Model Selection

Molecular Phylogenetics Course 2019

Kevin Gori

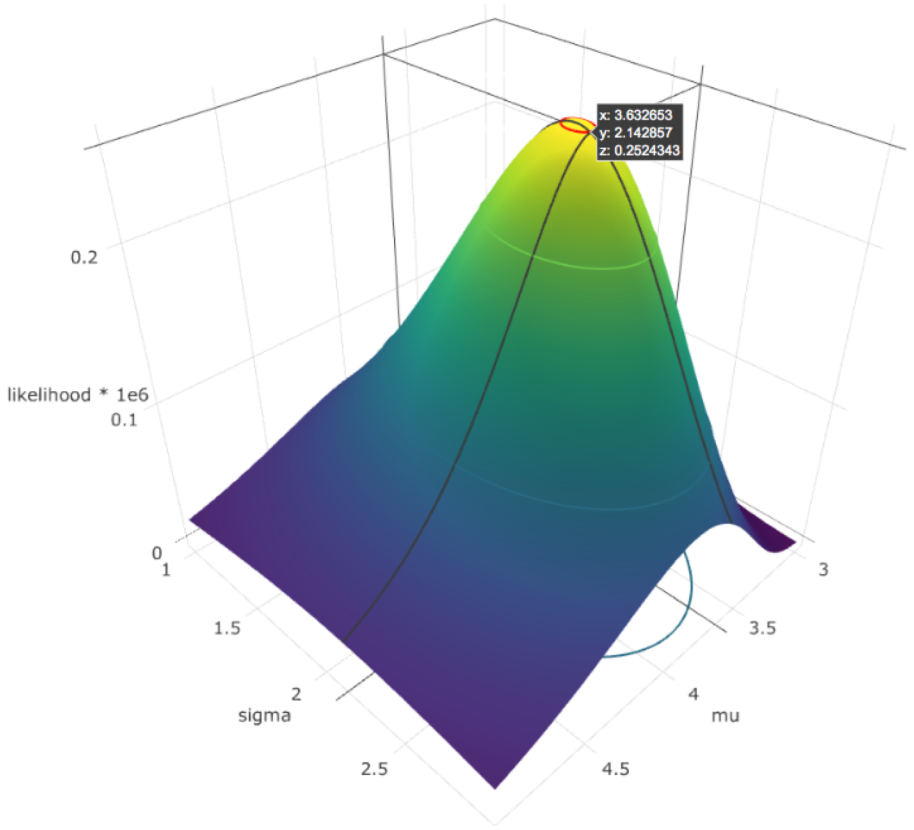
kcg25@cam.ac.uk

Outline

- Recap Likelihood vs. Probability
- Recap Phylogenetic Likelihood
  - substitution models
  - pruning algorithm
  - rate variation among sites
  - ascertainment bias correction
- Model selection
  - bias-variance trade-off
  - Likelihood ratio tests
  - Information Criteria
- Practicals

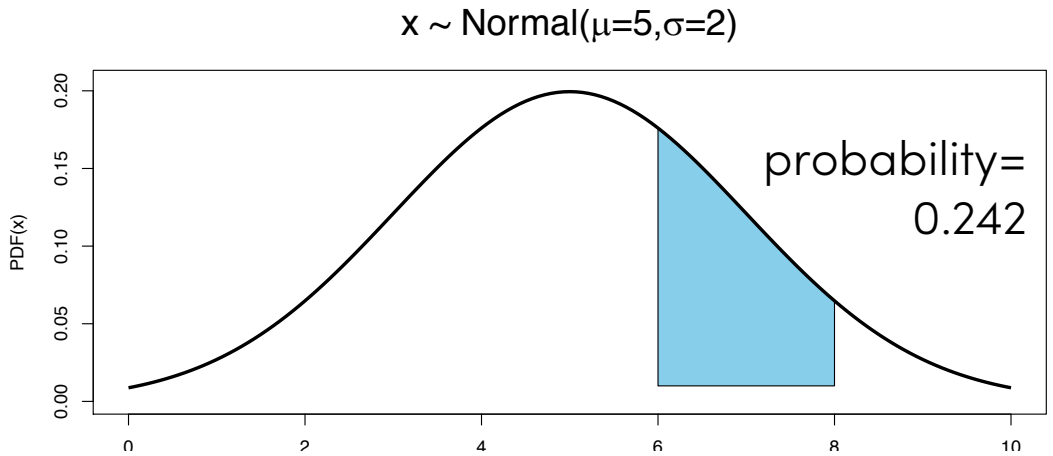
Likelihood

Recap

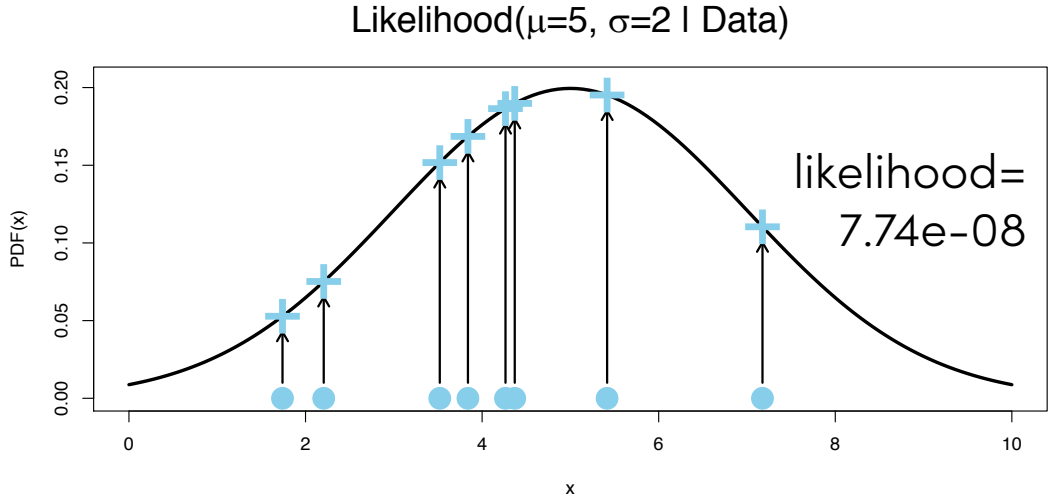


Probability and Likelihood

- Probability distribution:
  - With what probability do we expect to see new data with a particular value (range)

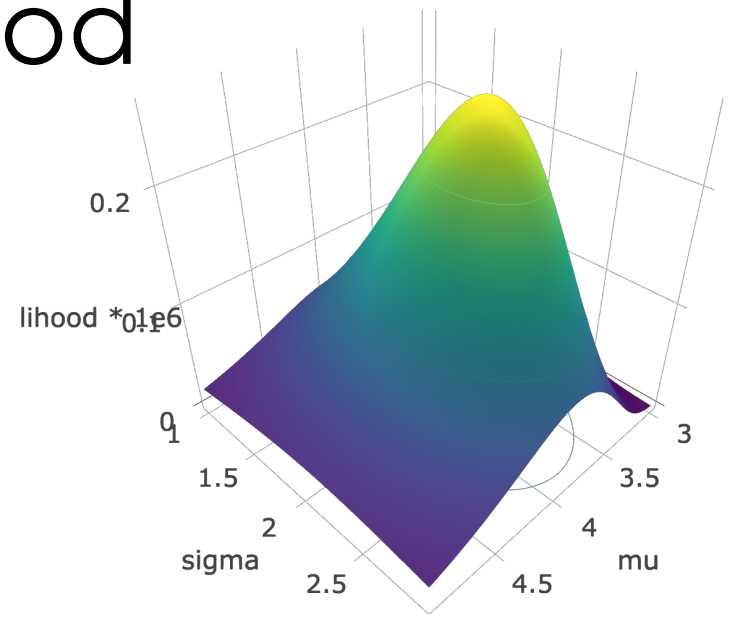


- Likelihood:
  - For given data, how likely is it that it was generated by a distribution with these parameters

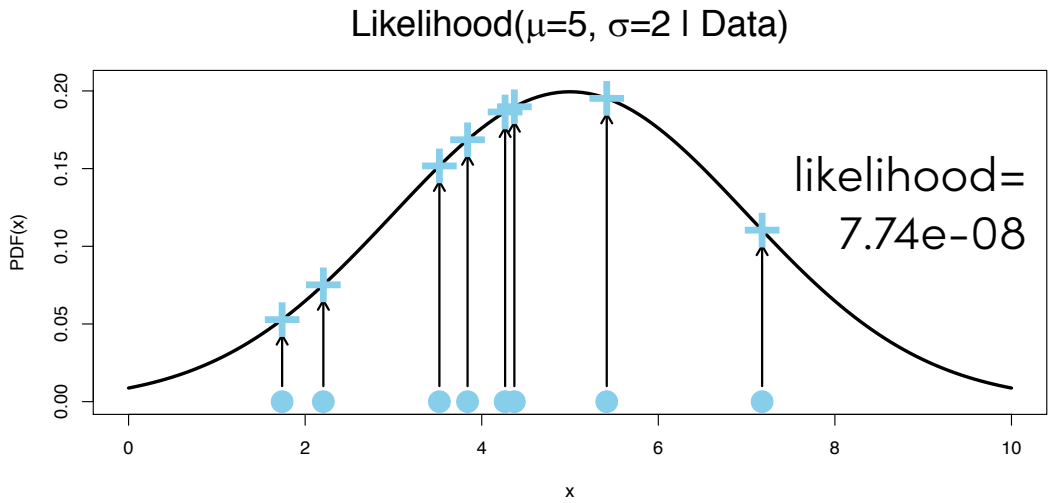


Probability and Likelihood

- Probability distribution:
  - With what probability do we expect to see new data with a particular value (range)

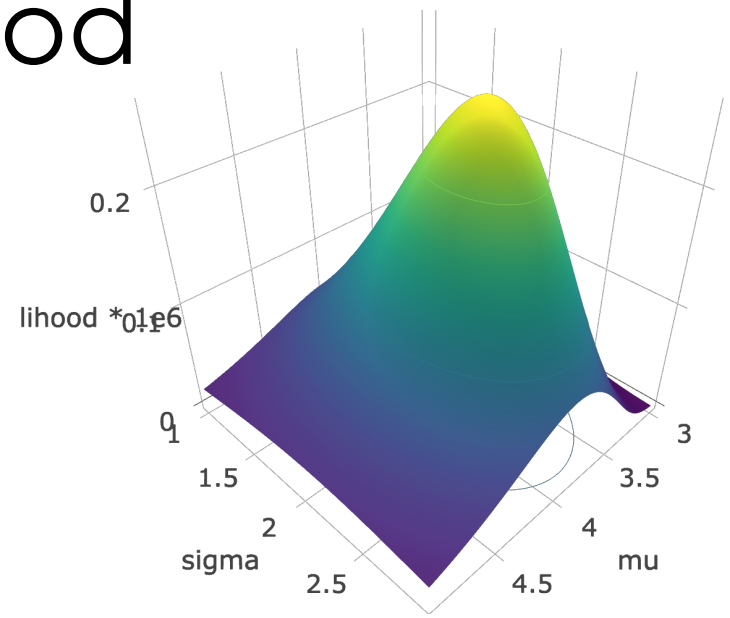


- Likelihood:
  - For given data, how likely is it that it was generated by a distribution with these parameters

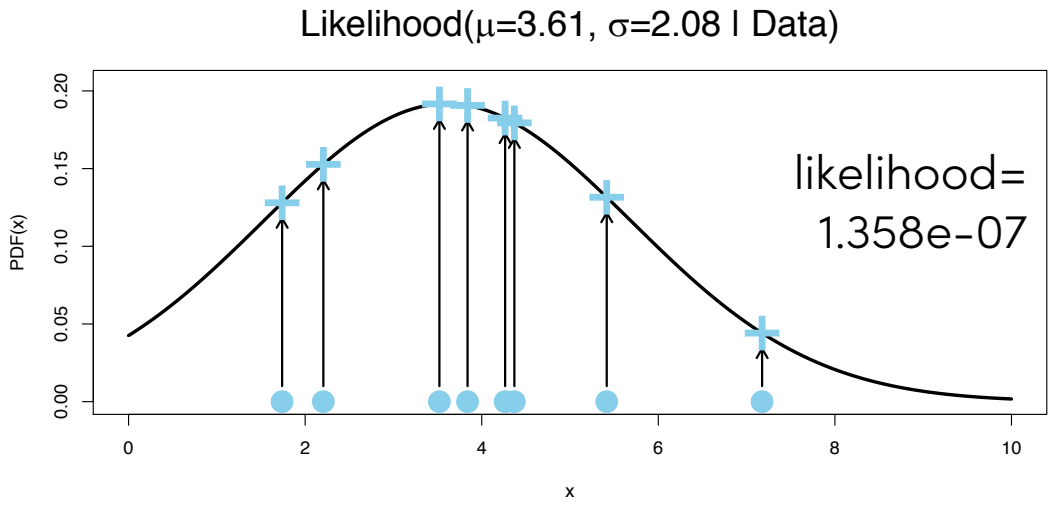


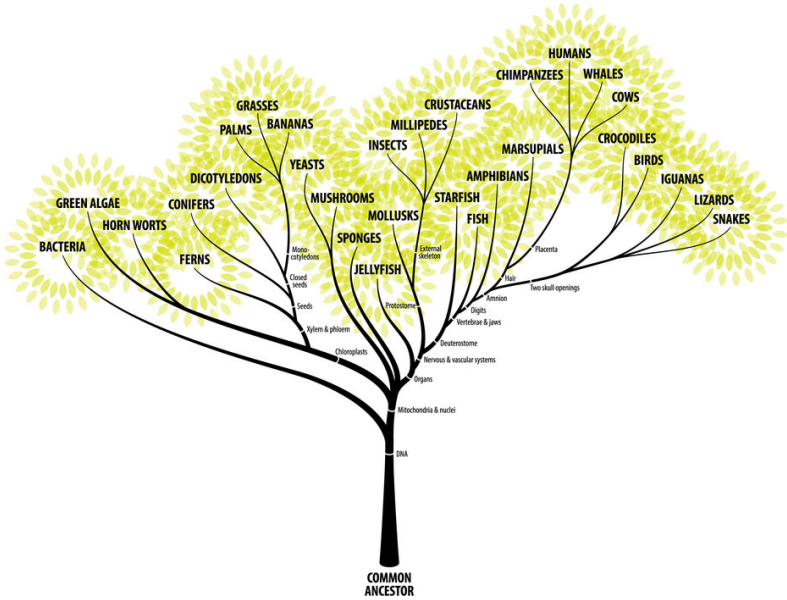
Probability and Likelihood

- Probability distribution:
  - With what probability do we expect to see new data with a particular value (range)



- Likelihood:
  - For given data, how likely is it that it was generated by a distribution with these parameters





Phylogenetic Likelihood

Substitution Models

- Model substitution process as a Continuous-time Markov Chain
- Defined by [n×n] Q-matrix
  - n is alphabet size, i.e. DNA: n=4

Substitution Models

Q_{JC69} Jukes and Cantor

	T	C	A	G
T	-3λ	λ	λ	λ
C	λ	-3λ	λ	λ
A	λ	λ	-3λ	λ
G	λ	λ	λ	-3λ

Diagonal entry ensures rows sum to 0

- Model substitution process as a Continuous-time Markov Chain
- Defined by [n×n] Q-matrix
  - n is alphabet size, i.e. DNA: n=4
- Probability of substitution after time t with rate r:

$$P_{ij} = \exp(Q \times t \times r)_{ij}$$

Substitution Models

Q_{K80} Kimura's ts/tv model

	T	C	A	G
T	-	α	β	β
C	α	-	β	β
A	β	β	-	α
G	β	β	α	-

Diagonal entry ensures rows sum to 0

- Model substitution process as a Continuous-time Markov Chain
- Defined by [n×n] Q-matrix
  - n is alphabet size, i.e. DNA: n=4
- Probability of substitution after time t with rate r:

$$P_{ij} = \exp(Q \times t \times r)_{ij}$$

Substitution Models

Q_{F81} Felsenstein's unequal base frequencies

	T	C	A	G
T	-	π_C	π_A	π_G
C	π_T	-	π_A	π_G
A	π_T	π_C	-	π_G
G	π_T	π_C	π_A	-

Diagonal entry ensures rows sum to 0

- Model substitution process as a Continuous-time Markov Chain
- Defined by [n×n] Q-matrix
  - n is alphabet size, i.e. DNA: n=4
- Probability of substitution after time t with rate r:

$$P_{ij} = \exp(Q \times t \times r)_{ij}$$

Substitution Models

Q_{HKY85} Hasegawa, *et al.*

	T	C	A	G
T	-	$\pi_C\alpha$	$\pi_A\beta$	$\pi_G\beta$
C	$\pi_T\alpha$	-	$\pi_A\beta$	$\pi_G\beta$
A	$\pi_T\beta$	$\pi_C\beta$	-	$\pi_G\alpha$
G	$\pi_T\beta$	$\pi_C\beta$	$\pi_A\alpha$	-

Diagonal entry ensures rows sum to 0

- Model substitution process as a Continuous-time Markov Chain
- Defined by [n×n] Q-matrix
  - n is alphabet size, i.e. DNA: n=4
- Probability of substitution after time t with rate r:

$$P_{ij} = \exp(Q \times t \times r)_{ij}$$

Substitution Models

Q_{GTR} General time reversible model

	T	C	A	G
T	-	$\pi_C a$	$\pi_A b$	$\pi_G c$
C	$\pi_T a$	-	$\pi_A d$	$\pi_G e$
A	$\pi_T b$	$\pi_C d$	-	$\pi_G f$
G	$\pi_T c$	$\pi_C e$	$\pi_A f$	-

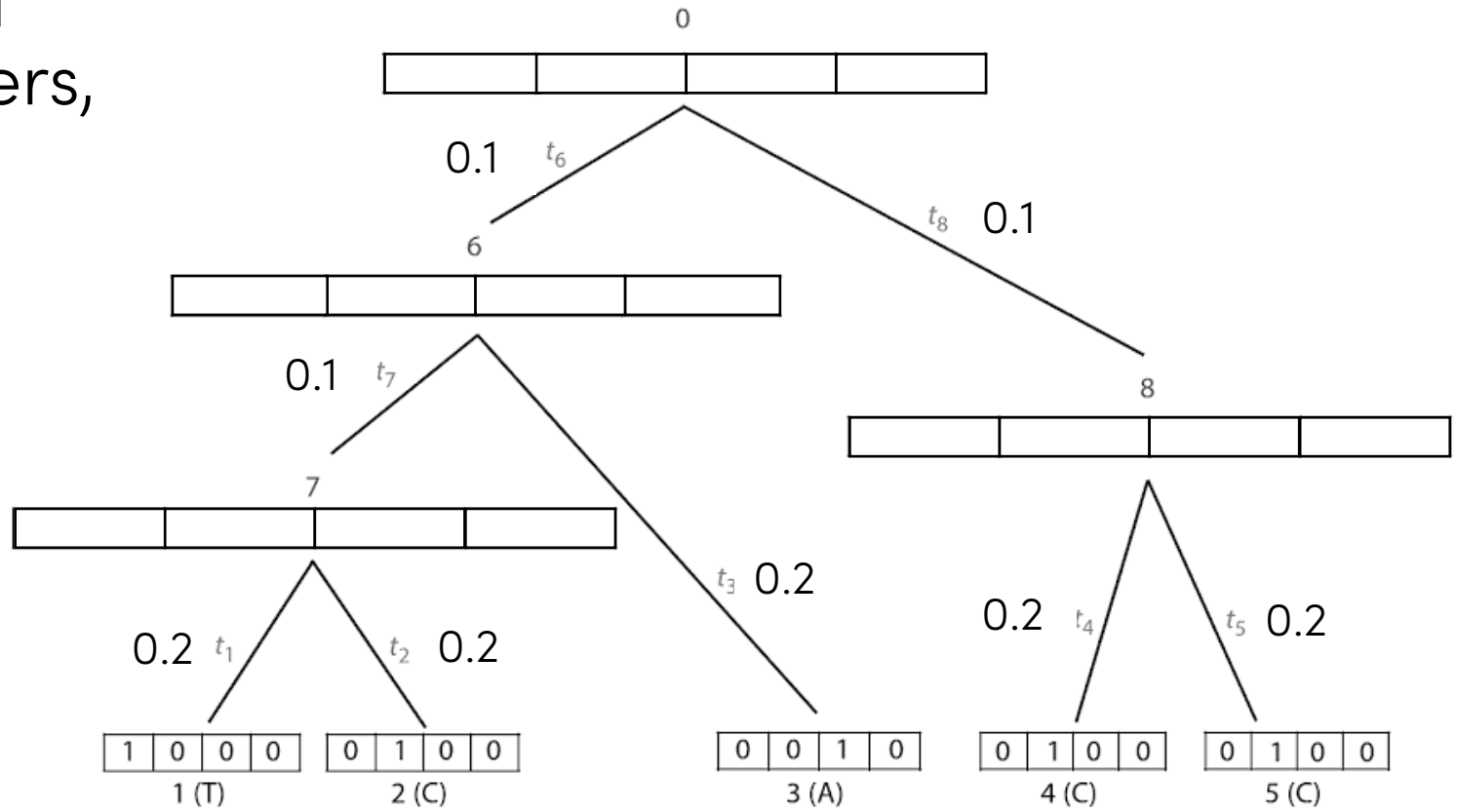
Diagonal entry ensures rows sum to 0

- Model substitution process as a Continuous-time Markov Chain
- Defined by [n×n] Q-matrix
  - n is alphabet size, i.e. DNA: n=4
- Probability of substitution after time t with rate r:

$$P_{ij} = \exp(Q \times t \times r)_{ij}$$

Pruning algorithm

- Computes likelihood of tree + model parameters, given sequence data
- Initialise conditional probabilities at leaves:
  - $Prob_i = 1$ if i = base
    0 otherwise

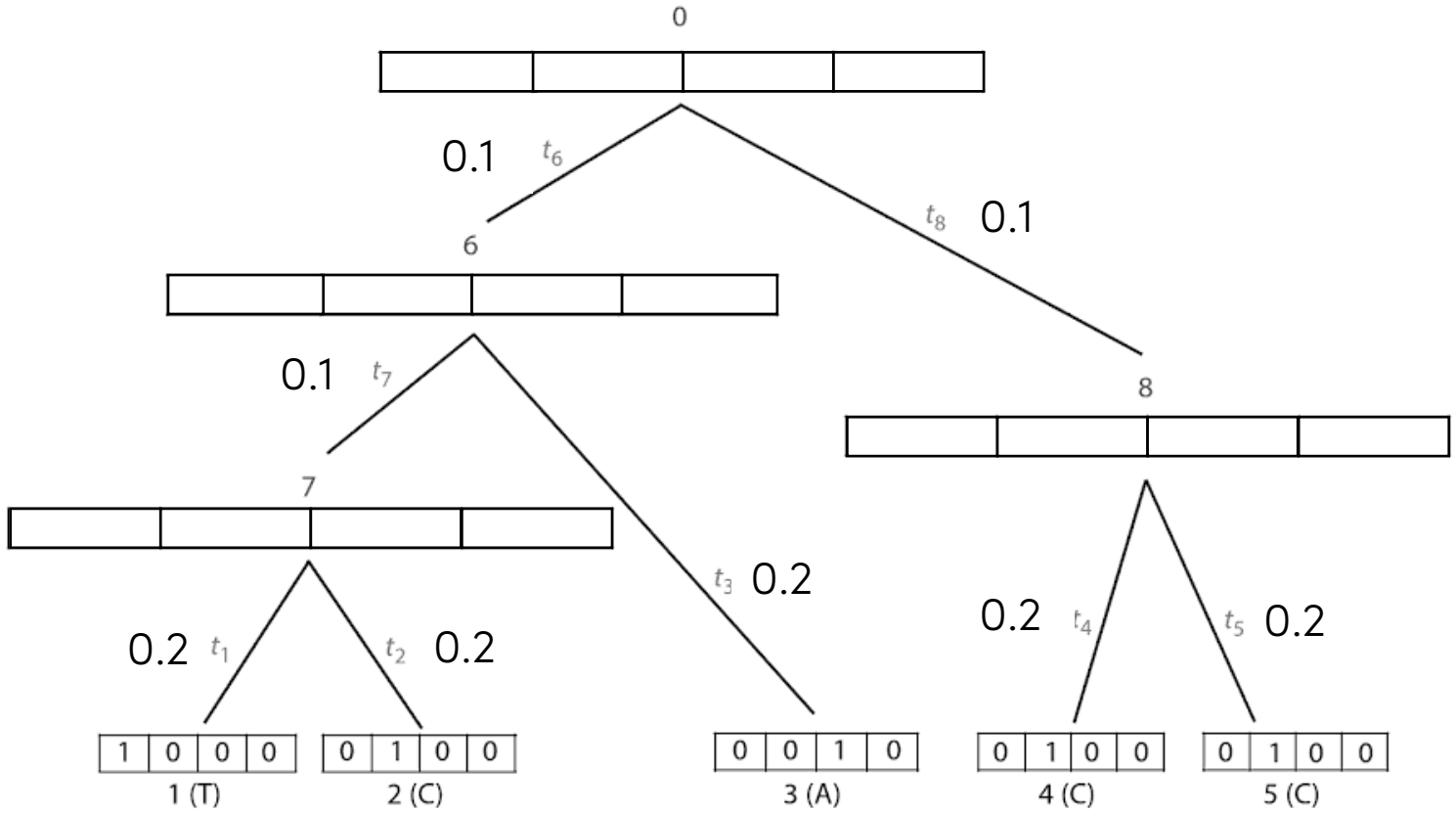


Yang, 2014

Pruning algorithm

- Postorder traversal:
  - Visit descendents before ancestors
- Inner conditionals:

  $[P(t_a) \times L(a)] \otimes [P(t_b) \times L(b)]$



× : Matrix multiplication (dot product)

⊗: Elementwise multiplication (Hadamard product)

Yang, 2014

Pruning algorithm

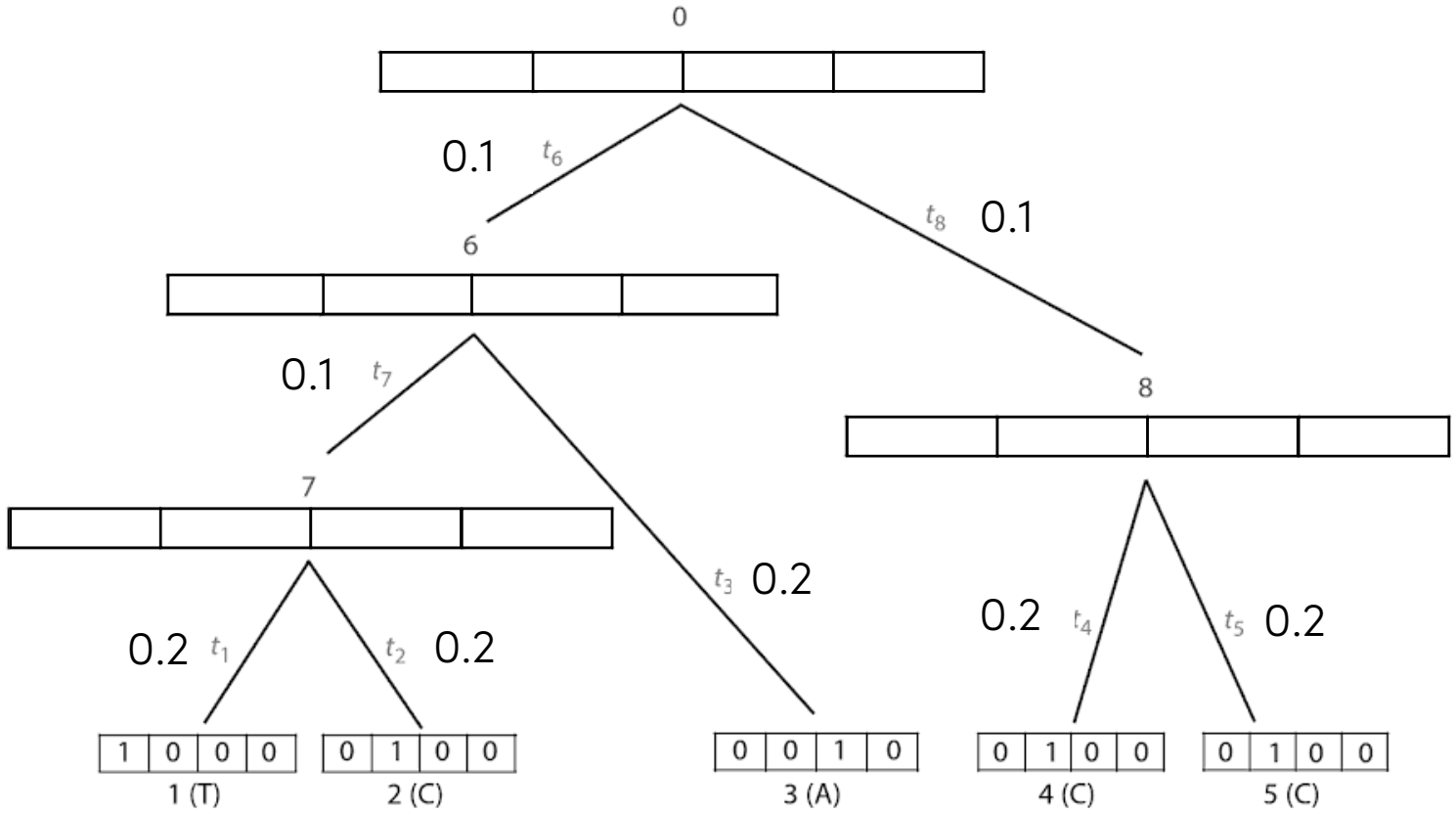
- Postorder traversal:
  - Visit descendents before ancestors
- Inner conditionals:

$[P(t_a) \times L(a)] \otimes [P(t_b) \times L(b)]$

Model = K80(α:2; β:1)

$$P(0.1) = \begin{bmatrix} 0.906563 & 0.045855 & 0.023791 & 0.023791 \\ 0.045855 & 0.906563 & 0.023791 & 0.023791 \\ 0.023791 & 0.023791 & 0.906563 & 0.045855 \\ 0.023791 & 0.023791 & 0.045855 & 0.906563 \end{bmatrix}$$

$$P(0.2) = \begin{bmatrix} 0.825092 & 0.084274 & 0.045317 & 0.045317 \\ 0.084274 & 0.825092 & 0.045317 & 0.045317 \\ 0.045317 & 0.045317 & 0.825092 & 0.084274 \\ 0.045317 & 0.045317 & 0.084274 & 0.825092 \end{bmatrix}$$



× : Matrix multiplication (dot product)

⊗: Elementwise multiplication (Hadamard product)

Yang, 2014

Pruning algorithm

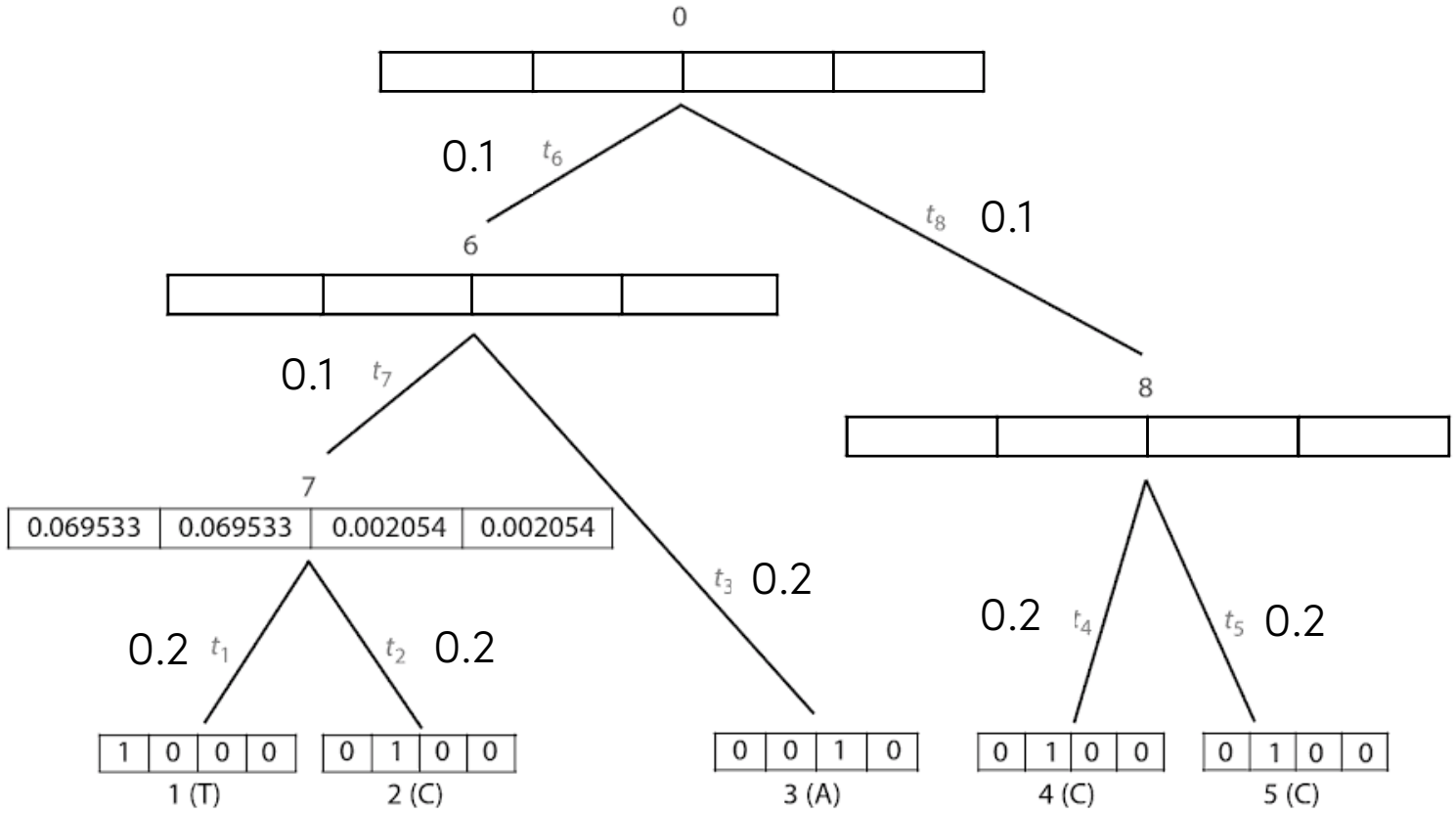
- Postorder traversal:
  - Visit descendents before ancestors
- Inner conditionals:

$[P(t_a) \times L(a)] \otimes [P(t_b) \times L(b)]$

Model = K80(α:2; β:1)

$$P(0.1) = \begin{bmatrix} 0.906563 & 0.045855 & 0.023791 & 0.023791 \\ 0.045855 & 0.906563 & 0.023791 & 0.023791 \\ 0.023791 & 0.023791 & 0.906563 & 0.045855 \\ 0.023791 & 0.023791 & 0.045855 & 0.906563 \end{bmatrix}$$

$$P(0.2) = \begin{bmatrix} 0.825092 & 0.084274 & 0.045317 & 0.045317 \\ 0.084274 & 0.825092 & 0.045317 & 0.045317 \\ 0.045317 & 0.045317 & 0.825092 & 0.084274 \\ 0.045317 & 0.045317 & 0.084274 & 0.825092 \end{bmatrix}$$



× : Matrix multiplication (dot product)

⊗: Elementwise multiplication (Hadamard product)

Yang, 2014

Pruning algorithm

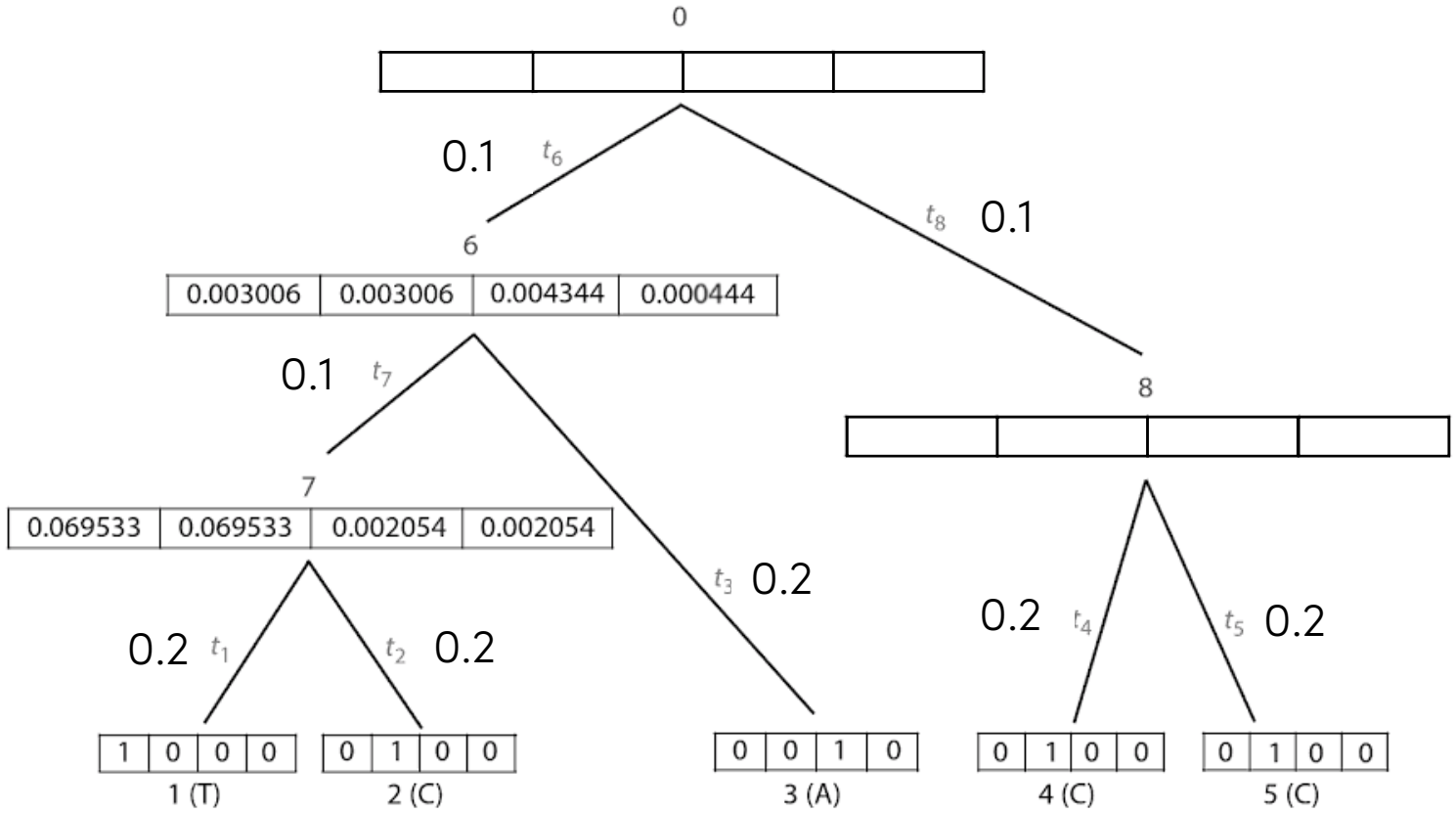
- Postorder traversal:
  - Visit descendents before ancestors
- Inner conditionals:

$[P(t_a) \times L(a)] \otimes [P(t_b) \times L(b)]$

Model = K80(α:2; β:1)

$$P(0.1) = \begin{bmatrix} 0.906563 & 0.045855 & 0.023791 & 0.023791 \\ 0.045855 & 0.906563 & 0.023791 & 0.023791 \\ 0.023791 & 0.023791 & 0.906563 & 0.045855 \\ 0.023791 & 0.023791 & 0.045855 & 0.906563 \end{bmatrix}$$

$$P(0.2) = \begin{bmatrix} 0.825092 & 0.084274 & 0.045317 & 0.045317 \\ 0.084274 & 0.825092 & 0.045317 & 0.045317 \\ 0.045317 & 0.045317 & 0.825092 & 0.084274 \\ 0.045317 & 0.045317 & 0.084274 & 0.825092 \end{bmatrix}$$



× : Matrix multiplication (dot product)

⊗: Elementwise multiplication (Hadamard product)

Yang, 2014

Pruning algorithm

- Postorder traversal:
  - Visit descendents before ancestors

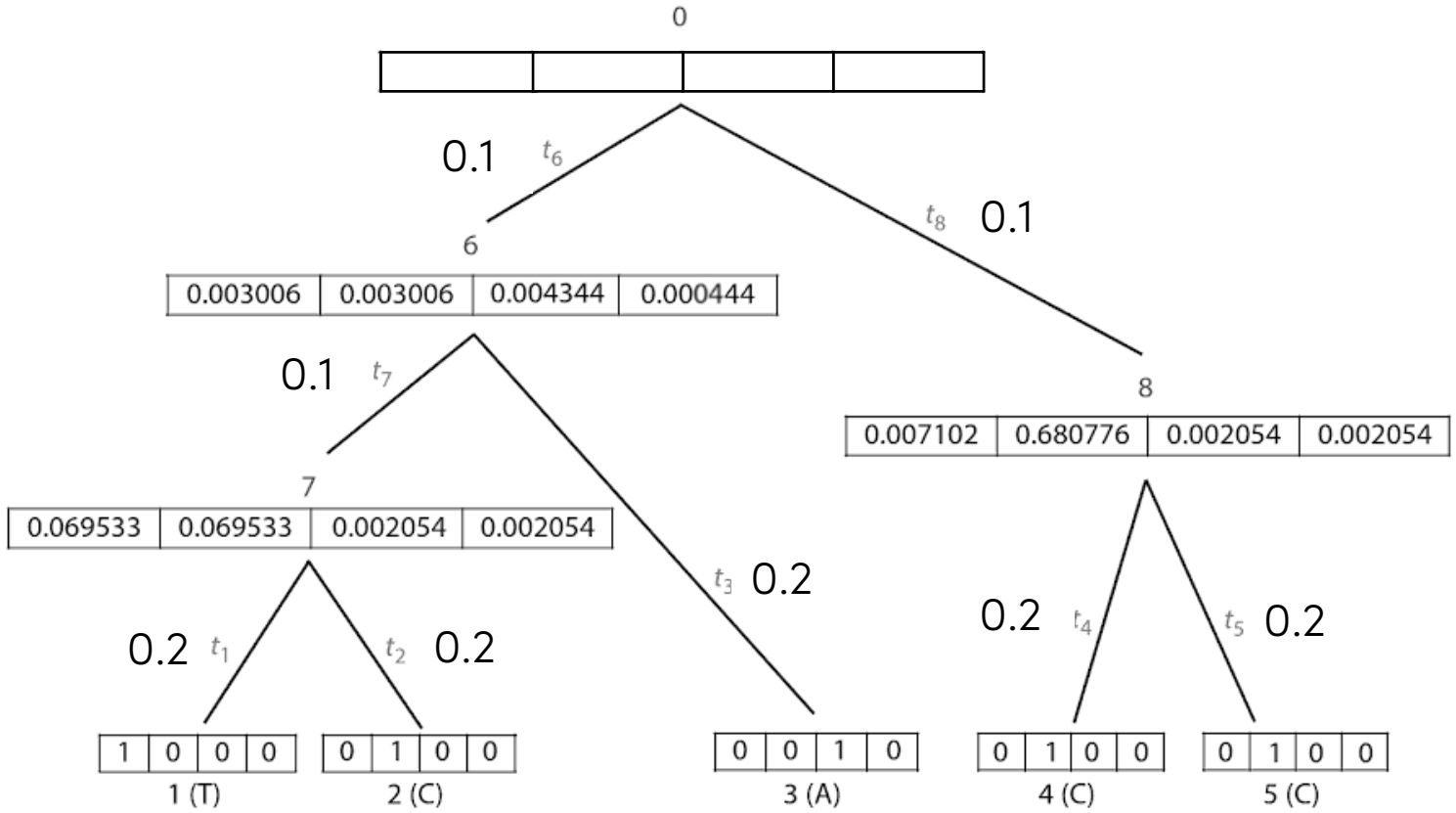
- Inner conditionals:

$[P(t_a) \times L(a)] \otimes [P(t_b) \times L(b)]$

Model = K80(α:2; β:1)

$$P(0.1) = \begin{bmatrix} 0.906563 & 0.045855 & 0.023791 & 0.023791 \\ 0.045855 & 0.906563 & 0.023791 & 0.023791 \\ 0.023791 & 0.023791 & 0.906563 & 0.045855 \\ 0.023791 & 0.023791 & 0.045855 & 0.906563 \end{bmatrix}$$

$$P(0.2) = \begin{bmatrix} 0.825092 & 0.084274 & 0.045317 & 0.045317 \\ 0.084274 & 0.825092 & 0.045317 & 0.045317 \\ 0.045317 & 0.045317 & 0.825092 & 0.084274 \\ 0.045317 & 0.045317 & 0.084274 & 0.825092 \end{bmatrix}$$



× : Matrix multiplication (dot product)

⊗: Elementwise multiplication (Hadamard product)

Yang, 2014

Pruning algorithm

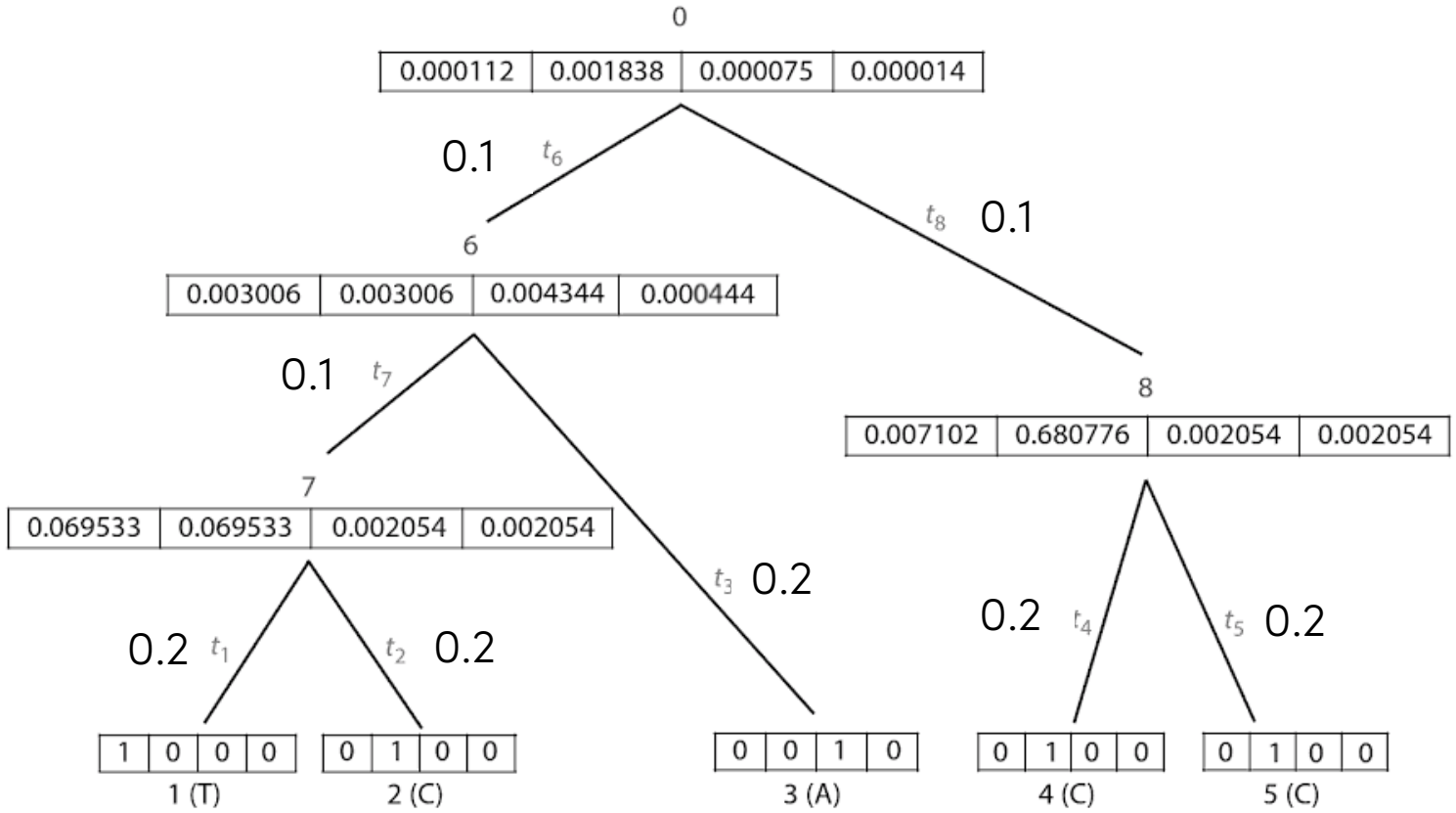
- Postorder traversal:
  - Visit descendents before ancestors
- Inner conditionals:

$[P(t_a) \times L(a)] \otimes [P(t_b) \times L(b)]$

Model = K80(α:2; β:1)

$$P(0.1) = \begin{bmatrix} 0.906563 & 0.045855 & 0.023791 & 0.023791 \\ 0.045855 & 0.906563 & 0.023791 & 0.023791 \\ 0.023791 & 0.023791 & 0.906563 & 0.045855 \\ 0.023791 & 0.023791 & 0.045855 & 0.906563 \end{bmatrix}$$

$$P(0.2) = \begin{bmatrix} 0.825092 & 0.084274 & 0.045317 & 0.045317 \\ 0.084274 & 0.825092 & 0.045317 & 0.045317 \\ 0.045317 & 0.045317 & 0.825092 & 0.084274 \\ 0.045317 & 0.045317 & 0.084274 & 0.825092 \end{bmatrix}$$



× : Matrix multiplication (dot product)

⊗: Elementwise multiplication (Hadamard product)

Yang, 2014

Pruning algorithm

- Postorder traversal:
  - Visit descendents before ancestors

- Inner conditionals:

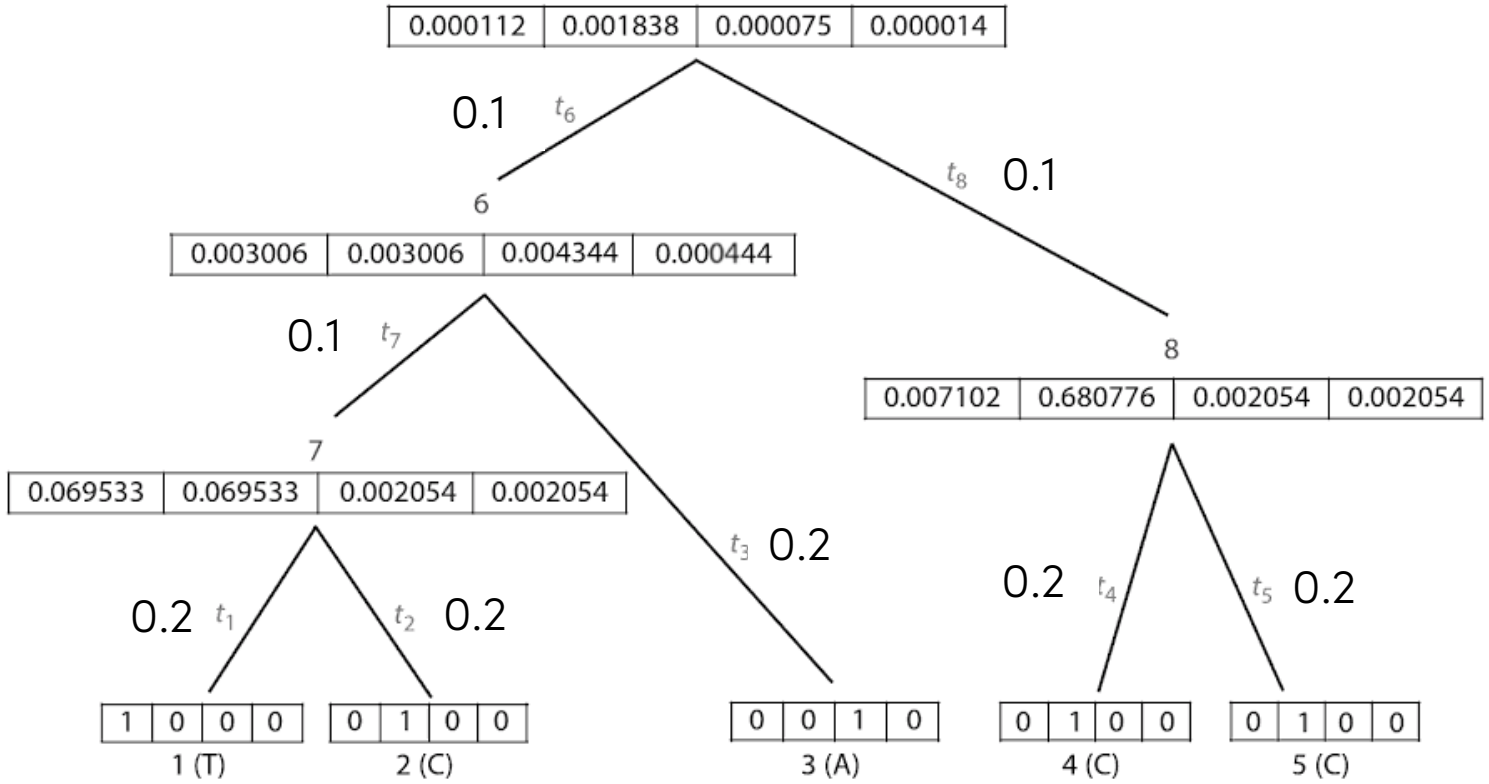
  $[P(t_a) \times L(a)] \otimes [P(t_b) \times L(b)]$

Model = K80(α:2; β:1)

$$P(0.1) = \begin{bmatrix} 0.906563 & 0.045855 & 0.023791 & 0.023791 \\ 0.045855 & 0.906563 & 0.023791 & 0.023791 \\ 0.023791 & 0.023791 & 0.906563 & 0.045855 \\ 0.023791 & 0.023791 & 0.045855 & 0.906563 \end{bmatrix}$$

$$P(0.2) = \begin{bmatrix} 0.825092 & 0.084274 & 0.045317 & 0.045317 \\ 0.084274 & 0.825092 & 0.045317 & 0.045317 \\ 0.045317 & 0.045317 & 0.825092 & 0.084274 \\ 0.045317 & 0.045317 & 0.084274 & 0.825092 \end{bmatrix}$$

logLik = log(π × L(0)) = -7.582



× : Matrix multiplication (dot product)

⊗: Elementwise multiplication (Hadamard product)

Yang, 2014

Models of Rate Heterogeneity

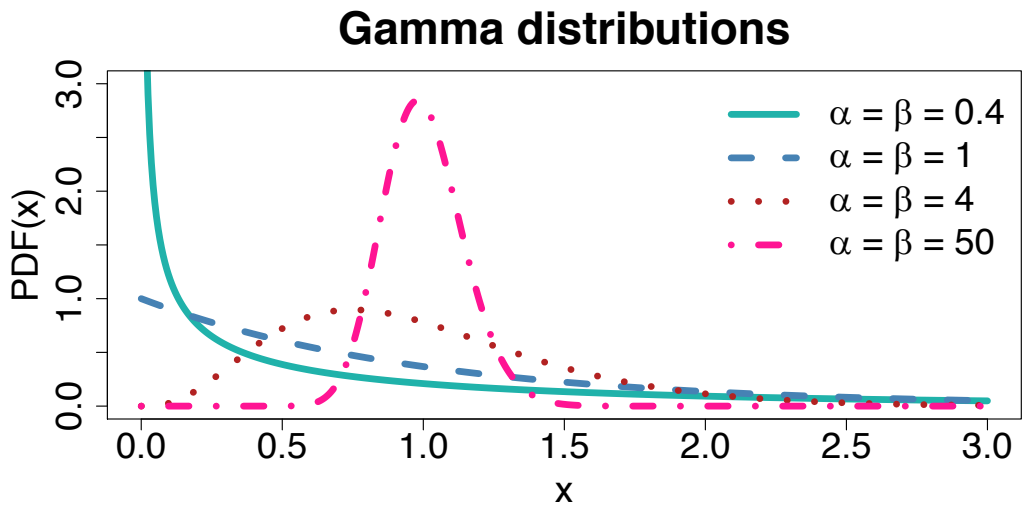
Sites do not all evolve at the same rate

- Selective constraints
- Chemical properties (e.g. CG → TG at methylated CpG)

Models of Rate Heterogeneity

- Gamma Distributed Rates (+G)

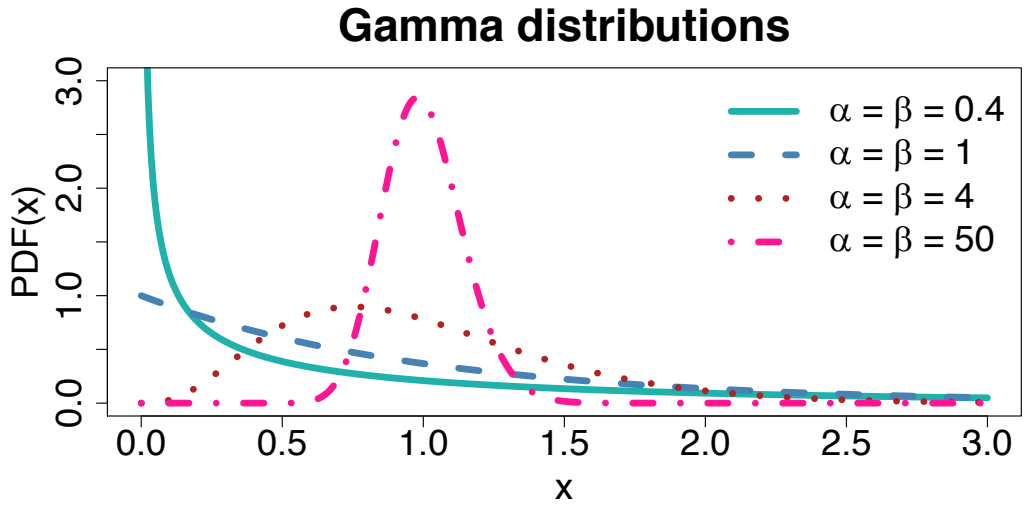
- Method
  - Compute tree likelihood for *n* discrete rate categories (e.g. *n*=4)
  - Use Pij = $\exp(Q \times t \times r_n)$
  - Likelihood at root is weighted average of *n* likelihoods

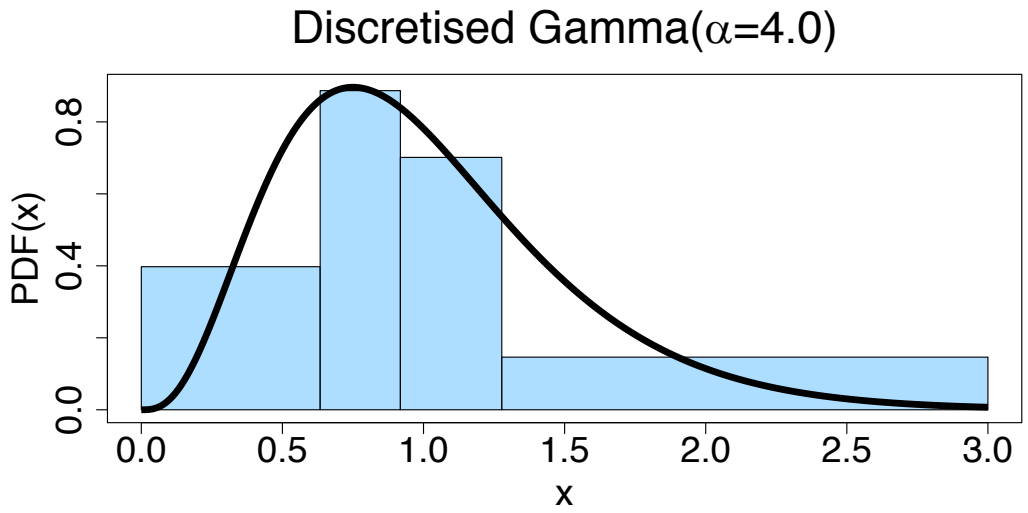


Models of Rate Heterogeneity

- Gamma Distributed Rates (+G)

- Method
  - Compute tree likelihood for *n* discrete rate categories (e.g. *n*=4)
  - Use Pij = exp(Q×t×r_n)
  - Likelihood at root is weighted average of *n* likelihoods





Models of Rate Heterogeneity

- Invariant Sites Model (+I)
- Estimate proportion of invariant sites: p_0
- Two rate categories: $r_0 = 0$, $r_1 = 1 / (1 - p_0)$
- Site likelihoods:

  $(1-p_0) \times L(r_1 \mid \text{site})$ if site is variable

  $p_0 + (1-p_0) \times L(r_1 \mid \text{site})$ if site is invariant

Ascertainment Bias Correction

- We threw away all our invariant sites!
  - Restriction enzyme digest sites
  - Morphological characters
  - SNPs called from Whole Genome Sequencing

- So what?
  - Data appears to have unrealistically high substitution rate
  - Branch lengths are inflated
  - Might mislead tree topology estimation

Ascertainment Bias Correction

Solution:

- Likelihood is adjusted by using conditional probability

$$P(A|B) = P(A,B)/P(B)$$

- Estimate is made conditional on the site data being variable

Likelihood of Tree (T) and Parameters (Ω), for Variable (V) site Data (D):

$$L(T,\Omega|D,V) \propto P(D,V|T,\Omega)$$

Conditional Likelihood:

$$L(T,\Omega|D) \propto$$

$$P(D|T,\Omega,V) = \frac{P(D,V|T,\Omega)}{P(V)}$$

$$P(V) = 1 - P(not\ V)$$

Lewis, 2001

Ascertainment Bias Correction

Solution:

- Likelihood is adjusted by using conditional probability

$$P(A|B) = P(A,B)/P(B)$$

- Estimate is made conditional on the site data being variable

Corrected log-likelihood:

$$logL_c(\Omega) = logL(\Omega) + \delta(\Omega)$$

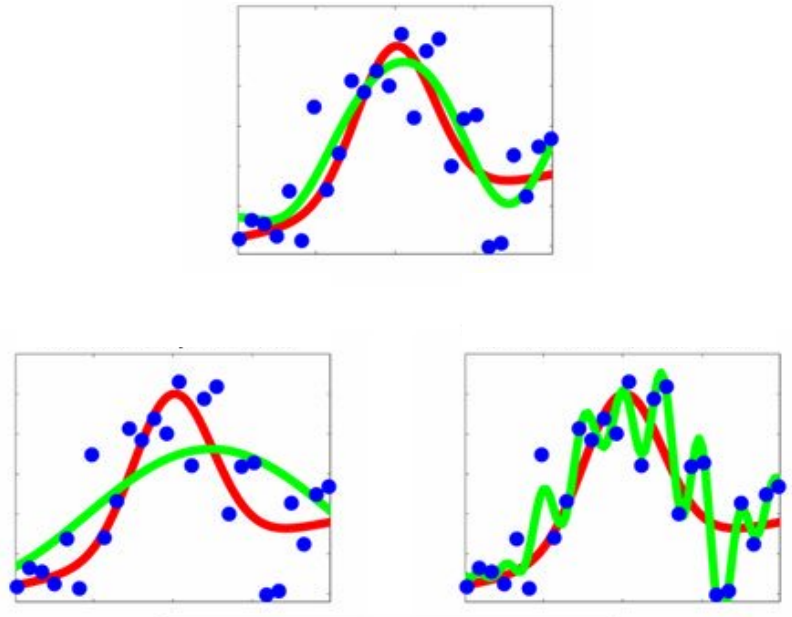
$$\delta(\Omega) = -nlog(1 - P(not\ V))$$

Nucleotide data?

- Use additional δ for each nucleotide

Tamuri and Goldman, 2017

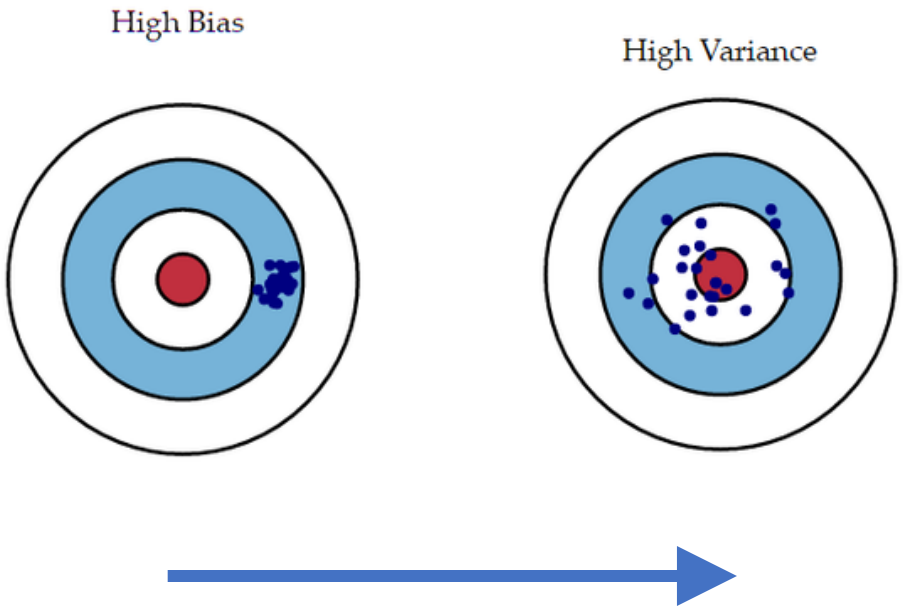
Model Selection



Model Selection

- Why does it matter which model?
  - Want a model complex enough that it captures interesting features of the data
  - Simple enough that each parameter can be estimated with high precision
- Even if not directly interested in the model
  - don't want poor model choice to affect the parts of the analysis we are interested in

Bias-Variance Trade-off

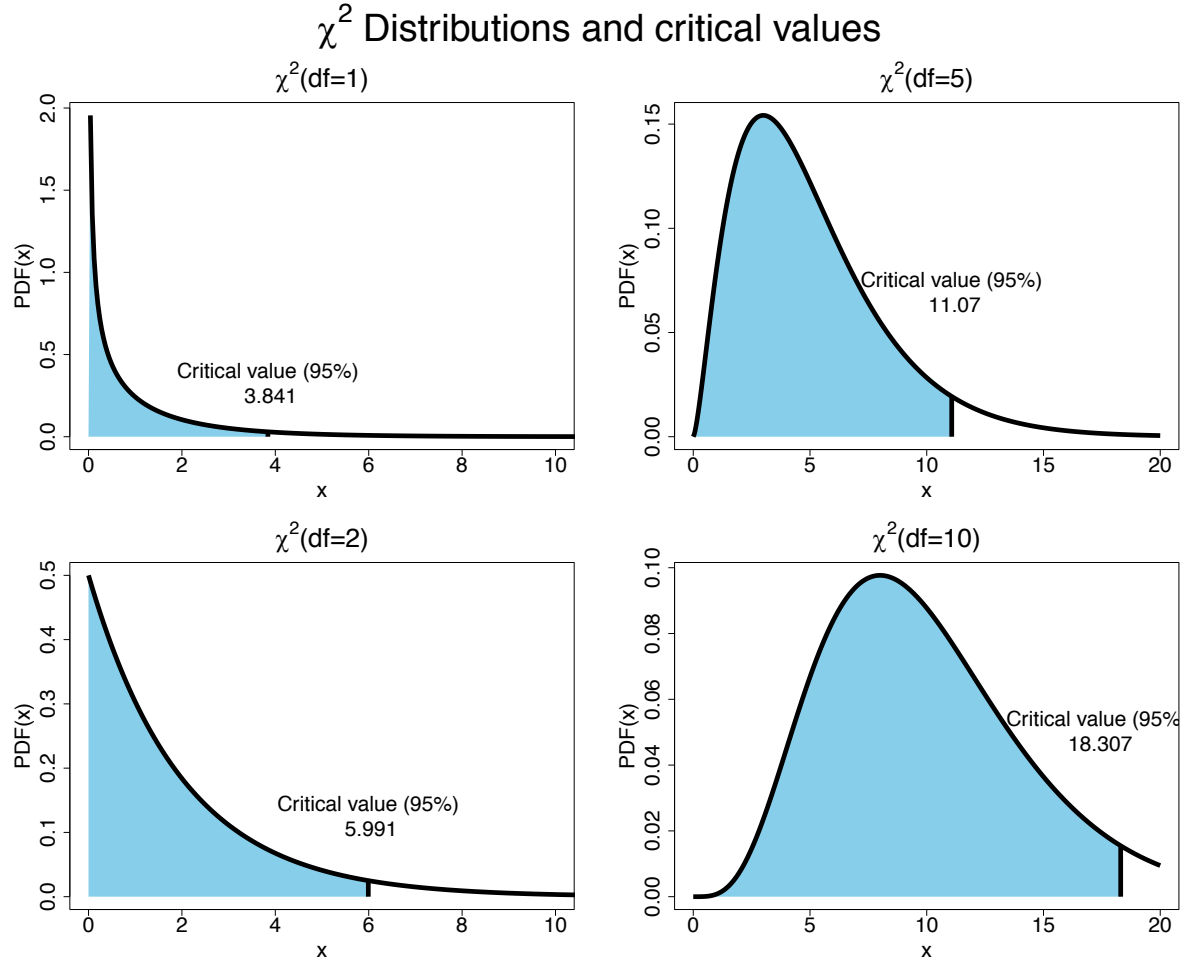


Techniques

- Likelihood Ratio Test
- Akaike Information Criterion (AIC)
- Bayes Information Criterion (BIC)
- All examples of the parsimony principle:
  - Prefer models with fewer parameters
  - Only introduce new parameters if there is a meaningful improvement in model fit

LRT – Likelihood Ratio Test

- Compare two models:
  - H_0: Null (simpler model)
  - H_1: Alternative (more parameters)
- If models are nested:
  - $\Delta = [\ \log L(H_1) - \log L(H_0)\ ]$
  - p_i = number of parameters in model i
  - $2\,\Delta \sim \chi^2(df = p_1 - p_0)$
- Reject H_1 if $2\,\Delta <$ critical value



AIC – Akaike Information Criterion

- Penalise likelihood for each extra parameter, p

$$\mathbf{AIC} = 2p - 2 \times \log L$$

AICc – Akaike Information Criterion

- Penalise likelihood for each extra parameter, p

$$\mathbf{AIC} = 2p - 2 \times \log L$$

- AICc accounts for number of data points, n

$$\mathbf{AICc} = AIC + \frac{2p(p+1)}{n-p-1}$$

BIC – Bayes Information Criterion

- Penalise likelihood for each extra parameter, p

$$\mathbf{AIC} = 2p - 2 \times \log L$$

- AICc accounts for number of data points, n

$$\mathbf{AICc} = AIC + \frac{2p(p+1)}{n-p-1}$$

- BIC penalises extra parameters more strongly than AIC

$$\mathbf{BIC} = \log(n) \times p - 2 \times \log L$$

Example

Model	sites	params	logL	LRT	AIC	AICc	BIC
JC	2852	60	-40020.222		80160.444	80163.066	80517.790
K80	2852	61	-36393.919	7252.606	72909.838	72912.549	73273.140
HKY85	2852	64	-36163.107	461.624	72454.213	72457.199	72835.383

$\chi^2(1)$ critical value: 3.84
$\chi^2(3)$ critical value: 7.81

Practicals

Model Selection by Hand

Use the alignment file **primate-mtDNA.fas**

1. Estimate a parsimony tree:
   - `iqtree -s primate-mtDNA.fas -te PARS -m HKY -pre static`
2. Calculate the likelihood using this tree for the following models:
   - `iqtree -s primate-mtDNA.fas -te static.treefile -m JC -pre jc`
   - `iqtree -s primate-mtDNA.fas -te static.treefile -m JC+G4 -pre jcg`
   - `iqtree -s primate-mtDNA.fas -te static.treefile -m K80 -pre k80`
   - `iqtree -s primate-mtDNA.fas -te static.treefile -m K80+G4 -pre k80g`
   - `iqtree -s primate-mtDNA.fas -te static.treefile -m HKY -pre hky`
   - `iqtree -s primate-mtDNA.fas -te static.treefile -m HKY+G4 -pre hkyg`
   - `iqtree -s primate-mtDNA.fas -te static.treefile -m GTR -pre gtr`
   - `iqtree -s primate-mtDNA.fas -te static.treefile -m GTR+G4 -pre gtrg`
3. Fill in a table of LRT, AIC, AICc and BIC values, and choose the best model

Automatic Model Selection

- Use the file **taz_mini.fas**. This file contains 5000 SNPs found in Tasmanian Devil Facial Tumour cancers. Only variable sites are included, so ascertainment bias correction is needed.
- Run IQtree's automated model selection procedure

```
iqtree -s taz_mini.fas -pre model_finder
```

- What are the features of the chosen model? Do you agree with the choice?

Description of IQtree's models:

http://www.iqtree.org/doc/Substitution-Models

References

Pruning algorithm:

Felsenstein, J. (1973). Maximum-likelihood and minimum-steps methods for evolutionary trees from data on discrete characters. Syst. Zool. 22:240-249

Ascertainment Bias Correction:

Felsenstein, J. (1992). Phylogenies from restriction sites: A maximum-likelihood approach. Evolution 46:159-173

Lewis, P. O. (2001). A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology* 50 (6): 913-25.

Tamuri, Asif, and Nick Goldman. (2017). Avoiding Ascertainment Bias in the Maximum Likelihood Inference of Phylogenies Based on Truncated Data. *bioRxiv*.

Everything:

Felsenstein, J. (2004) Inferring Phylogenies. Sinauer Associates Publishing.

Yang, Z. (2014). Molecular Evolution. A statistical approach. Oxford University Press