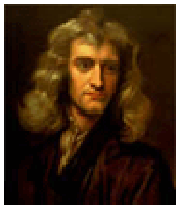
MANCHESTER 1824

The University of Manchester

Modelling heterogeneity in nucleotide sequence evolution

Simon Whelan

Supported by:



Isaac Newton Institute

Talk outline

i. Introduction: what is spatial and temporal heterogeneity?

ii. A temporal hidden Markov model of sequence evolution

iii. Characterizing heterogeneity in real sequence data

iv. Heterogeneity and the genetic code

Talk outline

i. *Introduction: what is spatial and temporal heterogeneity?*

ii. *A temporal hidden Markov model of sequence evolution*

iii. *Characterizing heterogeneity in real sequence data*

iv. *Heterogeneity and the genetic code*

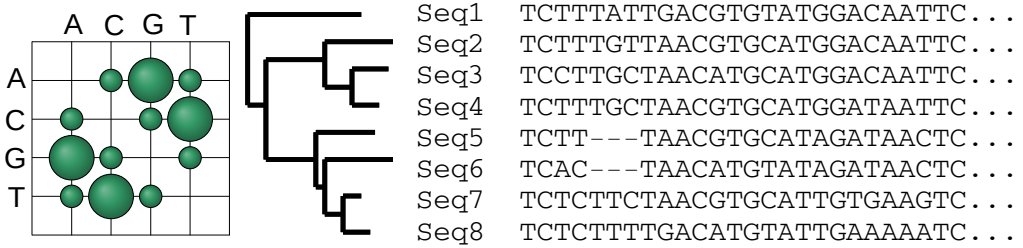
Why worry about heterogeneity?

Popular models of sequence evolution

GTR family for nucleotide substitutions

Empirical models for amino acid substitutions (WAG; mtREV)

Heterogeneity: rate variation between sites (Γ-distribution)



Heterogeneity and systematic error

Heterogeneity can cause popular models to go wrong

Misleading estimates of evolutionary relationships

Model misspecification can confuse inferences about process

Heterogeneity is what makes evolution interesting!

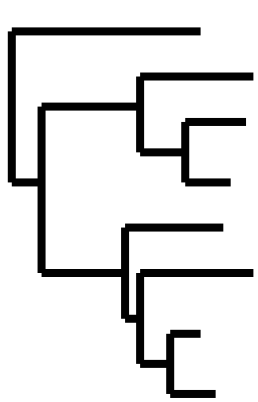
Can be the result of molecular adaptation or environmental changes

Provide an understanding of biological diversity

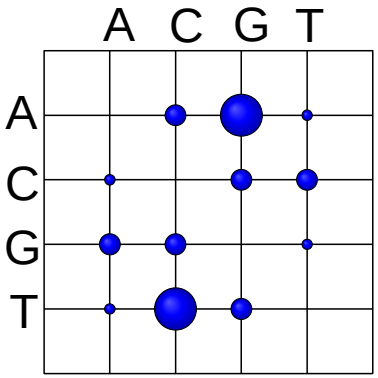
Allows dating of important evolutionary events

Spatial heterogeneity in sequence evolution

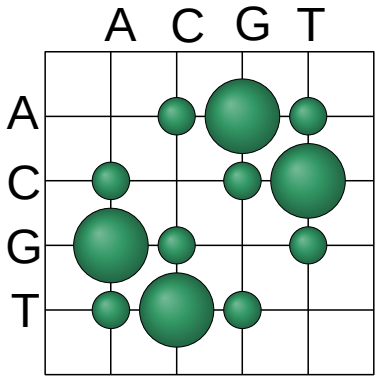
Also known as pattern heterogeneity



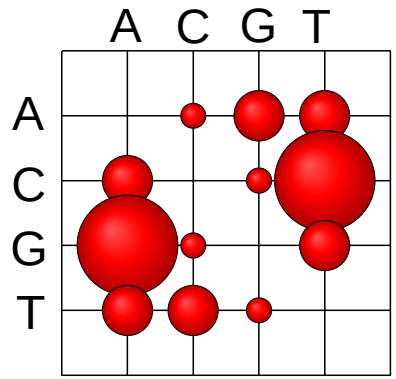
```
Seq1   TCTTTATTGACGTGTATGGACAATTC...
Seq2   TCTTTGTTAACGTGCATGGACAATTC...
Seq3   TCCTTGCTAACATGCATGGACAATTC...
Seq4   TCTTTGCTAACGTGCATGGATAATTC...
Seq5   TCTT---TAACGTGCATAGATAACTC...
Seq6   TCAC---TAACATGTATAGATAACTC...
Seq7   TCTCTTCTAACGTGCATTGTGAAGTC...
Seq8   TCTCTTTTGACATGTATTGAAAAATC...
```



Rate = 0.5



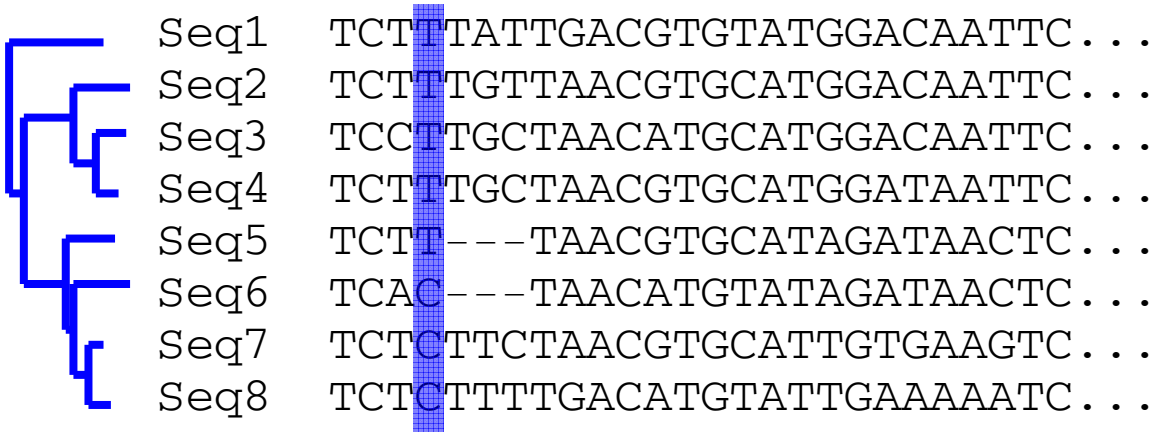
Rate = 1.0

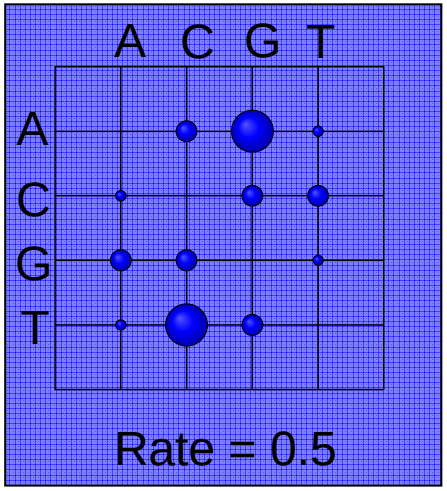


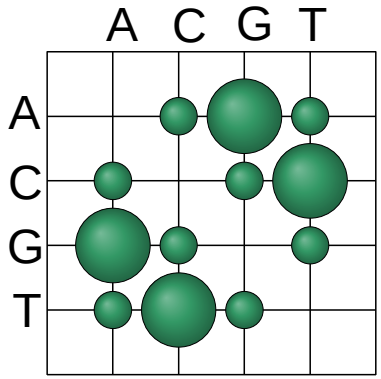
Rate = 2.0

Spatial heterogeneity in sequence evolution

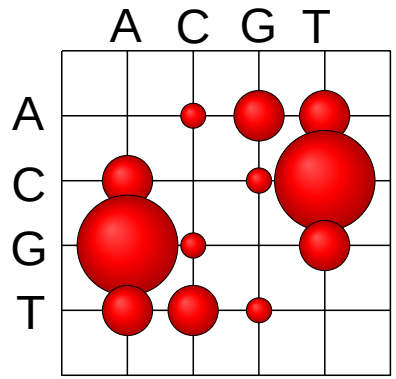
Also known as pattern heterogeneity







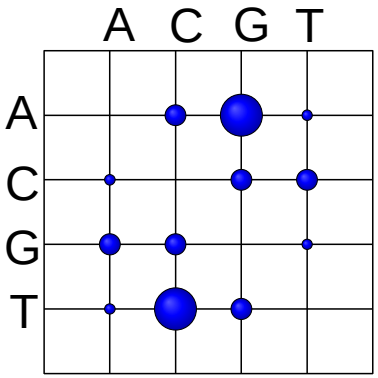
Rate = 1.0



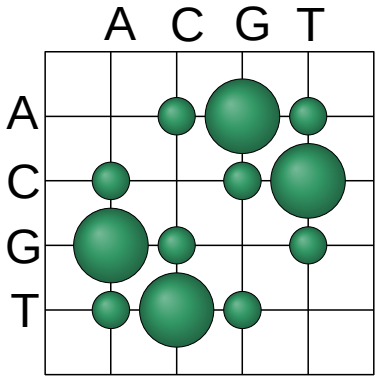
Rate = 2.0

Spatial heterogeneity in sequence evolution

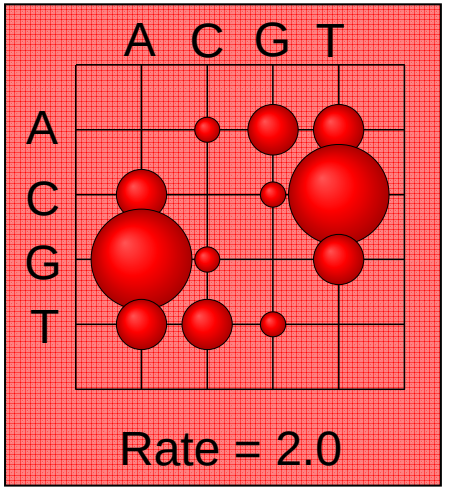
Also known as pattern heterogeneity



Rate = 0.5

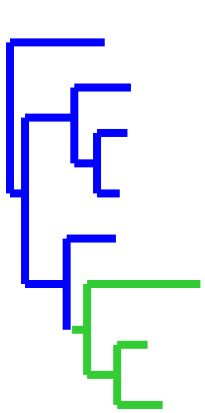


Rate = 1.0

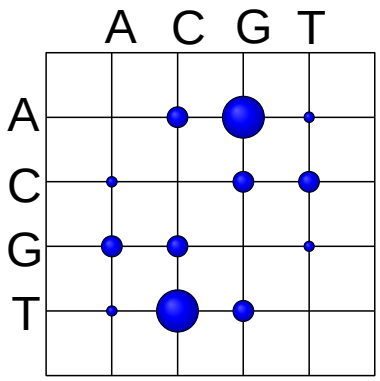


Rate = 2.0

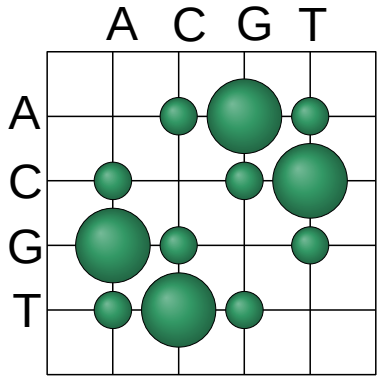
Temporal heterogeneity in sequence evolution



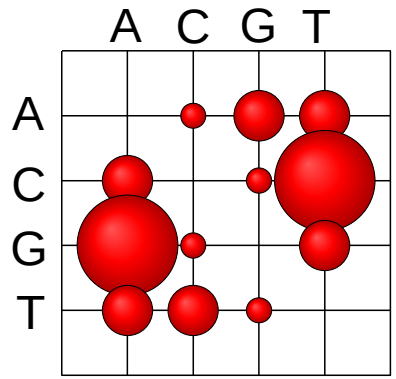
```
Seq1   TCTTTATTGACGTGTATGGACAATTC...
Seq2   TCTTTGTTAACGTGCATGGACAATTC...
Seq3   TCCTTGCTAACATGCATGGACAATTC...
Seq4   TCTTTGCTAACGTGCATGGATAATTC...
Seq5   TCTT---TAACGTGCATAGATAACTC...
Seq6   TCAC---TAACATGTATAGATAACTC...
Seq7   TCTCTTCTAACGTGCATTGTGAAGTC...
Seq8   TCTCTTTTGACATGTATTGAAAATC...
```



Rate = 0.5

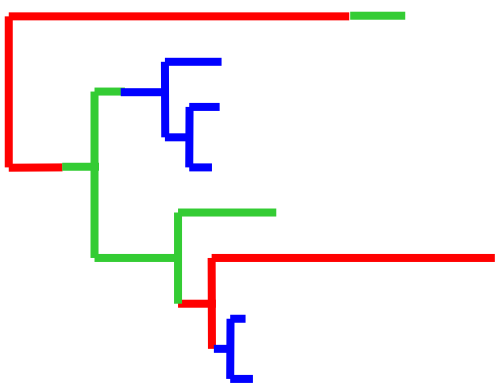


Rate = 1.0

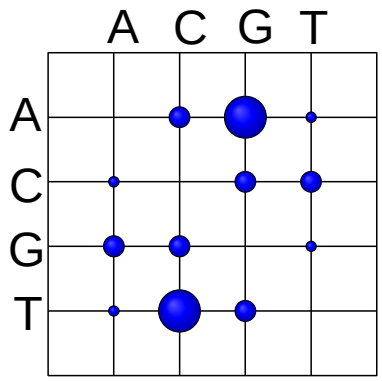


Rate = 2.0

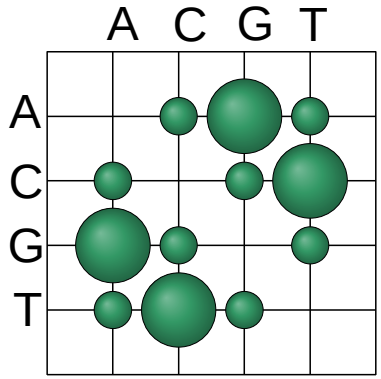
Temporal heterogeneity in sequence evolution



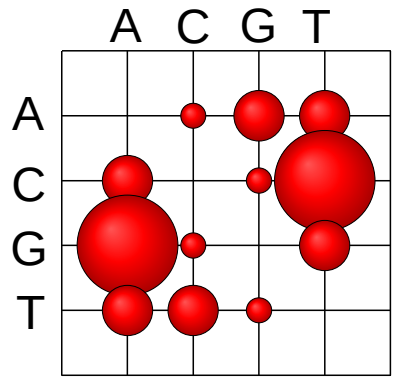
```
Seq1   TCTTTATTGACGTGTATGGACAATTC...
Seq2   TCTTTGTTAACGTGCATGGACAATTC...
Seq3   TCCTTGCTAACATGCATGGACAATTC...
Seq4   TCTTTGCTAACGTGCATGGATAATTC...
Seq5   TCTT---TAACGTGCATAGATAACTC...
Seq6   TCAC---TAACATGTATAGATAACTC...
Seq7   TCTCTTCTAACGTGCATTGTGAAGTC...
Seq8   TCTCTTTTGACATGTATTGAAAAATC...
```



Rate = 0.5



Rate = 1.0

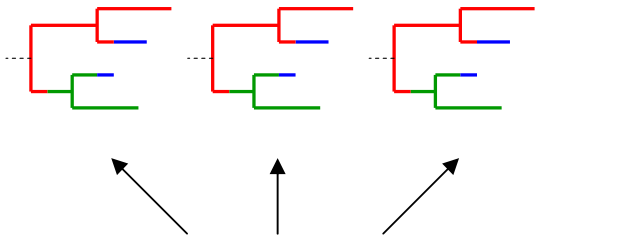


Rate = 2.0

Different forms of temporal heterogeneity

Fixed effect temporal heterogeneity

Each site has the same temporal heterogeneity



```
Seq1   TCTTTATTGACGTGTATGGACAATTCTCTTTAACGTGC
Seq2   TCTTTGTTAACGTGCATGGACAATTCTCTTTAACGTGC
Seq3   TCCTTGCTAACATGCATGGACAATTCTCTCTAACGTGC
Seq4   TCTTTGCTAACGTGCATGGATAATTCTCTCTGACATGT
```

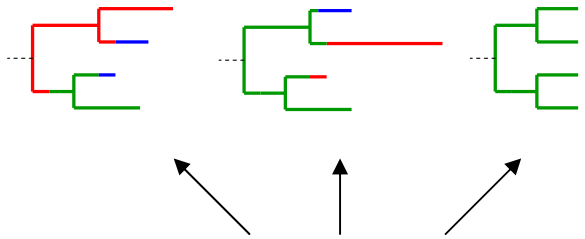
Biological causes include %GC-content variation and overall changes in selection

Can distinguish between temporal and spatial heterogeneity

Suitable for analysis under the general Markov model (Barry and Hartigan 1987)

Random effect temporal heterogeneity

Each site has a randomly chosen type of temporal heterogeneity



```
Seq1   TCTTTATTGACGTGTATGGACAATTCTCTTTAACGTGC
Seq2   TCTTTGTTAACGTGCATGGACAATTCTCTTTAACGTGC
Seq3   TCCTTGCTAACATGCATGGACAATTCTCTCTAACGTGC
Seq4   TCTTTGCTAACGTGCATGGATAATTCTCTCTGACATGT
```

Biological causes include changes in molecular structure, and the genetic code (see later)

Spatial and temporal heterogeneity become intertwined

Few models widely available for inference

Talk outline

i. *Introduction: what is spatial and temporal heterogeneity?*

ii. *A temporal hidden Markov model of sequence evolution*

iii. *Characterizing heterogeneity in real sequence data*

iv. *Heterogeneity and the genetic code*

Temporal hidden Markov models (THMMs)

Purpose of model

Describe random effect spatial and temporal heterogeneity

Allow simple likelihood computation (reversible; stationary; i.i.d.)

Previous incarnations

Mostly examine temporal and spatial rate variation

Covarion model of Tuffley and Steel and its progeny

Other names from phylogenetics and computer science include:

- Markov modulated Markov processes (models)
- Switching processes
- Covarion-like

Substitution classes

There are $1,\ldots,g$ separate HKY substitution processes, each representing a hidden class in a HMM

The k^{th} hidden class is defined by rate matrix $\mathbf{M}^k$:

$$\mathbf{M}^k = \mu^k \begin{bmatrix} - & \tilde{\pi}^k_{\mathrm{C}} & \kappa^k \tilde{\pi}^k_{\mathrm{G}} & \tilde{\pi}^k_{\mathrm{T}} \\ \tilde{\pi}^k_{\mathrm{A}} & - & \tilde{\pi}^k_{\mathrm{G}} & \kappa^k \tilde{\pi}^k_{\mathrm{T}} \\ \kappa^k \tilde{\pi}^k_{\mathrm{A}} & \tilde{\pi}^k_{\mathrm{C}} & - & \tilde{\pi}^k_{\mathrm{T}} \\ \tilde{\pi}^k_{\mathrm{A}} & \kappa^k \tilde{\pi}^k_{\mathrm{C}} & \tilde{\pi}^k_{\mathrm{G}} & - \end{bmatrix}$$

$\left\{\tilde{\pi}^k_{\mathrm{A}}, \tilde{\pi}^k_{\mathrm{C}}, \tilde{\pi}^k_{\mathrm{G}}, \tilde{\pi}^k_{\mathrm{T}}\right\}$ = nucleotide distribution of hidden class k

μ^k = rate of hidden class k

κ^k = transition/transversion rate ratio of hidden class k

Note: Subscripts refer to observable states. Superscripts refer to hidden classes

Temporal heterogeneity: a switching model

A reversible Markov model describing the switching rate between hidden classes

This process defined by g x g rate matrix $\mathbf{C}$

$$\mathbf{C} = \begin{bmatrix} - & \rho^{1,2}\tilde{\pi}^{2} & \cdots & \rho^{1,g}\tilde{\pi}^{g} \\ \rho^{1,2}\tilde{\pi}^{1} & - & & \rho^{2,g}\tilde{\pi}^{g} \\ \vdots & & \ddots & \\ \rho^{1,g}\tilde{\pi}^{1} & \rho^{2,g}\tilde{\pi}^{2} & & - \end{bmatrix}$$

$\rho^{k,l}$ = exchangeability between hidden classes k and l

$\tilde{\pi}^{1}, \tilde{\pi}^{2}, \ldots, \tilde{\pi}^{g}$ = probability of a hidden class

Note: Subscripts refer to observable states. Superscripts refer to hidden classes

Defining a THMM for DNA substitution

The 4*g* x 4*g* instantaneous rate matrix is:

$$Q_{i,j}^{k,l} = \begin{cases} M_{i,j}^{k} & \text{for all } i \neq j \text{ and } k = l \\ \tilde{\pi}_j^l C^{k,l} & \text{for all } i = j \text{ and } k \neq l \\ 0 & \text{for all } i \neq j \text{ and } k \neq l \end{cases}$$

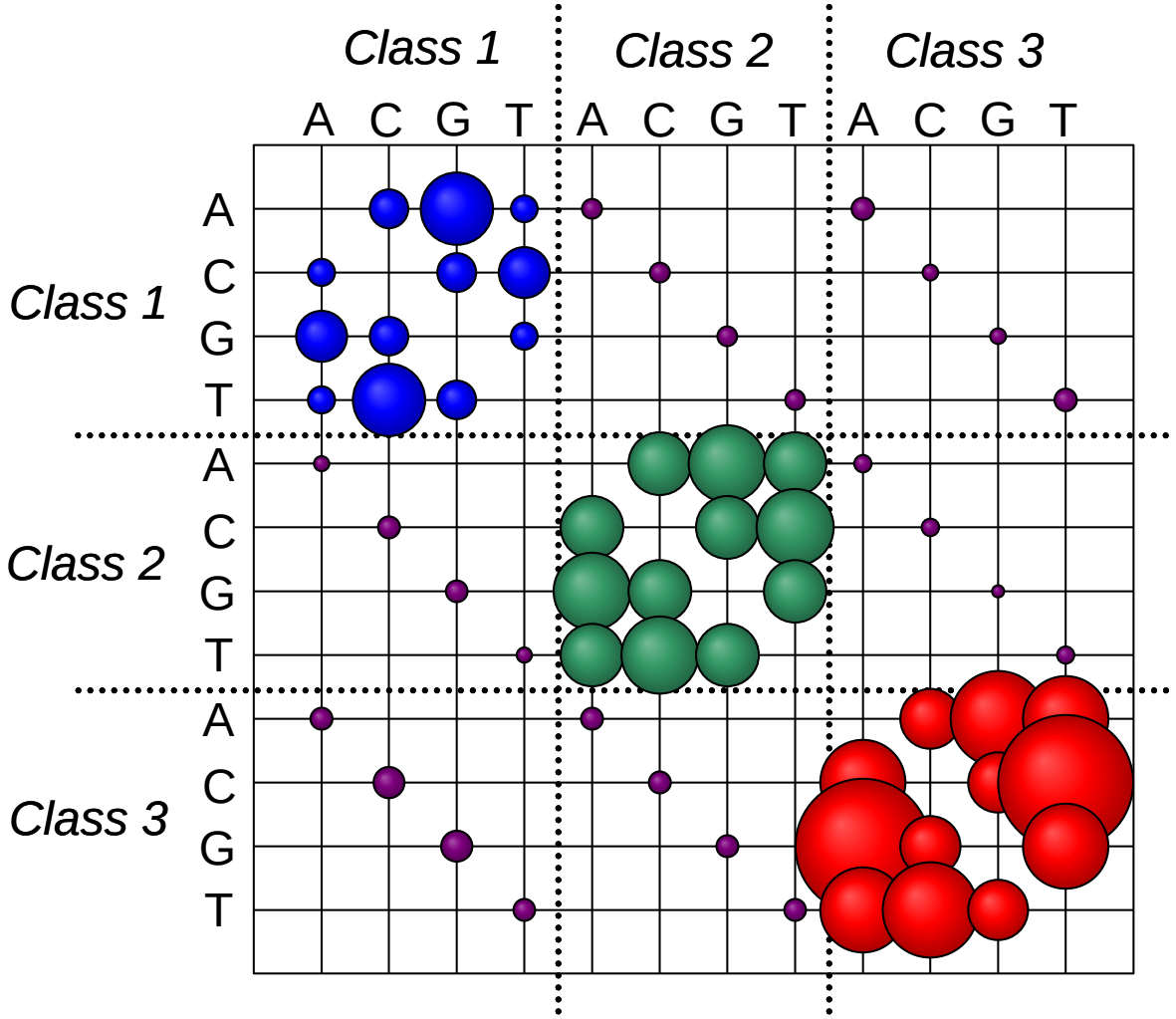
$Q_{i,j}^{k,l}$ = changes between observable states i, j and hidden classes k, l

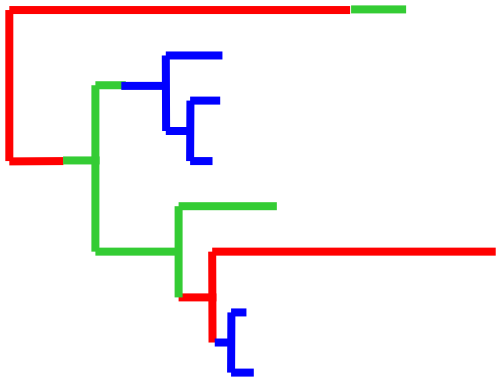
Equilibrium distribution is $\pi_i^k = \tilde{\pi}^k \tilde{\pi}_i^k$

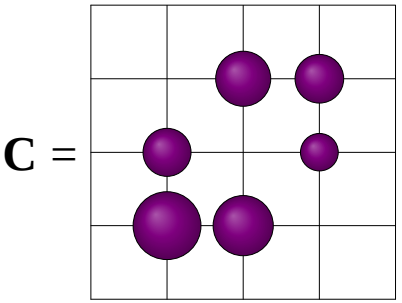
Hidden classes and observable states do not change simultaneously

Note: Subscripts refer to observable states. Superscripts refer to hidden classes

THMMs for spatial and temporal heterogeneity



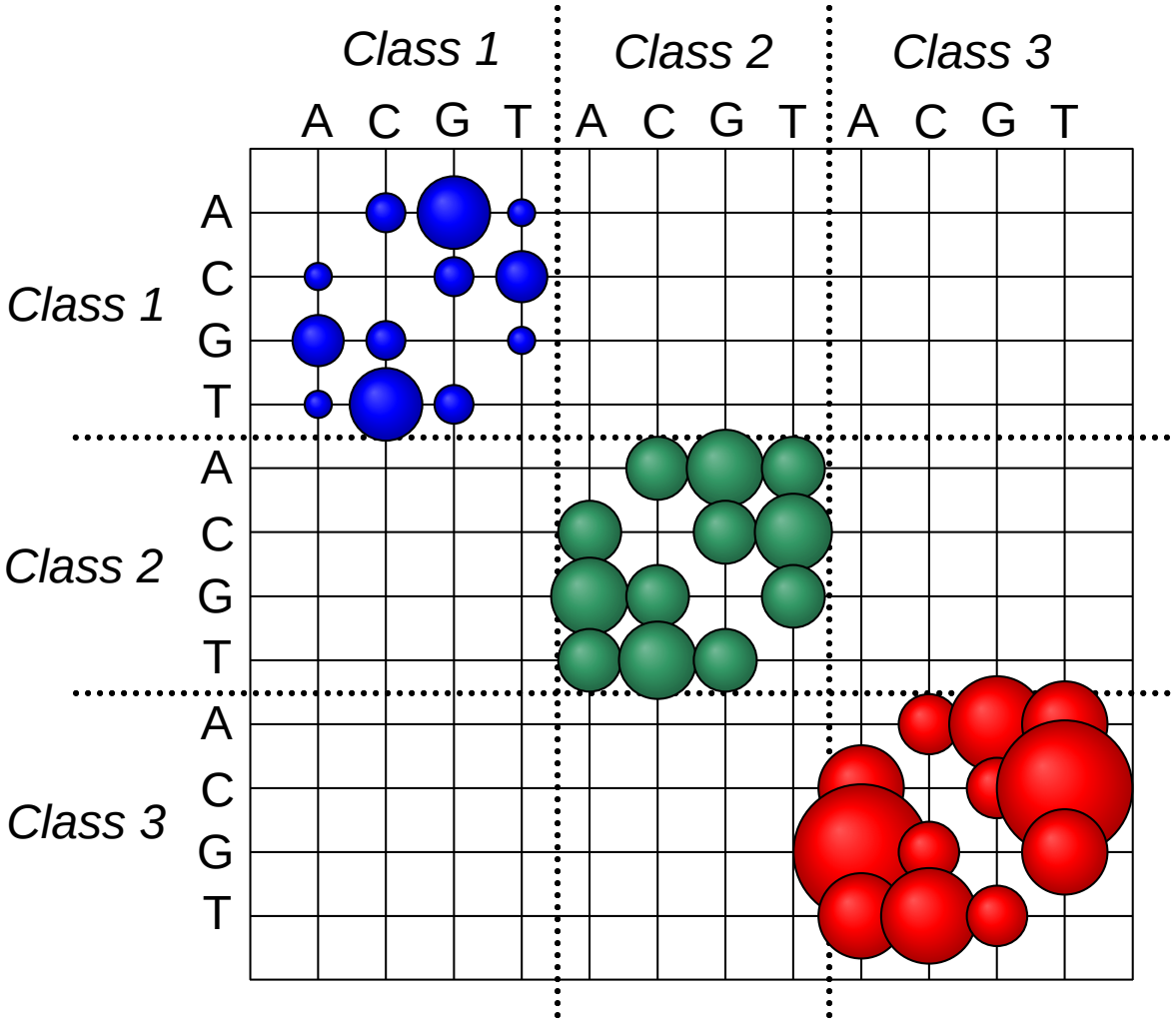




Rate of transitions between hidden classes relative to substitution rate 0.07

Note: Value proportional to bubble area

Mixture models are a special case of THMMs



Restricting all $\rho^{k,l}$ to zero results in a mixture model

Probability of different hidden classes accounted for by the equilibrium distribution at the root

Talk outline

i. *Introduction: what is spatial and temporal heterogeneity?*

ii. *A temporal hidden Markov model of sequence evolution*

iii. *Characterizing heterogeneity in real sequence data*

iv. *Heterogeneity and the genetic code*

Quantifying heterogeneity in sequence evolution

Research questions

How important are different types of heterogeneity in sequence evolution?

Can any factors predict the degree of evolutionary heterogeneity observed?

Experimental design

16 data sets examined

- An alignment from *groEL* (kindly provided by J Herbeck)
- 15 alignments from Pandit
- Trees estimated using Leaphy under GTR+Γ

Use THMM+Γ to investigate different types of heterogeneity

- Spatial heterogeneity in rate accounted for separately
- Maximum likelihood used to estimate all parameters

Quantifying spatial heterogeneity

Mixture model+Γ ($\rho^{k,l} = 0$) compared to HKY+Γ

Investigate relative importance of spatial heterogeneity in:

- Rates (μ^k to vary)
- Frequencies ($\tilde{\pi}^k$ to vary)
- Kappa (κ^k to vary)
- All (everything varies)

	Mixture model classes	
	2	3
Rates	2613.9	3360.0
Frequencies	8140.7	12331.2
Kappa	8307.9	9494.1
All	12567.0	18214.0

NB: ΔAIC of HKY cf. HKY+Γ = 70750.8

Values presented (ΔAIC)

Improvement in model fit relative to HKY+Γ

Summed across all 16 data sets

High values indicate better fit

Conclusions

Strong evidence for all types of spatial heterogeneity (even rate)

Evidence for 2 and 3 classes

Modelling one form of heterogeneity also captures other types

Quantifying temporal heterogeneity

THMM+Γ (1 ρ per data set) compared to HKY+Γ

	Mixture model classes		THMM classes		Temporal improvement	
	2	3	2	3	2	3
Rates	2613.9	3360.0	10025.3	10401.7	7411.4	7041.7
Frequencies	8140.7	12331.2	13029.1	18971.3	4888.4	6640.1
Kappa	8307.9	9494.1	11315.5	12479.2	3007.6	2985.1
All	12567.0	18214.0	18306.1	27587.6	5739.1	9373.6

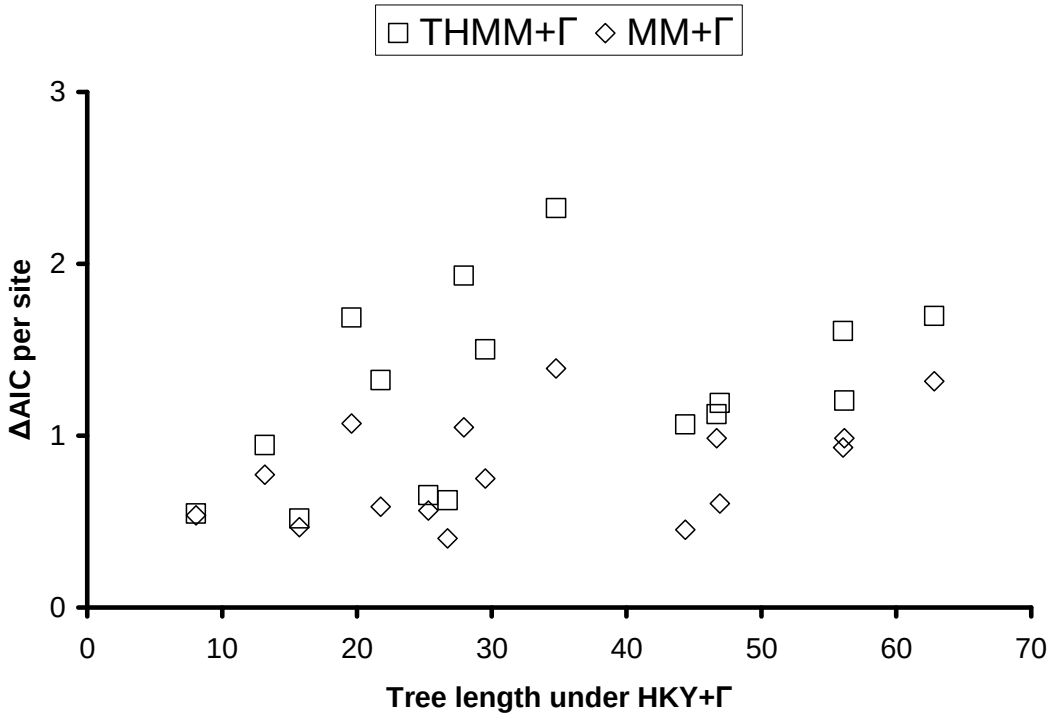
NB: ΔAIC of HKY cf. HKY+Γ = 70750.8

Conclusions

Strong evidence for all types of temporal heterogeneity

Evidence for both 2 and 3 classes

Heterogeneity and evolutionary divergence

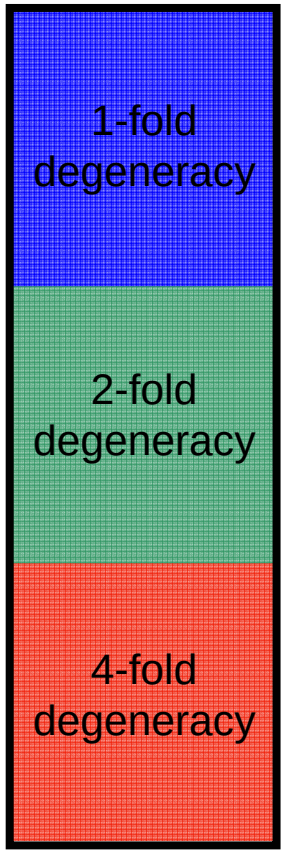


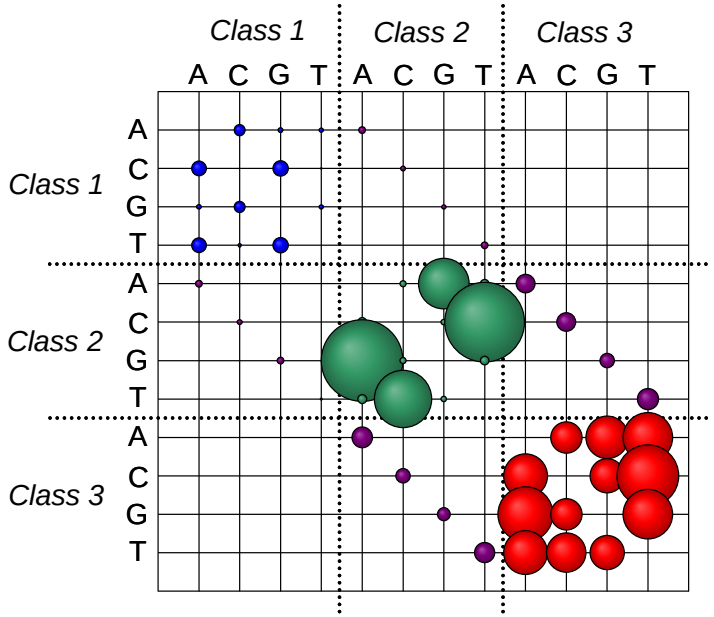
Weak correlation

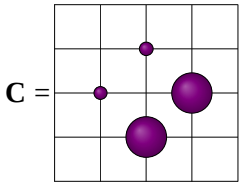
More spatial and temporal heterogeneity in distantly related sequences?

Evolutionary heterogeneity and the genetic code

First two codon positions







Rate of transitions between hidden classes relative to substitution rate

0.07

The effect of the genetic code

Staccato patterns of evolution

Introduces spatial heterogeneity over short times

Temporal heterogeneity over longer time-scales

Talk outline

i. *Introduction: what is spatial and temporal heterogeneity?*

ii. *A temporal hidden Markov model of sequence evolution*

iii. *Characterizing heterogeneity in real sequence data*

iv. *Heterogeneity and the genetic code*

Investigating the genetic code and evolution

Research questions

Does heterogeneity induced by the genetic code affect phylogenetic inference?

If so, to what extent are standard models led astray?

Generalising to other types of heterogeneity

Genetic code introduces complex dependencies in the sequence data

Results indicative of how dependencies affect all phylogenetic inference

Simulating under a codon model (M0 from PAML)

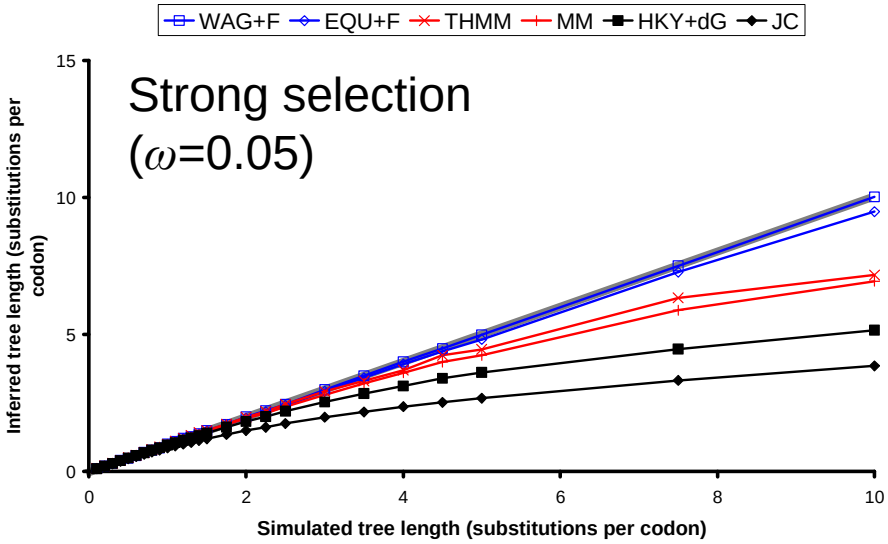
Two selective regimes are particularly interesting:

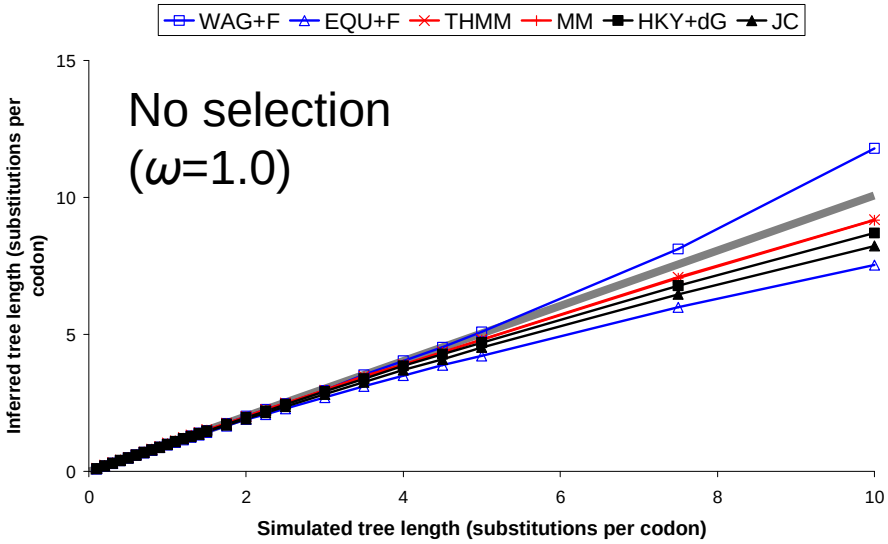
- Strong purifying selection (ω=0.05); High degree of dependency between sites.
- No purifying selection (ω=1.0); Few dependencies between sites

Other model parameters:

- 50 simulations of sequence length 500 for all model conditions
- Transition/transversion rate ratio: κ = 2.5
- Each codon position has different nucleotide composition (F3x4 model)

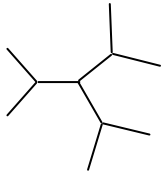
Tree length estimates





Simulation conditions

All branches equal on tree topology



Tree length varies

Models scaled to the same branch length units

Models examined

Standard DNA models:

- JC = All substitutions equally likely
- HKY+Γ = Most common factors included

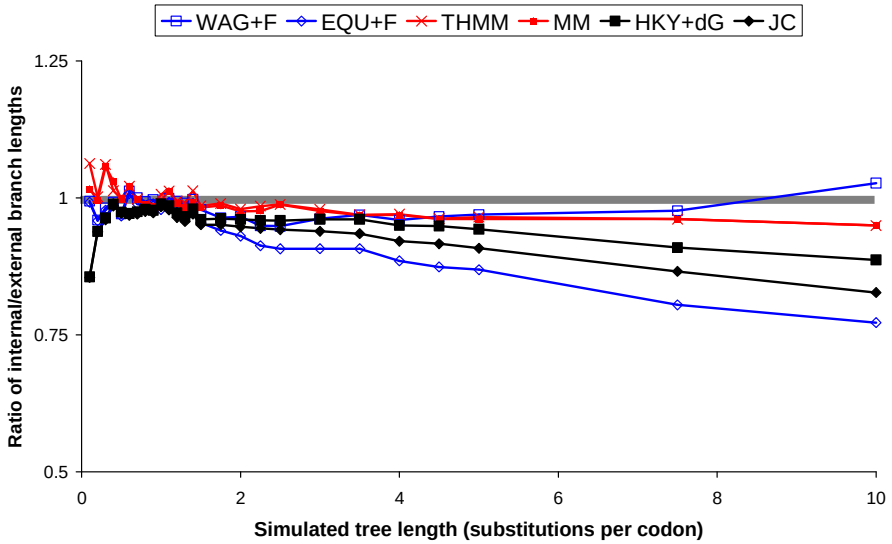
New DNA models:

- MM = Mixture model
- THMM = temporal hidden Markov model

Amino acid models:

- EQU = All substitutions equally likely
- WAG+F+Γ = Most common factors included

Internal and external branch lengths



Strong selection (high dependencies)

Internal branch lengths are underestimated

For divergent sequences this can be extreme

Amino acid models do well

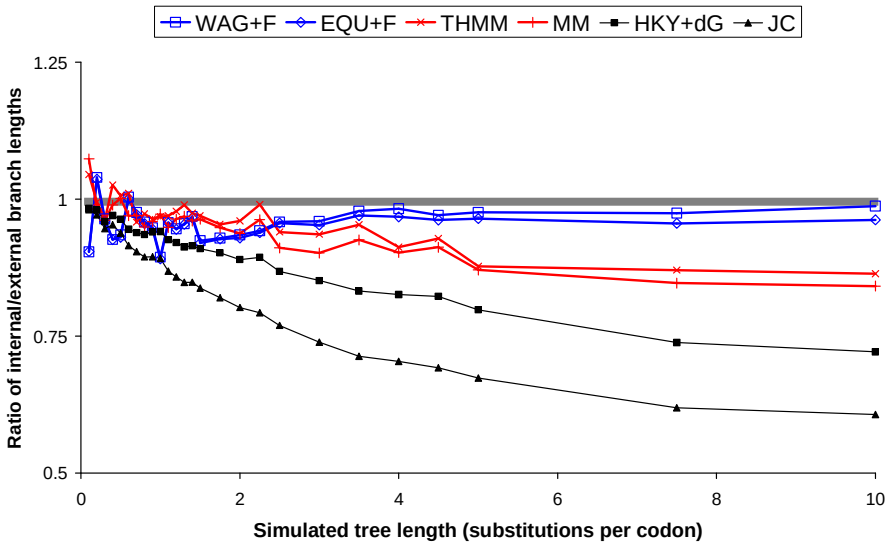
THMM is best nucleotide model, closely followed by MM

No selection (no dependencies)

Better estimates of internal branch lengths

Nucleotide models do well

Variable effects under amino acid models



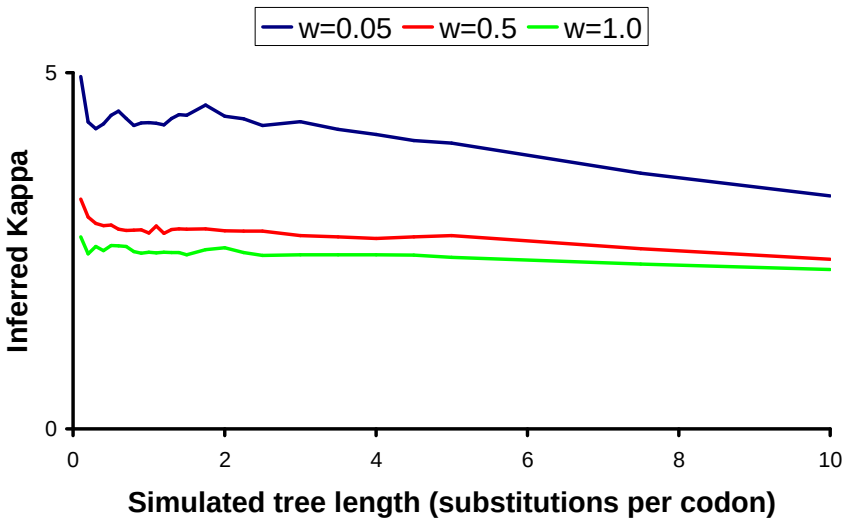
Tree length and parameter estimates

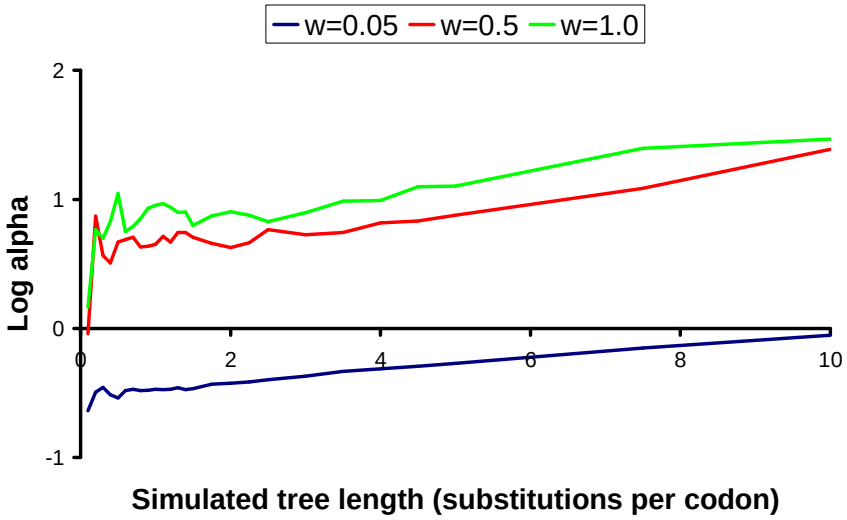
Parameter estimates from HKY+Γ

Evidence of non-Markov behaviour

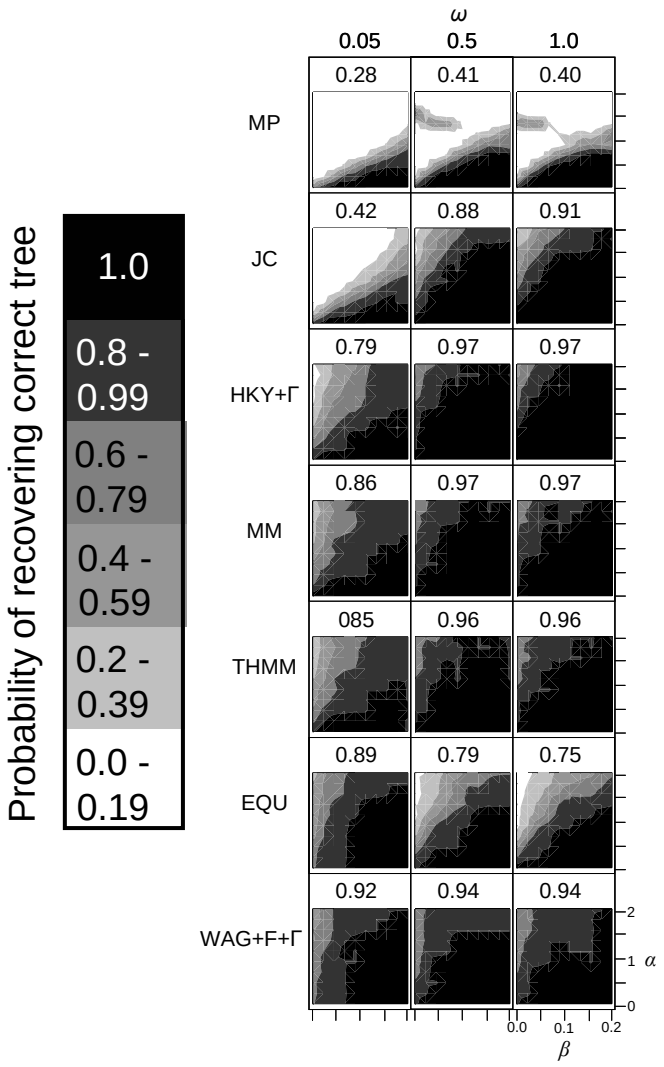
Strongest under strong purifying selection

Dependencies cause evolution to look different over different time-scales?

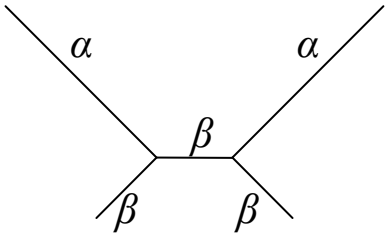




Tree estimation and the genetic code



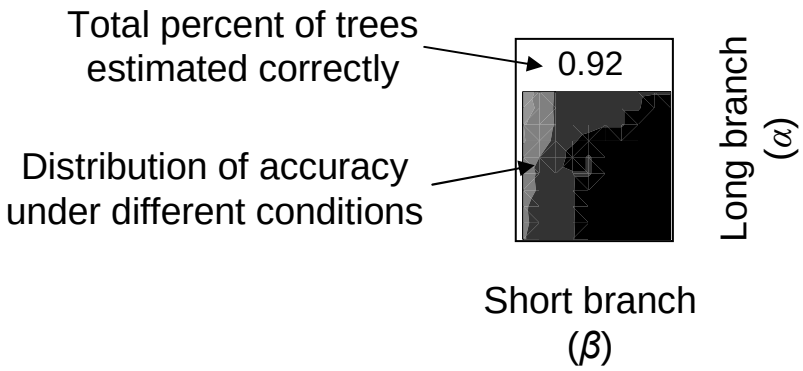
Simulation conditions



α = {0.2, 0.4, …, 2.0}

β = {0.02, 0.04, …, 0.2}

Measuring accuracy of tree estimation



Summary

Temporal hidden Markov models

- Generalisation of mixture models and covarion models
- Provide a generic description of spatial and temporal heterogeneity
- Can estimate evolutionary interesting parameters under difficult conditions

Heterogeneity in coding sequences

- Evolution in coding sequences is complex
- All aspects of nucleotide evolution exhibit temporal and spatial heterogeity
- Partly attributable to the genetic code

Heterogeneity and the genetic code

- The genetic code introduces systematic error to many popular models
- Error affects estimation of internal branches more than external branches
- Recoding data to remove dependencies can improve inference procedures

Simple models for a complex world?

Biology, heterogeneity and modelling sequence evolution

Can be difficult to distinguish and spatial and temporal heterogeneity

- Temporal heterogeneity can look like spatial heterogeneity
- How many classes of spatial heterogeneity are required to describe temporal heterogeneity?
- What biological conclusions can we draw by looking at one without the other?

Different models of evolution for different time scales

- The factors affecting the substitution process may be infinitely complex
- Underestimation increases with evolutionary distance
- Are more complex models required for longer time-scales?

Biologically explicit and generic models of sequence evolution

- Γ-distributed rate heterogeneity models, MMs, and THMMs are generic models
- Codon models and ‘free energy’ models explicitly describe biological phenomena
- What’s the role of different types of model?