



Modelling the evolution of protein coding sequences sampled from measurably evolving populations

Matthew Goode, Stephane Guindon and

Allen Rodrigo

The Bioinformatics Institute (New Zealand),

Department of Statistics, University of Auckland

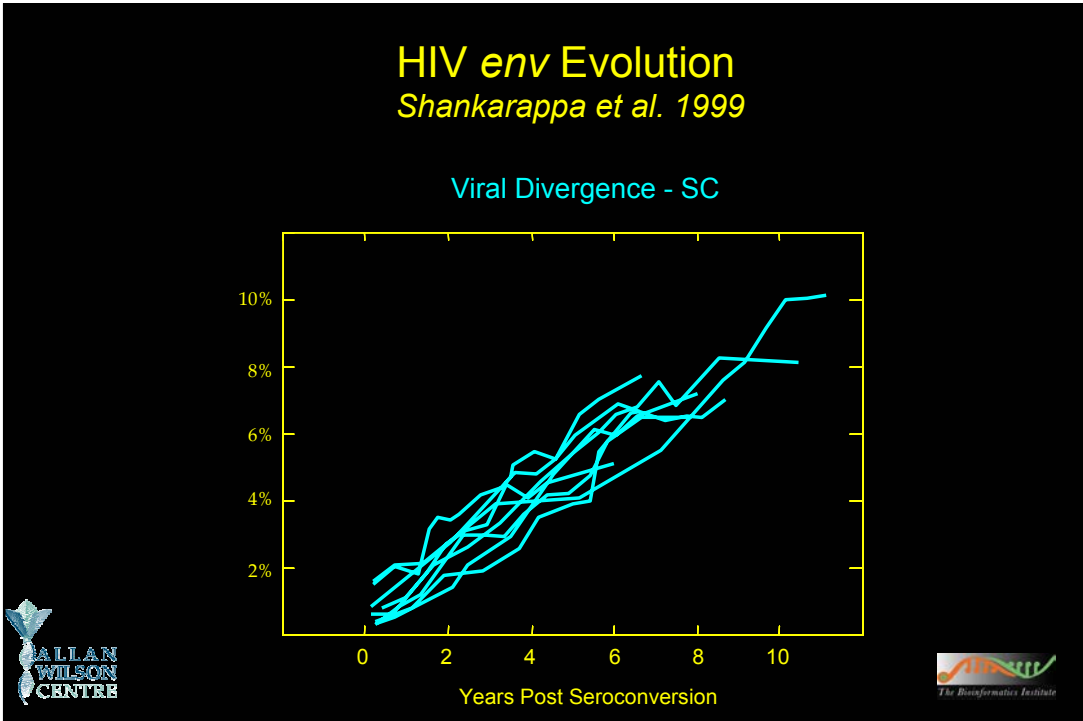
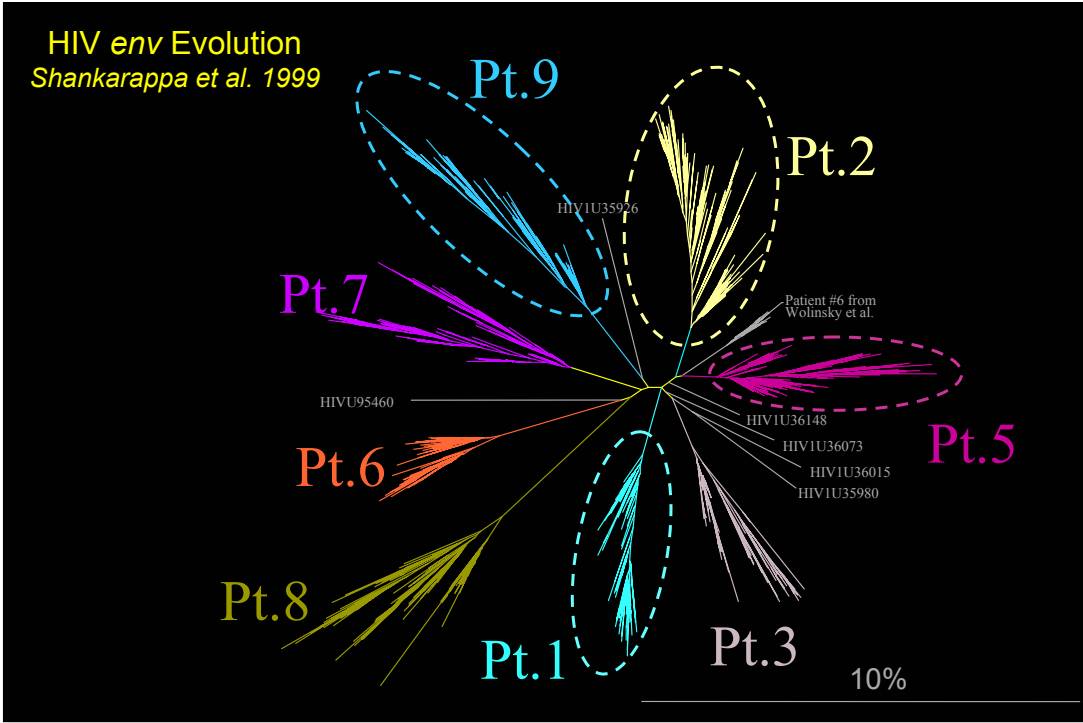
and

Allan Wilson Centre for Molecular Ecology and Evolution

Outline

- Measurably Evolving Populations
- Modelling codon evolution
- Conclusions

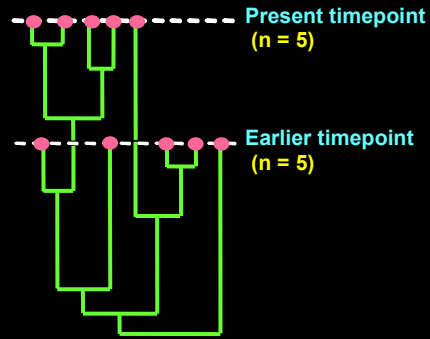




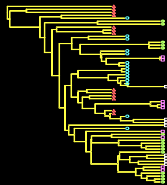
Measurably evolving populations (MEPs)

MEP: Any population evolving fast enough so that a statistically significant accumulation of substitutions between serially sampled sequences can be detected.

- Rapidly evolving pathogens, e.g., .HIV, FIV, Influenza.
- Ancient DNA: so far mostly mitochondrial, e.g.
 - Adelie penguins
 - Pleistocene bears

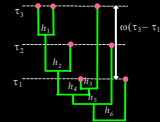


Developments in the Analysis of MEPs



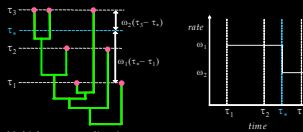
Reconstructing serial genealogies using sUPGMA (Drummond & Rodrigo, 2000)

1: Uniform Rate (TipDate: Rambaut, 2000)



- Estimates uniform rate ω over entire sampling period.
- Strict molecular clock.
- Use ML to optimize branch lengths, estimate parameters h, ω .
- Maximize $L(h, \omega) = P(D | T, h, \omega)$

Multiple Rates with Dated Tips (MRDT) (Drummond et al. 2001)



Multiple rates: sampling times not coincident with rate changes.

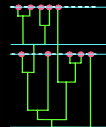
The s-coalescent in a likelihood framework (Rodrigo & Felsenstein, 1998)

• The likelihood function:

$$P(D | N, \tau, \omega) = \sum_{g \in \mathcal{G}} P(D | G, \omega) P(G | N, \tau)$$

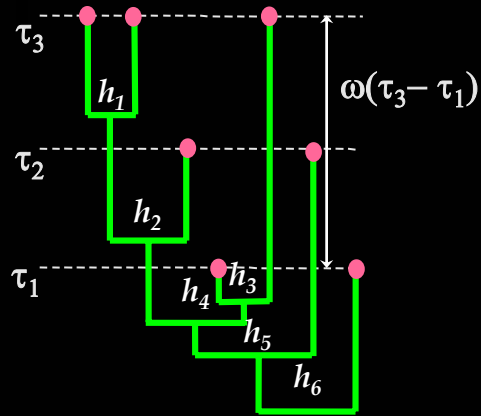
where $\omega = \mu / \tau$

$$P(G | N, \tau) = \prod_{v=2}^n \frac{2}{2N\tau} \exp\left(-\frac{v(v-1)}{2N\tau} \tau_v\right) \times \exp\left(-\frac{(n_1 - c)(n_1 - c - 1)}{2N\tau} (\tau_1 - \tau_2)\right) \times \prod_{v=1}^{c-1} \frac{2}{2N\tau} \exp\left(-\frac{v(v-1)}{2N\tau} \tau_v\right)$$



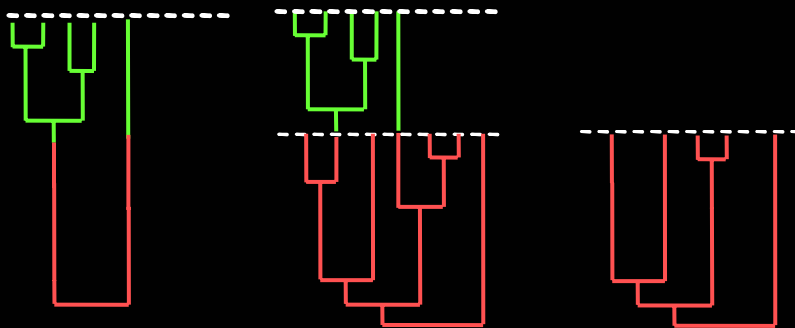
Estimating evolutionary rates

Single Rate with Date Tips (SRDT)



- Estimates uniform rate (ω) over entire sampling period.
- Strict molecular clock.
- Use ML to optimize branch lengths, estimate parameters h, ω .
- Maximise $L(h, \omega) = P(D | T, h, \omega)$;

Measurably evolving populations (MEPs)



Modelling Codon Evolution

- The ratio of the rate of nonsynonymous substitutions (d_N) to the rate of synonymous substitutions (d_S)
- This ratio is symbolised by ω
 - $\omega = d_N/d_S$



Modelling Codon Evolution

- Codons evolve under different selective regimes
- Positive, diversifying selection
 - $d_N > d_S$ $\omega > 1$
- Negative, purifying selection
 - $d_N < d_S$ $\omega < 1$
- Neutrality
 - $d_N = d_S$ $\omega = 1$

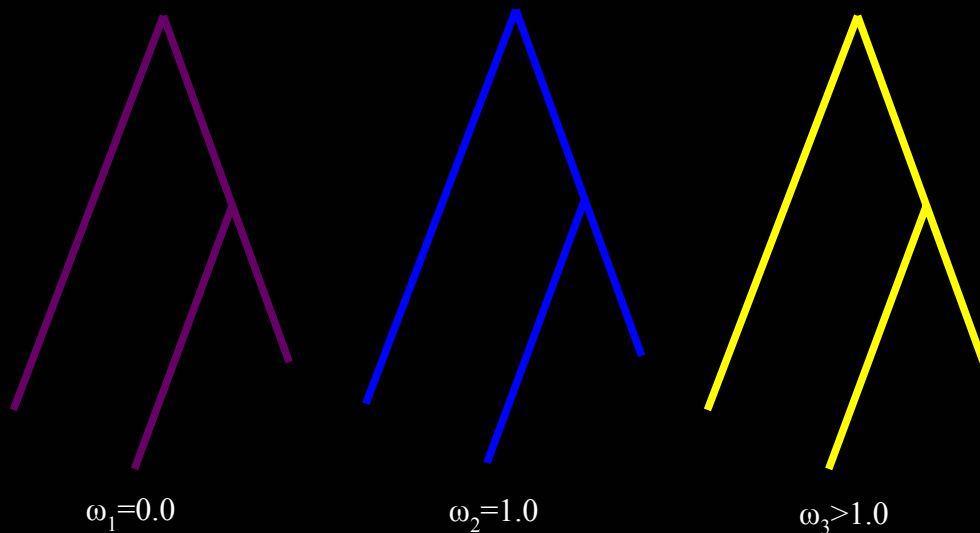


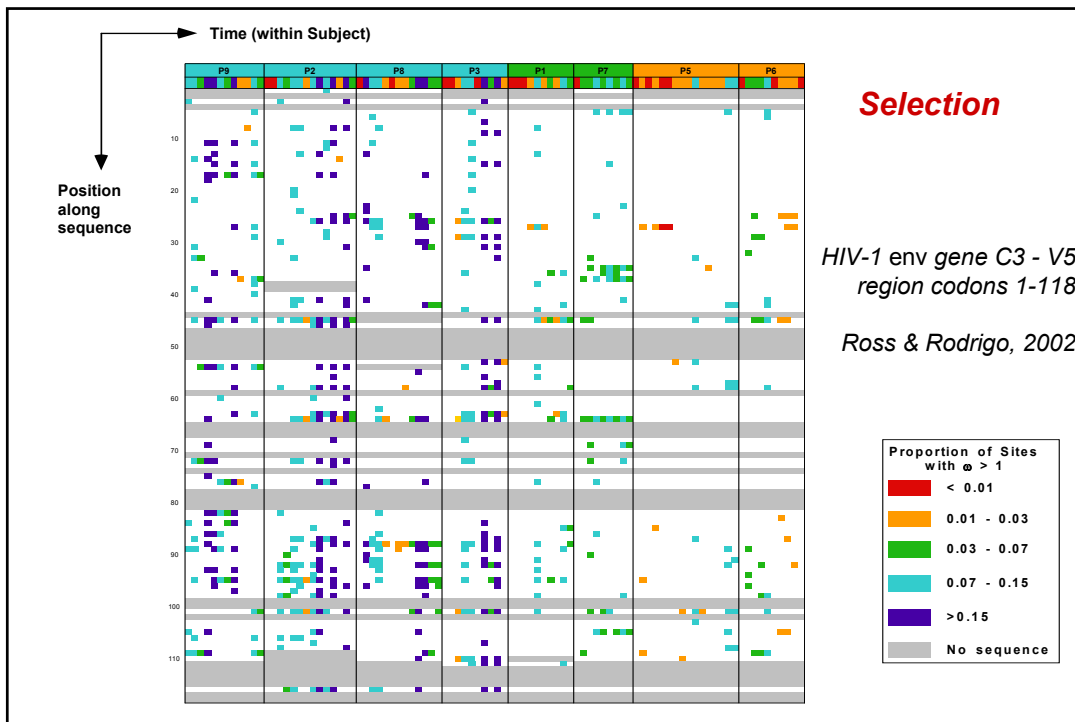
Codon Evolution

- In Nielsen & Yang Codon Model M2, a particular site is assumed to evolve under one, and only one, of the three classes
- With probabilities p_0 , p_1 , p_2 , for $\omega=0$, $\omega=1$, $\omega>1$ respectively.
- Across the tree a site never changes selection class



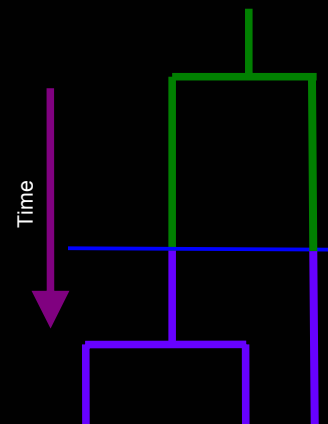
Nielsen & Yang's (1998) model





A new model of codon evolution for serially sampled sequences

- Substitution model changes at an *a priori* specified timepoint.
- A site is allowed to be in different selection classes before and after the split.
- Instantaneous rate matrix and transition probabilities change across split, but still easy to calculate likelihood.



Single split

		After Split		
		Negative	Neutral	Positive
Before Split	Negative	p1	p2	p3
	Neutral	p4	p5	p6
	Positive	p7	p8	p9



		After Split		
		Negative	Neutral	Positive
Before Split	Negative	p_0^2	$p_0 p_1$	$p_0 p_2$
	Neutral	$p_1 p_0$	p_1^2	$p_1 p_2$
	Positive	$p_2 p_0$	$p_2 p_1$	p_2^2

*Estimation &
Hierarchical
Likelihood Ratio
Tests*

Mean
effect

$$p(w_b=i) = p(w_a=i)$$

$$p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j)$$

		After Split (q)		
		Negative	Neutral	Positive
Before Split (p)	Negative	p0q0	p0q1	p0q2
	Neutral	p1q0	p1q1	p1q2
	Positive	p2q0	p2q1	p2q2

Estimation & Hierarchical Likelihood Ratio Tests

Mean effect

$$p(w_b=i) = p(w_a=i)$$

$$p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j)$$

Main effects

$$p(w_b=i) \neq p(w_a=i)$$

$$p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j)$$

		After Split		
		Negative	Neutral	Positive
Before Split	Negative	p1	p2	p3
	Neutral	p4	p5	p6
	Positive	p7	p8	p9

Estimation & Hierarchical Likelihood Ratio Tests

Mean effect

$$p(w_b=i) = p(w_a=i)$$

$$p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j)$$

Main effects

$$p(w_b=i) \neq p(w_a=i)$$

$$p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j)$$

Interaction effects

$$p(w_b=i) \neq p(w_a=i)$$

$$p(w_b=i, w_a=j) \neq p(w_a=i)p(w_b=j)$$

		After Split		
		Negative	Neutral	Positive
Before Split	Negative	p1	p2	p3
	Neutral	p4	p5	p6
	Positive	p7	p8	p9

Estimation & Hierarchical Likelihood Ratio Tests

Mean effect $\left\{ \begin{array}{l} p(w_b=i) = p(w_a=i) \\ p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j) \end{array} \right. \leftarrow \text{Not Nielsen \& Yang!}$

Main effects $\left\{ \begin{array}{l} p(w_b=i) \neq p(w_a=i) \\ p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j) \end{array} \right.$

Interaction effects $\left\{ \begin{array}{l} p(w_b=i) \neq p(w_a=i) \\ p(w_b=i, w_a=j) \neq p(w_a=i)p(w_b=j) \end{array} \right.$

		After Split		
		Negative	Neutral	Positive
Before Split	Negative	p0	0	0
	Neutral	0	p1	0
	Positive	0	0	p2

Estimation & Hierarchical Likelihood Ratio Tests

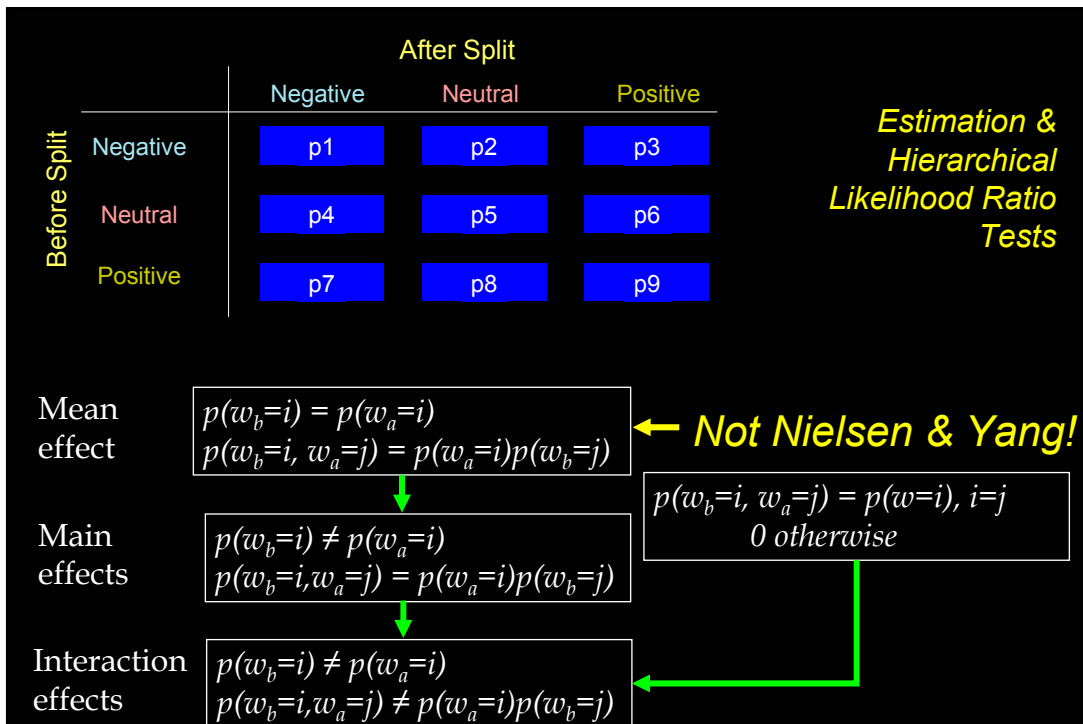
Mean effect $\left\{ \begin{array}{l} p(w_b=i) = p(w_a=i) \\ p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j) \end{array} \right. \leftarrow \text{Not Nielsen \& Yang!}$

Main effects $\left\{ \begin{array}{l} p(w_b=i) \neq p(w_a=i) \\ p(w_b=i, w_a=j) = p(w_a=i)p(w_b=j) \end{array} \right.$

Interaction effects $\left\{ \begin{array}{l} p(w_b=i) \neq p(w_a=i) \\ p(w_b=i, w_a=j) \neq p(w_a=i)p(w_b=j) \end{array} \right.$

$$p(w_b=i, w_a=j) = p(w=i), i=j$$

$$0 \text{ otherwise}$$



An HIV-1 env example

- HIV-1 envelope (*env*) sequences (60 sequences of 660 bases) from infected patient.
- sampled at days **0**, **214**, **671**, **699** and **1005**.
- Monotherapy with zidovudine was initiated after day **409**.

Likelihood Ratio Test

- Estimate parameters (ρ 's, ω 's, and κ 's) using maximum likelihood for NY-M2 and fully saturated model.
 - 8 degrees of freedom difference between models
- ML Estimates
 - Nielsen-Yang M2
 - Log Likelihood -2873.4
 - Saturated Model
 - Log Likelihood -2855.8



Parameter estimates

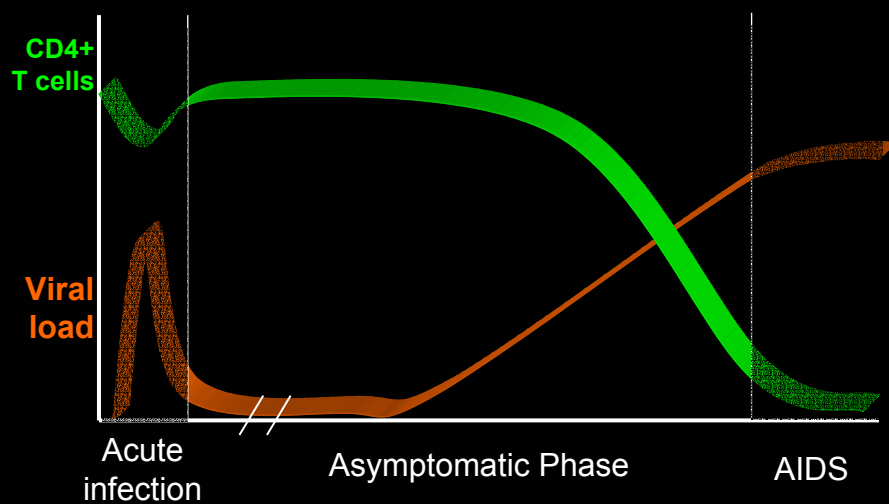
	$\omega_{after}=0$	$\omega_{after}=1$	$\omega_{after}=\infty$	Marginal p (before)
$\omega_{before}=0$	0.425	0.065	0.000	0.490
$\omega_{before}=1$	0.368	0.000	0.000	0.368
$\omega_{before}=7.9$	0.139	0.000	0.003	0.142
Marginal p (after)	0.932	0.065	0.003	1.000

Future work

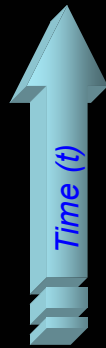
- ❑ To date, most methods on changing evolutionary parameters deal with lineage-independent changes.
- ❑ This is suitable for species (and higher taxa) phylogenies.
 - ❑ Forces that influence rates of evolution may act differently in different lineages
- ❑ Not necessarily suitable for intraspecific phylogenies.
 - ❑ External influences act on the population as a whole.
 - ❑ Also true for some taxonomic phylogenies



Progression of HIV Infection



Changing models as a function of time



Model of
evolution, $\mathbf{Q}(t)$ or
mutation rate, $\mu(t)$

Changing models of evolution as a function of time (commutable models)

Commutable models of evolution $\mathbf{Q}(t) \times \mathbf{Q}(t') = \mathbf{Q}(t') \times \mathbf{Q}(t)$

If \mathbf{Q} changes as a function of time, we can calculate the transition probabilities as:

$$\mathbf{P}_N(T) = e^{\int \mathbf{Q}(t) dt}$$

*Rodrigo et al. (2008)
Phil Trans Roy Soc B*

Conclusions

- We have developed a codon model of evolution that permits:
 - Changes to the ratio of non-synonymous to synonymous substitution rates over time.
 - Different proportions of sites in each selective class.
- The model is based on a simultaneous change in rate across all lineages.
 - Consequently, it is better for intraspecific phylogenies than interspecific phylogenies.



Acknowledgements

- David Bryant
- Alexei Drummond
- Joseph Heled
- Howard Ross

