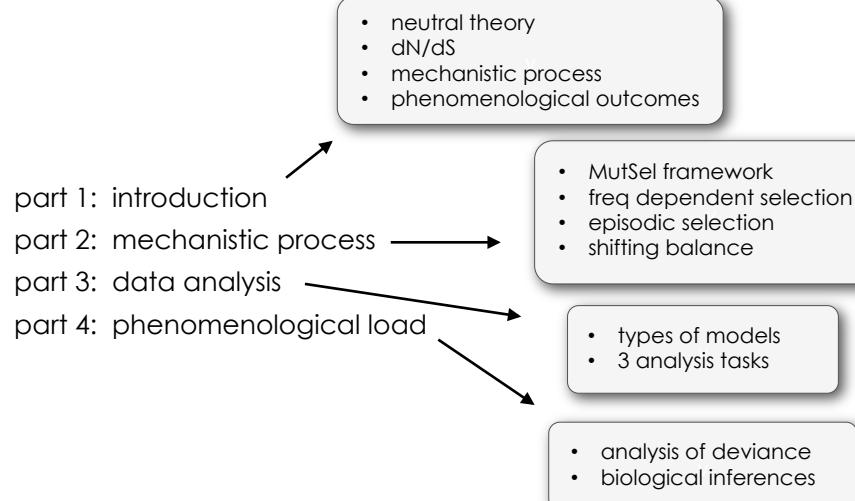


codon substitution models and the analysis of natural selection pressure

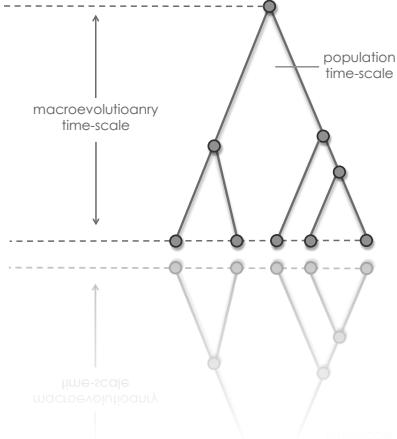
Joseph P. Bielawski
Department of Biology
Department of Mathematics & Statistics
Dalhousie University



The goals and the plan



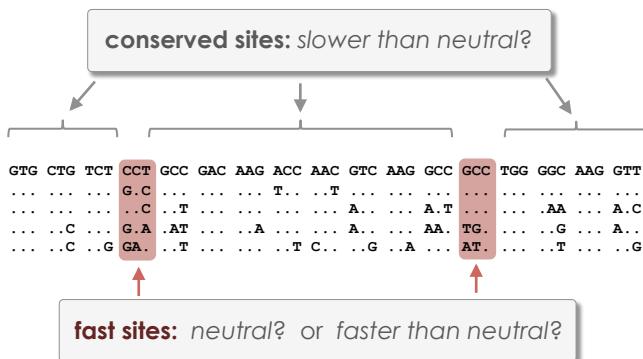
part 1: introduction



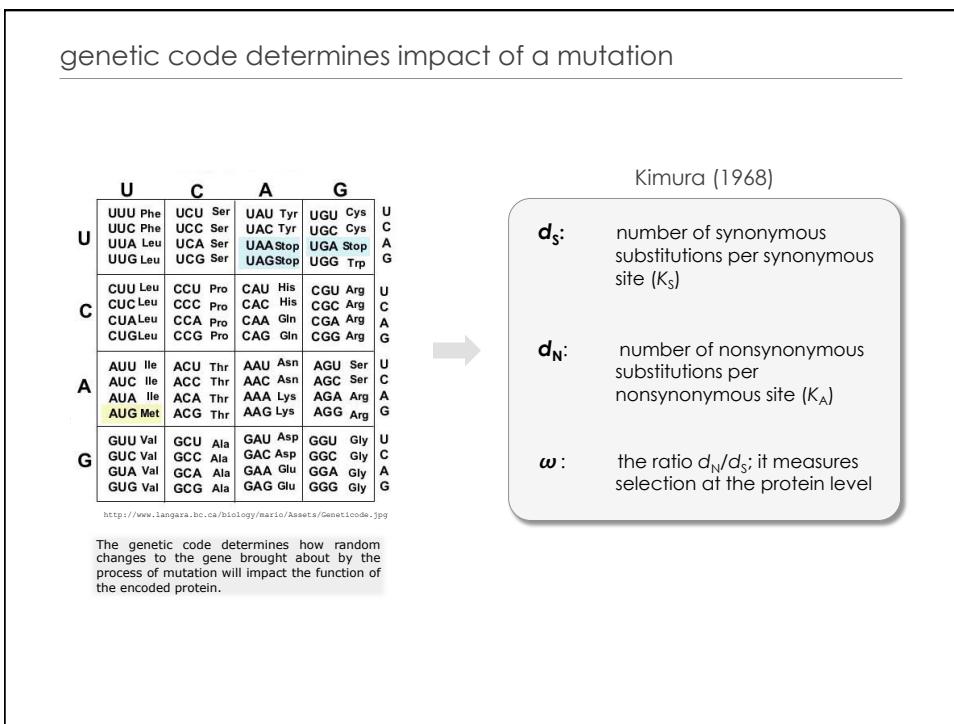
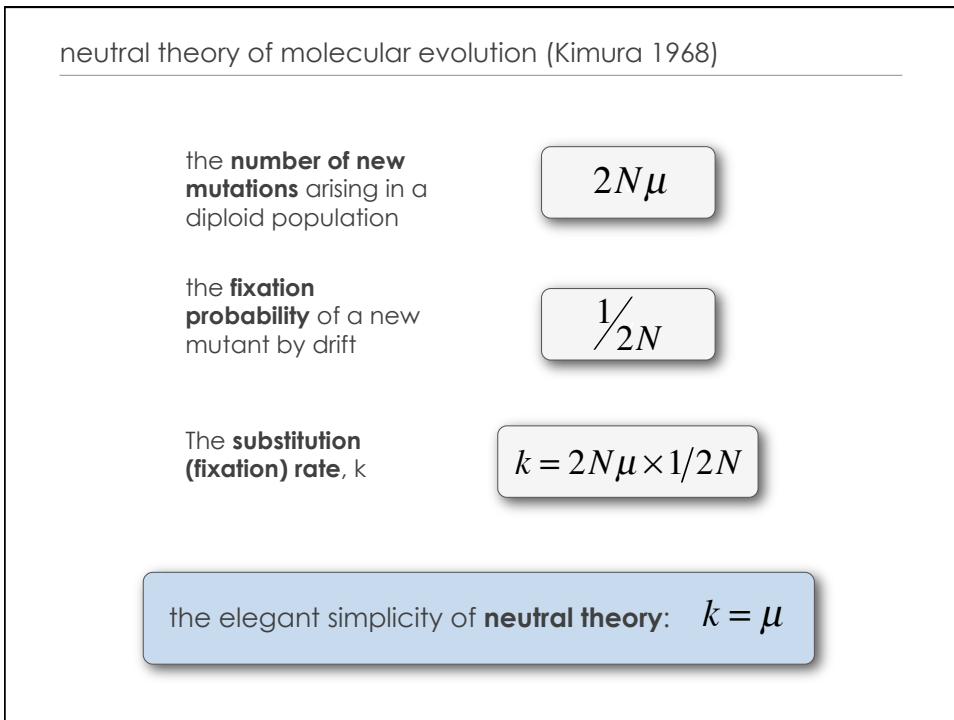
evolutionary rate depends on intensity of selection

selectively constrained = slower than neutral (drift alone)

adaptive divergence = faster than neutral (drift alone)



What is the neutral expectation?



an index of selection pressure

rate ratio	mode	example
$dN/dS < 1$	purifying (negative) selection	histones
$dN/dS = 1$	Neutral Evolution	pseudogenes
$dN/dS > 1$	Diversifying (positive) selection	MHC, Lysin

an index of selection pressure

Why use d_N and d_S ? (Why not use raw counts?)

example of counts:

- 300 codon gene from a pair of species
- 5 synonymous differences
- 5 nonsynonymous differences

$$5/5 = 1$$

why don't we conclude that rates are equal (i.e., neutral evolution)?

the genetic code & mutational opportunities

Relative proportion of different types of mutations in hypothetical protein coding sequence.				
Type	Expected number of changes (proportion)			
	All 3 Positions	1 st positions	2 nd positions	3 rd positions
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
Nonsynonymous	392 (71)	166 (91)	176 (96)	57 (27)
nonsense	23 (4)	9 (5)	7 (4)	7 (4)

Modified from Li and Graur (1991). Note that we assume a hypothetical model where all codons are used equally and that all types of point mutations are equally likely.

Why do we use d_N and d_S ?

same example, but using d_N and d_S :

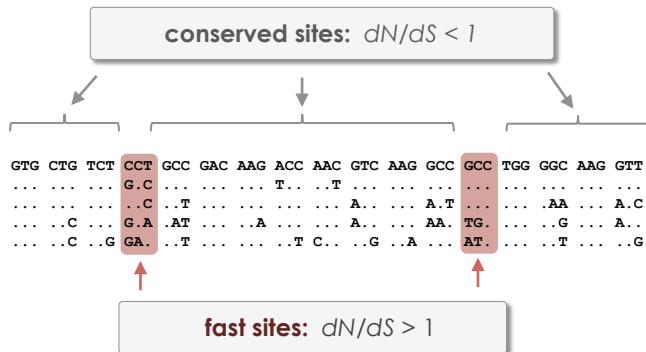
$$\text{Synonymous sites} = 25.5\% \\ S = 300 \times 3 \times 25.5\% = 229.5$$

$$\text{Nonsynonymous sites} = 74.5\% \\ N = 300 \times 3 \times 74.5\% = 670.5$$

$$\text{So, } d_S = 5/229.5 = 0.0218 \\ d_N = 5/670.5 = 0.0075$$

$$d_N/d_S (\omega) = 0.34, \text{ purifying selection !!!}$$

an index of selection pressure acting on the protein



conclusion: dN differs from dS due to the effect of selection on the protein.

mutational opportunity vs. physical site

Relative proportion of different types of mutations in hypothetical protein coding sequence.

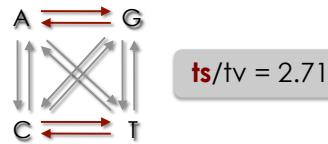
Type	Expected number of changes (proportion)			
	All 3 Positions	1 st positions	2 nd positions	3 rd positions
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
Nonsynonymous	392 (71)	166 (91)	176 (96)	57 (27)
nonsense	23 (4)	9 (5)	7 (4)	7 (4)

Note that by framing the counting of sites in this way we are using a "mutational opportunity" definition of the sites. Thus, a synonymous or non-synonymous site is not considered a physical entity!

Note that we assume a hypothetical model where all codons are used equally and that all types of point mutations are equally likely.

real data have biases (*Drosophila GstD1* gene)

transitions vs. transversions:



preferred vs. un-preferred codons:

partial codon usage table for the *GstD* gene of *Drosophila*

Phe F TTT	0	Ser S TCT	0	Tyr Y TAT	1	Cys C TGT	0
TTC	27	TCC	15	TAC	22	TGC	6
Leu L TTA	0	TCA	0 *** * TAA	TAA	0 *** * TGA	TGA	0
TTG	1	TCG	1 TAG	TAG	0 Trp W TGG	TGG	8
Leu L CTT	2	Pro P CCT	1 His H CAT	CAT	0 Arg R CGT	CGT	1
CTC	2	CCC	15	CAC	4 CGC	7	
CTA	0	CCA	3 Gln Q CAA	CAA	0 CGA	CGA	0
CTG	29	CCG	1 CAG	CAG	14 CGG	CGG	0

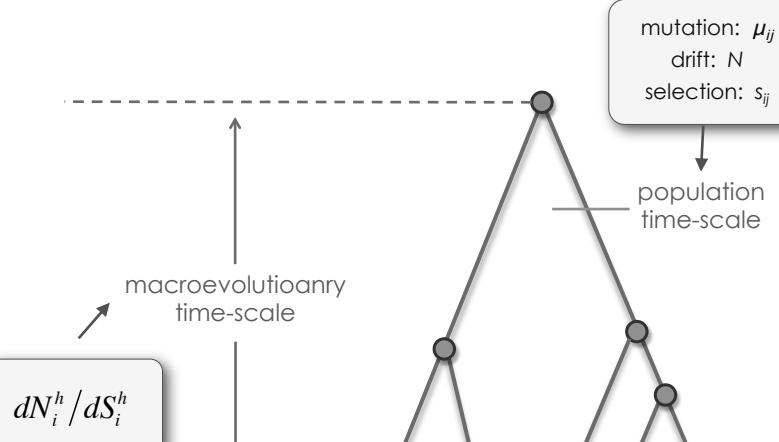
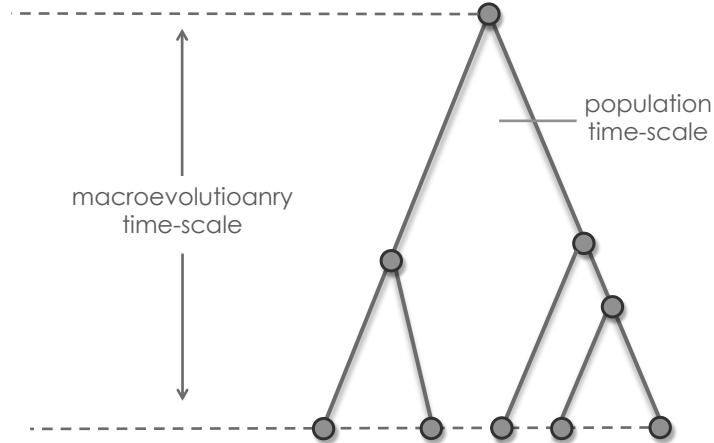
an index of selection pressure acting on the protein

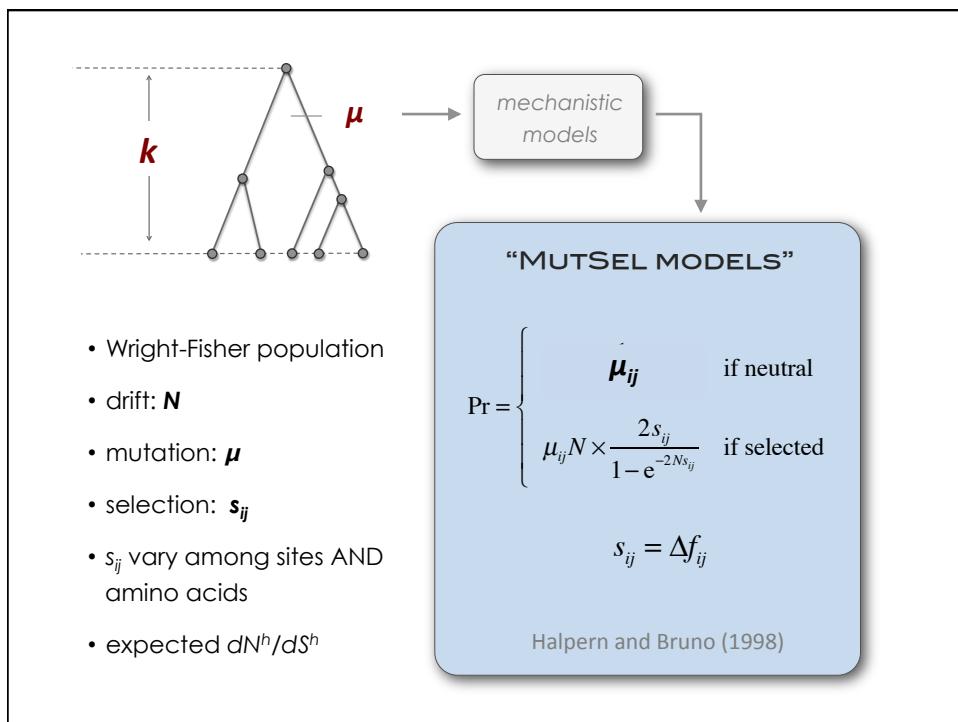
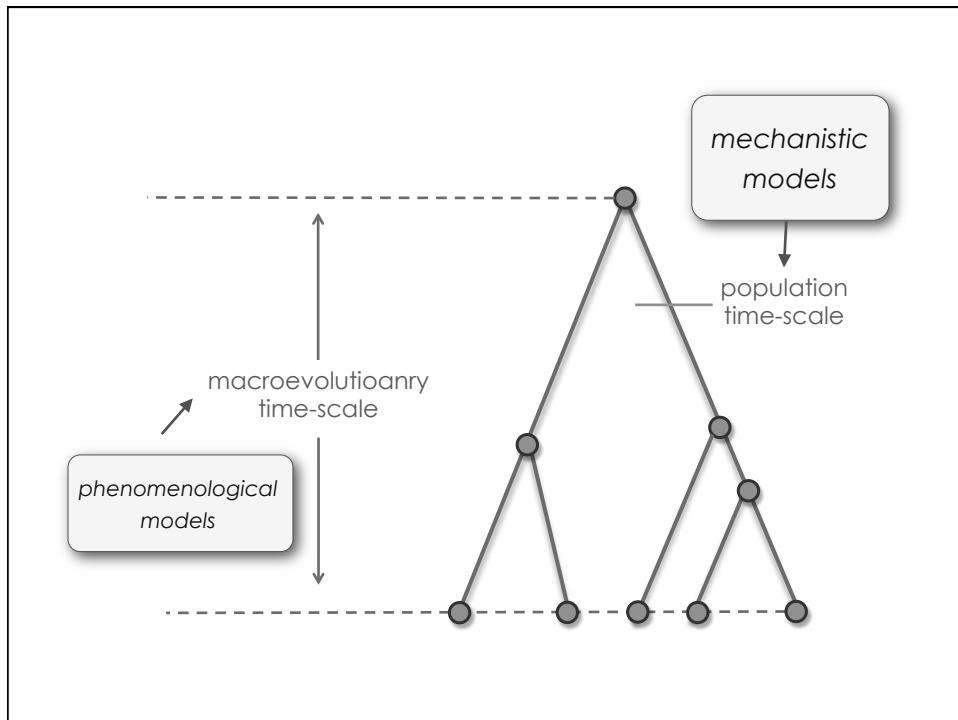
$$\omega = \frac{dN}{dS}$$

Don't worry: we will improve upon the counting method later in this lecture via likelihood!

correcting dS and dN for underlying mutational process of the DNA makes them **sensitive to assumptions about the process of evolution!**

reconciling evolutionary time scales





fixation probability with selection

population genetics at a single codon site (h)

fitness coefficients

$$f^h = \langle f_1, \dots, f_{61} \rangle$$

selection coefficients

$$s_{ij}^h = f_j^h - f_i^h$$

fixation probability (Kimura, 1962)

$$\Pr(s_{ij}^h) = \frac{2s_{ij}^h}{1 - e^{-2Ns_{ij}^h}}$$

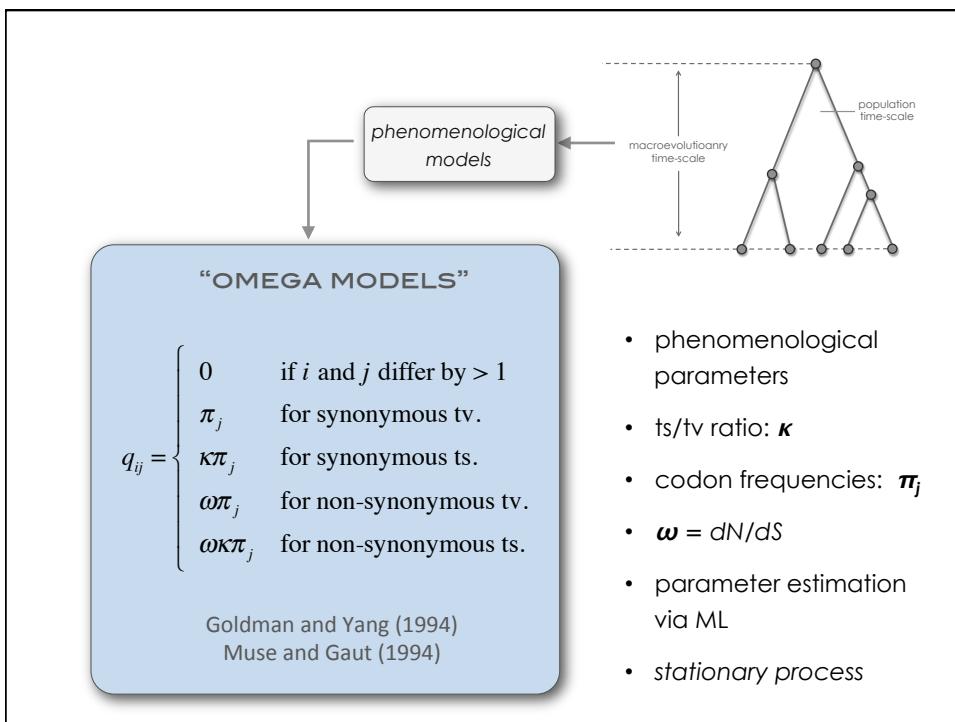
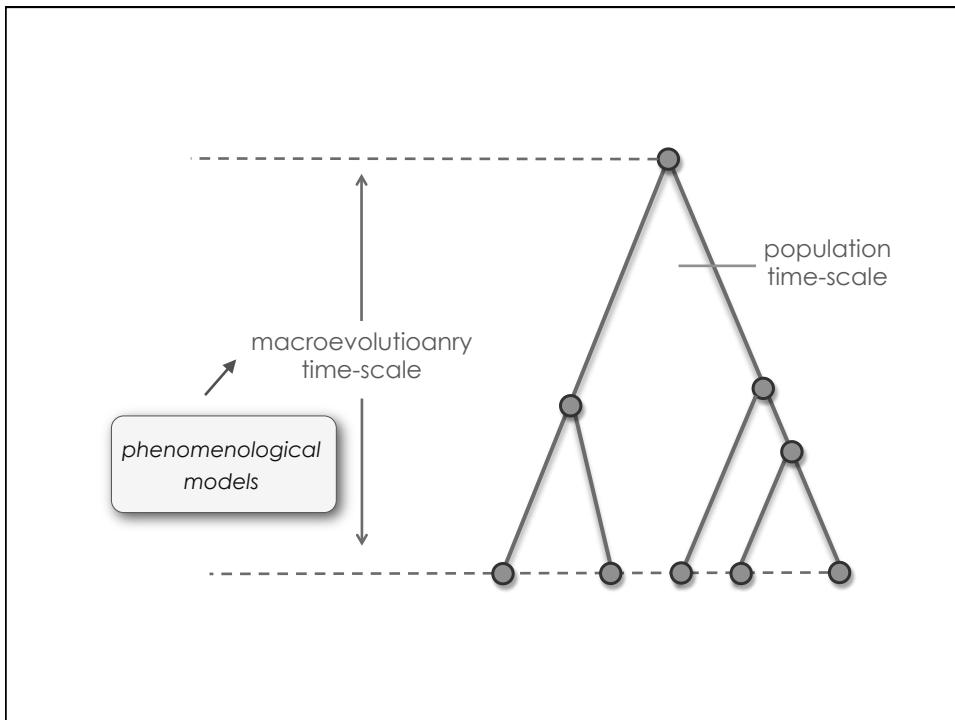
fixation probability with selection

MutSel: selection favours amino acids with higher fitness (if N is large enough)



realism: fitness expected to differ among sites and amino acids according to protein function

the cost of realism: too complex to fit such a model to real data (but simplified versions will allow new ways of data analysis)



the instantaneous rate matrix, Q , is very big: 61×61

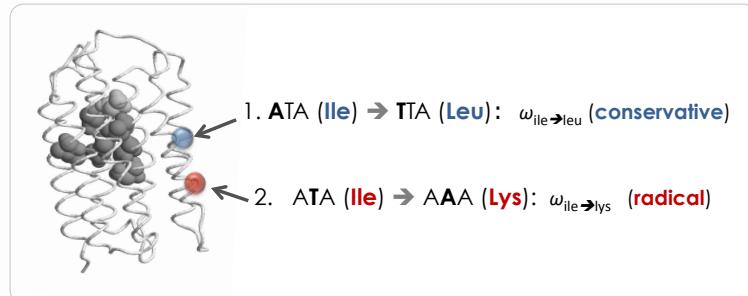
phenomenological codon models: just a few parameters are needed to cover the 3721 transitions between codons!

From codon below:	to codon below:							
	TTT (Phe)	TTC (Phe)	TTA (Leu)	TTG (Leu)	CTT (Leu)	CTC (Leu)	...	GGG (Gly)
TTT (Phe)	---	$\kappa\pi_{TTC}$	$\omega\pi_{TTA}$	$\omega\pi_{TTG}$	$\omega\kappa\pi_{TTT}$	0	...	0
TTC (Phe)	$\kappa\pi_{TTT}$	---	$\omega\pi_{TTA}$	$\omega\pi_{TTG}$	0	$\omega\kappa\pi_{CTC}$...	0
TTA (Leu)	$\omega\pi_{TTT}$	$\omega\pi_{TTC}$	---		0	0	...	0
TTG (Leu)	$\omega\pi_{TTT}$	$\omega\pi_{TTC}$	$\kappa\pi_{TTA}$	---	0	0	...	0
CTT (Leu)	$\omega\kappa\pi_{TTT}$	0	0	0	---	$\kappa\pi_{CTC}$...	0
CTC (Leu)	0	$\omega\kappa\pi_{TTC}$	0	0	$\kappa\pi_{TTT}$	---	...	0
...
GGG (Gly)	0	0	0	0	0	0	0	---

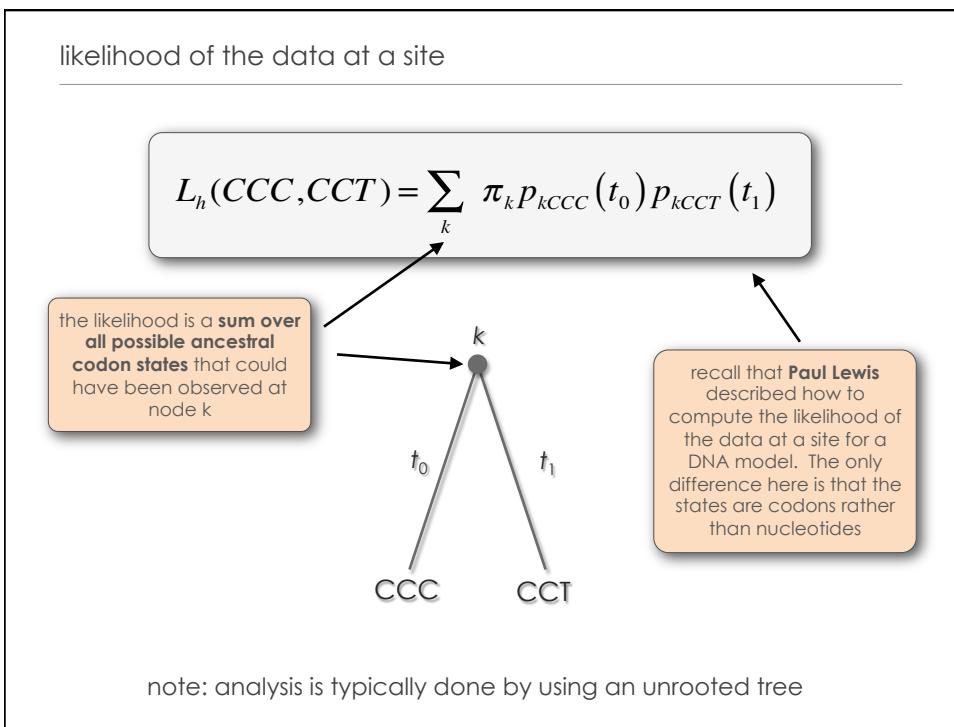
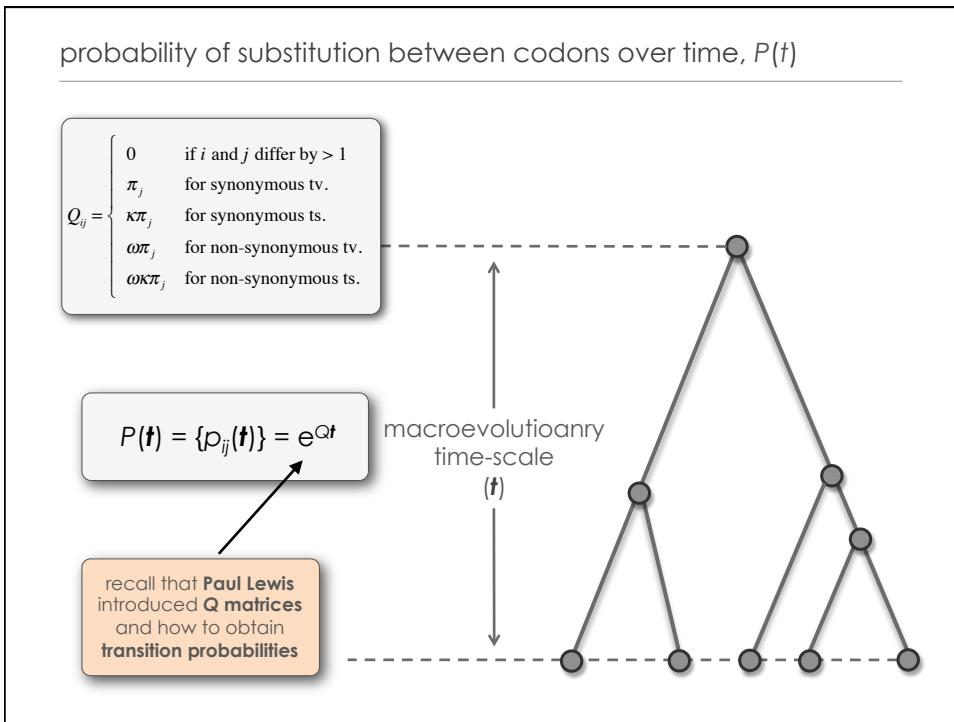
* This is equivalent to the codon model of Goldman and Yang (1994). Parameter ω is the ratio d_N/d_S , κ is the transition/transversion rate ratio, and π_i is the equilibrium frequency of the target codon (i).

substitution probability with selection

intentional simplification: all amino acid substitutions have the same ω !



contradiction? selection should favour amino acids with higher fitness.



likelihood of the data at all sites

The likelihood of observing the entire sequence alignment is the product of the probabilities at each site.

Paul Lewis
covered this with
the "AND" rule in
his likelihood
lecture

$$L = L_1 \times L_2 \times L_3 \times \dots \times L_N = \prod_{h=1}^N L_h$$

see Paul Lewis's
lecture slides for
more about
likelihoods vs. log-
likelihoods

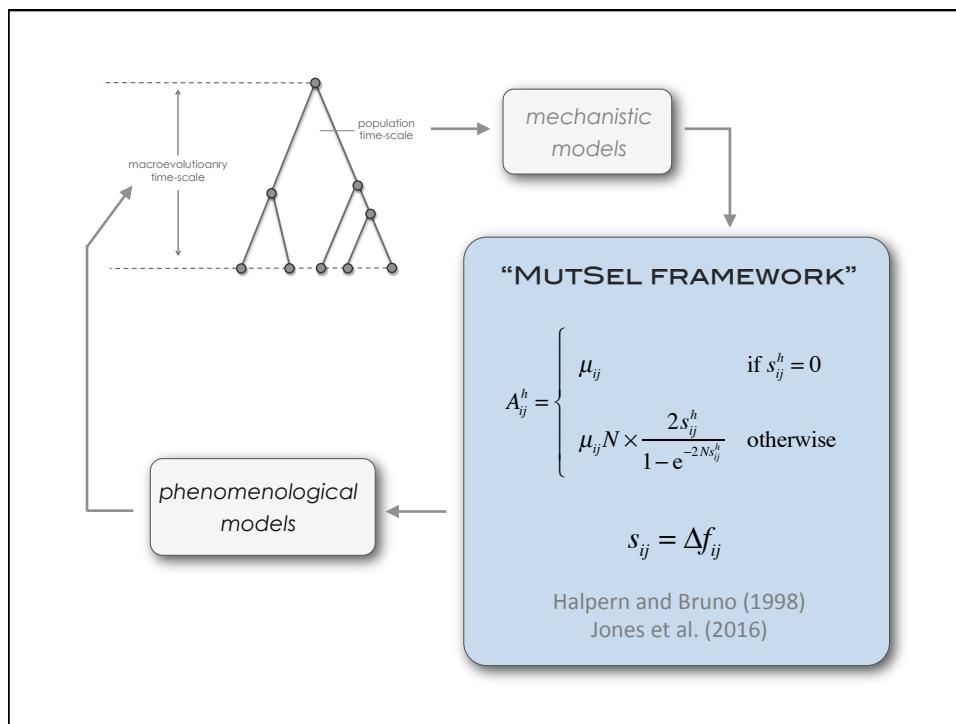
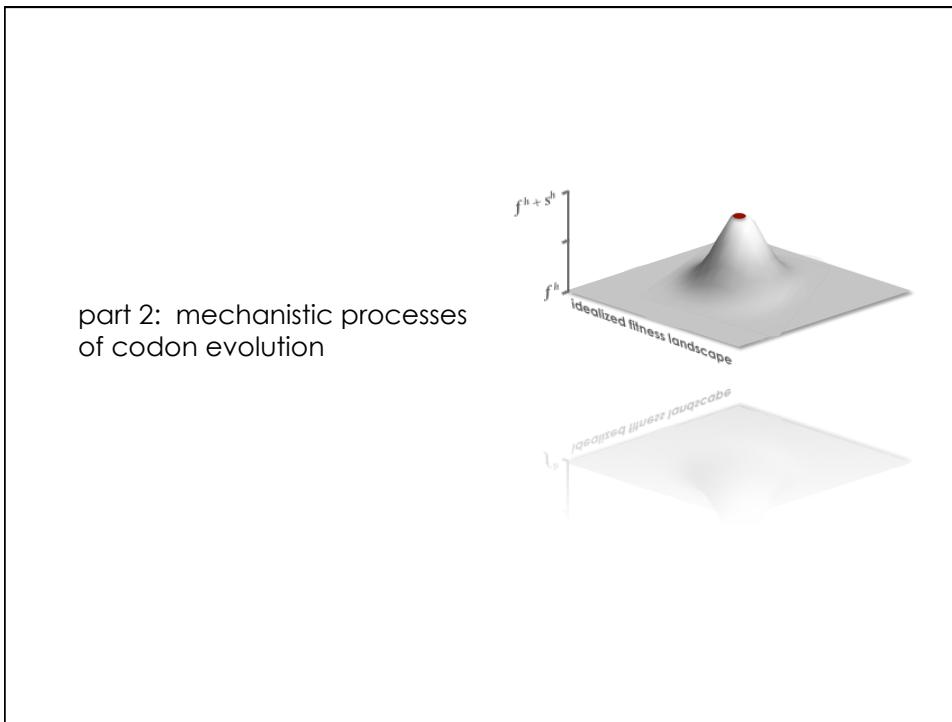
The log likelihood is a sum over all sites.

$$\ell = \ln\{L\} = \ln\{L_1\} + \ln\{L_2\} + \ln\{L_3\} + \dots + \ln\{L_N\} = \sum_{h=1}^N \ln\{L_h\}$$

we made some progress...

1. we are now being explicit about **phenomenological and mechanistic models**
2. we are more **cautious about mechanistic interpretation** of phenomenological parameters
3. we have learned how to **connect evolutionary mechanisms to the substitution process**
4. we introduced the *idea* that we can **compute expectations** from mechanistic parameters

Lets look at some mechanism of evolution and "see" what we should expect!



site-specific MutSel rate matrix

MUTSEL RATE MATRIX

$$A_{ij}^h = \begin{cases} \mu_{ij} & \text{if } s_{ij}^h = 0 \\ \mu_{ij}N \times \frac{2s_{ij}^h}{1 - e^{-2Ns_{ij}^h}} & \text{otherwise} \end{cases}$$

- MutSel time-scale is infinitesimal compared to substitution scale
- MutSel probabilities approximate the instantaneous site-specific rate matrix, A
- μ_{ij} = nucleotide GTR process (before the effect of selection)

site-specific MutSel rate matrix

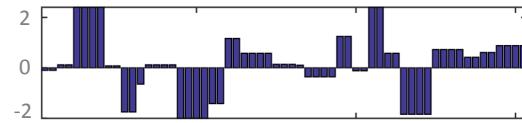
two explicit ways to reconcile **population genetics** and **macroevolution**:

1. map fitness to equilibrium frequencies
2. macroevolution index of selection intensity

1. fitness coefficients map to stationary codon frequencies

fitness
coefficients

$$f^h = \langle f_1, \dots, f_{61} \rangle$$



codon
frequencies

$$\pi^h = \langle \pi_1, \dots, \pi_{61} \rangle$$



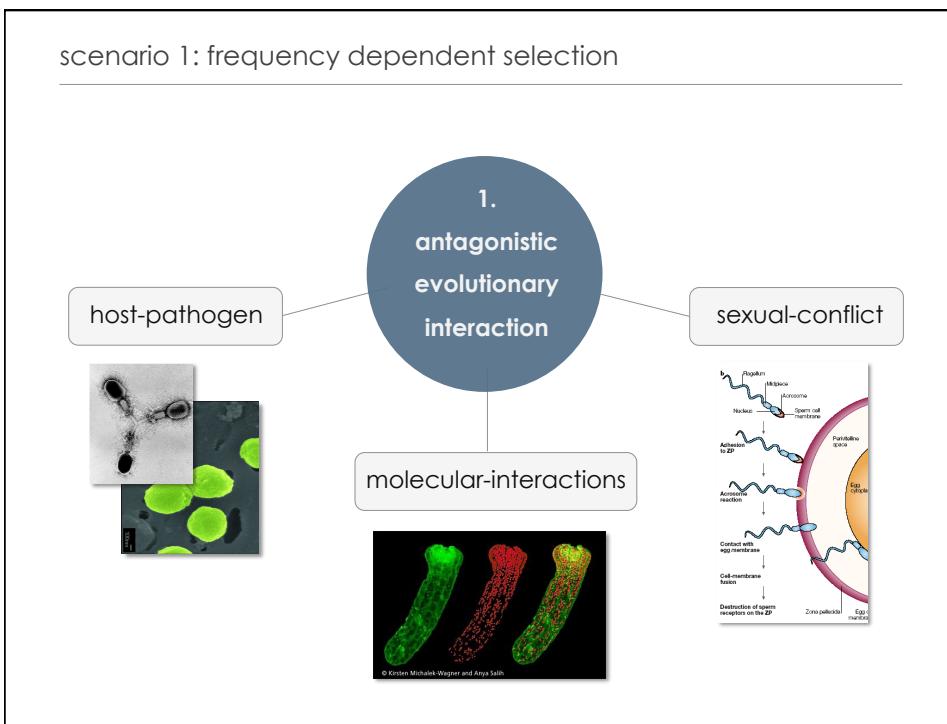
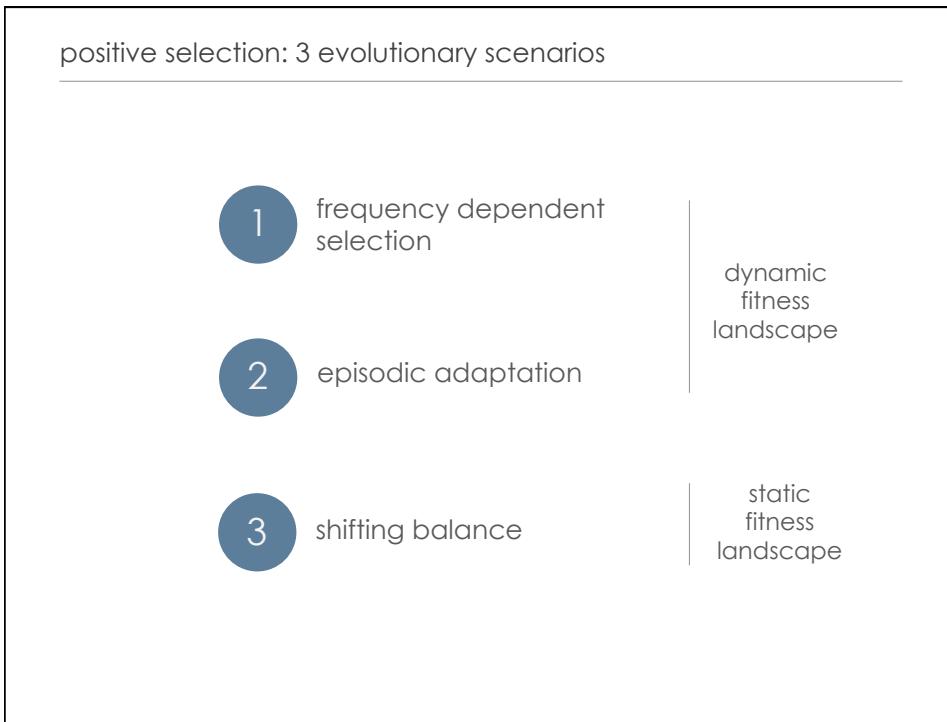
2. from fitness coefficients to dN/dS

MUTSEL RATE MATRIX

$$dN^h / dS^h = \frac{E[\text{evolution w/ selection}]}{E[\text{evolution by drift alone}]}$$

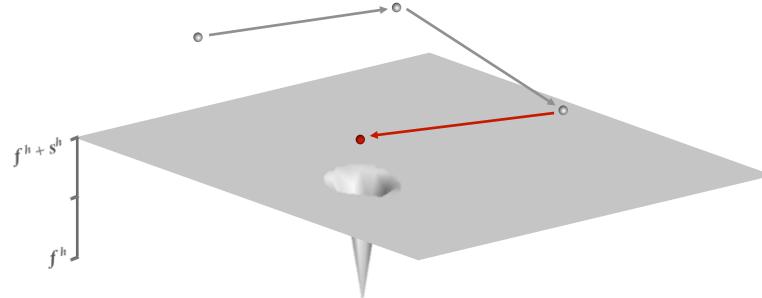
$$dN^h / dS^h = \frac{\sum_{i \neq j} \pi_i^h A_{ij}^h I_N}{\sum_{i \neq j} \pi_i^h \mu_{ij} I_N}$$

- $dN/dS = \omega$ when matrix A^h is replaced by matrix Q of model M0
- dN/dS is an analog of ω under MutSel



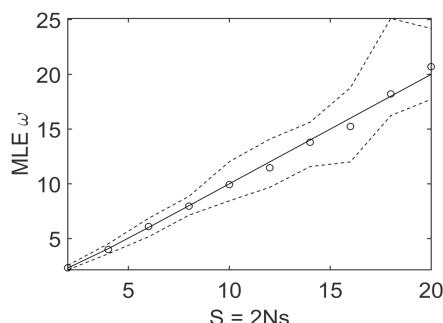
frequency-dependent selection: MutSelM0

1. amino acid at a site has f^h ; all others have $f^h + s$
2. fitness values swap when a substitution occurs



MutSelM0: (1) and (2) above imply Markov chain properties with the same rate matrix \mathcal{Q} as **codon model M0**

frequency-dependent selection: MutSelM0



generating process:

MutSelM0

expectation = dN^h/dS^h

symbol = —

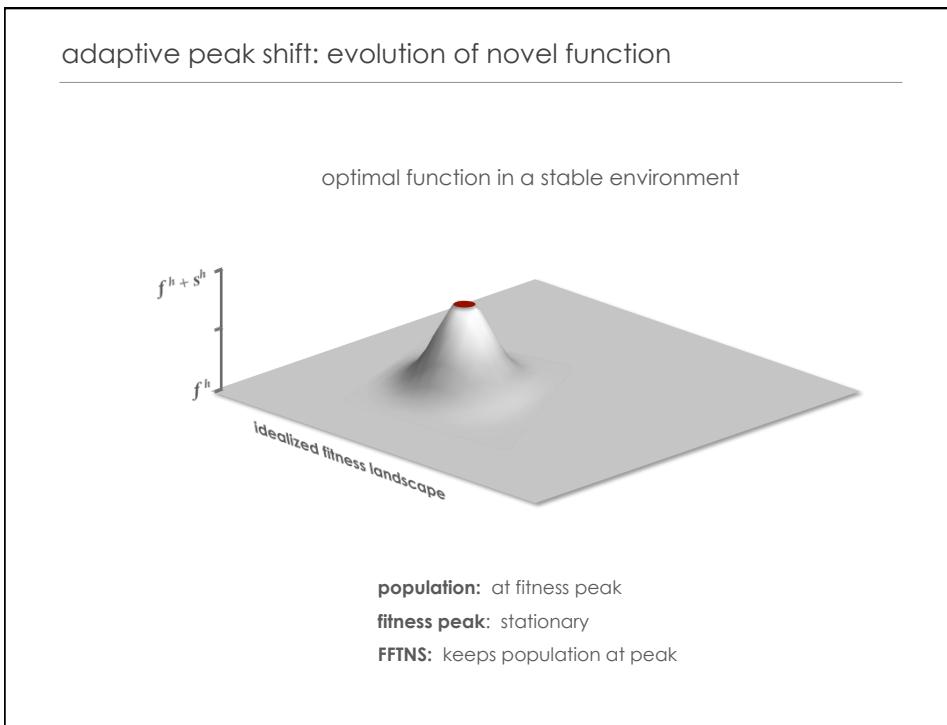
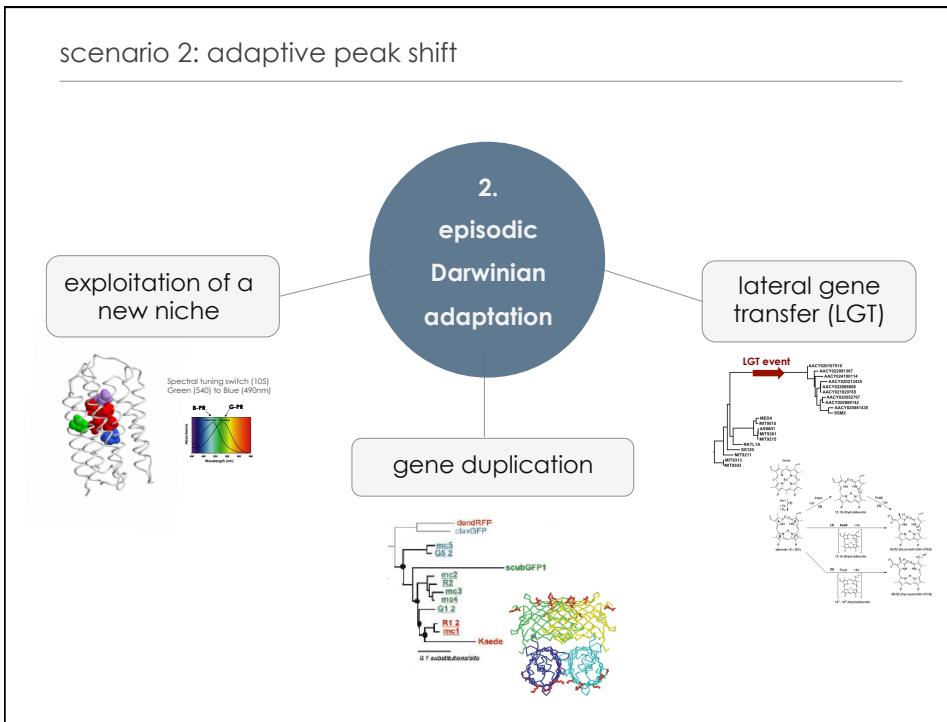
fitted model:

model M0

inference = MLE ω

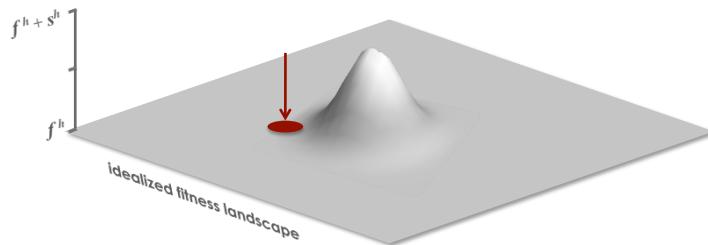
symbol = ○

conclusion: phenomenological codon models assume frequency-dependent selection



adaptive peak shift: evolution of novel function

sub-optimal function in a novel environment



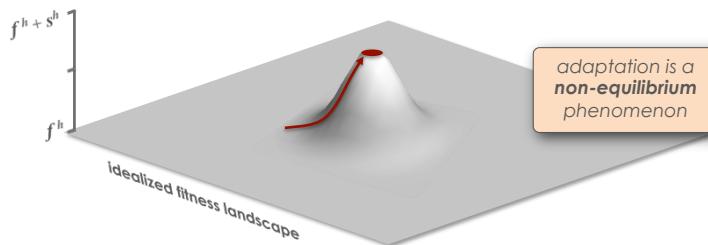
population: lower fitness

fitness peak: moving

FFTNS: increase population mean fitness
(non-stationary process)

adaptive peak shift: evolution of novel function

episodic adaptive evolution of a novel function



population: returns to peak

fitness peak: stabilized

FFTNS: increases population mean
fitness until at peak

adaptive peak shift: MutSelES model

BIOLOGY LETTERS
rsbl.royalsocietypublishing.org

Research

Cite this article: dos Reis M. 2015 How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the Fisher–Wright mutation–selection framework. *Biol. Lett.* 11: 20141031. <http://dx.doi.org/10.1098/rsbl.2014.1031>

Received: 8 December 2014
Accepted: 16 March 2015

Molecular evolution

How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the Fisher–Wright mutation–selection framework

Mario dos Reis
Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK

First principles of population genetics are used to obtain formulae relating the non-synonymous to synonymous substitution rate ratio to the selection coefficients acting at codon sites in protein-coding genes. Two theoretical cases are discussed and two examples from real data (a chloroplast gene and a virus polymerase) are given. The formulae give much insight into the dynamics of non-synonymous substitutions and may inform the development of methods to detect adaptive evolution.

4. The non-synonymous rate during adaptive evolution

adaptive peak shift: MutSelES

conclusion : episodic models “work” because $\omega > 1$ is a consequence of a system moving towards a new fitness peak.

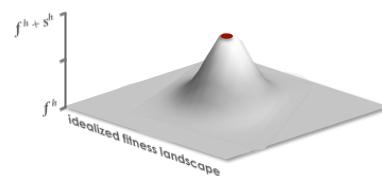
conclusion : episodic models “work” because they are sensitive to non-stationary behavior

Scenario 3: non-adaptive evolution

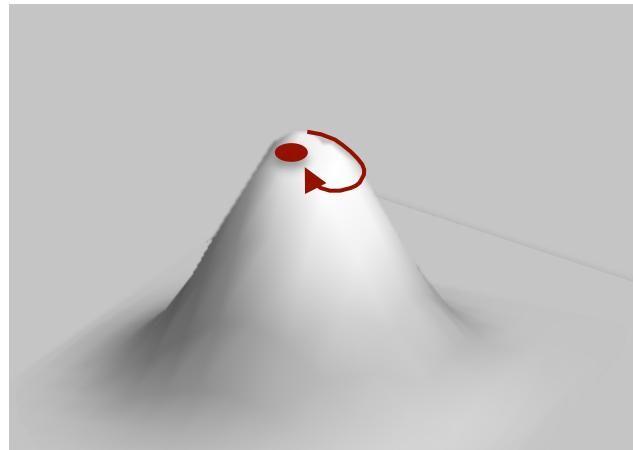
3. fitness
coefficients are
constant
(fixed-peak)

Spielman and Wilke (2015)

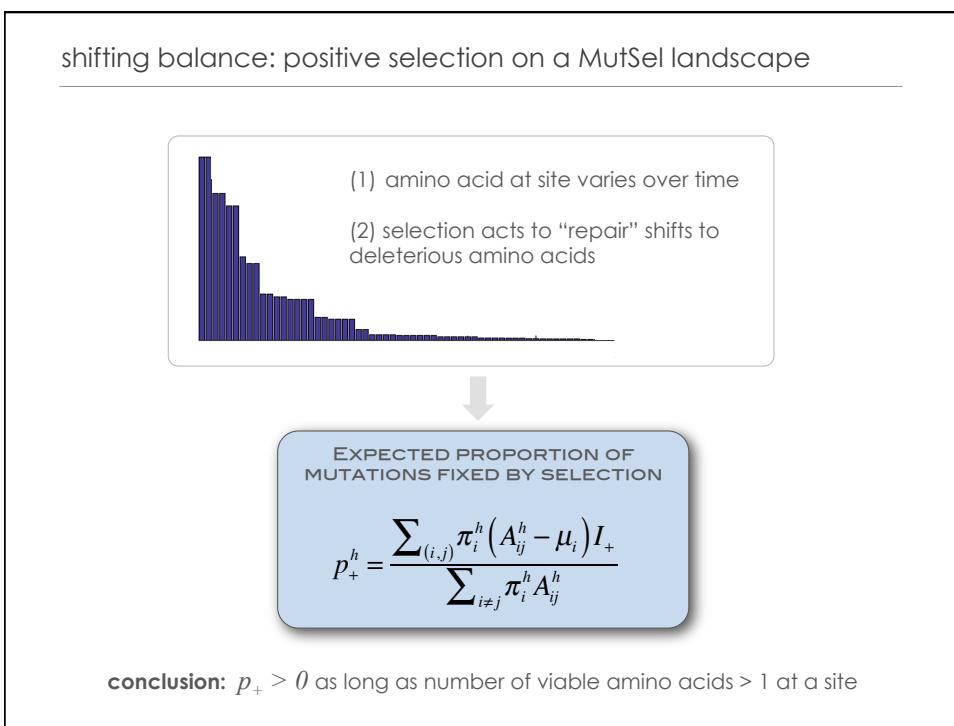
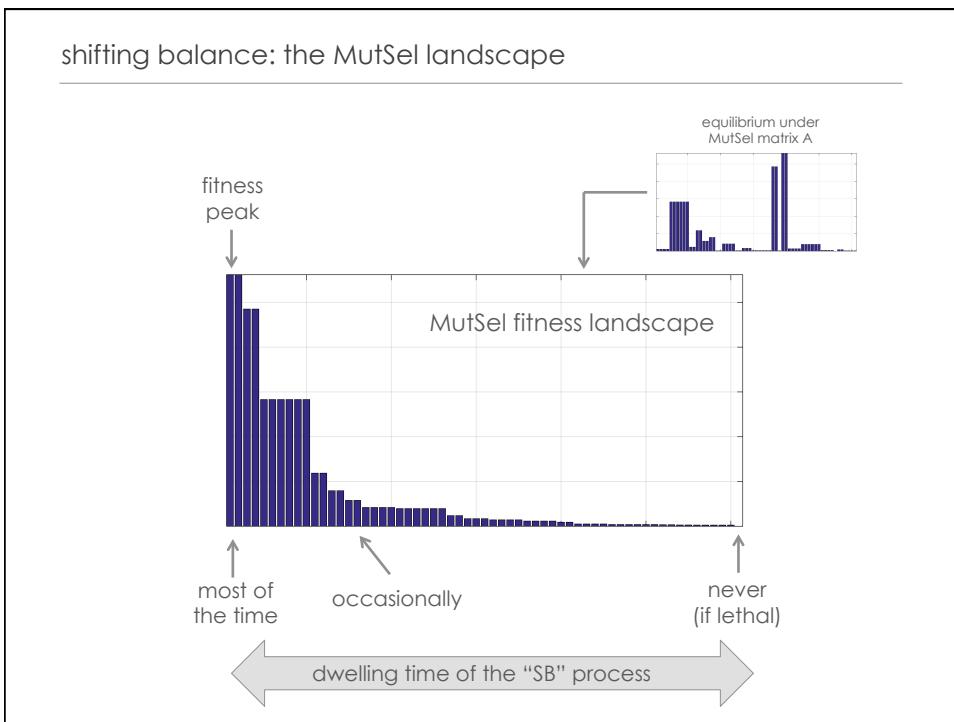
- dN/dS must be ≤ 1 when fitness coefficients are fixed.
- positive selection is not possible on a stationary fitness peak



shifting balance: movement around peak

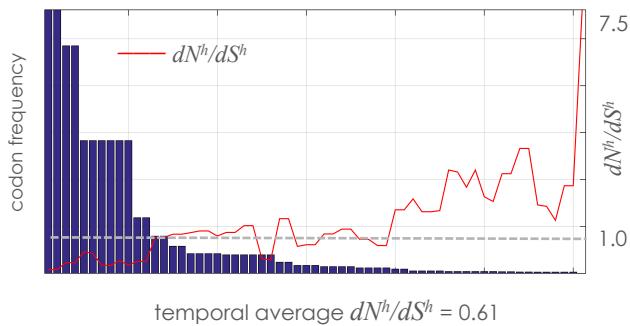


mutation and **drift** can move a pop_± off a fitness peak



shifting balance: the MutSel landscape

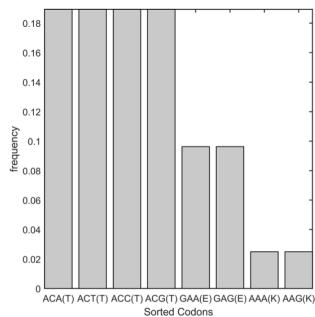
dN^h/dS^h depends on the current amino acid



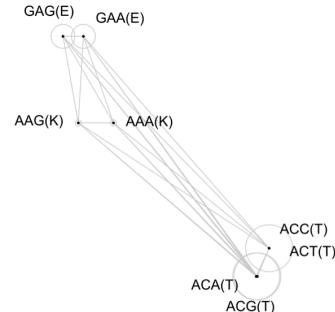
conclusion: positive selection operates on a stationary fitness peak in the same way as when there is an adaptive peak shift

landscapes have unique structures

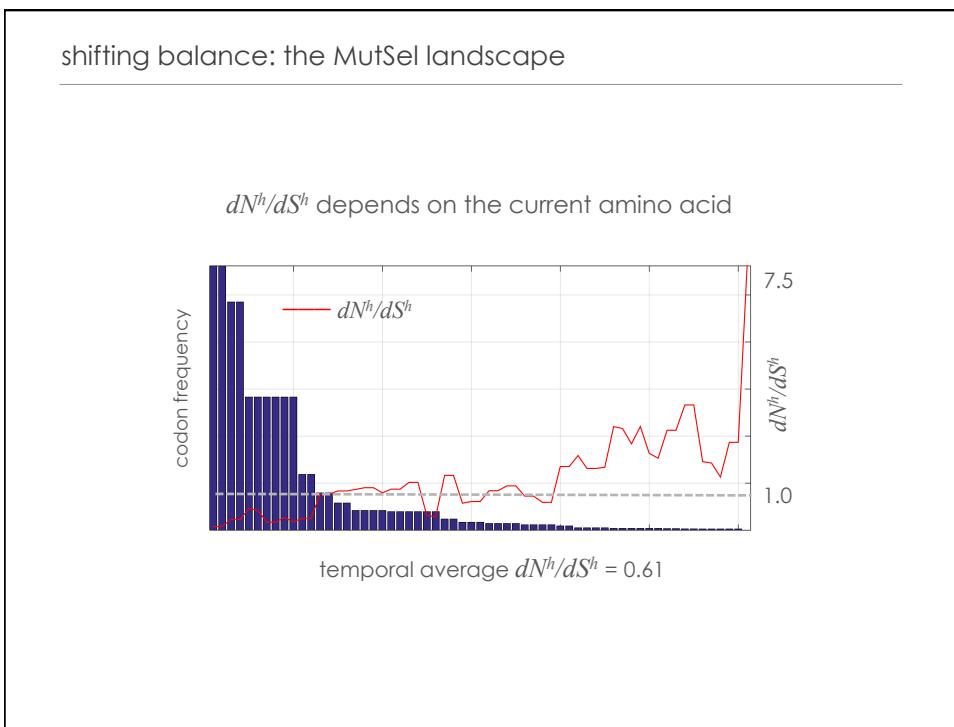
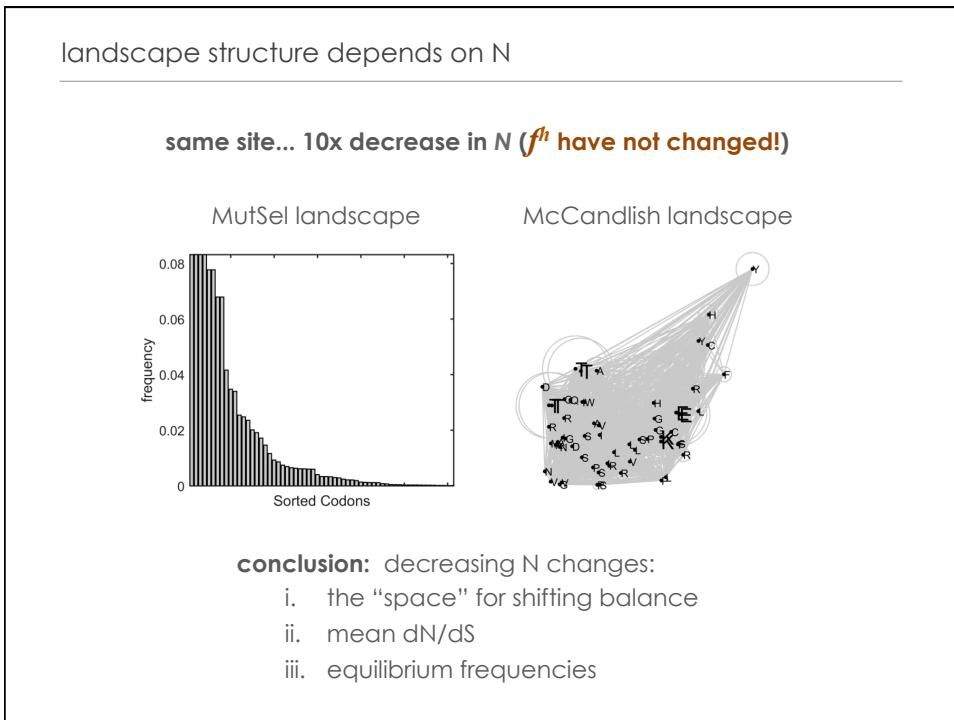
MutSel landscape

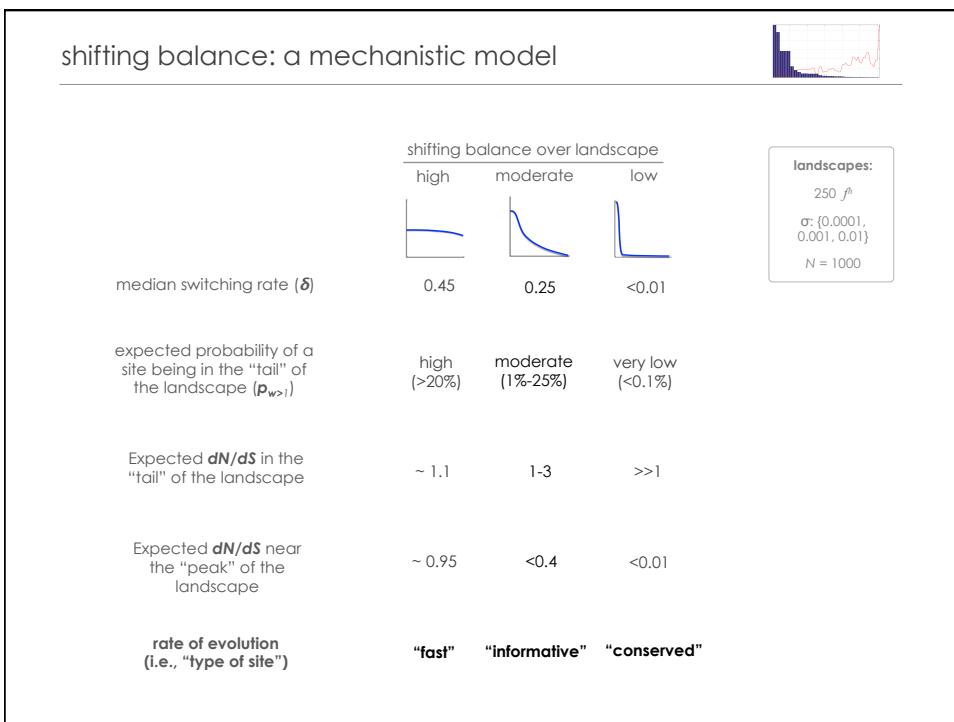
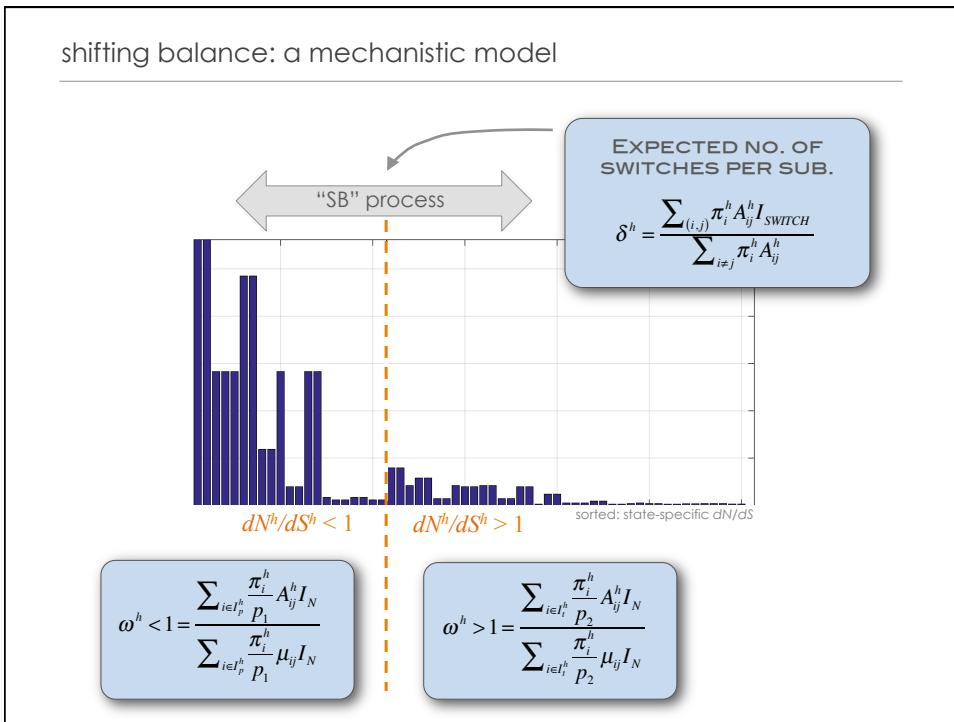


McCandlish landscape



conclusion: A population can get to a sub-optimal codon (E) by drift and reside there for some time (b/c moving between T and E requires changes ≥ 2 codons).





gene sequences

```

human
cow
rabbit
rat
opossum

GTG CTG TCT CCT GCC GAC AAG AAC GTC AAG GCC GCC TGG GGC AAG GTT GGC GCG CAC
... . . . . .G.C . . . . .T . . . . .A . . . . .A.T . . . . .AA . . . .A.C . . . .AGC . .
... . . . . .C . . . . .T . . . . .A . . . . .A . . . . .AA . . . .TG . . . .G . . . .A . . . .T . . . .GC . . . .T
... . . . . .C . . . . .G.A . . . . .AT . . . . .A . . . . .A . . . . .AA . . . .G . . . .A . . . .T . . . .GC . . . .T
... . . . . .C . . . . .G . . . . .GA . . . . .T . . . . .C . . . . .G . . . .A . . . .AT . . . . .T . . . . .G . . . .A . . . .GC . . . .

GCT GGC GAG TAT GGT GCG GAG GCC CTG GAG AGG ATG TTC CTG TCC TTC CCC ACC ACC AAG
... . . . . .A . . . . .CT . . . . .C . . . . .A . . . . .T . . . . .G . . . . .AG . . . . .C . . . . .G . . .
... . . . . .G . . . . .C . . . . .C . . . . .G . . . . .T . . . . .GG . . . . .C . . . . .A . . . . .T . . . . .A
... . . . . .G . . . . .T . . . . .A . . . . .C . . . . .A . . . . .A . . . . .C . . . . .GCT . . . . .G . . . . .A
... . . . . .C . . . . .T . . . . .CC . . . . .C . . . . .CA . . . . .T . . . . .A . . . . .T . . . . .CC . . . . .A . . . . .CC . . . . .C . . . . .T . . . . .A

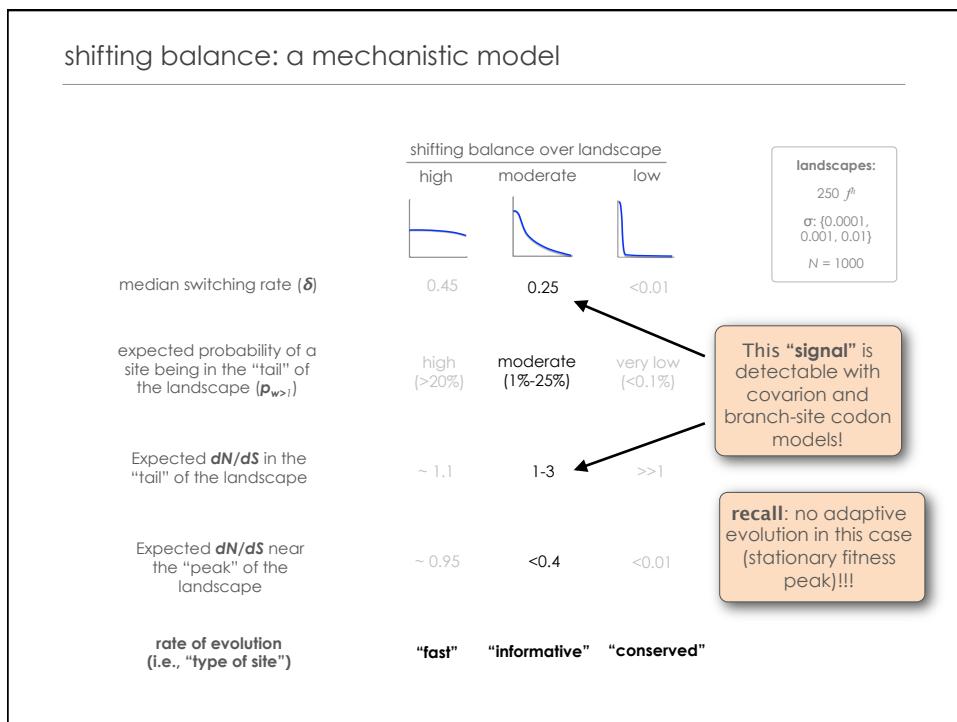
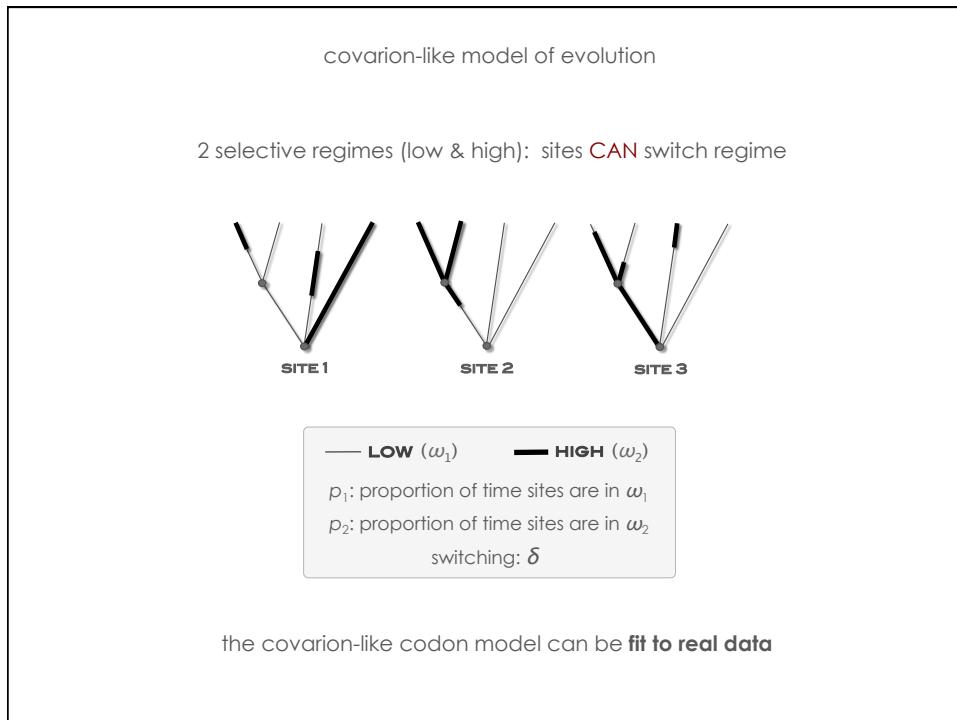
ACC TAC TTC CCG CAC TTC GAC CTG AGC CAC GGC TCT GCC CAG GTT AAG GGC CAC GGC AAG
... . . . . .C . . . . .C . . . . .G . . . . .C . . . . .G . . . . .C . . . . .G . . .
... . . . . .C . . . . .C . . . . .T.C . . . . .C . . . . .AG . . . . .A.C . . . . .A.C . . . . .C . . .
... . . . . .T.T . . . . .A.T . . . . .T.G.A . . . . .C . . . . .C . . . . .C . . . . .CT . . . . .T
... . . . . .T . . . . .C . . . . .TC . . . . .C . . . . .C . . . . .A.C . . . . .T . . . . .T . . . . .A

```

covarion-like model of evolution

$$Q = \left[\begin{array}{|c|c|} \hline \text{evolutionary regime 1:} & \text{switching process:} \\ \omega_1 = \text{low} & \omega_1 \rightarrow \omega_2 \\ (\text{"near the peak"}) & \\ \hline \text{switching process:} & \text{evolutionary regime 2:} \\ \omega_2 \rightarrow \omega_1 & \omega_1 = \text{high} \\ & (\text{"in the tail"}) \\ \hline \end{array} \right]$$

Guindon et al., 2004
Jones et al. 2016



summary

- standard codon models (single ω) assume frequency dependent selection, which yields a persistent $dN/dS > 1$
- episodic adaptive evolution leads to transient $dN/dS > 1$ (non-stationary process, with ω upwardly biased)
- MutSel landscapes can be complex and a site can reside at a sub-optimal state for extended periods of time
- protein evolution on a static fitness landscape has temporal dynamics that include positive selection
- rate variation among sites reflects the interplay between mutation, drift, and selection (i.e., shifting balance dynamics)