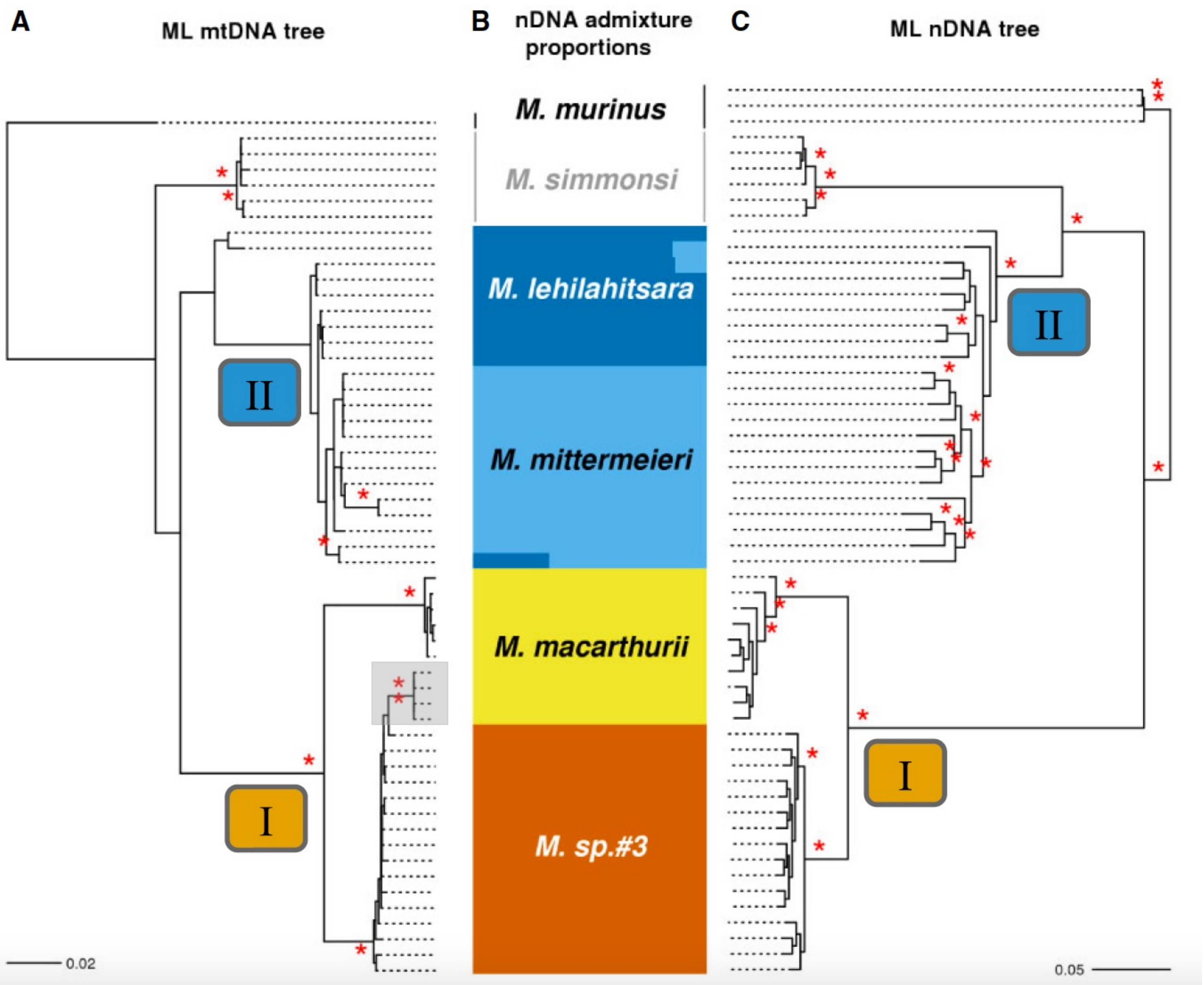
Phylogenetics: A Maximum Likelihood Approach

George P. Tiley
University of Antananarivo

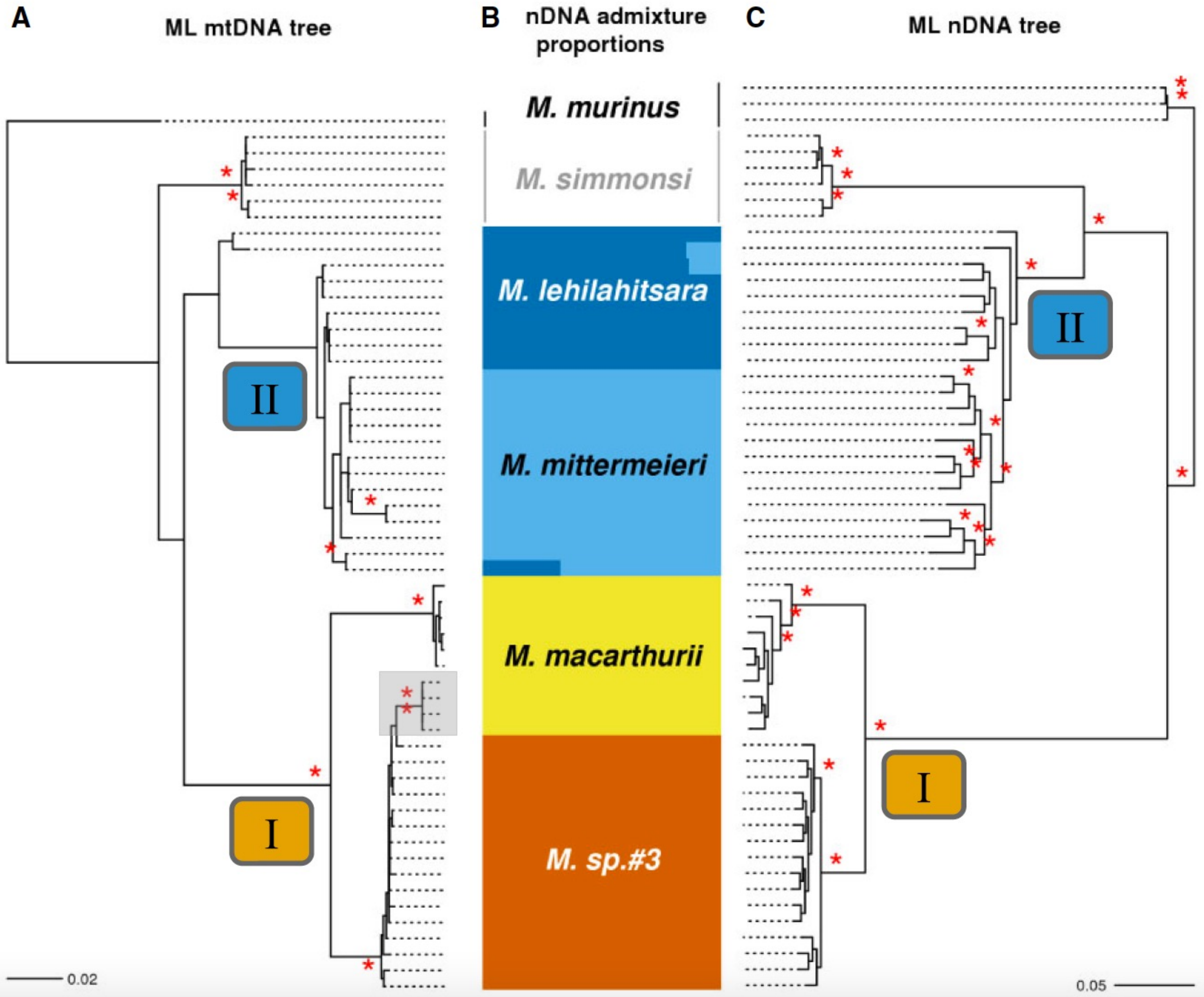
DBEV Phylogenomics Workshop
7 March 2022

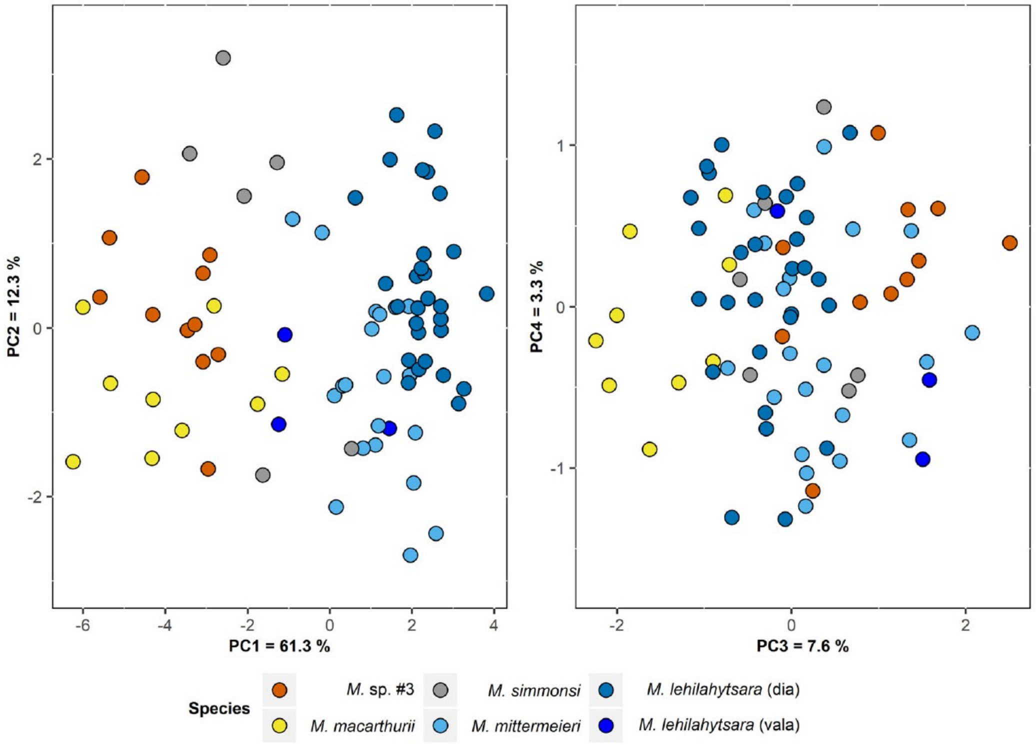
Motivation: Why Phylogenies?

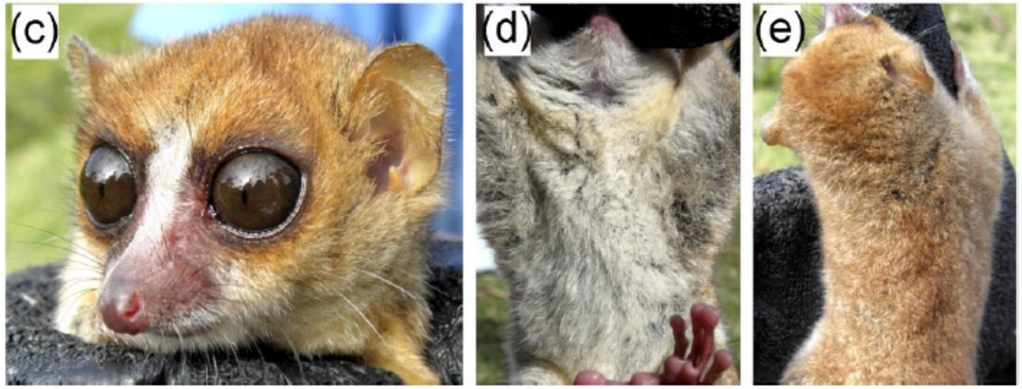


Poelstra et al. 2021

Motivation: Why *Molecular* Phylogenies?







Poelstra et al. 2021; Schüßler et al. 2020

Motivation: Why *Molecular* Phylogenies

Valuable for taxonomy

Informative about trait evolution

Used for divergence time estimation

Understand biogeographic patterns

Many More!

Learning Goals

Explain terminology

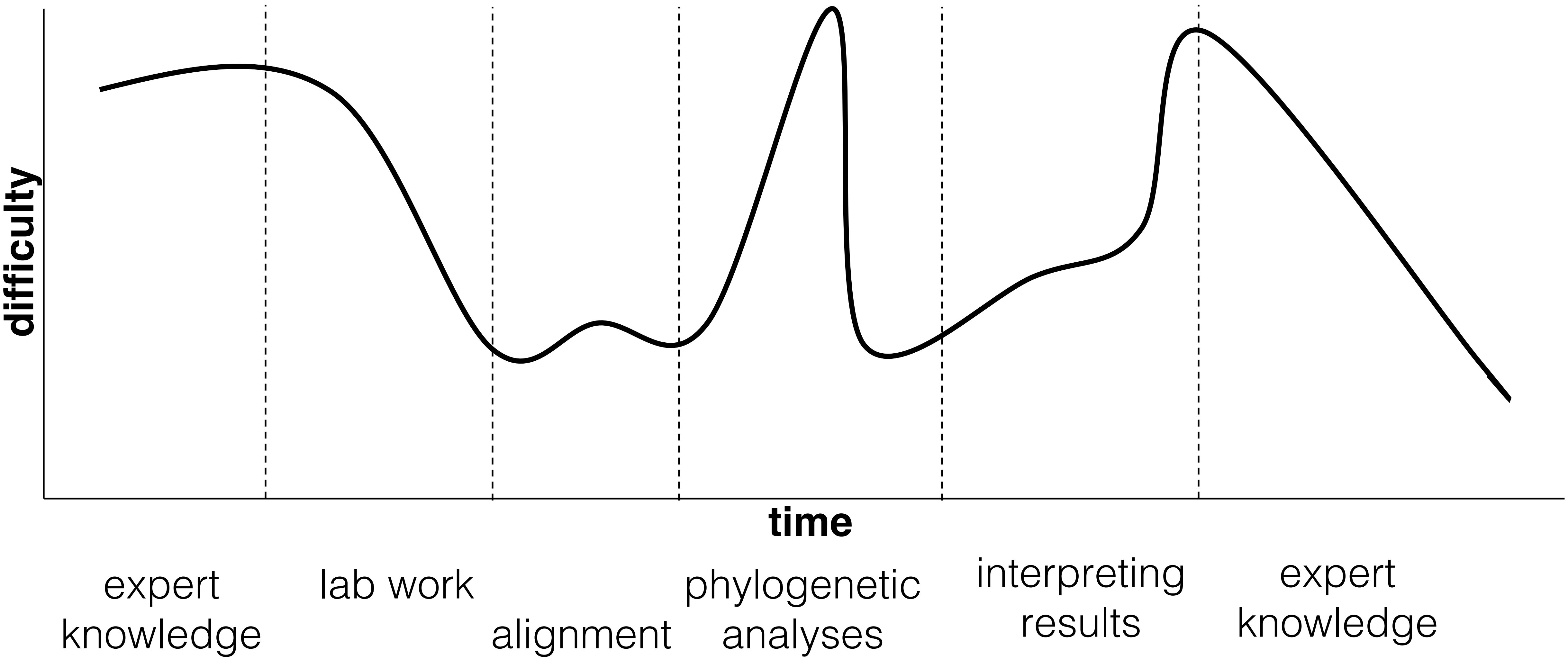
Primer on probability and likelihood

Models of molecular evolution

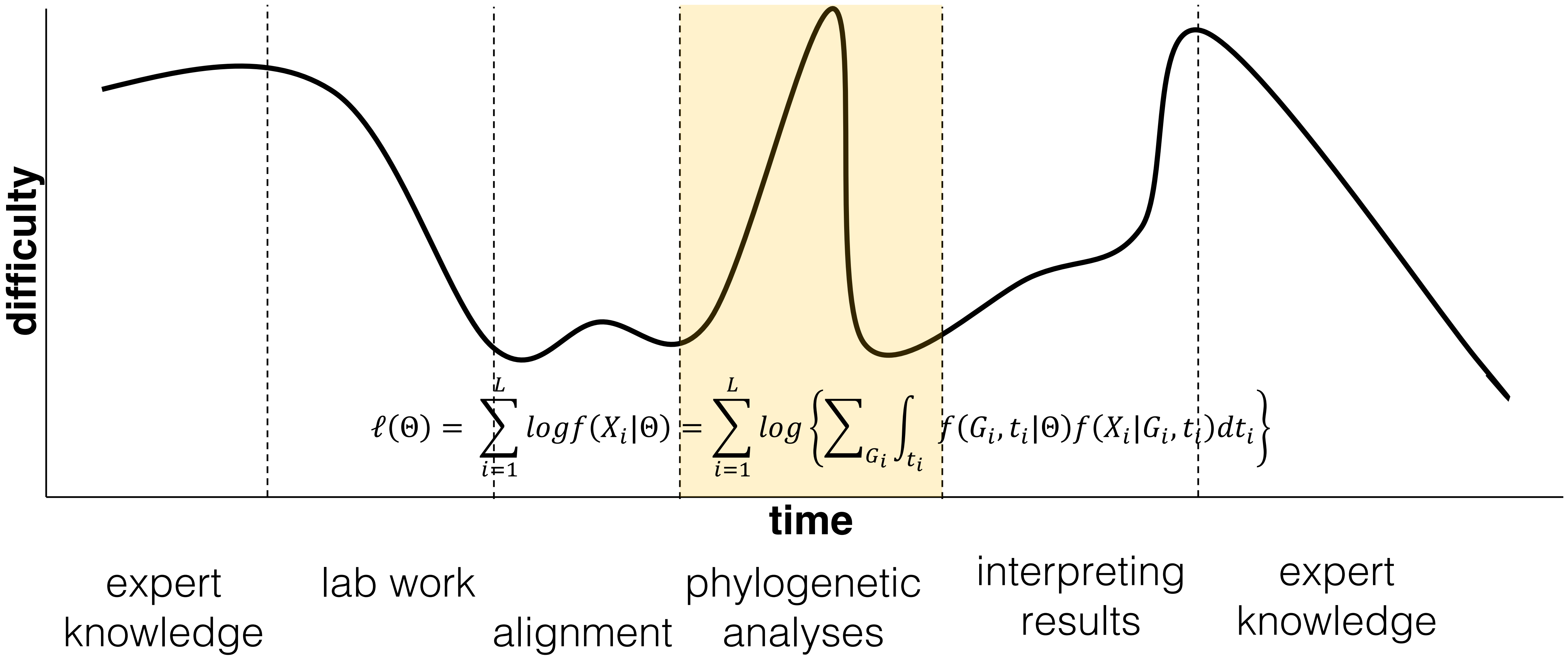
How to select a model

Application of models for phylogenetic estimation

Learning Goals



Learning Goals



$$\ell(\Theta) = \sum_{i=1}^{L} log f(X_i|\Theta) = \sum_{i=1}^{L} log\left\{\sum_{G_i}\int_{t_i} f(G_i,t_i|\Theta)f(X_i|G_i,t_i)dt_i\right\}$$

Learning Goals

Explain terminology

Primer on probability and likelihood

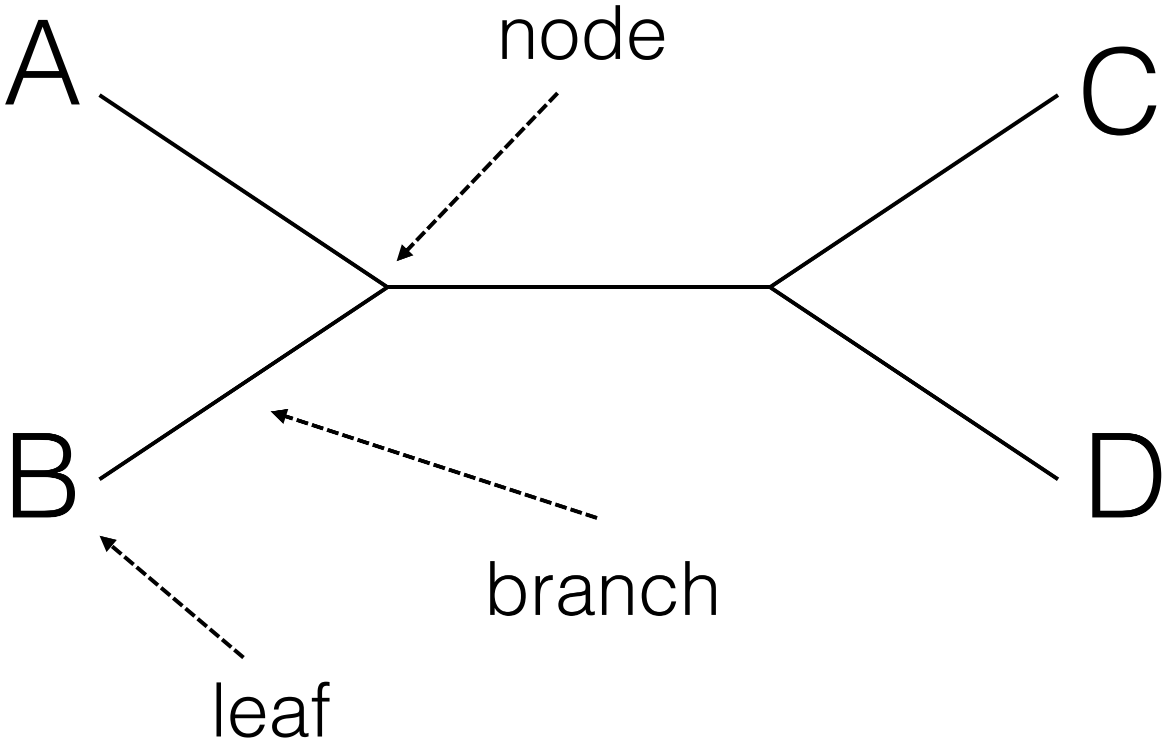
Models of molecular evolution

How to select a model

Application of models for phylogenetic estimation

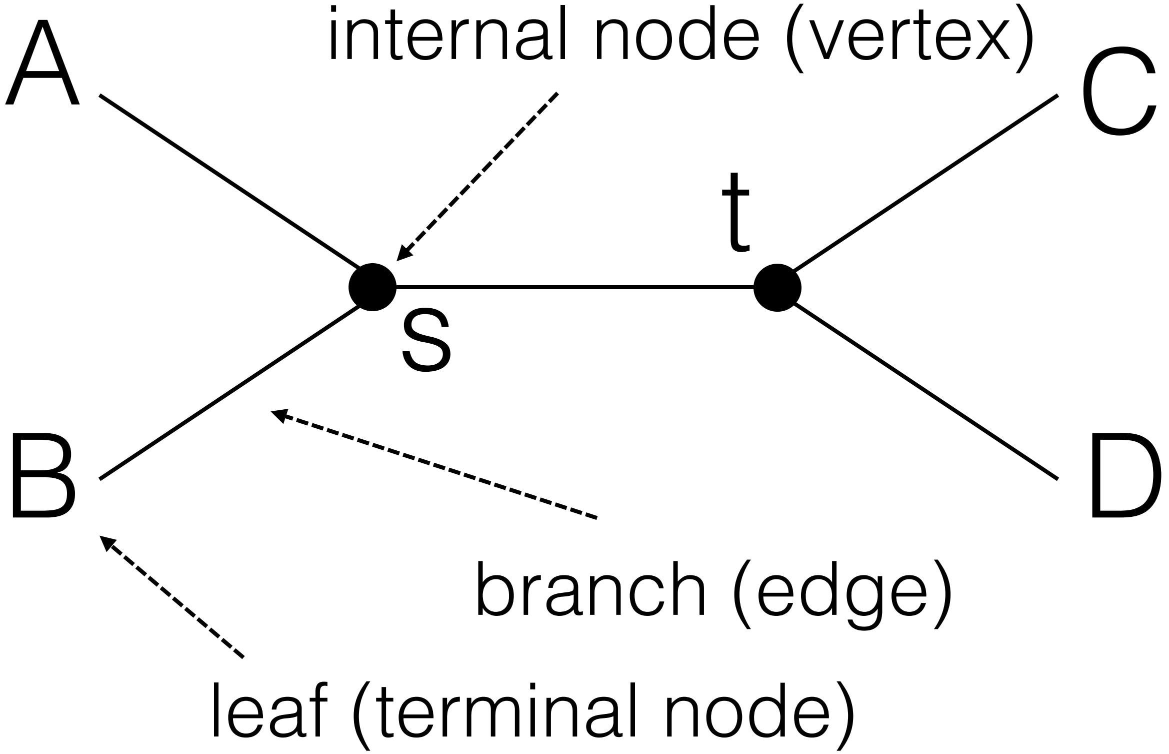
Terminology

Unrooted Tree



Terminology

Unrooted Tree

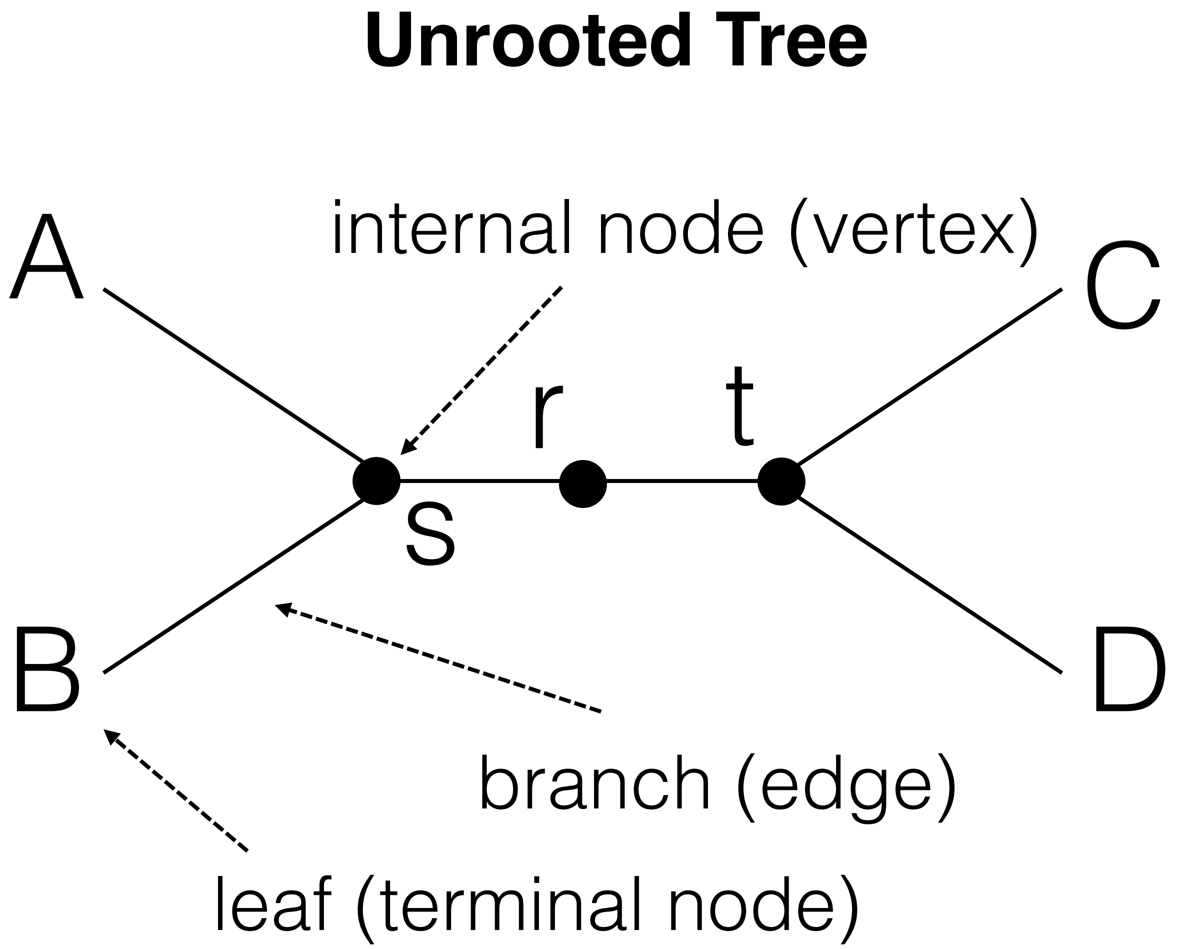


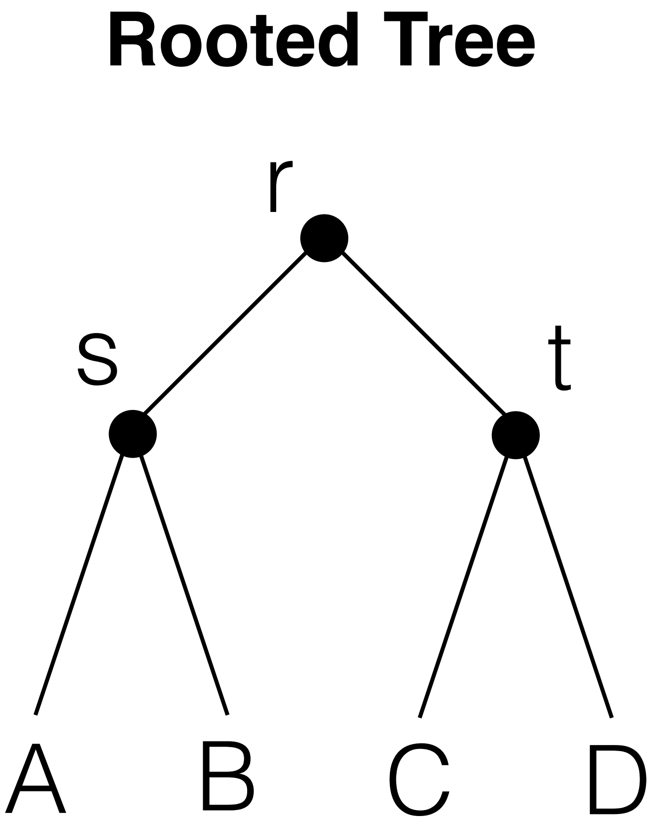
internal node (vertex)

branch (edge)

leaf (terminal node)

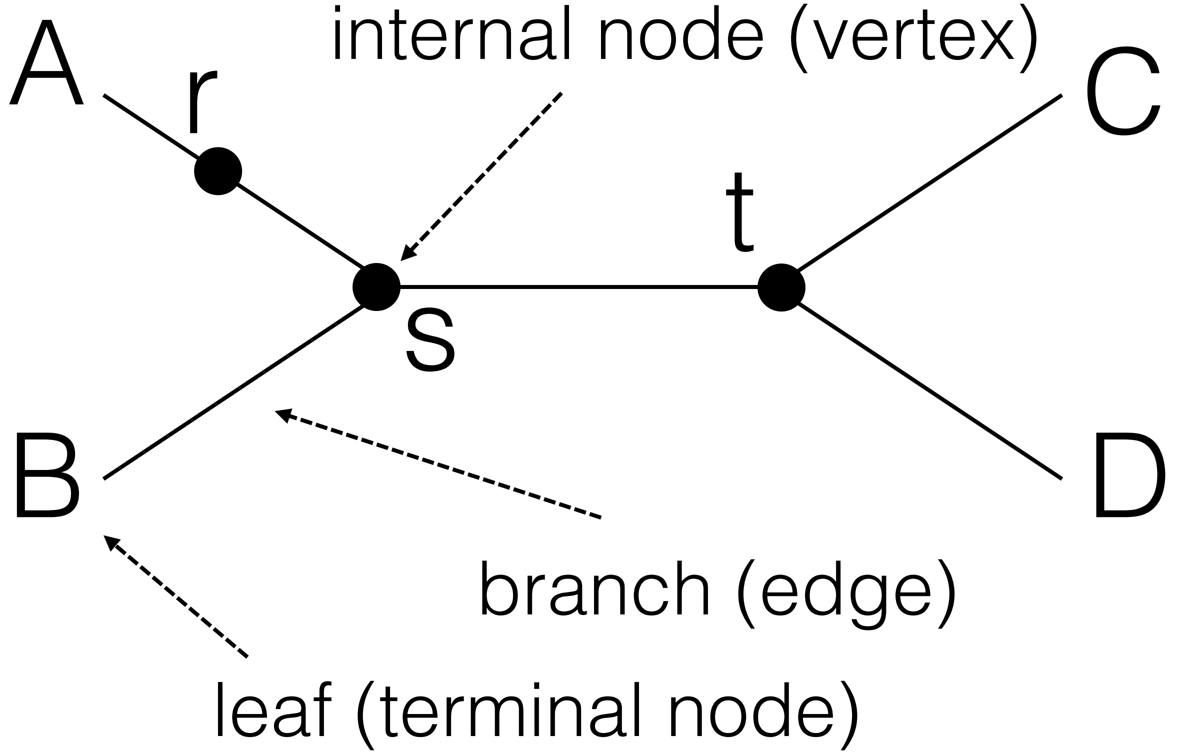
Terminology





Terminology

Unrooted Tree

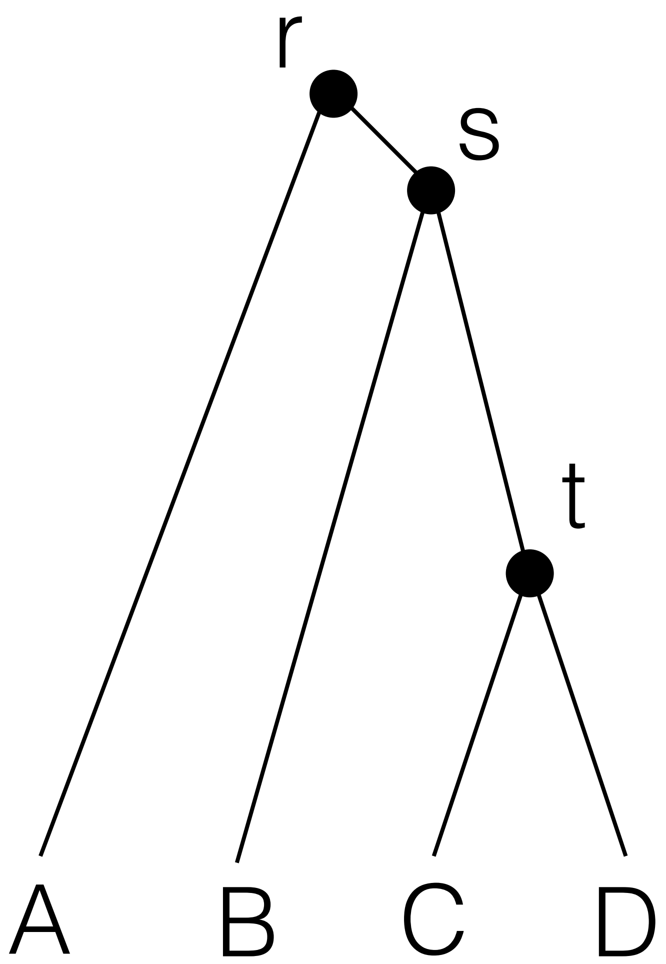


internal node (vertex)

branch (edge)

leaf (terminal node)

Rooted Tree



Learning Goals

Explain terminology

Primer on probability and likelihood

Models of molecular evolution

How to select a model

Application of models for phylogenetic estimation

Probability and Likelihood

Rules of probability

Combining multiple *independent* events

AND
(x)

OR
(+)

Probability and Likelihood

Rules of probability

Combining multiple *independent* events

AND
(x)

Probability I roll a 1 and 2?

OR
(+)

Probability I roll a 1 or 2?

Probability and Likelihood

Rules of probability

Combining multiple *independent* events

AND
(x)

Probability I roll a 1 and 2?

Pr{1 and 2} = 1/10 x 1/10 = 1/100

OR
(+)

Probability I roll a 1 or 2?

Pr{1 or 2} = 1/10 + 1/10 = 2/10 = 1/5

Probability and Likelihood

Difference between probability and likelihood?

Consider I rolled a 2 and 1 and 1 and 4. Is this a surprising result?

$$\Pr\{2,1,1,4 \mid \text{10-sided die}\} = 1/10 \times 1/10 \times 1/10 \times 1/10 = 1/10000$$

Probability and Likelihood

Difference between probability and likelihood?

Consider I rolled a 2 and 1 and 1 and 4. Is this a surprising result?

Pr{2,1,1,4 | 10-sided die} = 1/10 x 1/10 x 1/10 x 1/10 = 1/10000

Maybe I used a 4-sided die. Is this less surprising?

Pr{2,1,1,4 | 4-sided die} = 1/4 x 1/4 x 1/4 x 1/4 = 1/256

Probability and Likelihood

Difference between probability and likelihood?

Consider I rolled a 2 and 1 and 1 and 4. Is this a surprising result?

Pr{2,1,1,4 | 10-sided die} = 1/10 x 1/10 x 1/10 x 1/10 = 1/10000

Maybe I used a 4-sided die. Is this less surprising?

Pr{2,1,1,4 | 4-sided die} = 1/4 x 1/4 x 1/4 x 1/4 = 1/256

These are models!

Probability and Likelihood

Likelihood – probability of observed data with respect to a particular model

Likelihood(4-sided die) = Pr{2,1,1,4 | 4-sided die} = ¼ x ¼ x ¼ x ¼

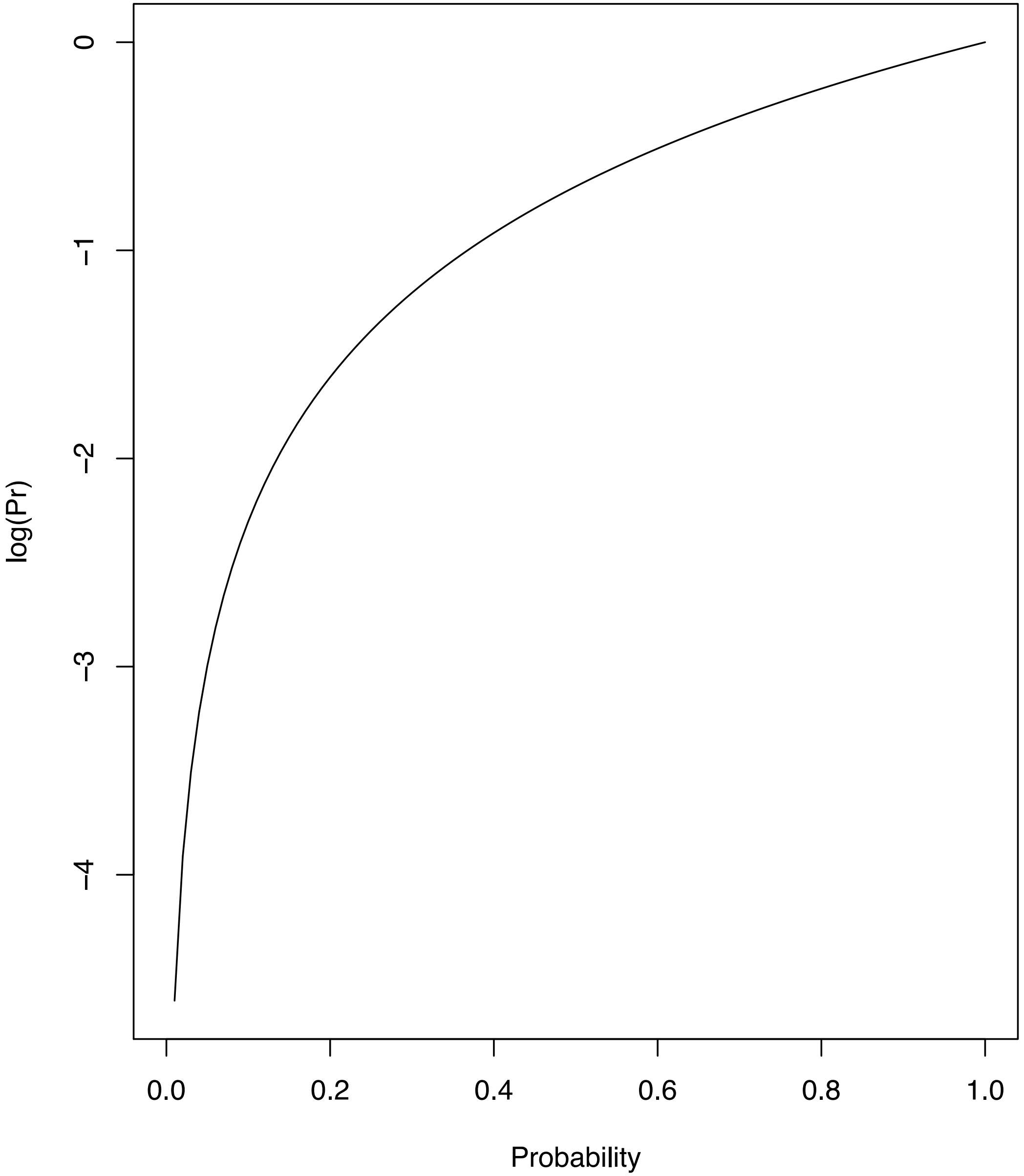
Likelihood(4-sided die) = $f(X|4-sided\ die)\ =\ \prod_{h=1}^{n} f(x_h|4-sided\ die)$

Likelihood(θ) = $f(X|\theta)\ =\ \prod_{h=1}^{n} f(x_h|\theta)$

$\mathrm{L}(\theta) = f(X|\theta)\ =\ \prod_{h=1}^{n} f(x_h|\theta)$

Probability and Likelihood

Likelihood – why do we always see *log*?



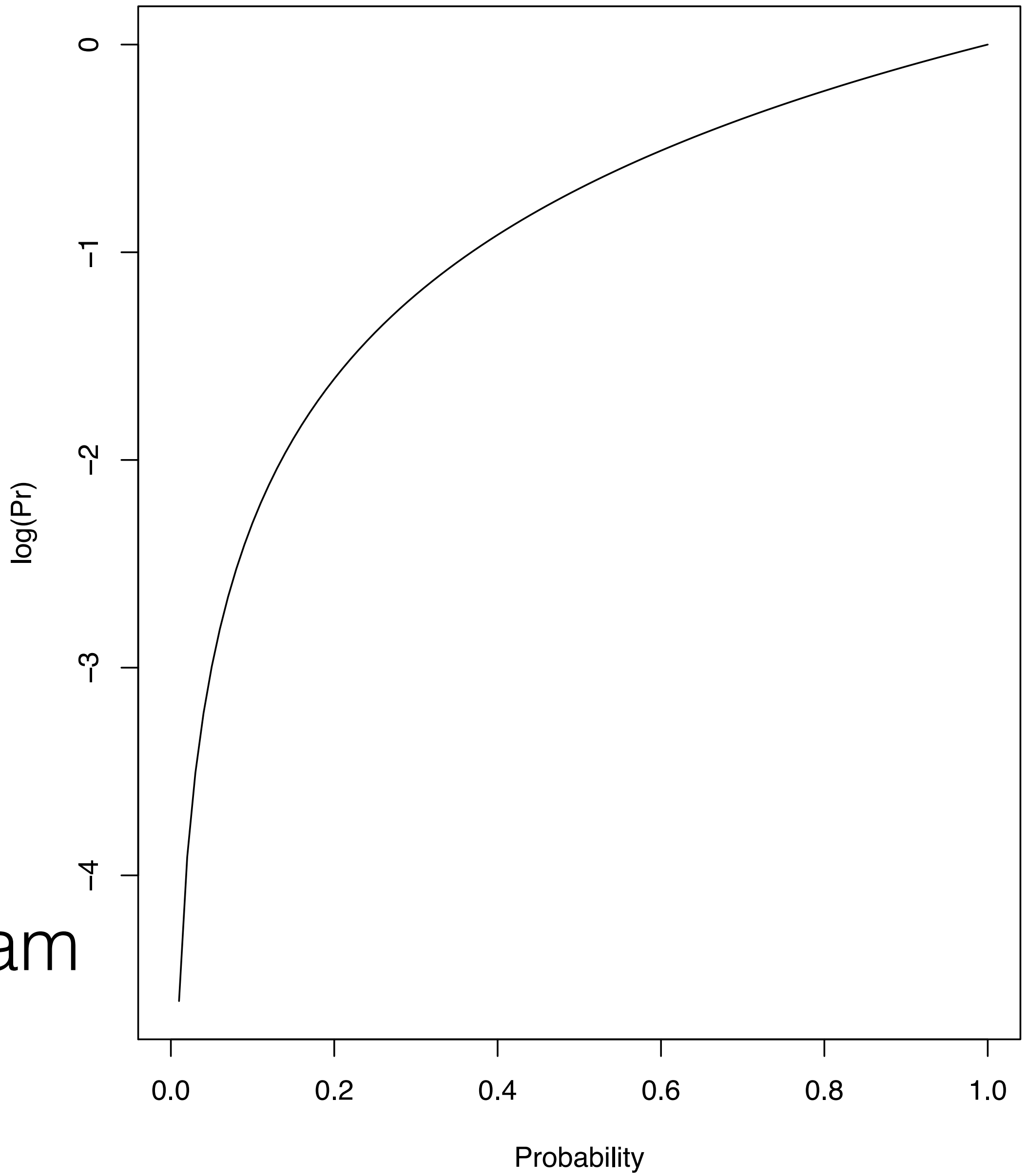
Probability and Likelihood

Likelihood – why do we always see *log*?

Many probabilities we calculate with phylogenies will be very small

Computers do not store small numbers accurately, usually to 16 decimal places

Thus we always see log(L) from a program where log(x) = ln(x) = 1/e^x



Probability and Likelihood

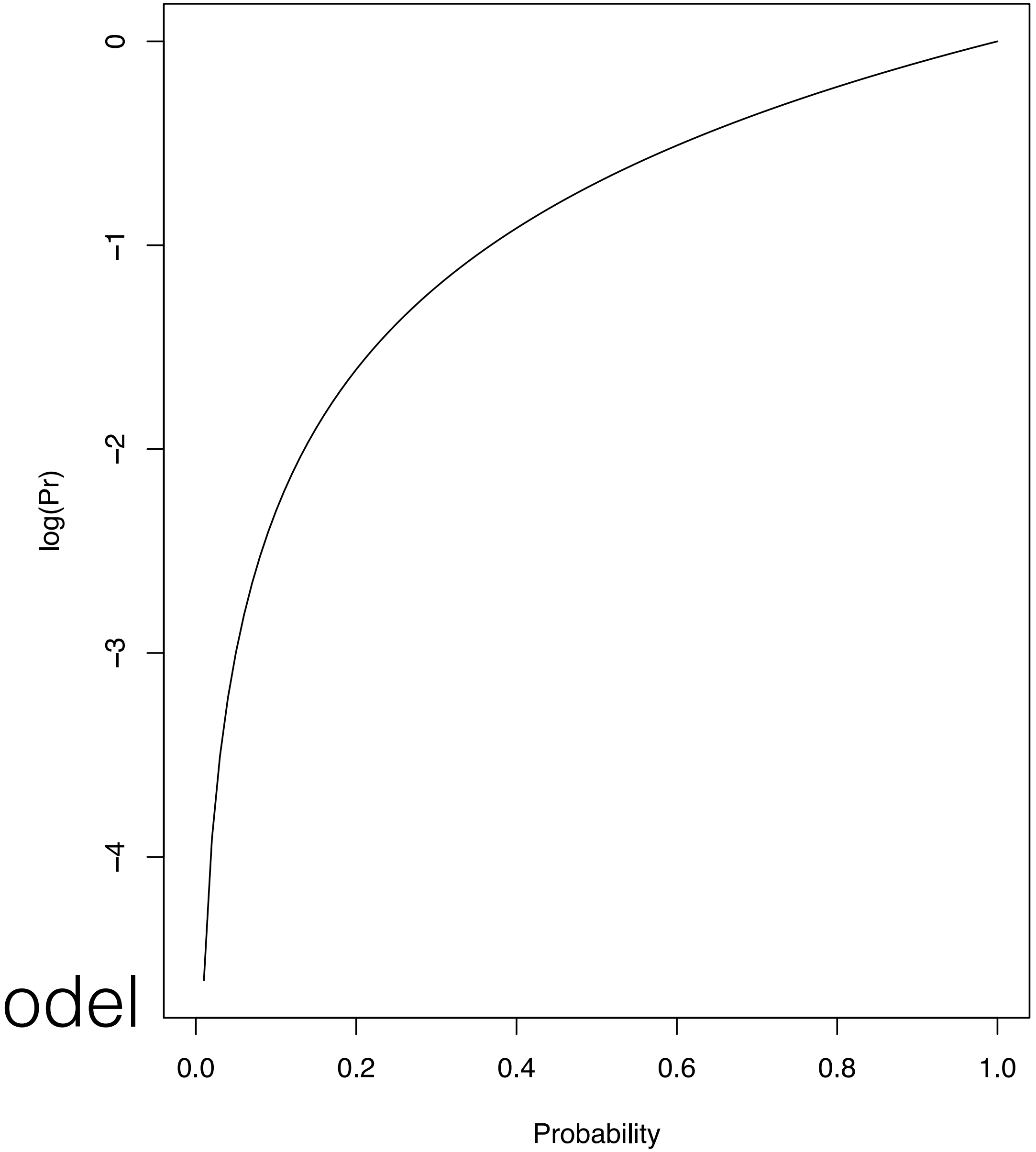
Likelihood – logs have rules

$$log(x \times y) = log(x) + \log(y)$$

Pr{2,1,1,4 | 10-sided die} = 1/10000
log(Pr{2,1,1,4 | 10-sided die}) = -9.21

Pr{2,1,1,4 | 4-sided die} = 1/256
log(Pr{2,1,1,4 | 4-sided die}) = -5.55

The *log-likelihood* is *maximized* for the model that surprises us the least



Probability and Likelihood

Likelihood – probability of observed data with respect to a particular model

Likelihood(4-sided die) = Pr{2,1,1,4 | 4-sided die} = ¼ x ¼ x ¼ x ¼

Likelihood(4-sided die) = $f(X|4 - sided\ die) \ = \prod_{h=1}^{n} f(x_h|4 - sided\ die)$

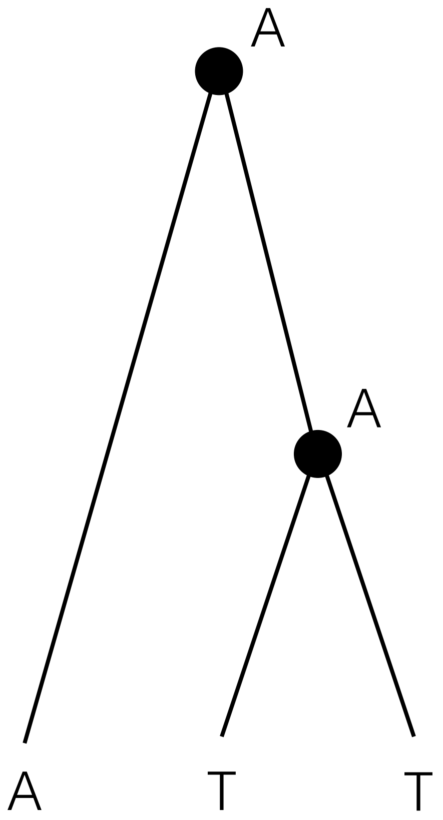
Likelihood(θ) = $f(X|\theta) \ = \prod_{h=1}^{n} f(x_h|\theta)$

L(θ) = $f(X|\theta) \ = \prod_{h=1}^{n} f(x_h|\theta)$

$$\ell = \log\{L(\theta)\} = \sum_{h=1}^{n} log\{f(x_h|\theta)\}$$

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)



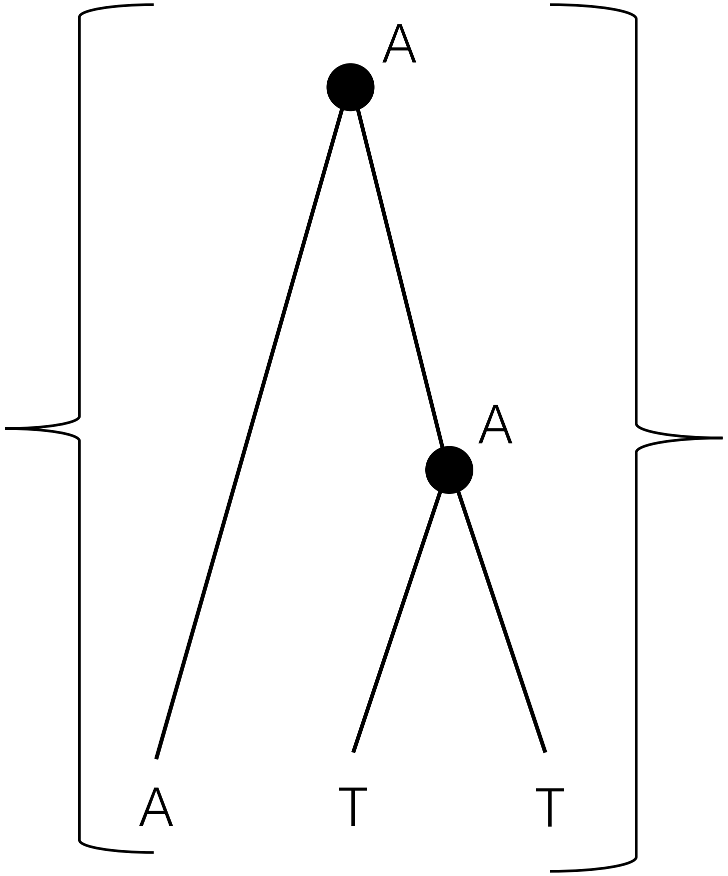
Pretend we know the ancestral states

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

Pretend we know the ancestral states

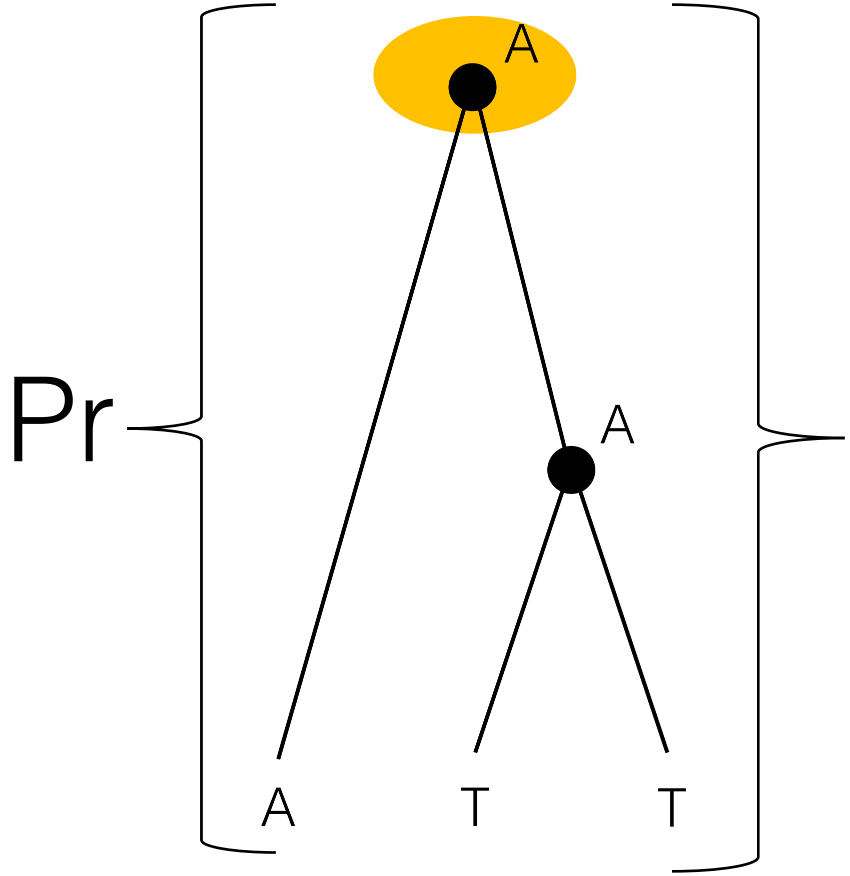
Pr



= Pr(A) x Pr(AA) x Pr(AA) x Pr(AT) x Pr(AT)

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

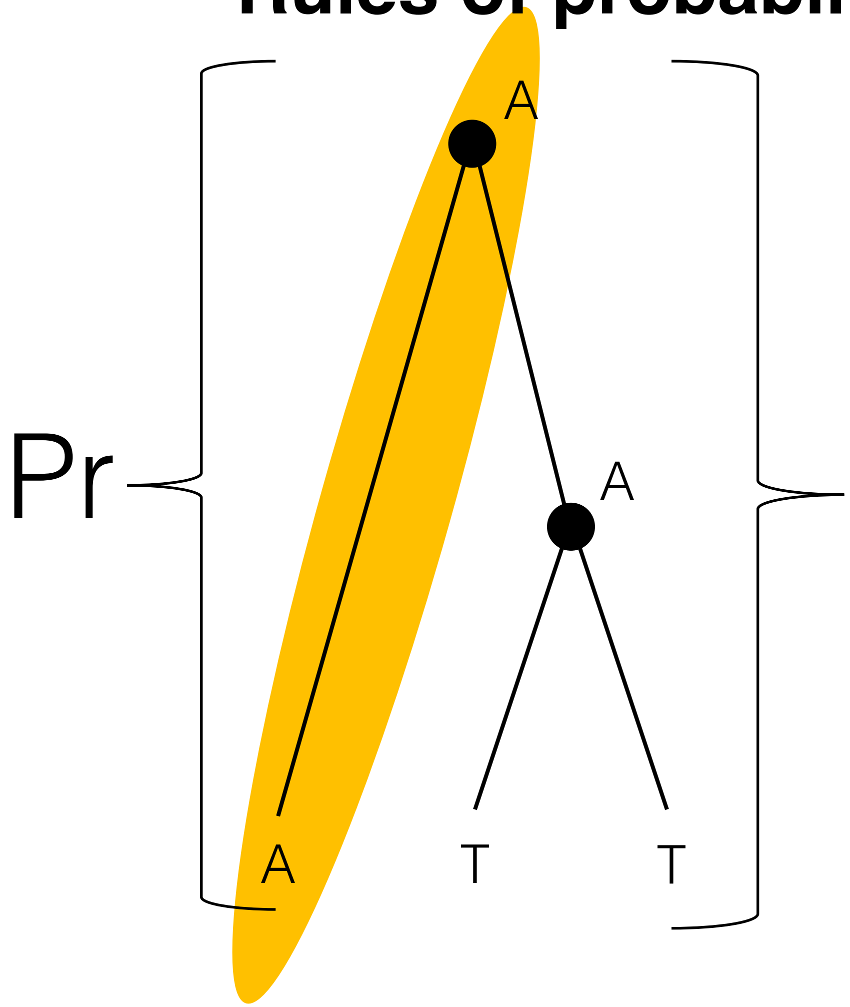


Pretend we know the ancestral states

= Pr(A) x Pr(AA) x Pr(AA) x Pr(AT) x Pr(AT)

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)



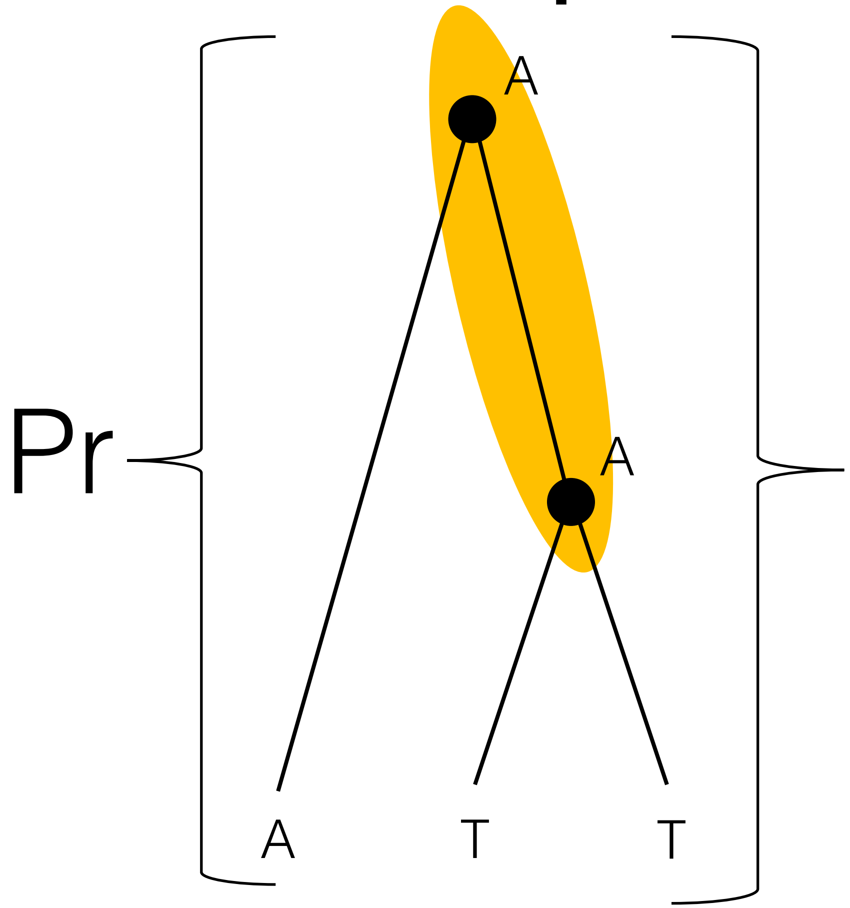
Pretend we know the ancestral states

= Pr(A) x Pr(AA) x Pr(AA) x Pr(AT) x Pr(AT)

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

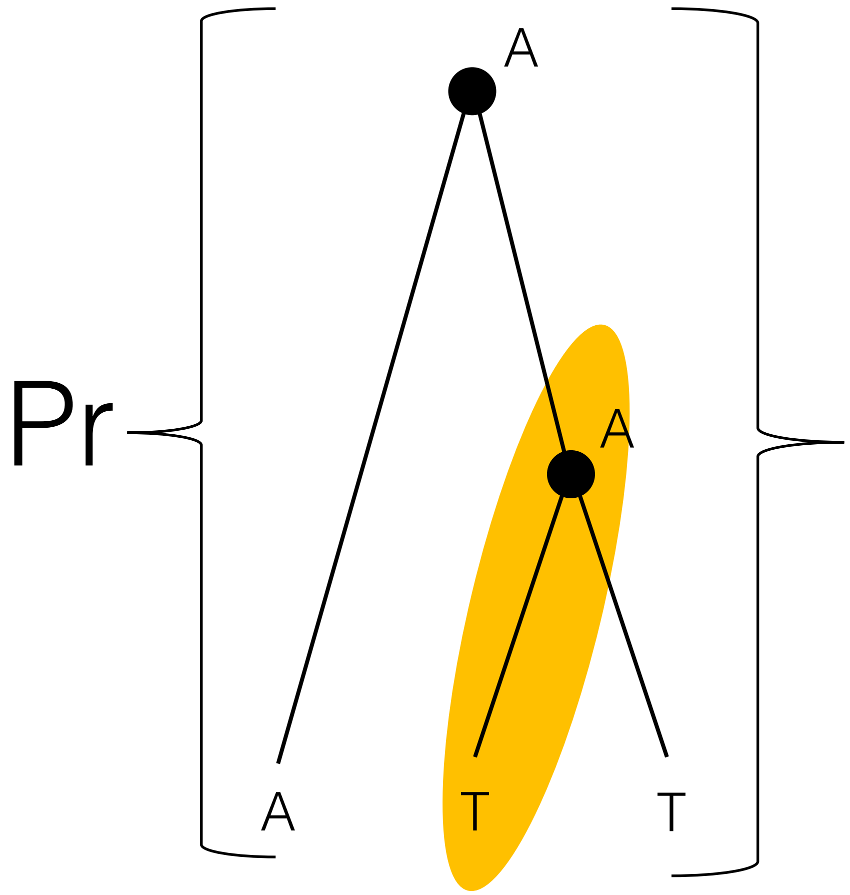
Pretend we know the ancestral states



= Pr(A) x Pr(AA) x Pr(AA) x Pr(AT) x Pr(AT)

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

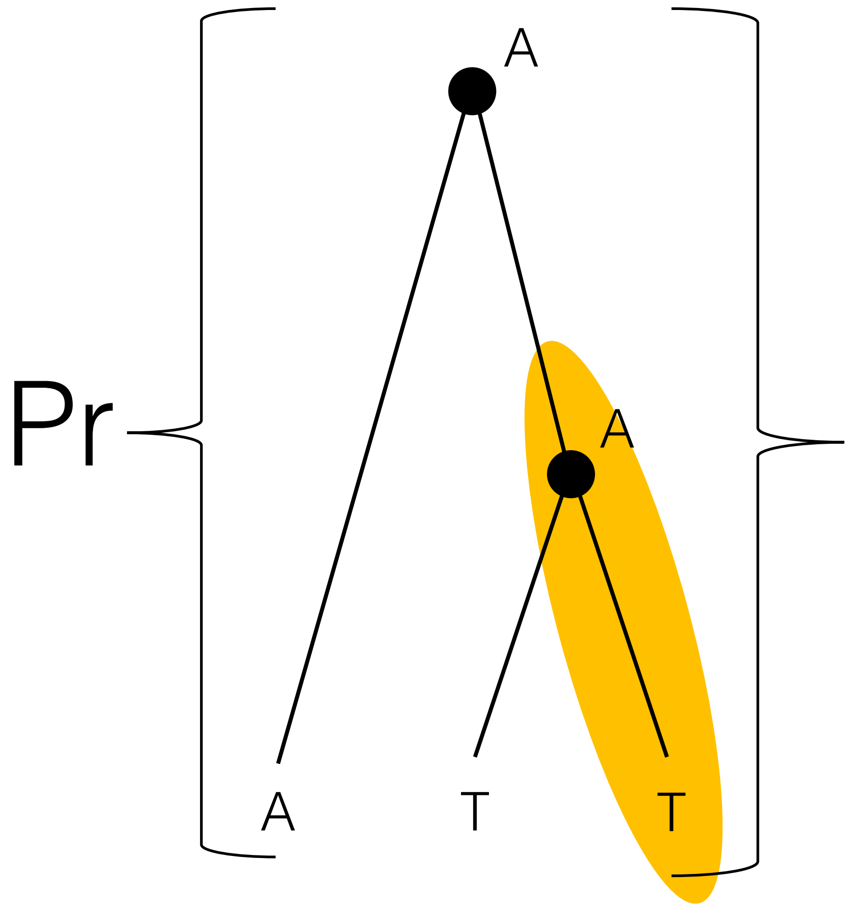


Pretend we know the ancestral states

= Pr(A) x Pr(AA) x Pr(AA) x Pr(AT) x Pr(AT)

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)



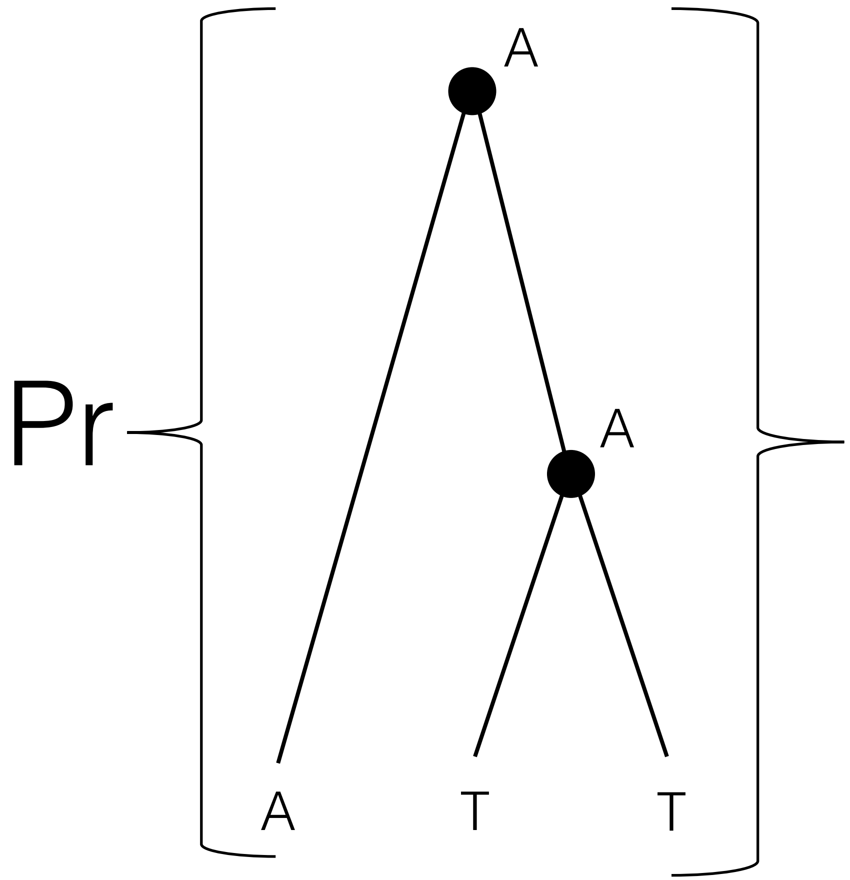
Pretend we know the ancestral states

= Pr(A) x Pr(AA) x Pr(AA) x Pr(AT) x Pr(AT)

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

Pretend we know the ancestral states



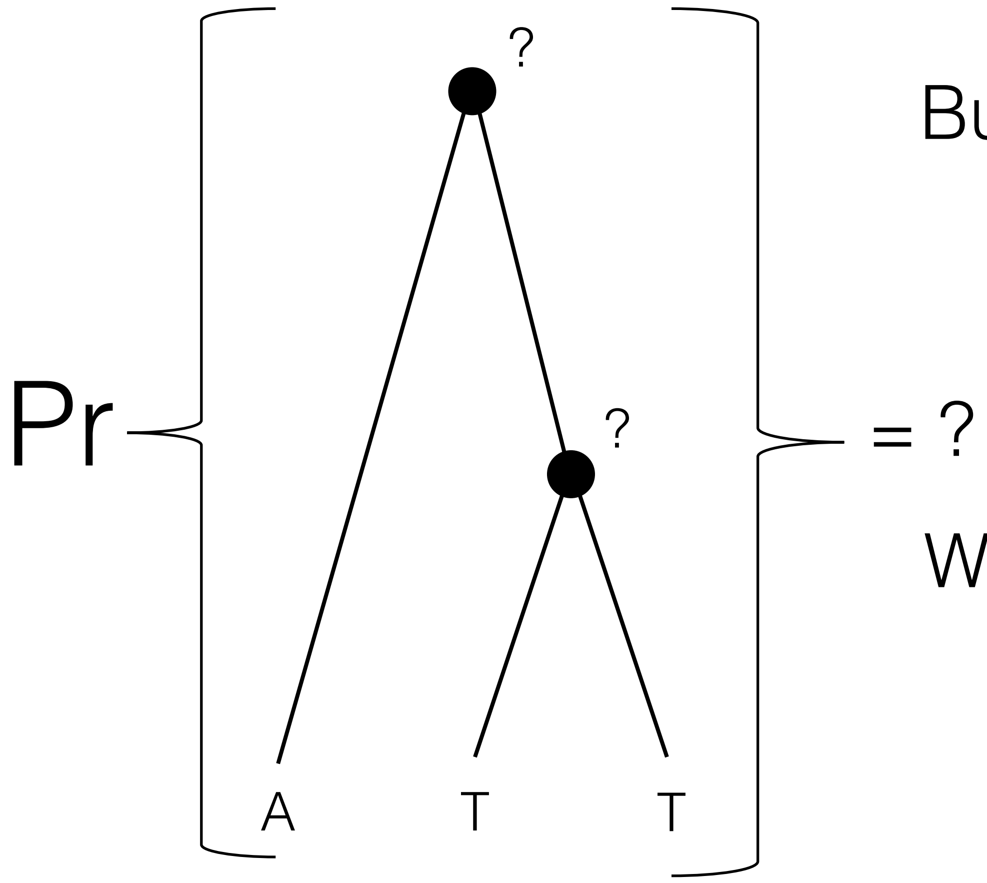
= Pr(A) x Pr(AA) x Pr(AA) x Pr(AT) x Pr(AT)

equilibrium frequency

transition probabilities

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

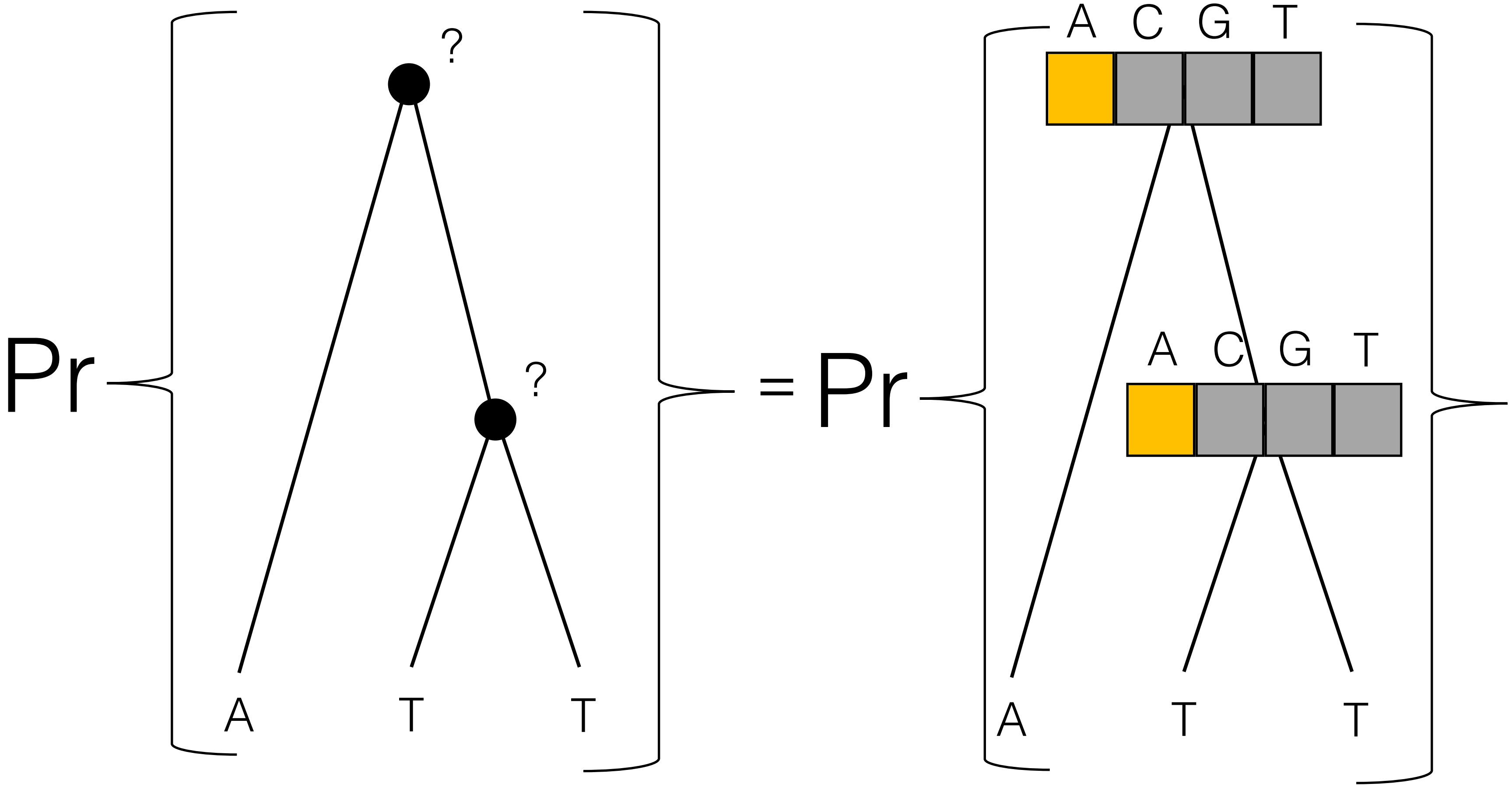


But in nature we do not know the ancestral states!

We get around this problem by integrating over the possible states at each node

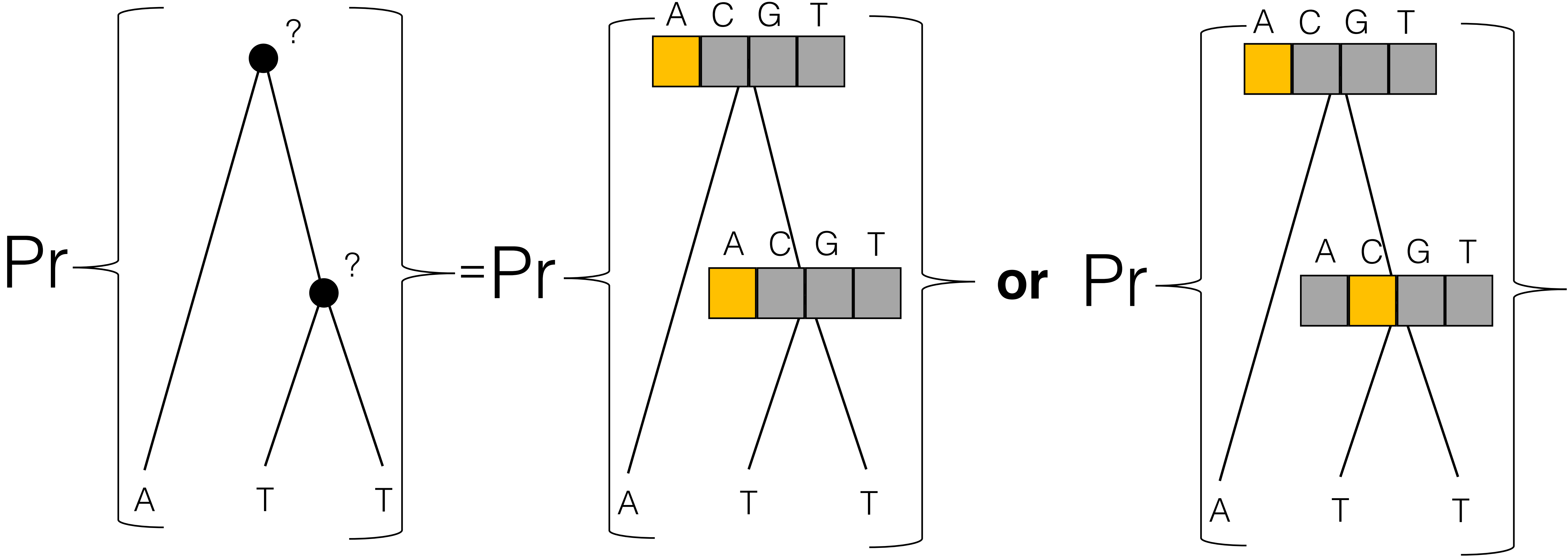
Probability and Likelihood

Rules of probability applied to phylogeny (DNA)



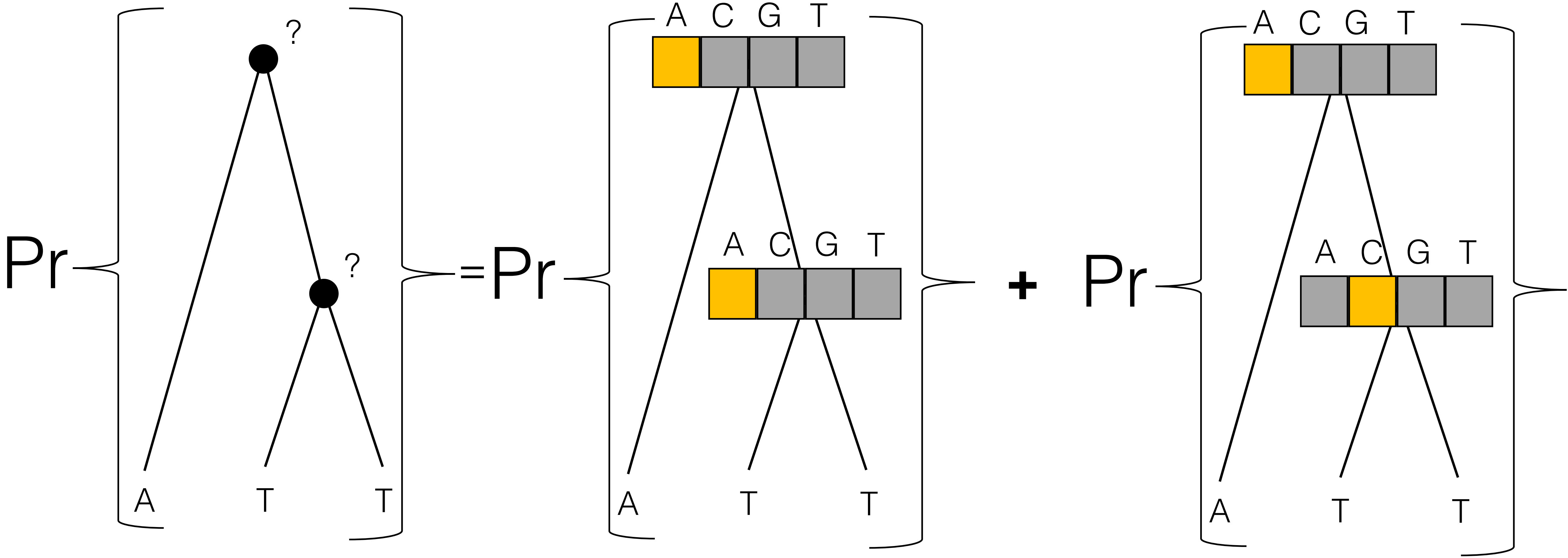
Probability and Likelihood

Rules of probability applied to phylogeny (DNA)



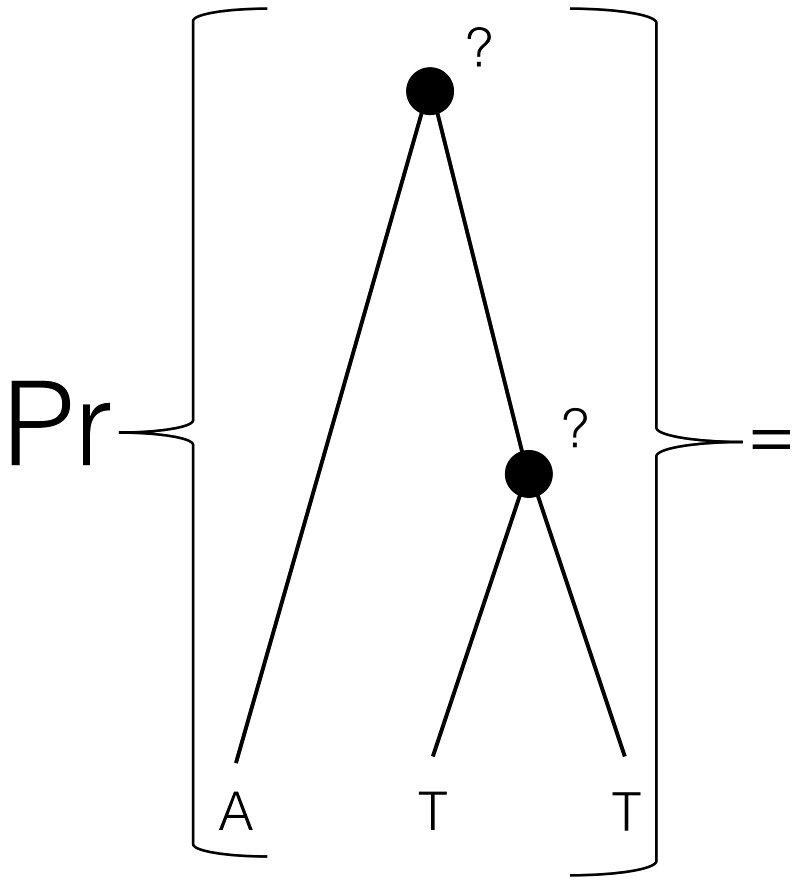
Probability and Likelihood

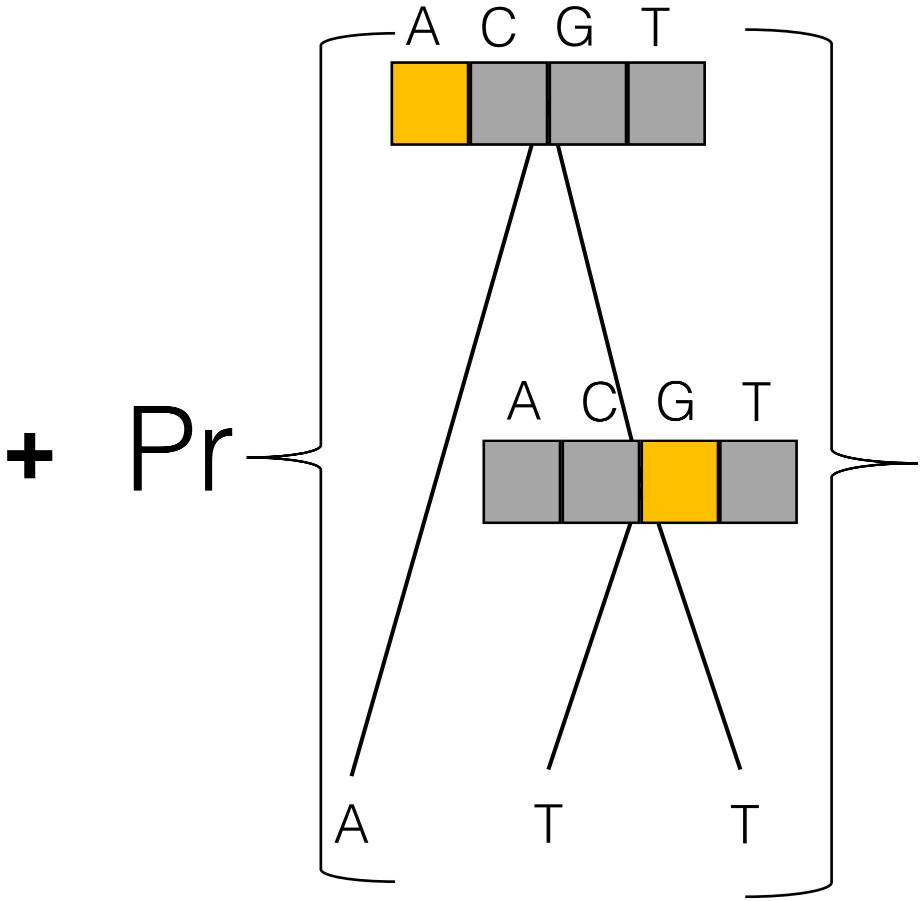
Rules of probability applied to phylogeny (DNA)



Probability and Likelihood

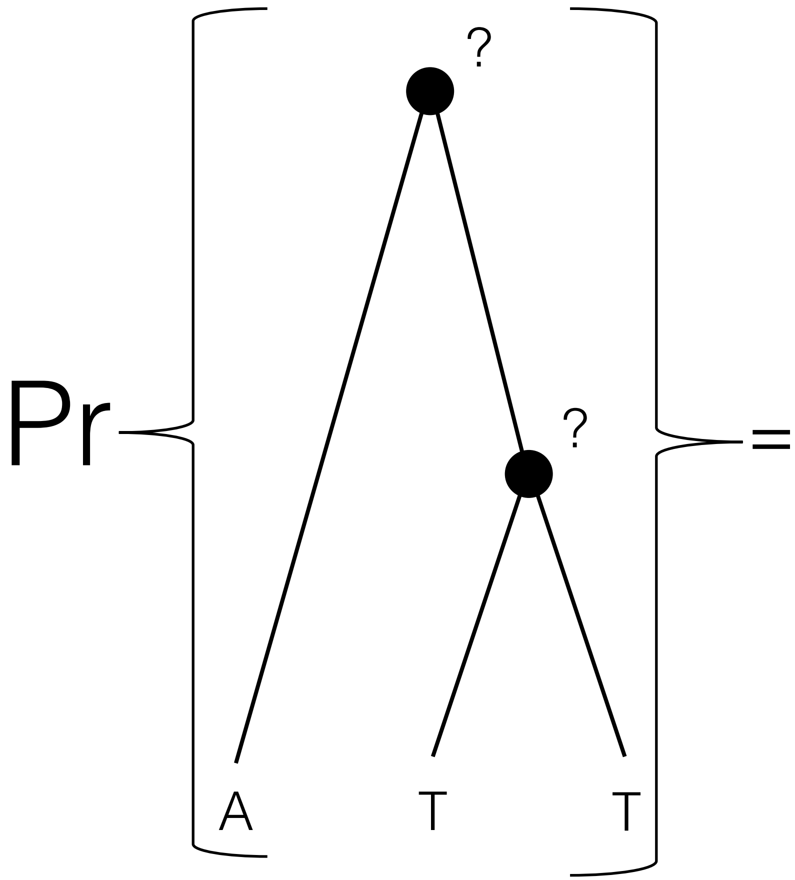
Rules of probability applied to phylogeny (DNA)

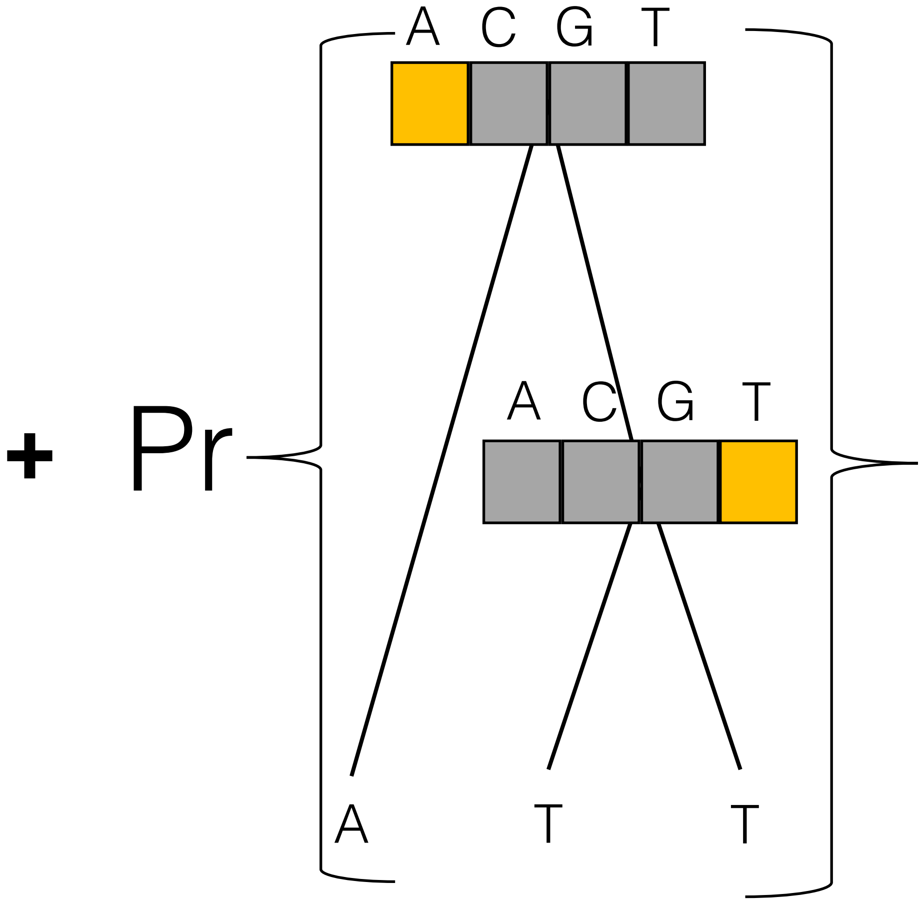




Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

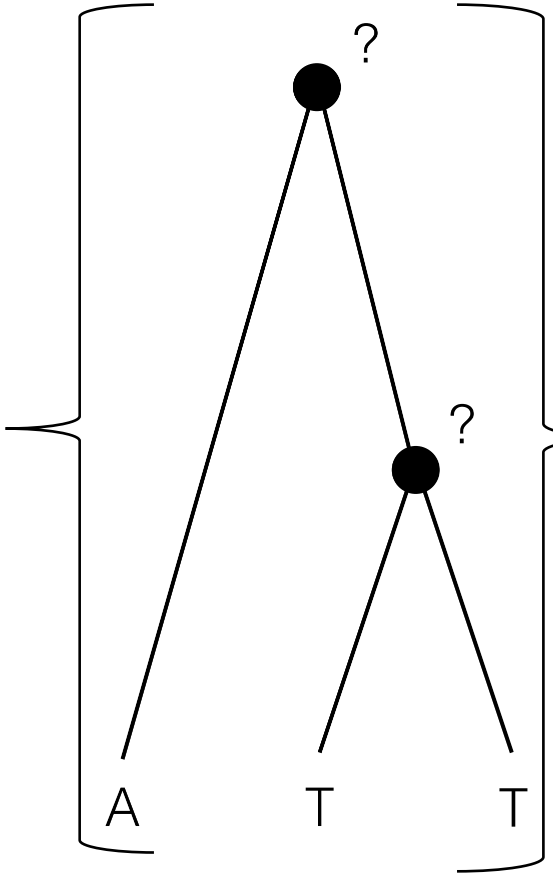




Probability and Likelihood

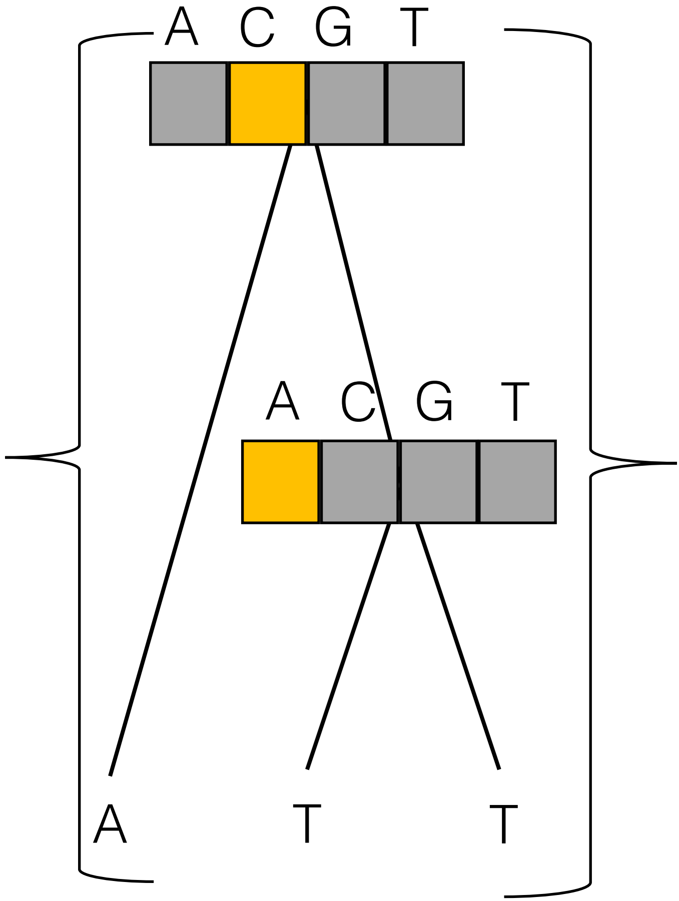
Rules of probability applied to phylogeny (DNA)

Pr



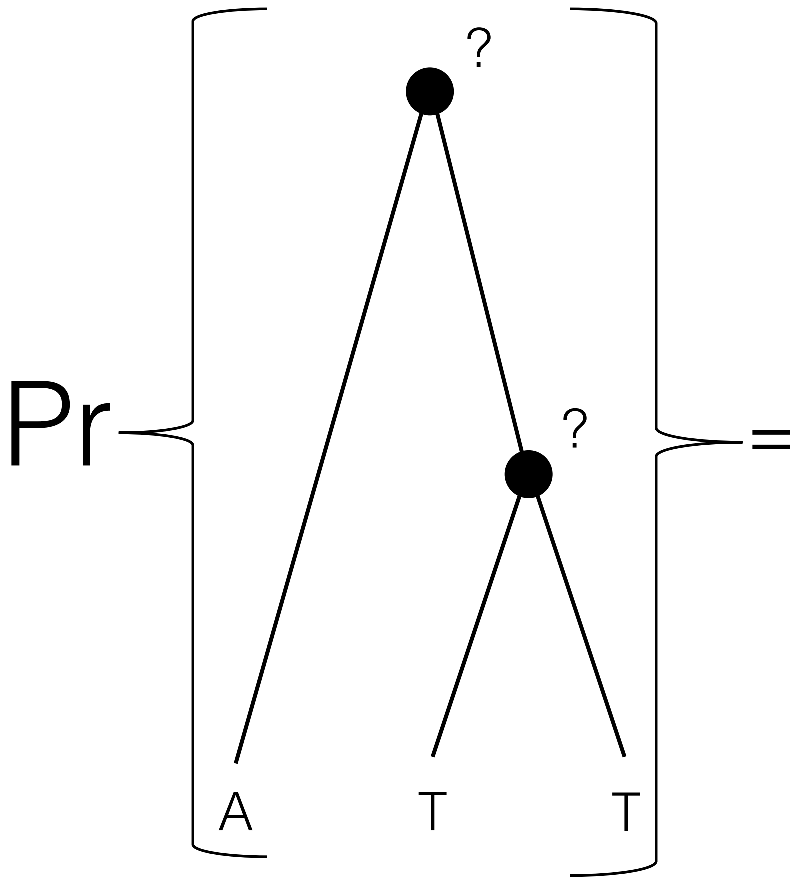
=

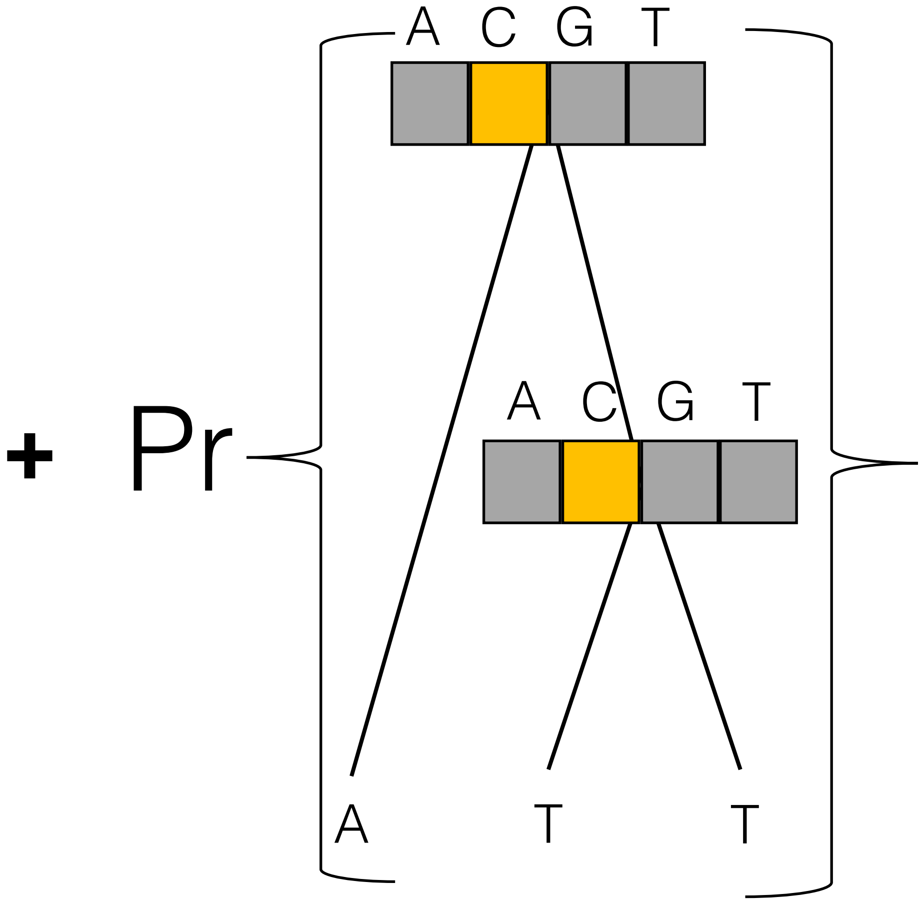
+ Pr



Probability and Likelihood

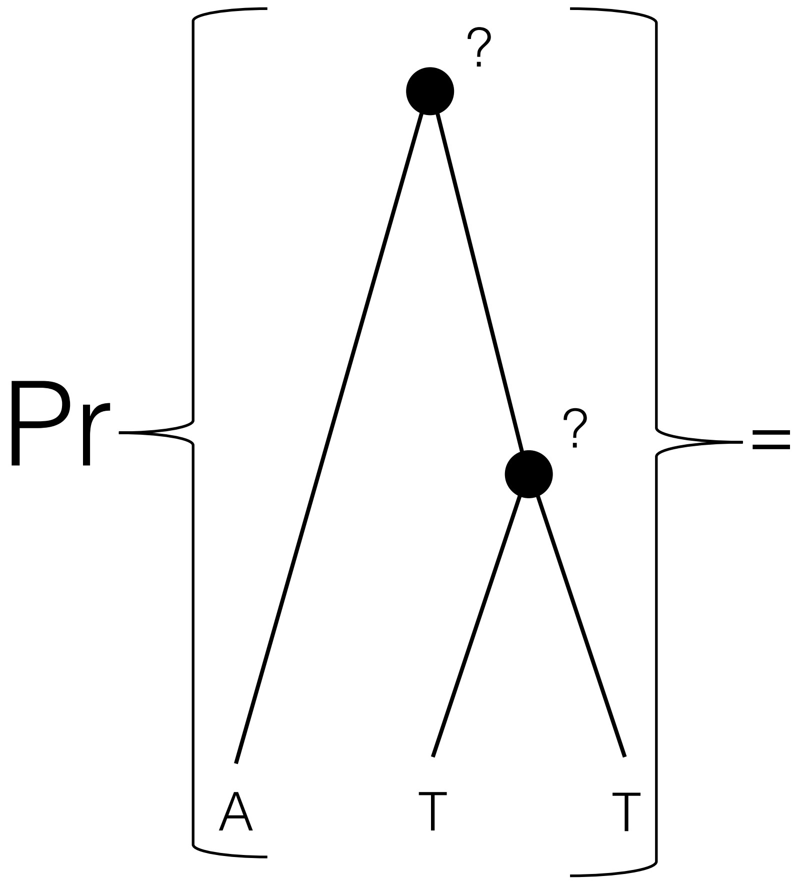
Rules of probability applied to phylogeny (DNA)

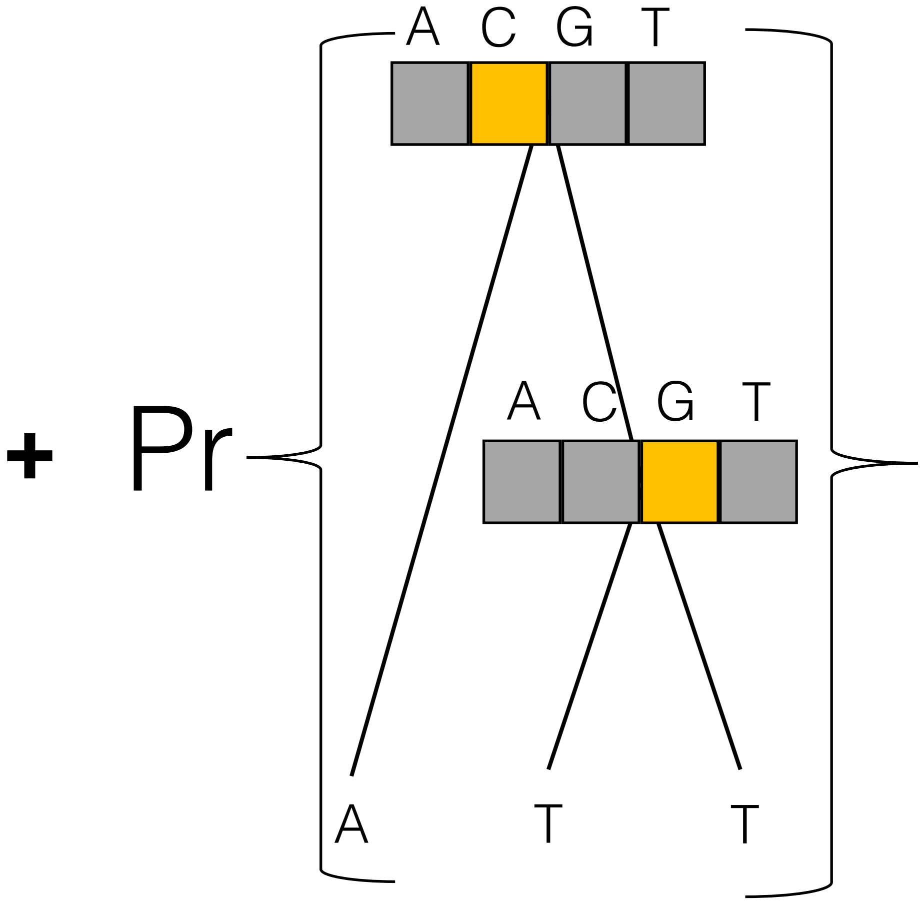




Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

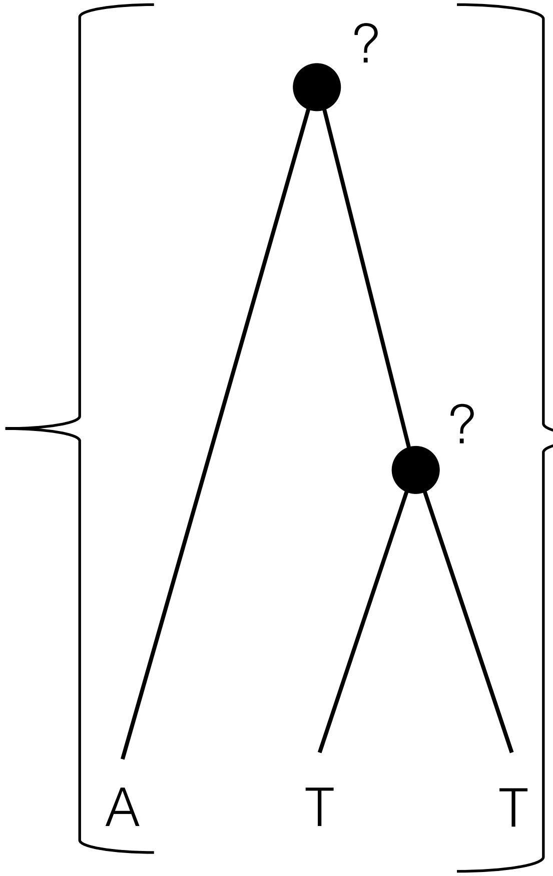




Probability and Likelihood

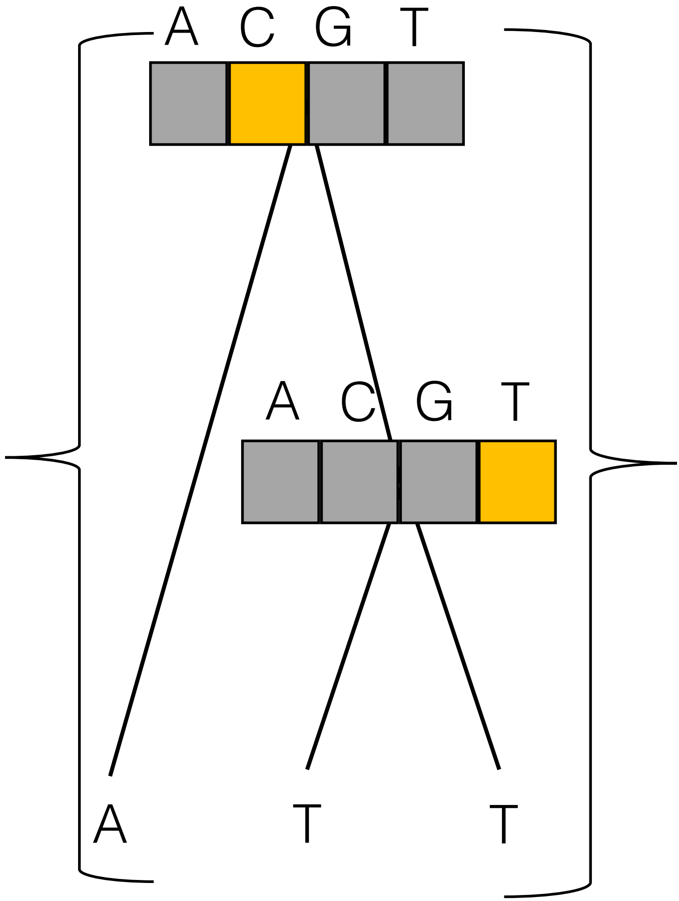
Rules of probability applied to phylogeny (DNA)

Pr



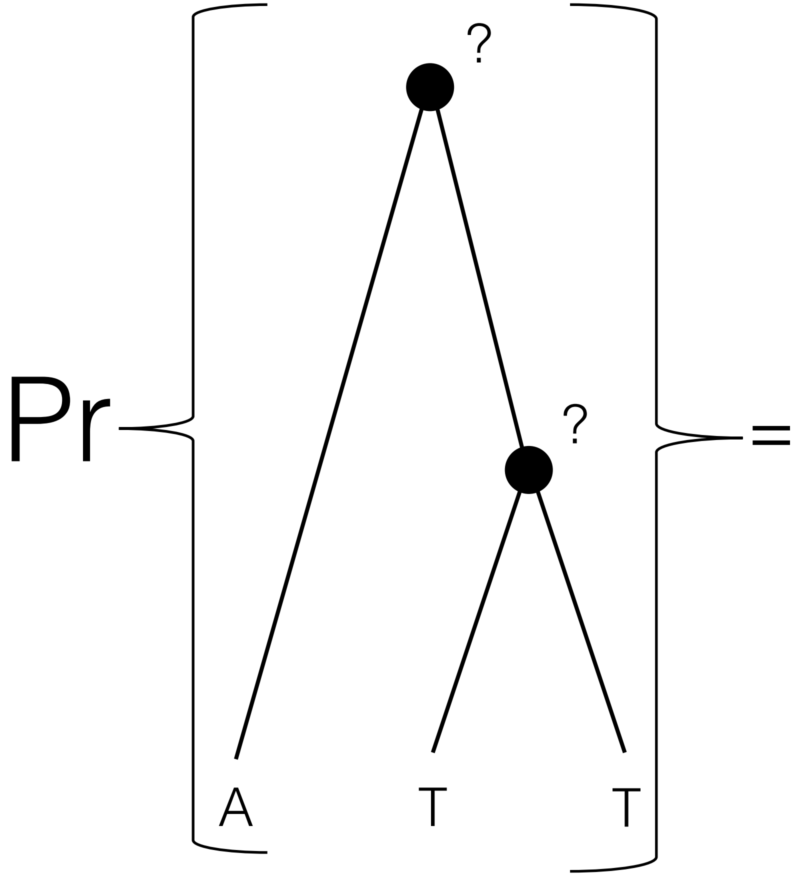
=

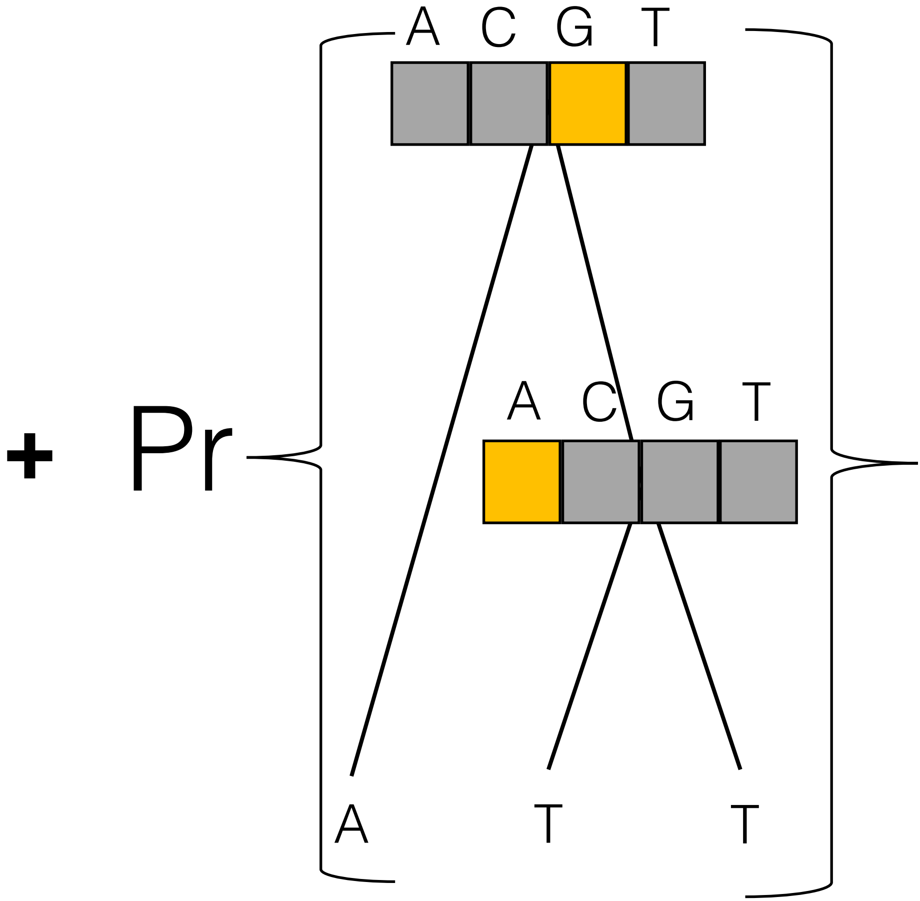
+ Pr



Probability and Likelihood

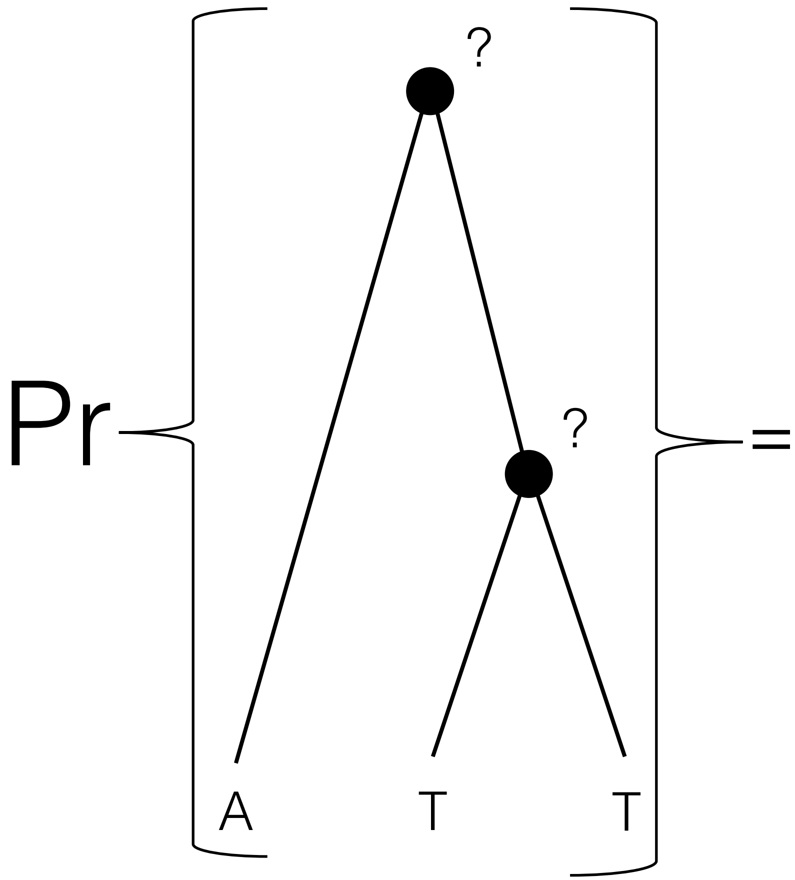
Rules of probability applied to phylogeny (DNA)

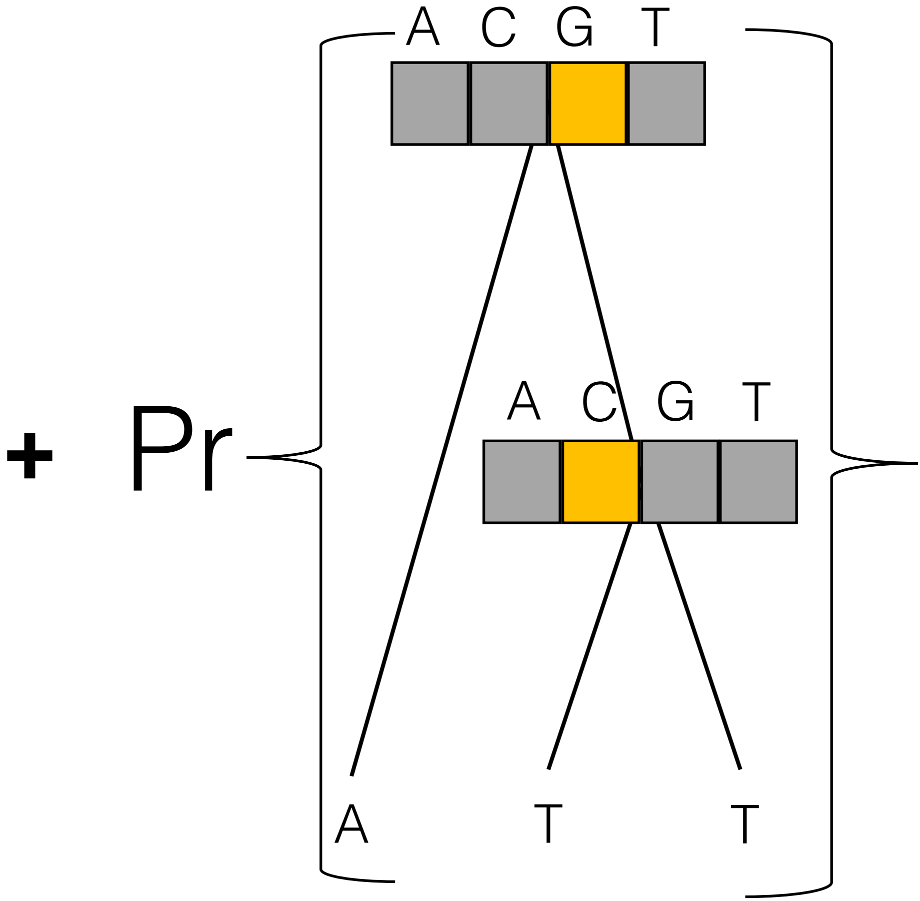




Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

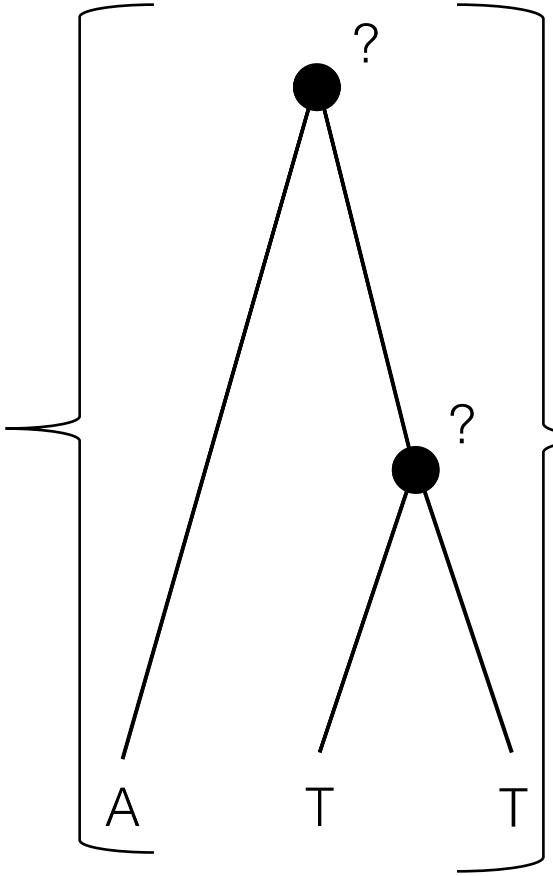




Probability and Likelihood

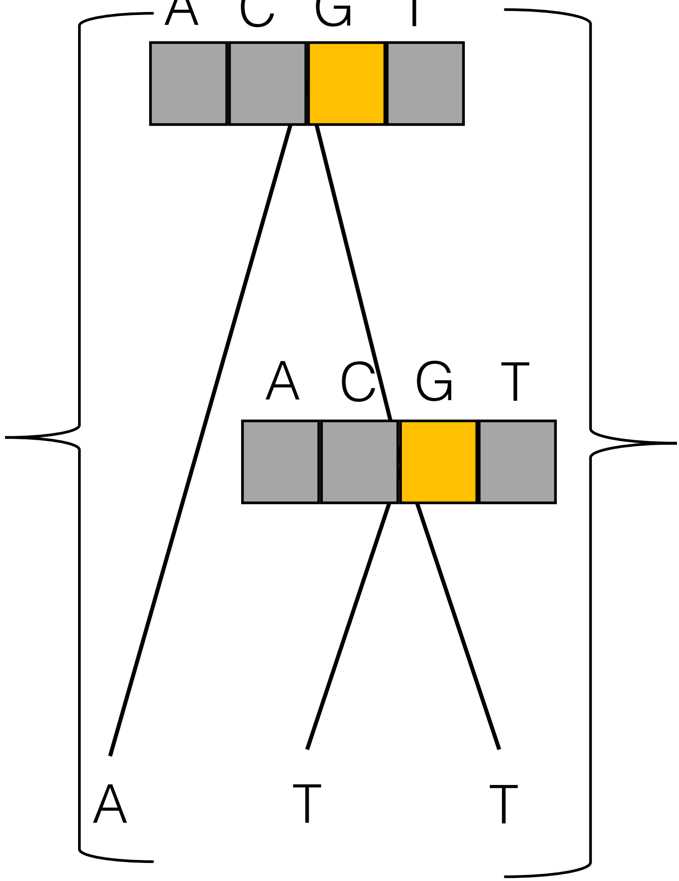
Rules of probability applied to phylogeny (DNA)

Pr



=

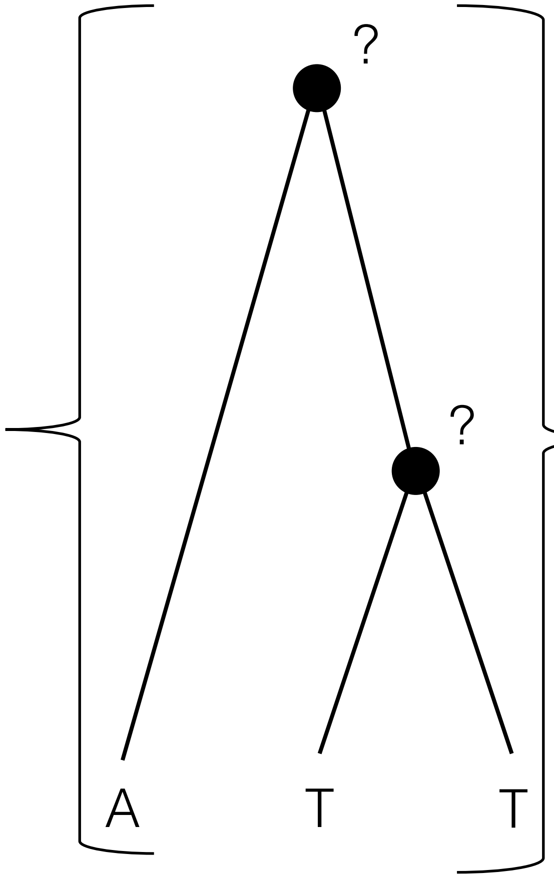
+ Pr



Probability and Likelihood

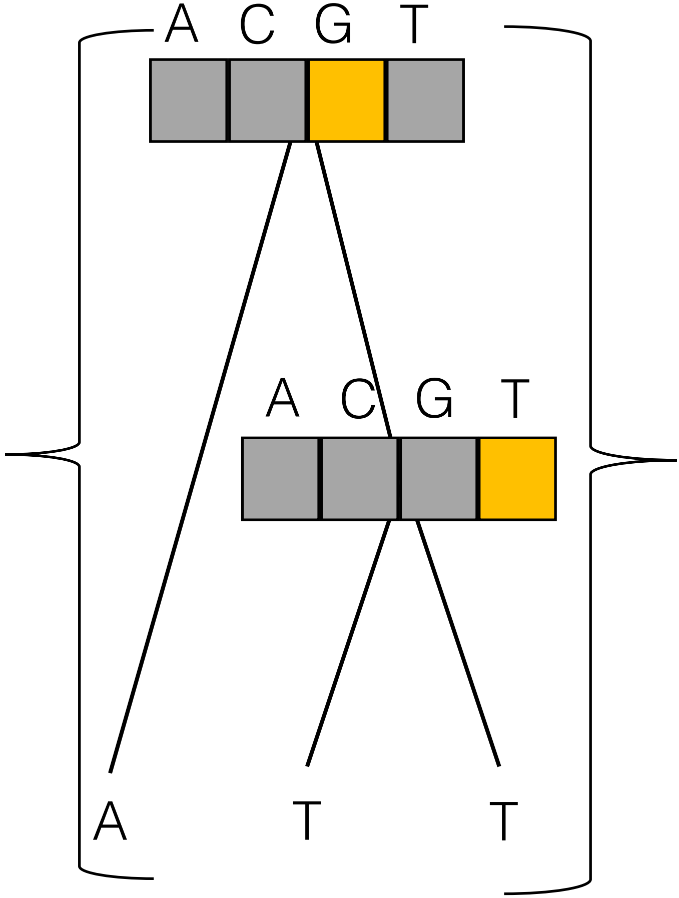
Rules of probability applied to phylogeny (DNA)

Pr



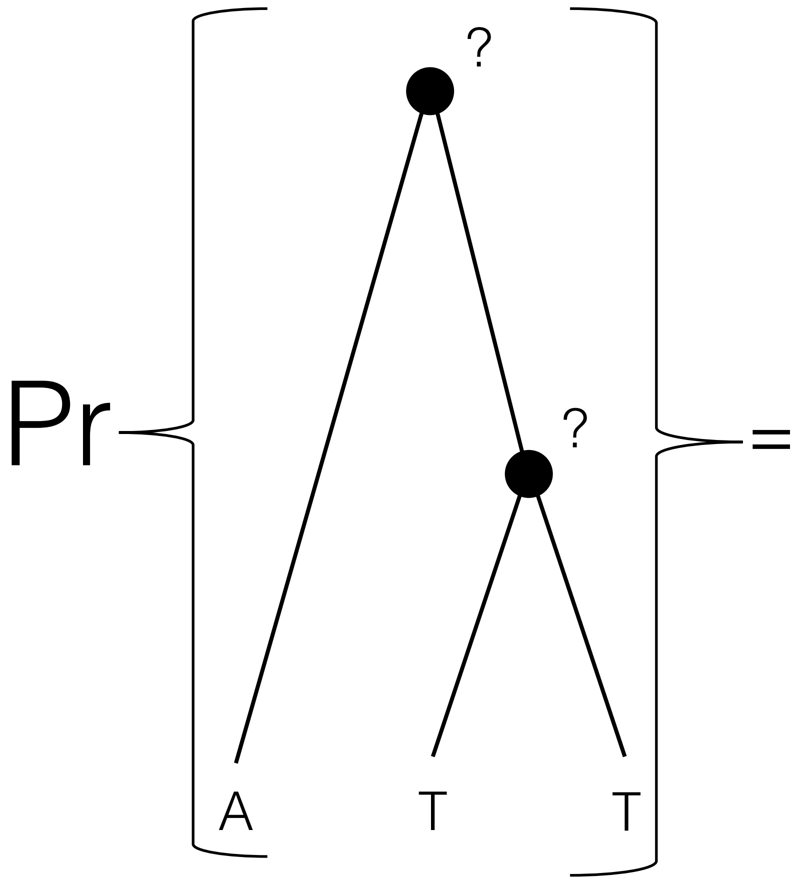
=

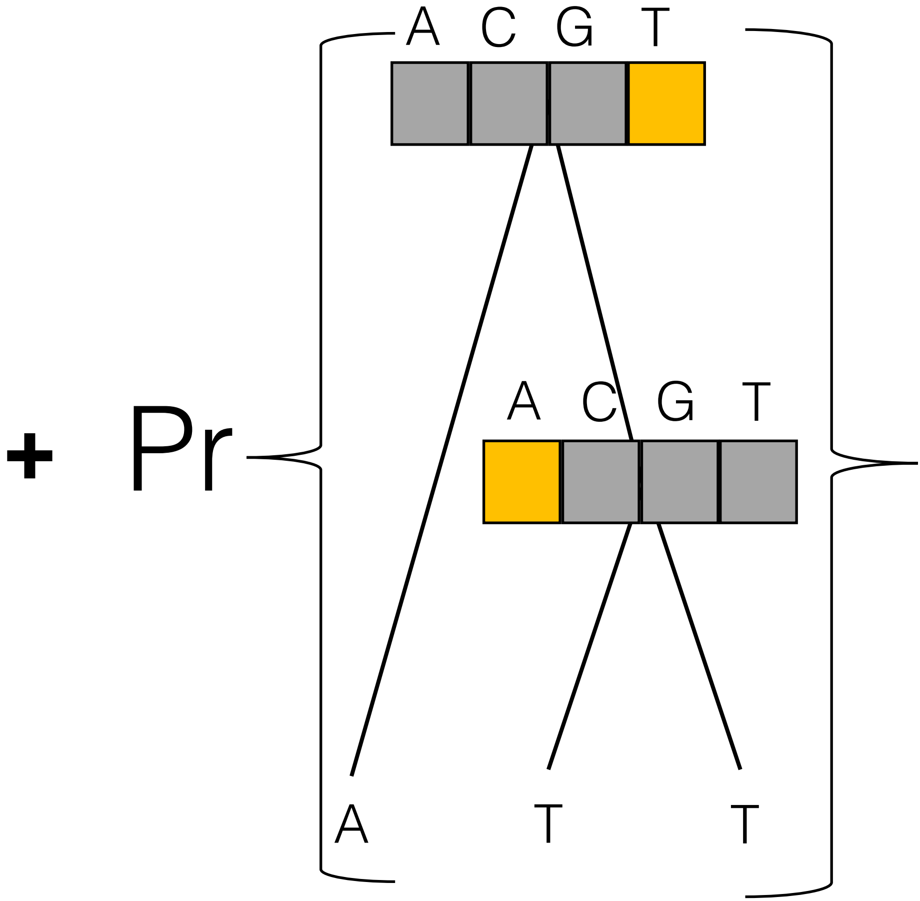
+ Pr



Probability and Likelihood

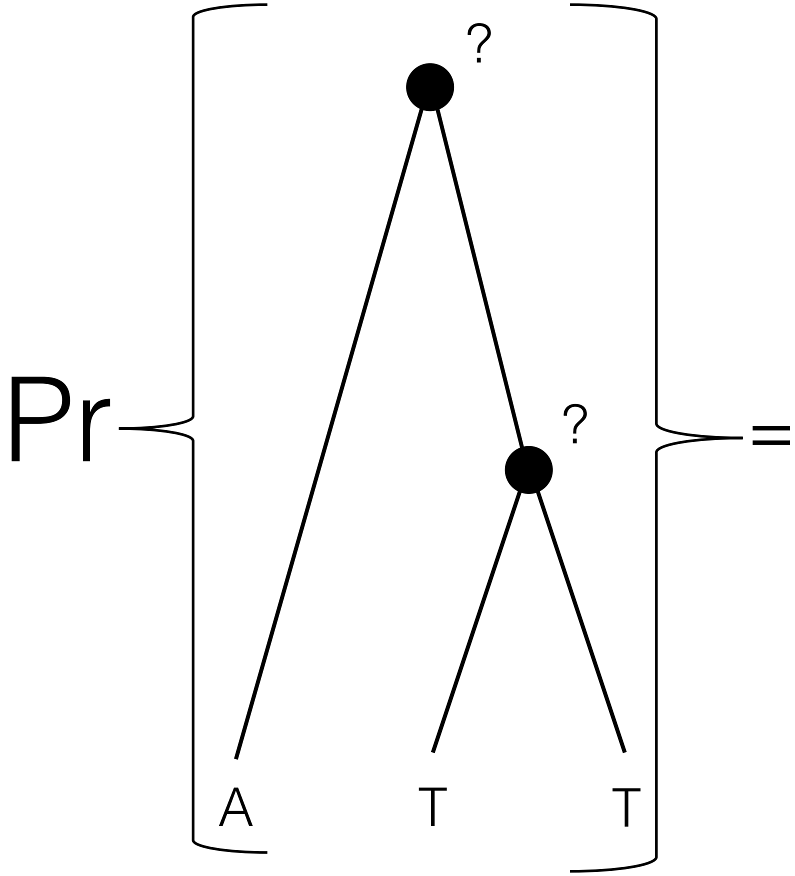
Rules of probability applied to phylogeny (DNA)

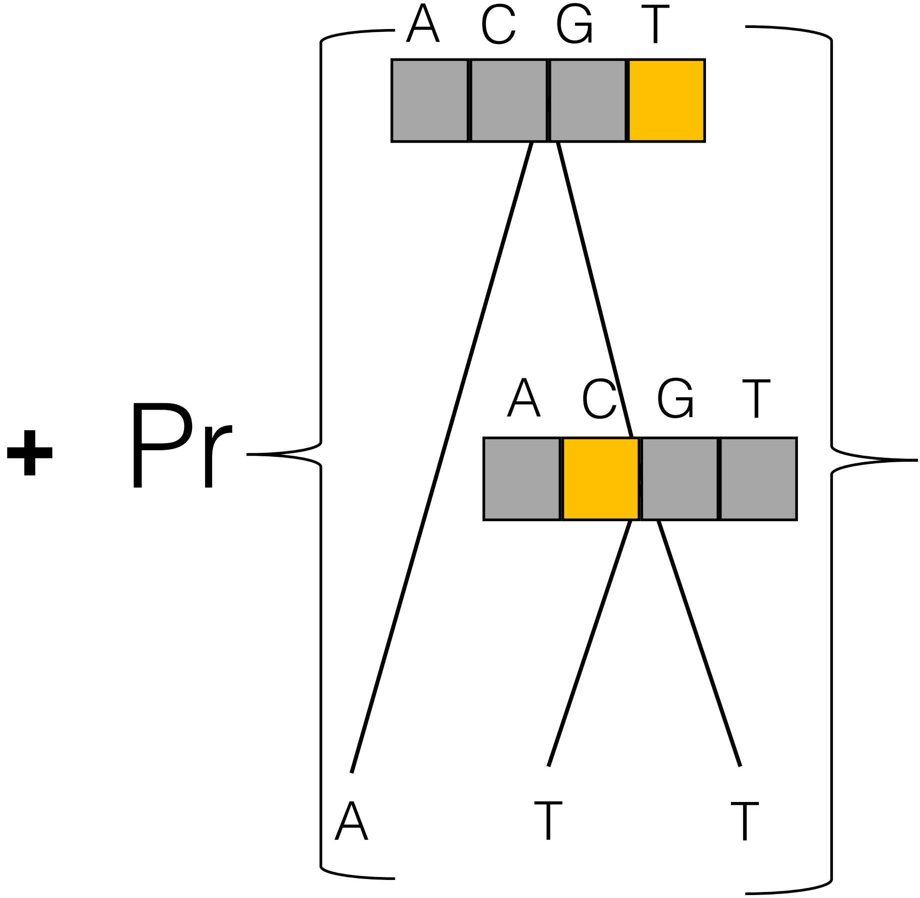




Probability and Likelihood

Rules of probability applied to phylogeny (DNA)

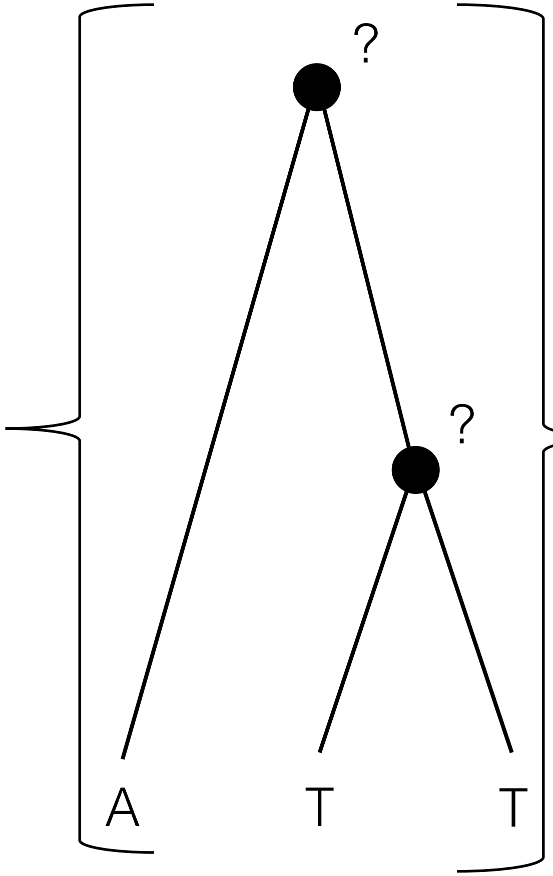




Probability and Likelihood

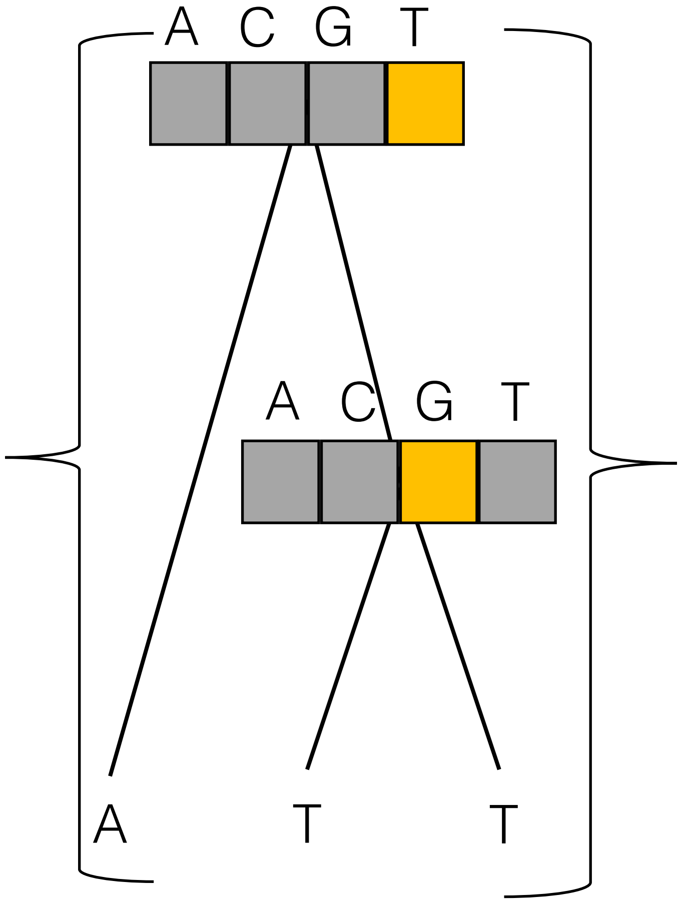
Rules of probability applied to phylogeny (DNA)

Pr



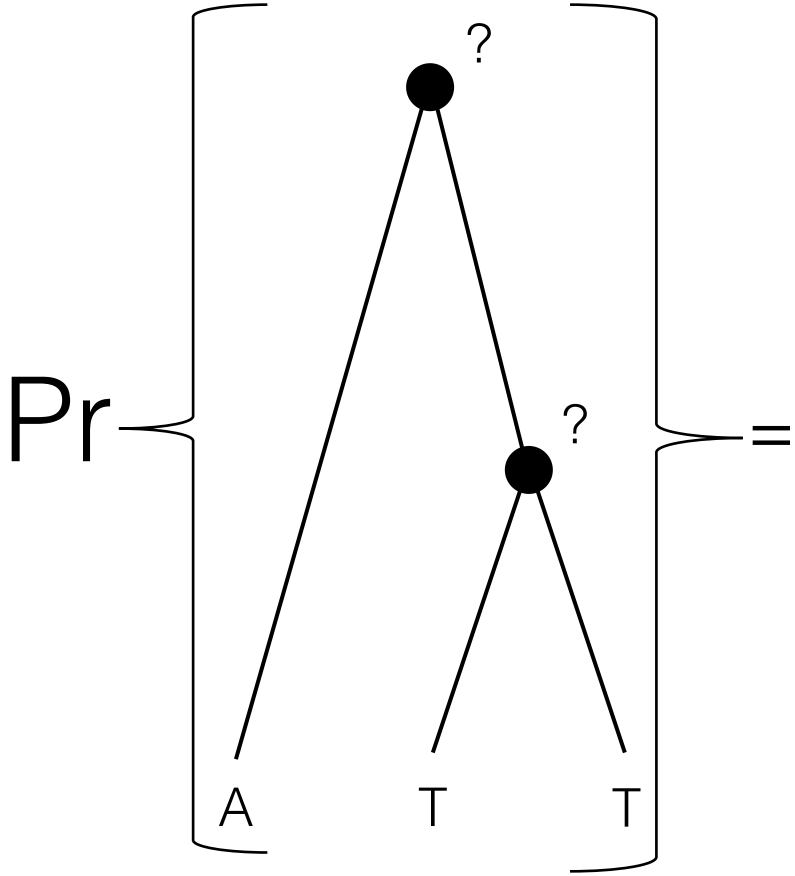
=

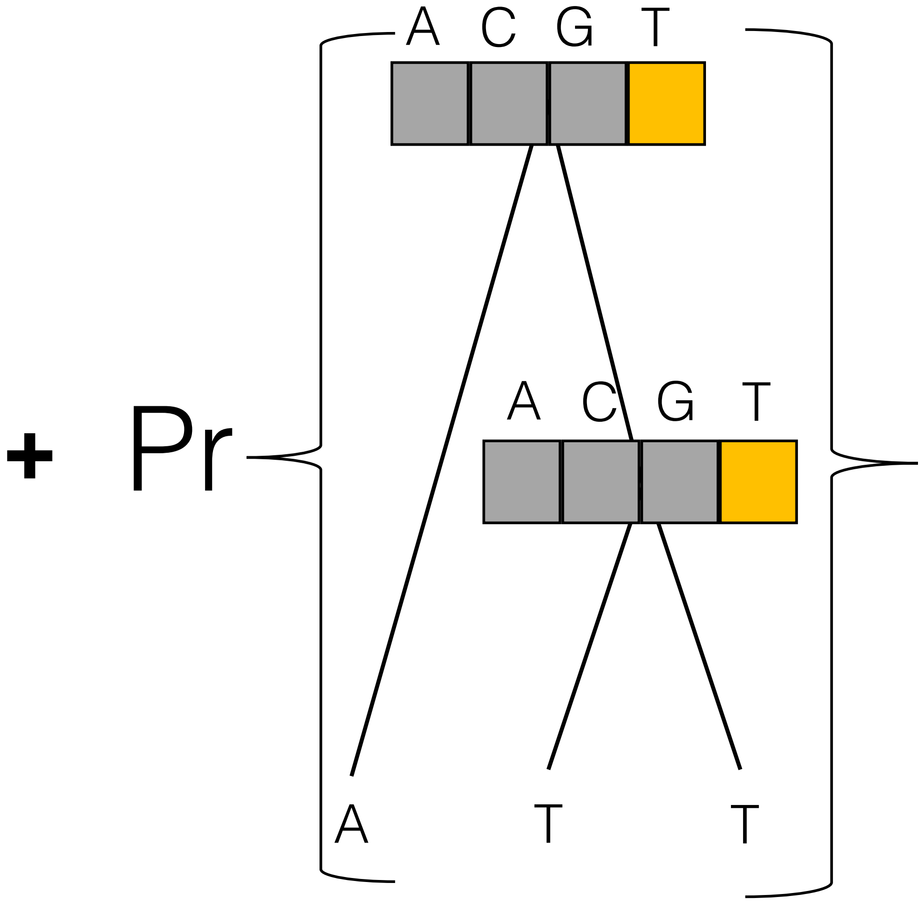
+ Pr

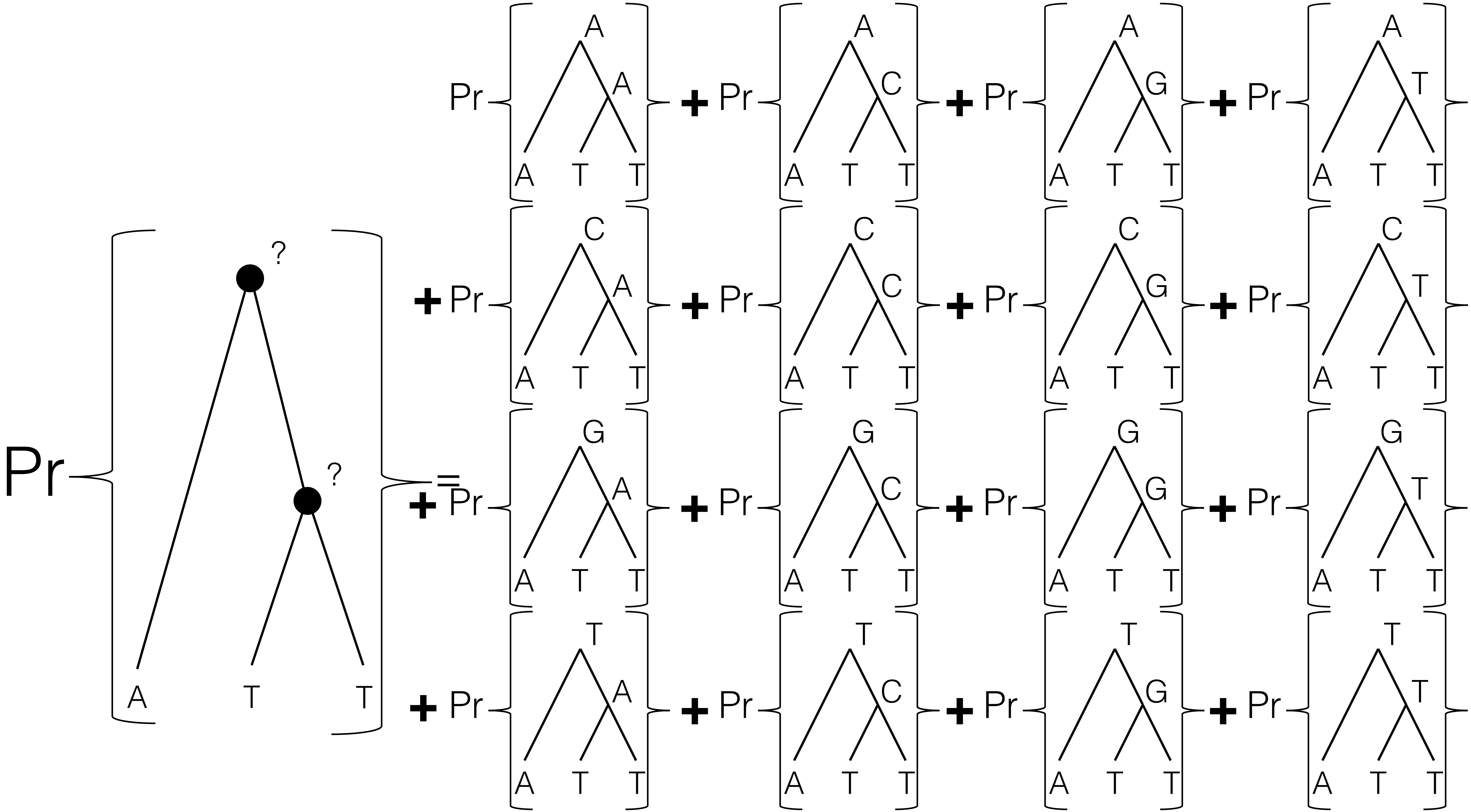


Probability and Likelihood

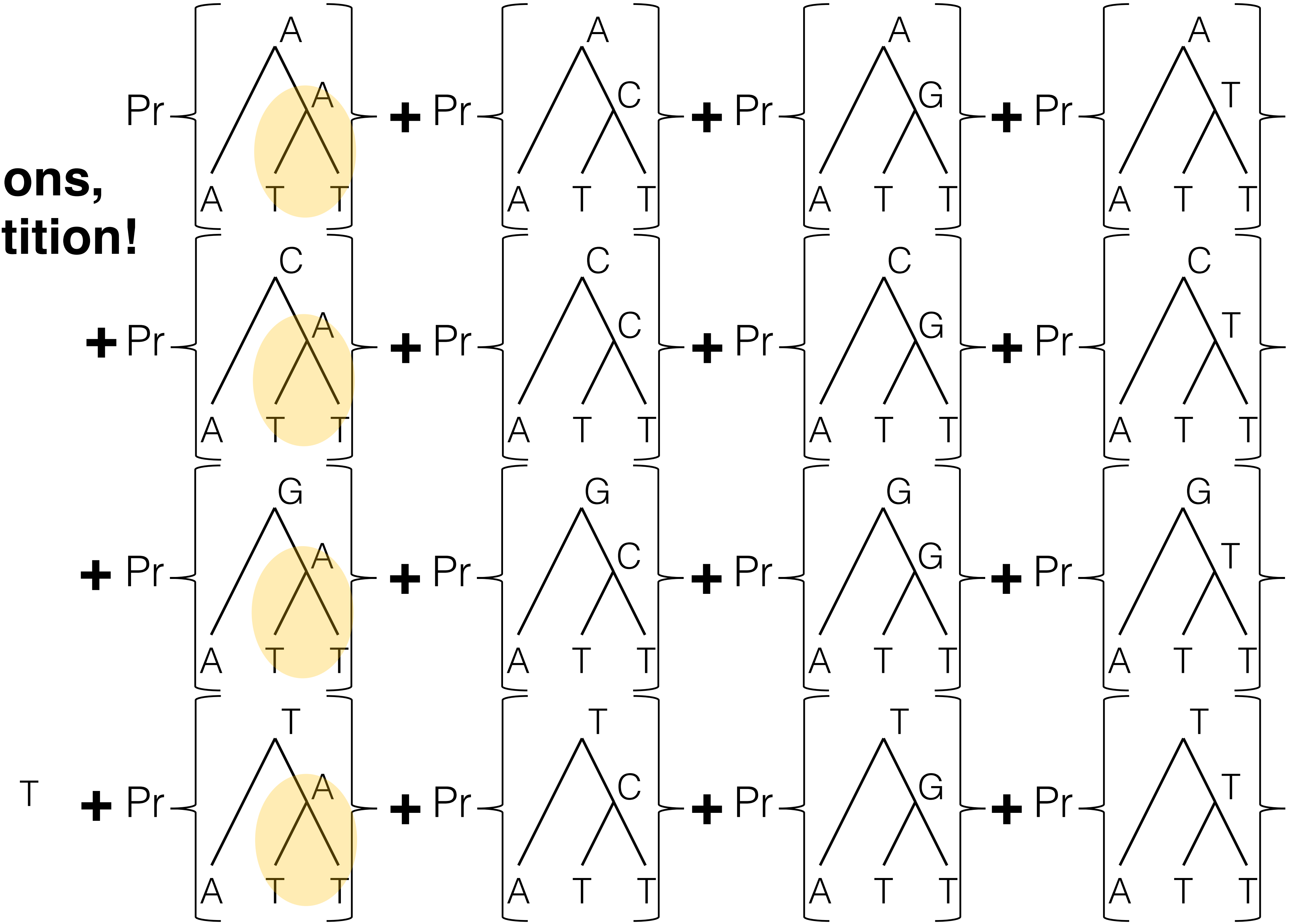
Rules of probability applied to phylogeny (DNA)






Pr
?
?
A
T
T
=
Pr
A
A
A T T
+
Pr
A
C
A T T
+
Pr
A
G
A T T
+
Pr
A
T
A T T
+
Pr
C
A
A T T
+
Pr
C
C
A T T
+
Pr
C
G
A T T
+
Pr
C
T
A T T
+
Pr
G
A
A T T
+
Pr
G
C
A T T
+
Pr
G
G
A T T
+
Pr
G
T
A T T
+
Pr
T
A
A T T
+
Pr
T
C
A T T
+
Pr
T
G
A T T
+
Pr
T
T
A T T

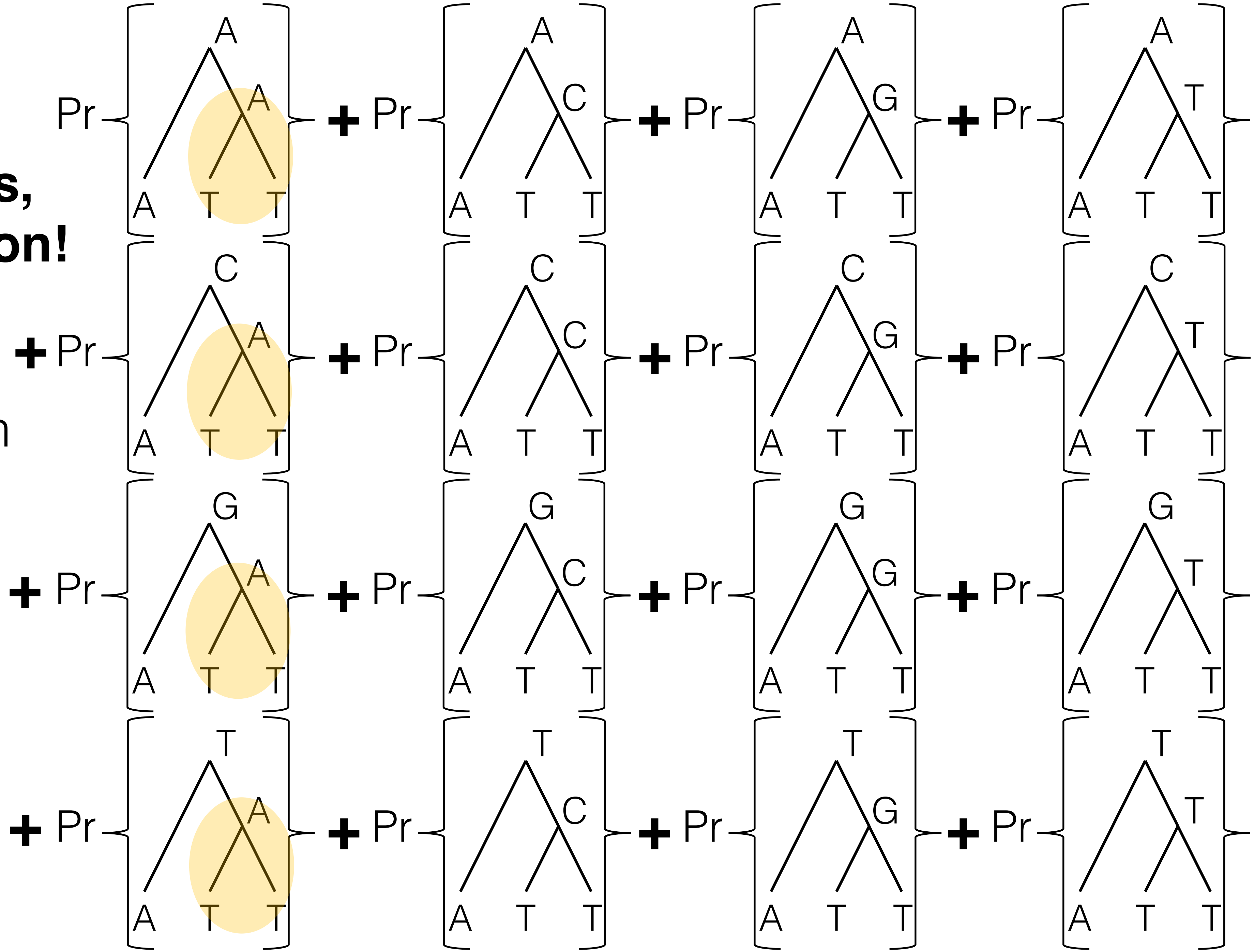
Many calculations, but a lot of repetition!



Many calculations, but a lot of repetition!

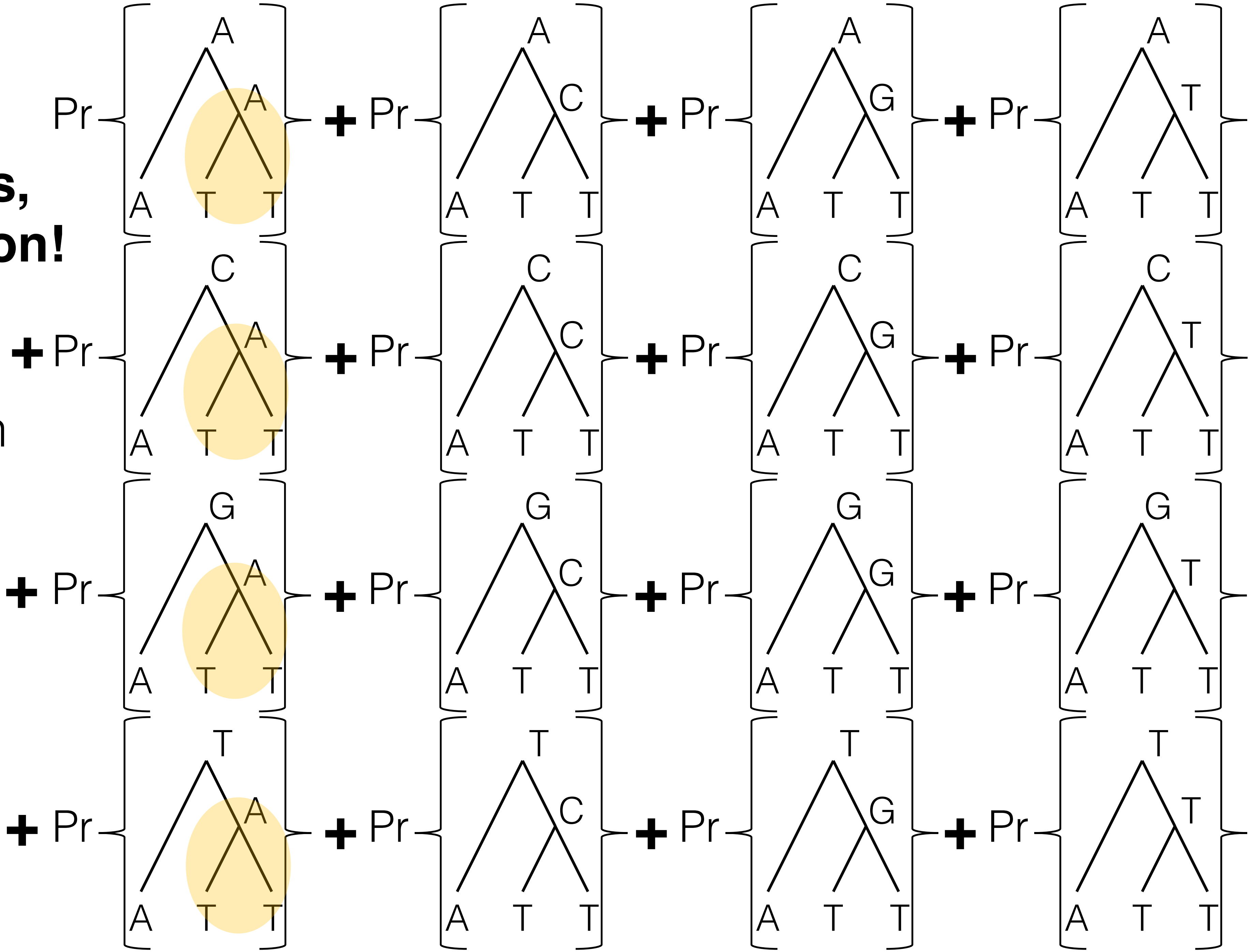
Efficient calculation of *likelihoods* achieved with the *pruning algorithm* (Felsenstein 1981)

T


Pr A A A T T + Pr A C A T T + Pr A G A T T + Pr A T A T T
+ Pr C A A T T + Pr C C A T T + Pr C G A T T + Pr C T A T T
+ Pr G A A T T + Pr G C A T T + Pr G G A T T + Pr G T A T T
+ Pr T A A T T + Pr T C A T T + Pr T G A T T + Pr T T A T T

Many calculations, but a lot of repetition!

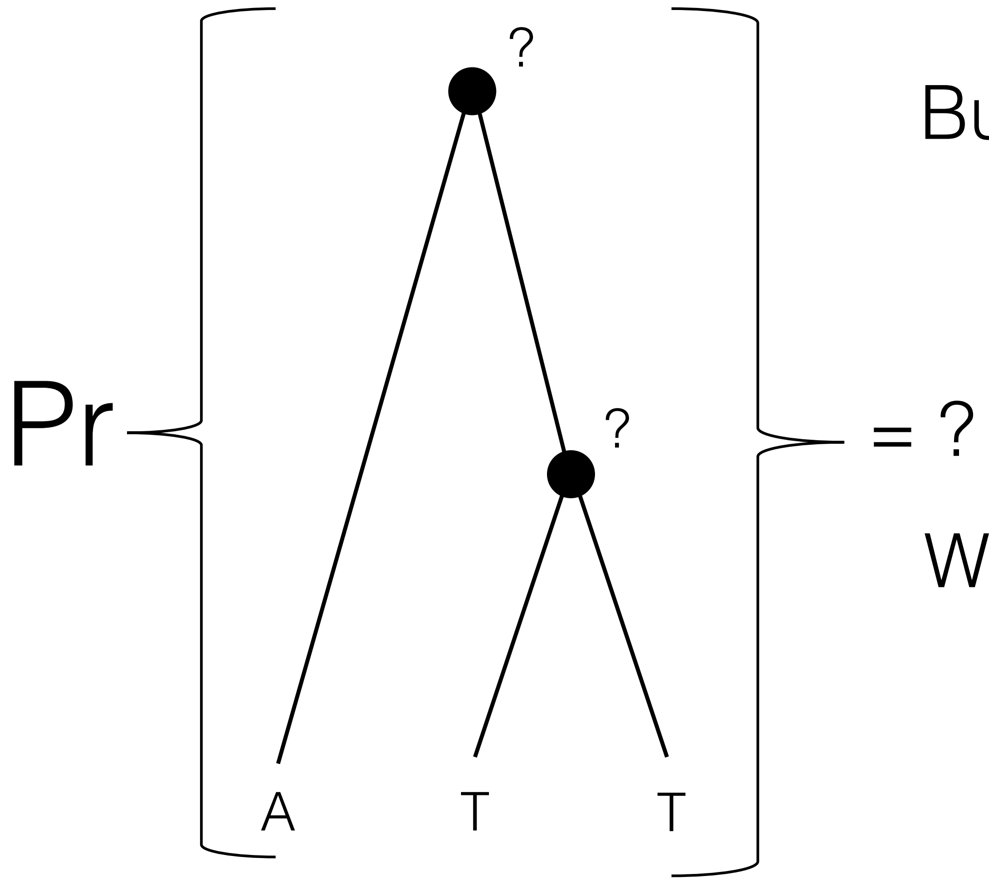
Efficient calculation of *likelihoods* achieved with the *pruning algorithm* (Felsenstein 1981)


Pr A A A T T + Pr A C A T T + Pr A G A T T + Pr A T A T T
+ Pr C A A T T + Pr C C A T T + Pr C G A T T + Pr C T A T T
+ Pr G A A T T + Pr G C A T T + Pr G G A T T + Pr G T A T T
+ Pr T A A T T + Pr T C A T T + Pr T G A T T + Pr T T A T T

T

Probability and Likelihood

Rules of probability applied to phylogeny (DNA)



But in nature we do not know the ancestral states!

We get around this problem by integrating over the possible states at each node

Learning Goals

Explain terminology

Primer on probability and likelihood

Models of molecular evolution

How to select a model

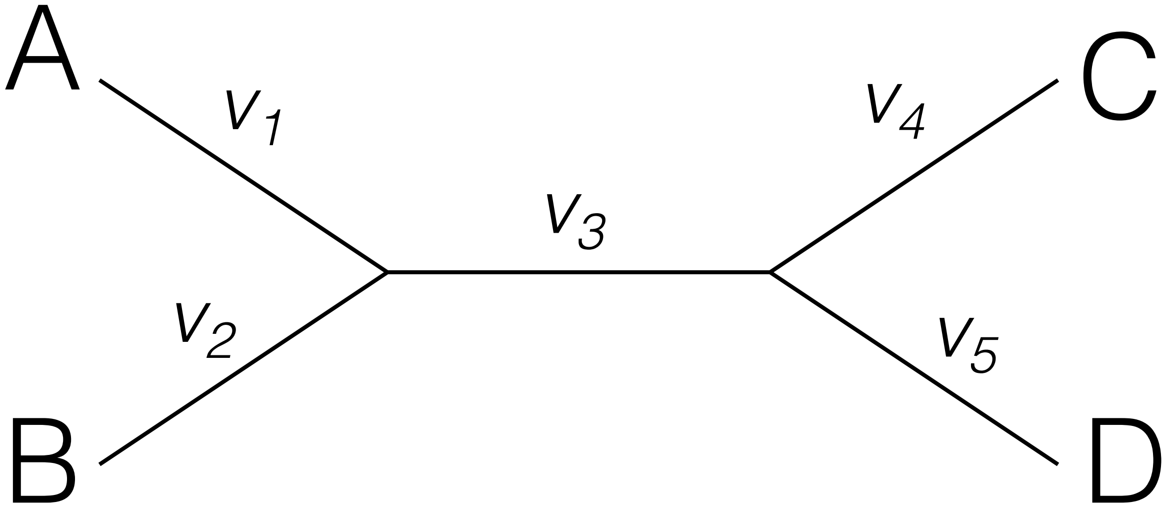
Application of models for phylogenetic estimation

Models of molecular evolution

We discussed the general likelihood function for a tree, but where do the conditional probabilities come from?

Models of molecular evolution

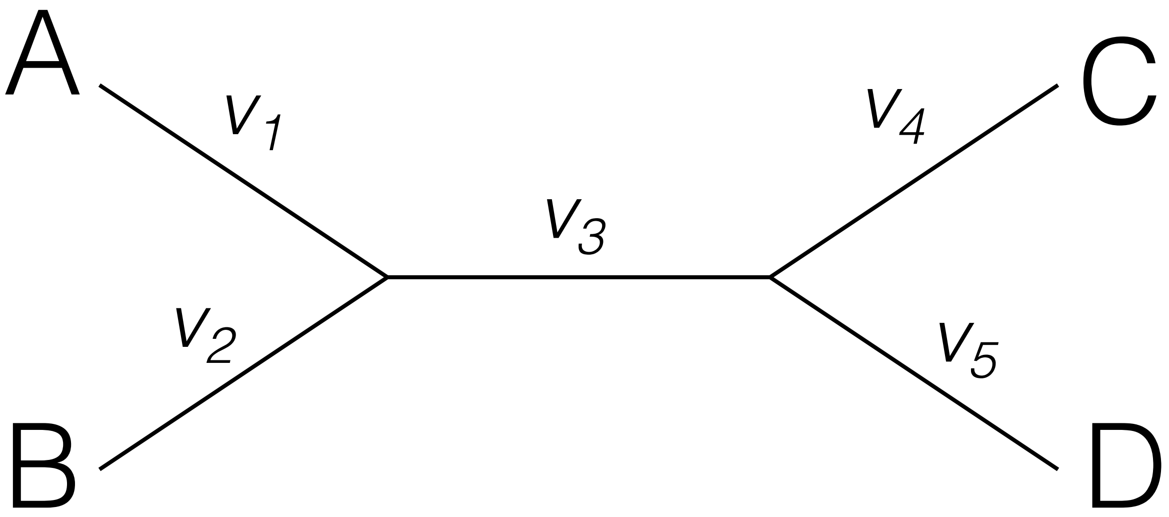
Topology and branch lengths



v_i – length of branch *i* measured in expected number of substitutions per site

Models of molecular evolution

Topology and branch lengths



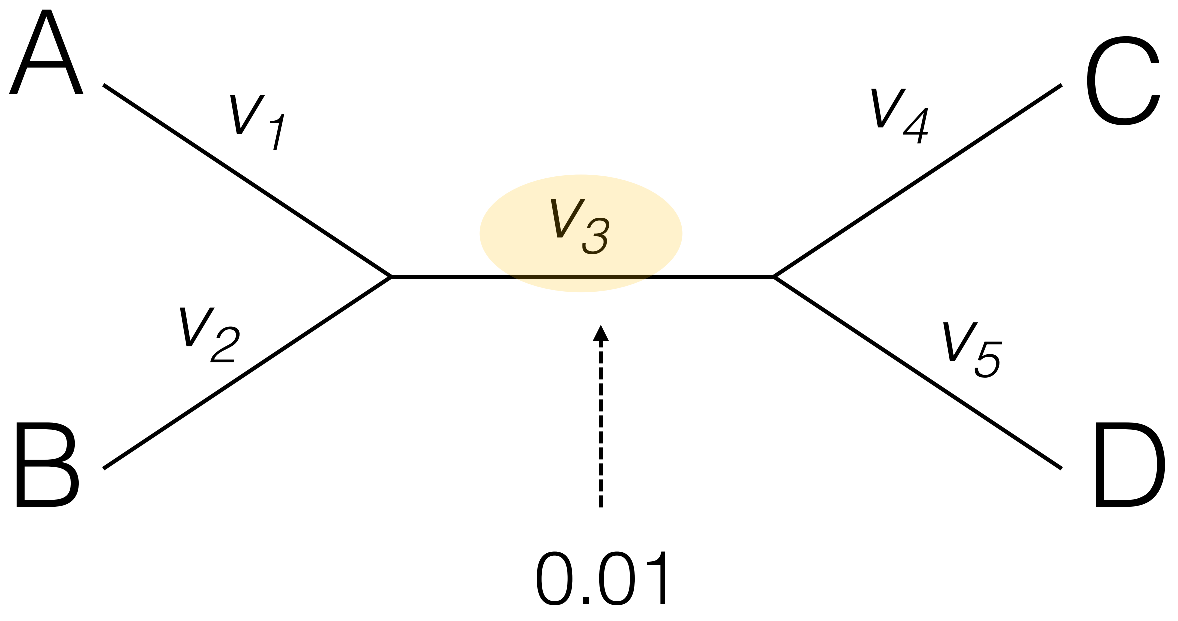
Transition rate matrix

$$
\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & -3\alpha & \alpha & \alpha & \alpha \\
C & \alpha & -3\alpha & \alpha & \alpha \\
G & \alpha & \alpha & -3\alpha & \alpha \\
T & \alpha & \alpha & \alpha & -3\alpha
\end{array}
$$

v_i– length of branch *i* measured in **expected number of substitutions per site**

Models of molecular evolution

Branch lengths (evolutionary distance) – the confounded measurement of rate and time



We expect 1 substitution for every 100 sites

Models of molecular evolution

Branch lengths = evolutionary distance
- These are the confounded measurement of rate and time

Consider a genome of 10,000bp

$$\frac{1\ substitution}{1\ million\ years} \times 100\ million\ years = \frac{100\ substitutions}{10{,}000bp} = 0.01$$

$$\frac{10\ substitutions}{1\ million\ years} \times 10\ million\ years = \frac{100\ substitutions}{10{,}000bp} = 0.01$$

Understanding absolute time requires *clock models*

Models of molecular evolution

Transition rate matrix = describes how one base changes into another per unit of evolutionary distance *t*

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & -3\alpha & \alpha & \alpha & \alpha \\ C & \alpha & -3\alpha & \alpha & \alpha \\ G & \alpha & \alpha & -3\alpha & \alpha \\ T & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

corrects for *multiple hits*

Time-reversible

C → G = G → C

Markov model

what happens in between the start and end state does not matter

Models of molecular evolution

Let's derive the *transition probabilities* using the easiest rate matrix, Juke and Cantor (1969)

$$
\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & -3\alpha & \alpha & \alpha & \alpha \\
C & \alpha & -3\alpha & \alpha & \alpha \\
G & \alpha & \alpha & -3\alpha & \alpha \\
T & \alpha & \alpha & \alpha & -3\alpha
\end{array}
$$

Models of molecular evolution

Let's derive the *transition probabilities* using the easiest model, Juke and Cantor (1969)

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & -3\alpha & \alpha & \alpha & \alpha \\ C & \alpha & -3\alpha & \alpha & \alpha \\ G & \alpha & \alpha & -3\alpha & \alpha \\ T & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

The equilibrium frequencies (π_i) are probability of starting in state *i*

$\pi_A = \pi_C = \pi_G = \pi_T = 0.25$

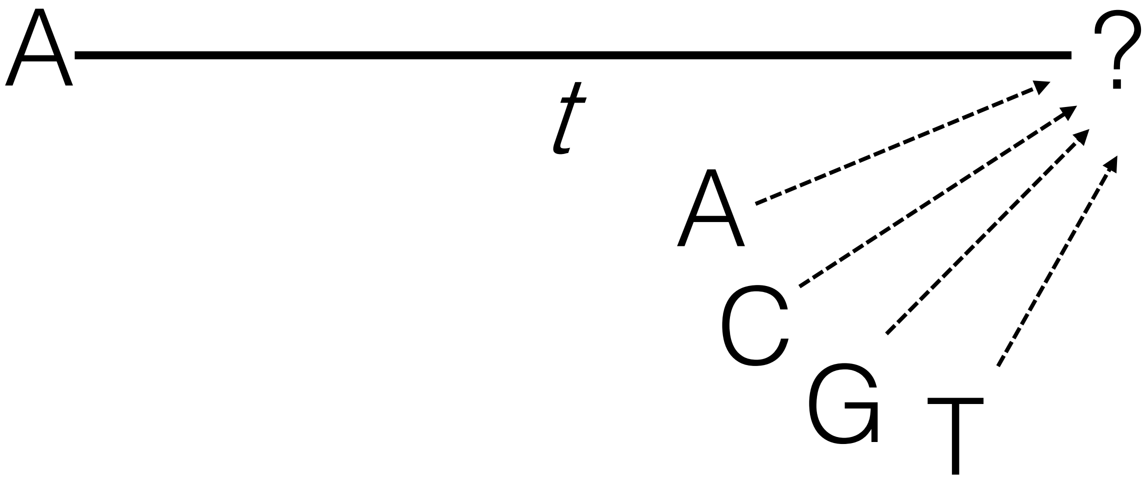
A is changing into C at rate α, G at rate α, and T at rate α. But remember, the process is reversible, so A stays an A with rate *-3*α

Models of molecular evolution

Consider a single site that starts as an A

And some amount of time t passes

Site A can become anything at rate μ

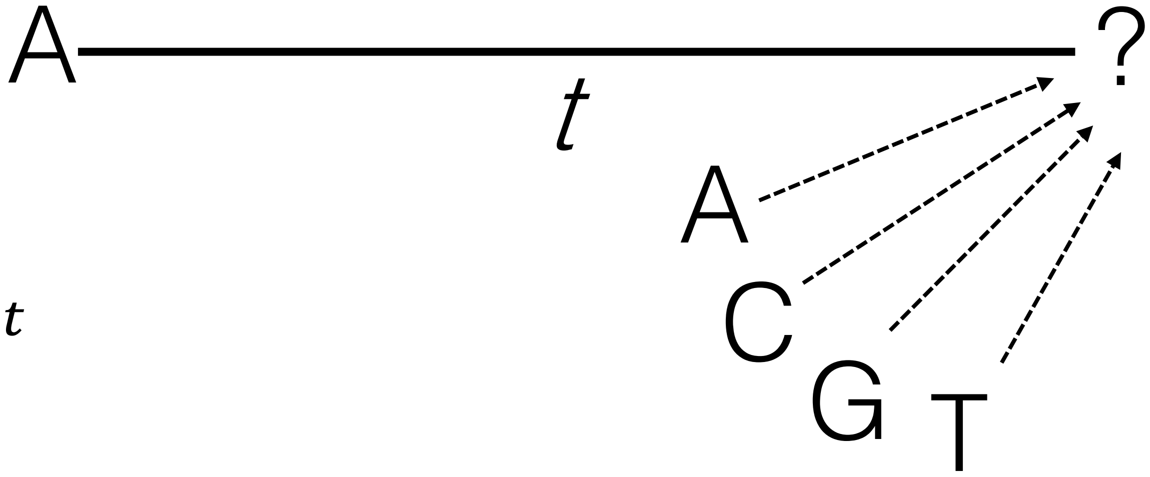


Becuase rate μ represents *any* base change, a change from A to a specific base, α, is $\frac{1}{4}\mu$

We could rewrite as $\mu = 4\alpha$

Models of molecular evolution

The probability of something with rate μ over time t is given by the **Poisson distribution**!



$$P(nothing\ happens) = e^{-\mu t}$$

$$P(at\ least\ one\ thing\ happens) = 1 - e^{-\mu t}$$

$$P(the\ thing\ that\ happened\ was\ a\ T) = {}^{1}/_{4}$$

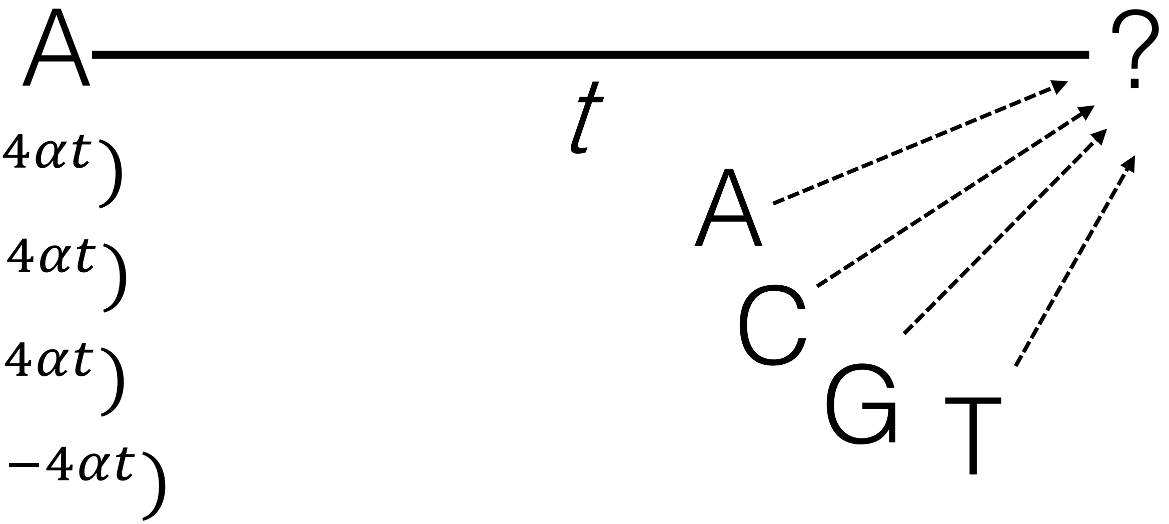
$$P(end\ with\ T|\ start\ with\ A) = \tfrac{1}{4}\times(1 - e^{-\mu t})$$

$$= \tfrac{1}{4}\times(1 - e^{-4\alpha t})$$

remember! $\mu = 4\alpha$

Models of molecular evolution

What do all of the probability calculations look like?



$$P(end\ with\ T|\ start\ with\ A) = \frac{1}{4}\times(1-e^{-4\alpha t})$$

$$P(end\ with\ G|\ start\ with\ A) = \frac{1}{4}\times(1-e^{-4\alpha t})$$

$$P(end\ with\ C|\ start\ with\ A) = \frac{1}{4}\times(1-e^{-4\alpha t})$$

$$P(end\ with\ A|\ start\ with\ A) = \frac{1}{4}\times(1+3e^{-4\alpha t})$$

That -3α in our rate matrix ensures the probabilities sum to 1

$4\alpha t$ is our branch length. We can now change the values of the model parameters to find the most likely tree!

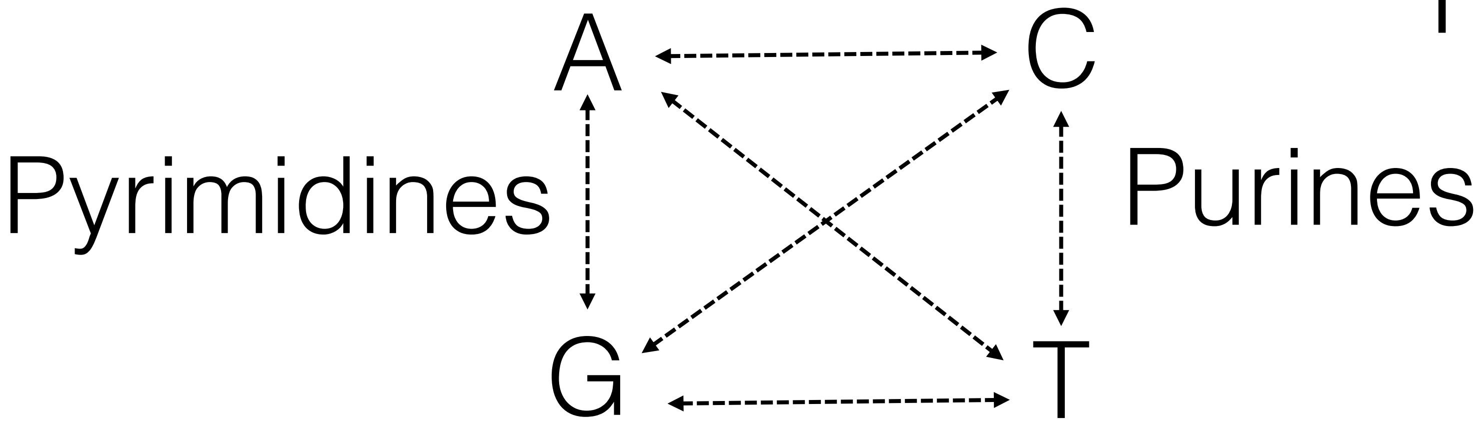
Models of molecular evolution

Models can be more complex!

Some types of substitutions might occur at a different rate than others

Hasegawa, Kishino, and Yano (1985) assumed transitions would be more frequent than transversions

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



Models of molecular evolution

Models can be more complex!

Maybe let all substitution types have their own rate and let the data decide?

GTR (Tavaré 1986)

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

Models of molecular evolution

There are many more models in existence, but most are between JC69 and GTR.

People used to spend a lot of time trying to pick the best one for their data. Now, programs like RAxML always use GTR or IQTREE can a automate the model selection process for you.

There are models for amino acids and codons too, which bring up additional theoretical complexities not covered here.

Models of molecular evolution

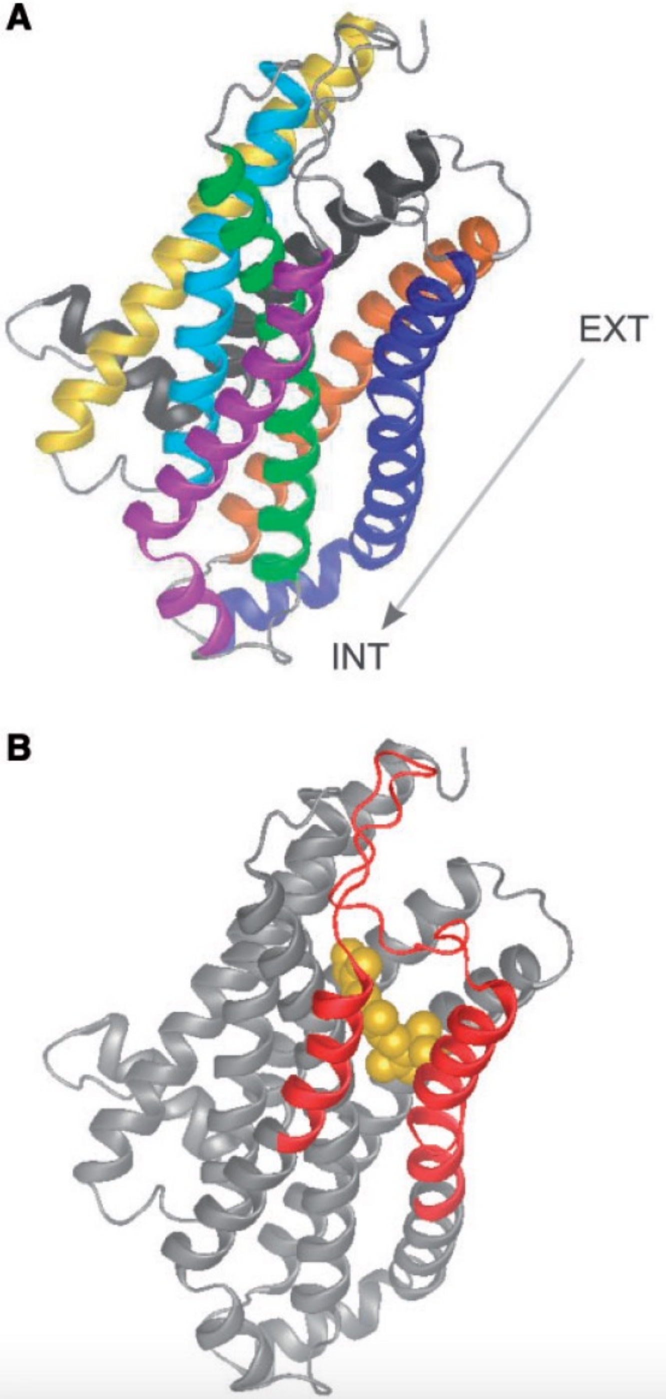
All models can incorporate rate variation though

Rate variation – some site evolve quickly and some evolve slowly

Models of molecular evolution

All models can incorporate rate variation though

Rate variation – some site evolve quickly and some evolve slowly

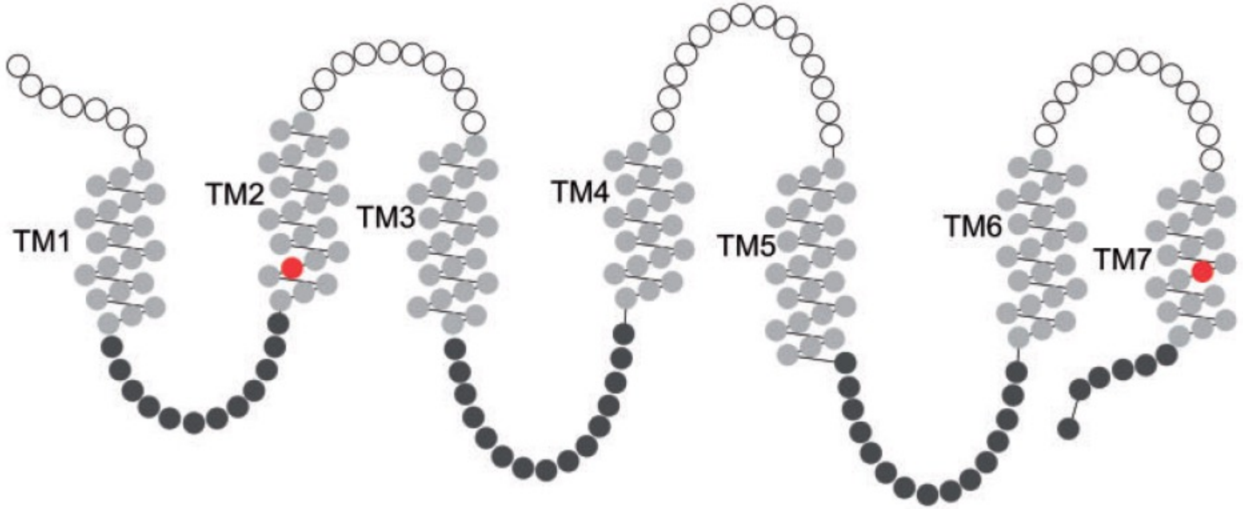


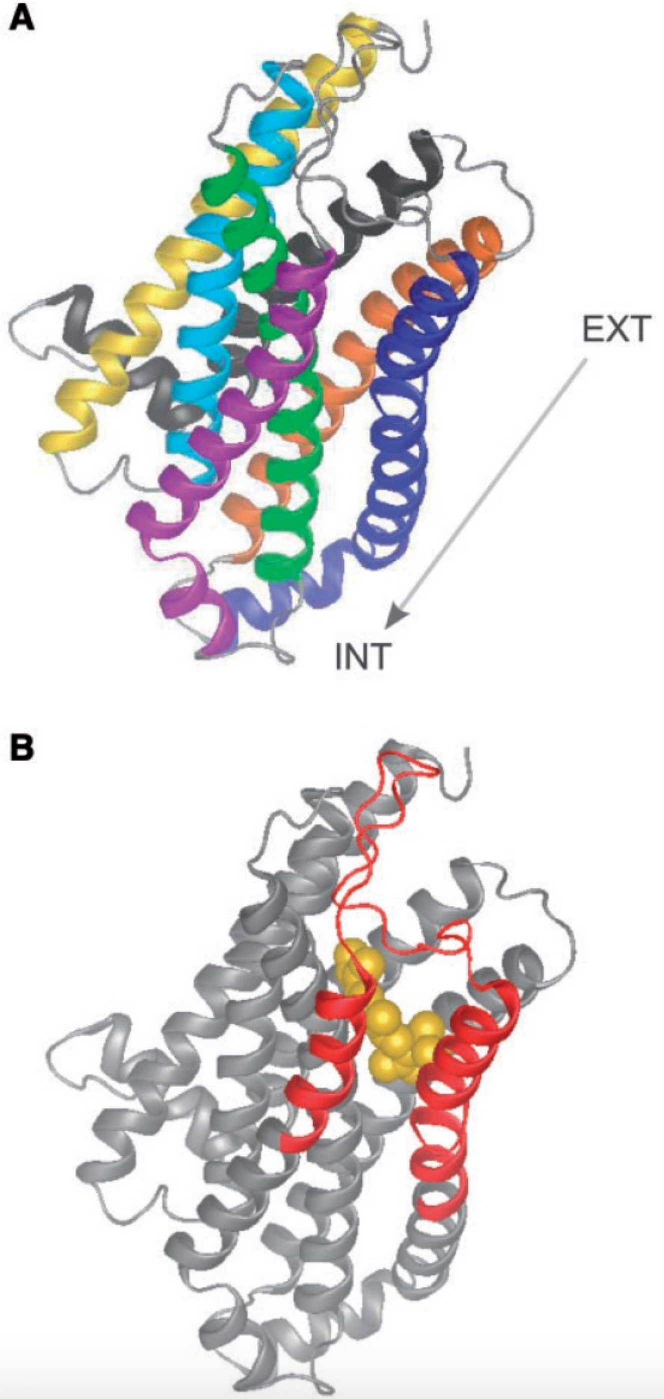
Yoder et al. 2015

Models of molecular evolution

All models can incorporate rate variation though

Rate variation – some site evolve quickly and some evolve slowly





Yoder et al. 2015

Models of molecular evolution

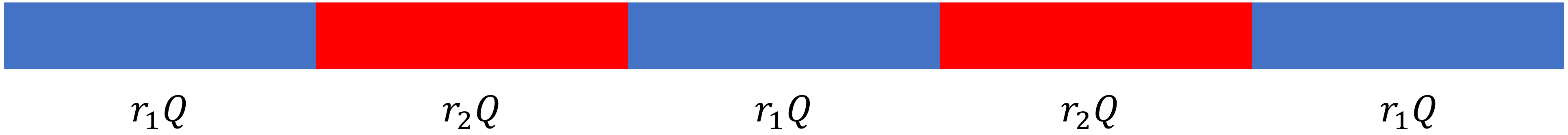
All models can incorporate rate variation though

Rate variation – some site evolve quickly and some evolve slowly

Models of molecular evolution

A site specific approach

Let different regions have different rates *(r)*

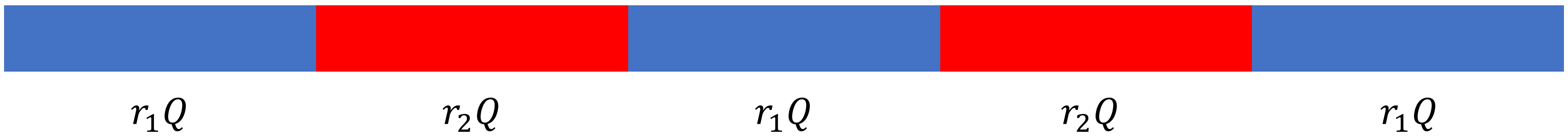


r_1Q r_2Q r_1Q r_2Q r_1Q

Models of molecular evolution

A site specific approach

Let different regions have different rates *(r)*



Our JC69 probability matrix will now look like:

For slow regions

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-r_1 4\alpha t}$$

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-r_1 4\alpha t}$$

For fast regions

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-r_2 4\alpha t}$$

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-r_2 4\alpha t}$$

Models of molecular evolution

A site specific approach

Pros

Convenient because we only have to estimate the likelihood of each site once

Cons

Requires *a priori* specification of regions (partitions)

Can still fit poorly for many sites

Models of molecular evolution

A mixture model approach

Integrate over multiple rates at each site

Consider we have 4 rates and *n* sites.

$$L(D|\theta) = \prod_{i=1}^{n} \left[\frac{1}{4} Pr(D_i|r_1) + \frac{1}{4} Pr(D_i|r_2) + \frac{1}{4} Pr(D_i|r_3) + \frac{1}{4} Pr(D_i|r_4)\right]$$

Models of molecular evolution

A mixture model approach

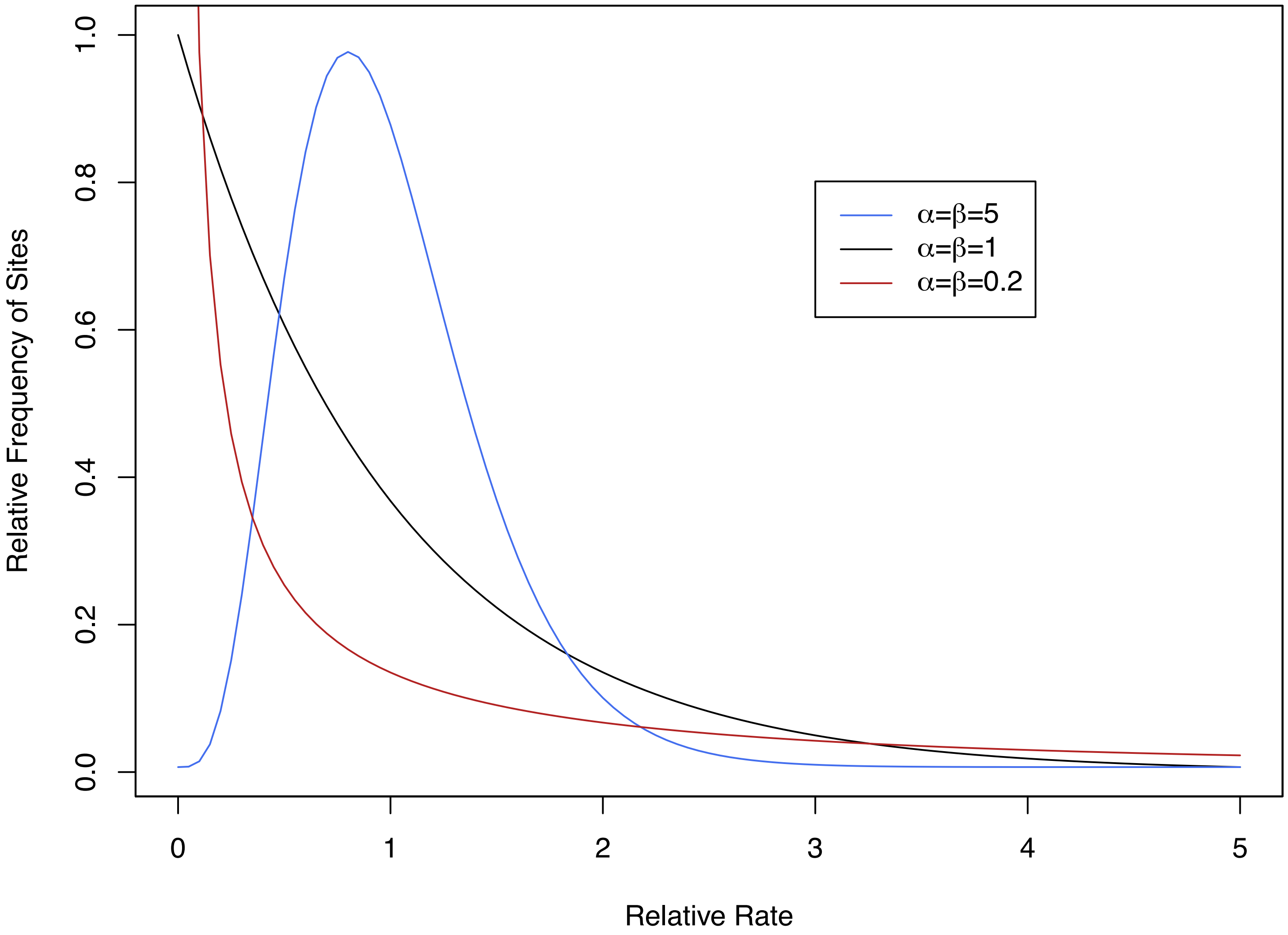
Where do those rates come from?

Models of molecular evolution

A mixture model approach

Assume rates follow a Gamma distribution

Higher values of alpha imply less rate variation



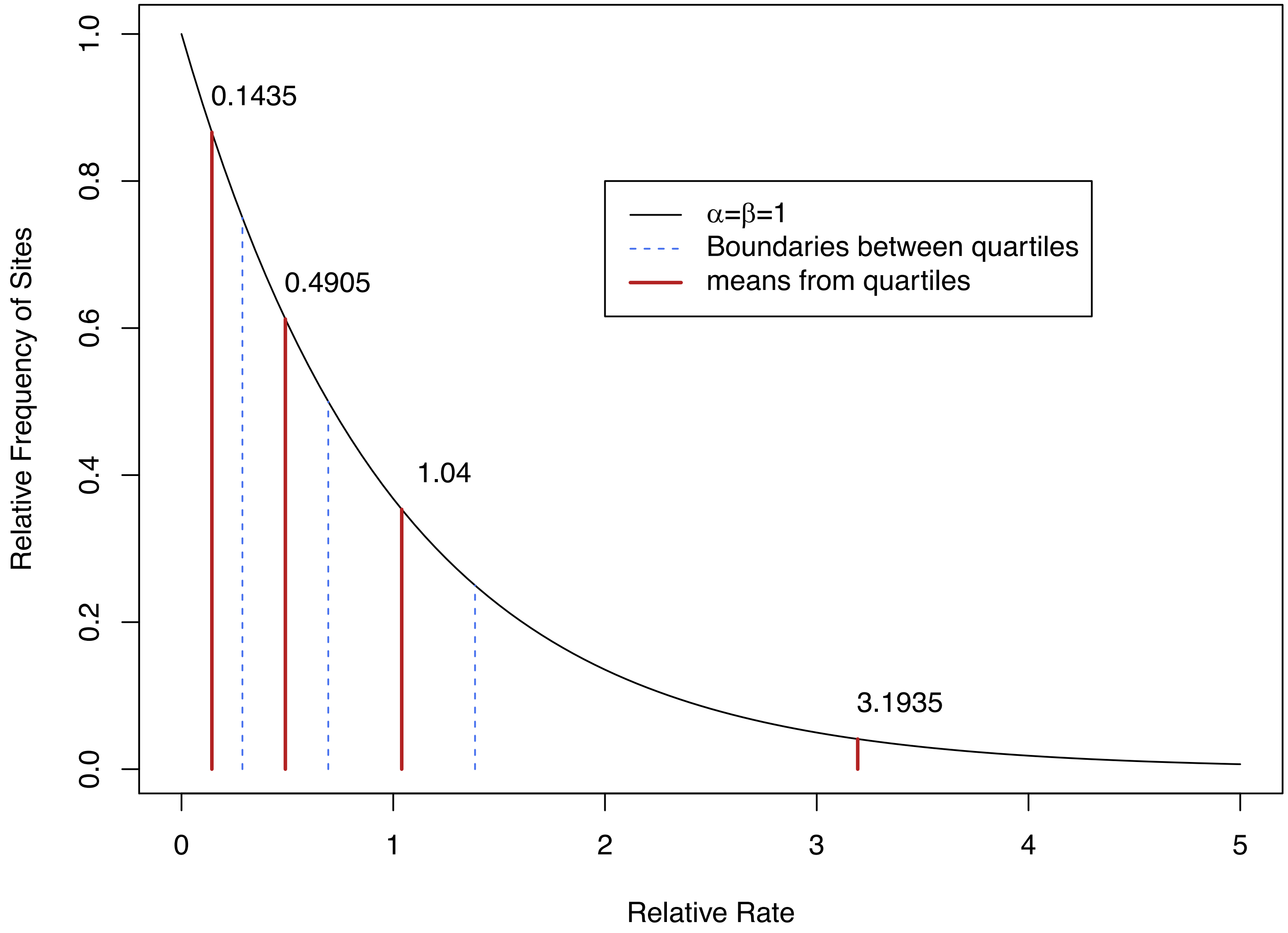
Yang 1994

Models of molecular evolution

A mixture model approach

We get discrete rates by splitting the distribution into quantiles (often 4)

The rates are then the means of those quantiles



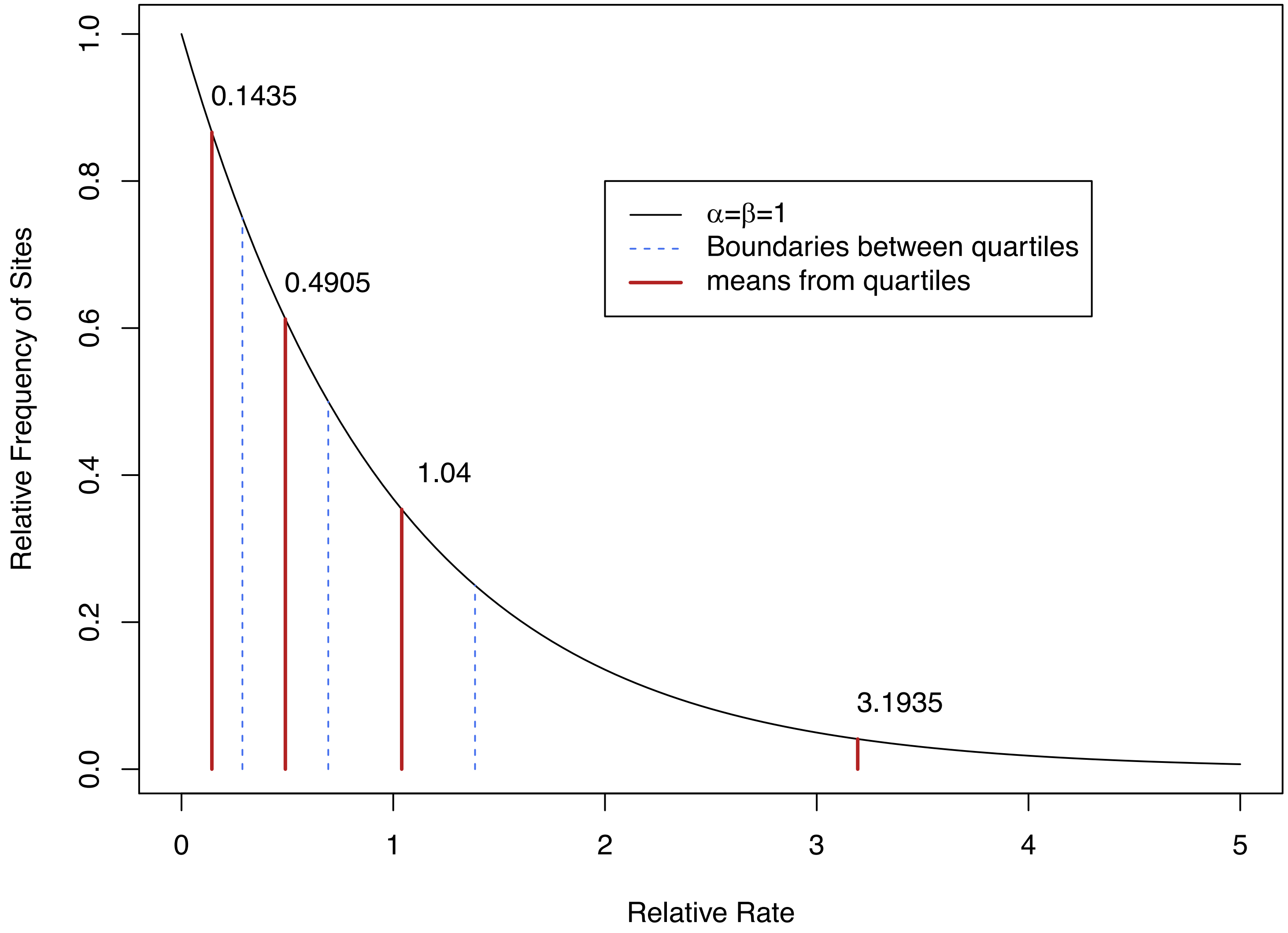
Yang 1994

Models of molecular evolution

A mixture model approach

We get discrete rates by splitting the distribution into quantiles (often 4)

The rates are then the means of those quantiles



Yang 1994

Models of molecular evolution

A mixture model approach

It is also possible to fit a model of rate heterogeneity where we do not assume the gamma distribution (free-rate model; Yang 1995)

The rate classes are approximated from the data directly

Estimating the rate classes requires having sufficient data and the gamma distribution is still widely used

Learning Goals

Explain terminology

Primer on probability and likelihood

Models of molecular evolution

How to select a model

Application of models for phylogenetic estimation

Model Selection

The Akaike Information Criterion (AIC)

$$AIC = 2k - 2ln(L)$$

The number of parameters

The maximum likelihood estimate

Model Selection

The Akaike Information Criterion (AIC)

$$AIC = 2k - 2ln(L)$$

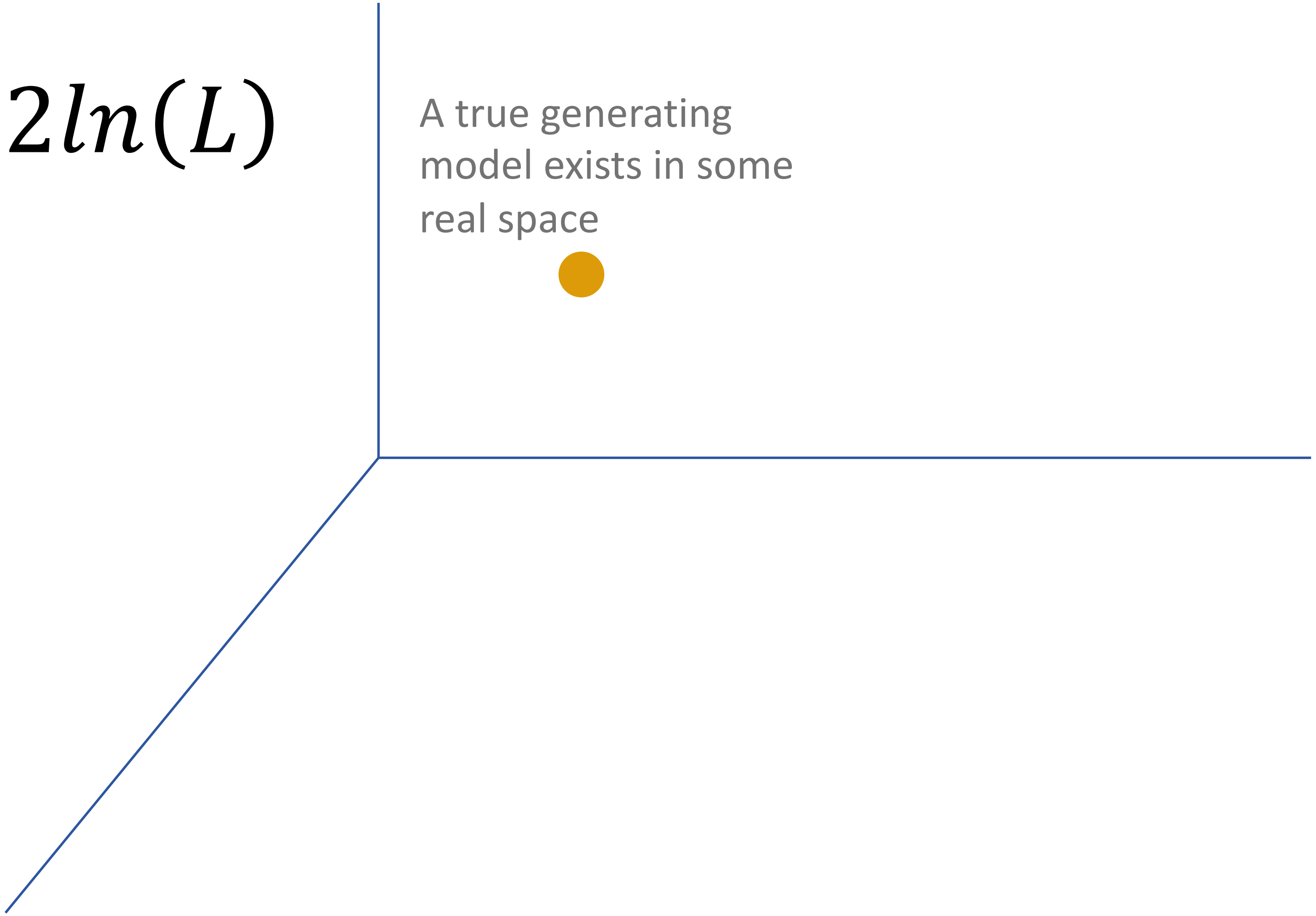
The number of parameters

The maximum likelihood estimate

Goal: Find the best (least worst) model among the set of possible models

Model Selection

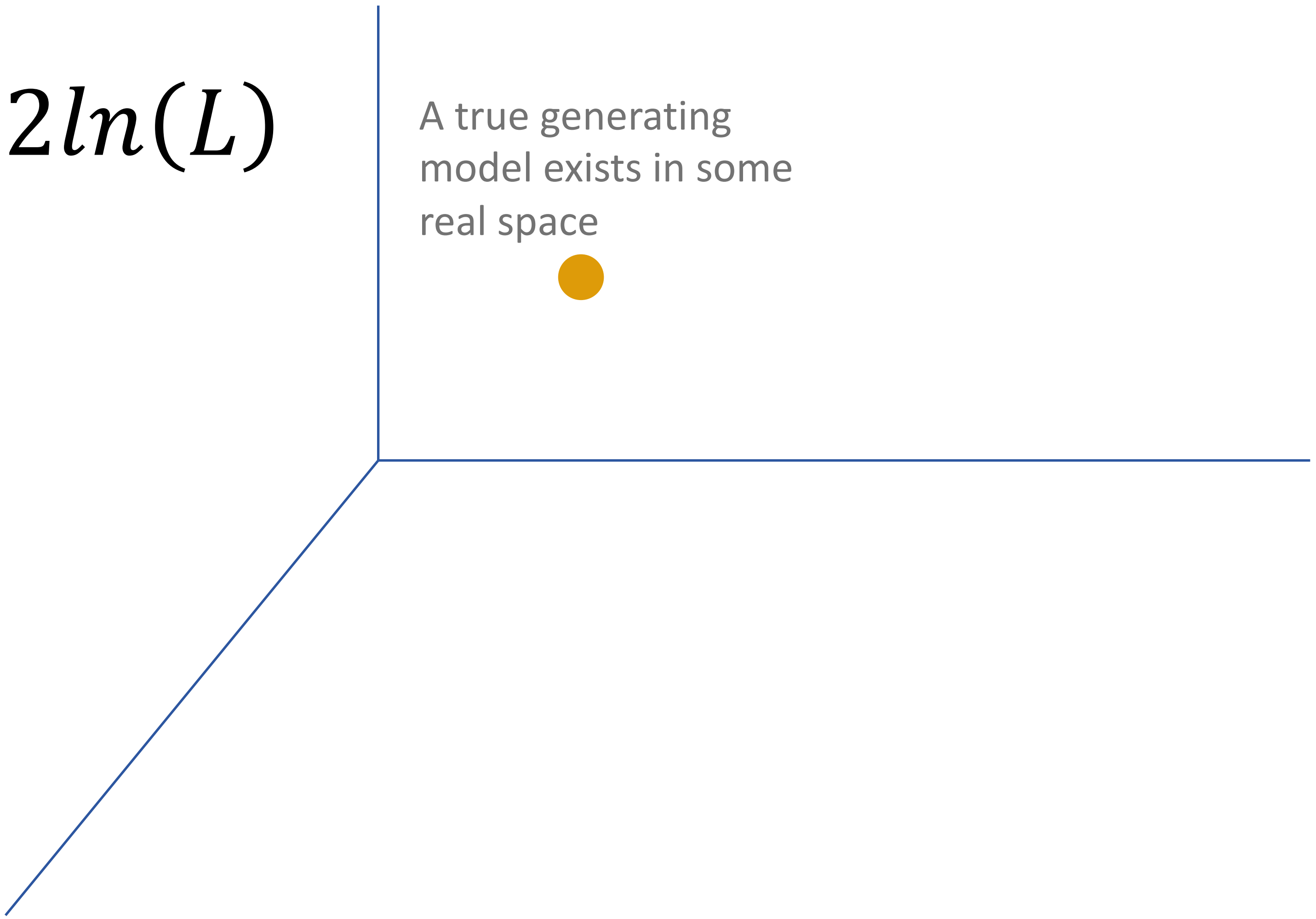
$$AIC = 2k - 2ln(L)$$



A true generating model exists in some real space

Model Selection

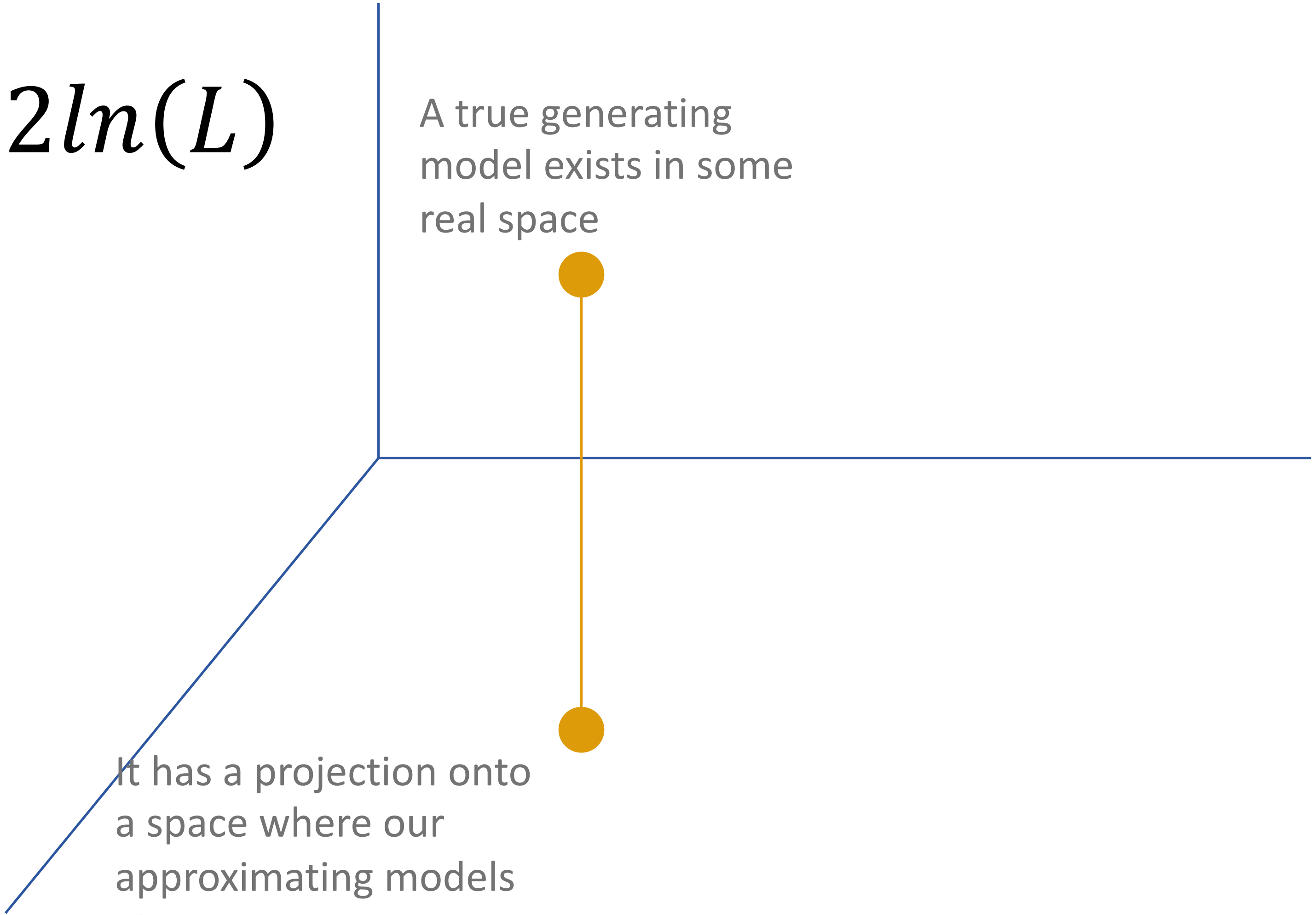
$$AIC\ = 2k - 2ln(L)$$



A true generating model exists in some real space

Model Selection

$$AIC = 2k - 2ln(L)$$

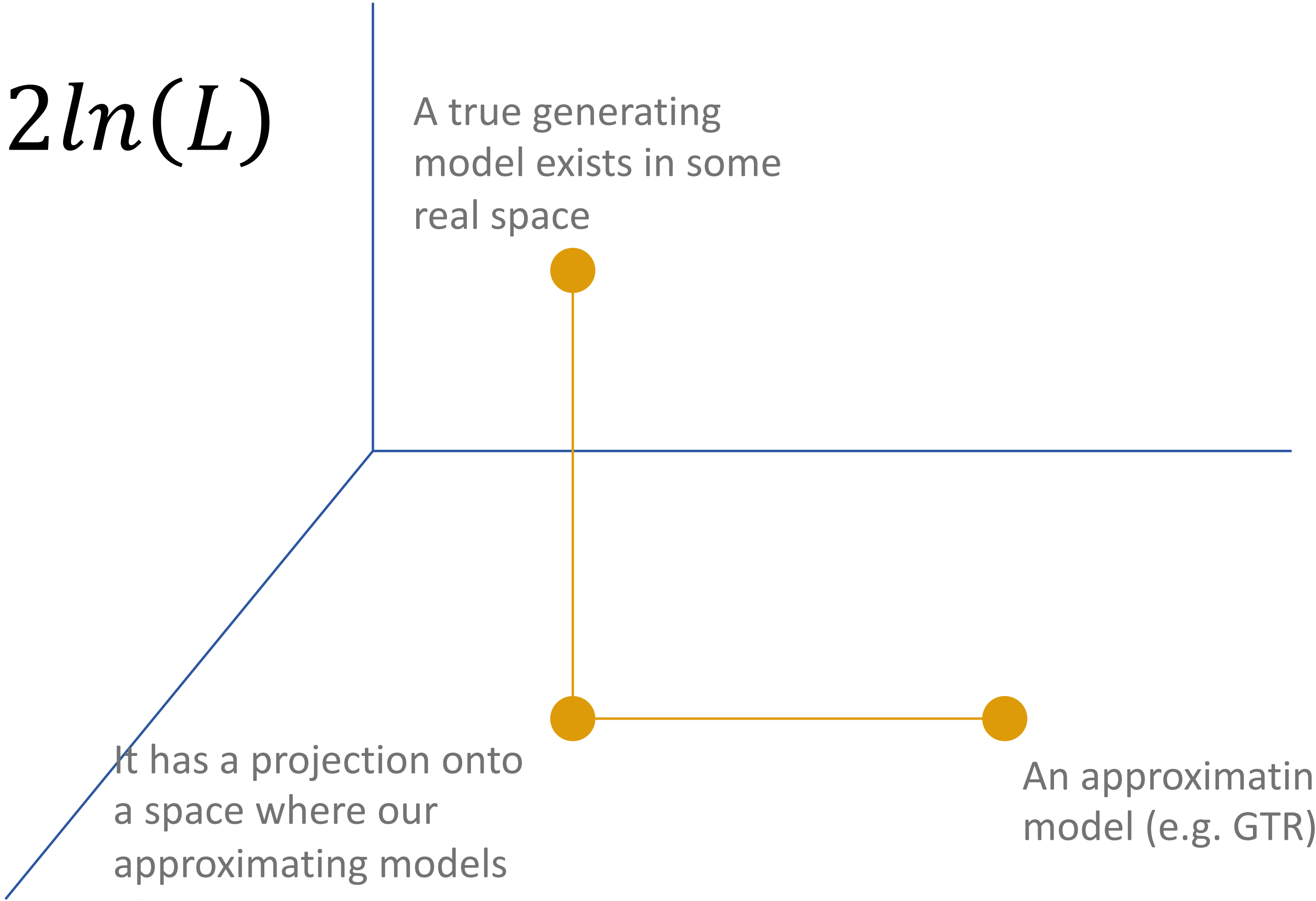


A true generating model exists in some real space

It has a projection onto a space where our approximating models also exist

Model Selection

$$AIC = 2k - 2ln(L)$$



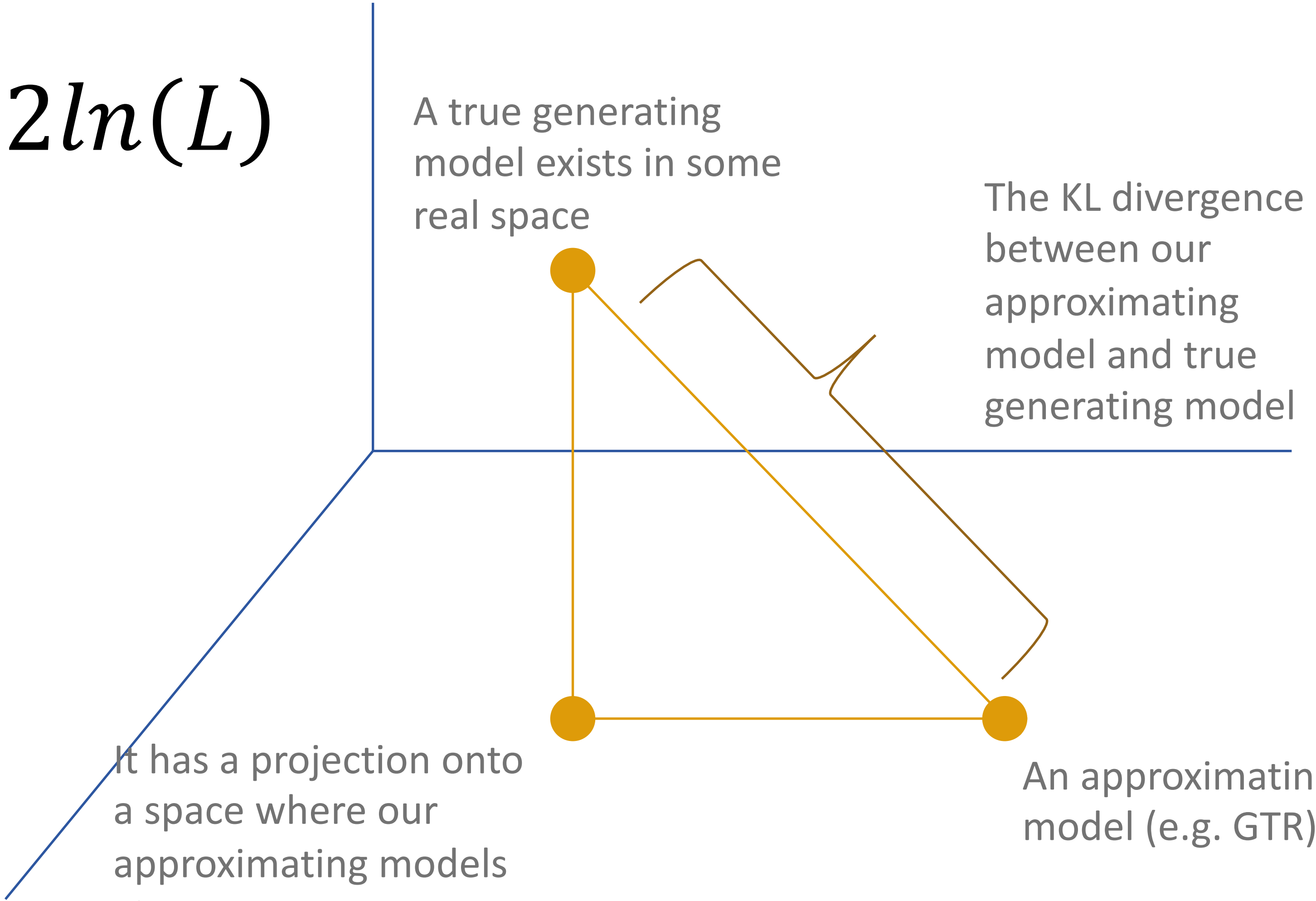
A true generating model exists in some real space

It has a projection onto a space where our approximating models also exist

An approximating model (e.g. GTR)

Model Selection

$$AIC\ = 2k - 2ln(L)$$



A true generating model exists in some real space

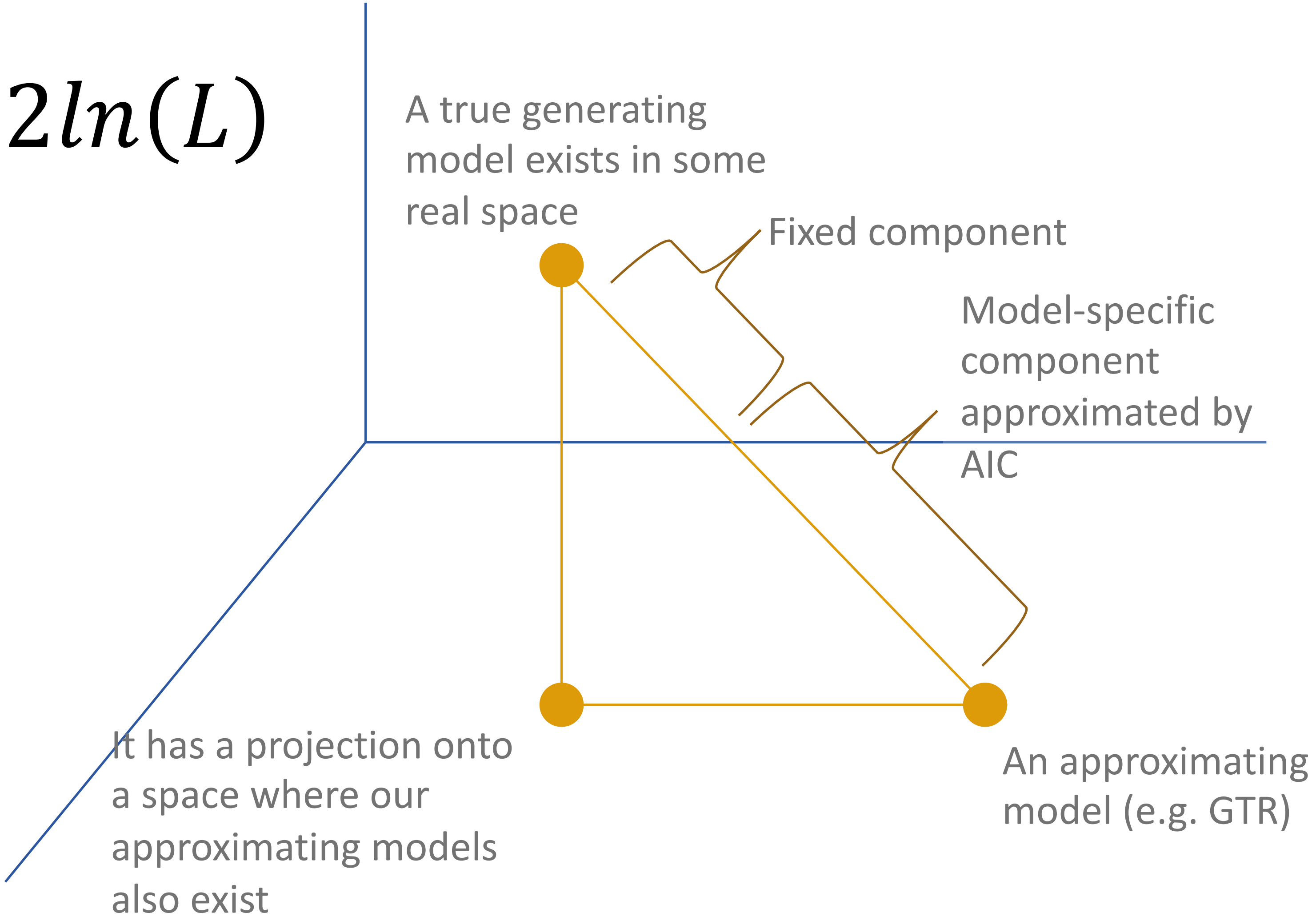
The KL divergence between our approximating model and true generating model

It has a projection onto a space where our approximating models also exist

An approximating model (e.g. GTR)

Model Selection

$$AIC = 2k - 2ln(L)$$



A true generating model exists in some real space

Fixed component

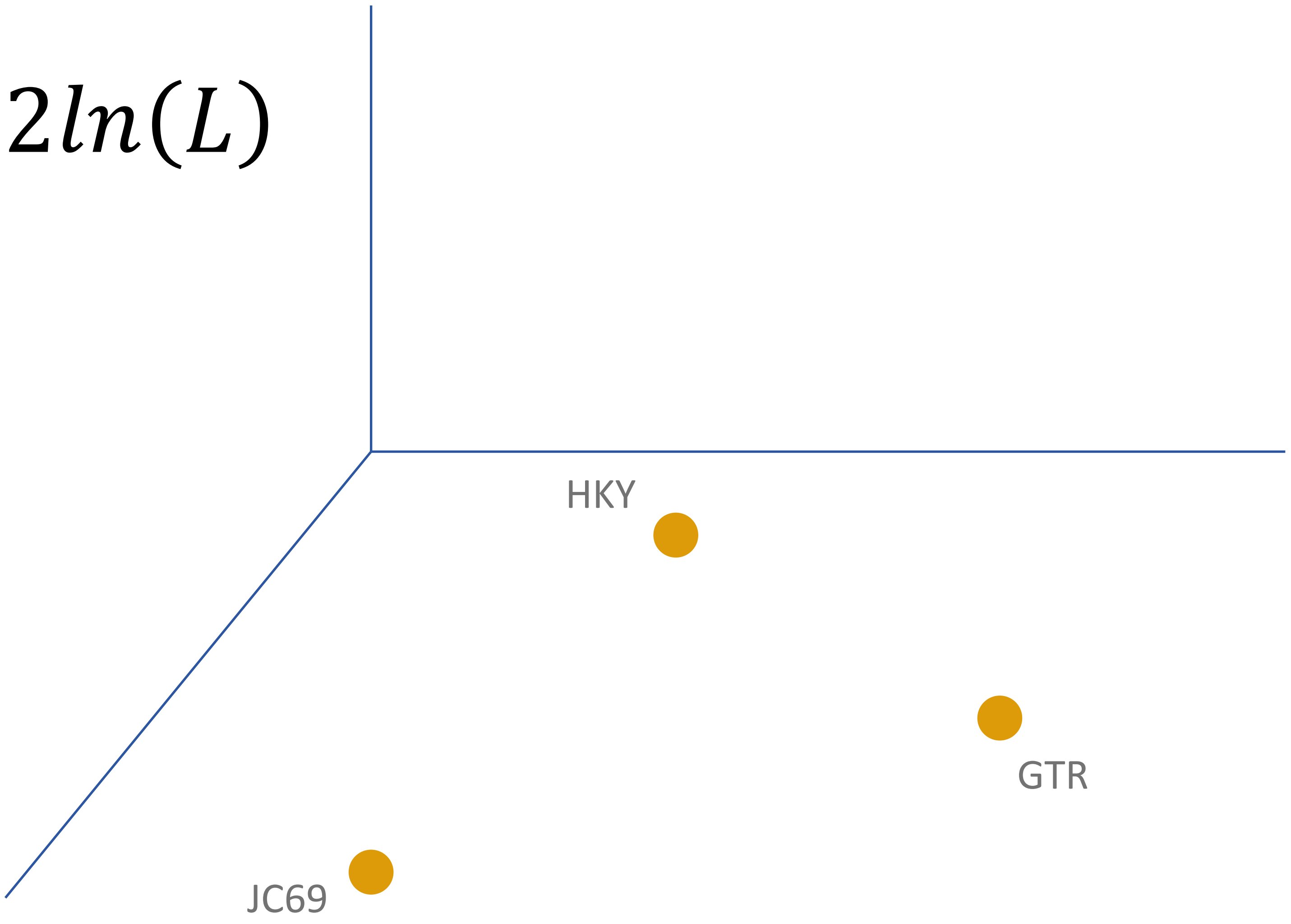
Model-specific component approximated by AIC

It has a projection onto a space where our approximating models also exist

An approximating model (e.g. GTR)

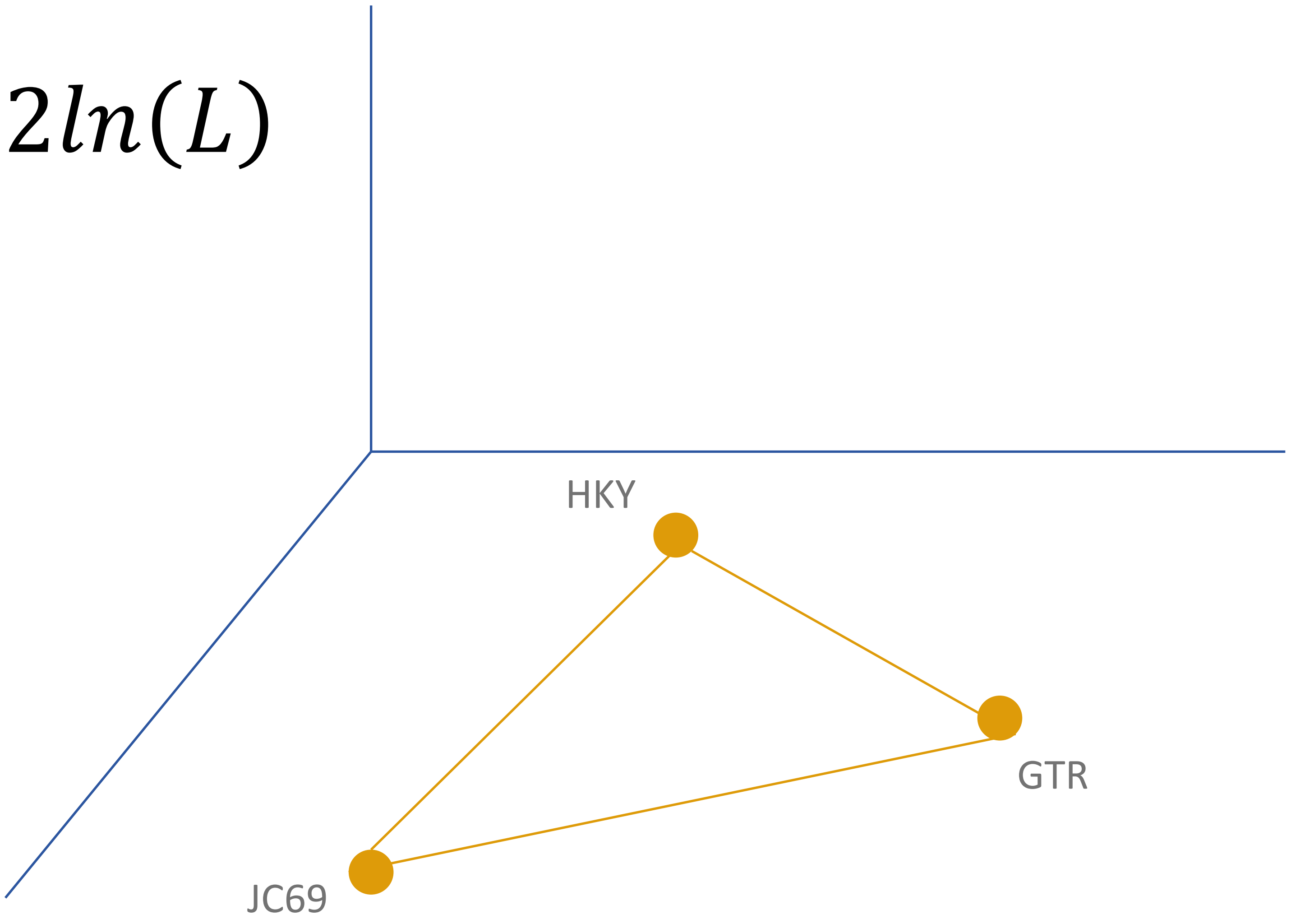
Model Selection

$$AIC\ = 2k - 2ln(L)$$



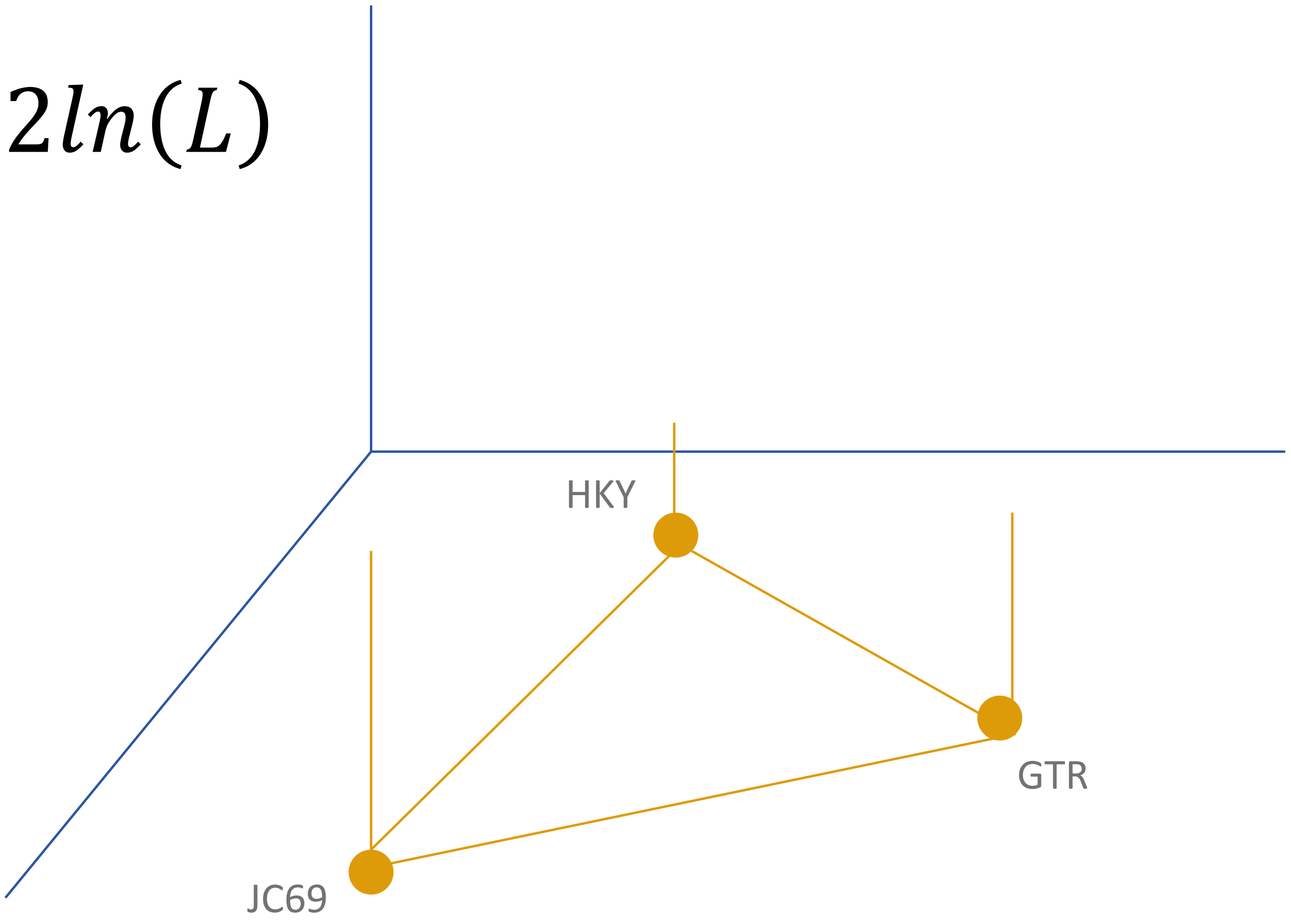
Model Selection

$$AIC = 2k - 2ln(L)$$



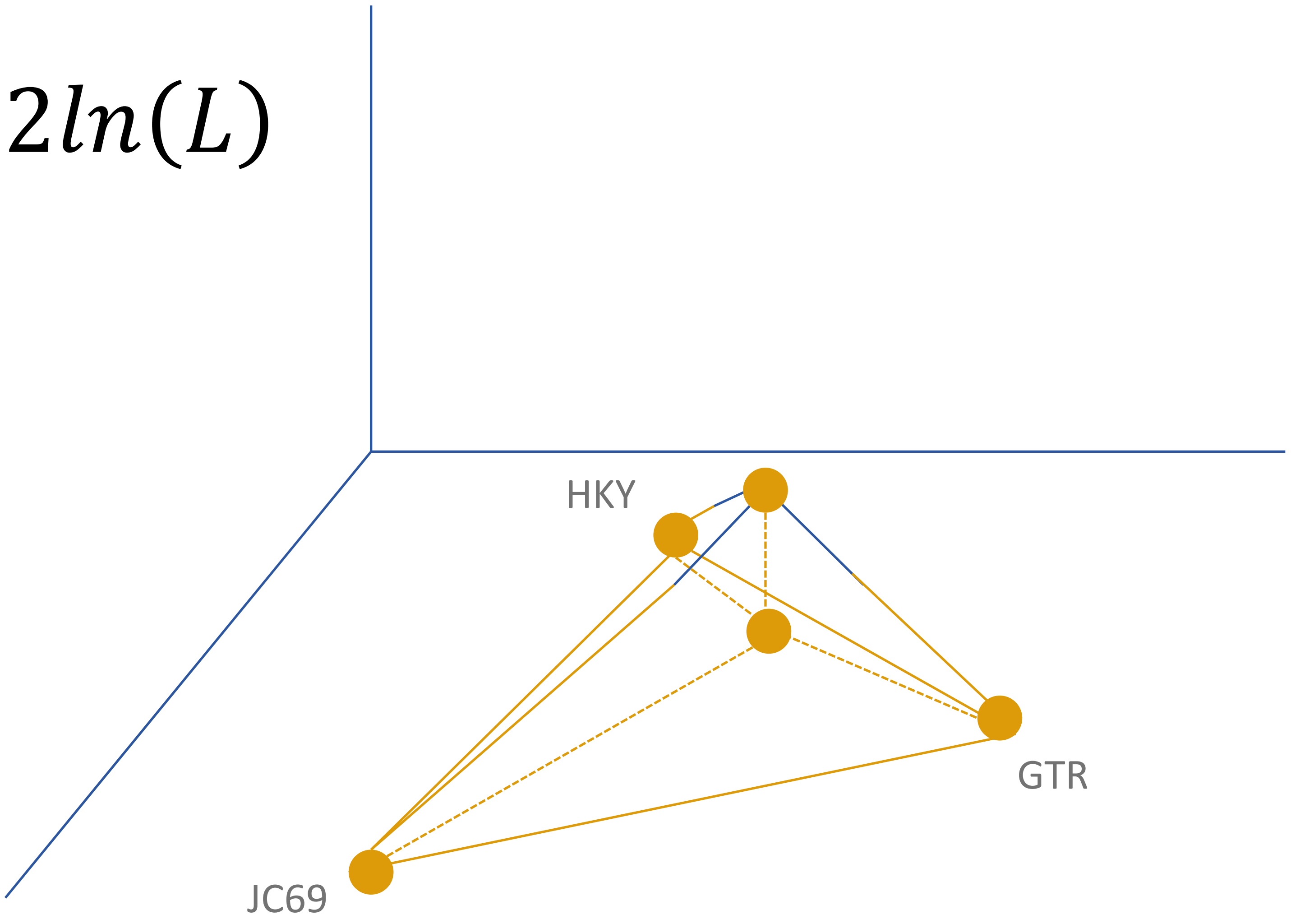
Model Selection

$$AIC = 2k - 2ln(L)$$



Model Selection

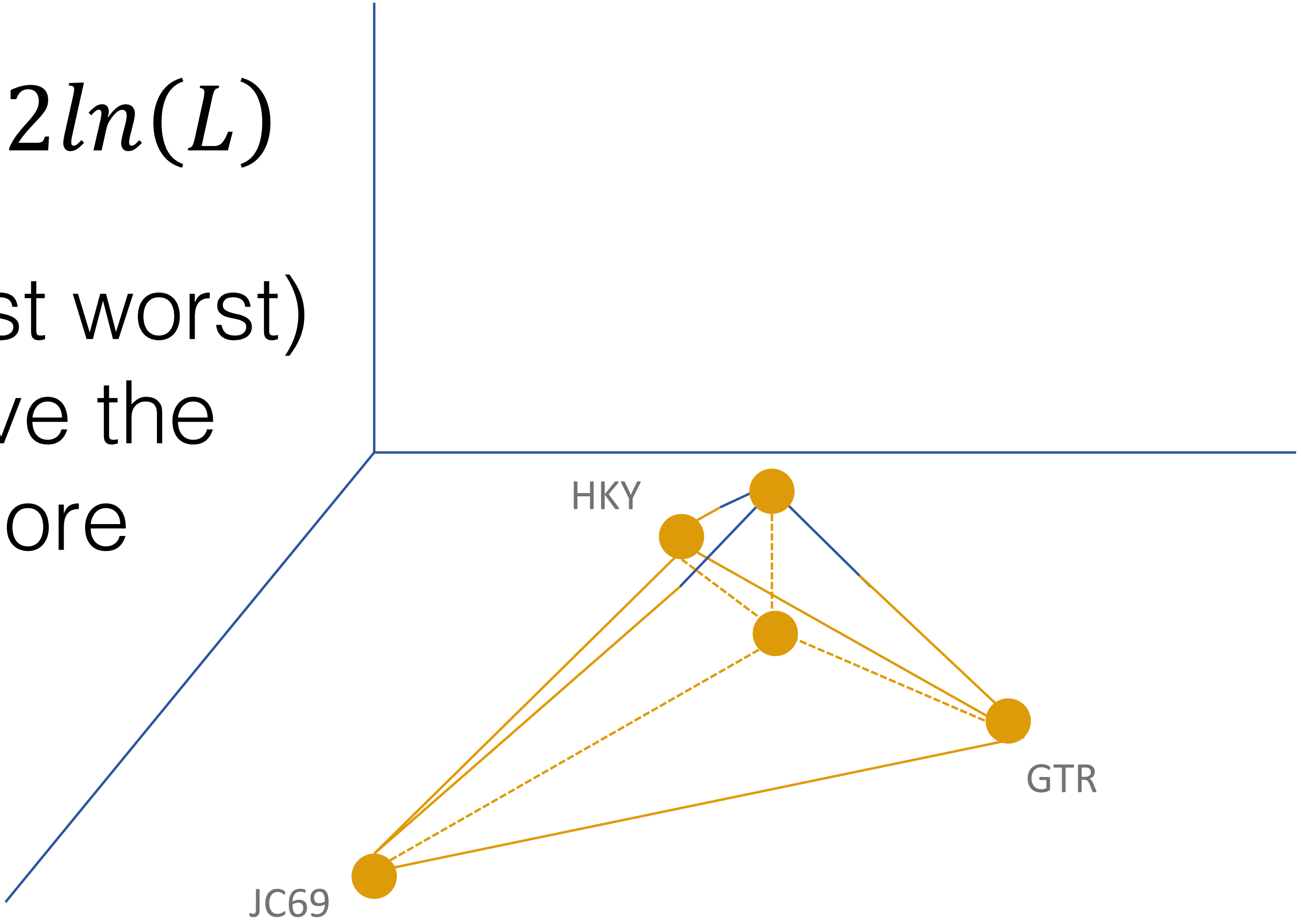
$$AIC = 2k - 2ln(L)$$



Model Selection

$$AIC = 2k - 2ln(L)$$

The best (least worst) model will have the lowest AIC score



Model Selection

$$AIC\ = 2k - 2ln(L)$$

Model	k	ln(L)	AIC	ΔAIC
JC69	1	-533214	1066430	16825
GTR	8	-524800	1049608	3
HKY	5	-524800	1049605	0

The best model will have the lowest AIC score

Usually a difference of 2 AIC points is accepted

Model Selection

Alternatives to AIC

AIC Corrected for small sample size

$$AICc \ = 2k - 2ln(L) + \frac{2k^2 + 2k}{n - k - 1}$$

Bayesian information criterion

$$BIC \ = kln(n) - 2ln(L)$$

Not a lot of agreement about which is best. Phylogenetics often uses the AICc. Programs will often return all three.

Learning Goals

Explain terminology

Primer on probability and likelihood

Models of molecular evolution

How to select a model

Application of models for phylogenetic estimation

Application

It is now possible to generate many loci for many species

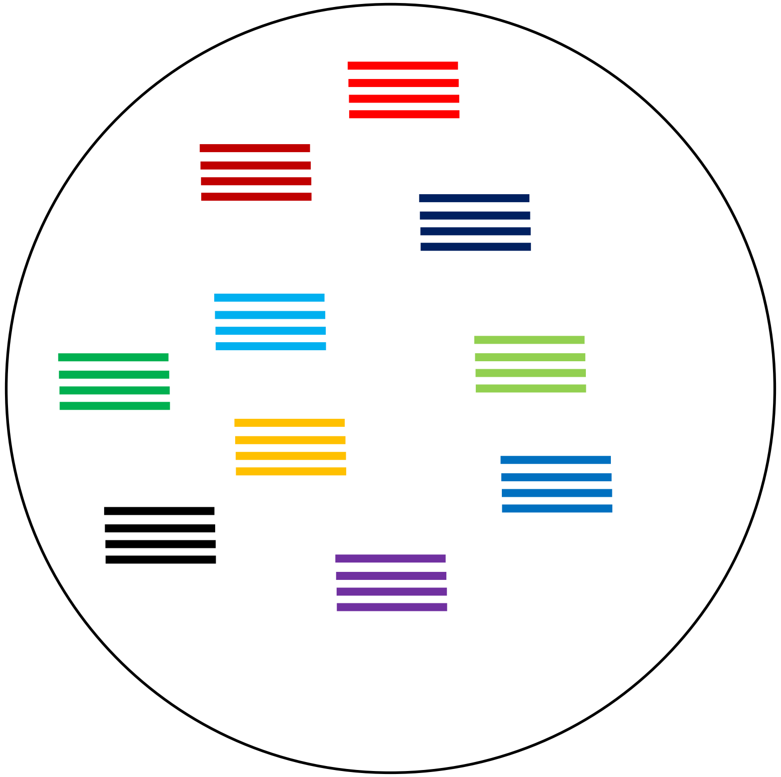
We will cover different ways of generating sequence data on Wednesday

The good news is that if you can analyze 1 locus well, you can do it for 100 or many more

Application

Concatenated maximum likelihood with partitioned mixture models

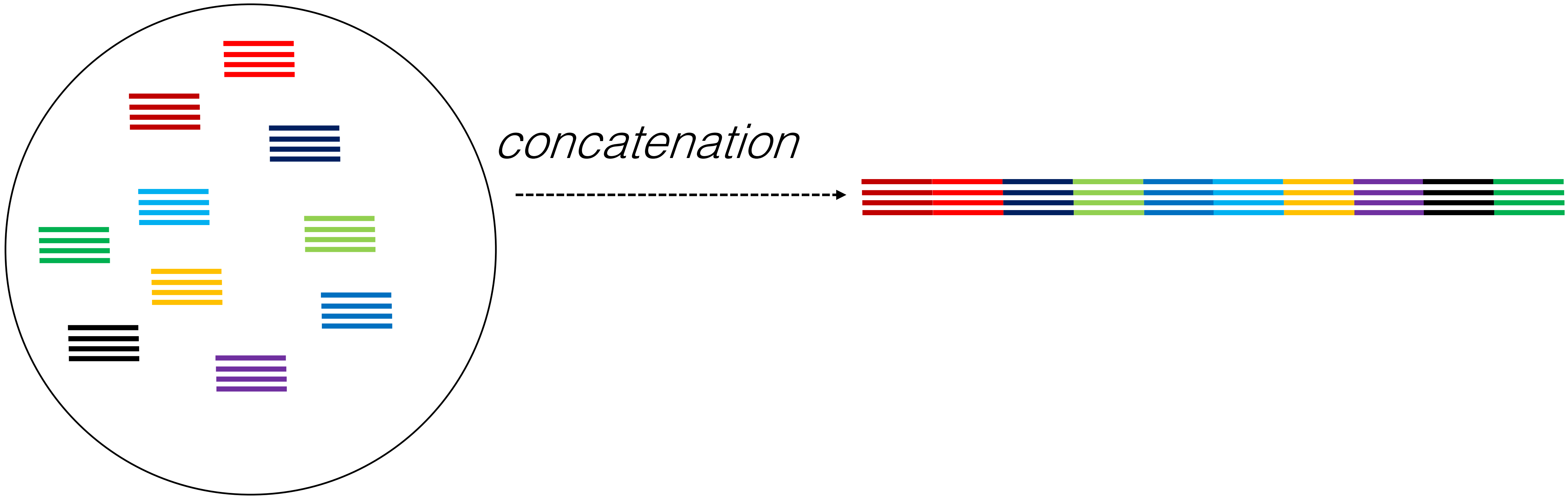
10 loci (genes) for 4 species



Application

Concatenated maximum likelihood with partitioned mixture models

10 loci (genes) for 4 species

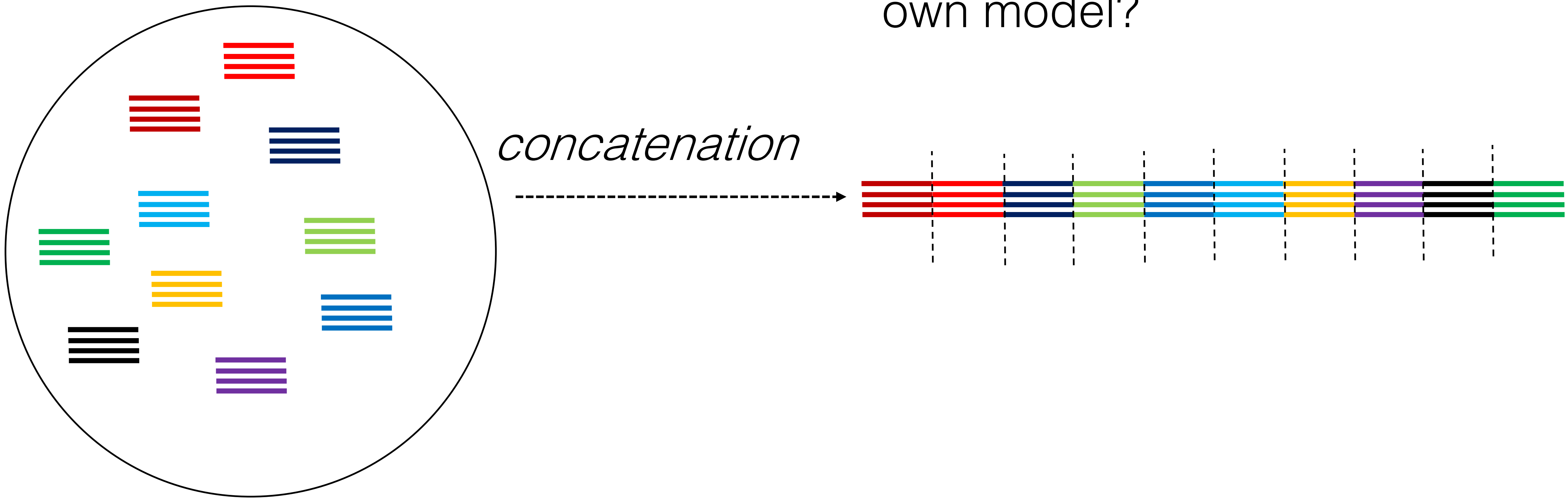


Application

Concatenated maximum likelihood with partitioned mixture models

10 loci (genes) for 4 species

Should each gene have its own model?



Application

Concatenated maximum likelihood with partitioned mixture models

Used in almost every phylogenomics paper to some extent

Some notable controversies

Application

Conflict between models based on the same data!

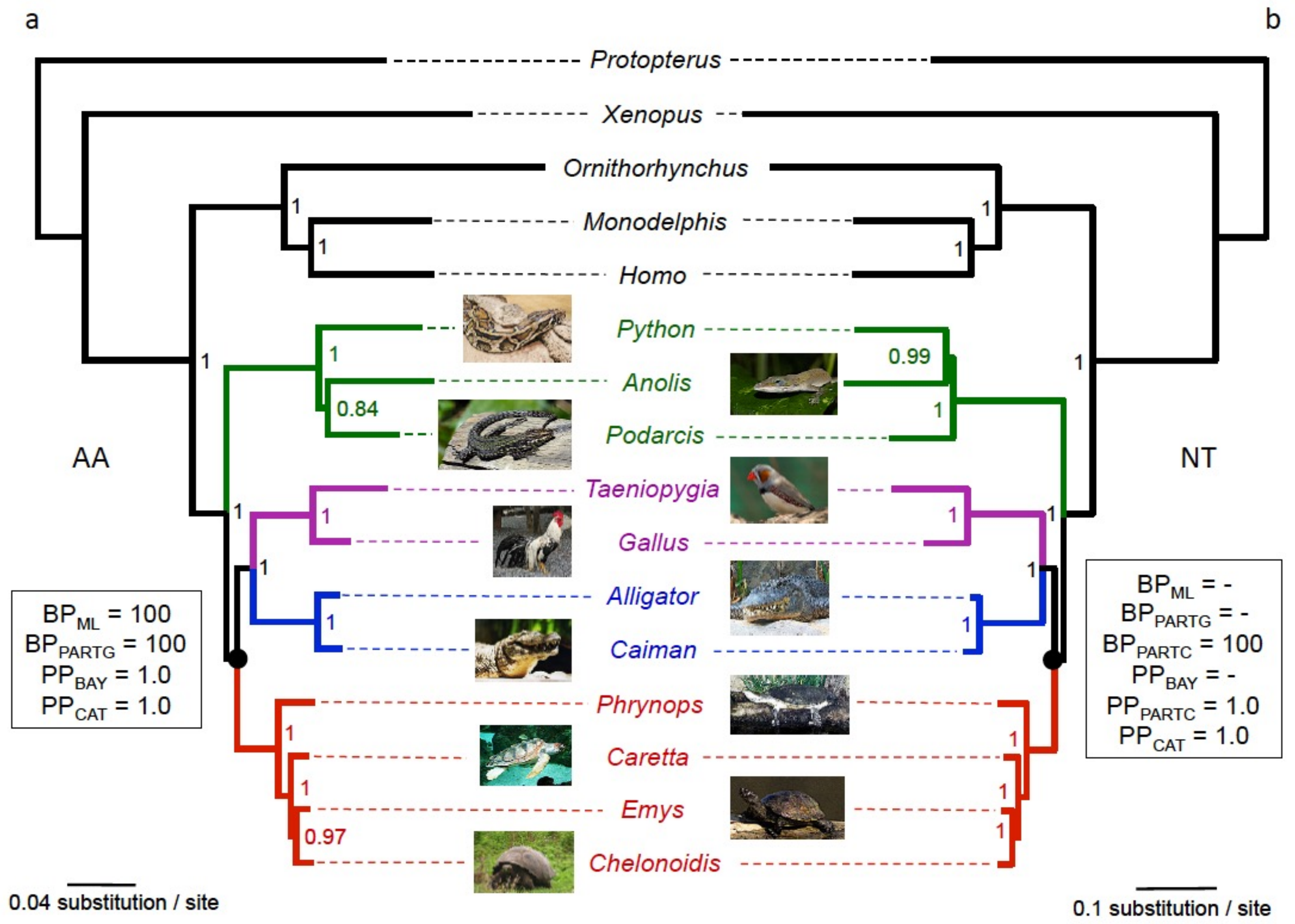


Figure 1 Phylogenetic relationships of amniotes as inferred from analyses of the 248-gene dataset. (a) Bayesian consensus topology obtained from analyses of the amino-acid dataset (62,342 sites) under the CAT-GTR + G4 mixture model. **(b)** Bayesian consensus topology obtained from analyses of the complete nucleotide dataset (187,026 sites) under the CAT-GTR + G4 mixture model. The nodal values indicate the clade Bayesian posterior probability (PP). Statistical support values obtained with different methods, models and data partitions detailed in Table 1 are reported in boxes for turtles plus archosaurs. Note the relative incongruence between the two trees concerning the position of *Python*. All pictures are from Wikimedia Commons, except for *Chelonoidis* from Y. Chiari. Please note also that the taxonomy of Galapagos turtles being currently revised, the appropriate species name for the *Chelonoidis* specimen included here might be *Chelonoidis* sp.

Chiari et al. 2012

Application

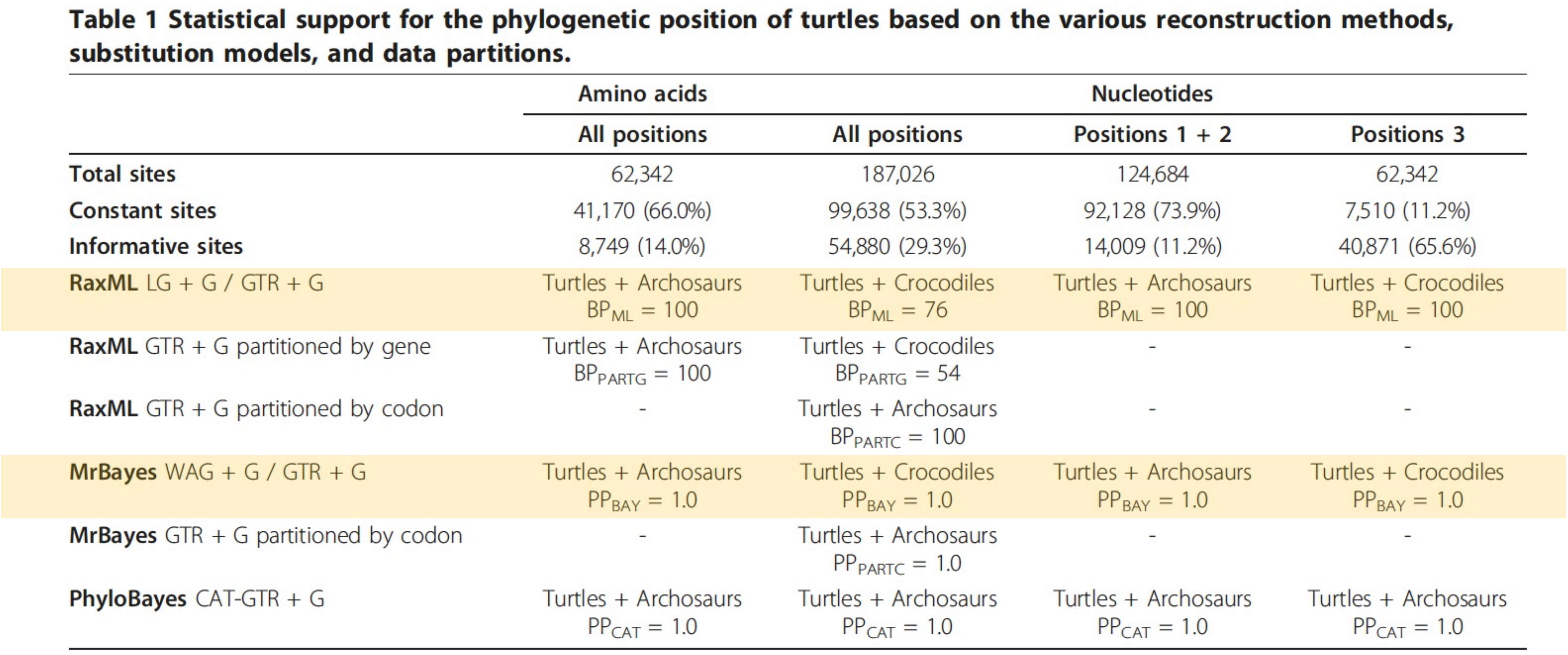


Table 1 Statistical support for the phylogenetic position of turtles based on the various reconstruction methods, substitution models, and data partitions.

	Amino acids	Nucleotides		
	All positions	**All positions**	**Positions 1 + 2**	**Positions 3**
Total sites	62,342	187,026	124,684	62,342
Constant sites	41,170 (66.0%)	99,638 (53.3%)	92,128 (73.9%)	7,510 (11.2%)
Informative sites	8,749 (14.0%)	54,880 (29.3%)	14,009 (11.2%)	40,871 (65.6%)
RaxML LG + G / GTR + G	Turtles + Archosaurs $BP_{ML} = 100$	Turtles + Crocodiles $BP_{ML} = 76$	Turtles + Archosaurs $BP_{ML} = 100$	Turtles + Crocodiles $BP_{ML} = 100$
RaxML GTR + G partitioned by gene	Turtles + Archosaurs $BP_{PARTG} = 100$	Turtles + Crocodiles $BP_{PARTG} = 54$	-	-
RaxML GTR + G partitioned by codon	-	Turtles + Archosaurs $BP_{PARTC} = 100$	-	-
MrBayes WAG + G / GTR + G	Turtles + Archosaurs $PP_{BAY} = 1.0$	Turtles + Crocodiles $PP_{BAY} = 1.0$	Turtles + Archosaurs $PP_{BAY} = 1.0$	Turtles + Crocodiles $PP_{BAY} = 1.0$
MrBayes GTR + G partitioned by codon	-	Turtles + Archosaurs $PP_{PARTC} = 1.0$	-	-
PhyloBayes CAT-GTR + G	Turtles + Archosaurs $PP_{CAT} = 1.0$	Turtles + Archosaurs $PP_{CAT} = 1.0$	Turtles + Archosaurs $PP_{CAT} = 1.0$	Turtles + Archosaurs $PP_{CAT} = 1.0$

Chiari et al. 2012

Application

The optimal partitioning for a phylogeny of Malpighiales is not obvious

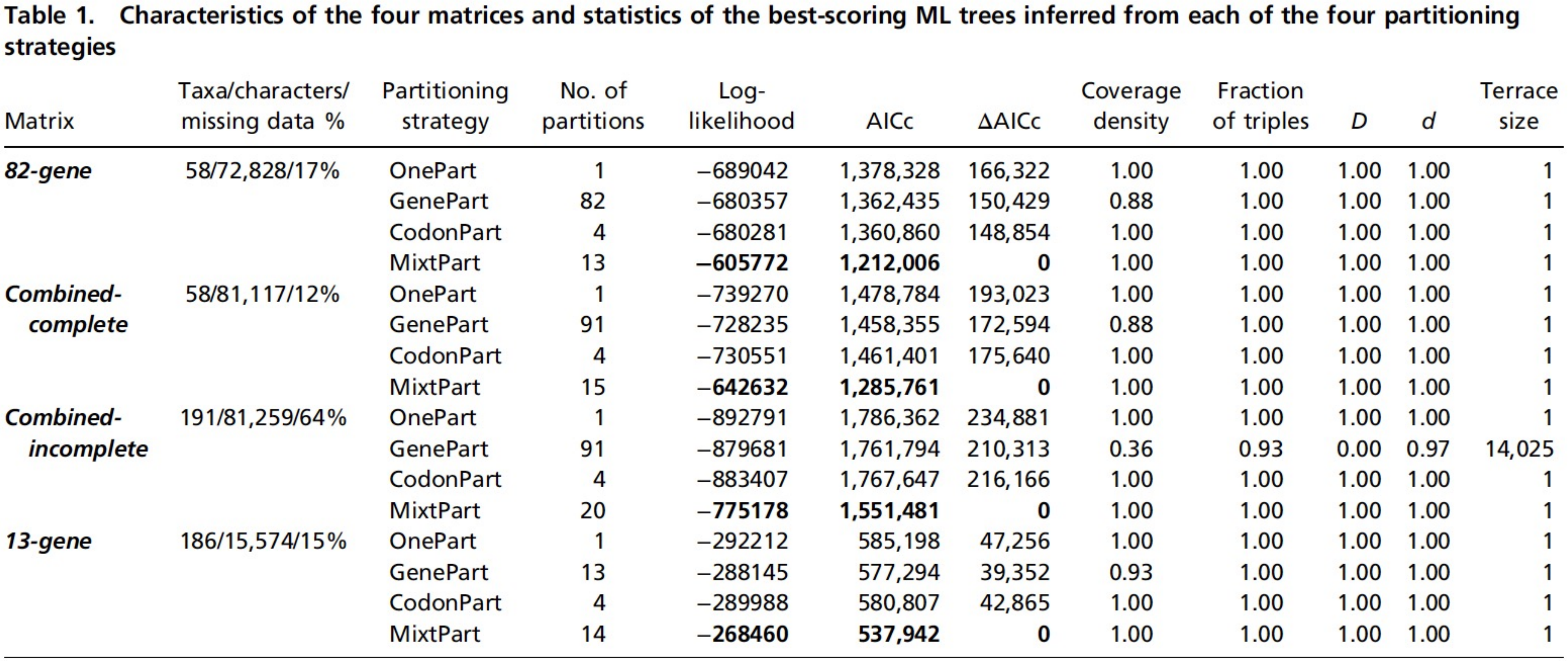


Table 1. Characteristics of the four matrices and statistics of the best-scoring ML trees inferred from each of the four partitioning strategies

Matrix	Taxa/characters/ missing data %	Partitioning strategy	No. of partitions	Log-likelihood	AICc	ΔAICc	Coverage density	Fraction of triples	*D*	*d*	Terrace size
82-gene	58/72,828/17%	OnePart	1	−689042	1,378,328	166,322	1.00	1.00	1.00	1.00	1
		GenePart	82	−680357	1,362,435	150,429	0.88	1.00	1.00	1.00	1
		CodonPart	4	−680281	1,360,860	148,854	1.00	1.00	1.00	1.00	1
		MixtPart	13	**−605772**	**1,212,006**	**0**	1.00	1.00	1.00	1.00	1
Combined-complete	58/81,117/12%	OnePart	1	−739270	1,478,784	193,023	1.00	1.00	1.00	1.00	1
		GenePart	91	−728235	1,458,355	172,594	0.88	1.00	1.00	1.00	1
		CodonPart	4	−730551	1,461,401	175,640	1.00	1.00	1.00	1.00	1
		MixtPart	15	**−642632**	**1,285,761**	**0**	1.00	1.00	1.00	1.00	1
Combined-incomplete	191/81,259/64%	OnePart	1	−892791	1,786,362	234,881	1.00	1.00	1.00	1.00	1
		GenePart	91	−879681	1,761,794	210,313	0.36	0.93	0.00	0.97	14,025
		CodonPart	4	−883407	1,767,647	216,166	1.00	1.00	1.00	1.00	1
		MixtPart	20	**−775178**	**1,551,481**	**0**	1.00	1.00	1.00	1.00	1
13-gene	186/15,574/15%	OnePart	1	−292212	585,198	47,256	1.00	1.00	1.00	1.00	1
		GenePart	13	−288145	577,294	39,352	0.93	1.00	1.00	1.00	1
		CodonPart	4	−289988	580,807	42,865	1.00	1.00	1.00	1.00	1
		MixtPart	14	**−268460**	**537,942**	**0**	1.00	1.00	1.00	1.00	1

Based on whole-chloroplast genomes

Xi et al. 2012

Application

But it can affect the boostrap support of some major clades!

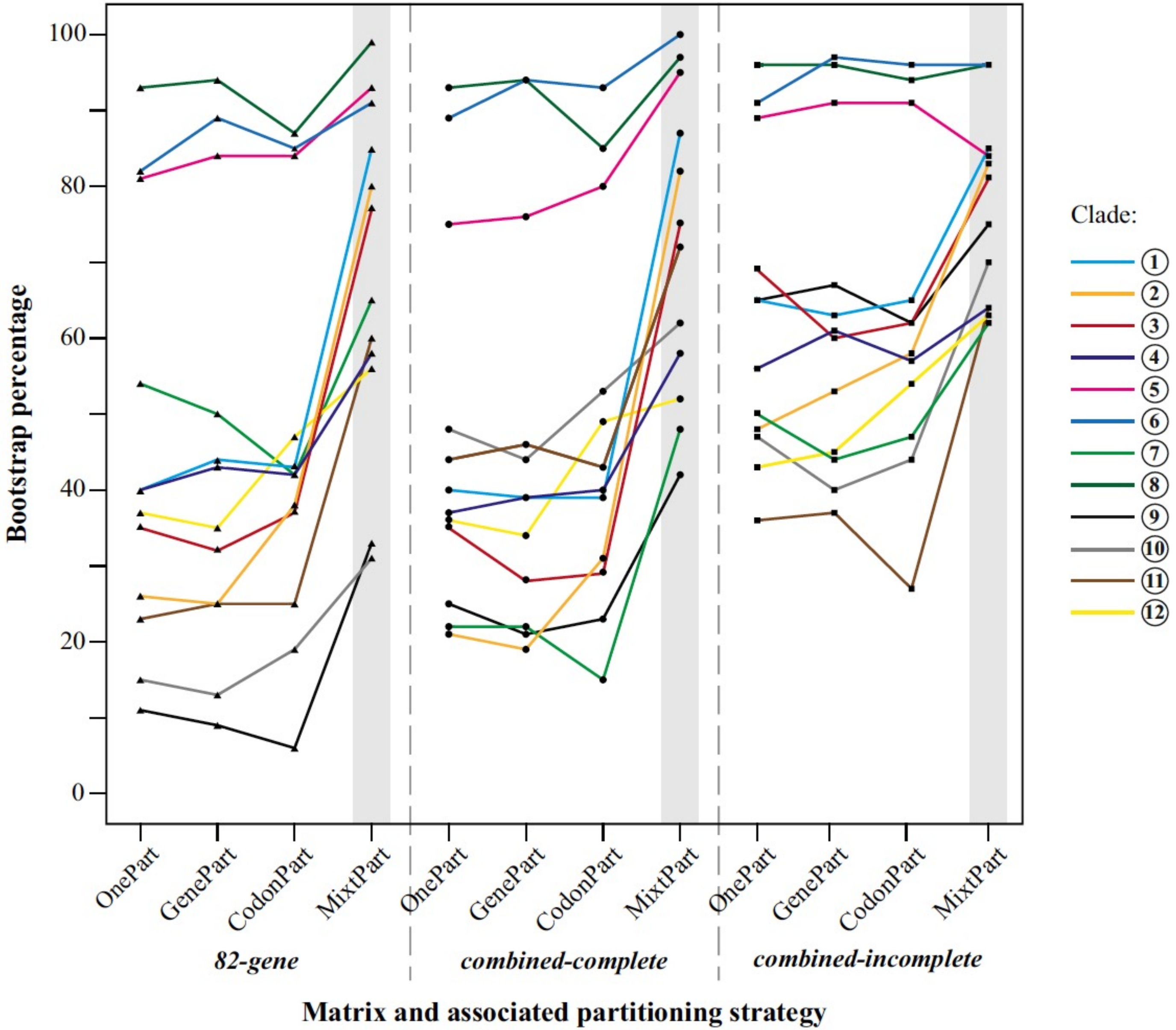


Fig. 3. ML BPs of the 12 additional clades we identified in Malpighiales (Fig. 1) inferred from three matrices and four partitioning strategies. The MixtPart partitioning strategy is highlighted in gray.

Xi et al. 2012

Application

But it can affect the bootstrap support of some major clades!

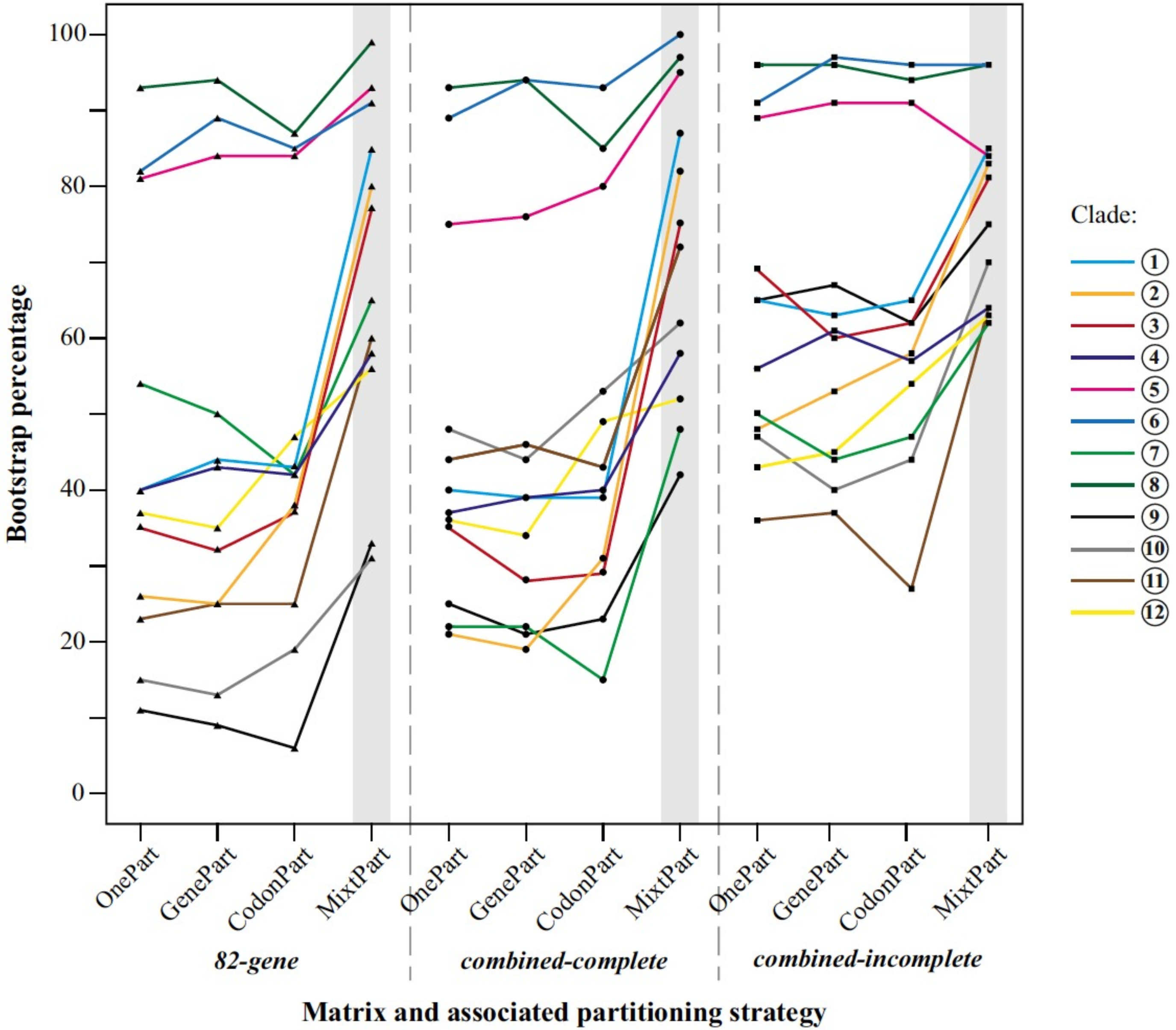
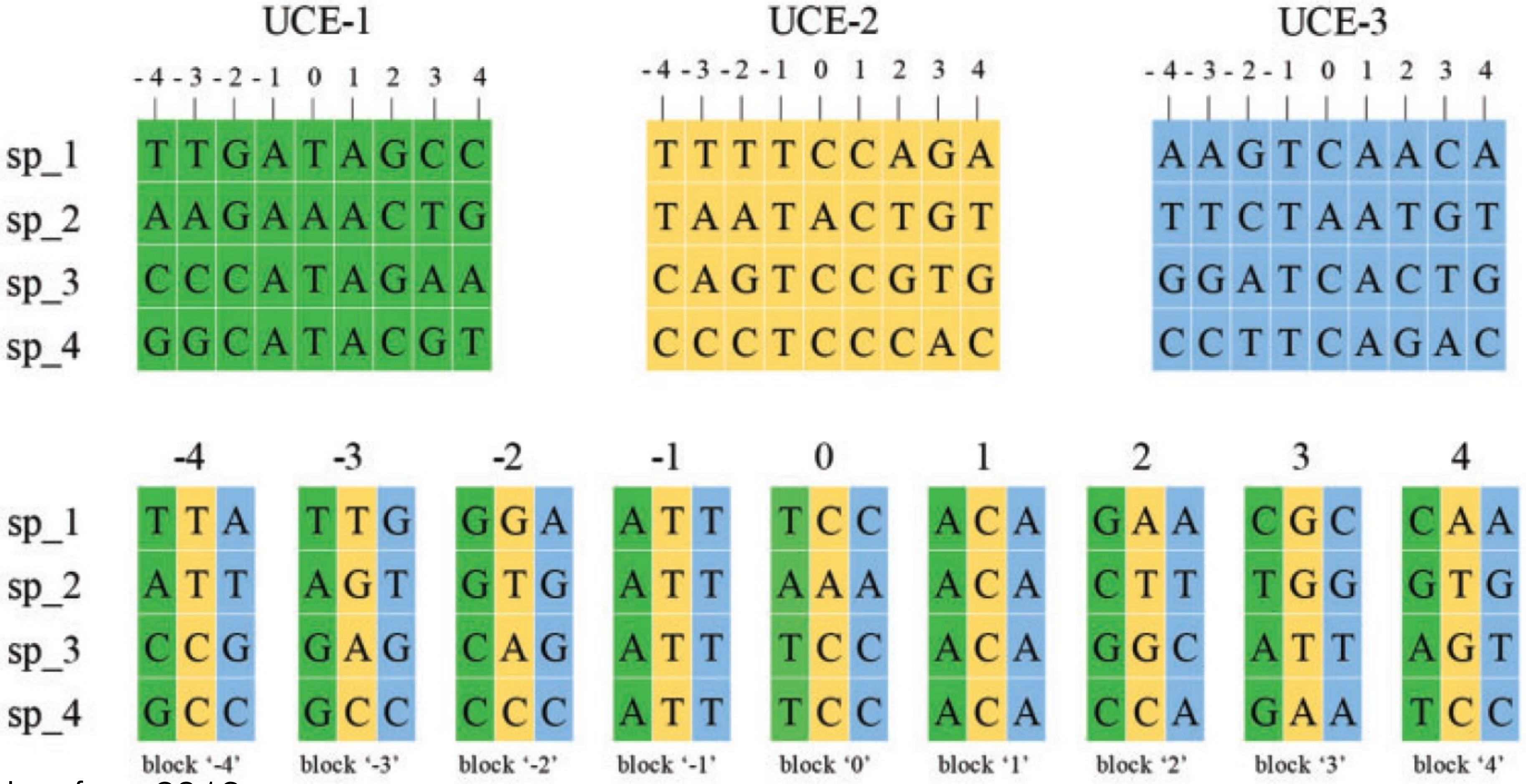


Fig. 3. ML BPs of the 12 additional clades we identified in Malpighiales (Fig. 1) inferred from three matrices and four partitioning strategies. The MixtPart partitioning strategy is highlighted in gray.

Xi et al. 2012

Application

Target enrichment data is conserved in the middle and variable on the ends



UCE-1

	-4	-3	-2	-1	0	1	2	3	4
sp_1	T	T	G	A	T	A	G	C	C
sp_2	A	A	G	A	A	A	C	T	G
sp_3	C	C	C	A	T	A	G	A	A
sp_4	G	G	C	A	T	A	C	G	T

UCE-2

	-4	-3	-2	-1	0	1	2	3	4
sp_1	T	T	T	T	C	C	A	G	A
sp_2	T	A	A	T	A	C	T	G	T
sp_3	C	A	G	T	C	C	G	T	G
sp_4	C	C	C	T	C	C	C	A	C

UCE-3

	-4	-3	-2	-1	0	1	2	3	4
sp_1	A	A	G	T	C	A	A	C	A
sp_2	T	T	C	T	A	A	T	G	T
sp_3	G	G	A	T	C	A	C	T	G
sp_4	C	C	T	T	C	A	G	A	C

	-4	-3	-2	-1	0	1	2	3	4
sp_1	TTA	TTG	GGA	ATT	TCC	ACA	GAA	CGC	CAA
sp_2	ATT	AGT	GTG	ATT	AAA	ACA	CTT	TGG	GTG
sp_3	CCG	GAG	CAG	ATT	TCC	ACA	GGC	ATT	AGT
sp_4	GCC	GCC	CCC	ATT	TCC	ACA	CCA	GAA	TCC
	block ‘-4’	block ‘-3’	block ‘-2’	block ‘-1’	block ‘0’	block ‘1’	block ‘2’	block ‘3’	block ‘4’

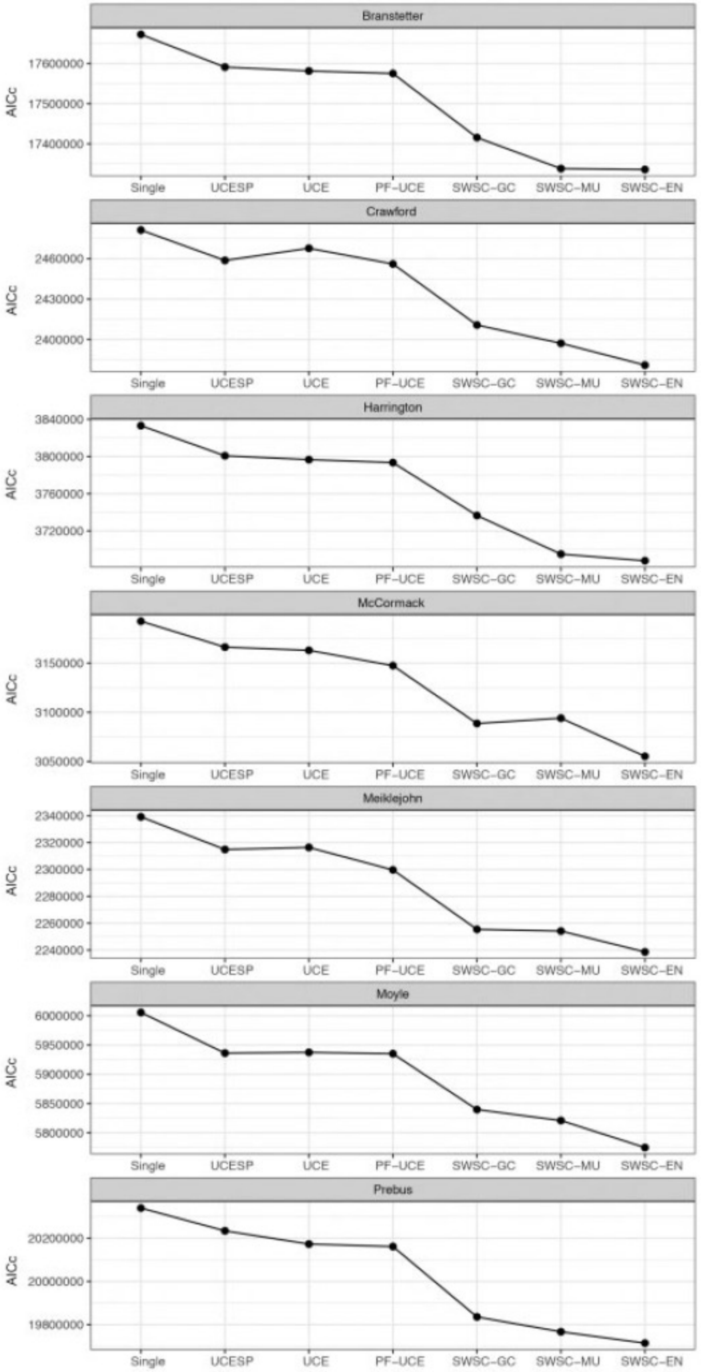
Tagliacollo and Lanfear 2018

Application

Target enrichment data is conserved in the middle and variable on the ends

More sophisticated partitioning can lead to much better likelihood scores

Not clear if it will cause a different biological interpretation of the results though



Tagliacollo and Lanfear 2018

Application

Modern software now automate this model selection process for you!

This includes selecting among different types of substitution models and potentially different partitioning strategies.

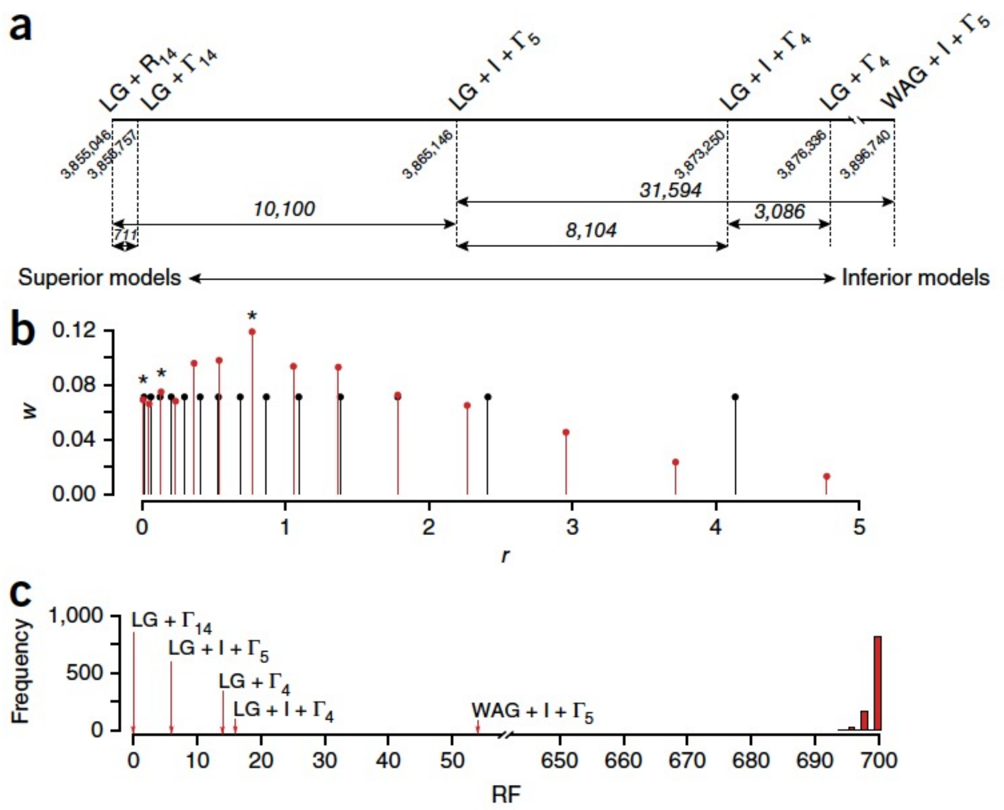


Figure 2 | Advantages provided by ModelFinder. (**a**) BIC scores of selected models of SE, given the alignment of bacterial and archaeal amino acids used by Wu *et al.*[19]. Models are listed above the thick horizontal line. Numbers along the line are BIC scores, and those in italics denote ΔBIC. (**b**) r_i and w_i values obtained under the R_{14} model of RHAS (red lines and balls) and the Γ_{14} model of RHAS (black lines and balls) for the alignment analyzed by Wu *et al.*[19]. Stars indicate local peaks in the R_{14} model. (**c**) RF distances between the most likely tree inferred under various models of SE. For comparison, a histogram with the distribution of 1,000 RF distances is included; each of these distances was obtained by comparing the most likely tree inferred under the LG + R_{14} model of SE to a randomly generated tree with the same number of leaves.

Kalyaanamoorthy et al. 2017

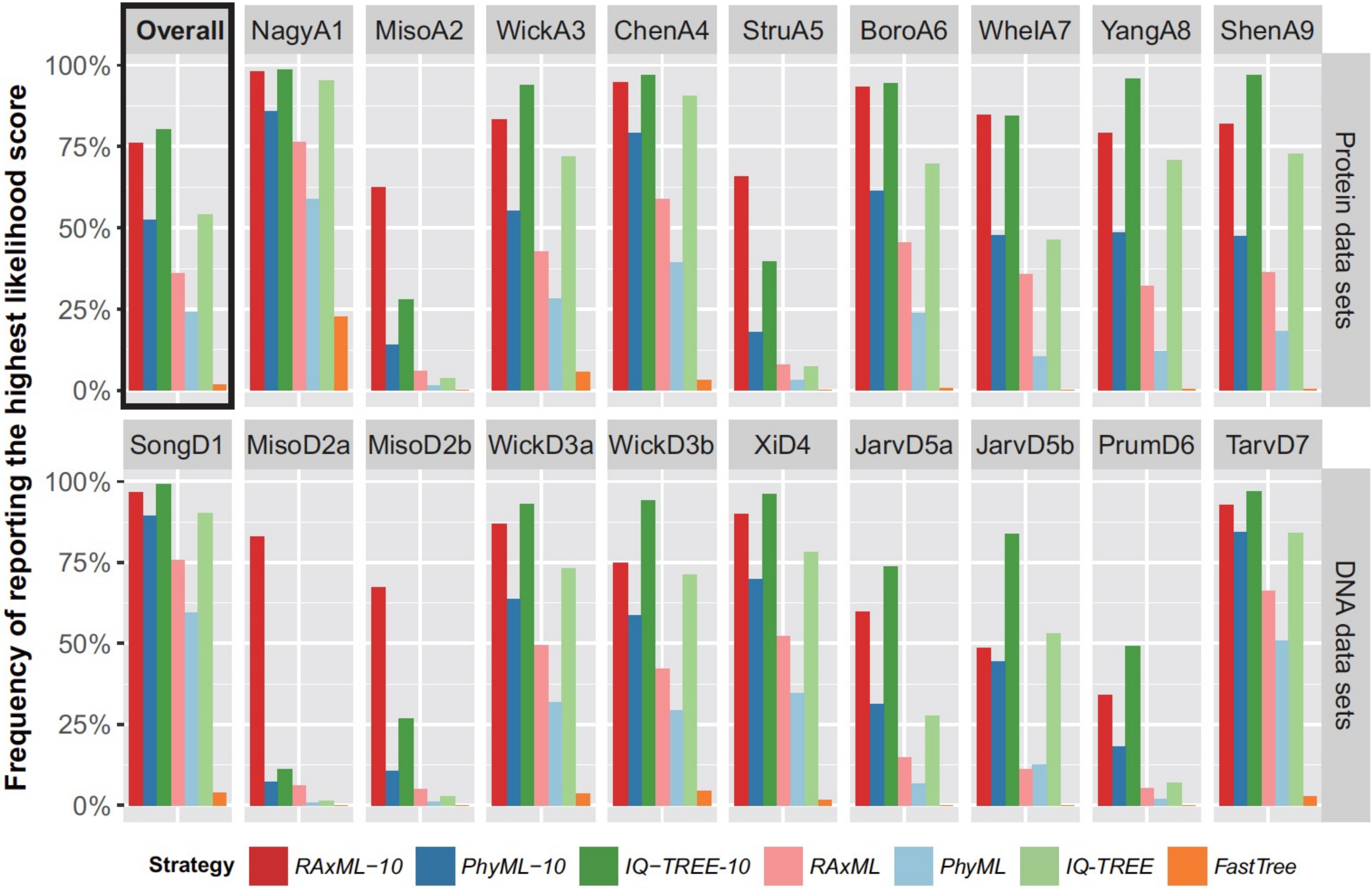
Application

IQTREE2 (Minh et al. 2020)

There are many software packages to choose from, why will we use IQTREE today?

Fast
Automated model selection
Well-annotated log files
Many nice features when analyzing many loci
Returns best likelihoods most frequently

Application



Zhou et al. 2018

End

When we come back we will analyze some data with IQTREE