How should we go about modeling this?

```
gorilla   GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan GGACTCCTTGAGAAATAAACTGCACACACTGG
```

Model parameters?

Time

Substitution rate

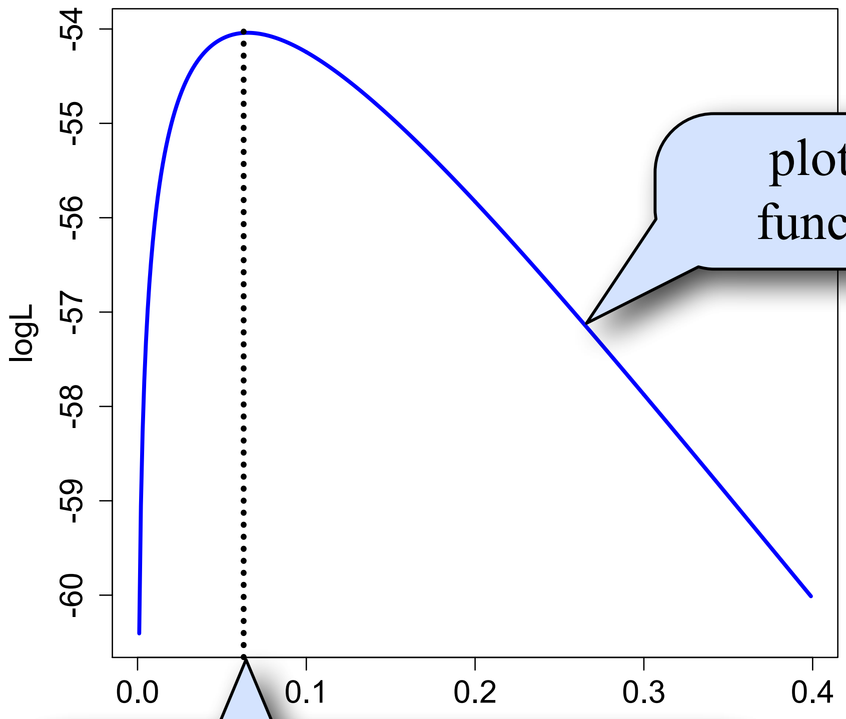
Can we observe time or subst. rate?

What *can* we observe?

Maximum likelihood estimation

First 32 nucleotides of the ψη-globin gene of gorilla and orangutan:

```
gorilla   GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan GGACTCCTTGAGAAATAAACTGCACACACTGG
```



plot of log-likelihood as a function of branch length *v*

maximum likelihood estimate (MLE) of *v* is 0.065259

Knowing that v = rate×time,

- Why does the curve drop as you move left of the MLE?
- Why does the curve drop as you move right of the MLE?

JC69 rate matrix

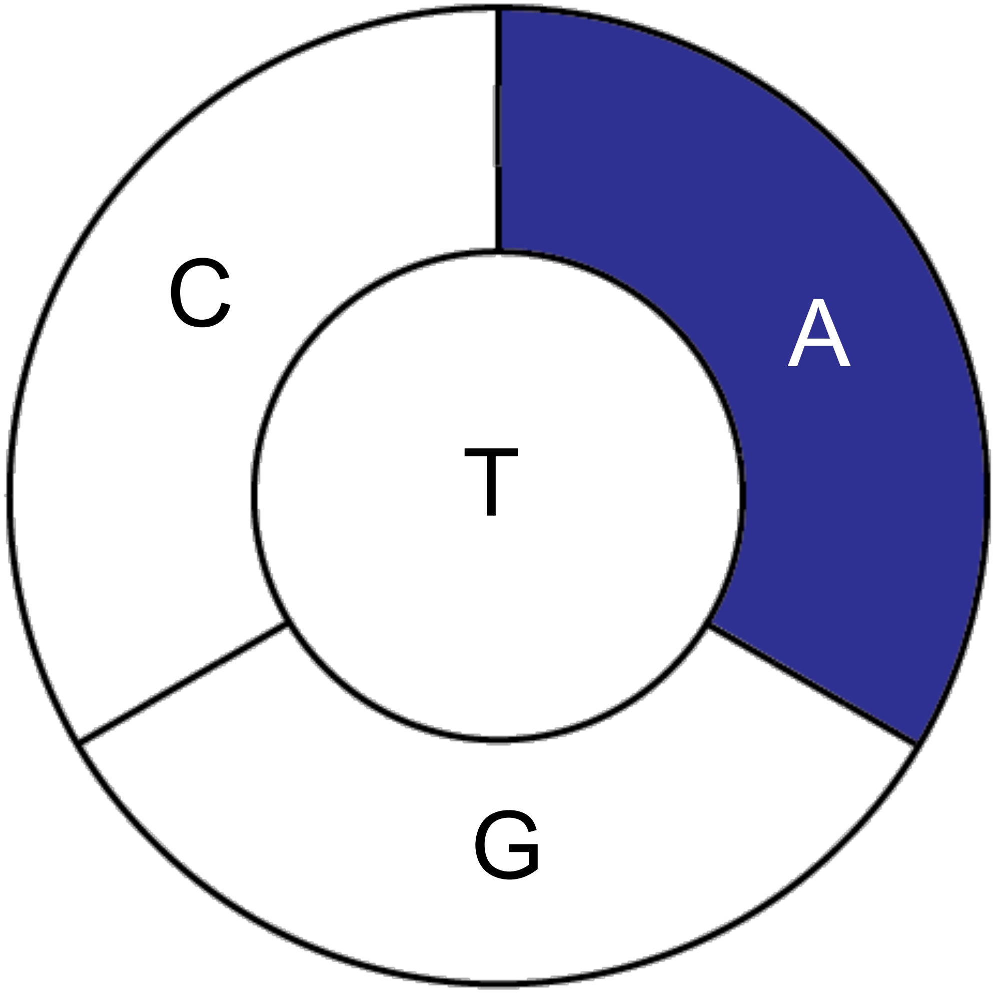
1 parameter:
α

What does this mean?
Why is this number negative?

$$\begin{array}{cc|cccc}
 & & & \text{To} & & \\
 & & A & C & G & T \\
\hline
 & A & -3\alpha & \alpha & \alpha & \alpha \\
\text{From} & C & \alpha & -3\alpha & \alpha & \alpha \\
 & G & \alpha & \alpha & -3\alpha & \alpha \\
 & T & \alpha & \alpha & \alpha & -3\alpha
\end{array}$$

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 *in* H. N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York.

Equilibrium Frequencies



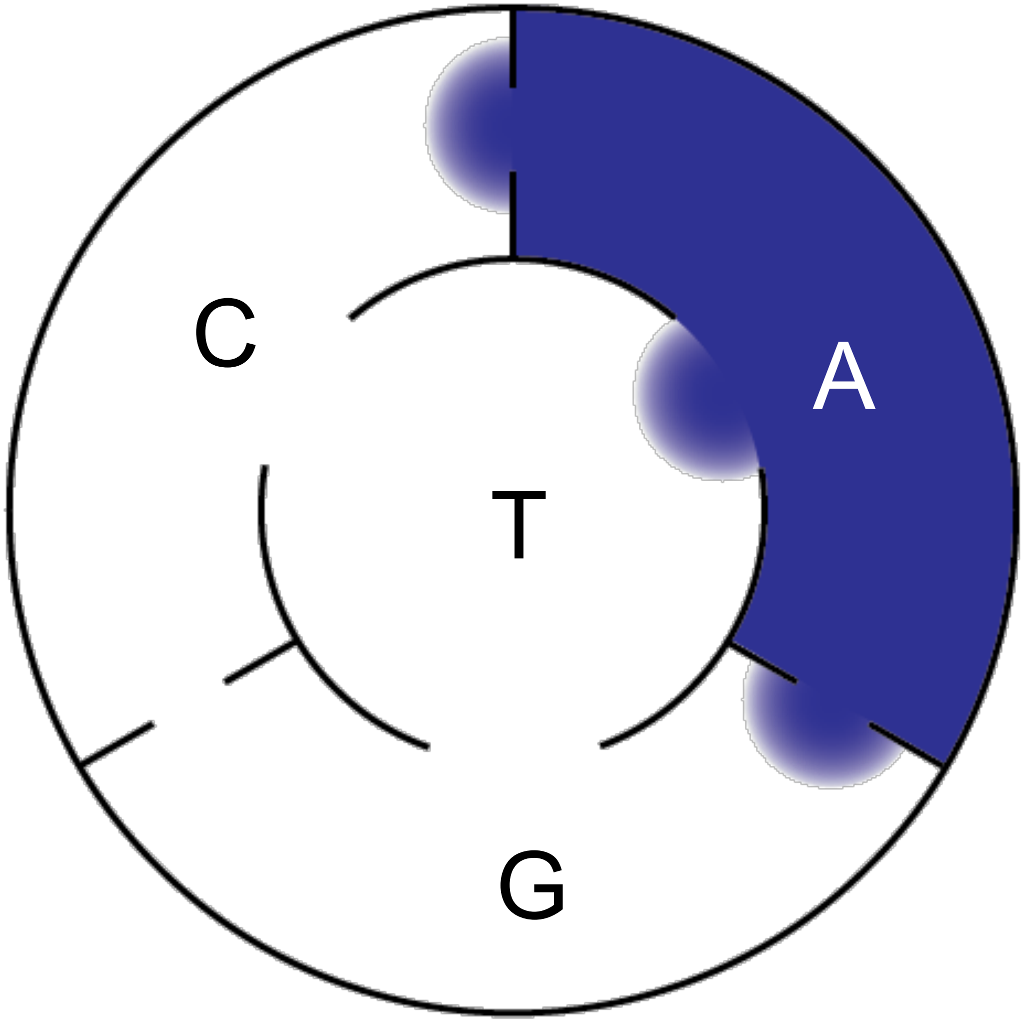
A sequence consisting only of A...

```
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
```

Perfume bottle broken, and perfume quickly fills the room

Equilibrium Frequencies

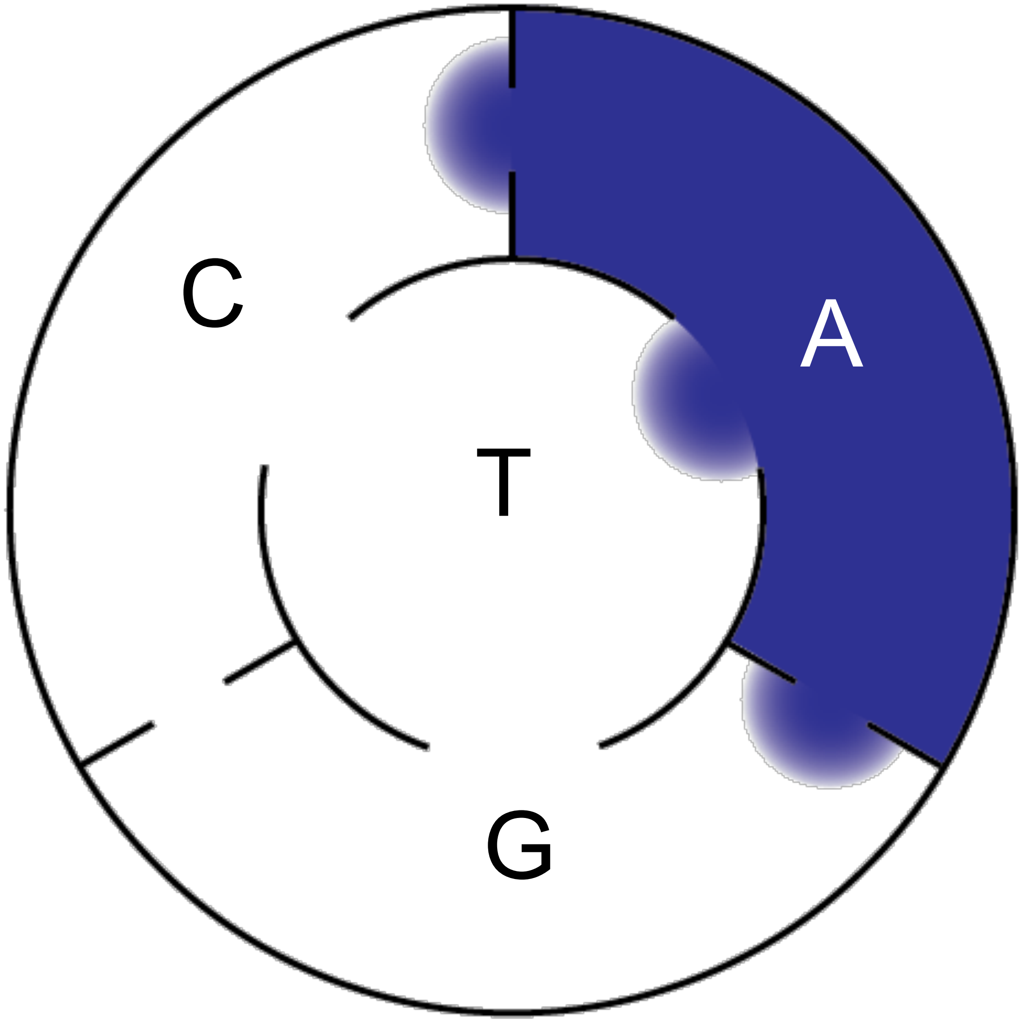
If all doors are suddenly opened, perfume will spread by diffusion to the other rooms...



The instant the doors open, the rate *away* from A is 3α (i.e. rate = -3α)

AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAAAAAAAAAAAAA

Equilibrium Frequencies

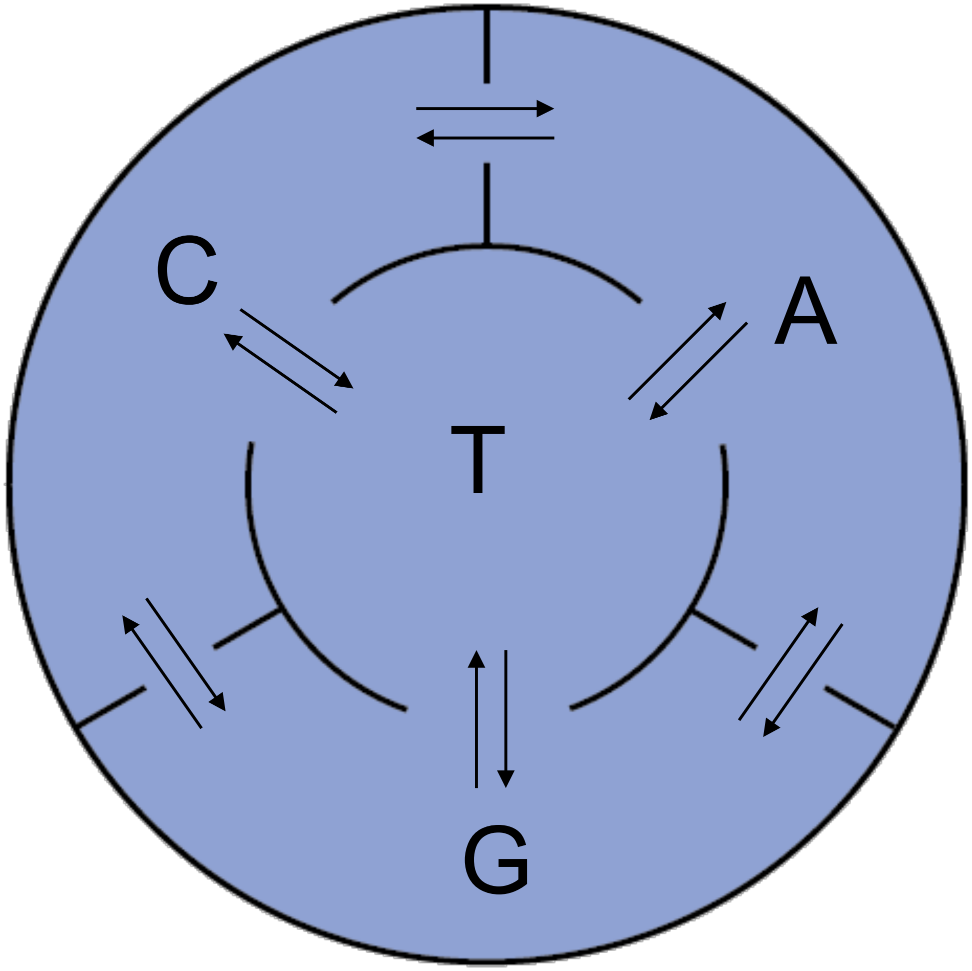


Sequence now contains a few Cs, Gs, and Ts...

```
AATAAAAAAAAAAAA
AAAAAAAAAAAAAAA
AAAACAAAAAATAAA
AAAAAAAAAAAAAAA
AAAAAAACGAAGAAA
AAAAAAAAAAAAAAA
AAATAAAAAAAAAAA
```

As perfume spreads by diffusion, the difference in concentration among rooms decreases...

Equilibrium Frequencies

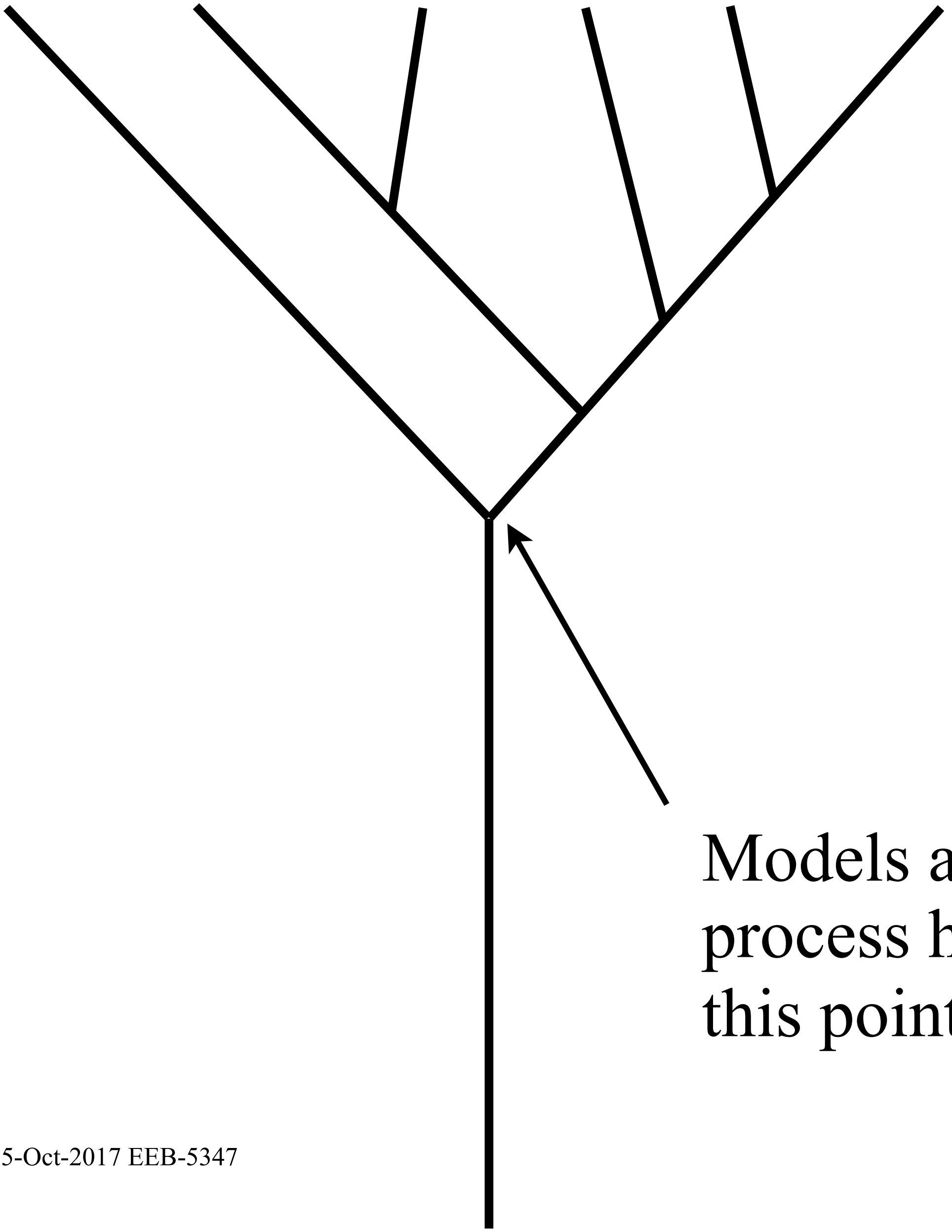


Sequence contains a mixture of about equal quantities A, C, G and T

```
CAGAATCGAGCAGCT
TGACTACGTCATGTG
GTTGCGCCGCAACGC
CATATACCGCCGACT
AGTTTGAGGGCGGTT
AGGGCTCGGTTCGTA
CATCGTATAAACATT
```

After a long time, equilibrium (=stationarity) is achieved.

Stationarity Assumed



Models assume *stationarity*, which means that the state frequencies do not change across the tree

Models assume that substitution process has reached equilibrium by this point

K80 (or K2P) rate matrix

2 parameters:
α
β

To

$$\begin{array}{cc|cccc}
 & & \text{A} & \text{C} & \text{G} & \text{T} \\
\hline
 & \text{A} & -\alpha - 2\beta & \beta & \alpha & \beta \\
\text{From} & \text{C} & \beta & -\alpha - 2\beta & \beta & \alpha \\
 & \text{G} & \alpha & \beta & -\alpha - 2\beta & \beta \\
 & \text{T} & \beta & \alpha & \beta & -\alpha - 2\beta
\end{array}$$

transition rate

transversion rate

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111-120.

K80 rate matrix
(looks different, but actually the same)

2 parameters:
κ
β

$$\begin{array}{c|cccc}
 & \text{A} & \text{C} & \text{G} & \text{T} \\
\hline
\text{A} & -\beta(\kappa+2) & \beta & \kappa\beta & \beta \\
\text{C} & \beta & -\beta(\kappa+2) & \beta & \kappa\beta \\
\text{G} & \kappa\beta & \beta & -\beta(\kappa+2) & \beta \\
\text{T} & \beta & \kappa\beta & \beta & -\beta(\kappa+2)
\end{array}$$

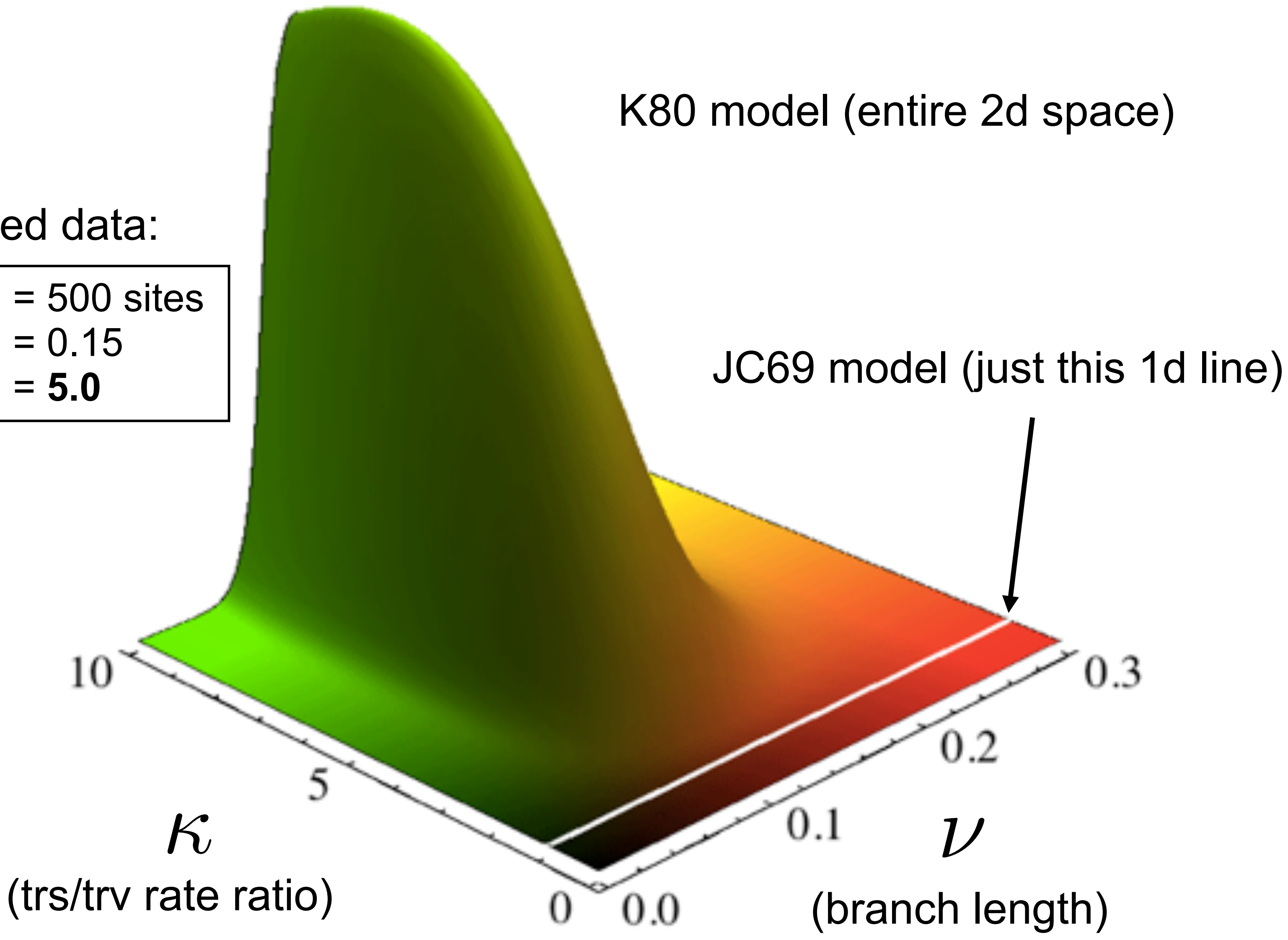
All I've done is define κ to be the *transition/transversion rate ratio* ⟶ $\kappa = \dfrac{\alpha}{\beta}$

Note: the K80 model is identical to the JC69 model if $\kappa = 1$ ($\alpha = \beta$)

Likelihood Surface when K80 true

Based on simulated data:

sequence length	= 500 sites
true branch length	= 0.15
true kappa	= **5.0**



K80 model (entire 2d space)

JC69 model (just this 1d line)

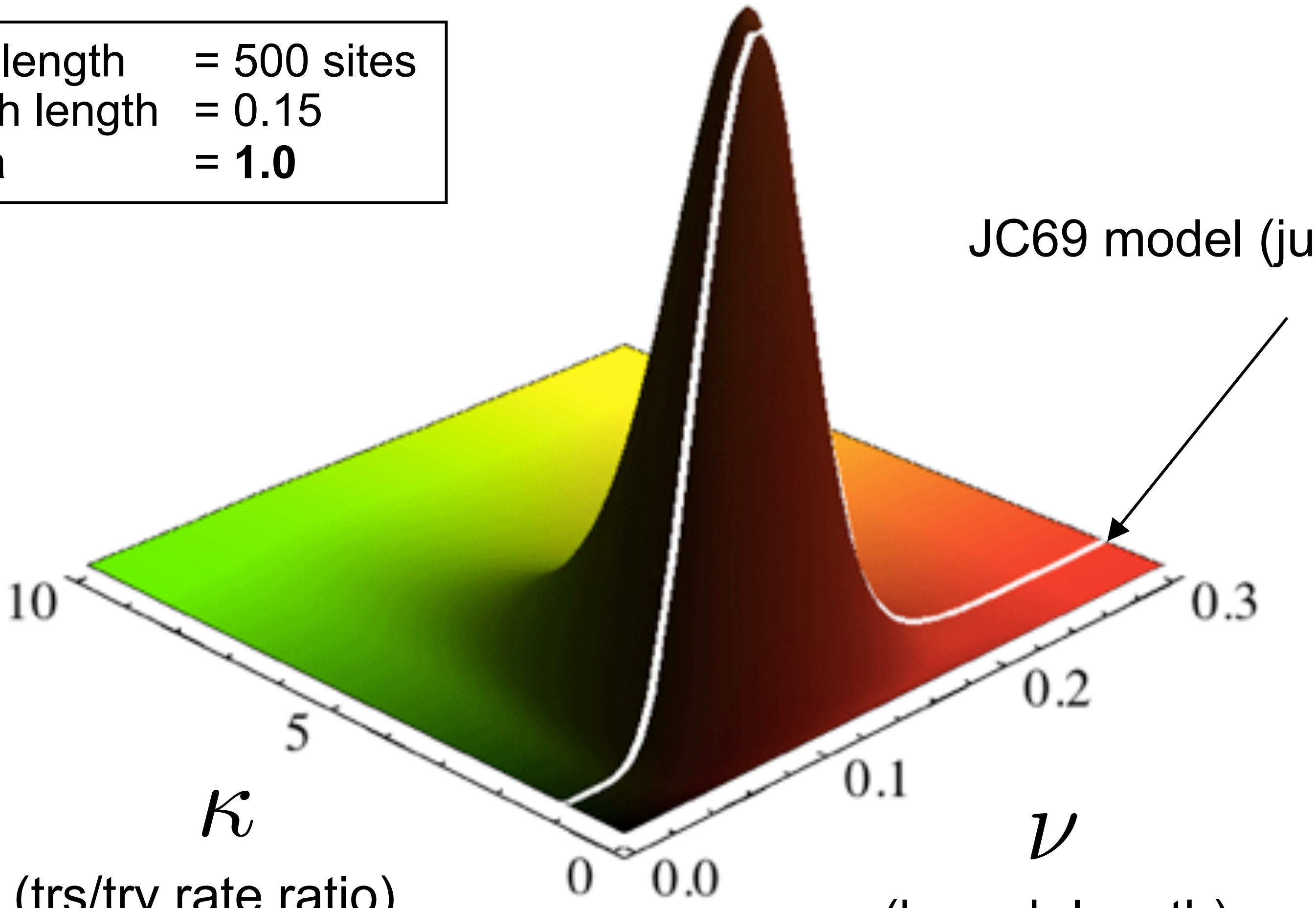
Likelihood Surface when JC true

Based on simulated data:

sequence length	= 500 sites
true branch length	= 0.15
true kappa	= **1.0**

K80 model (entire 2d space)

JC69 model (just this 1d line)



(trs/trv rate ratio)

(branch length)

F81 rate matrix

4 parameters:

μ

π_A

π_C

π_G

$$\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & -\mu(1-\pi_A) & \pi_C\mu & \pi_G\mu & \pi_T\mu \\
C & \pi_A\mu & -\mu(1-\pi_C) & \pi_G\mu & \pi_T\mu \\
G & \pi_A\mu & \pi_C\mu & -\mu(1-\pi_G) & \pi_T\mu \\
T & \pi_A\mu & \pi_C\mu & \pi_G\mu & -\mu(1-\pi_T)
\end{array}$$

Note: the F81 model is identical to the JC69 model if all base frequencies are equal

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368-376.

HKY85 rate matrix

5 parameters:
κ
β
π_A
π_C
π_G

$$\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & - & \pi_C\beta & \pi_G\beta\kappa & \pi_T\beta \\
C & \pi_A\beta & - & \pi_G\beta & \pi_T\beta\kappa \\
G & \pi_A\beta\kappa & \pi_C\beta & - & \pi_T\beta \\
T & \pi_A\beta & \pi_C\beta\kappa & \pi_G\beta & -
\end{array}$$

A dash means equal to negative sum of other elements on the same row

Note: the HKY85 model is identical to the F81 model if $\kappa = 1$. If, in addition, all base frequencies are equal, it is identical to JC69.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 21:160-174.

GTR rate matrix

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & - & \pi_C a\mu & \pi_G b\mu & \pi_T c\mu \\ C & \pi_A a\mu & - & \pi_G d\mu & \pi_T e\mu \\ G & \pi_A b\mu & \pi_C d\mu & - & \pi_T f\mu \\ T & \pi_A c\mu & \pi_C e\mu & \pi_G f\mu & - \end{array}$$

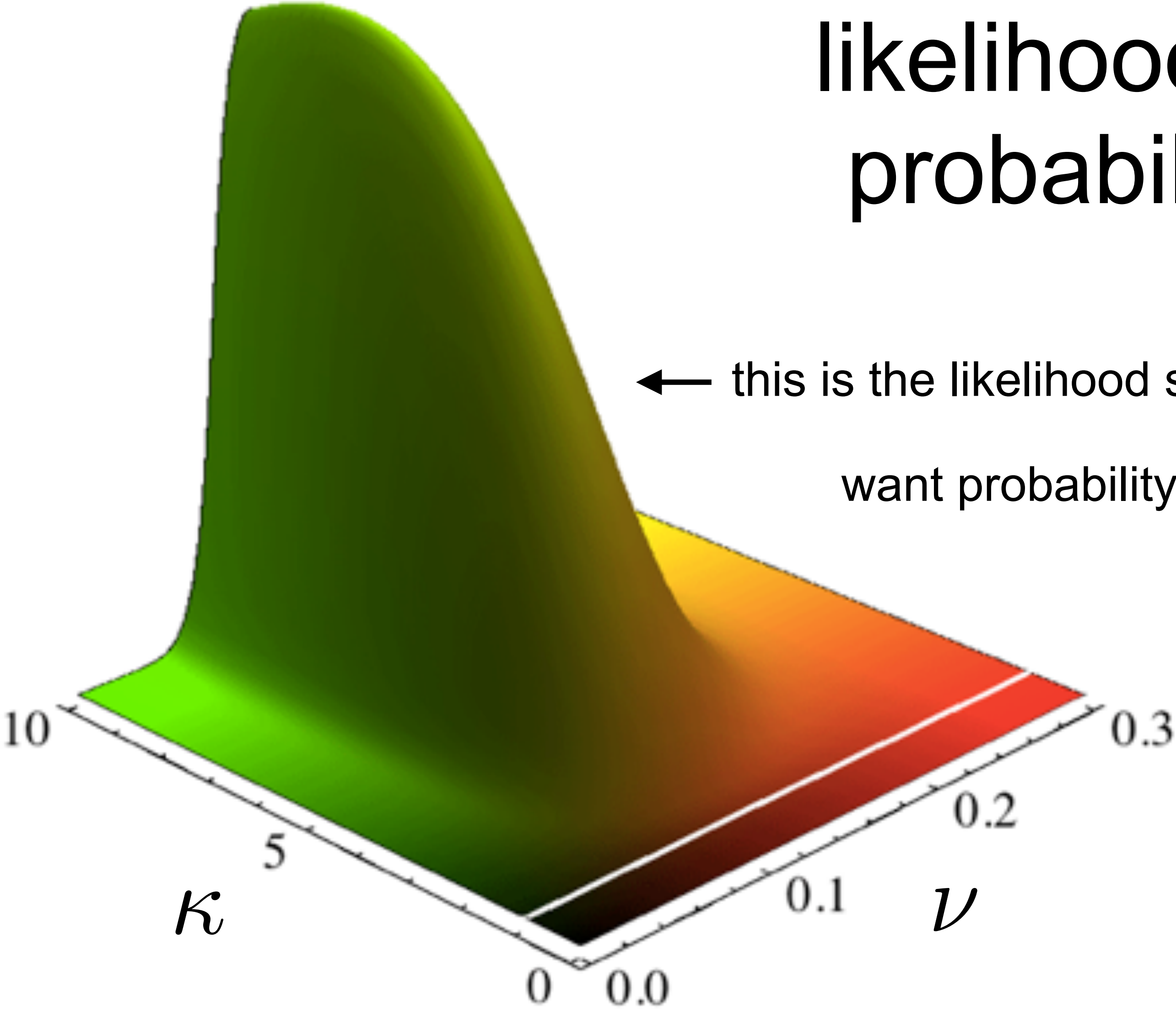
9 parameters:

π_A
π_C
π_G
a
b
c
d
e
μ

Identical to the F81 model if $a = b = c = d = e = f = 1$. If, in addition, all the base frequencies are equal, GTR is identical to JC69. If $a = c = d = f = \beta$ and $b = e = \kappa\beta$, GTR becomes the HKY85 model.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86-93.

So, how do you turn likelihoods into probabilities?

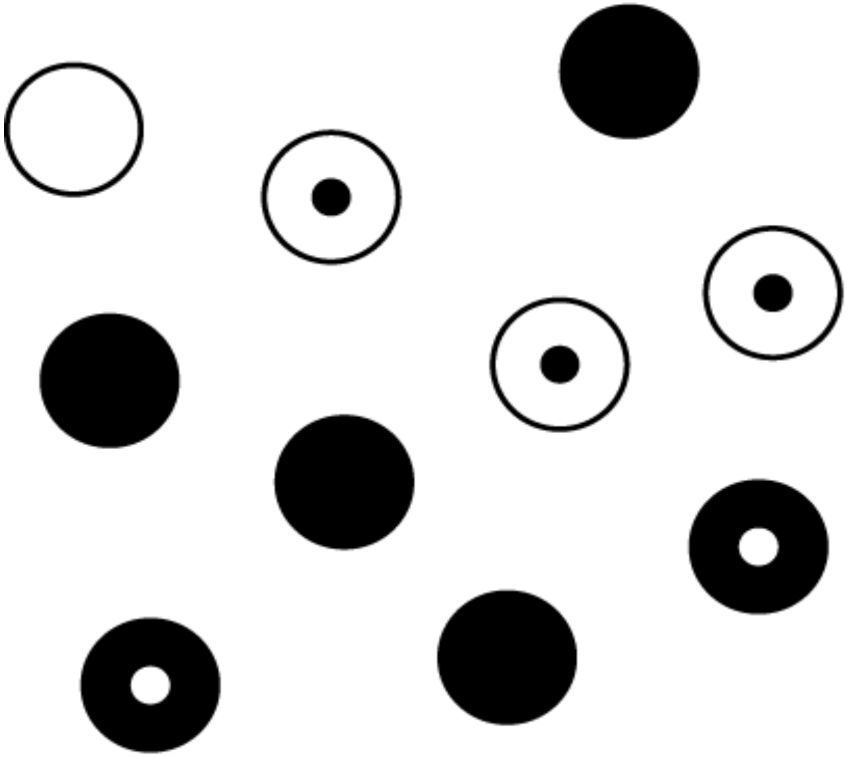


← this is the likelihood surface: Pr(data|κ,v)

want probability surface: Pr(κ,v|data)

Kinds of probabilities

B	**=**	**Black**	**S**	**=**	**Solid**
W	**=**	**White**	**D**	**=**	**Dotted**



Marginal probabilities:

Pr(B) = 0.6 Pr(S) = 0.5

Pr(W) = 0.4 Pr(D) = 0.5

Joint probabilities:

Pr(◙) = Pr(B, D) = 0.2

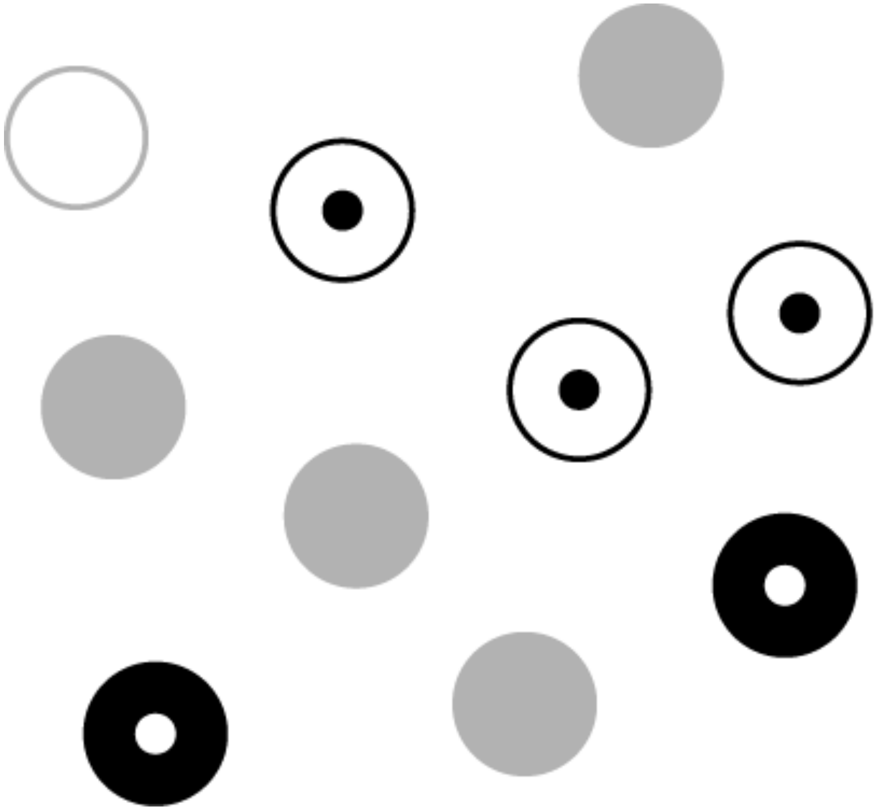
Pr(●) = Pr(B, S) = 0.4

Pr(⊙) = Pr(W, D) = 0.3

Pr(○) = Pr(W, S) = 0.1

Kinds of probabilities (continued)

Conditional probability



$$\Pr(B|D) = \frac{2}{5} = 0.4$$

Hide all solid marbles (leaving 5 with dot)

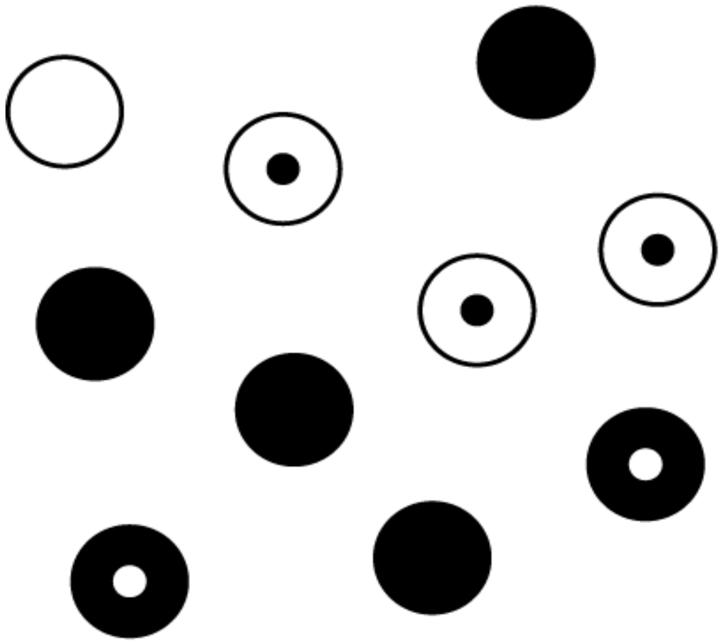
Of those left, 2 are black

Bayes' rule provides a way to calculate conditional probabilities

Pr(B, D)

$$\Pr(D)\,\Pr(B|D) = \Pr(B)\,\Pr(D|B)$$

$$\frac{1}{2} \times \frac{2}{5} = \frac{3}{5} \times \frac{1}{3}$$



$$\Pr(B|D) = \frac{\Pr(B)\,\Pr(D|B)}{\Pr(D)}$$

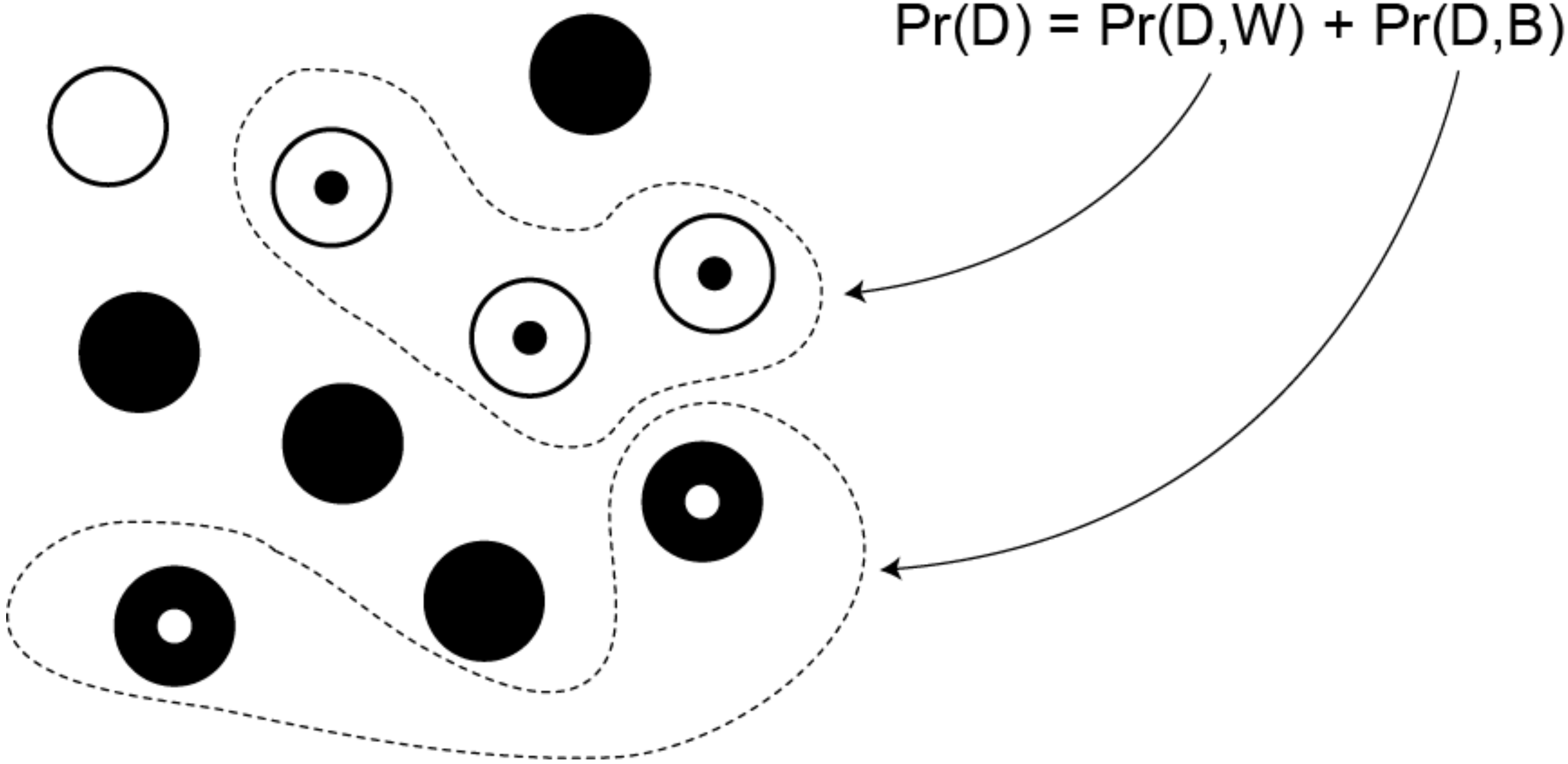
$$= \frac{\frac{3}{5} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{5}$$

Bayes' rule shows how to turn Pr(D|B) into Pr(B|D)

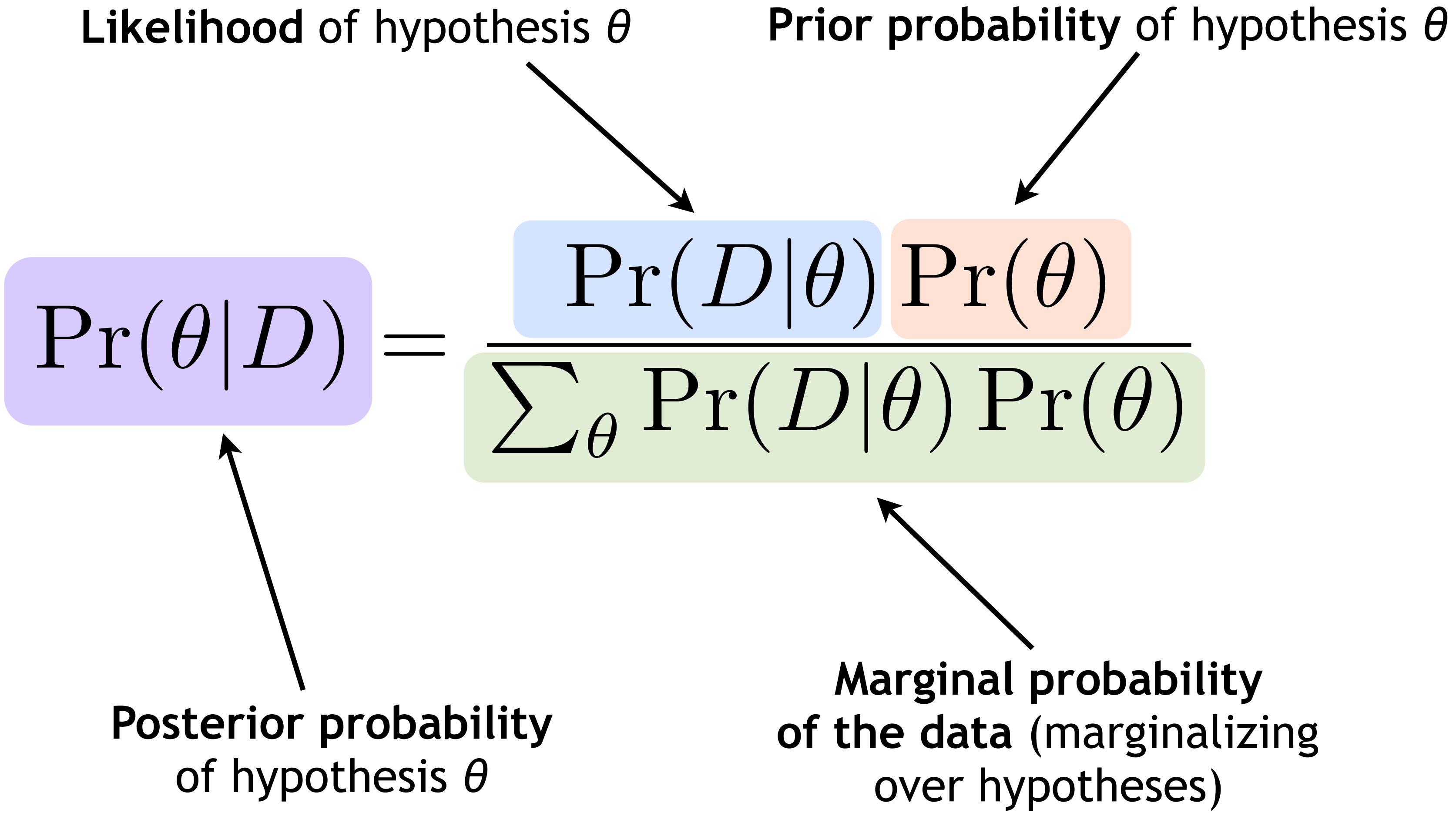
$$\Pr(B|D) = \frac{\Pr(B)\Pr(D|B)}{\Pr(D)}$$

The marginal probability of D is the sum of all joint probabilities involving D

$$\Pr(D) = \Pr(D,W) + \Pr(D,B)$$



Bayes' rule in statistics



$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

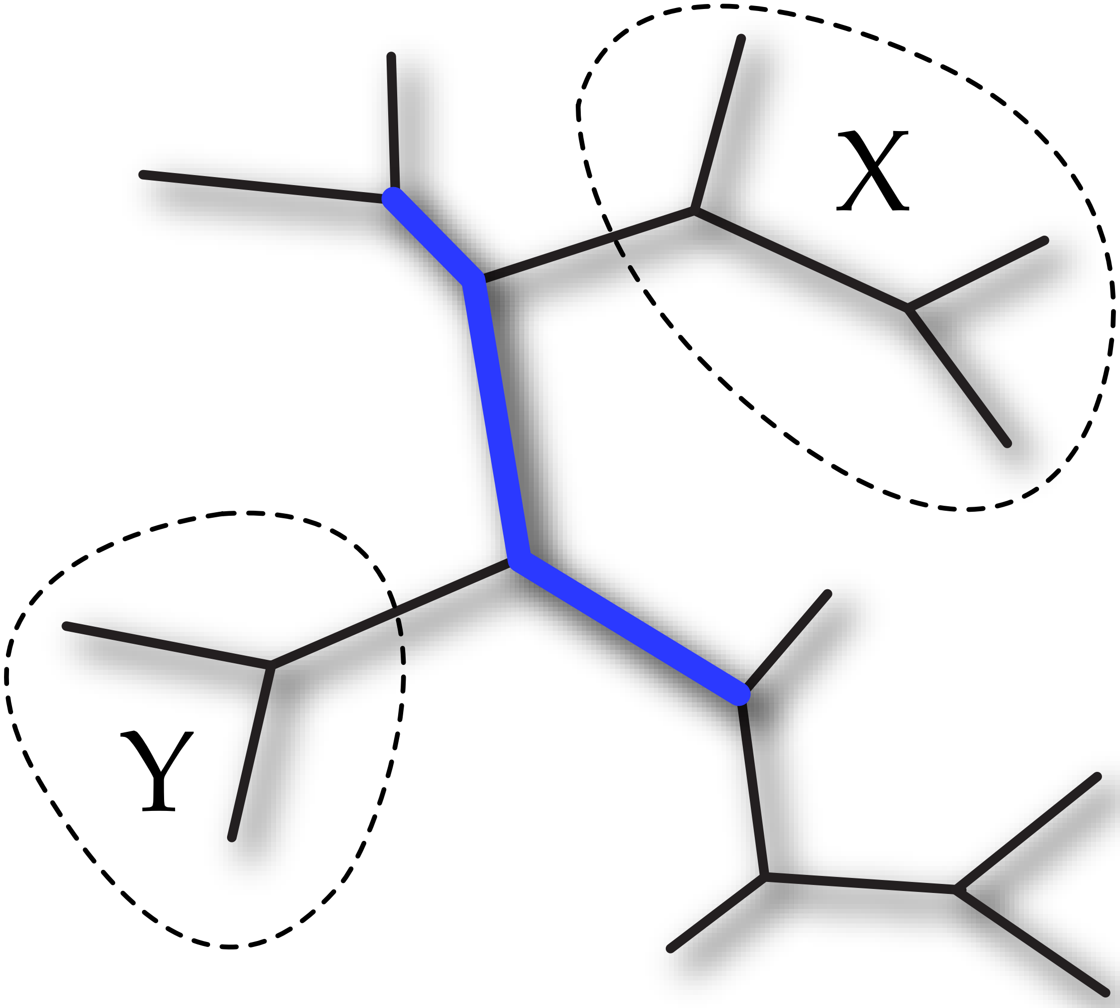
Likelihood of hypothesis θ

Prior probability of hypothesis θ

Posterior probability of hypothesis θ

Marginal probability of the data (marginalizing over hypotheses)

Moving through treespace

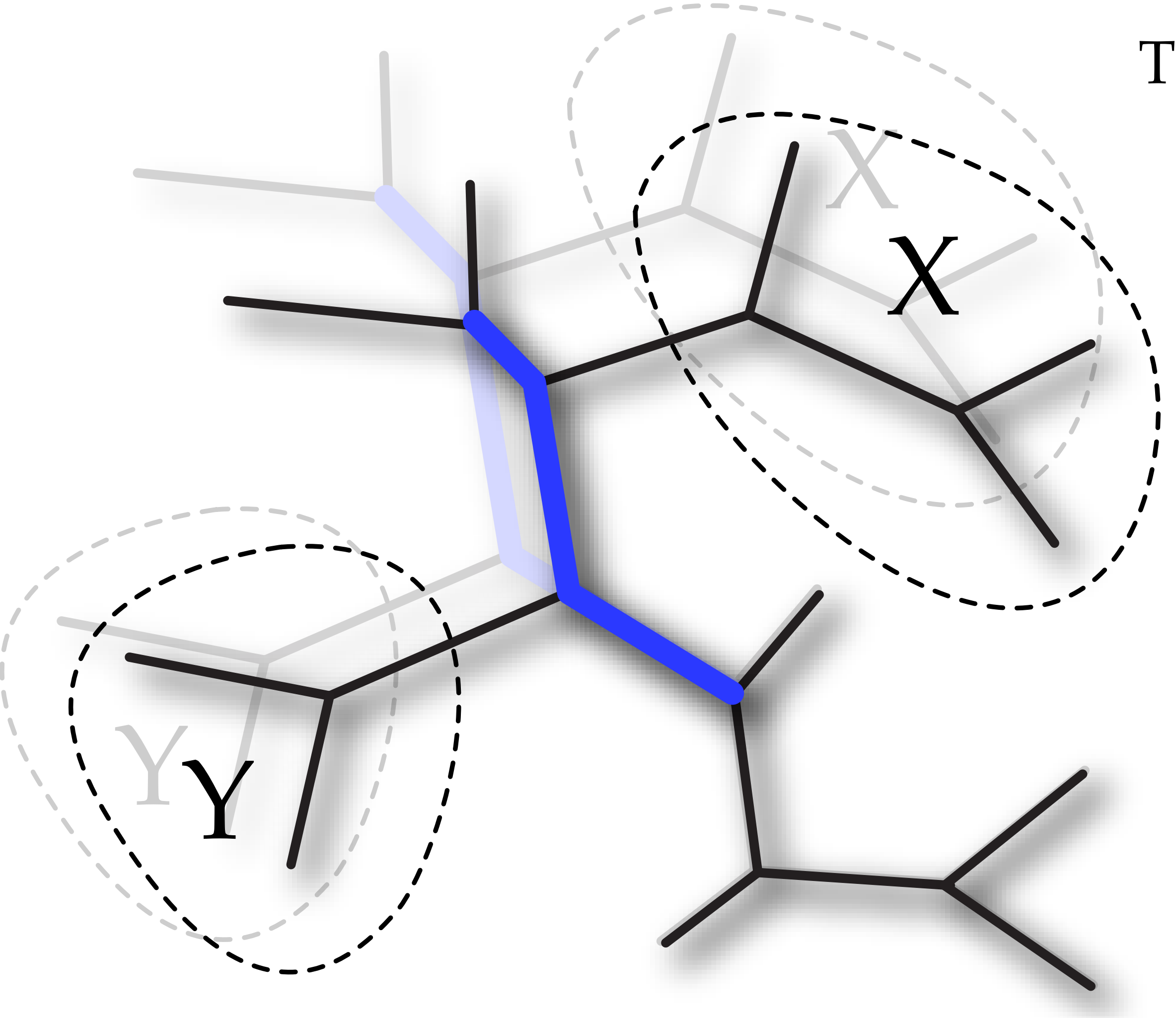


The Larget-Simon move

Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

*Larget, B., and D. L. Simon. 1999. Markov chain monte carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16: 750-759. See also: Holder et al. 2005. Syst. Biol. 54: 961-965.

Moving through treespace

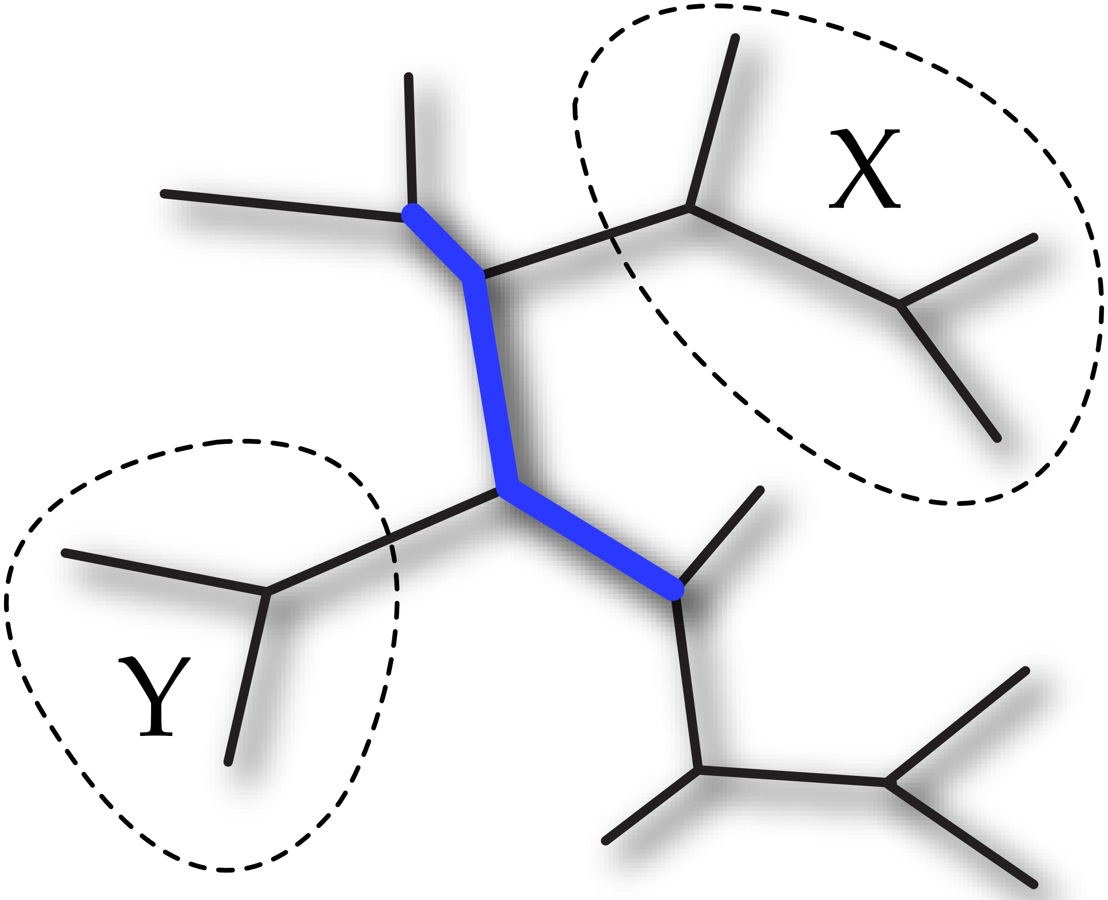


The Larget-Simon move

Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
Shrink or grow selected 3-edge segment by a random amount

Moving through treespace

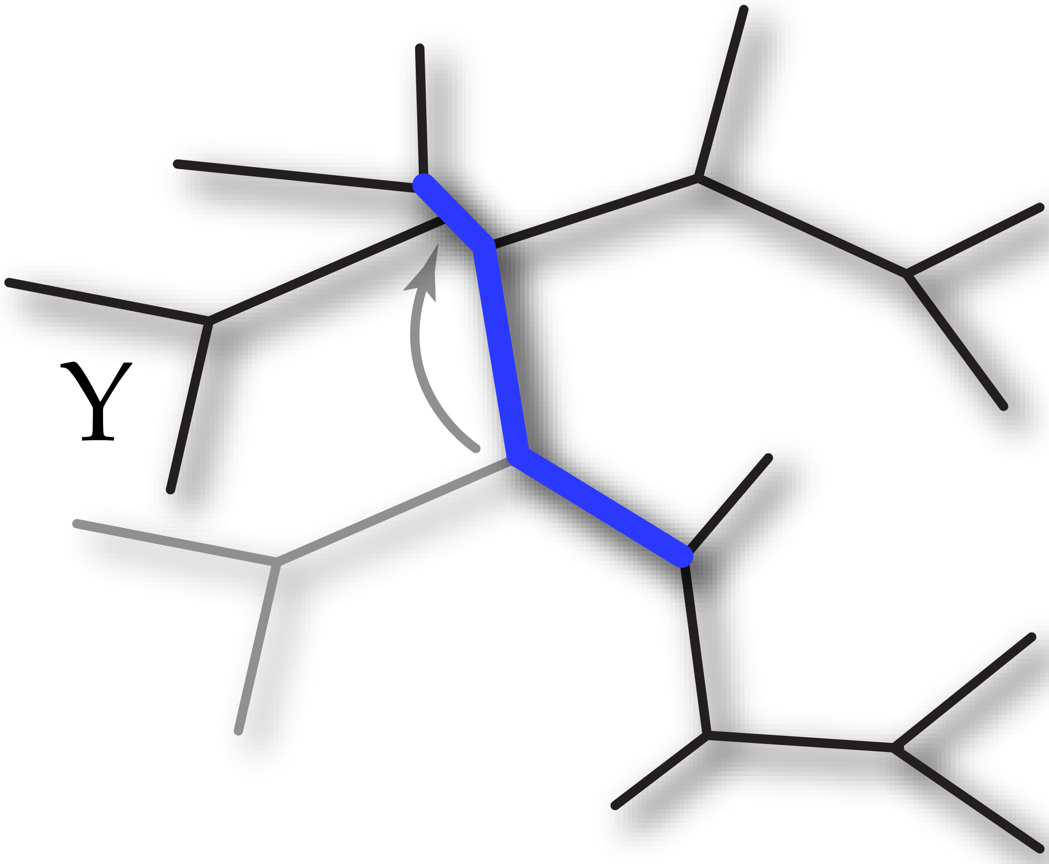


The Larget-Simon move

Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
Shrink or grow selected 3-edge segment by a random amount

Moving through treespace



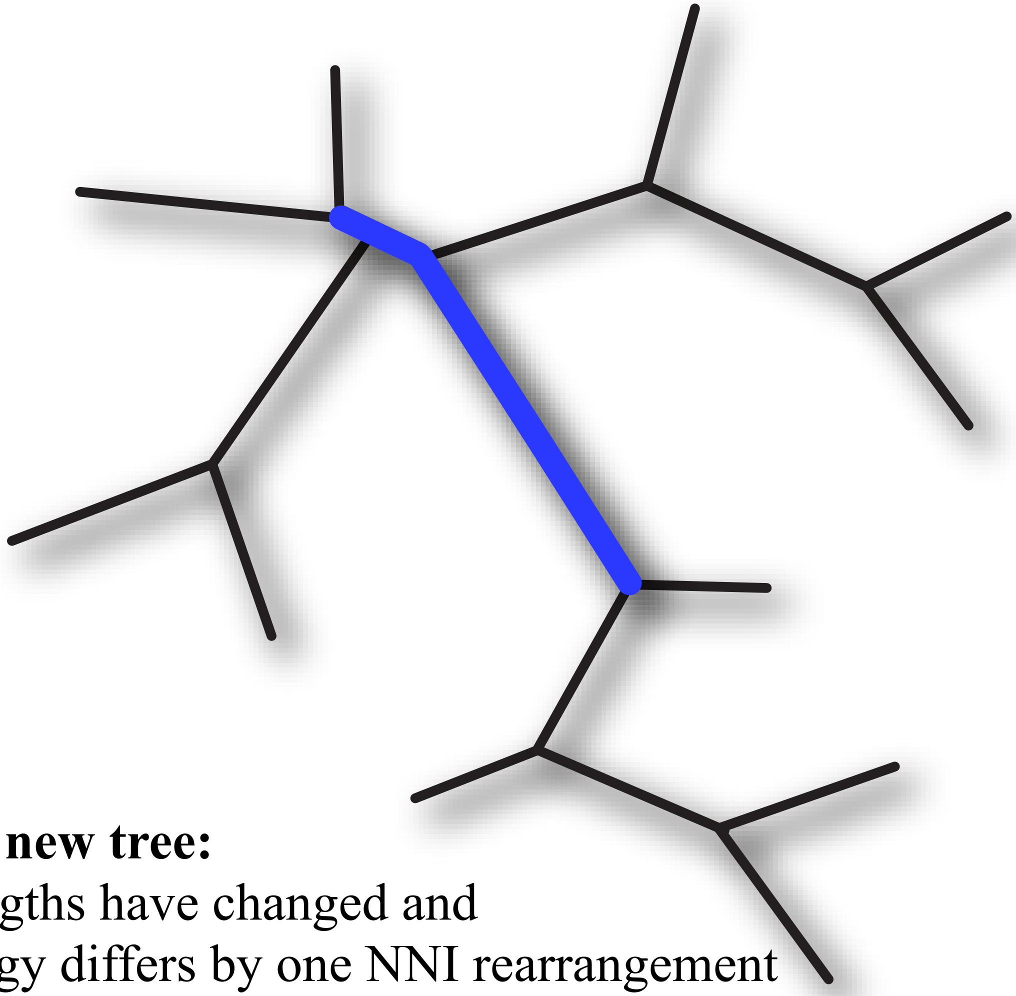
The Larget-Simon move

Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
Shrink or grow selected 3-edge segment by a random amount

Step 3:
Choose X or Y randomly, then reposition randomly

Moving through treespace



The Large-Simon move

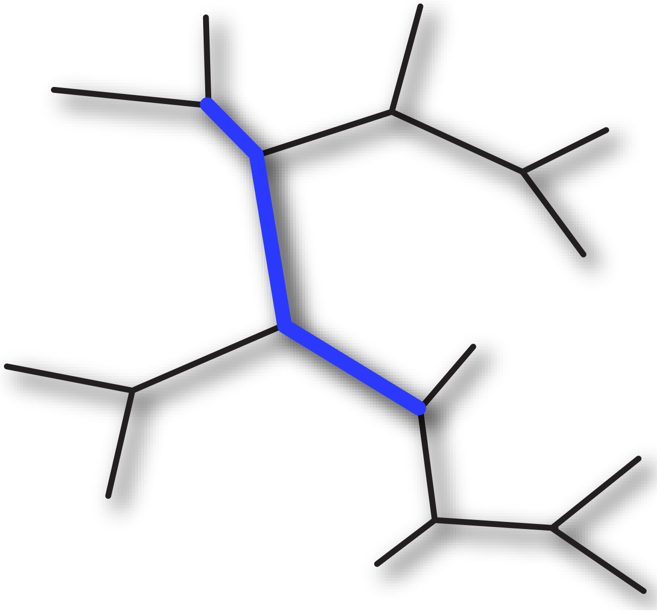
Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
Shrink or grow selected 3-edge segment by a random amount

Step 3:
Choose X or Y randomly, then reposition randomly

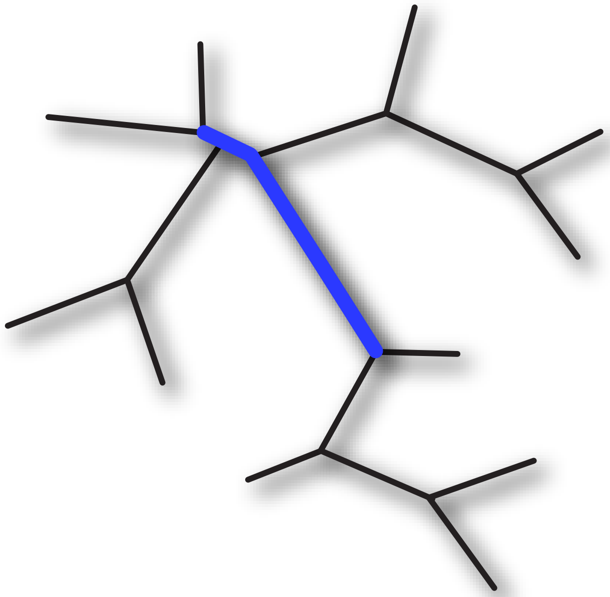
Proposed new tree:
3 edge lengths have changed and
the topology differs by one NNI rearrangement

Moving through treespace



Current tree

log-posterior = -34256



Proposed tree

log-posterior = -32519
(better, so accept)