Evolutionary models

Wytamma Wirth

- Postdoctoral researcher @ The Doherty
- Working with Sebastian Duchene
- PhD in pathology and epidemiology (JCU)
- Research interests:
  - Drivers of infectious diseases
  - Computational biology
  - Tool building
- @wytamma

Popular phylogenetic methods

Find the distance between a pair of aligned molecular sequences

Popular phylogenetic methods:

- Maximum parsimony
- Distance-based methods
- Maximum likelihood
- **Bayesian inference**

Model based

What are models?

A mathematical model is a stringently phrased hypothesis

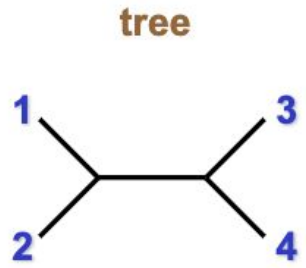
Models are not 1:1 representations of the world (and they shouldn’t be) they are abstractions that allow us to test our beliefs about the world.

The evolutionary models that we choose reflect our beliefs about the generative process of our data.

Be careful when selecting models!

Maximum parsimony

Taxon-1	ATATT
Taxon-2	ATCGT
Taxon-3	GCAGT
Taxon-4	GCCGT



Identifies the tree topology that can explain the sequence data, using the smallest number of inferred substitution events

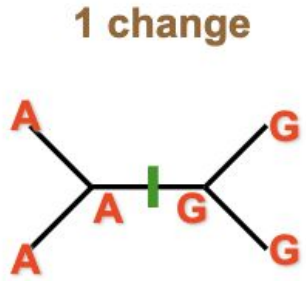
Not model based

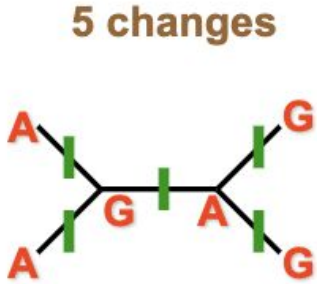
Now rarely used for analysing genetic data

Cannot estimate evolutionary rates or timescales

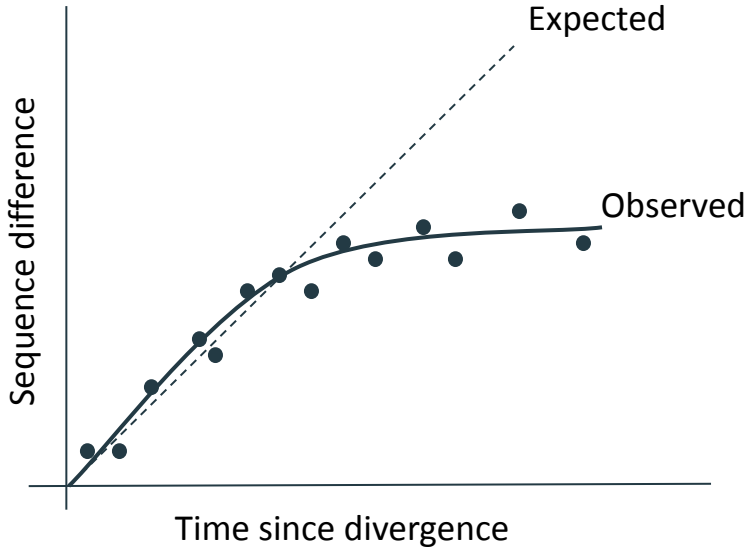
Does not correct for multiple substitutions at the same site

- This leads to a problem known as ‘long-branch attraction’
- Long branches in the tree tend to group together





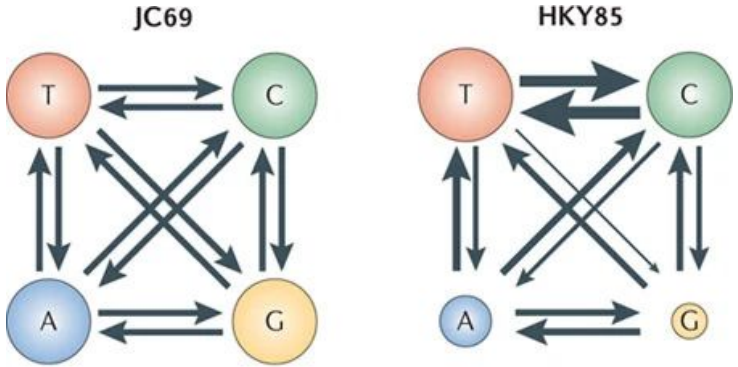
Multiple substitutions

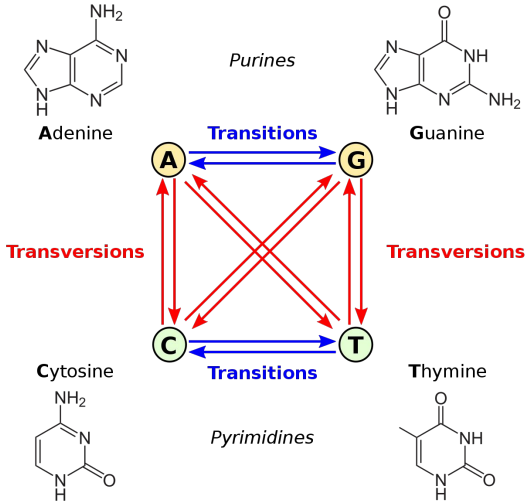


ACCGGTC
TCCCGTC
GCCGGTA

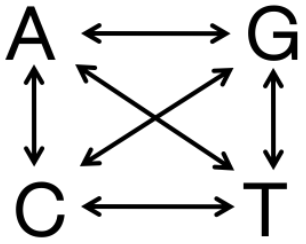
Nucleotide substitution models

- How nucleotides change to other nucleotides.
- Used for distance correction.
- Continuous-time Markov process





Rate Matrix



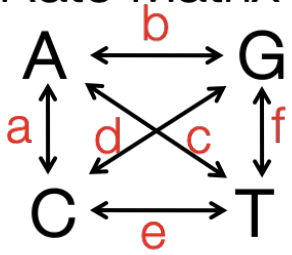
Base Frequencies

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

Site Rates

$+ I + G$

Rate Matrix



Base Frequencies

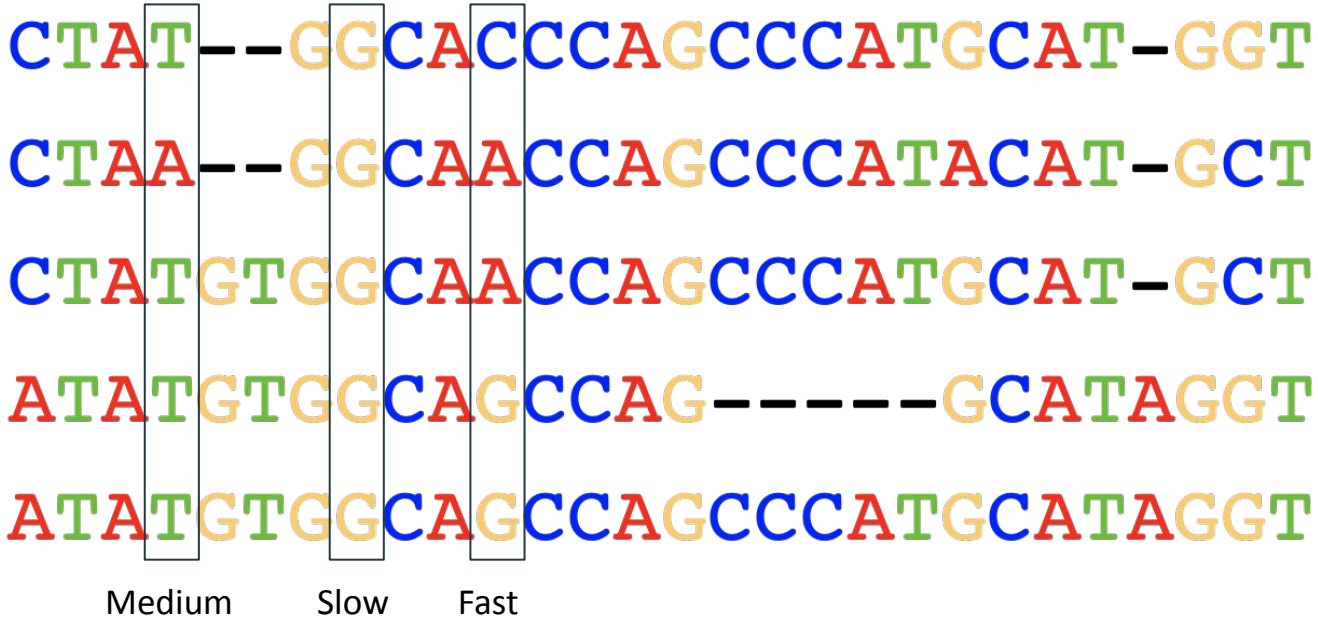
$\pi_A + \pi_C + \pi_G + \pi_T = 1$

Site Rates

$+ I + G$

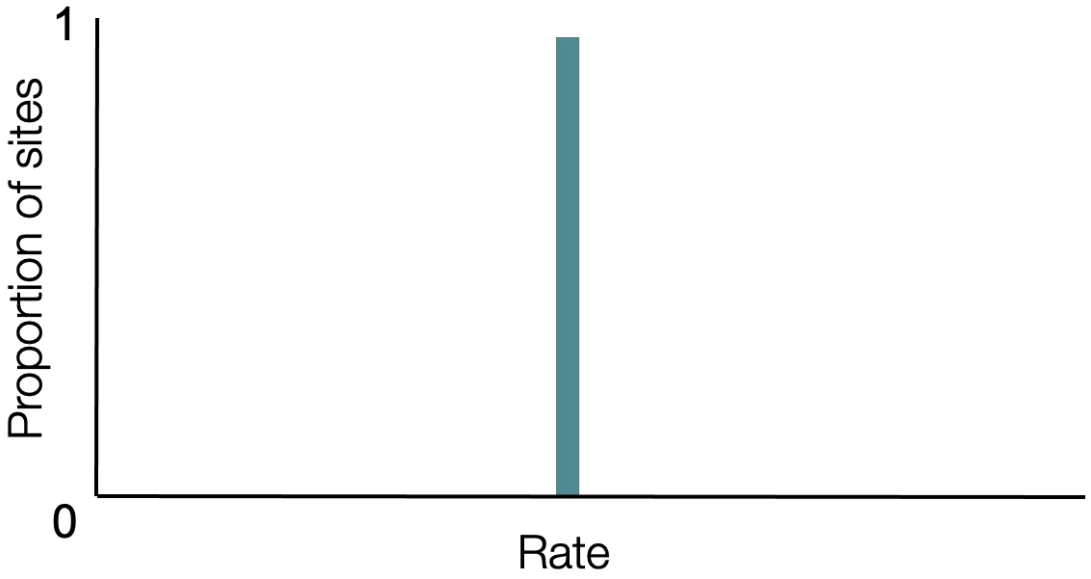
JC	HKY	GTR	GTR+I+G
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$
No I or G	No I or G	No I or G	I, G
0 free parameters	4 free parameters	8 free parameters	10 free parameters

Rate variation across sites



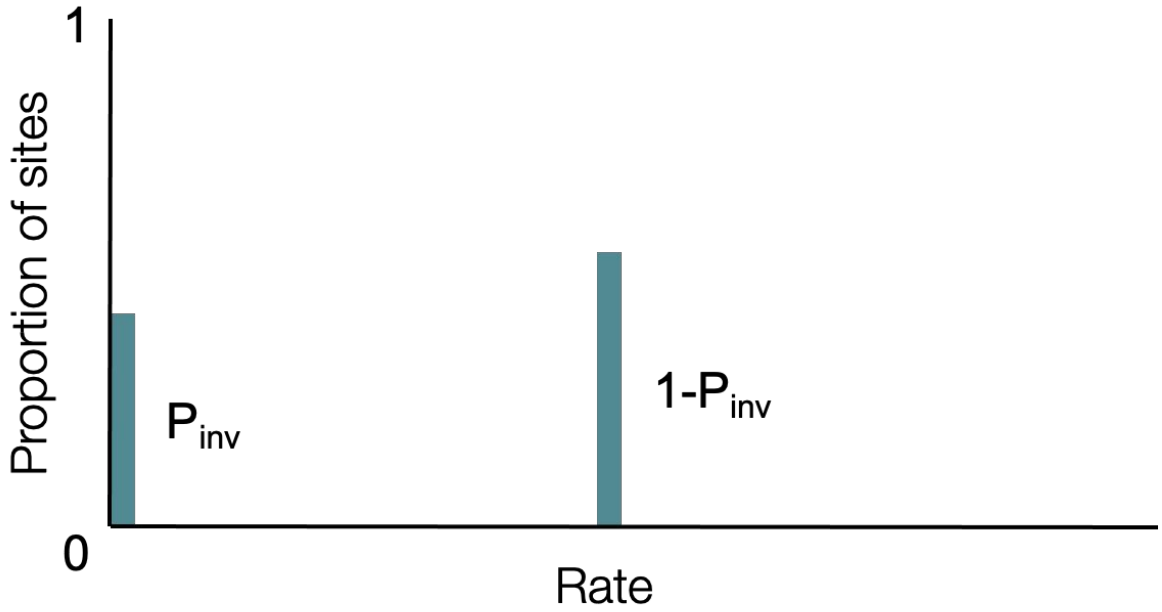
Rate variation across sites

Equal rates among sites (e.g., **JC**, **GTR**, **HKY** models)



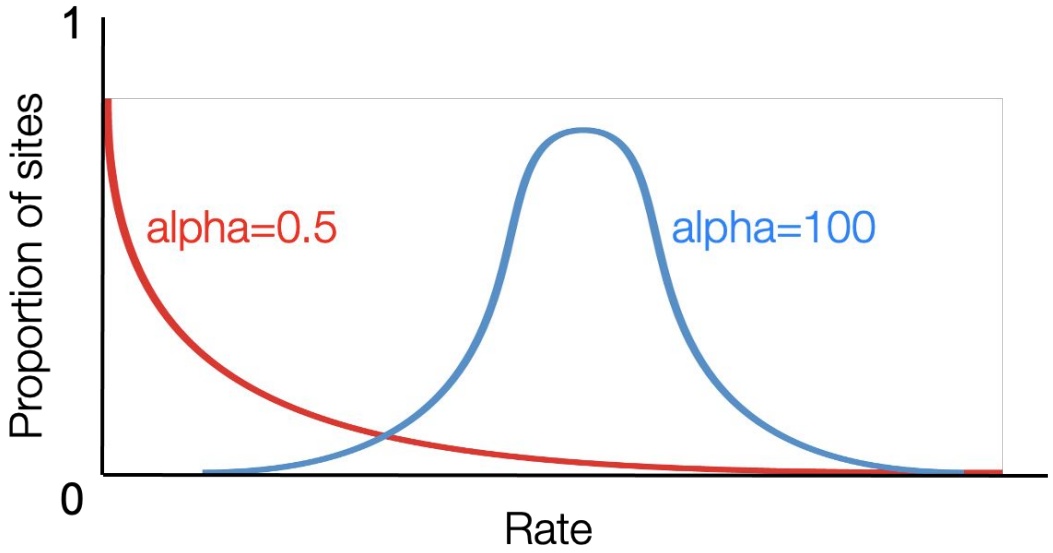
Rate variation across sites

- Proportion of invariable sites (e.g., JC+I, GTR+I, HKY+I models)



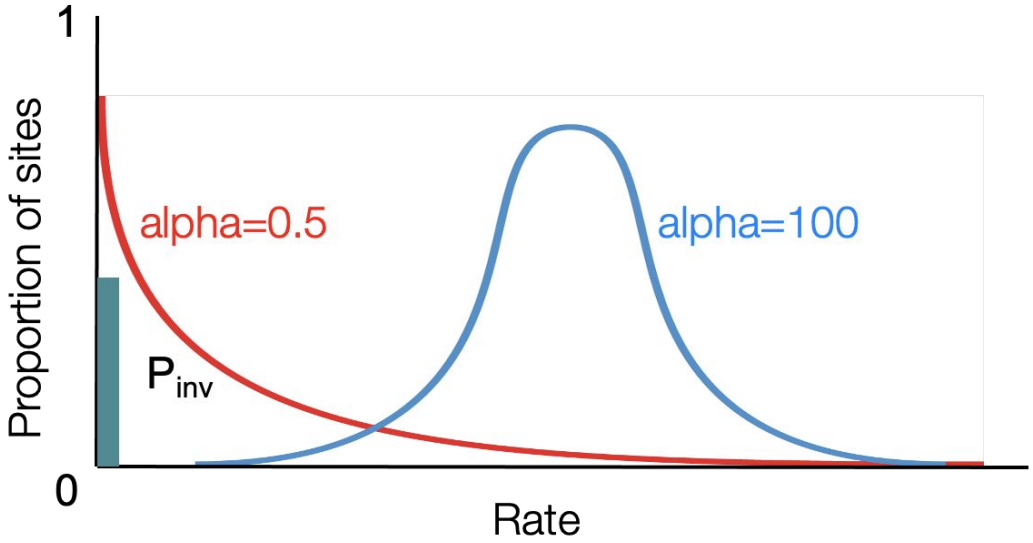
Rate variation across sites

•Gamma-distributed rate variation among sites
(e.g., JC+G, GTR+G, HKY+G models)



Rate variation across sites

•Gamma-distributed rate variation among sites and a proportion of invariable sites (e.g., JC+G+I, GTR+G+I, HKY+G+I models)



Amino acid substitution matrices

20x20 matrix of substitution probabilities

Too many parameters to estimate

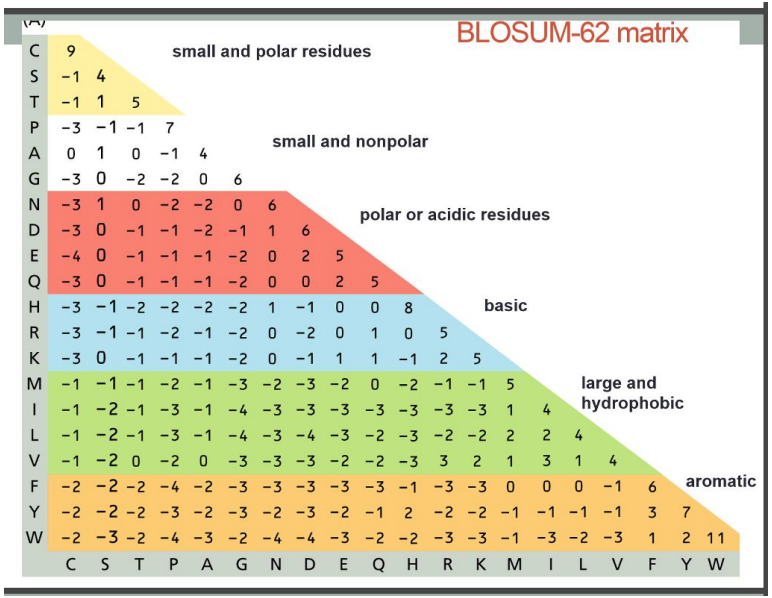
GTR model for DNA: 6 parameters

GTR model for proteins: 190 parameters

Estimate substitution probabilities using a large data set

Standard matrices:

- PAM, BLOSUM, EDSS, etc.



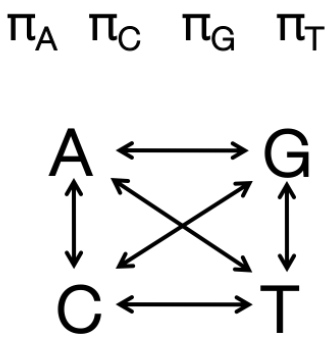
Fundamental assumptions

Stationary – base frequencies the same across the alignment

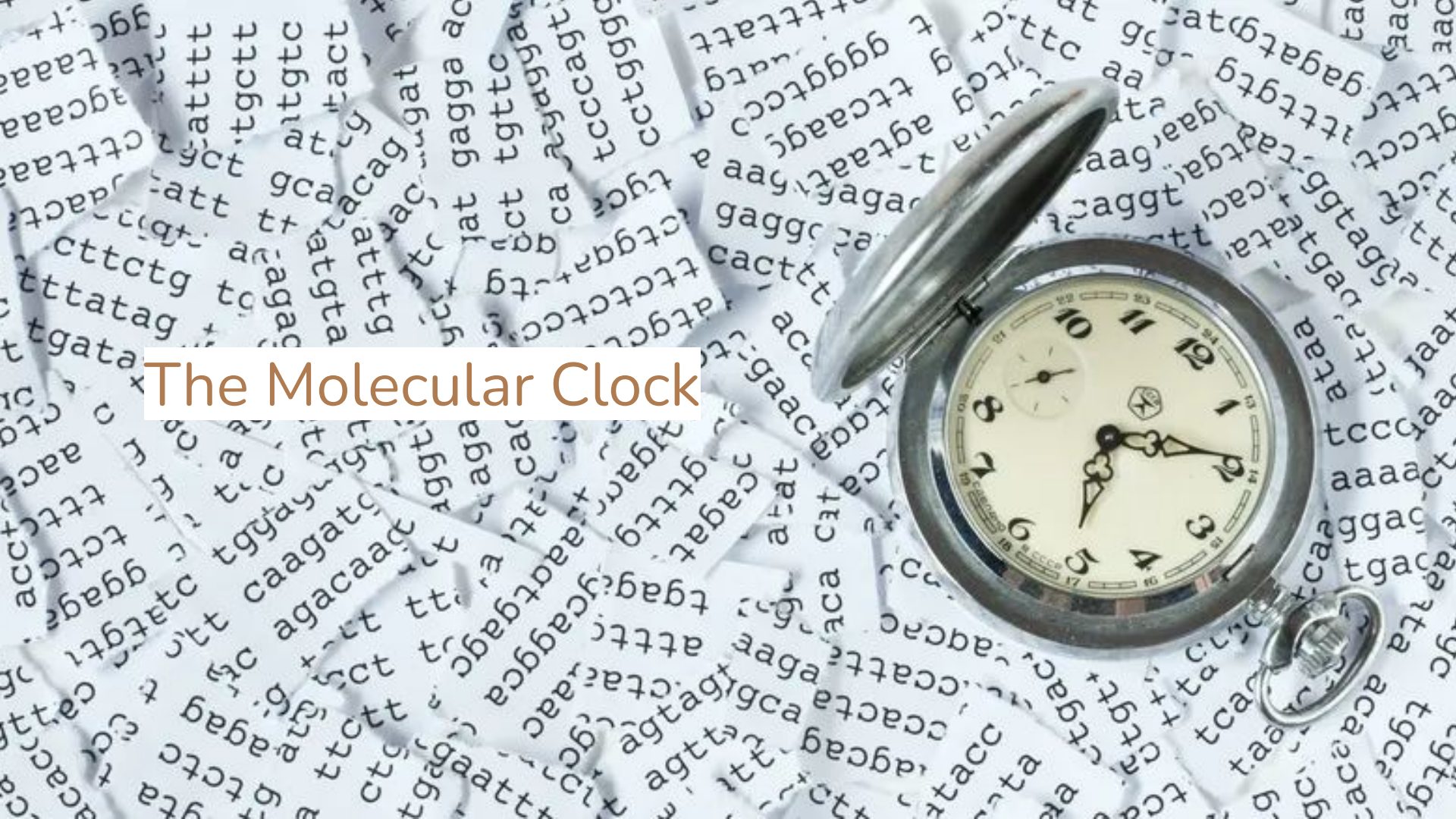
Reversible – probability of going from one to the other the same as going the other way

Homogenous – probabilities are the same across the alignment

Independent across sites - neighbouring sites do not influence



The Molecular Clock



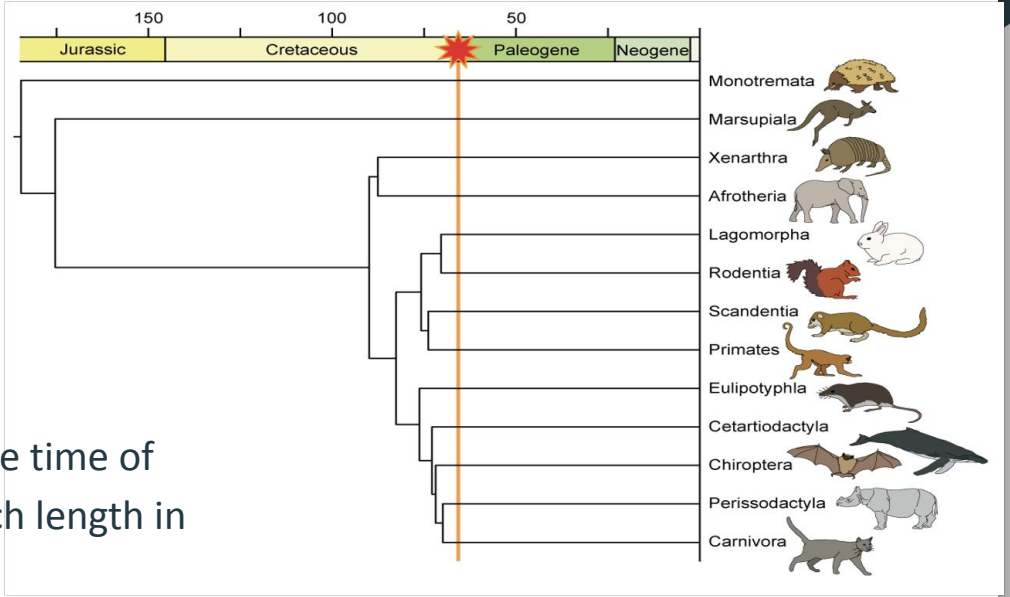
The Molecular Clock

Molecular clock refers to estimating evolutionary timescales

Understand evolution over millions of years or trace the spread of disease over months.

Estimate a chronogram

- The clock model is used to estimate the time of evolutionary divergence (i.e. the branch length in a chronogram).

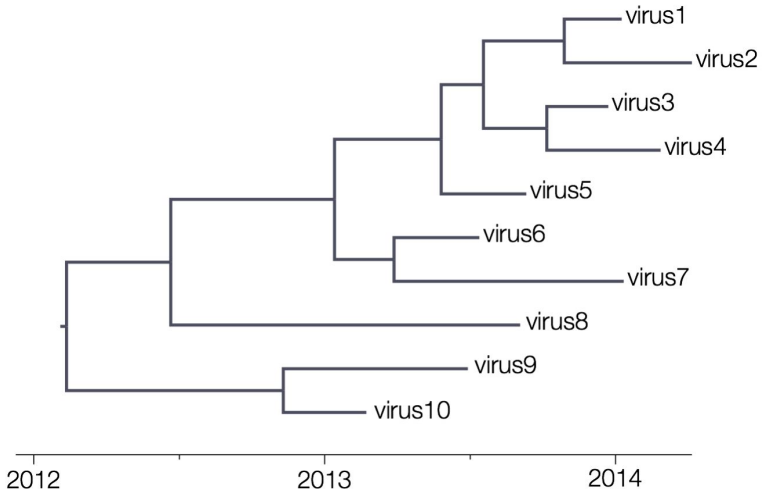


Strict clock

Every branch in a phylogenetic tree evolves according to the same evolutionary rate.

1 parameter model (conversion rate between branch lengths and evolutionary time)

Strict clocks typically are best suited for early outbreaks or populations with low structure or diversity

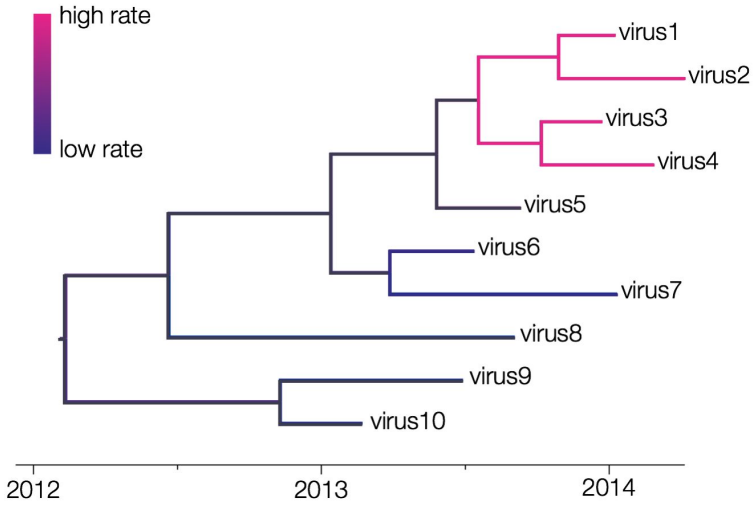


Local molecular clock

One or more specific clades in the tree does not evolve according to this global rate

Selected clades/lineages evolve according different evolutionary rates while rate constancy is assumed across the remainder of the tree

Used when there is structure in the population. Can be driven by biological characteristics. I.e. generation time



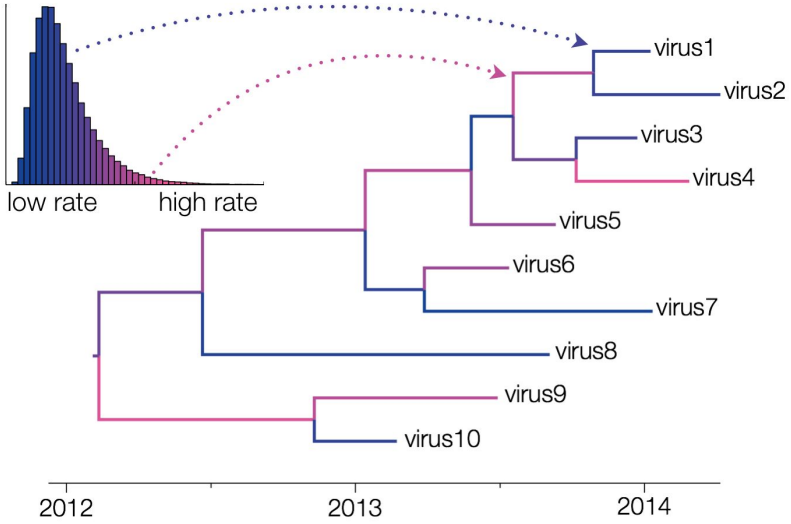
Bayesian relaxed clocks

Allow a different rate in each branch

Statistical models of rates among branches

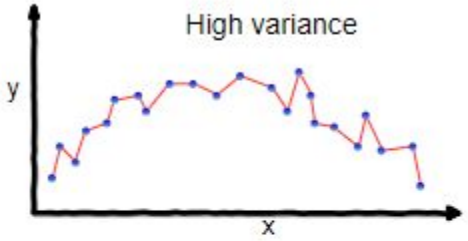
Rates can be auto-correlated or uncorrelated

- Autocorrelated: rates in neighbouring branches are related
- Uncorrelated: rates identically and independently distributed among branches

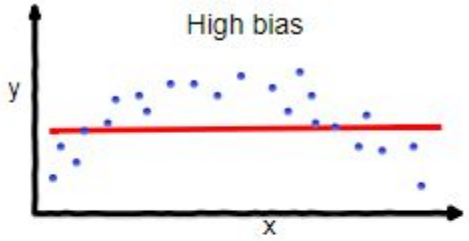


Model selection and averaging

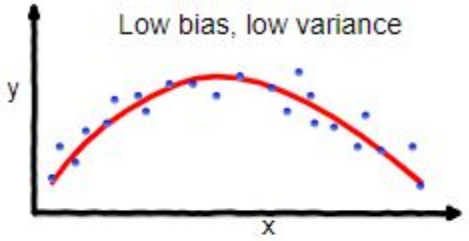
Model Selection



overfitting



underfitting



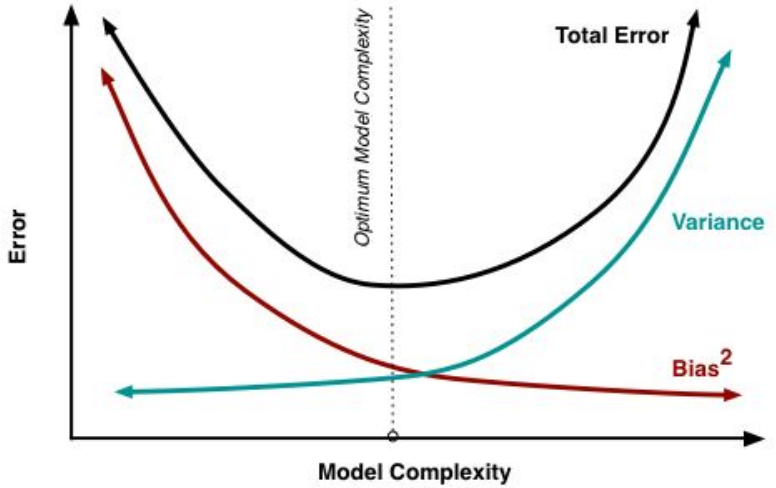
Good balance

Bias-variance tradeoff

Adding more parameters always improves the fit of the model to the observed data

But it doesn't necessarily improve the model!

Goal is to find the best balance between bias and variance



Model selection methods

Adding a parameter to the model:

- Is the improvement in likelihood worth the cost of adding a parameter?

Compare and rank different hypothesis

Model selection methods:

- Likelihood-ratio test (LRT)
  - Compare nested models
- Akaike information criterion (AIC)
  - Compare non-nested models
- Bayesian information criterion (BIC)
  - Stronger penalty on number of parameters (compared to AIC)

Bayes Factors

Ratio of the marginal likelihoods (denominator in Bayes formula) of two models.

Computationally intensive task and there are several ways to estimate marginal likelihoods:

- Stepping stone
- Path sampling
- Nested sampling

If BF is larger than 1, model M1 is favoured, and otherwise M2 is favoured.

$\log_{10}(BF)$ range			Interpretation
0	–	0.5	hardly worth mentioning
0.5	–	1.3	positive support
1.3	–	2.2	strong support
	>	2.2	overwhelming support

$$\frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Model averaging

The MCMC algorithms jumps between the different models

Models more appropriate for the data will be sampled more often than unsuitable models.

Results averaged across models