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
<b>a</b>	<i>A</i>	<i>T</i>	<i>A</i>	<i>T</i>	<i>A</i>	<i>C</i>
<b>b</b>	<i>A</i>	<i>T</i>	<i>C</i>	<i>T</i>	<i>A</i>	<i>C</i>
<b>c</b>	<i>G</i>	<i>T</i>	<i>C</i>	<i>G</i>	<i>A</i>	<i>C</i>
<b>d</b>	<i>T</i>	<i>T</i>	<i>C</i>	<i>G</i>	<i>T</i>	<i>C</i>

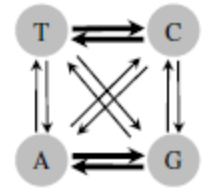
# And now... Imagine that we know the topology of the tree

Branch length  
(expected number  
of changes by  
amount of time) ←



$P(D|\theta)$

K80



$\theta = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}$

# And now... Imagine that we know the topology of the tree

**$P(D|\theta)$**       *“What is the probability (likelihood) of observing the data given the values of the set  $\theta$  of parameters”*

Which is the likelihood of position  $h$  given the length of the branches and my model of nucleotide evolution?

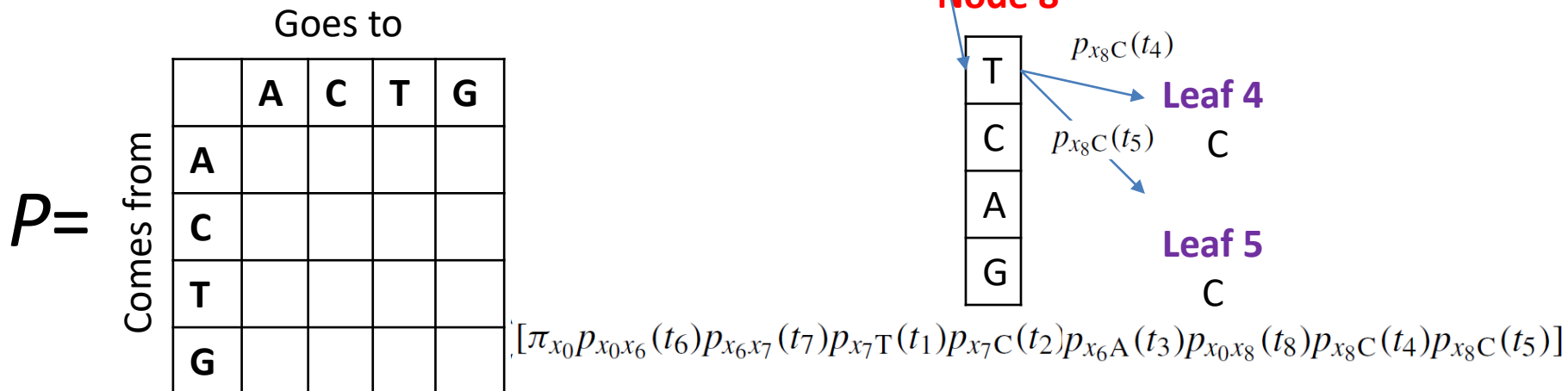
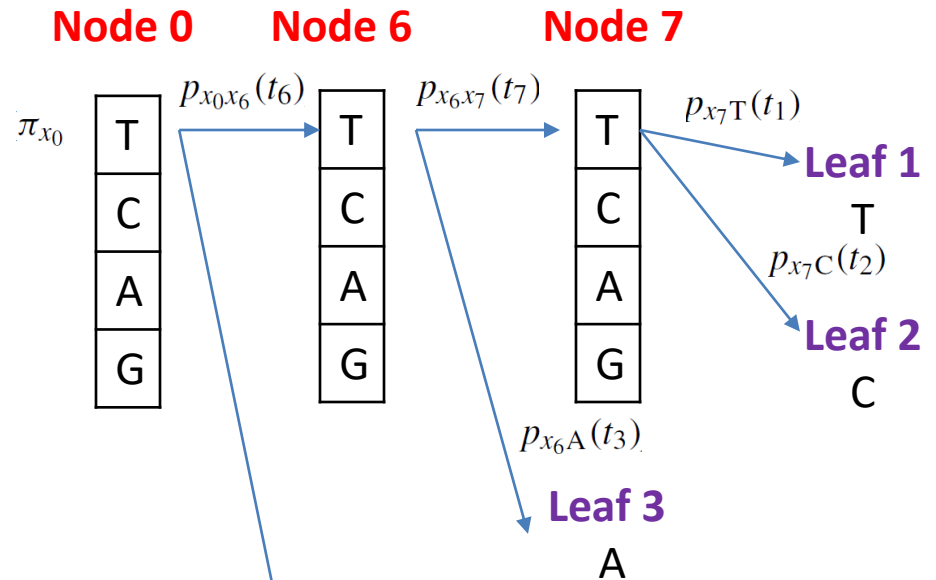
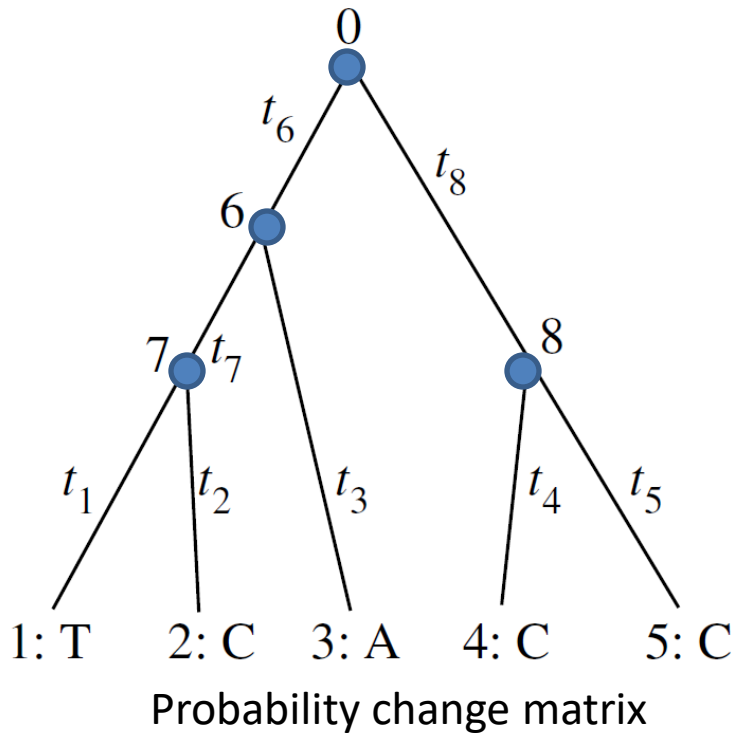
$$f(\mathbf{x}_h|\theta) = \sum_{x_0} \sum_{x_6} \sum_{x_7} \sum_{x_8} [\pi_{x_0} p_{x_0x_6}(t_6) p_{x_6x_7}(t_7) p_{x_7T}(t_1) p_{x_7C}(t_2) \\ \times p_{x_6A}(t_3) p_{x_0x_8}(t_8) p_{x_8C}(t_4) p_{x_8C}(t_5)].$$

# Take a moment to understand it...

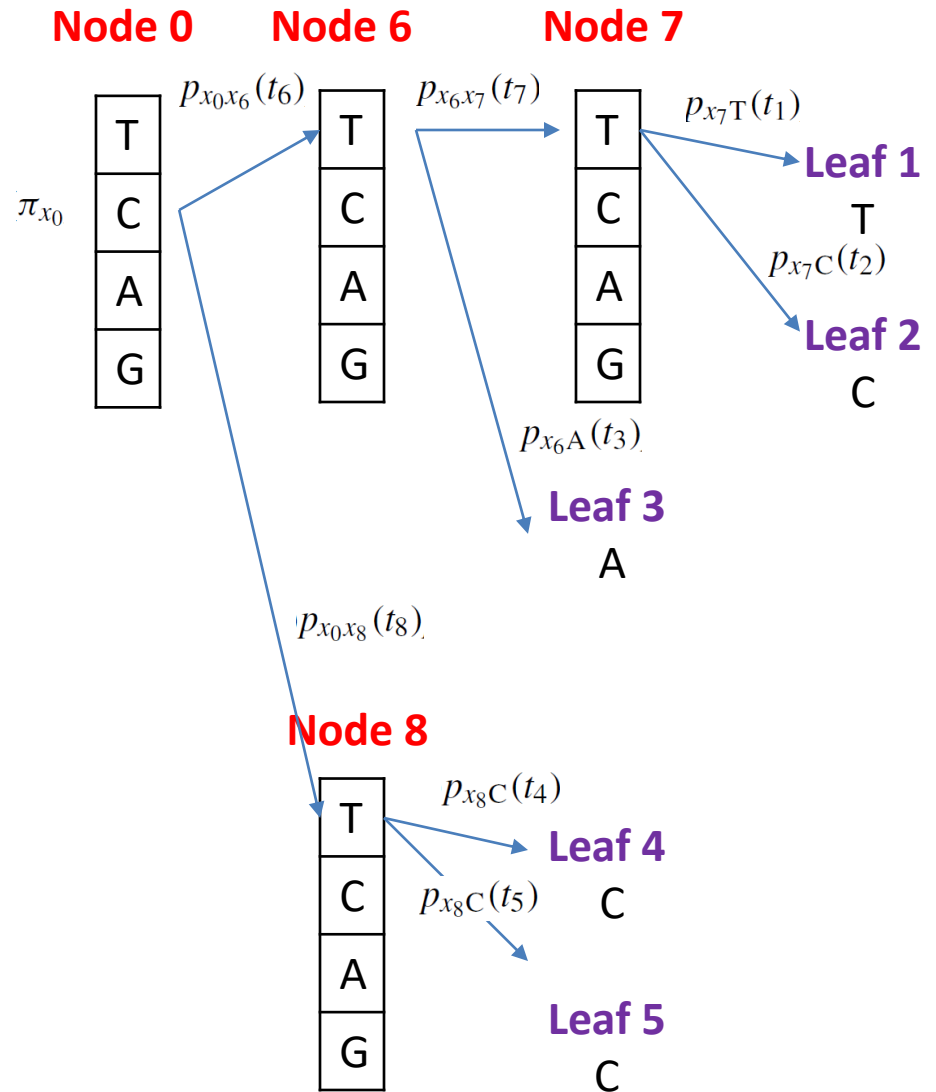
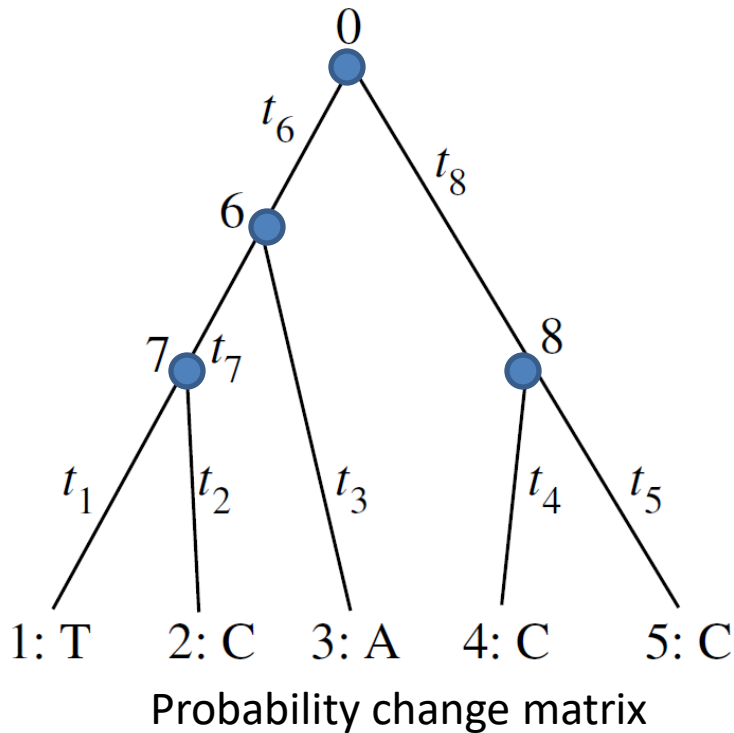


$$f(\mathbf{x}_h|\theta) = \sum_{x_0} \sum_{x_6} \sum_{x_7} \sum_{x_8} [\pi_{x_0} p_{x_0x_6}(t_6) p_{x_6x_7}(t_7) p_{x_7T}(t_1) p_{x_7C}(t_2) \\ \times p_{x_6A}(t_3) p_{x_0x_8}(t_8) p_{x_8C}(t_4) p_{x_8C}(t_5)].$$

# Take a moment to understand it...



# Take a moment to understand it...



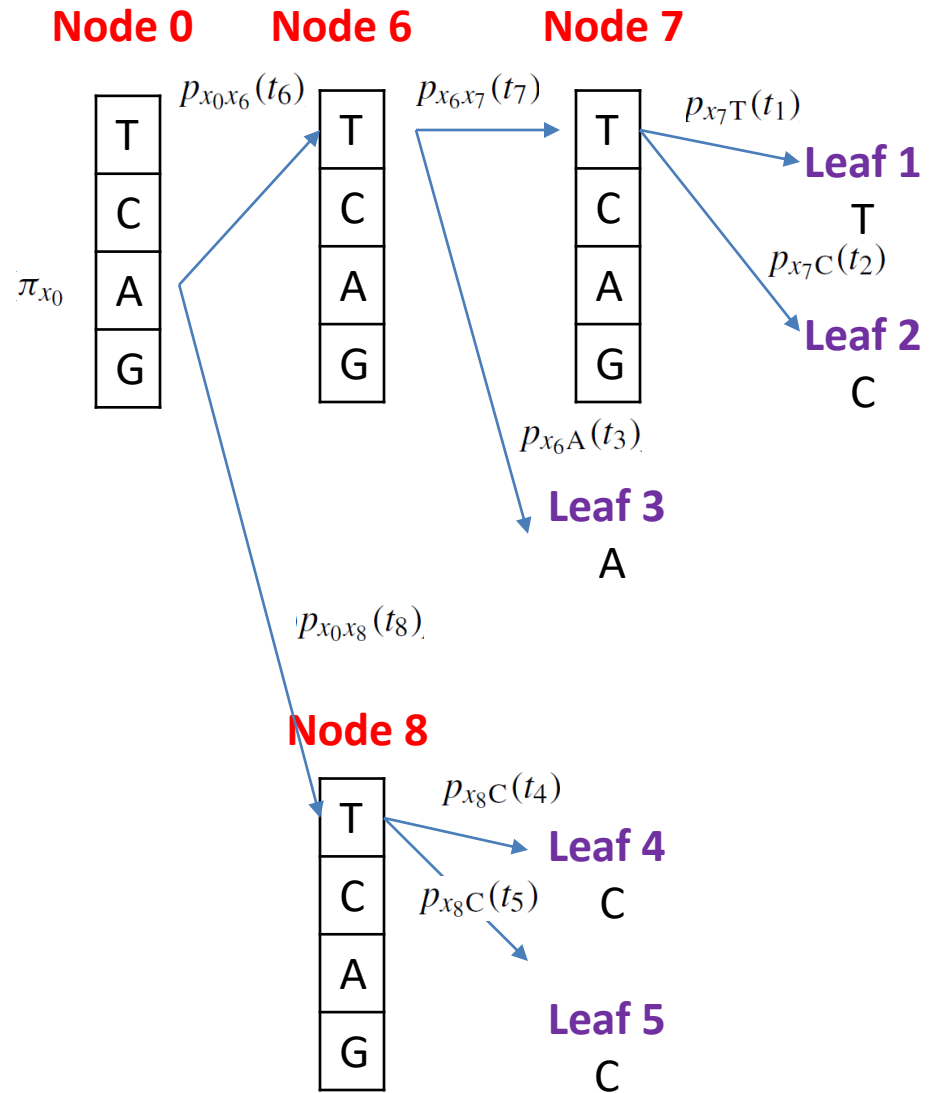
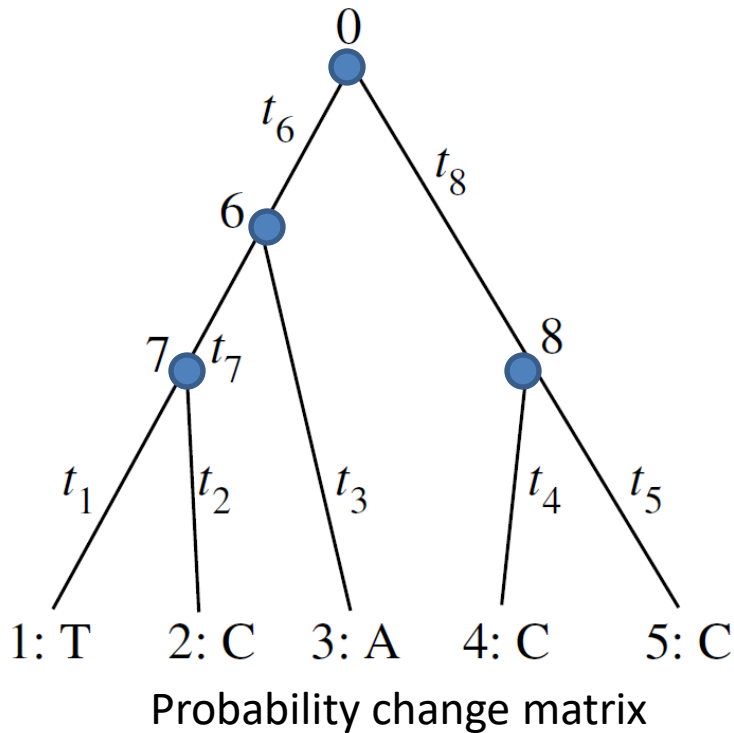
Comes from

Goes to

	A	C	T	G
A				
C				
T				
G				

$$[\pi_{x_0} p_{x_0 x_6}(t_6) p_{x_6 x_7}(t_7) p_{x_7 T}(t_1) p_{x_7 C}(t_2) p_{x_6 A}(t_3) p_{x_0 x_8}(t_8) p_{x_8 C}(t_4) p_{x_8 C}(t_5)]$$

# Take a moment to understand it...



Comes from

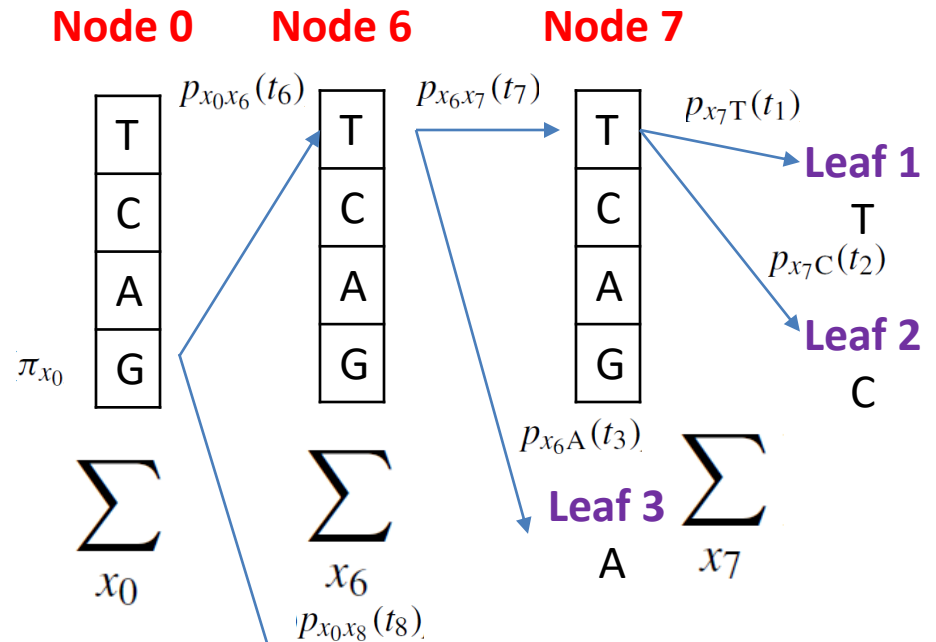
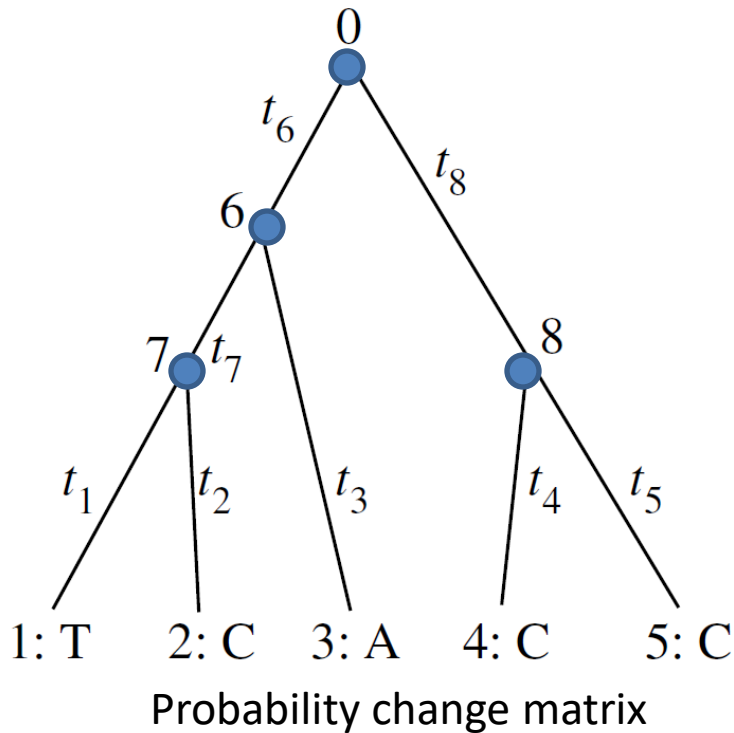
Goes to

	A	C	T	G
A				
C				
T				
G				

$$[\pi_{x_0} p_{x_0 x_6}(t_6) p_{x_6 x_7}(t_7) p_{x_7 T}(t_1) p_{x_7 C}(t_2) p_{x_6 A}(t_3) p_{x_0 x_8}(t_8) p_{x_8 C}(t_4) p_{x_8 C}(t_5)]$$



# Take a moment to understand it...

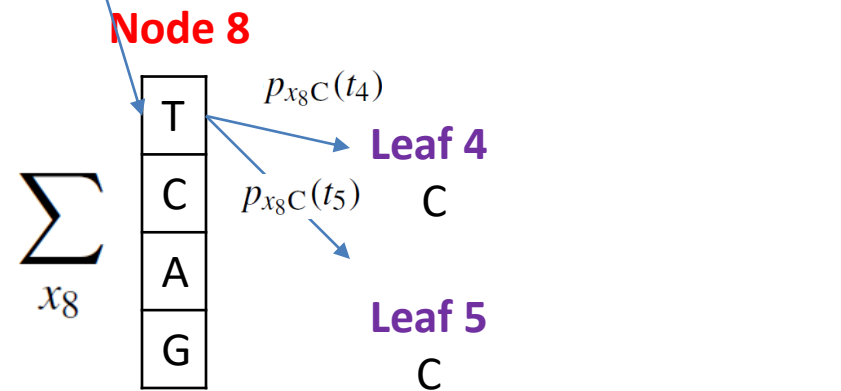


$P =$

Goes to

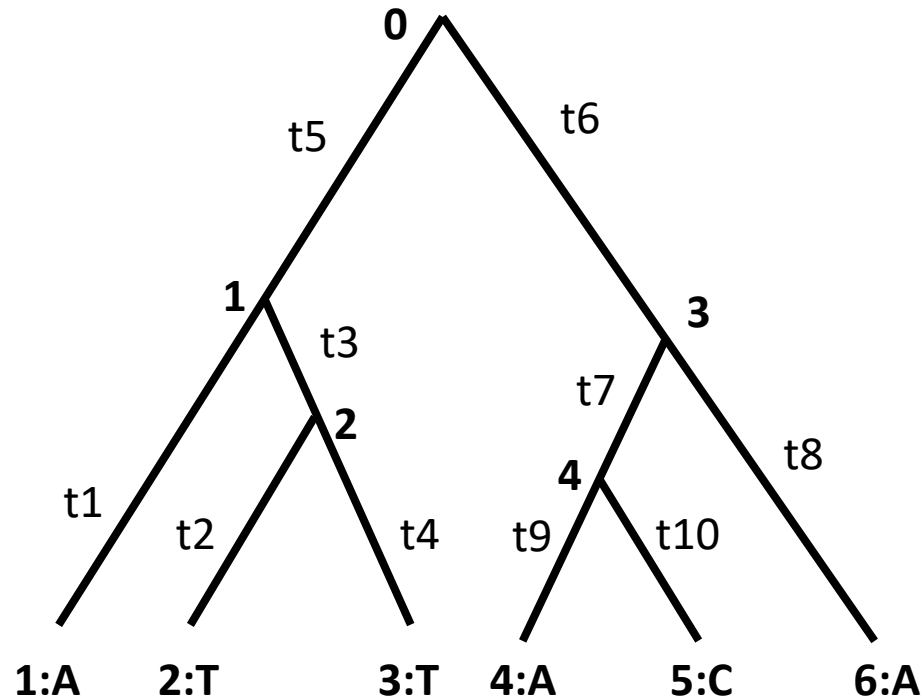
	A	C	T	G
A				
C				
T				
G				

Comes from



$$[\pi_{x_0} p_{x_0x_6}(t_6) p_{x_6x_7}(t_7) p_{x_7T}(t_1) p_{x_7C}(t_2) p_{x_6A}(t_3) p_{x_0x_8}(t_8) p_{x_8C}(t_4) p_{x_8C}(t_5)]$$

What would be the likelihood of this tree?

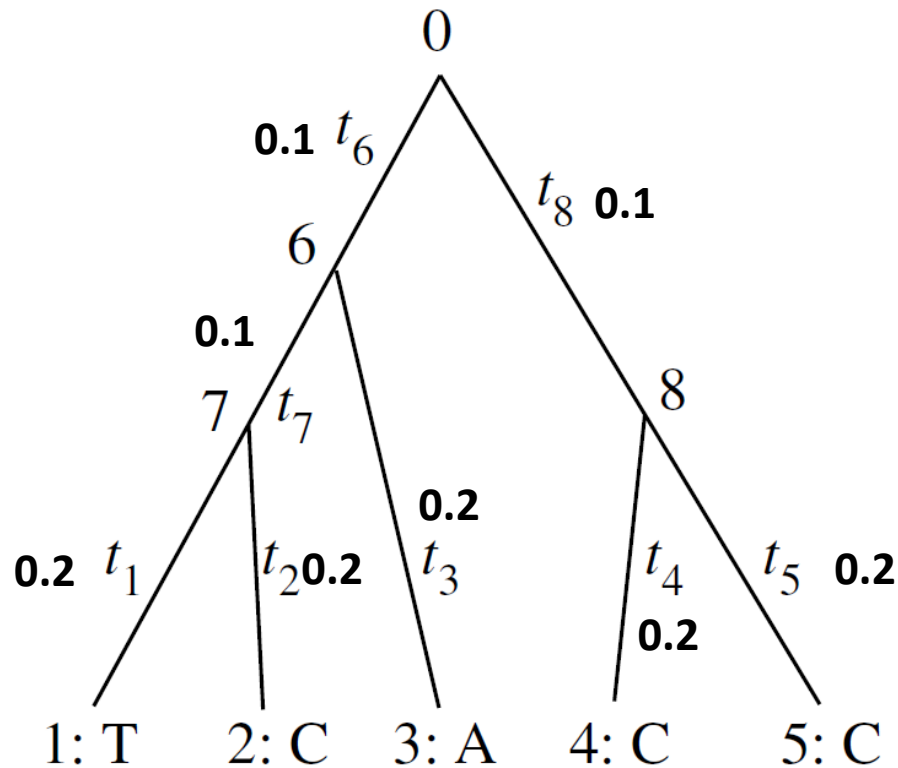




### Felselstein's pruning algorithm

$$\begin{aligned}
 f(\mathbf{x}_h|\theta) = & \sum_{x_0} \pi_{x_0} \left\{ \sum_{x_6} p_{x_0x_6}(t_6) \left[ \left( \sum_{x_7} p_{x_6x_7}(t_7) p_{x_7T}(t_1) p_{x_7C}(t_2) \right) p_{x_6A}(t_3) \right] \right\} \\
 & \times \left[ \sum_{x_8} p_{x_0x_8}(t_8) p_{x_8C}(t_4) p_{x_8C}(t_5) \right] \quad (4.3)
 \end{aligned}$$

And now... Imagine that we know the topology of the tree and the parameters!



$$P(0.1) = \begin{matrix} \text{T} \\ \text{C} \\ \text{A} \\ \text{G} \end{matrix} \begin{bmatrix} 0.906563 & 0.045855 & 0.023791 & 0.023791 \\ 0.045855 & 0.906563 & 0.023791 & 0.023791 \\ 0.023791 & 0.023791 & 0.906563 & 0.045855 \\ 0.023791 & 0.023791 & 0.045855 & 0.906563 \end{bmatrix},$$

$$P(0.2) = \begin{matrix} \text{T} \\ \text{C} \\ \text{A} \\ \text{G} \end{matrix} \begin{bmatrix} 0.825092 & 0.084274 & 0.045317 & 0.045317 \\ 0.084274 & 0.825092 & 0.045317 & 0.045317 \\ 0.045317 & 0.045317 & 0.825092 & 0.084274 \\ 0.045317 & 0.045317 & 0.084274 & 0.825092 \end{bmatrix}.$$

And now... Imagine that we know the topology of the tree and the parameters!

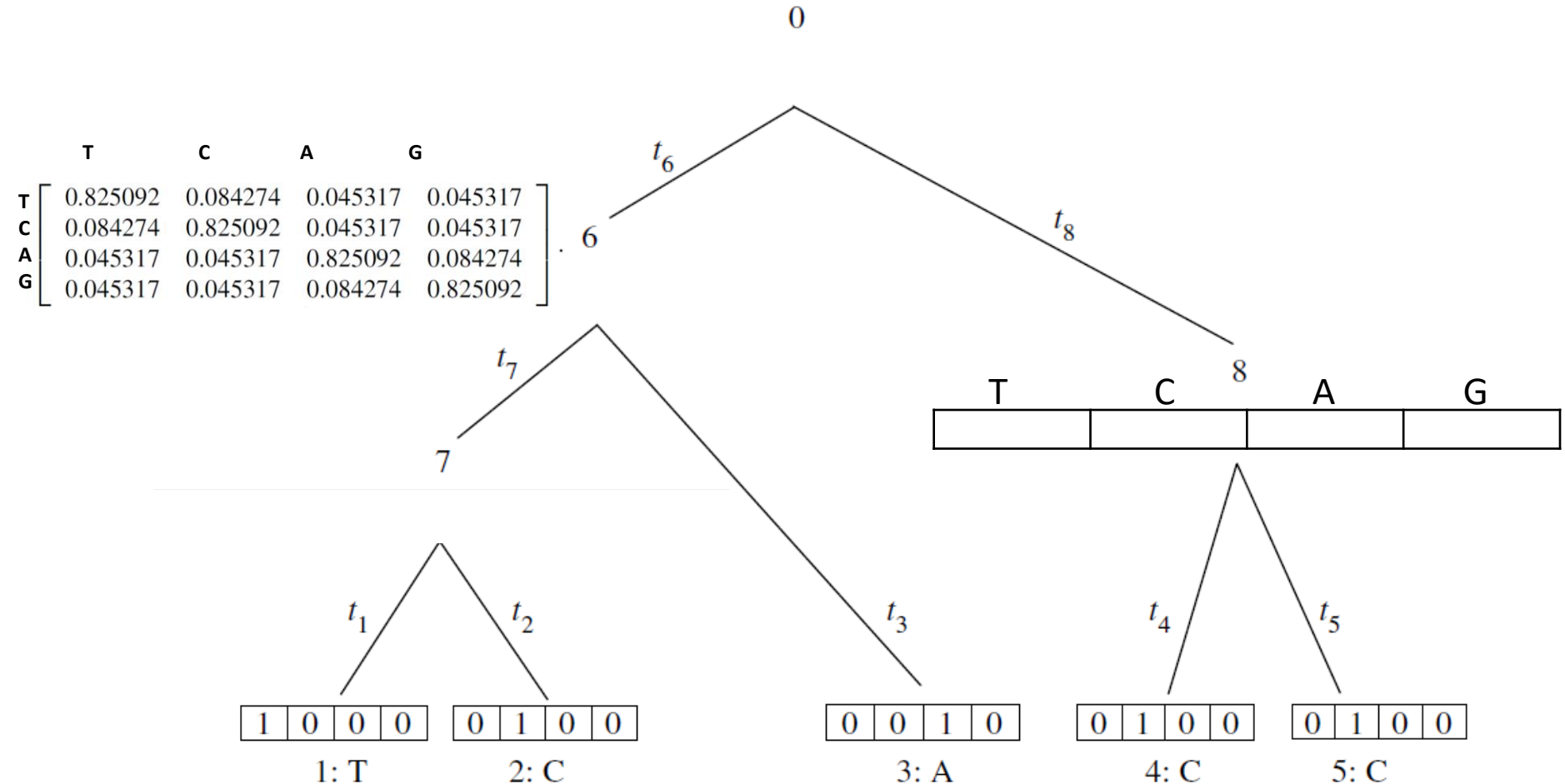
At the beginning of the node I was

Now I am

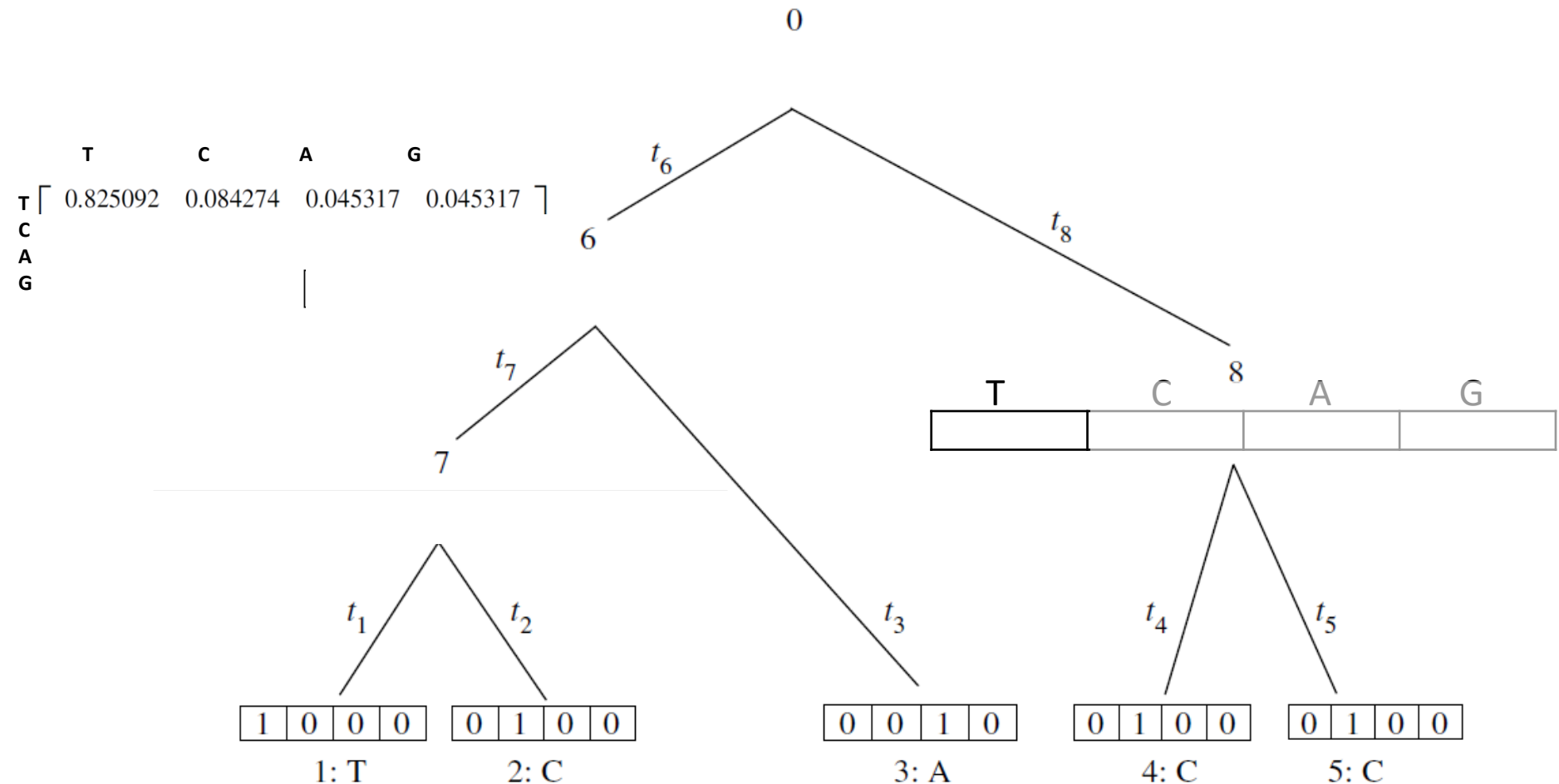
	T	C	A	G
T	0.825092	0.084274	0.045317	0.045317
C	0.084274	0.825092	0.045317	0.045317
A	0.045317	0.045317	0.825092	0.084274
G	0.045317	0.045317	0.084274	0.825092

And now... Imagine that we know the topology of the tree and the parameters!

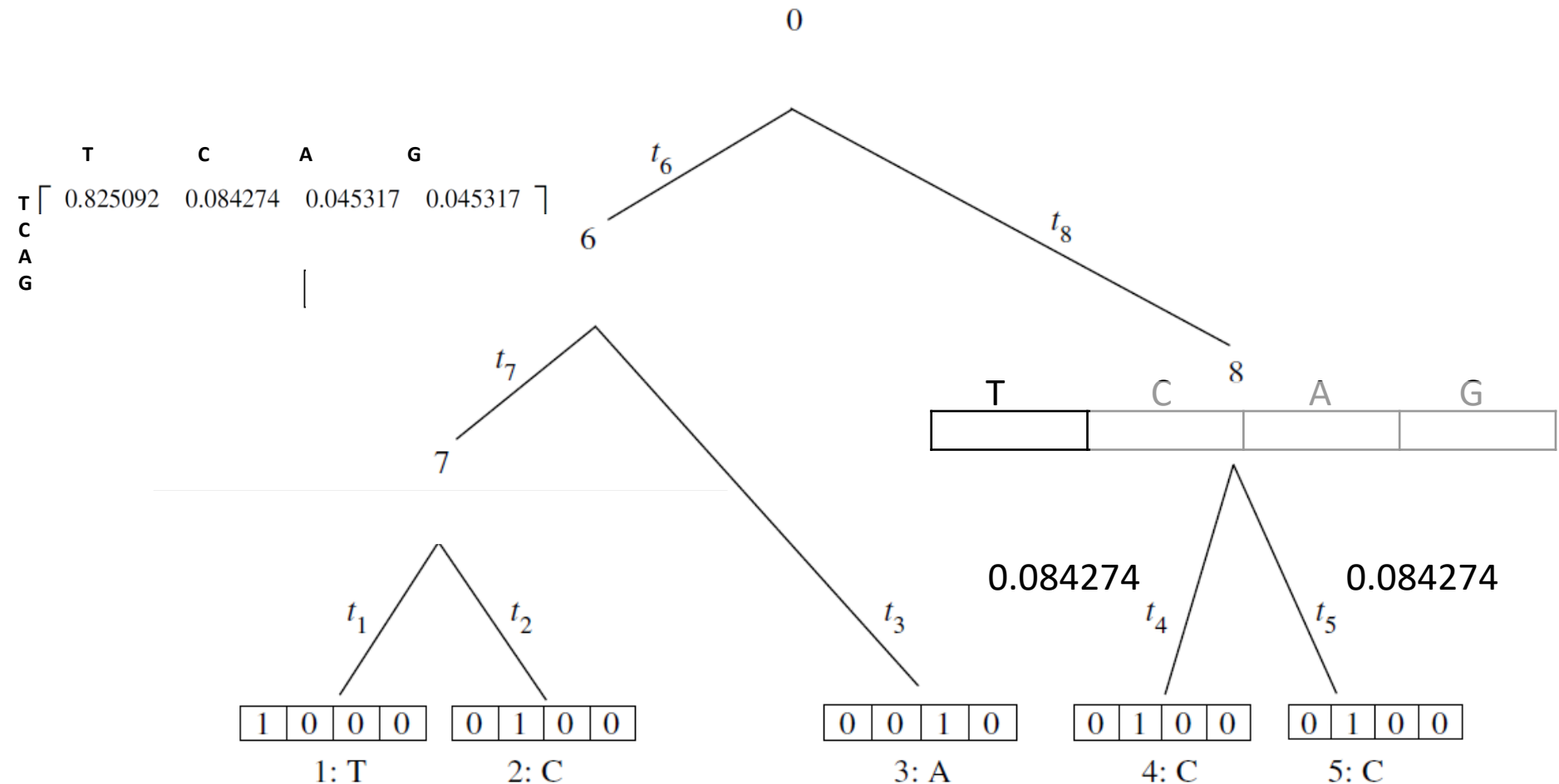
## The pruning algorithm for estimating the likelihood



And now... Imagine that we know the topology of the tree and the parameters!

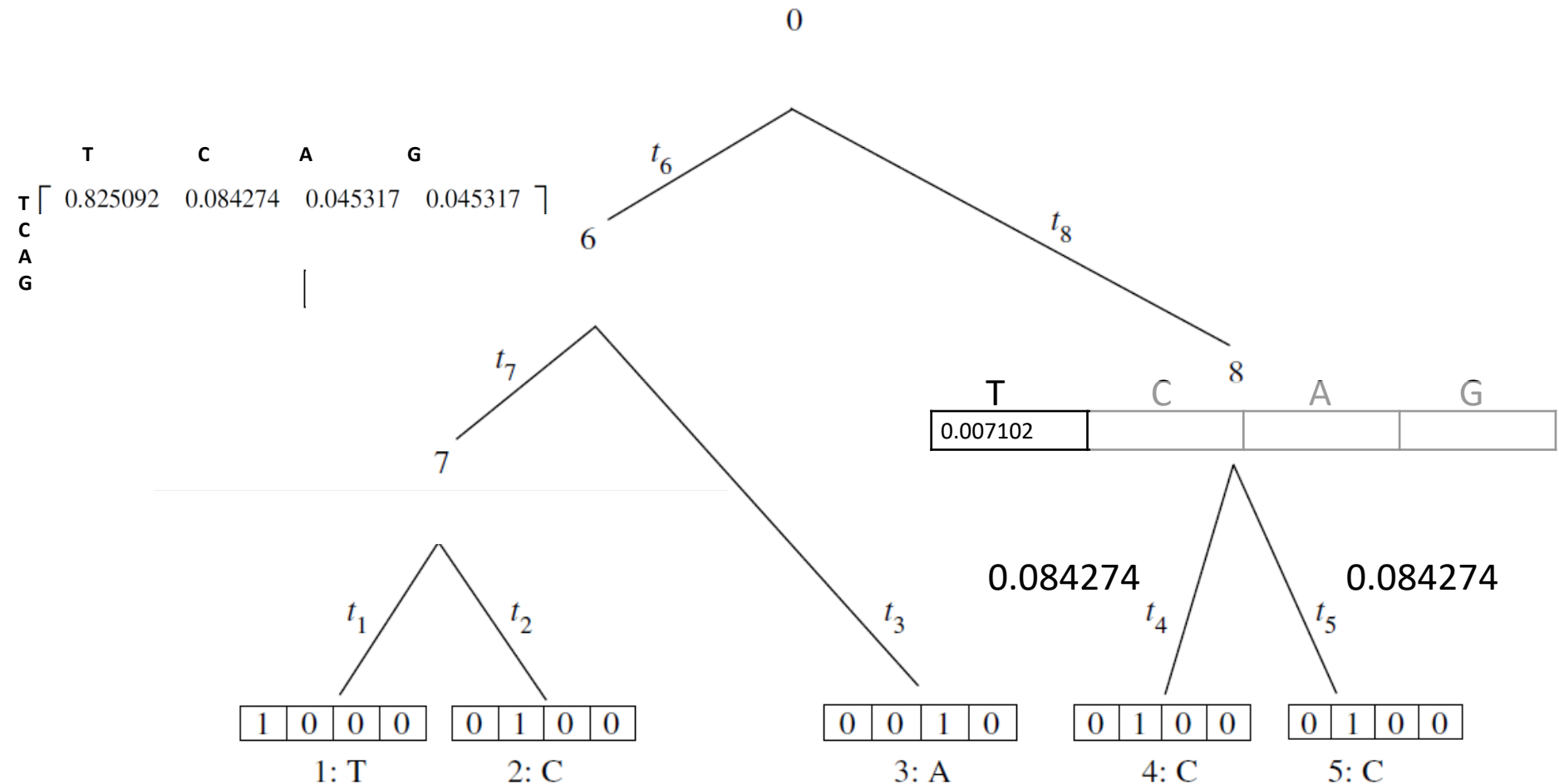


And now... Imagine that we know the topology of the tree and the parameters!

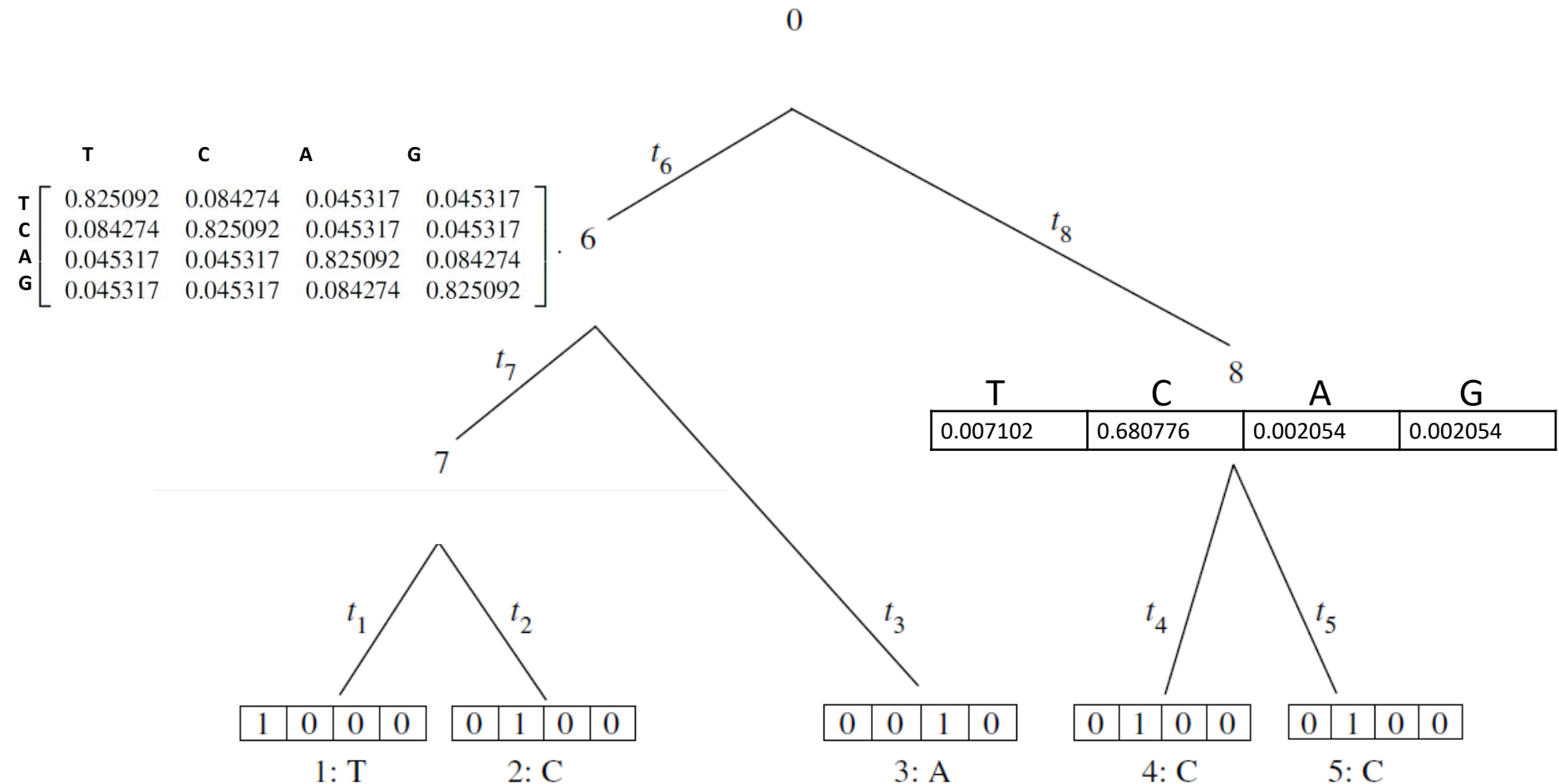




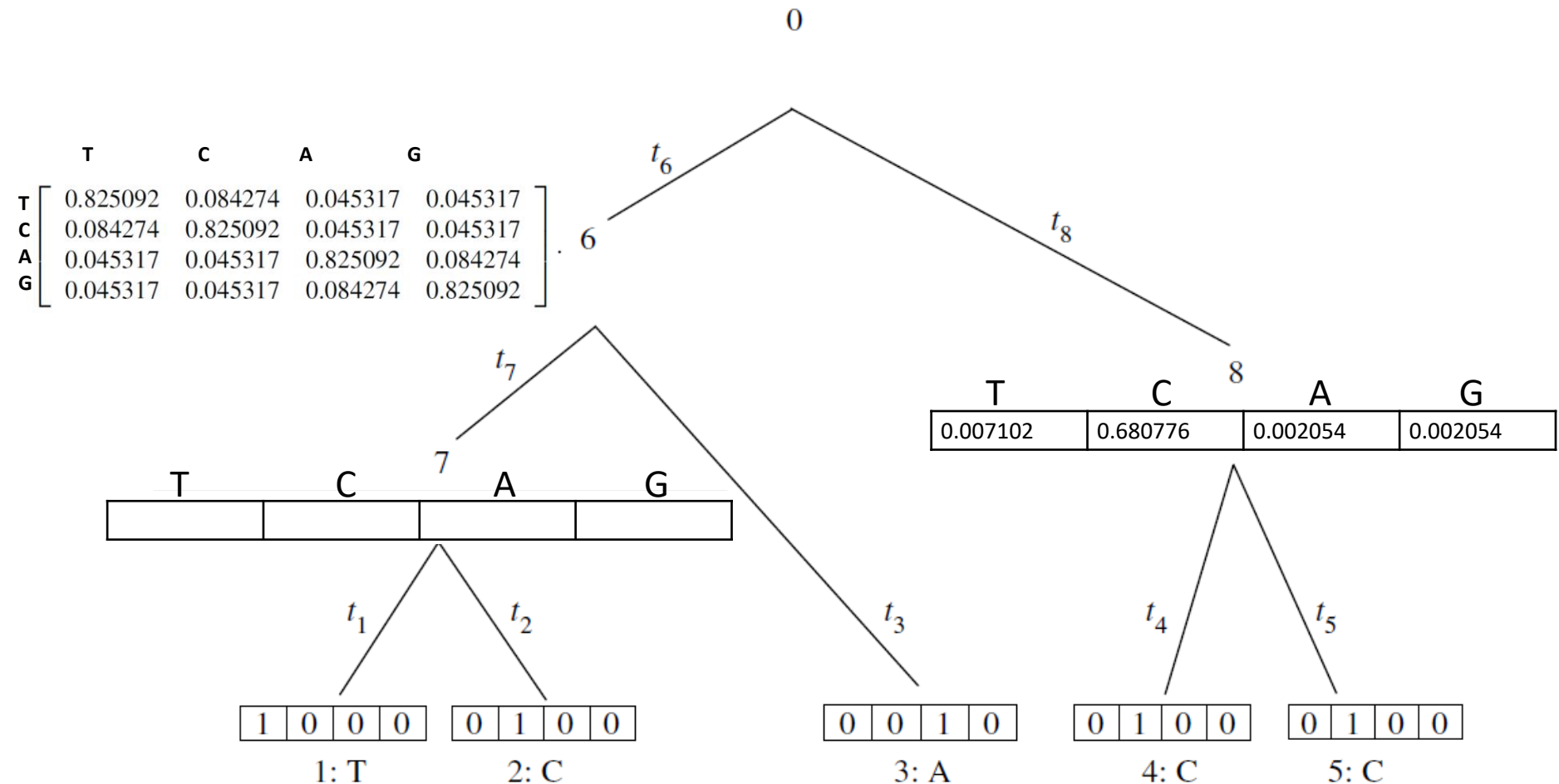
And now... Imagine that we know the topology of the tree and the parameters!



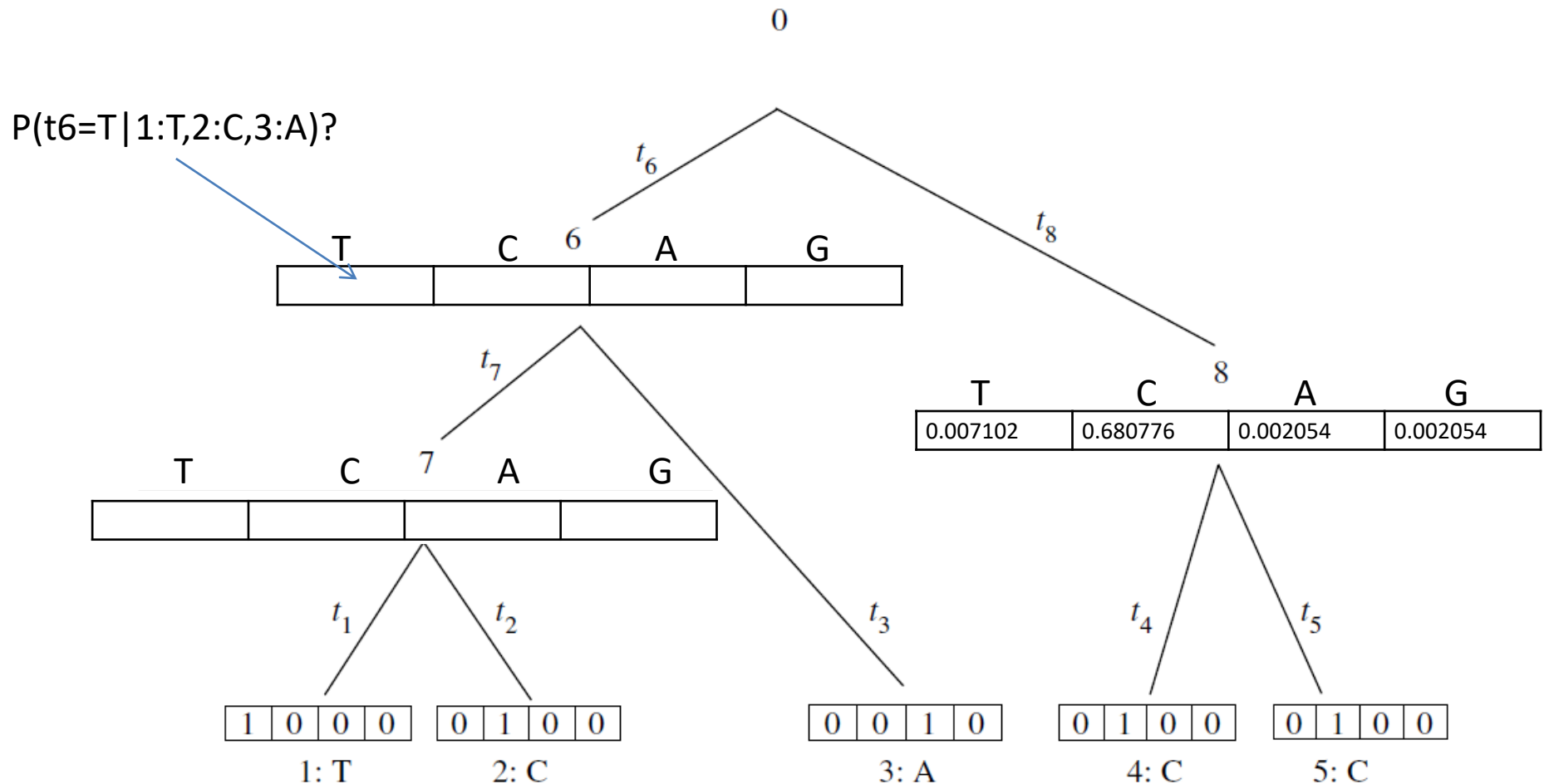
# And now... Imagine that we know the topology of the tree and the parameters!



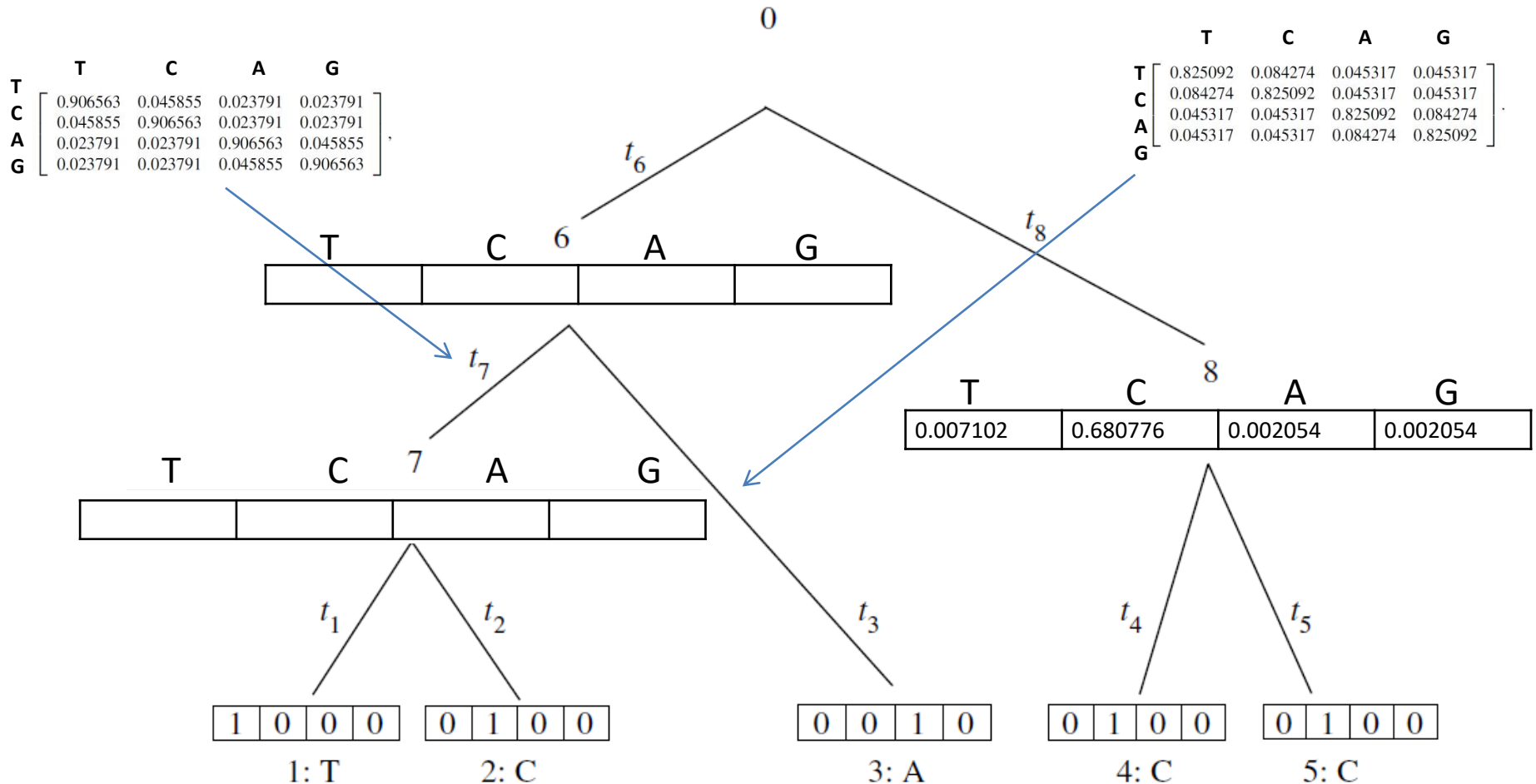
# And now... Imagine that we know the topology of the tree and the parameters!



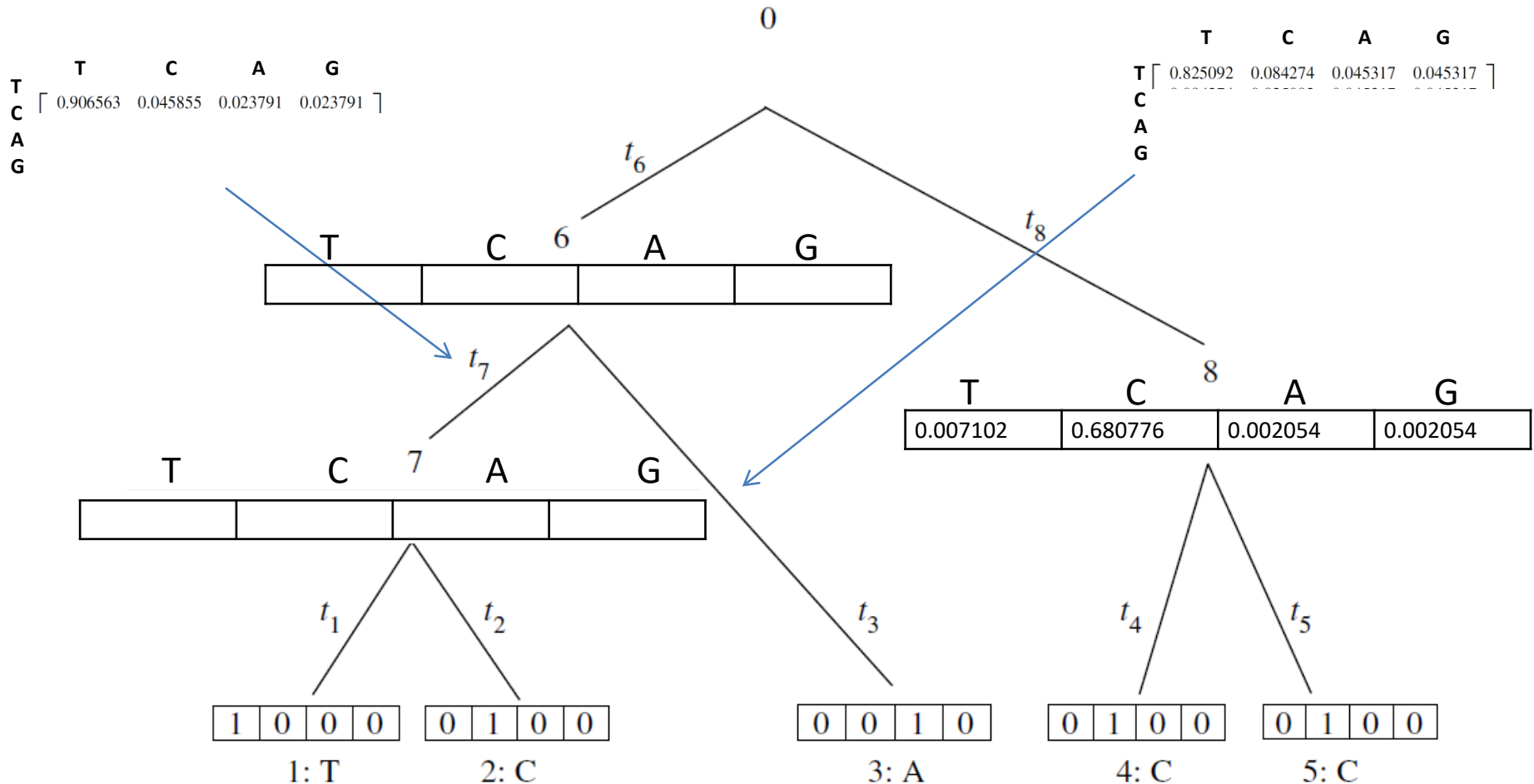
# And now... Imagine that we know the topology of the tree and the parameters!



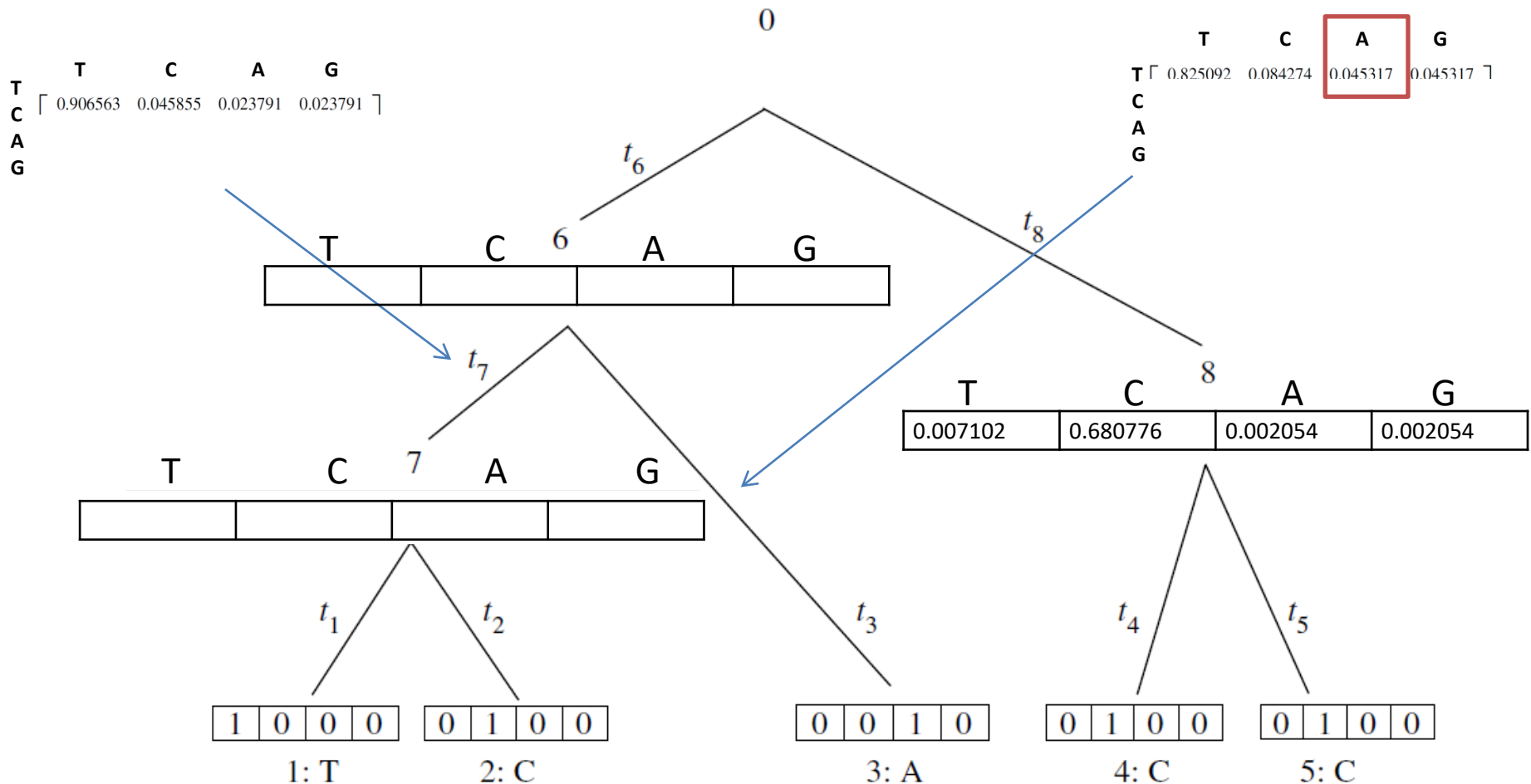
# And now... Imagine that we know the topology of the tree and the parameters!



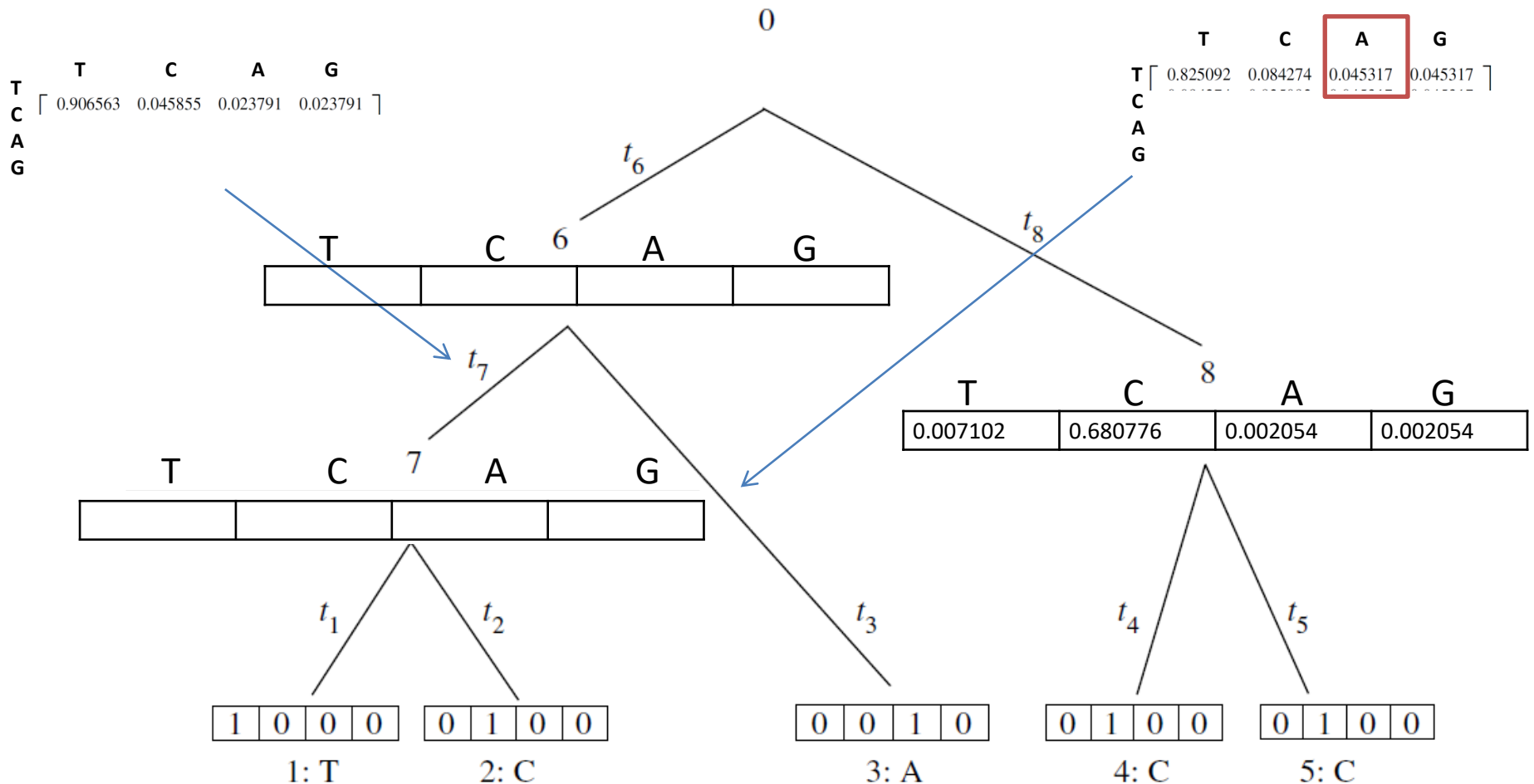
# And now... Imagine that we know the topology of the tree and the parameters!



# And now... Imagine that we know the topology of the tree and the parameters!



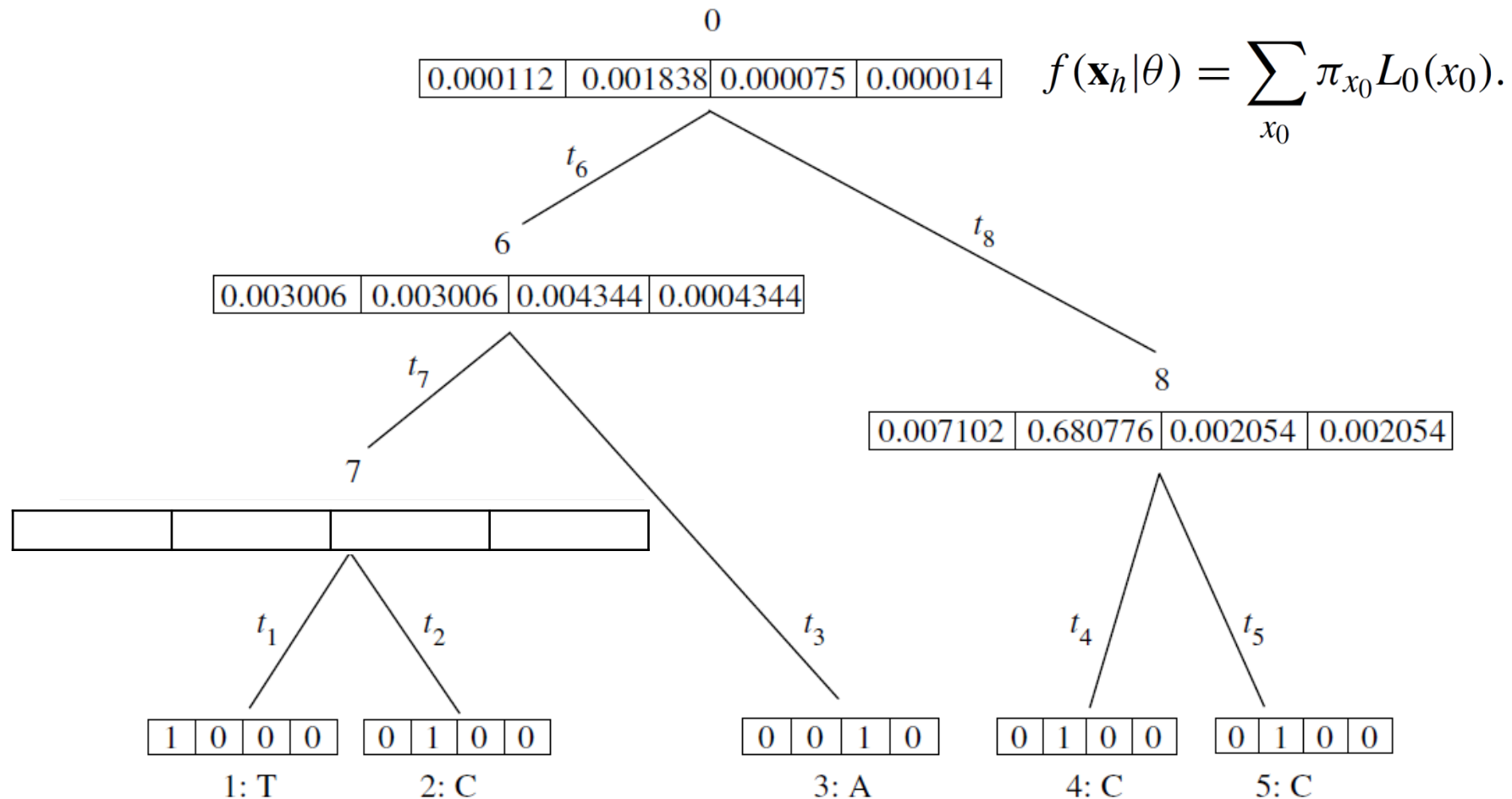
# And now... Imagine that we know the topology of the tree and the parameters!



$$P(t_6=T | 1:T, 2:C, 3:A) = ((0.906563 * 0.069533) + (0.045855 * 0.069533) + (0.023791 * 0.002054) + (0.023791 * 0.002054)) * 0.045317$$



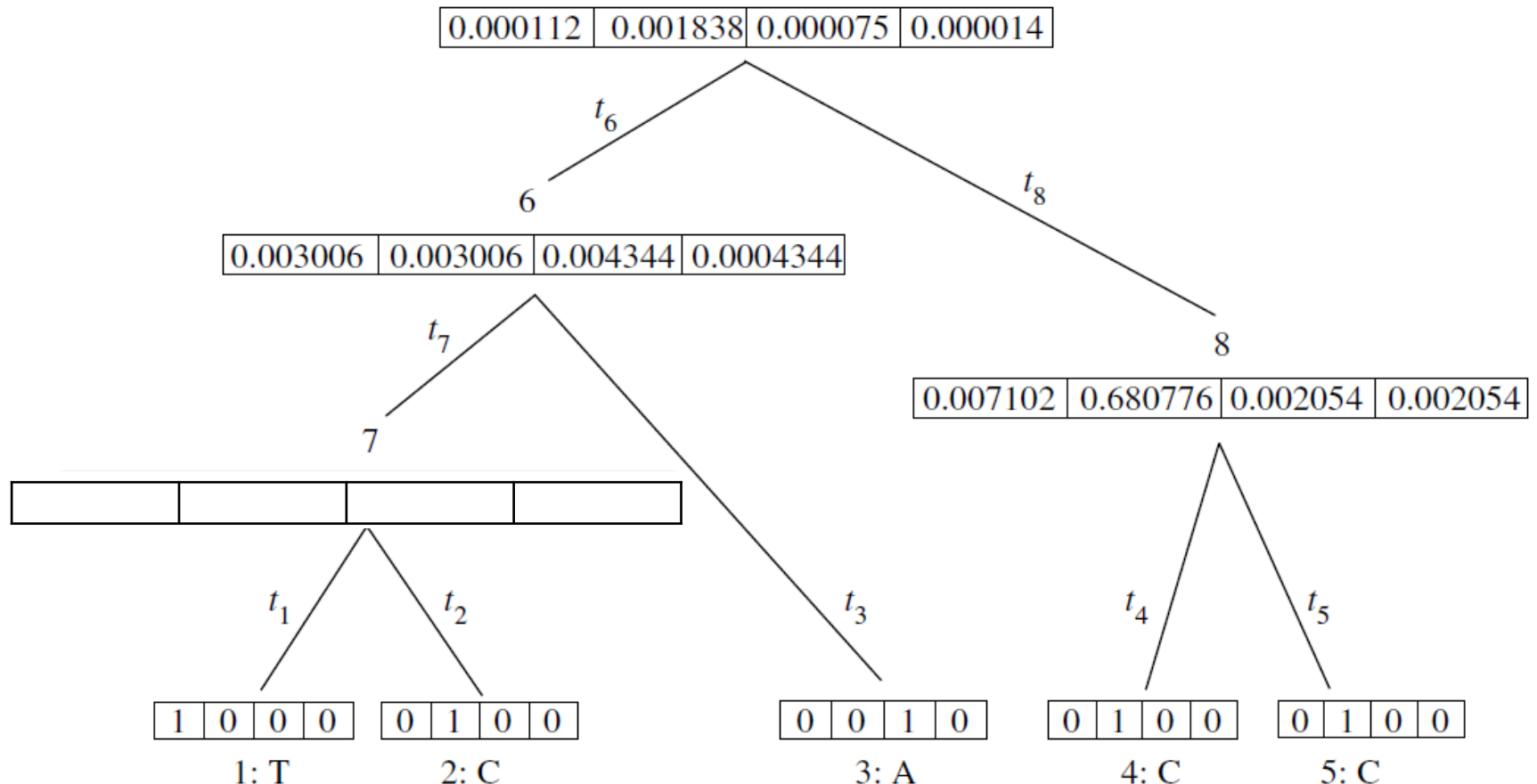
And now... Imagine that we know the topology of the tree and the parameters!



# And now... Imagine that we know the topology of the tree and the parameters!

$$f(X_h|\theta) = \frac{1}{4}(0.000112 + 0.001838 + 0.000075 + 0.000014)$$

0



# And now... Imagine that we know the topology of the tree

**$P(D|\theta)$**       *“What is the probability (likelihood) of observing the data given the values of the set of parameters”*

Which is the loglikelihood of all the positions given the length of the branches and my model of nucleotide evolution?

$$\ell = \log(L) = \sum_{h=1}^n \log\{f(\mathbf{x}_h|\theta)\}.$$

This is the parameter we want to maximize!

# What we want to do?

- Parameters
  - Branch length
  - Rates of substitution
- Topology
  - Relationship between OTUs

# What we want to do?

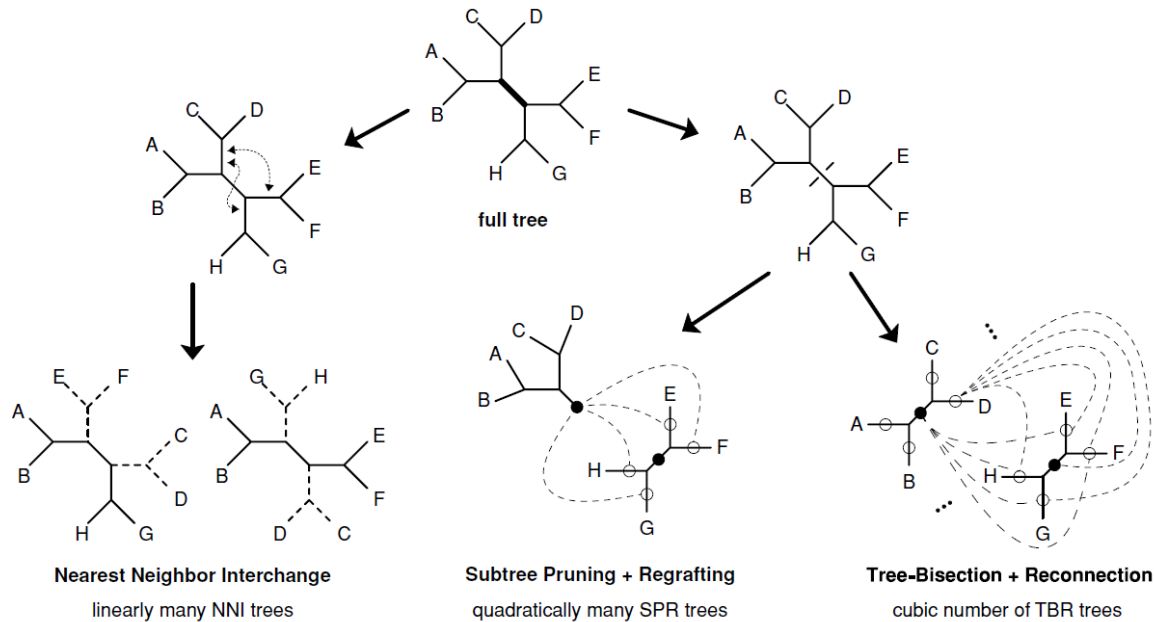
**Maximizing Parameters of a given tree**

**Propose a new tree**



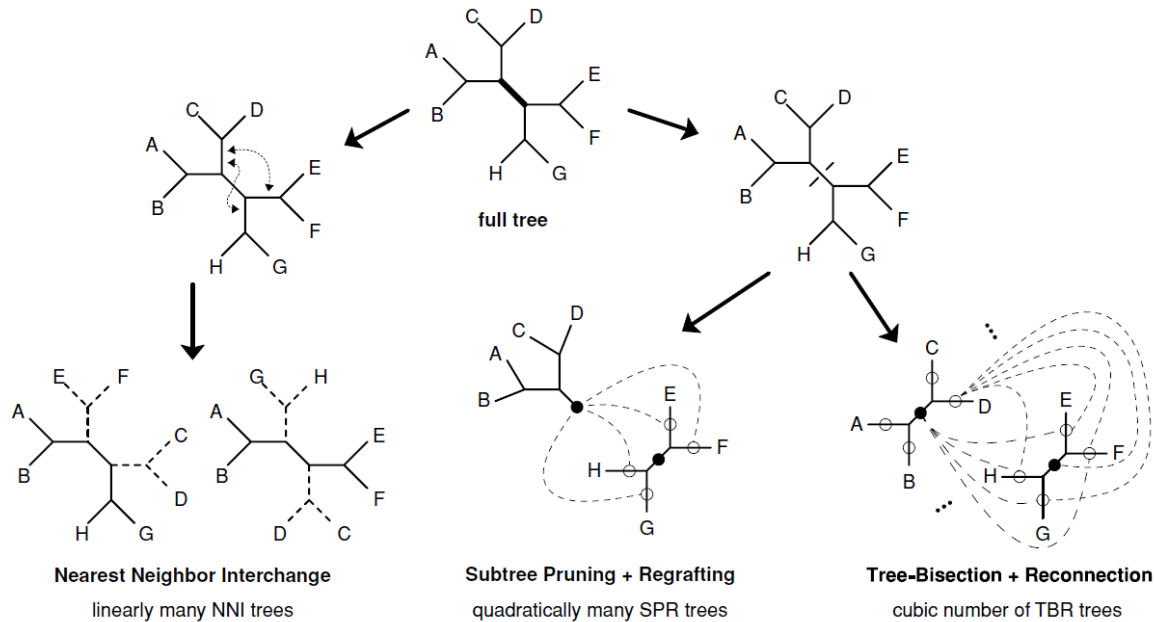
# What we want to do?

$$\ell = \log(L) = \sum_{h=1}^n \log\{f(\mathbf{x}_h|\theta)\}.$$



# What we want to do?

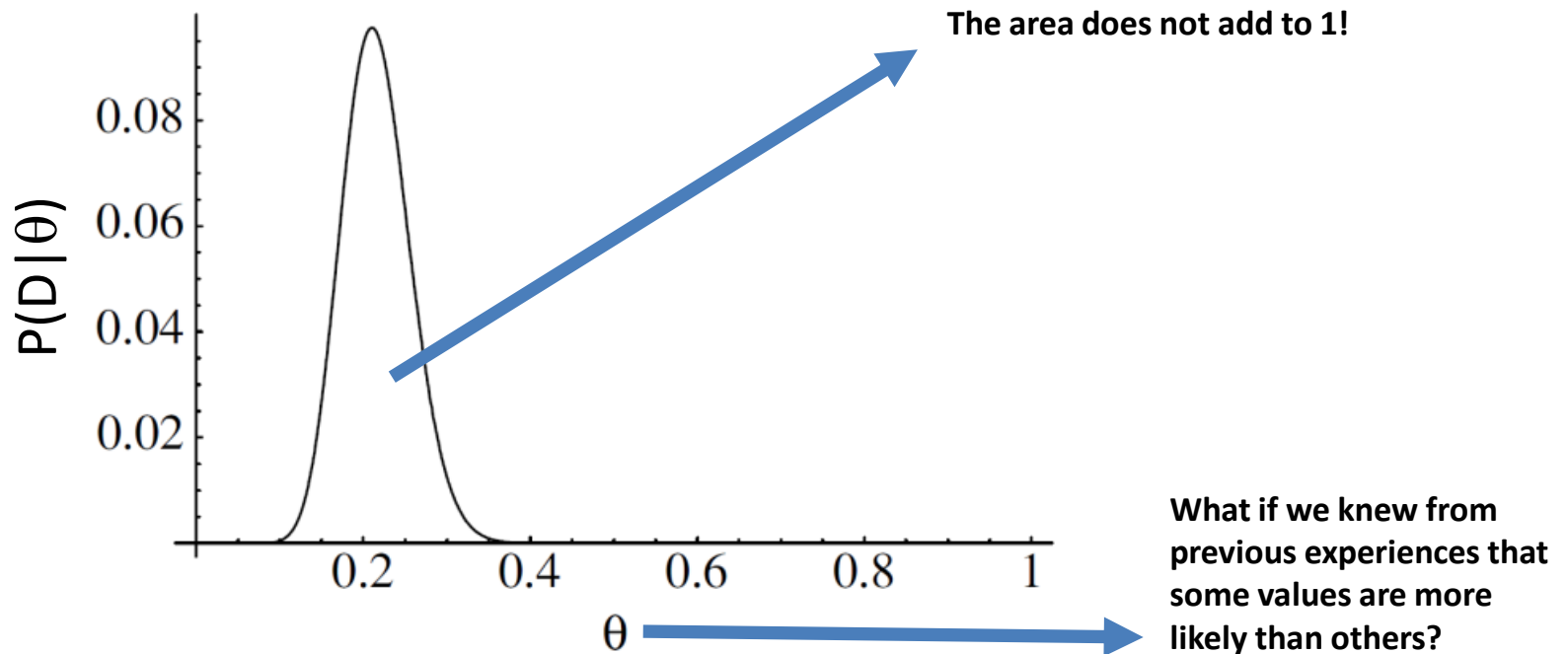
$$\ell = \log(L) = \sum_{h=1}^n \log\{f(\mathbf{x}_h|\theta)\}.$$



Which are the differences between maximum parsimony and maximum likelihood?

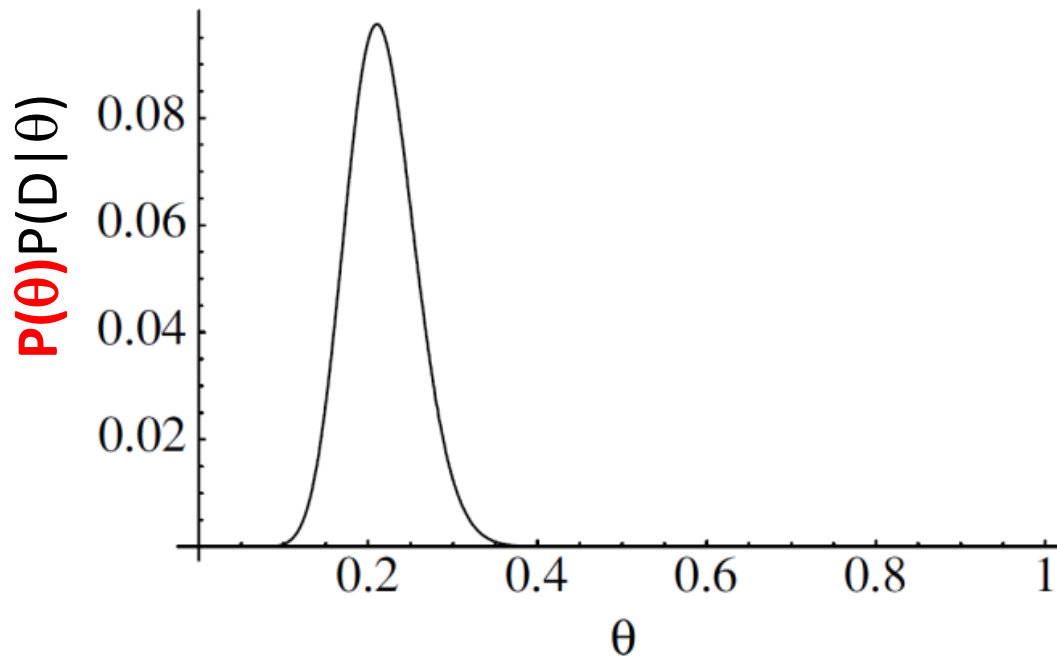


# Bayesian methods



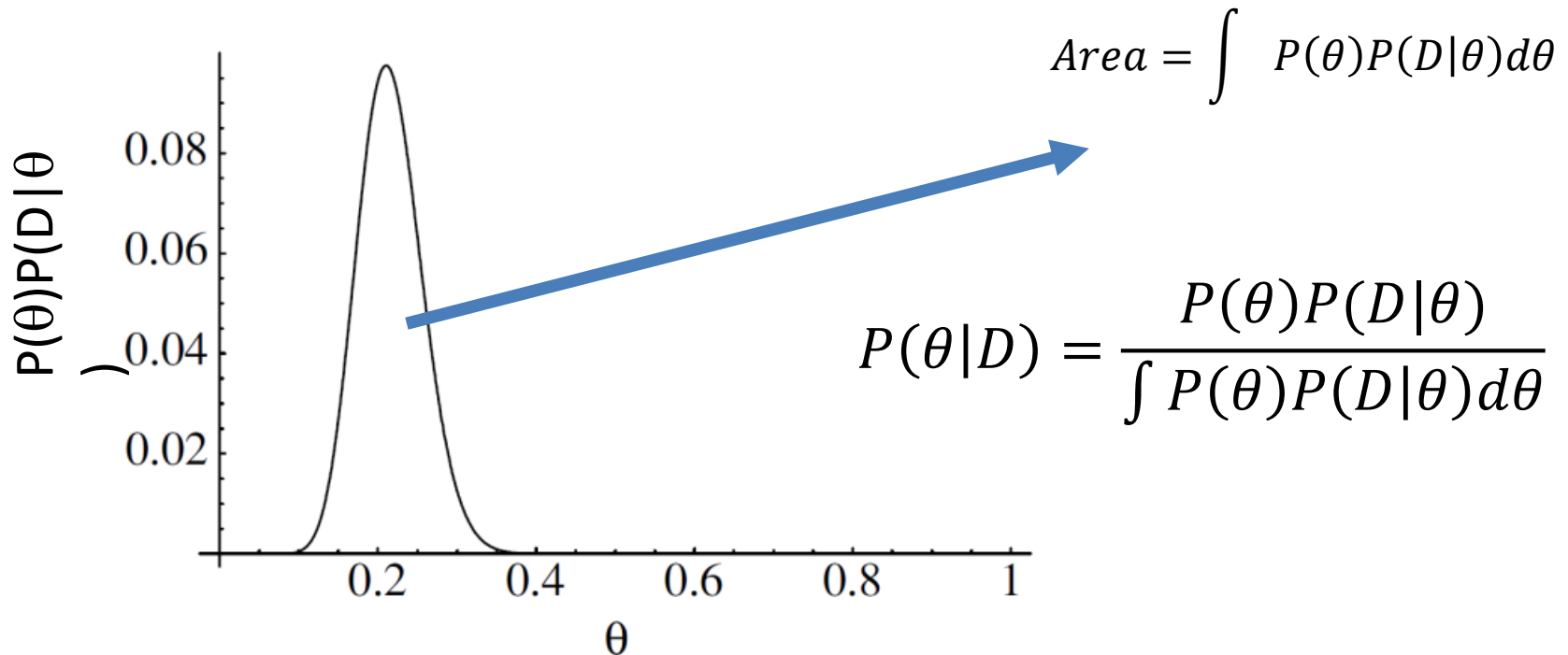
# Bayesian methods

What if we knew from previous experiences that some values are more likely than others?



# Bayesian methods

The area does not add to 1!



# Bayesian methods

## A formal way of expressing it...

$$P(D; \theta) = P(\theta)P(D|\theta) = P(D)P(\theta|D)$$

$$P(D) = \int P(\theta)P(D|\theta)d\theta$$

Prior knowledge  $\swarrow$

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

$\swarrow$  Posterior distribution of  $\theta$        $\searrow$  Normalizing constant

$\longrightarrow$  “our current knowledge is a mixture of what we previously knew updated with new data”

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta} \longrightarrow \text{“integrating over the marginals”}$$

# Bayesian methods

## Some extra definitions

		Topologies			Joint probabilities
		$\tau_A$	$\tau_B$	$\tau_C$	
Branch length vectors	$\mathbf{v}^A$	0.10	0.07	0.12	0.29
	$\mathbf{v}^B$	0.05	0.22	0.06	0.33
	$\mathbf{v}^C$	0.05	0.19	0.14	0.38
		0.20	0.48	0.32	Marginal probabilities

# Bayesian methods

Remember: Bayesian statistics

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}$$

$$\int P(\theta)P(D|\theta)d\theta$$

# Bayesian methods

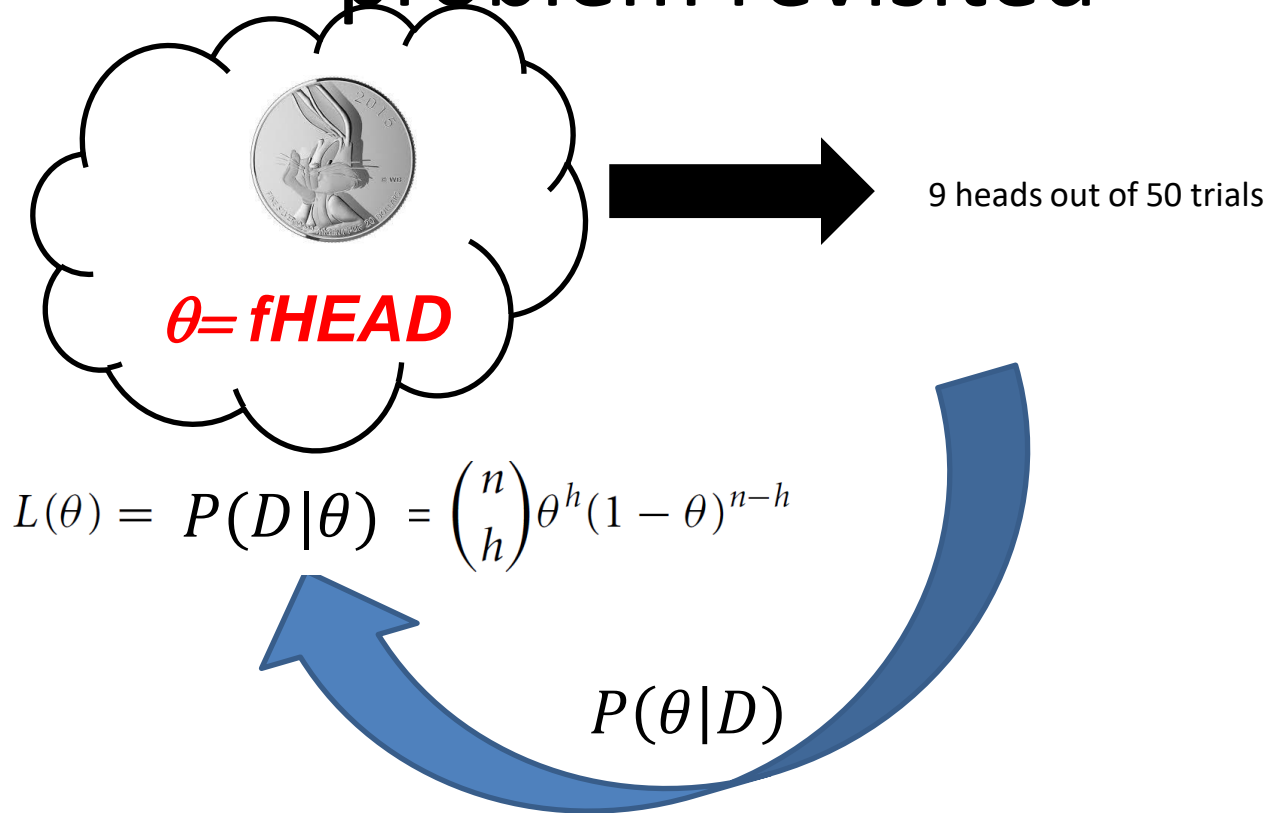
$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}$$


$$\int P(\theta)P(D|\theta)d\theta$$

**What if it does not exist?**

# Bayesian methods

## A continuous variable: The head problem revisited





# Bayesian methods

## How to get the Posterior?

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

Assume uniform prior knowledge ("the frequency can be any value between 0 and 1 with equal probability")

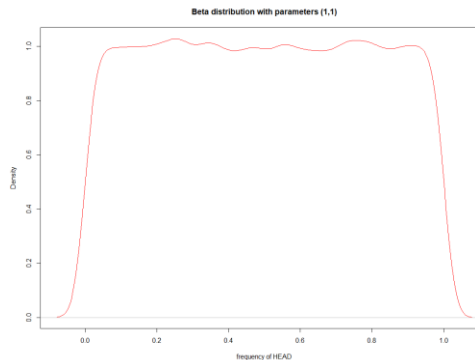
**Binomial Distribution**

$$P(y|\theta, n) = \frac{n!}{y!(n-y)!} \theta^y (1-\theta)^{n-y}$$

**Beta distribution**

$$P(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{\left( \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \right)}$$

$$\alpha = 1; \beta = 1$$



# Bayesian methods

Remember: Bayesian statistics

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = K \theta^{y+1-1} (1 - \theta)^{n-y+1-1}$$

**Beta distribution**  $P(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1} (1 - \theta)^{\beta-1}}{\left( \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \right)}$

“Beta distribution is a **conjugate** of the Binomial distribution”

# Bayesian methods

Remember: Bayesian statistics

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = K \theta^{y+1-1} (1-\theta)^{n-y+1-1}$$

**Beta distribution**  $P(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{\left( \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \right)}$

$$P(\theta|D) = B(y+1, n-y+1) \quad \text{Mean}(B(y+1, n-y+1)) = \frac{y+1}{n+2}$$

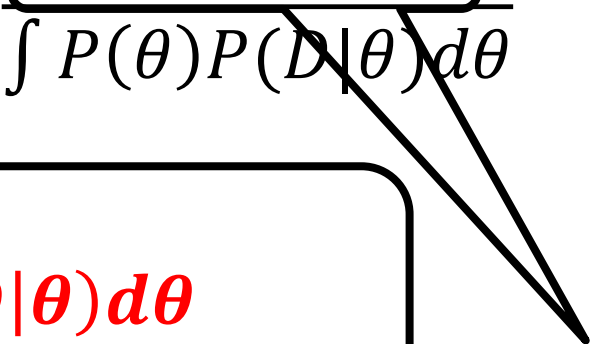
# Bayesian methods

Remember: Bayesian statistics

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}$$

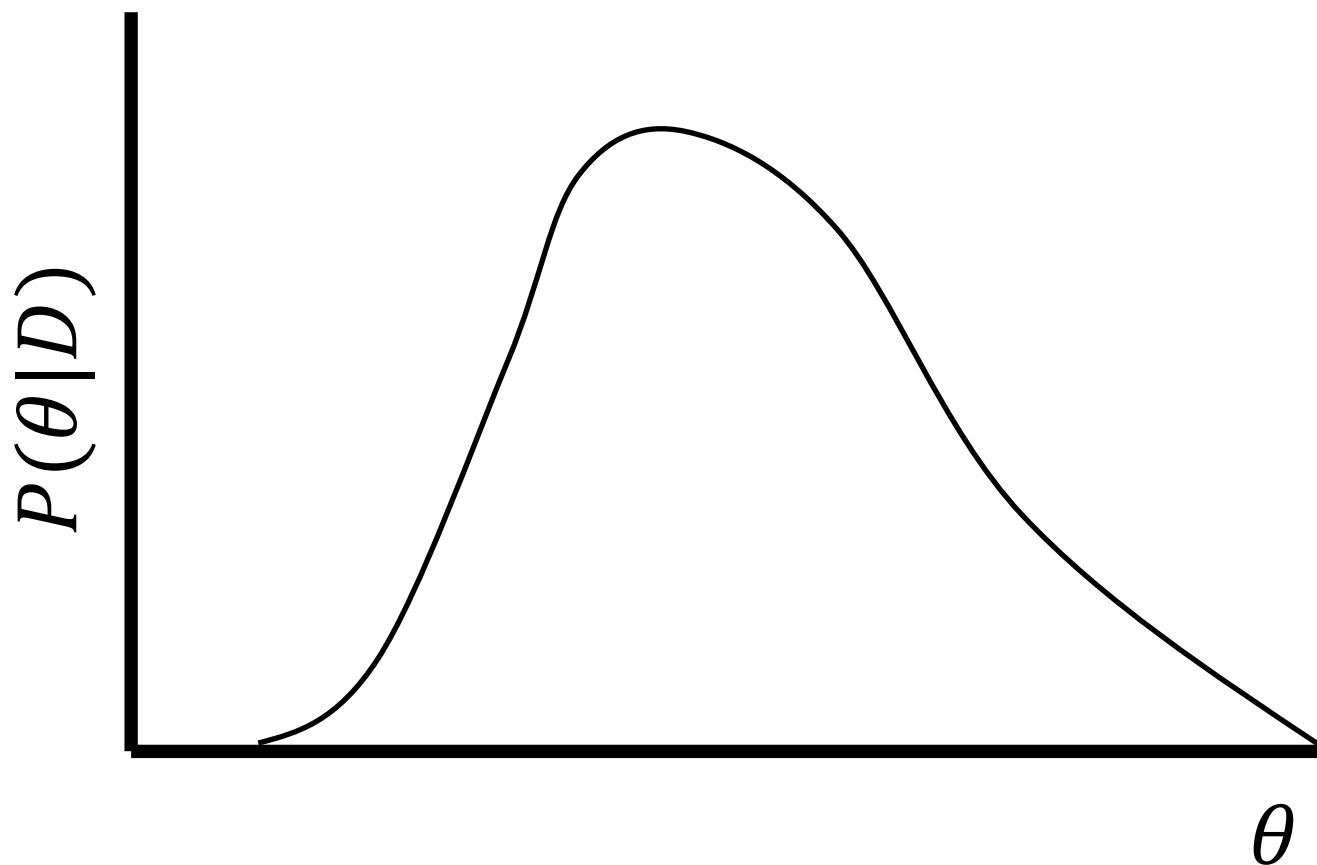

$$\int P(\theta)P(D|\theta)d\theta$$

What if it does not exist?



What if  
there is no  
conjugate?

# Bayesian methods

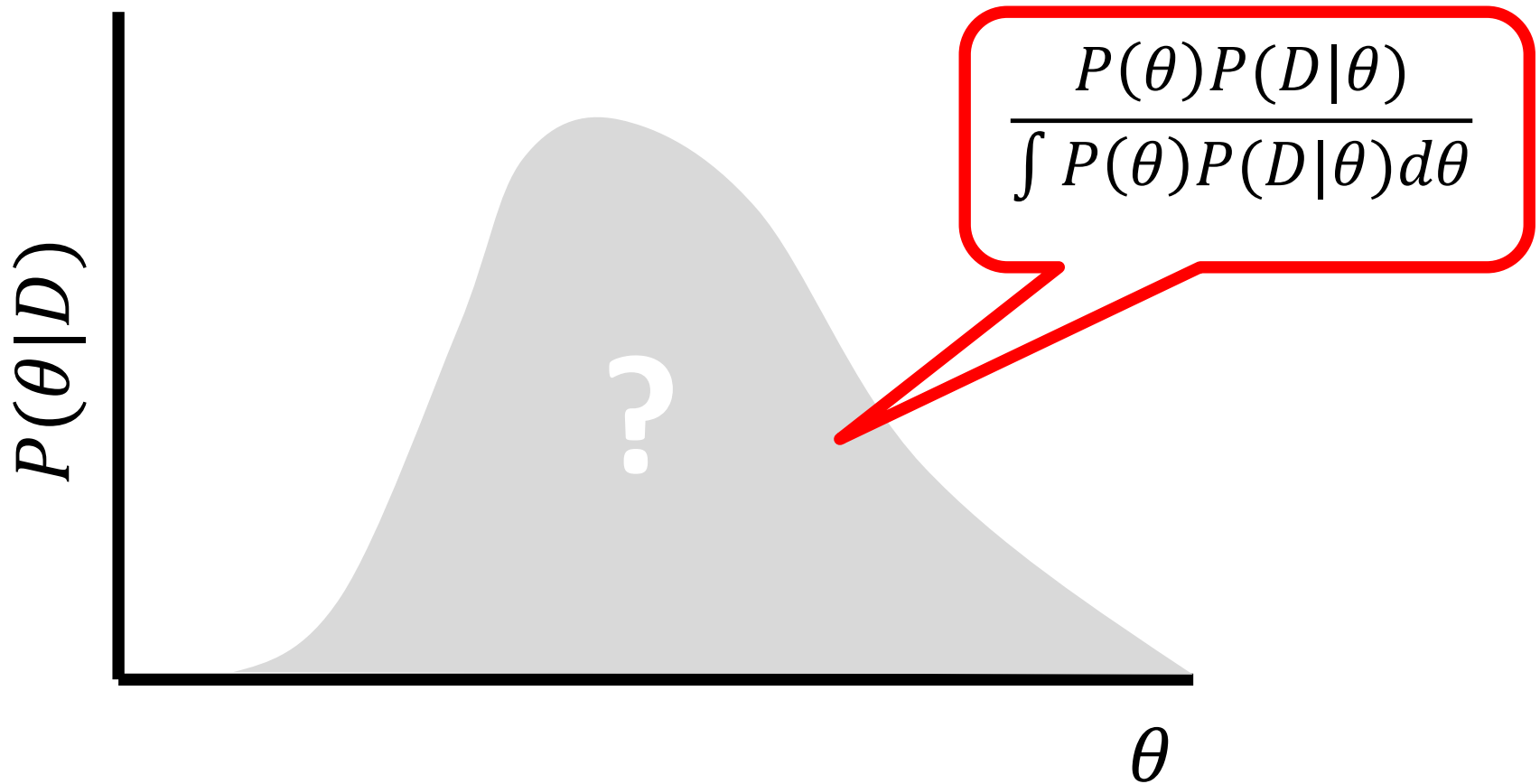


# Bayesian methods

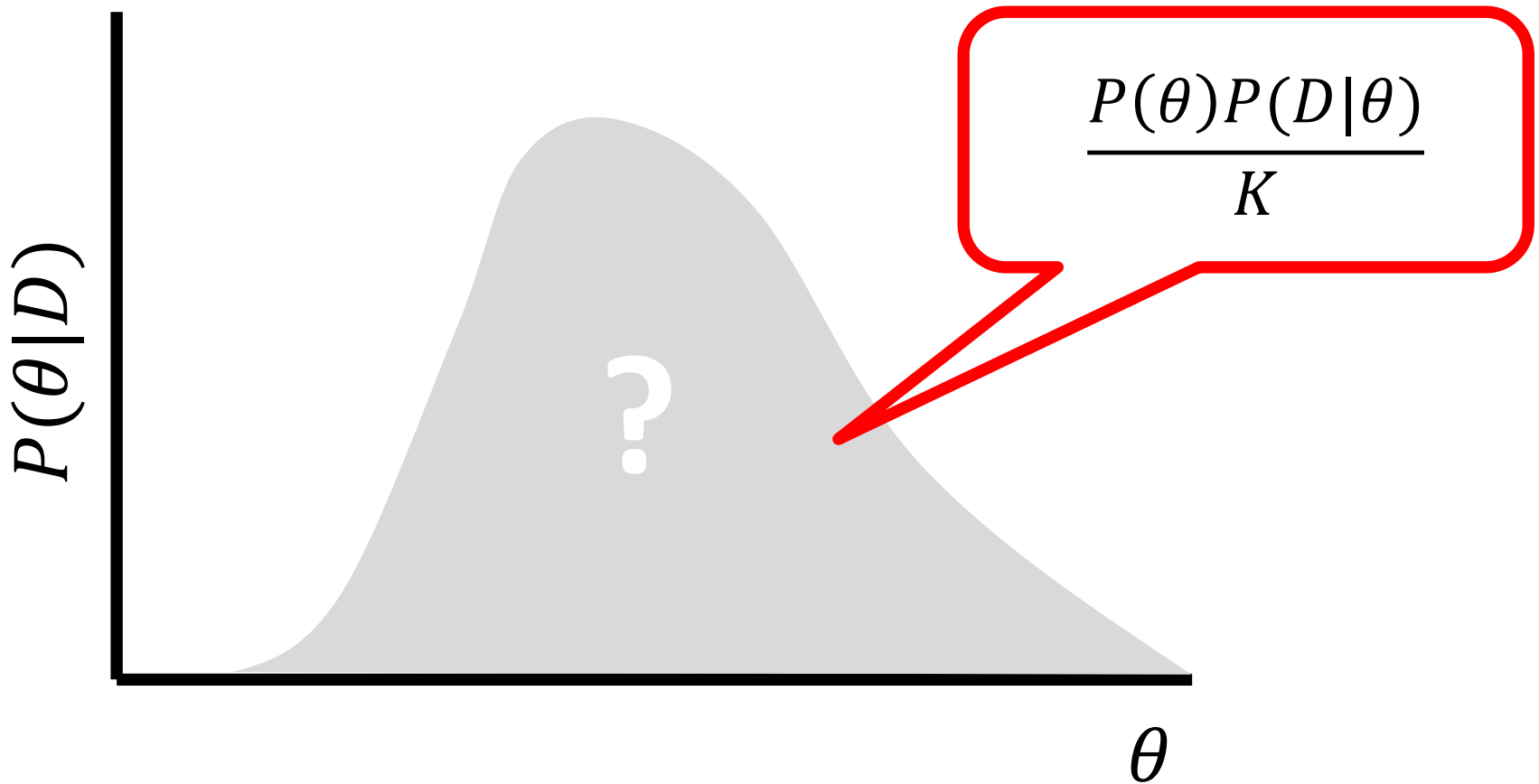
At least, is there any way we can generate samples from this (unknown) posterior distribution so we can broadly regenerate the (**unknown**) posterior distribution?



# Bayesian methods



# Bayesian methods





# Bayesian methods

How many times is more likely  $j$  than  $i$  in this (unknown) distribution?

$$r(i, j) = \frac{\frac{P(\theta = i)P(D|\theta = i)}{K}}{\frac{P(\theta = j)P(D|\theta = j)}{K}} = \frac{P(\theta = i)P(D|\theta = i)}{P(\theta = j)P(D|\theta = j)} = \frac{\pi(j)}{\pi(i)}$$

# Bayesian methods

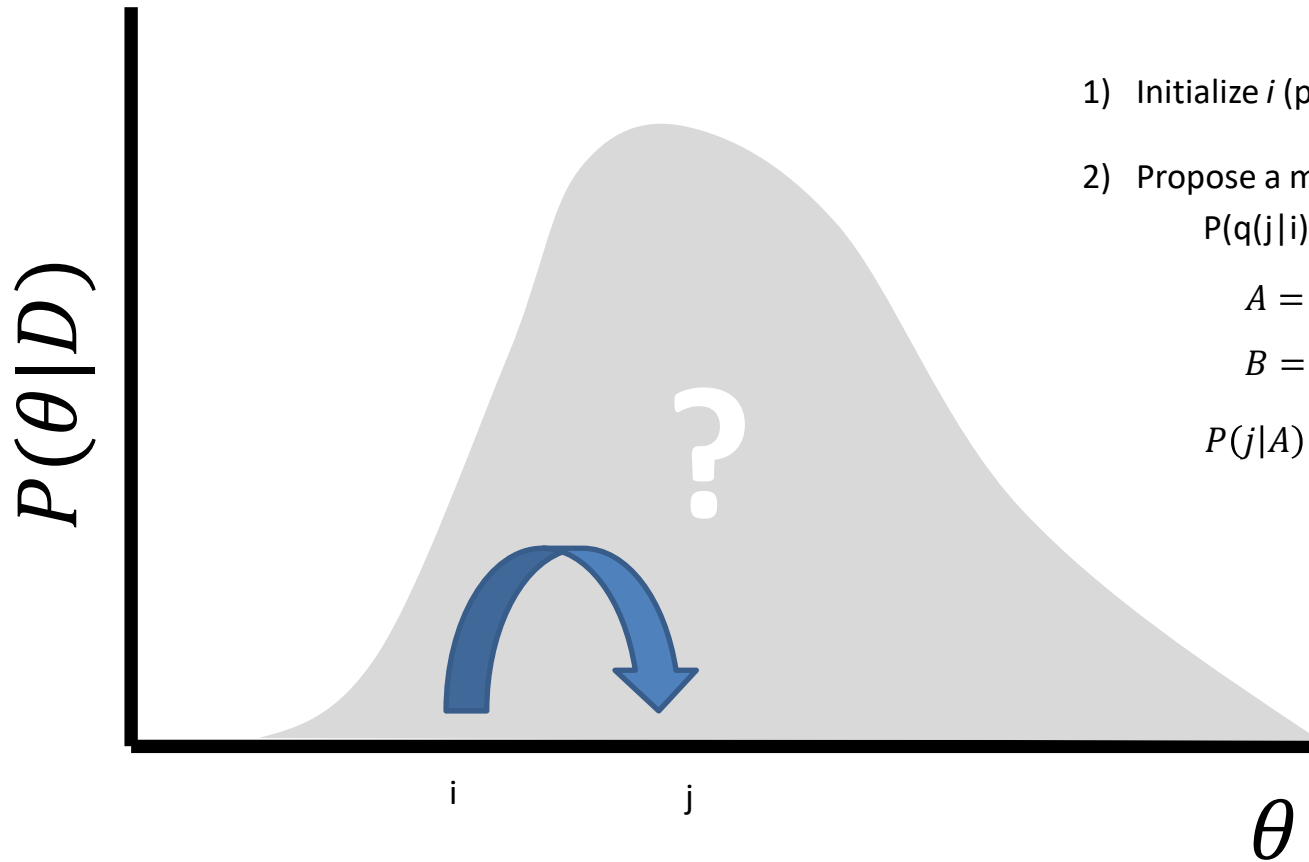
## Markov Chain Monte Carlo Methods: Metropolis algorithm



1) Initialize  $i$  (pick a value of  $\theta$  at random)

# Bayesian methods

## Markov Chain Monte Carlo Methods: Metropolis algorithm



1) Initialize  $i$  (pick a value of  $\theta$  at random)

2) Propose a movement  $q(\cdot|i)$

$$P(q(j|i))=P(q(i|j))$$

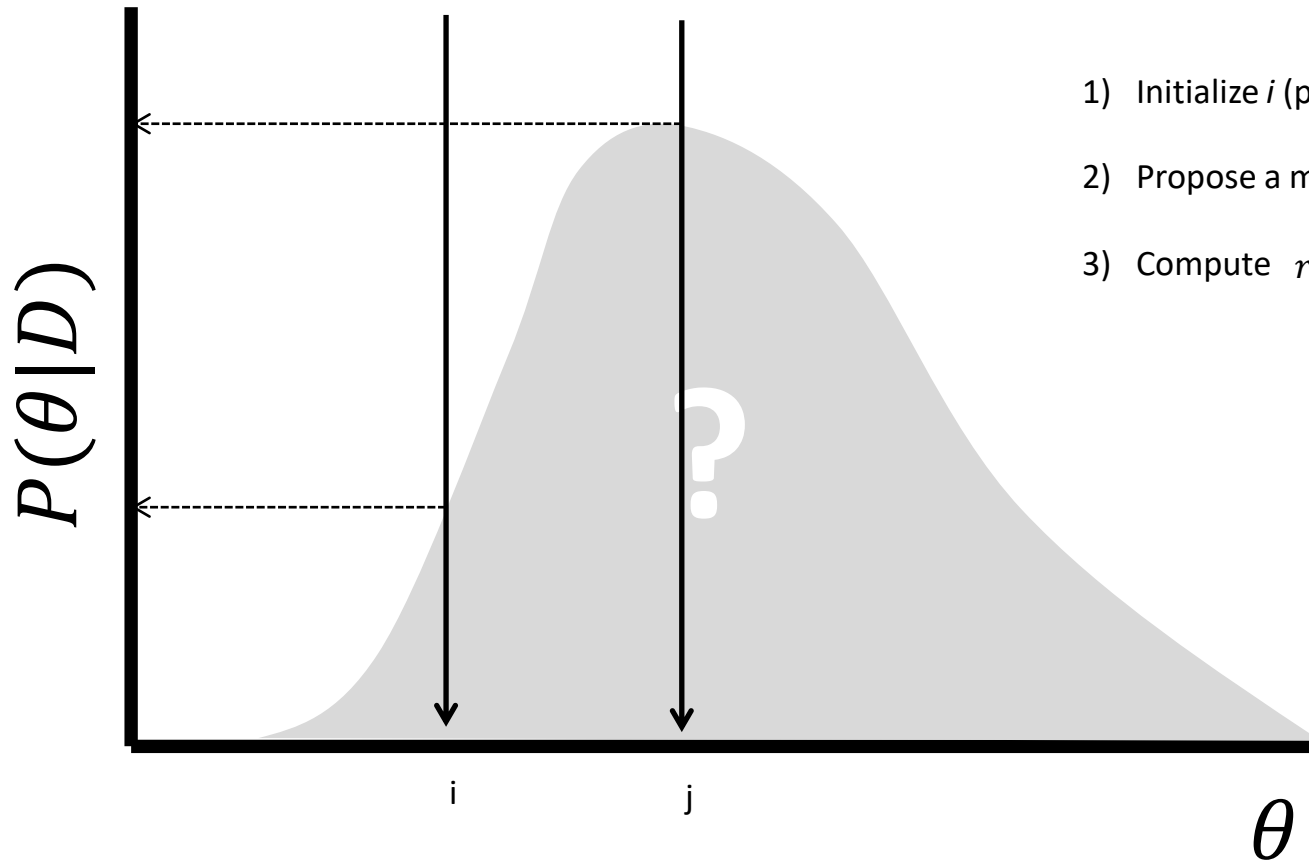
$$A = N(i, \sigma)$$

$$B = N(j, \sigma)$$

$$P(j|A) = p(i|B)$$

# Bayesian methods

## Markov Chain Monte Carlo Methods: Metropolis algorithm



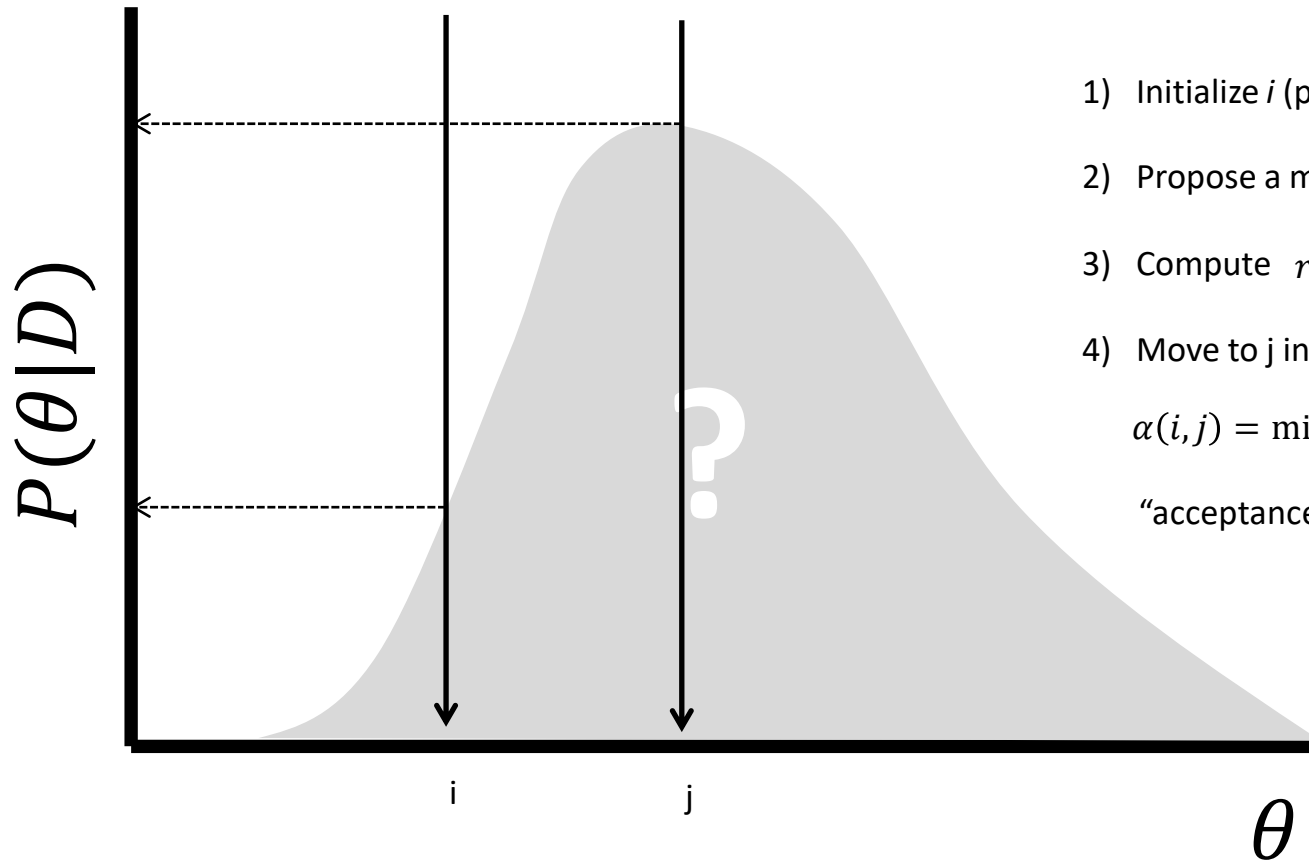
1) Initialize  $i$  (pick a value of  $\theta$  at random)

2) Propose a movement  $q(\cdot | i)$

3) Compute  $r(i, j) = \frac{\pi(j)}{\pi(i)}$

# Bayesian methods

## Markov Chain Monte Carlo Methods: Metropolis algorithm



1) Initialize  $i$  (pick a value of  $\theta$  at random)

2) Propose a movement  $q(\cdot | i)$

3) Compute  $r(i, j) = \frac{\pi(j)}{\pi(i)}$

4) Move to  $j$  in probability

$$\alpha(i, j) = \min(1, r(i, j))$$

“acceptance ratio”

# Bayesian methods

## Markov Chain Monte Carlo Methods: **Metropolis-Hastings**



1) Initialize  $i$  (pick a value of  $\theta$  at random)

2) Propose a movement  $q(\cdot|i)$

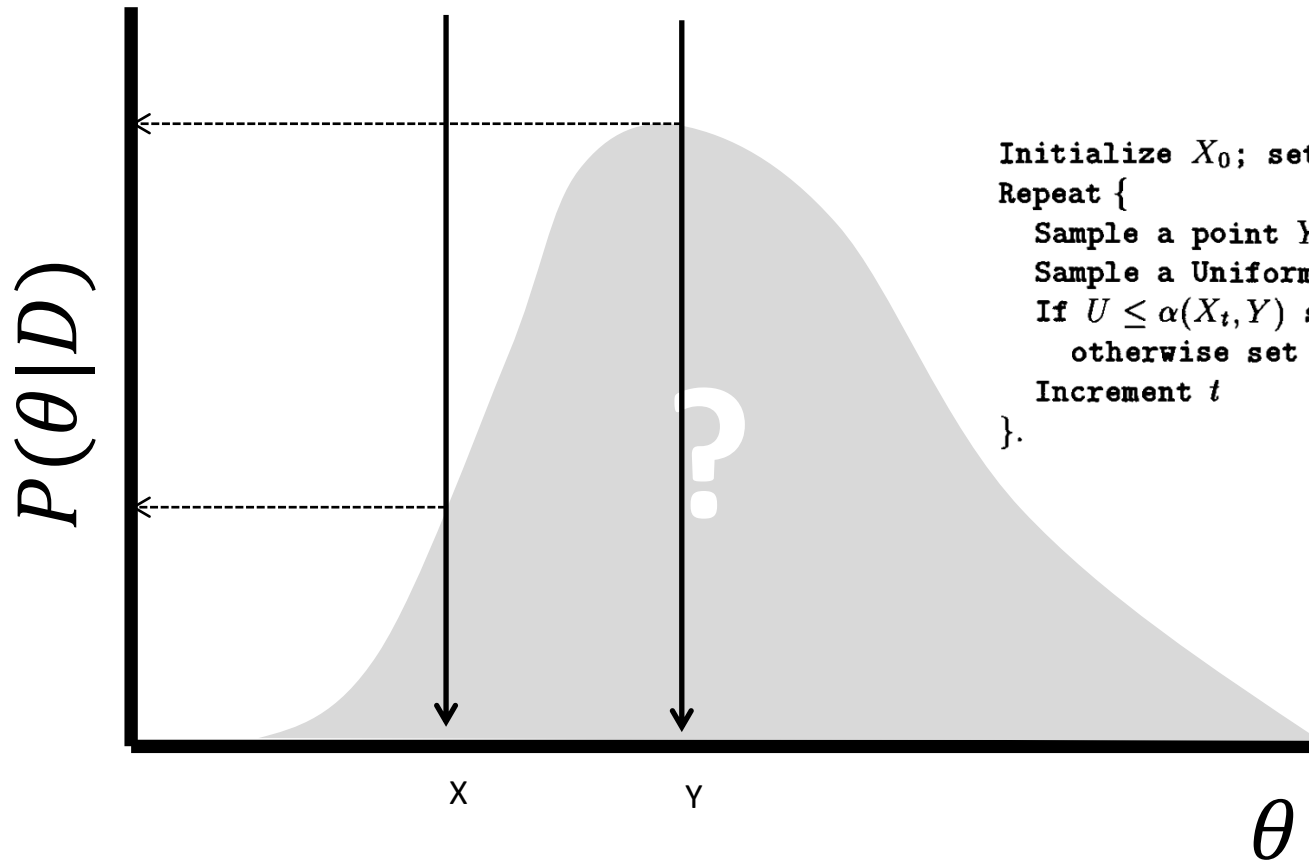
3) Compute  $r(i,j) = \frac{\pi(j)q(i|j)}{\pi(i)q(j|i)}$

4) Move to  $j$  in probability

$$\alpha(i,j) = \min(1, r(i,j))$$

# Bayesian methods

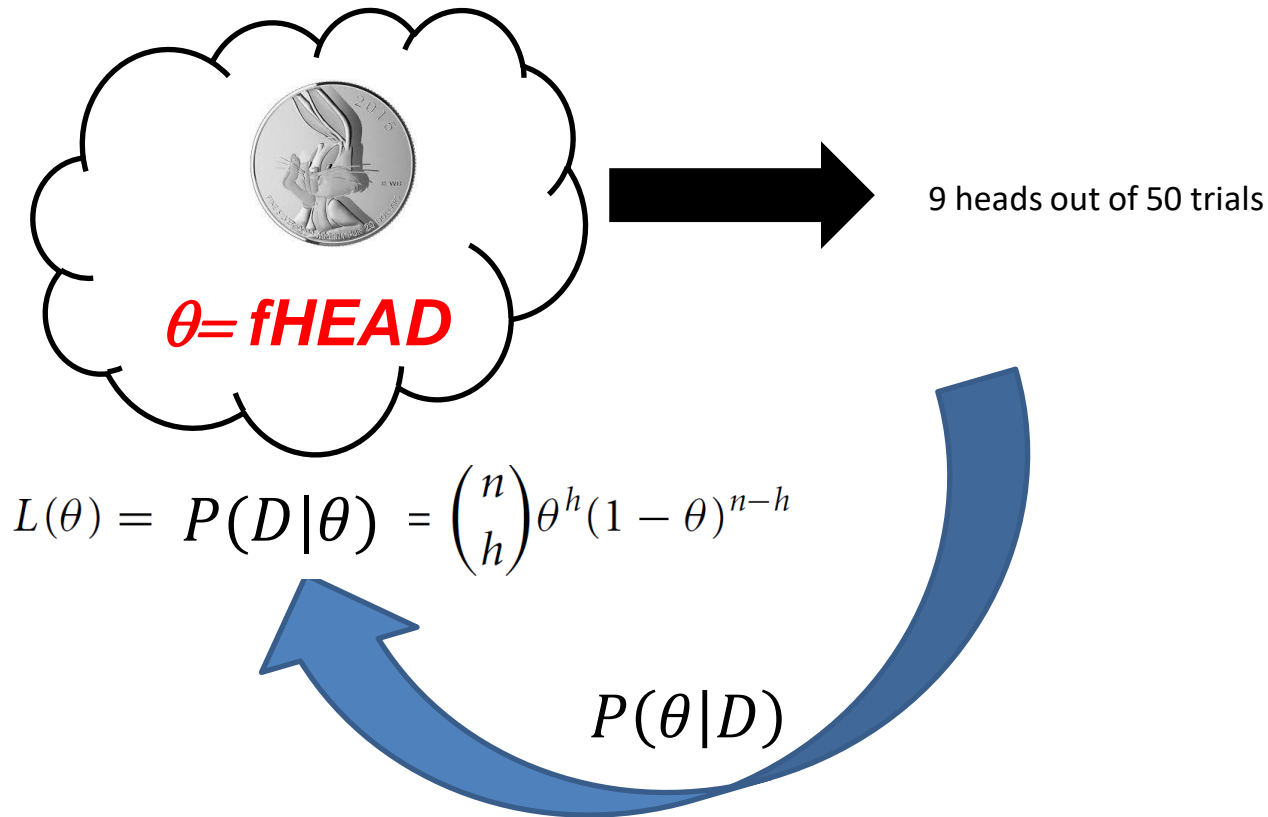
## Markov Chain Monte Carlo Methods: Metropolis-Hastings



```
Initialize  $X_0$ ; set  $t = 0$ .  
Repeat {  
  Sample a point  $Y$  from  $q(.|X_t)$   
  Sample a Uniform(0,1) random variable  $U$   
  If  $U \leq \alpha(X_t, Y)$  set  $X_{t+1} = Y$   
  otherwise set  $X_{t+1} = X_t$   
  Increment  $t$   
}.
```

# Bayesian methods

## The head problem revisited





# Bayesian methods

## Example of the Metropolis algorithm at work!

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{U(0,1)Binom(h|\theta, n)}{?}$$

$$Binom(h|\theta, n) = \frac{n!}{h! (n-h)!} \theta^h (1-\theta)^{n-h}$$

# Bayesian methods

## Example of the Metropolis algorithm at work!

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{U(0,1)Binom(h|\theta, n)}{?}$$

$$Binom(h|\theta, n) = \frac{n!}{h! (n-h)!} \theta^h (1-\theta)^{n-h} \quad \begin{array}{l} h = 9; \\ n = 50 \end{array}$$

$q(.|X_t) = N(X_t, 1)$

```
Initialize  $X_0$ ; set  $t = 0$ .  
Repeat {  
  Sample a point  $Y$  from  $q(.|X_t)$   
  Sample a Uniform(0,1) random variable  $U$   
  If  $U \leq \alpha(X_t, Y)$  set  $X_{t+1} = Y$   
  otherwise set  $X_{t+1} = X_t$   
  Increment  $t$   
}.
```


# Bayesian methods

## Example of the Metropolis algorithm at work!

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{U(0,1)Binom(h|\theta, n)}{?}$$

$$Binom(h|\theta, n) = \frac{n!}{h!(n-h)!} \theta^h (1-\theta)^{n-h} \quad \begin{array}{l} h = 9; \\ n = 50 \end{array}$$

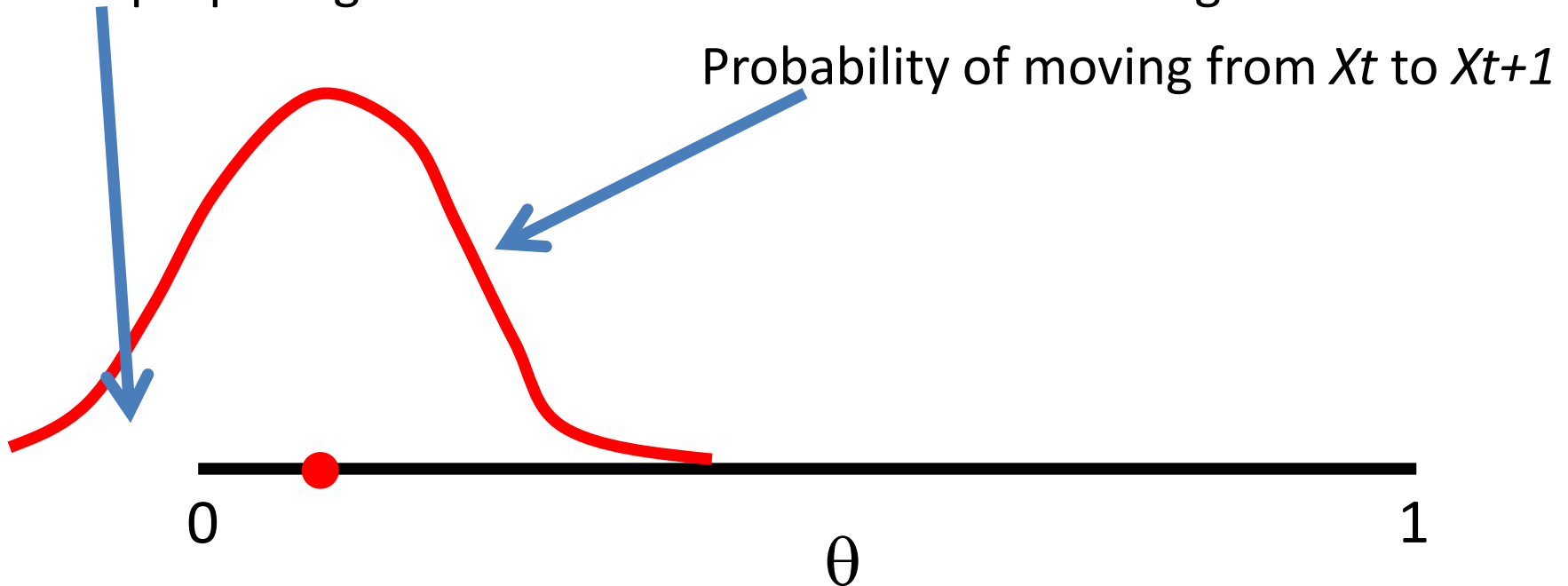
$q(.|X_t) = N(X_t, 1)$

```
Initialize  $X_0$ ; set  $t = 0$ .   $\theta = [0,1]$   
Repeat {  
  Sample a point  $Y$  from  $q(.|X_t)$   
  Sample a Uniform(0,1) random variable  $U$   
  If  $U \leq \alpha(X_t, Y)$  set  $X_{t+1} = Y$   
  otherwise set  $X_{t+1} = X_t$   
  Increment  $t$   
}.
```

# Bayesian methods

## A potential problem with this variable definition...

I am proposing movements to values out of the range!!!!



# Bayesian methods

## Example of the Metropolis algorithm at work!

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{U(0,1)Binom(h|\theta, n)}{?}$$

$$Binom(h|\theta, n) = \frac{n!}{h!(n-h)!} \theta^h (1-\theta)^{n-h}$$

$$h = 9;$$

$$n = 50$$

$$q(.|X_t) = N(X_t, 1)$$

Initialize  $X_0$ ; set  $t = 0$ .

Repeat {

Sample a point  $Y$  from  $q(.|X_t)$

Sample a Uniform(0,1) random variable  $U$

If  $U \leq \alpha(X_t, Y)$  set  $X_{t+1} = Y$

otherwise set  $X_{t+1} = X_t$

Increment  $t$

}.

$$\zeta = \log \left( \frac{\theta}{1-\theta} \right);$$

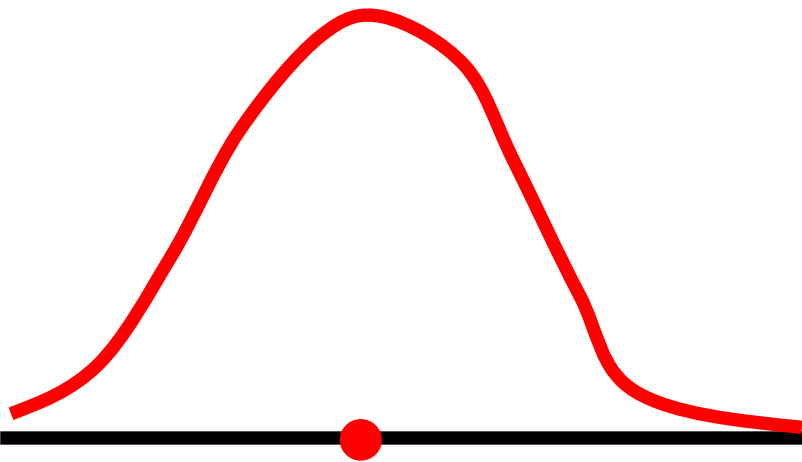
$$\theta = [0,1];$$

$$\zeta = (-\infty, \infty)$$

# Bayesian methods

Now there is no problem anymore!

$$\theta = \frac{e^{\zeta}}{1 + e^{\zeta}}$$



$\zeta$


# Bayesian methods

## Example of the Metropolis algorithm at work!

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{U(0,1)Binom(h|\theta, n)}{?}$$

$$Binom(h|\theta, n) = \frac{n!}{h!(n-h)!} \theta^h (1-\theta)^{n-h} \quad \begin{array}{l} h = 9; \\ n = 50 \end{array}$$

$$q(.|X_t) = N(X_t, 1)$$

Initialize  $X_0$ ; set  $t = 0$ .   $\zeta_0 = 0; \theta_0 = 0.5; P(9|0.5, 50) = 2.2e-06$

Repeat {

  Sample a point  $Y$  from  $q(.|X_t)$      $Y = -0.46; Y_\theta = 0.38; P(9|0.38, 50) = 0.0012$

  Sample a Uniform(0,1) random variable  $U$

  If  $U \leq \alpha(X_t, Y)$  set  $X_{t+1} = Y$

    otherwise set  $X_{t+1} = X_t$

  Increment  $t$

}

# Bayesian methods


## Example of the Metropolis algorithm at work!

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{U(0,1)Binom(h|\theta, n)}{?}$$

$$Binom(h|\theta, n) = \frac{n!}{h!(n-h)!} \theta^h (1-\theta)^{n-h} \quad \begin{array}{l} h = 9; \\ n = 50 \end{array}$$

$$q(.|X_t) = N(X_t, 1)$$

```

Initialize  $X_0$ ; set  $t = 0$ .   $\zeta_0 = 0; \theta_0 = 0.5; P(9|0.5, 50) = 2.2e-06$ 
Repeat {
  Sample a point  $Y$  from  $q(.|X_t)$      $Y = -0.46; Y_\theta = 0.38; P(9|0.38, 50) = 0.0012$ 
  Sample a Uniform(0,1) random variable  $U$      $U = 0.9661937$ 
  If  $U \leq \alpha(X_t, Y)$  set  $X_{t+1} = Y$ 
  otherwise set  $X_{t+1} = X_t$ 
  Increment  $t$ 
}.
  
```




# Bayesian methods

## Example of the Metropolis algorithm at work!

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{U(0,1)Binom(h|\theta, n)}{?}$$

$$Binom(h|\theta, n) = \frac{n!}{h!(n-h)!} \theta^h (1-\theta)^{n-h} \quad \begin{array}{l} h = 9; \\ n = 50 \end{array}$$

$$q(.|X_t) = N(X_t, 1)$$

Initialize  $X_0$ ; set  $t = 0$ .   $\zeta_0 = 0; \theta_0 = 0.5; P(9|0.5, 50) = 2.2e-06$

Repeat {

  Sample a point  $Y$  from  $q(.|X_t)$      $Y = -0.46; Y_\theta = 0.38; P(9|0.38, 50) = 0.0012$

  Sample a Uniform(0,1) random variable  $U$      $U = 0.9661937$

  If  $U \leq \alpha(X_t, Y)$  set  $X_{t+1} = Y$      $\alpha(\theta_0, Y_\theta) = \frac{0.0012}{2.2e-06} = 572.28$

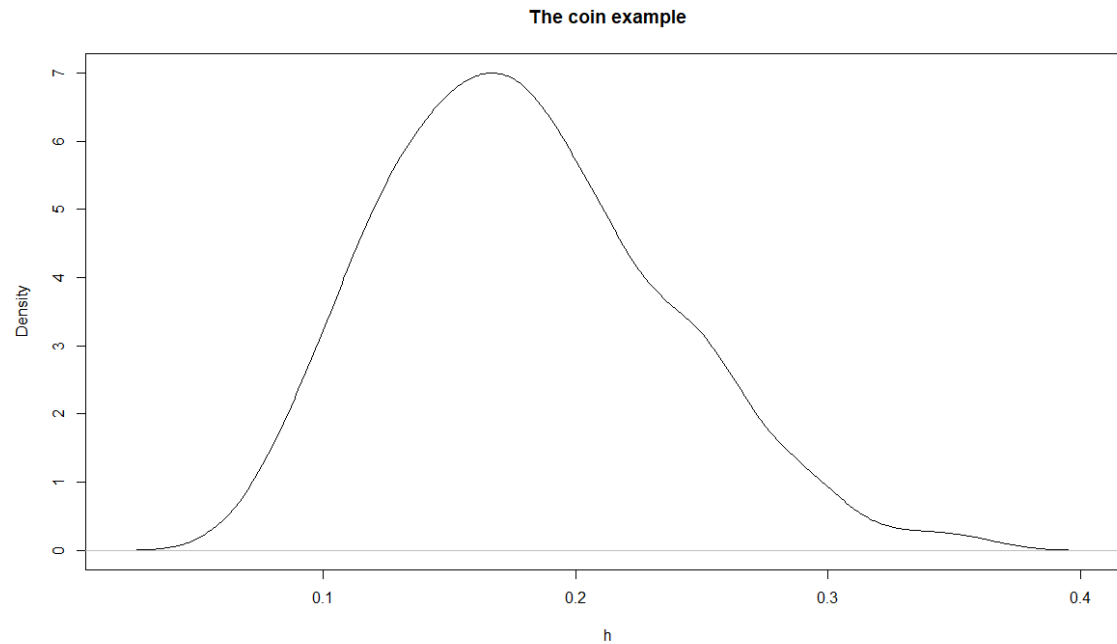
    otherwise set  $X_{t+1} = X_t$

  Increment  $t$

}.

$\zeta_1 = -0.46; \theta_1 = 0.38; P(9|0.38, 50) = 0.0012$

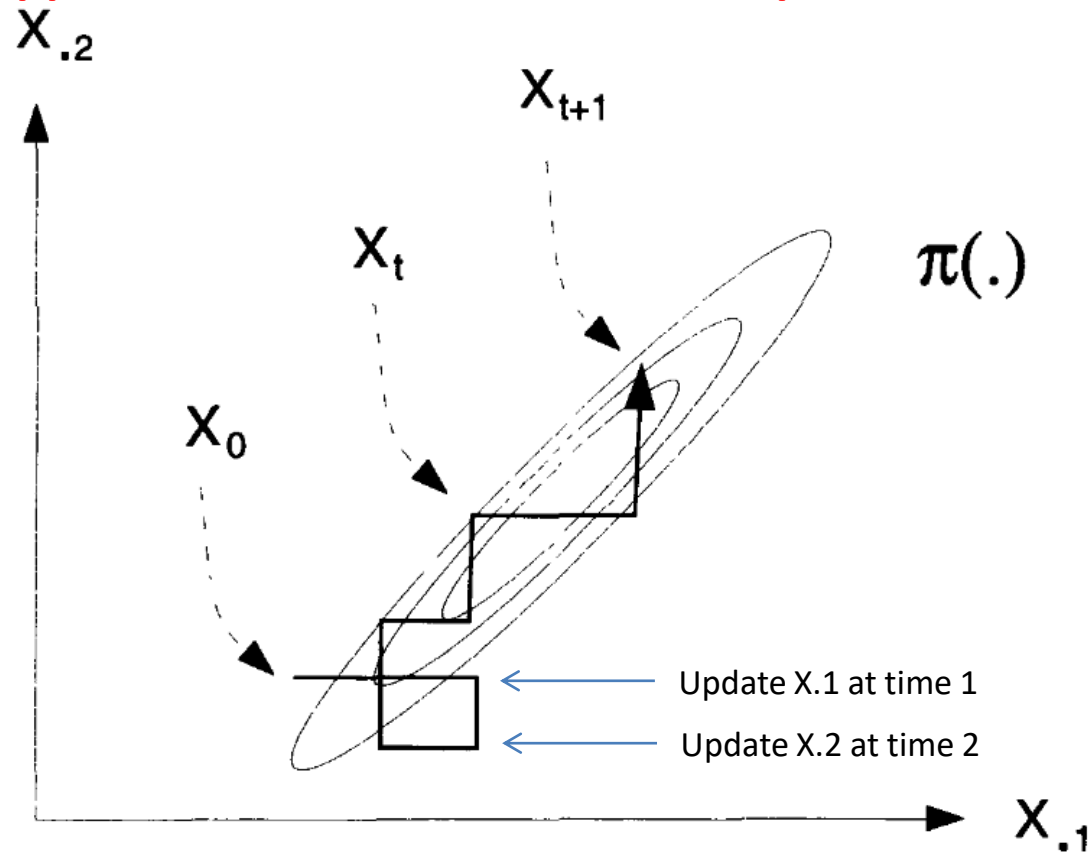
# Posterior distribution found by Metropolis



# Bayesian methods

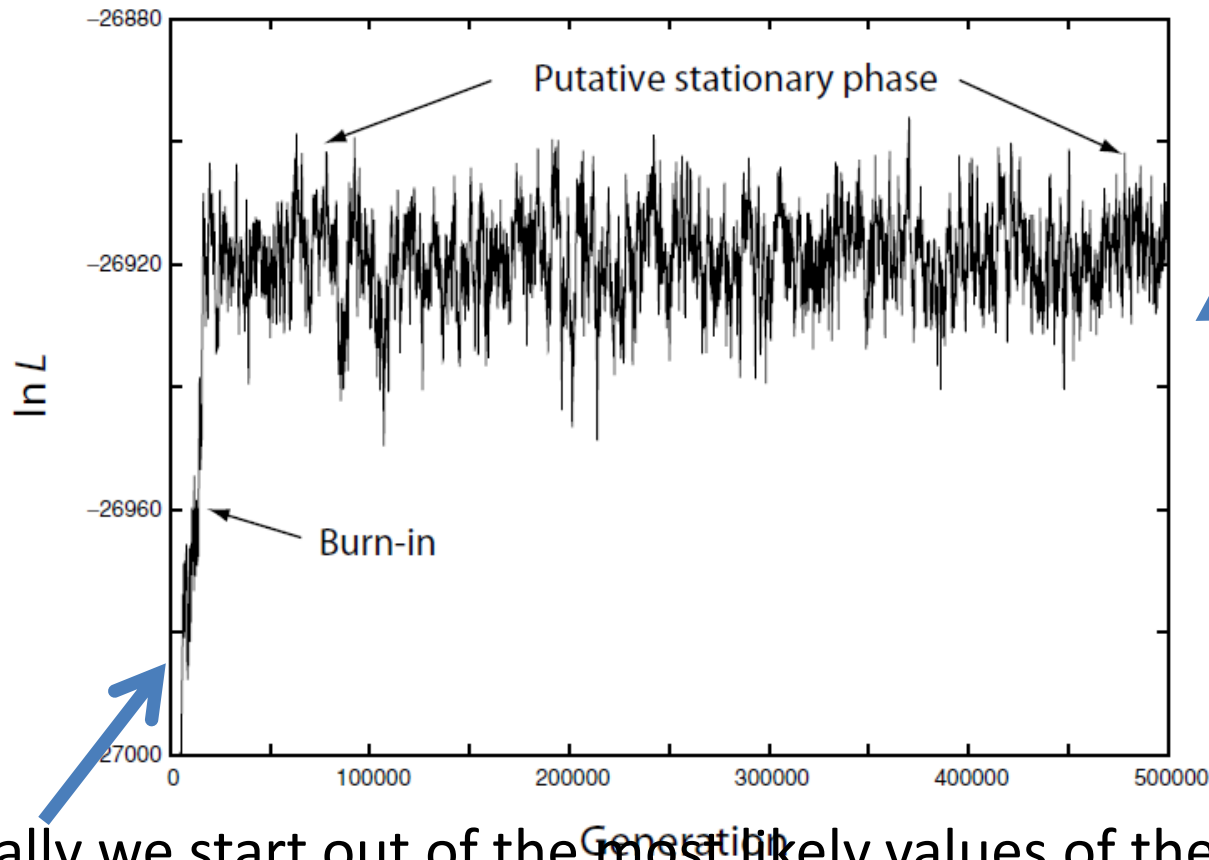
MCMC: Metropolis-Hastings single component update

**What happens if we have more than one parameter to estimate?**



# Bayesian methods

## MCMC in practice

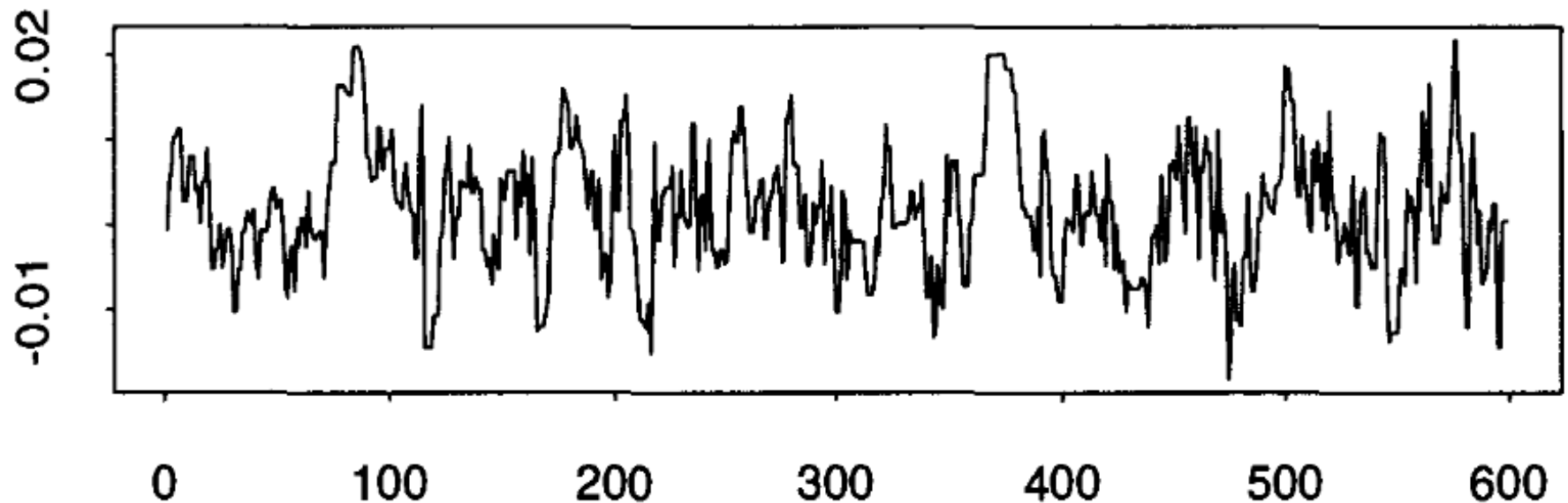


When to stop?

Usually we start out of the most likely values of the posterior

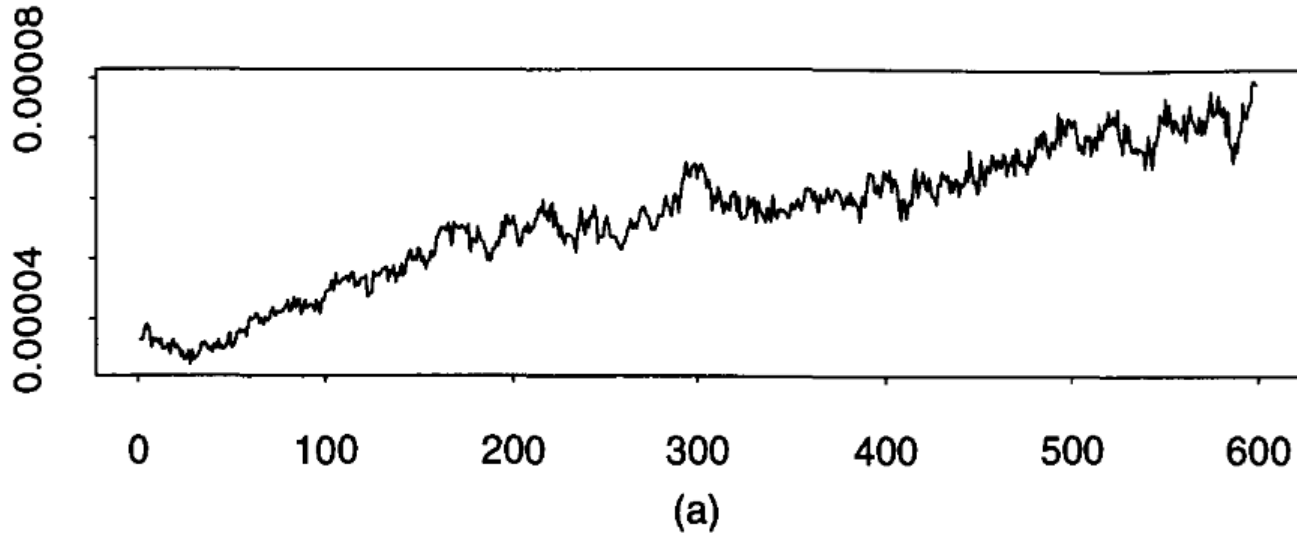
# Bayesian methods

## MCMC in practice: good mixing



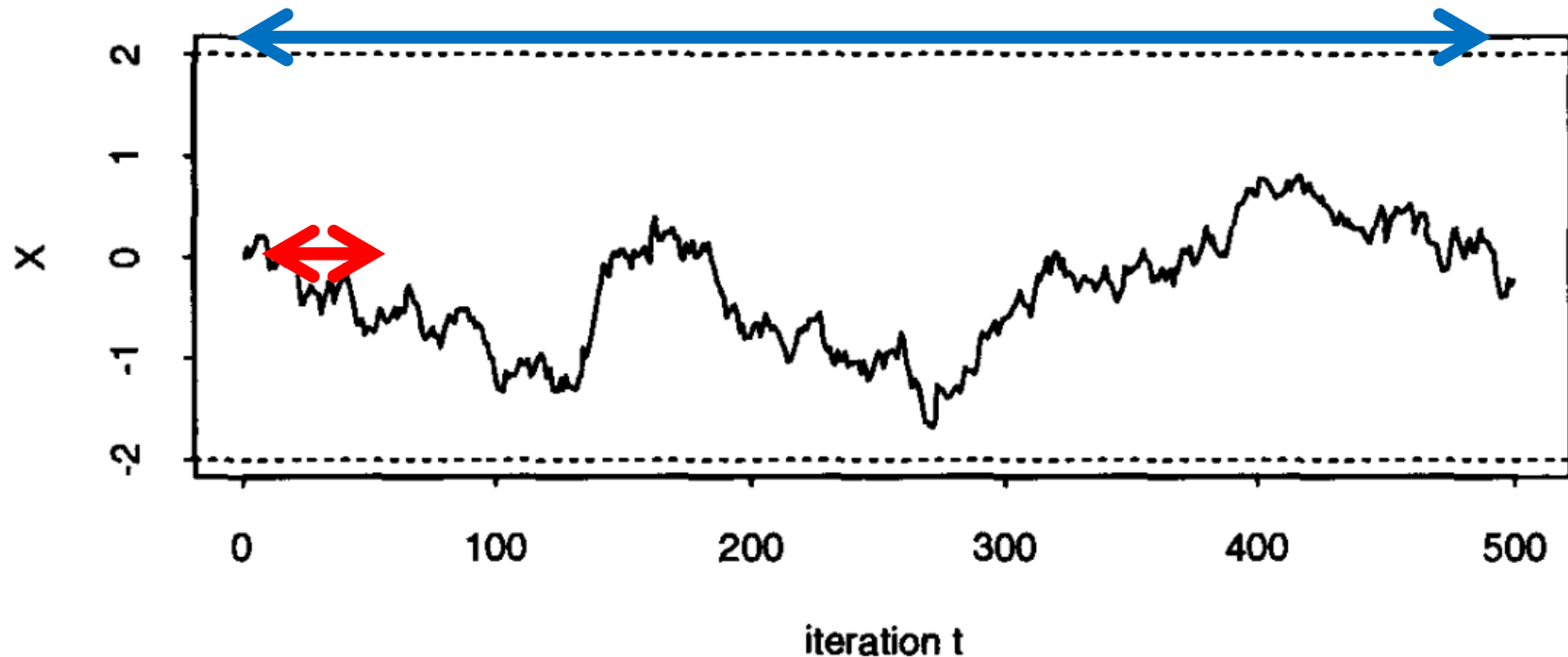
# Bayesian methods

## MCMC in practice: bad mixing



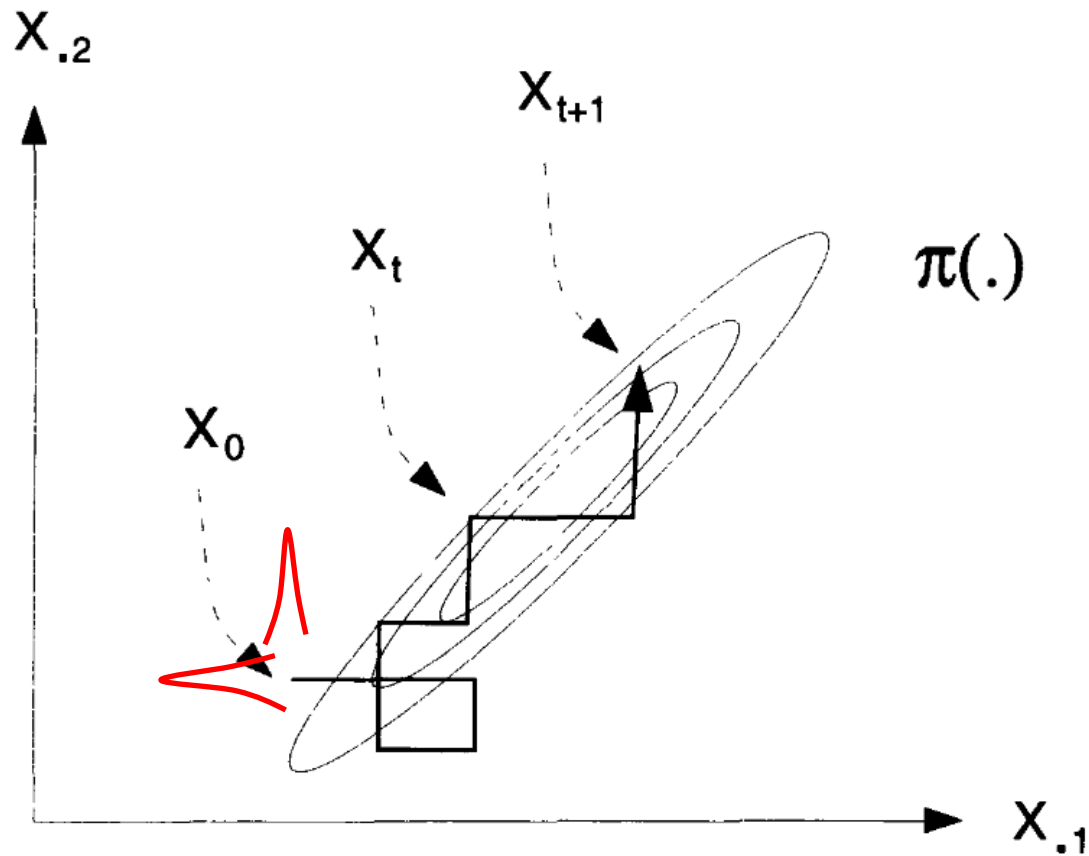
# Bayesian methods

## MCMC in practice: bad mixing



# Bayesian methods

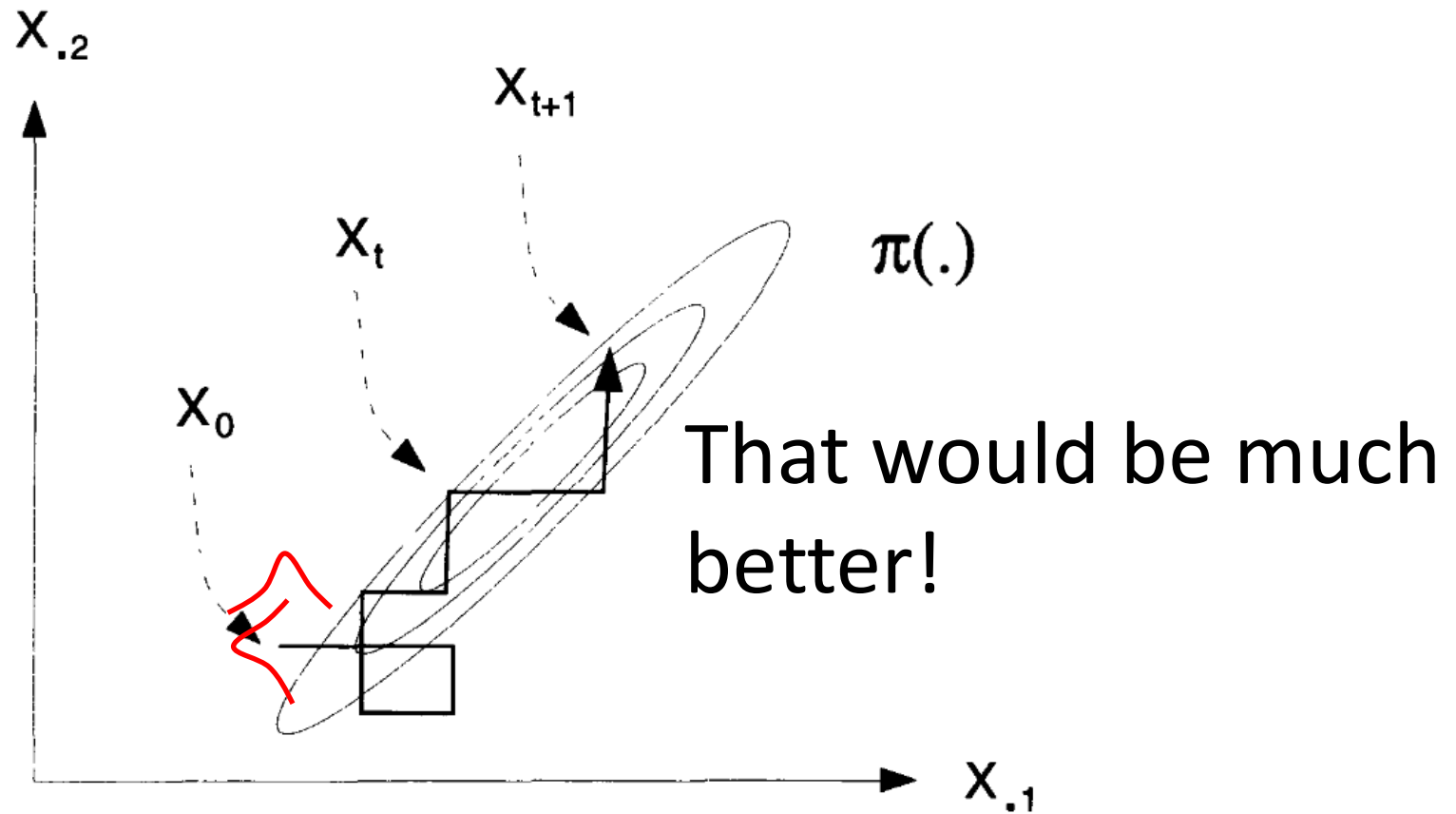
MCMC: Why poor mixing?





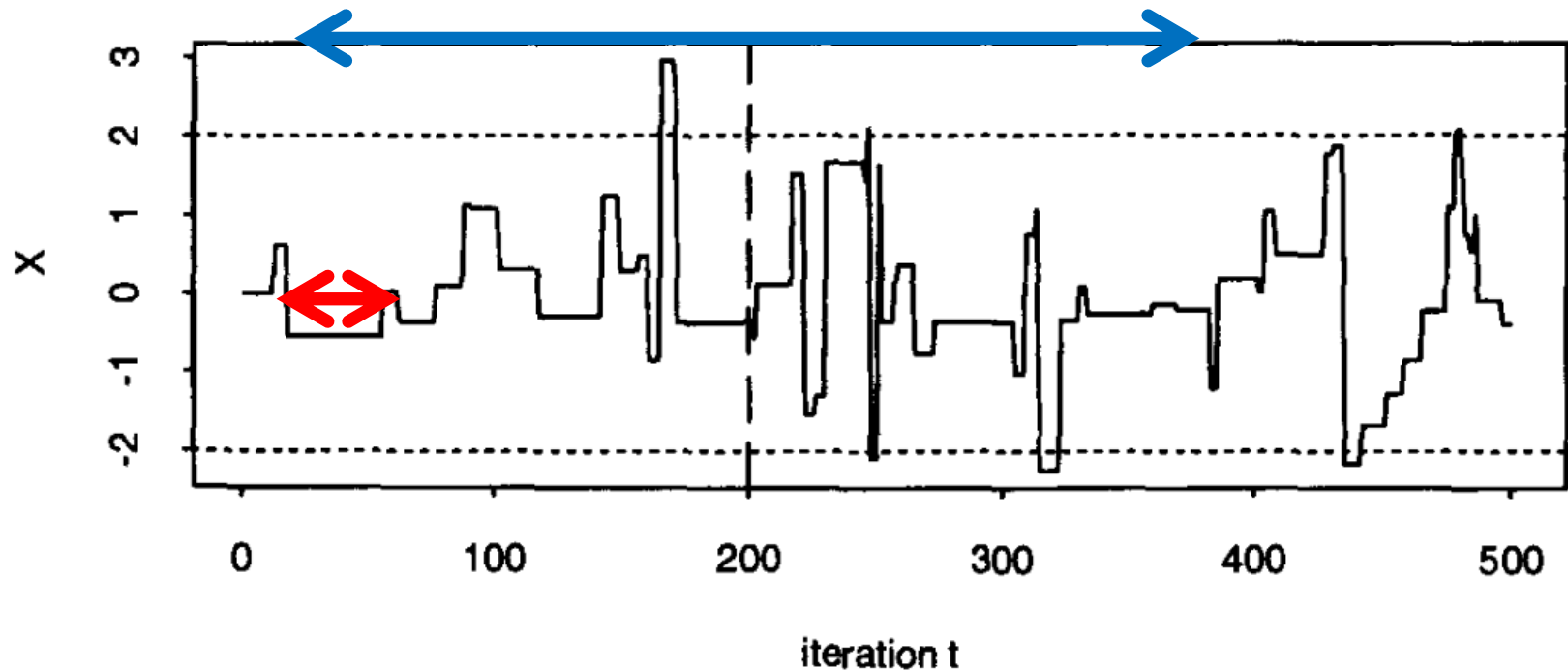
# Bayesian methods

MCMC: Why poor mixing?



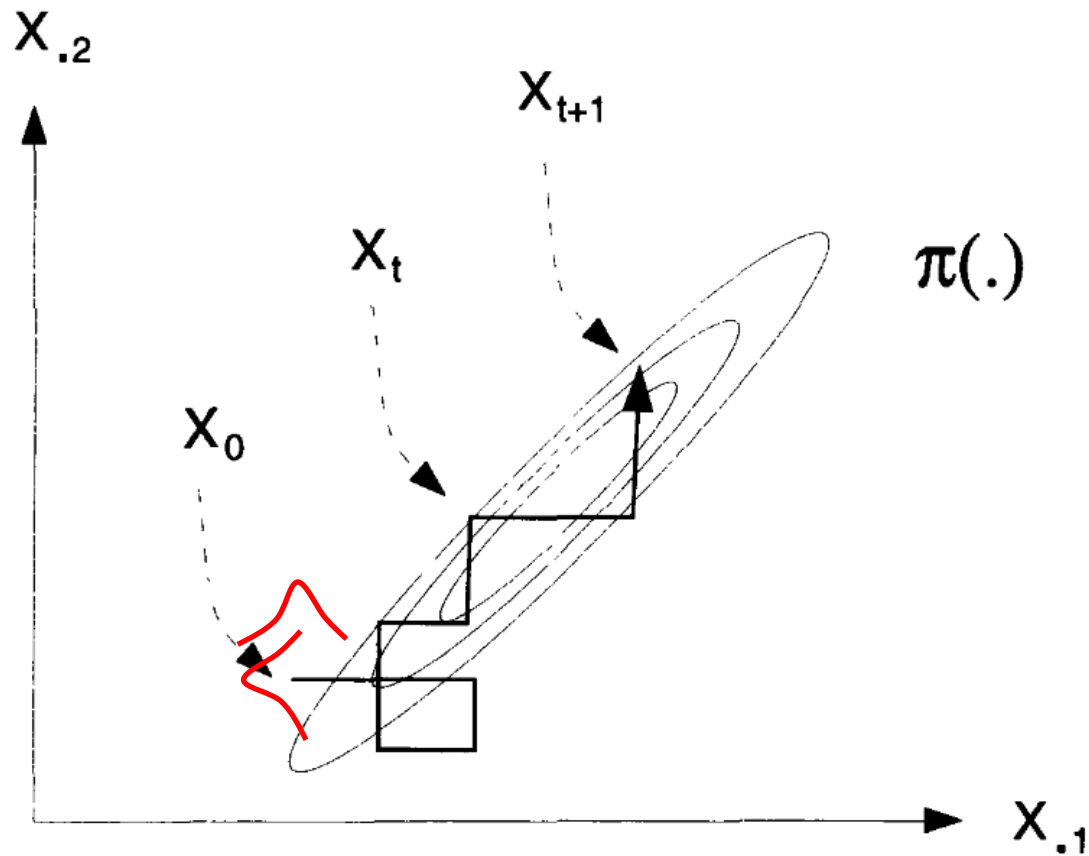
# Bayesian methods

## MCMC in practice: bad mixing



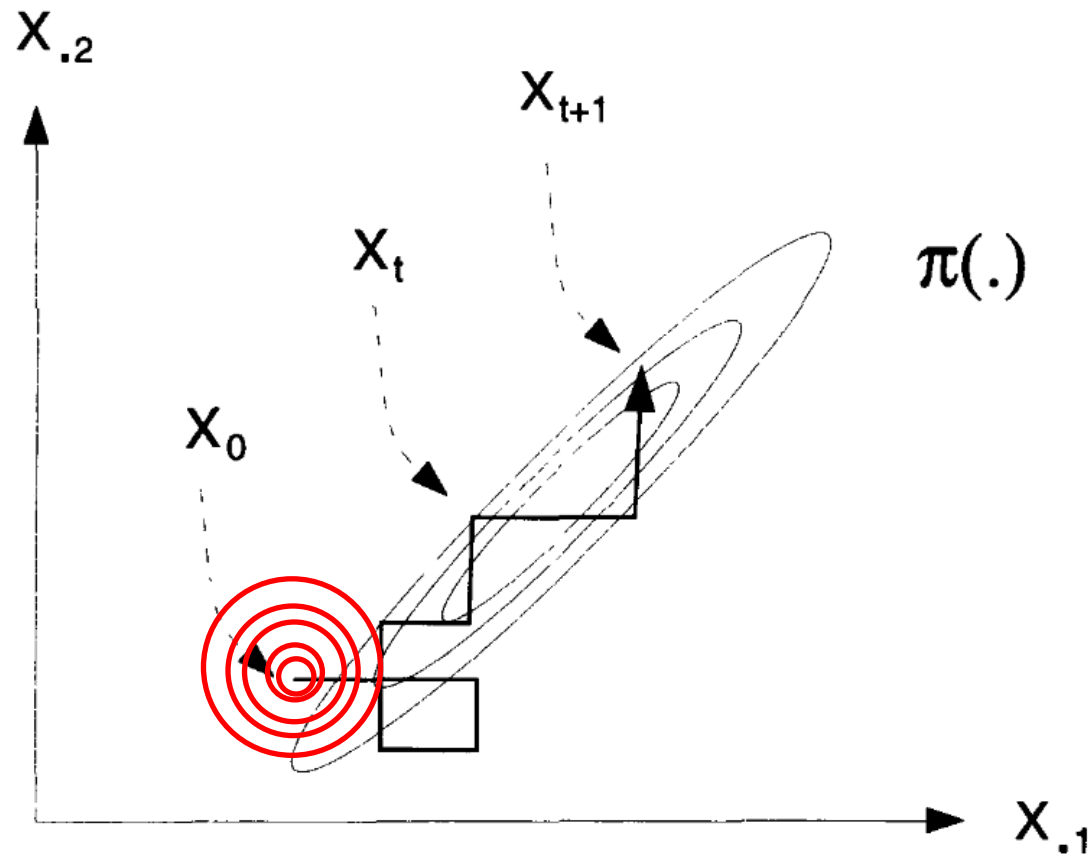
# Bayesian methods

MCMC: Why poor mixing?



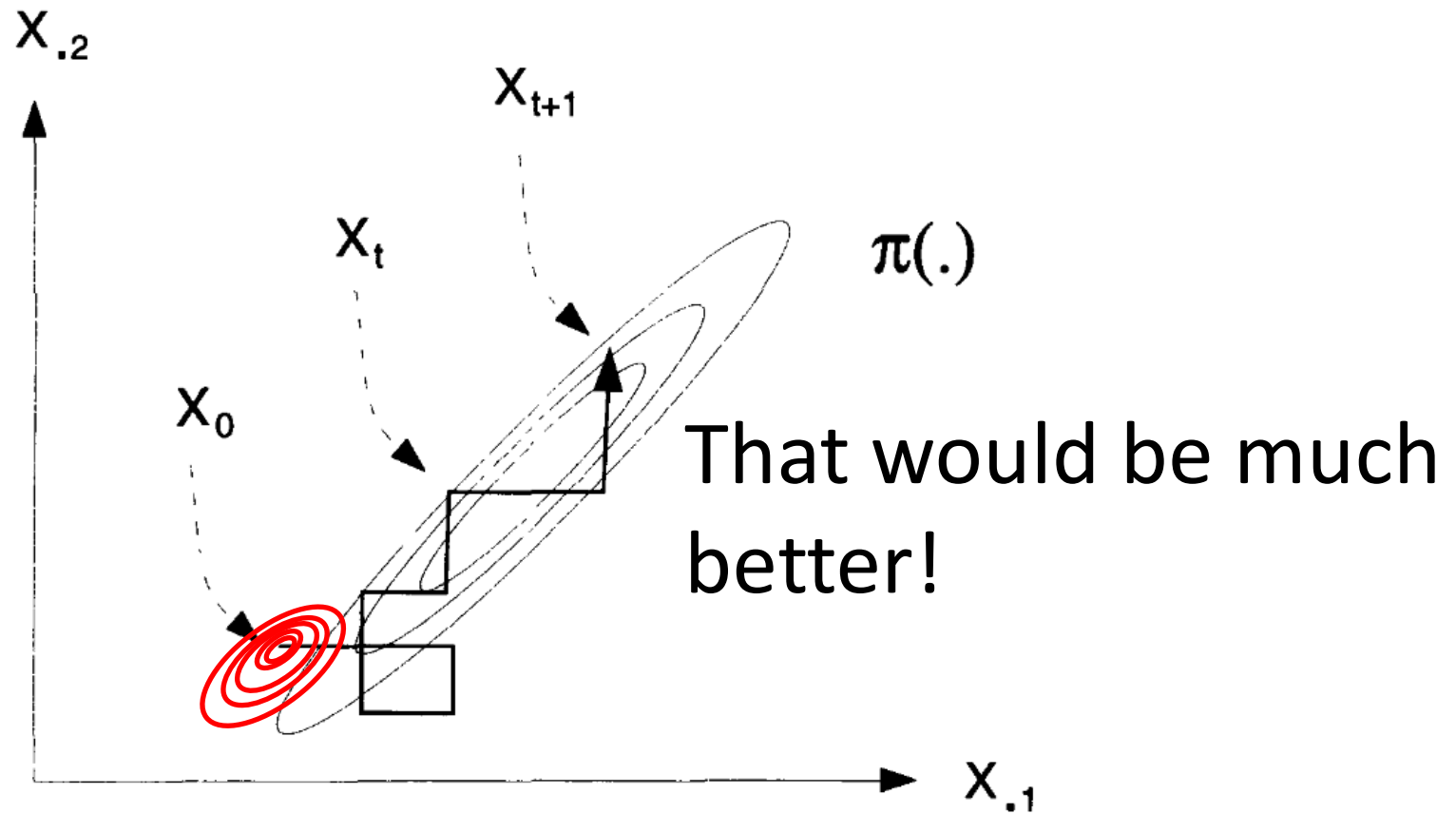
# Bayesian methods

MCMC: Why poor mixing?



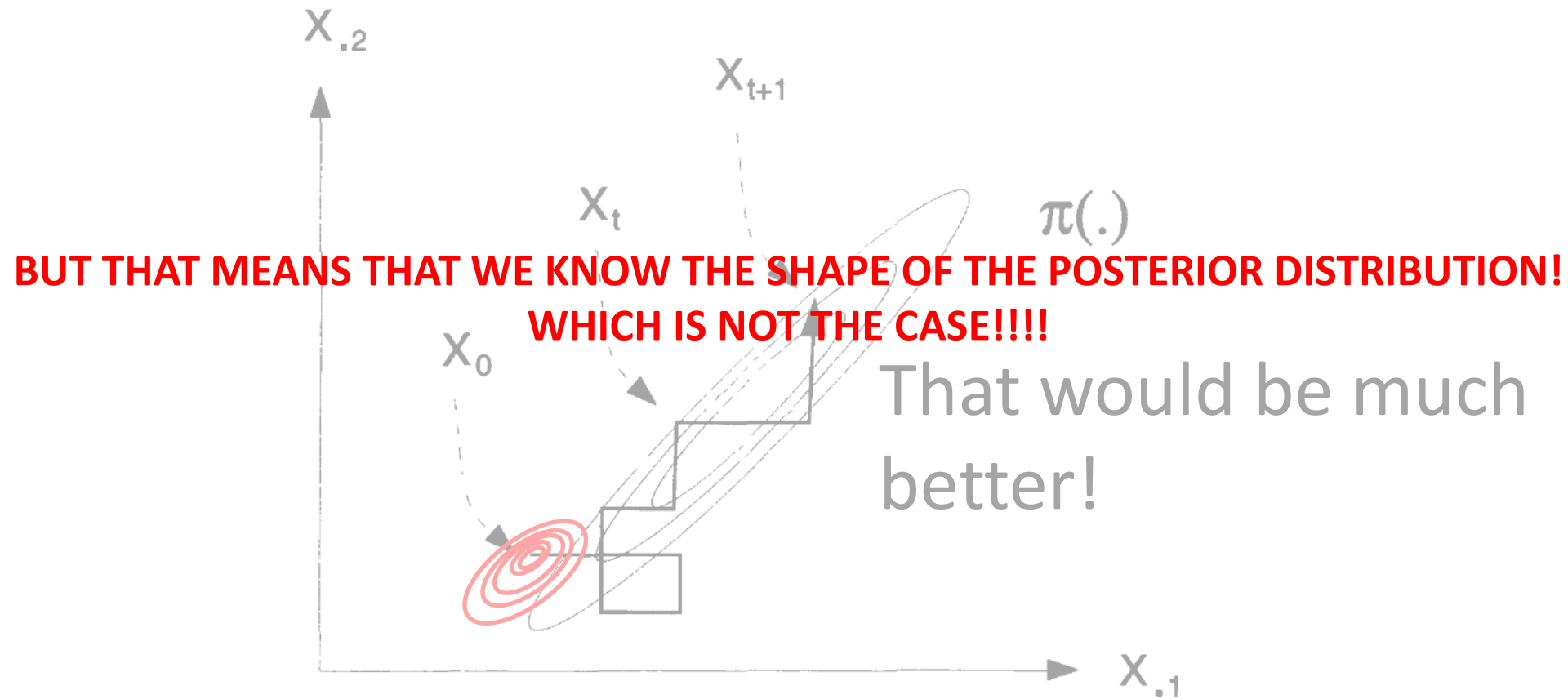
# Bayesian methods

MCMC: Why poor mixing?



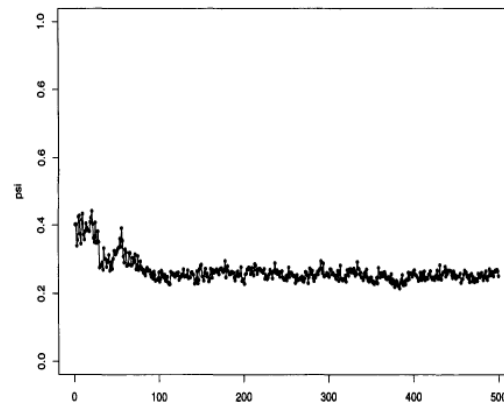
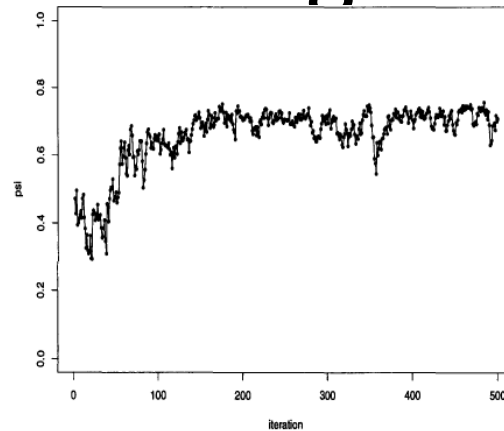
# Bayesian methods

MCMC: Why poor mixing?



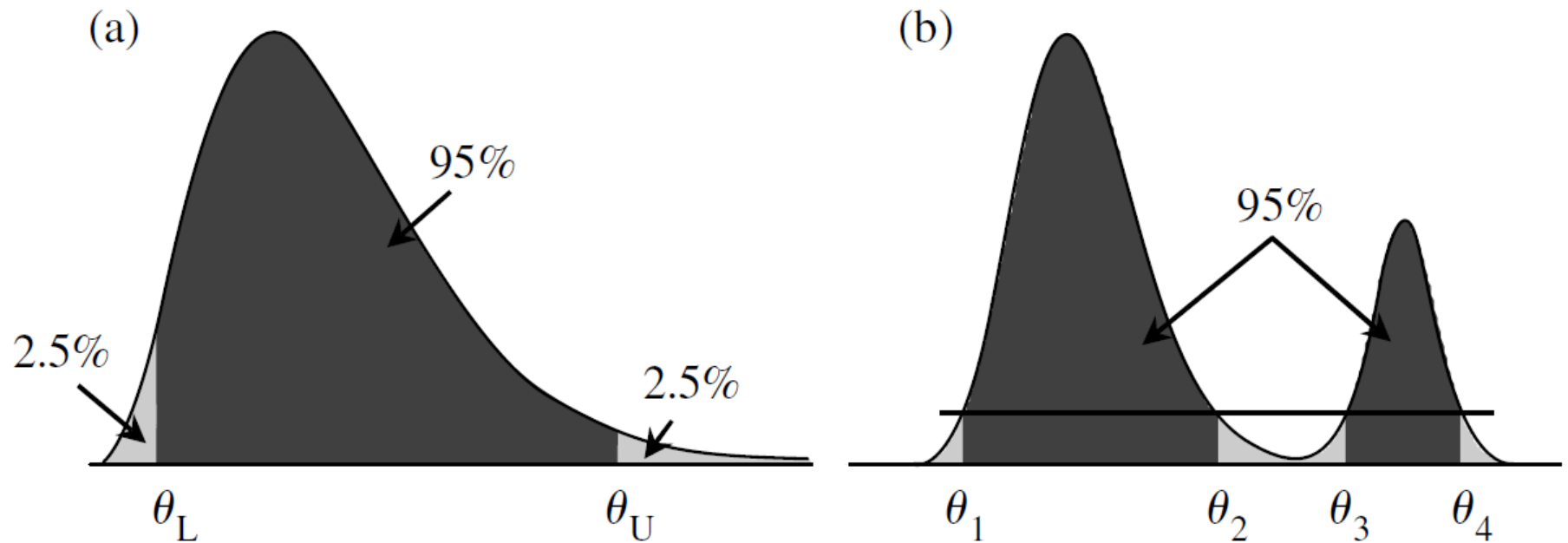
# Bayesian methods

## MCMC in practice: Lack of convergence



# Bayesian methods

## Yet another issue: multimodality





# Bayesian methods

## MCMC: Be careful



# Bayesian methods

## MCMC in practice

- The final **GOAL** of MCMC is to retrieve samples from the posterior distribution without having a close form of this distribution
- However, we start at a random point of the parameter space, which probably has a low probability given the data.
- We want independent samples from the posterior distribution. However, we are doing walks that depends from previous sampled values...

# Bayesian methods

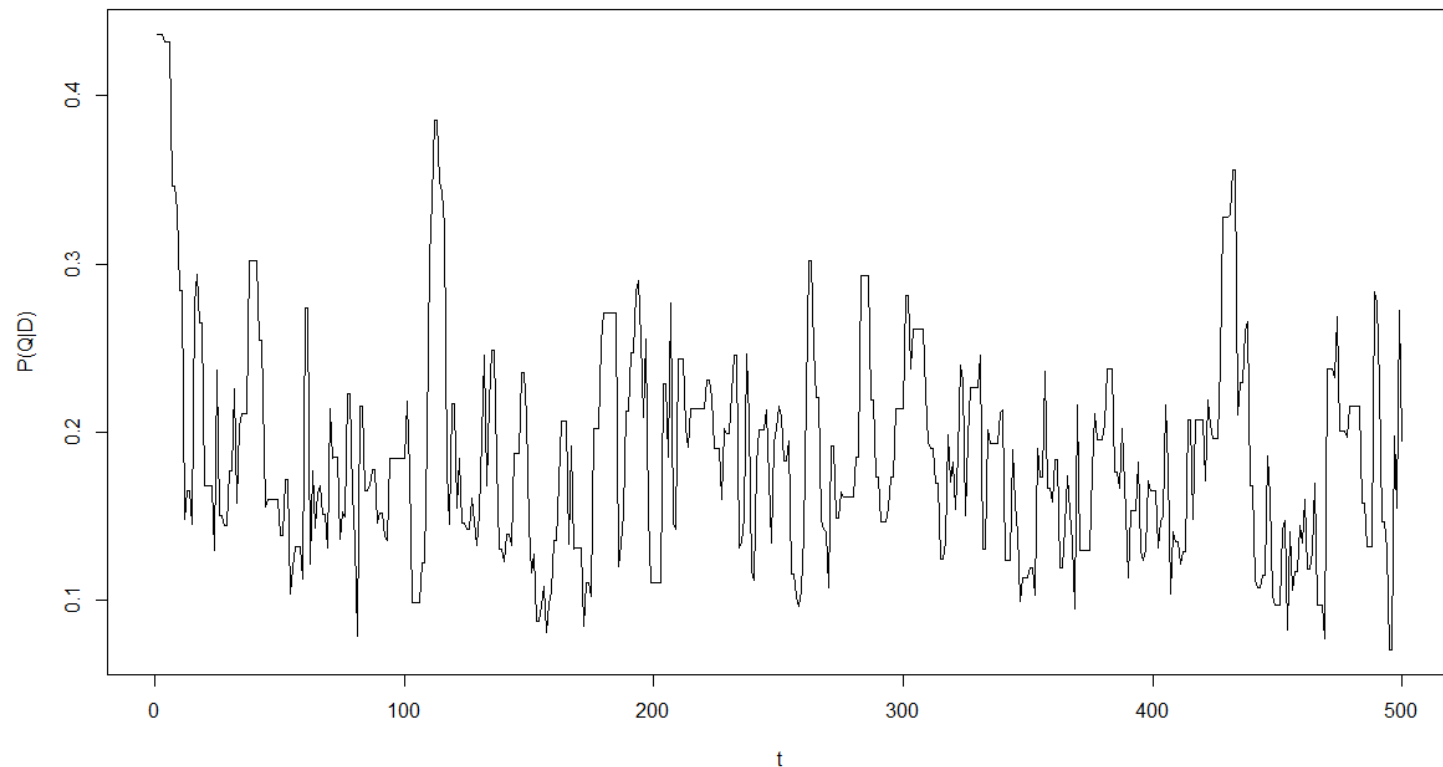
## MCMC: Some hints

- Define burn-in
- Run more than one MCMC chain and check convergence of the loglikelihood.
- Compute metrics of autocorrelation between  $t$  and  $t+k$ 
  - Is good the mixing?
  - Define lag between accepted runs

# Bayesian methods

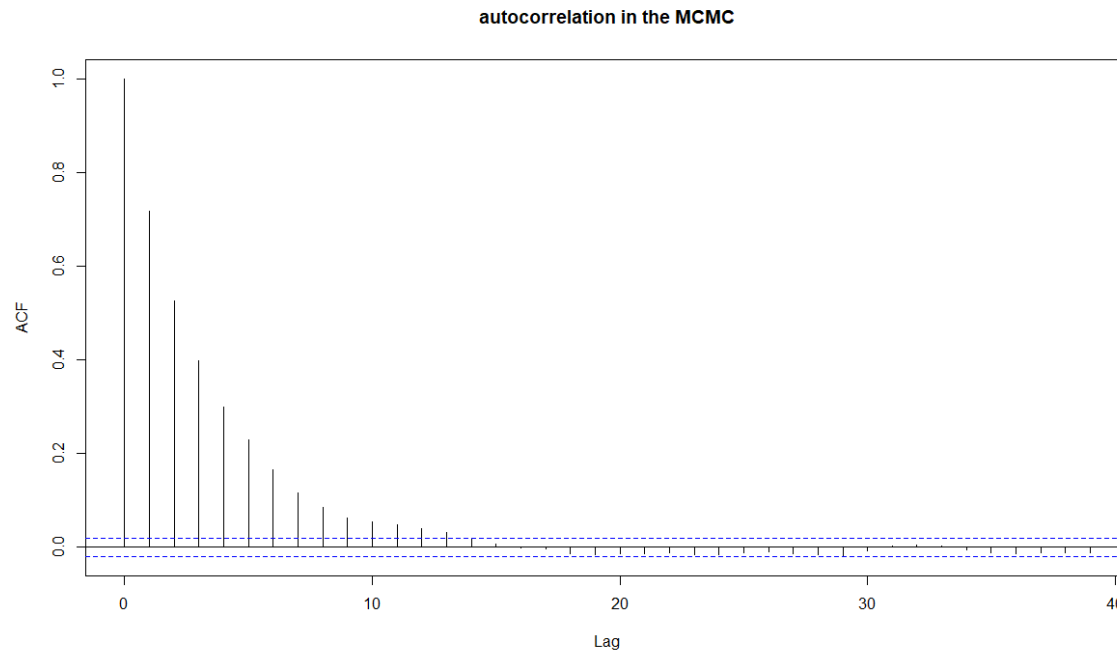
## Example of the Metropolis algorithm at work!

The coin example



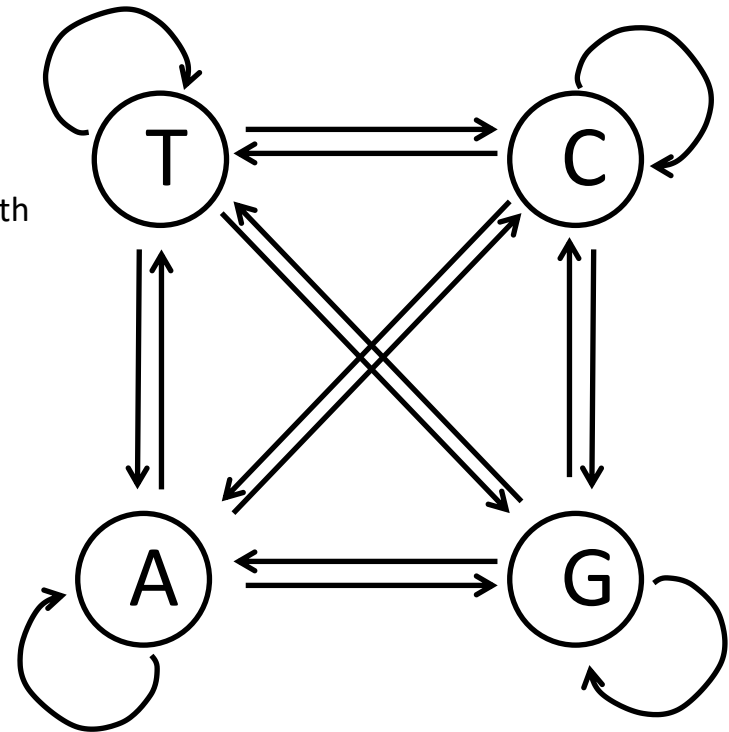
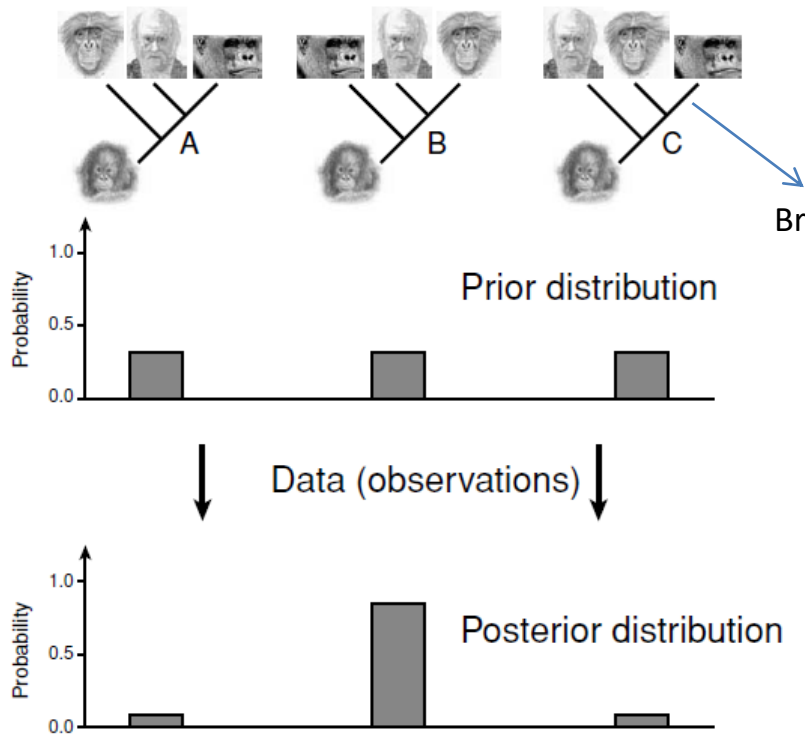
# Bayesian methods

## Example of the Metropolis algorithm at work



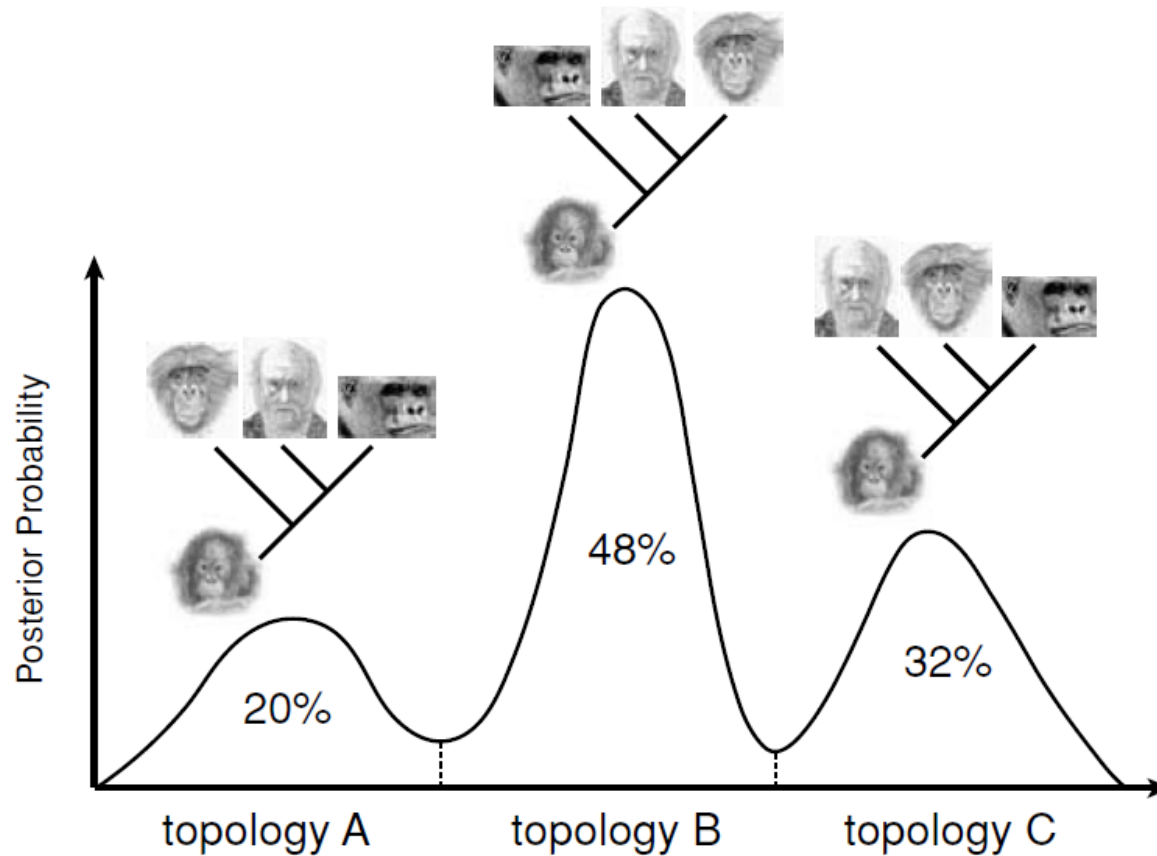
# Bayesian methods

## Bayesian phylogenetics



# Bayesian methods

## Bayesian phylogenetics

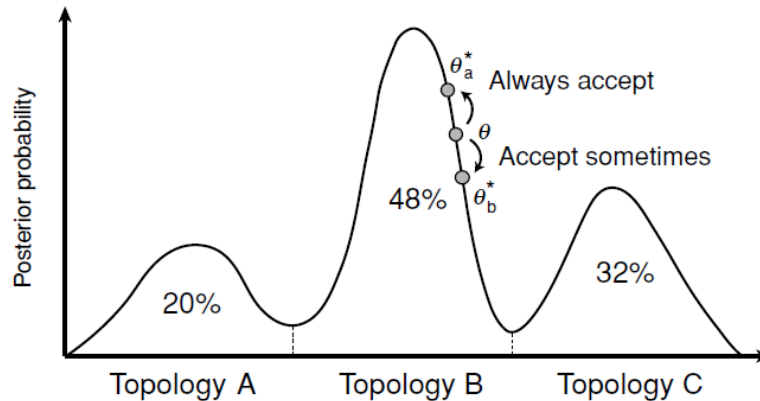


# Bayesian methods

## Bayesian phylogenetics

### Markov chain Monte Carlo steps

1. Start at an arbitrary point ( $\theta$ )
2. Make a small random move (to  $\theta^*$ )
3. Calculate height ratio ( $r$ ) of new state (to  $\theta^*$ ) to old state ( $\theta$ )
  - (a)  $r > 1$ : new state accepted
  - (b)  $r < 1$ : new state accepted with probability  $r$   
if new state rejected, stay in old state
4. Go to step 2



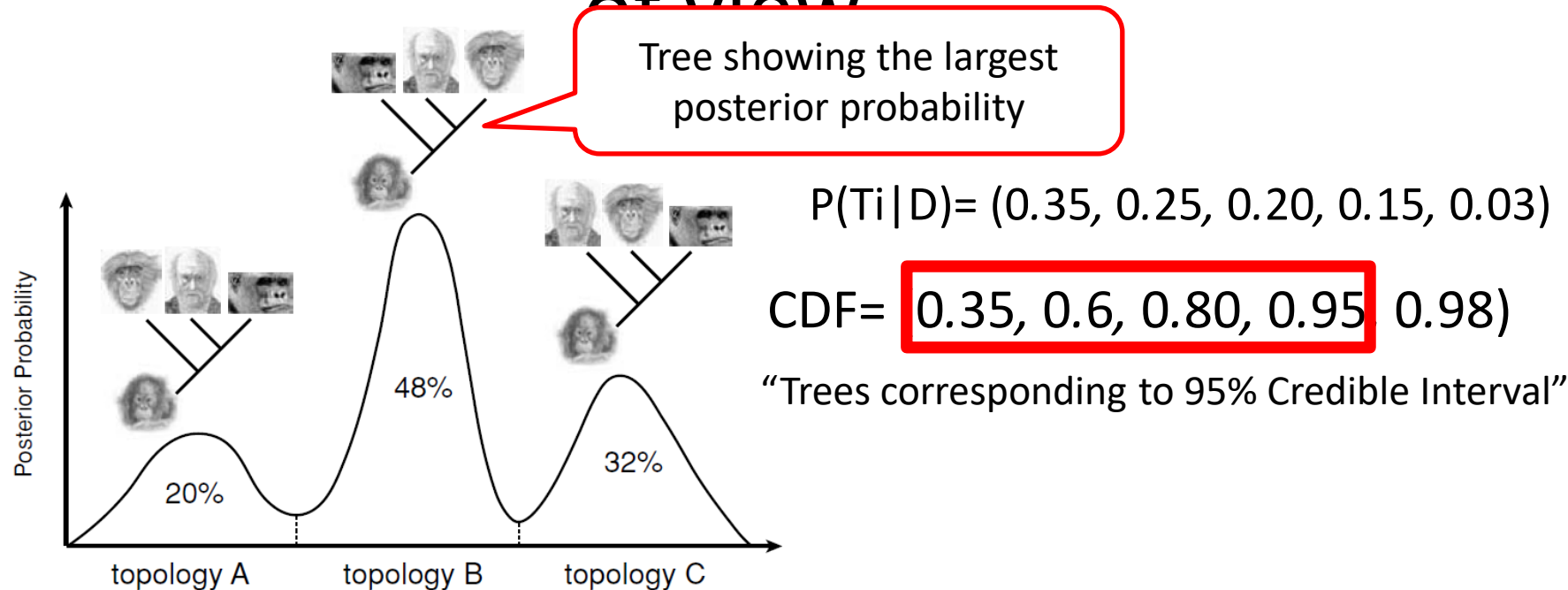


# Bayesian methods

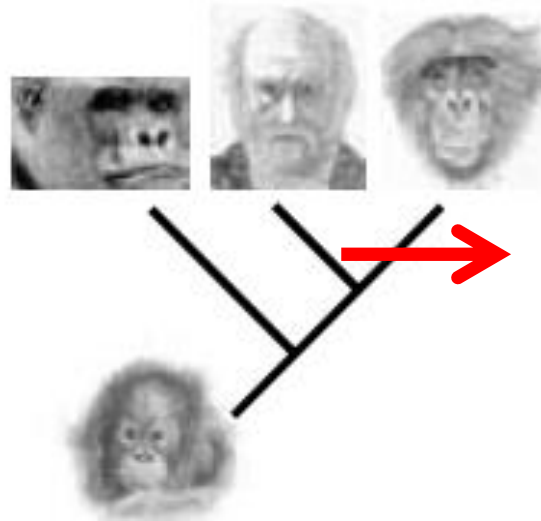
## Bayesian phylogenetics

1. Start with a random tree  $\tau$ , with random branch lengths  $\mathbf{b}$ , and random substitution parameters  $\theta$ .
2. In each iteration do the following:
  - a. Propose a change to the tree, by using tree rearrangement algorithms (such as NNI or SPR). This step may change branch lengths  $\mathbf{b}$  as well.
  - b. Propose changes to branch lengths  $\mathbf{b}$ .
  - c. Propose changes to parameters  $\theta$ .
  - d. Every  $k$  iterations, sample the chain: save  $\tau$ ,  $\mathbf{b}$ ,  $\theta$  to disk.
3. At the end of the run, summarize the results.

# Bayesian phylogenetics: how to summarize results? From a tree point of view



# Bayesian phylogenetics: how to summarize results? From a branch point of view



- Branch length
  - Mean, median, mode
  - 95% Credible interval
- How often this branch is retrieved?