

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

part 2: mechanistic process

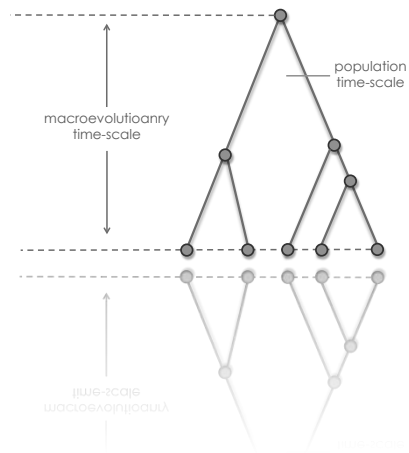
part 3: phenomenological
modeling

- neutral theory
- dN/dS
- mechanistic process
- phenomenological outcomes

- MutSel framework
- freq dependent selection
- episodic selection
- shifting balance

- types of models
- 3 analysis tasks
- phylogenomic example
- best practices

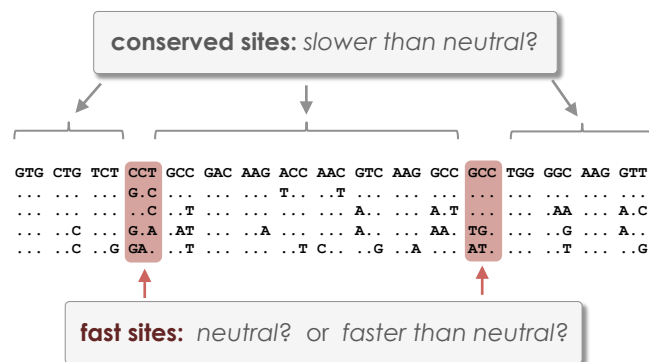
part 1: introduction



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?

neutral theory of molecular evolution (Kimura 1968)

the **number of new mutations** arising in a diploid population

$$2N\mu$$

the **fixation probability** of a new mutant by drift

$$1/2N$$

The **substitution (fixation) rate, k**

$$k = 2N\mu \times 1/2N$$

the elegant simplicity of **neutral theory**: $k = \mu$

genetic code determines impact of a mutation

	U	C	A	G	
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

<http://www.iangara.bc.ca/biology/mario/assets/Geneticcode.jpg>

The genetic code determines how random changes to the gene brought about by the process of mutation will impact the function of the encoded protein.

Kimura (1968)

d_S : number of synonymous substitutions per synonymous site (K_S)

d_N : number of nonsynonymous substitutions per nonsynonymous site (K_A)

ω : the ratio d_N/d_S ; it measures selection at the protein level

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 :

Synonymous sites = 25.5%

$$S = 300 \times 3 \times 25.5\% = 229.5$$

Nonsynonymous sites = 74.5%

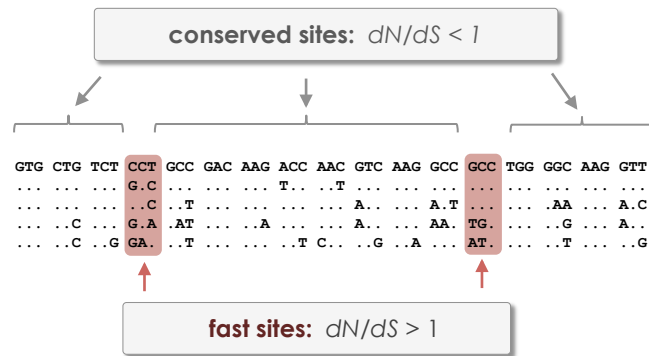
$$N = 300 \times 3 \times 74.5\% = 670.5$$

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

$$d_N = 5/670.5 = 0.0075$$

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

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

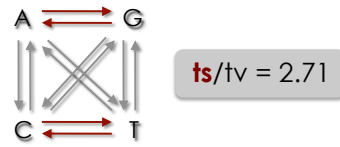
Relative proportion of different types of mutations in hypothetical protein coding sequence.				
Type	Expected number of changes (proportion)			
	All 3 Positions	1 st positions	2 nd positions	3 rd positions
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
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**:

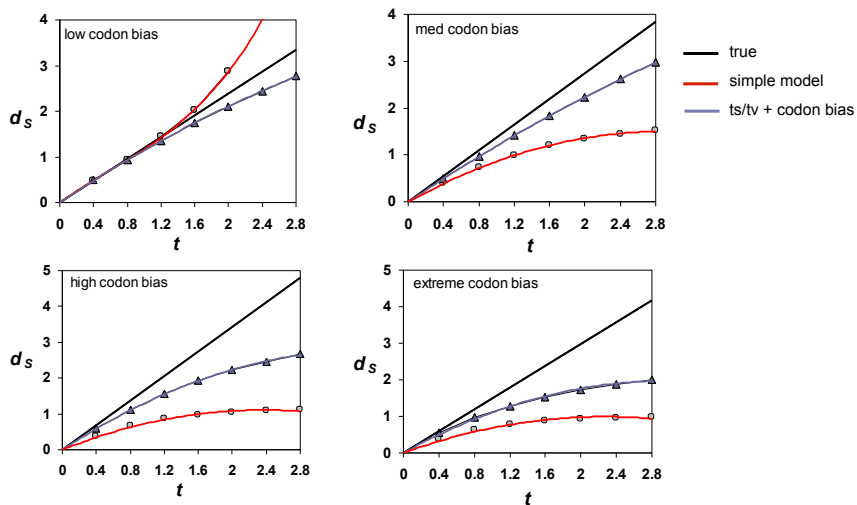


preferred vs. **un-preferred** codons:

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	0	*** * TGA	0
TTG	1	TCG	1	TAG	0	Trp W TGG	8
Leu L CTT	2	Pro P CCT	1	His H CAT	0	Arg R CGT	1
CTC	2	CCC	15	CAC	4	CGC	7
CTA	0	CCA	3	Gln Q CAA	0	CGA	0
CTG	29	CCG	1	CAG	14	CGG	0

uncorrected evolutionary bias leads to estimation bias



data from: Dunn, Bielawski, and Yang (2001) *Genetics*, 157: 295-305

recap

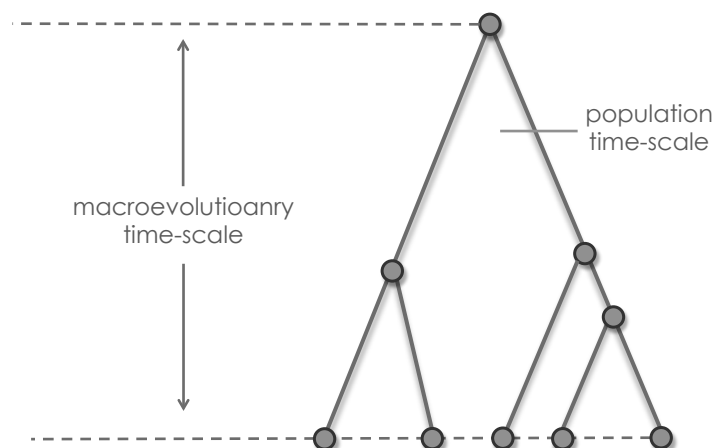
dS and dN must be corrected for BOTH the structure of genetic code and the underlying mutational process of the DNA

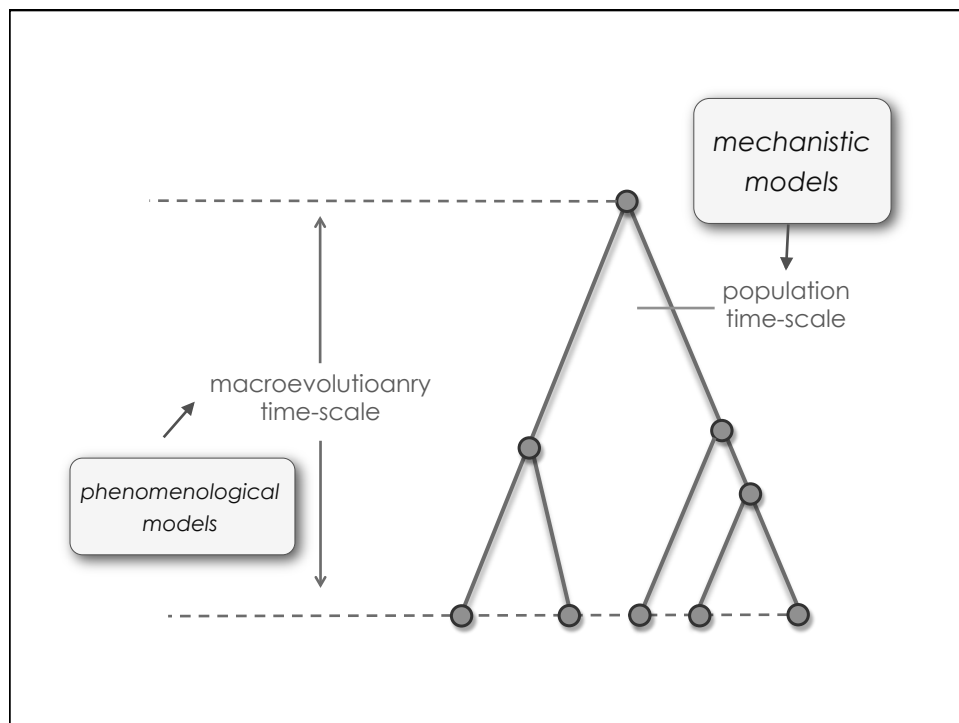
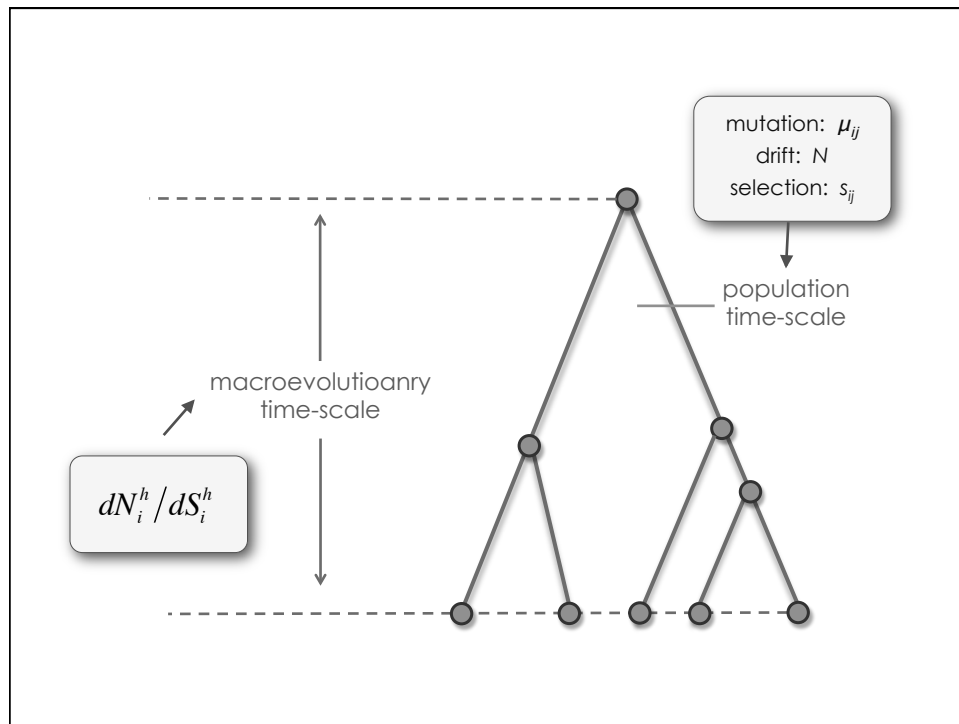
but, this can differ among lineages and genes!

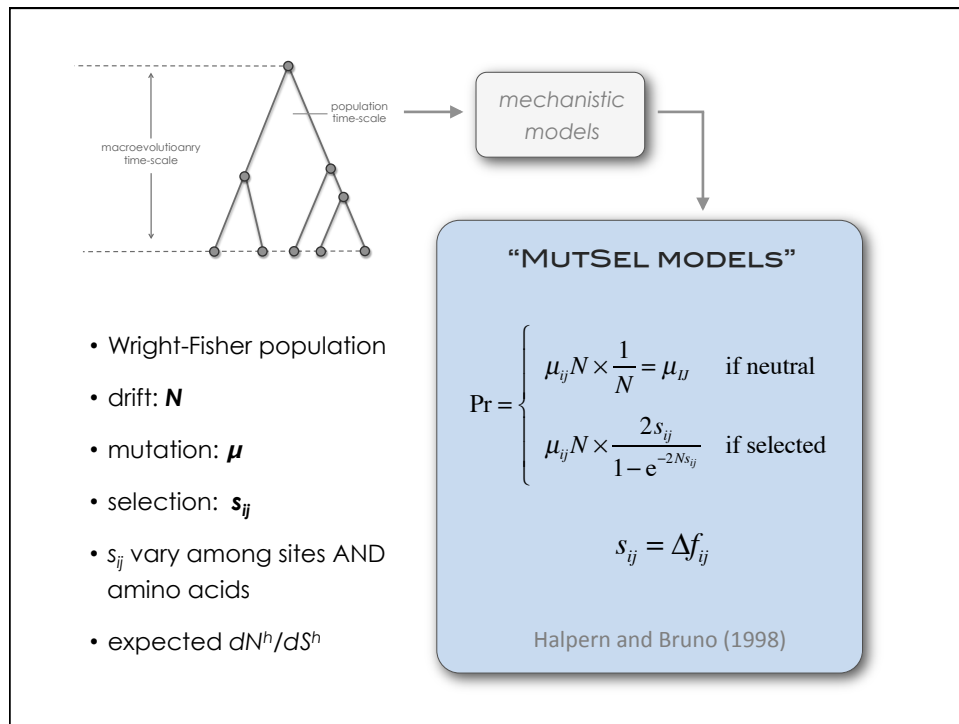
correcting dS and dN for underlying mutational process of the DNA makes them **sensitive to assumptions about the process of evolution!**

but, the process of evolution occurs at the population genetic level (micro-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

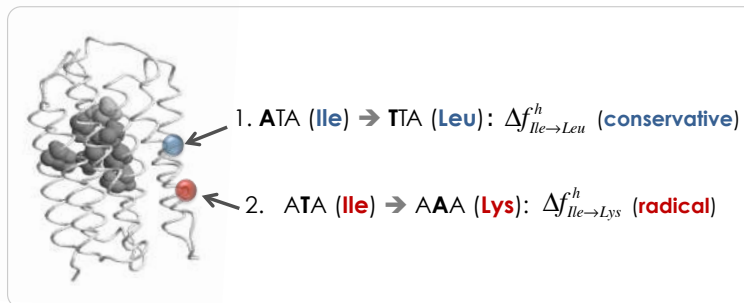
$$s_{ij}^h = f_j^h - f_i^h$$

fixation probability (Kimura, 1962)

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

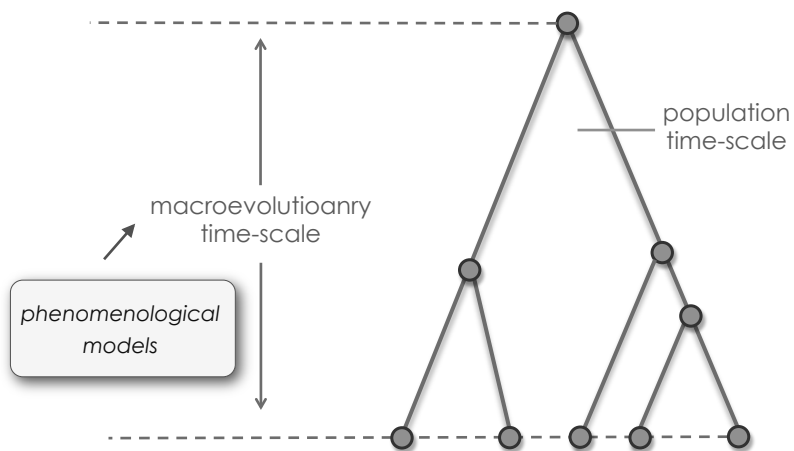
fixation probability with selection

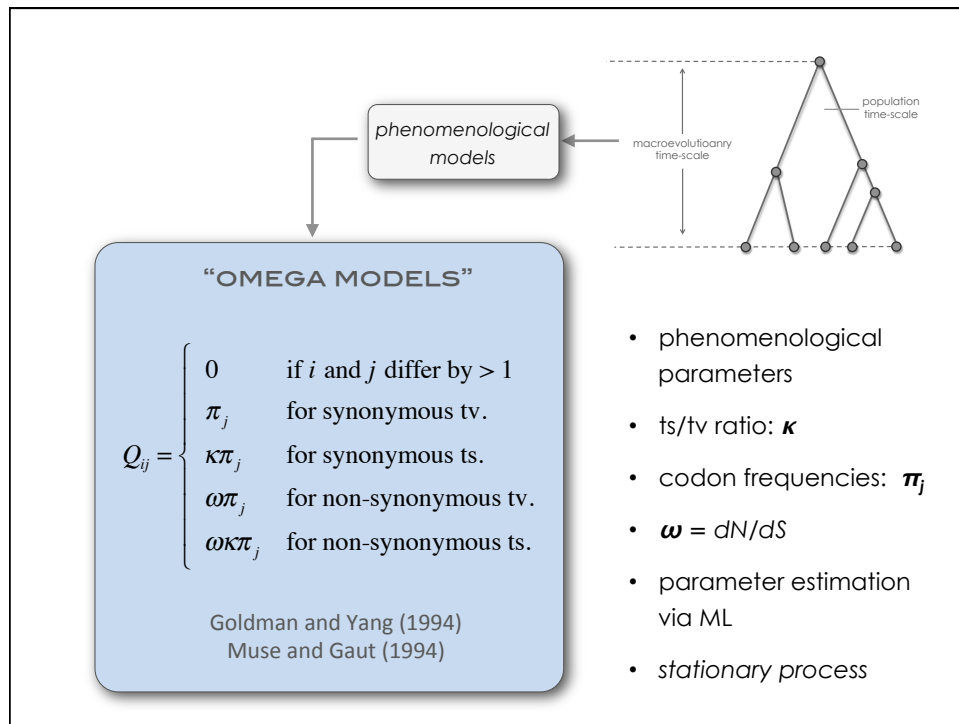
MutSel: selection favours amino acids with higher fitness (if N is large enough)



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





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

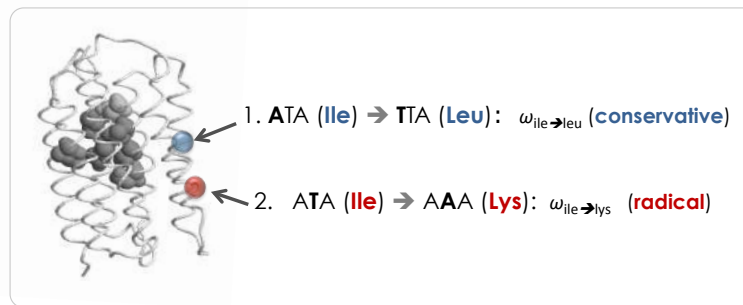
phenomenological codon models: just a few parameters are needed to cover the 3721 transitions between codons!

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	—

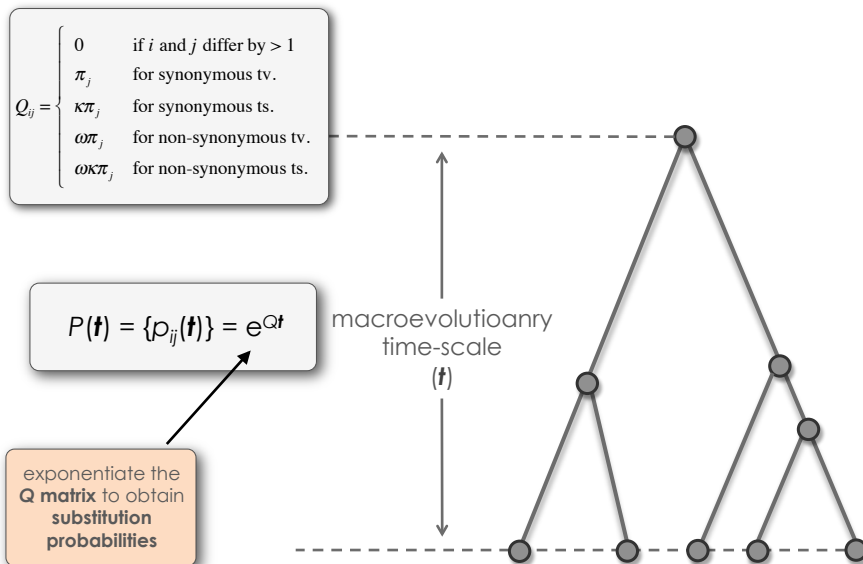
* This is equivalent to the codon model of Goldman and Yang (1994). Parameter ω is the ratio d_N/d_S , κ is the transition/transversion rate ratio, and π_i is the equilibrium frequency of the target codon (i).

substitution probability with selection

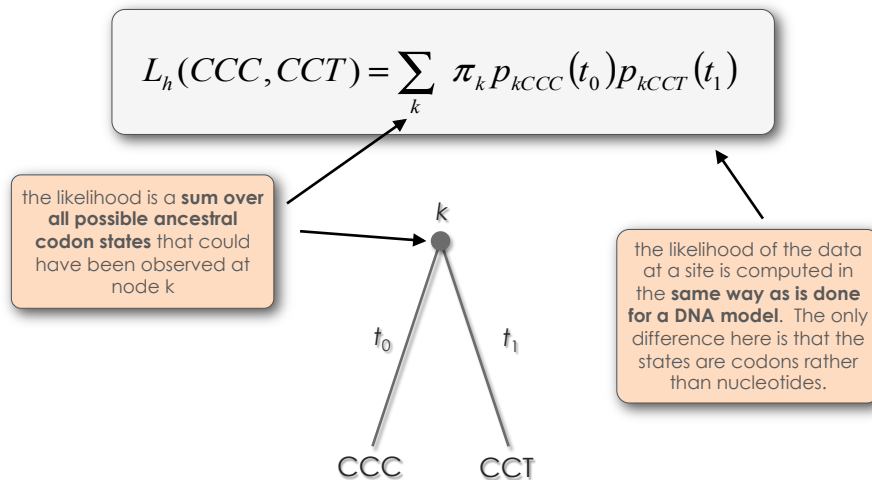
intentional simplification: all amino acid substitutions have the same ω !



contradiction? selection should favour amino acids with higher fitness.

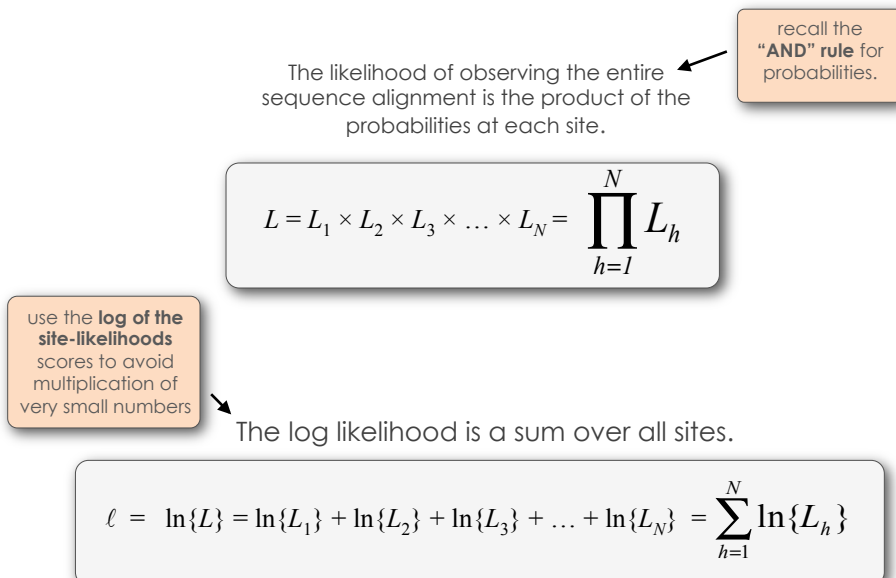
probability of substitution between codons over time, $P(t)$ 

likelihood of the data at a site



note: analysis is typically done by using an unrooted tree

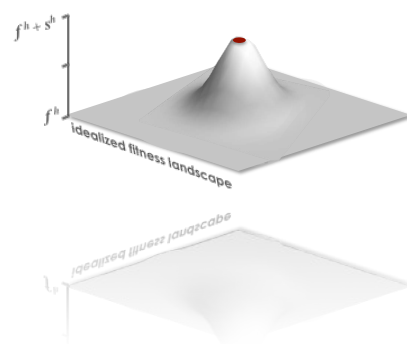
likelihood of the data at all sites



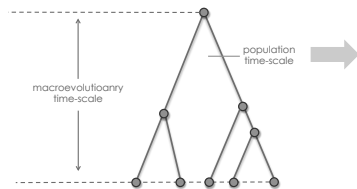
summary

- dN/dS is a measure of selection pressure that can be connected to a mechanistic process of population genetic evolution (MutSel models)
- dN/dS can be estimated from multi-sequence alignments as a parameter (ω) in a phenomenological model of sequence evolution
- estimates of dN/dS for real data must be corrected for the underlying process of evolution for those data
- estimates of dN/dS can be sensitive to assumptions about the underlying process of evolution
- phenomenological estimates of dN/dS are highly simplistic summaries of a much more complex evolutionary process

part 2: mechanistic processes
of codon evolution



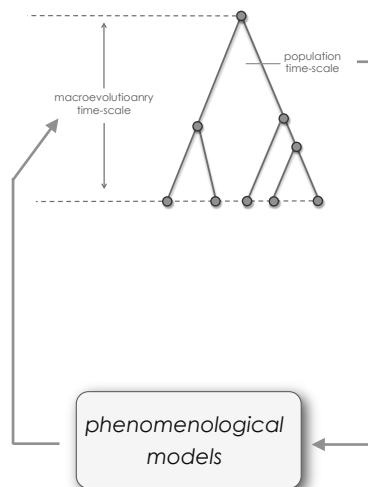
site-specific MutSel rate matrix



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)



"MUTSEL FRAMEWORK"

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

$$s_{ij} = \Delta f_{ij}$$

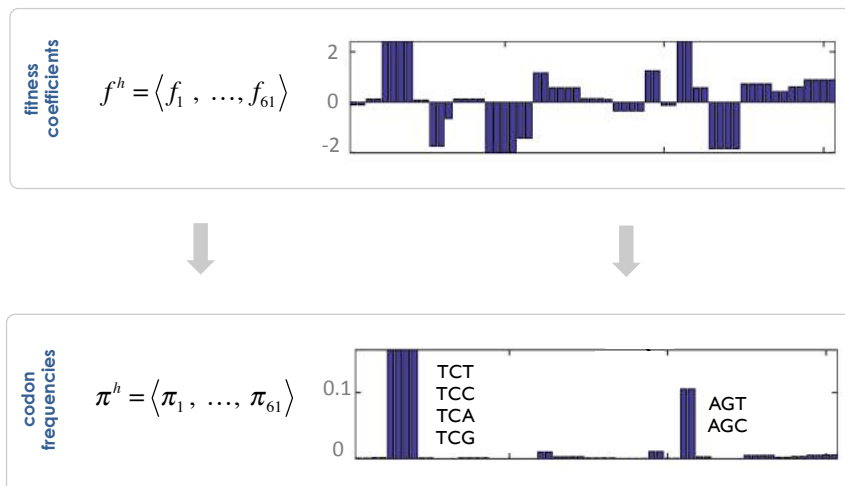
Halpern and Bruno (1998)
Jones et al. (2016)

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



2. from fitness coefficients to dN/dS

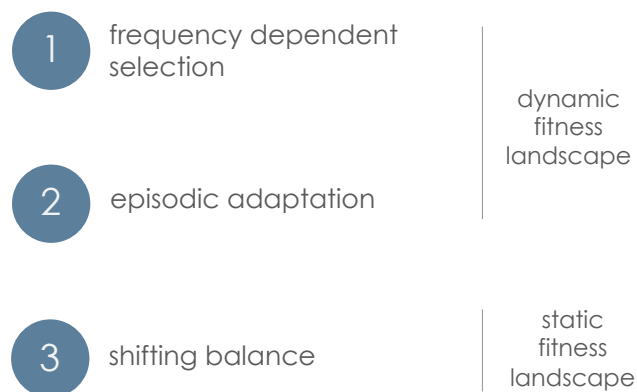
MUTSEL RATE MATRIX

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

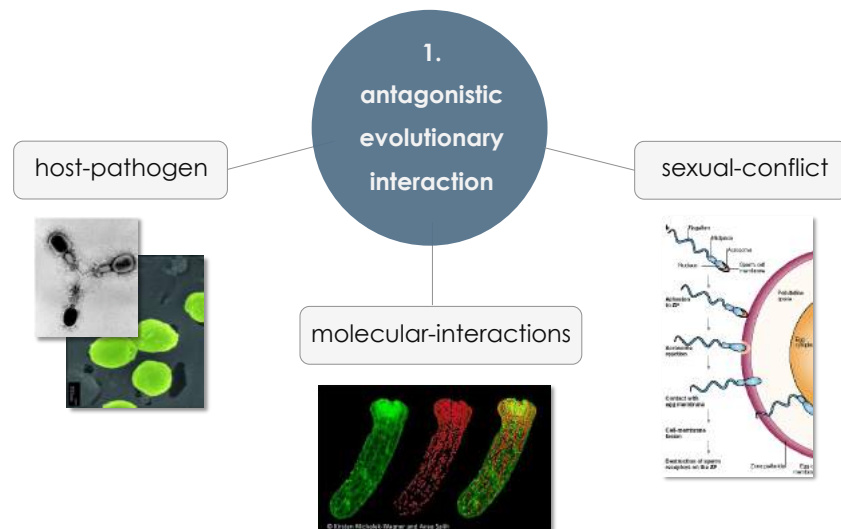
$$dN^h / dS^h = \frac{\sum_{i \neq j} \pi_i^h A_{ij}^h I_N}{\sum_{i \neq j} \pi_i^h \mu_{ij} I_N}$$

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

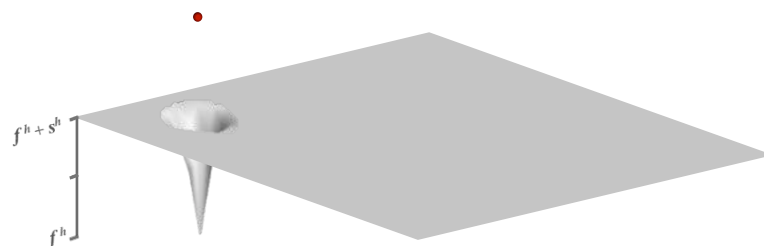
positive selection: 3 evolutionary scenarios



scenario 1: frequency dependent selection

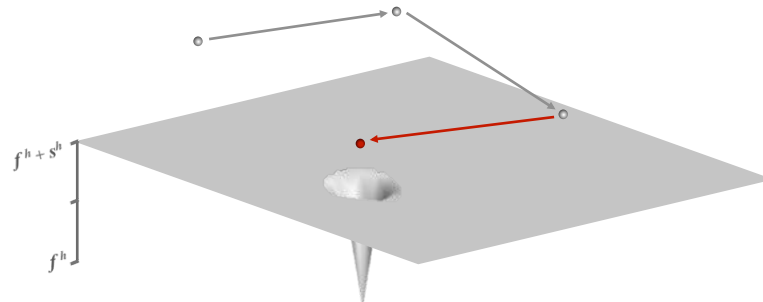


frequency-dependent adaptive landscape (weird)



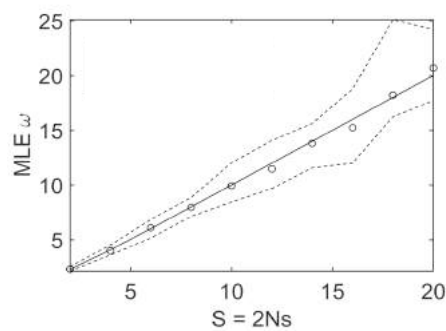
frequency-dependent selection: MutSelM0

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 Q as **codon model M0**

frequency-dependent selection: MutSelM0



generating process:

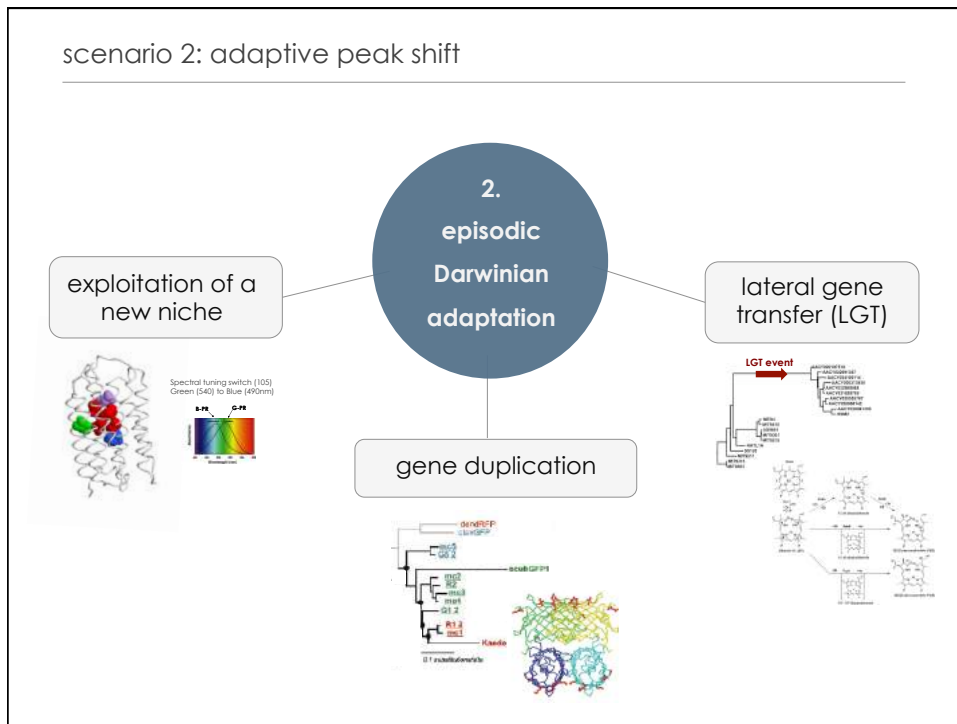
MutSelM0
 expectation = dN^h/dS^h
 symbol = —

fitted model:

model M0
 inference = MLE ω
 symbol = ○

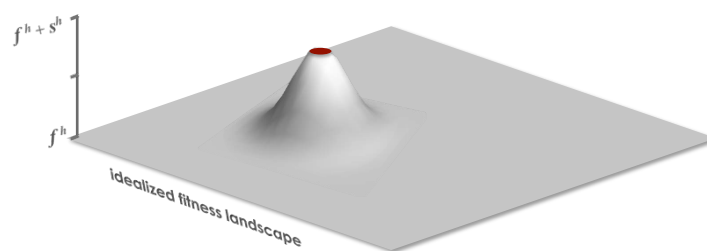
conclusion: phenomemological codon models
 assume frequency-dependent selection

scenario 2: adaptive peak shift



adaptive peak shift: evolution of novel function

optimal function in a stable environment



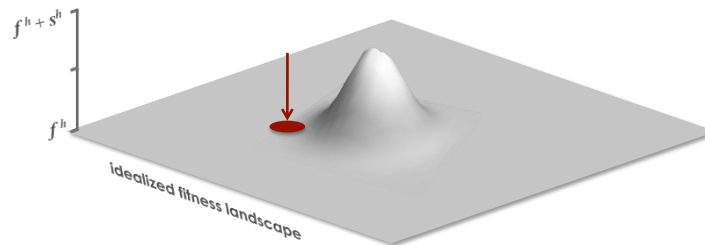
population: at fitness peak

fitness peak: stationary

FFNS: keeps population at peak

adaptive peak shift: evolution of novel function

sub-optimal function in a novel environment



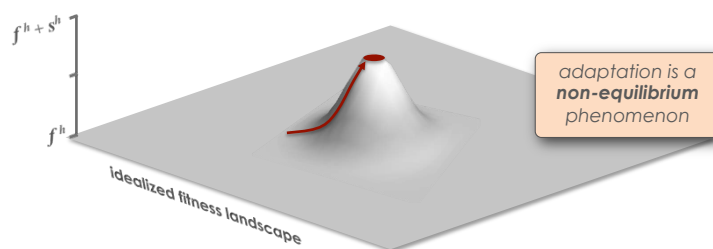
population: lower fitness

fitness peak: moving

FTNS: 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

FTNS: 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

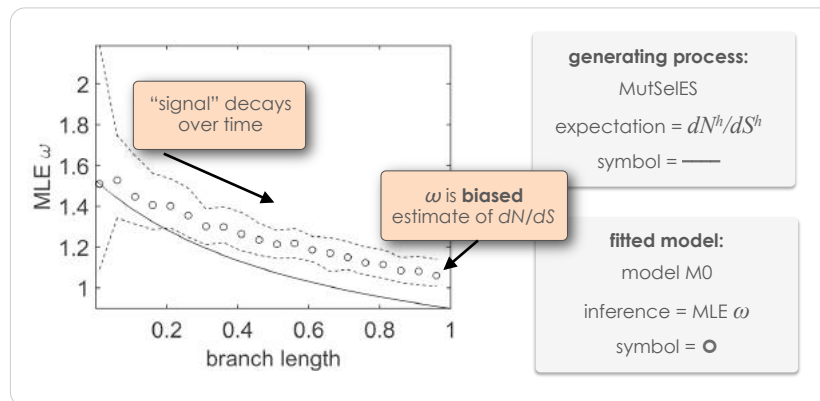
Mario dos Reis

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.

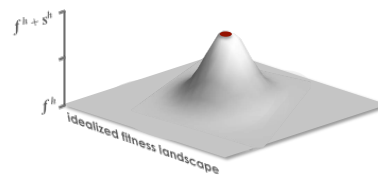
conclusion : episodic models “work” because they are sensitive to non-stationary behavior

Scenario 3: non-adaptive evolution

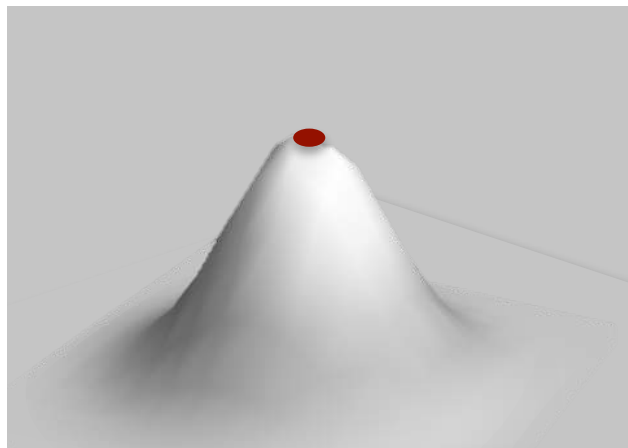
3. fitness
coefficients are
constant
(fixed-peak)

Spielman and Wilke (2015)

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

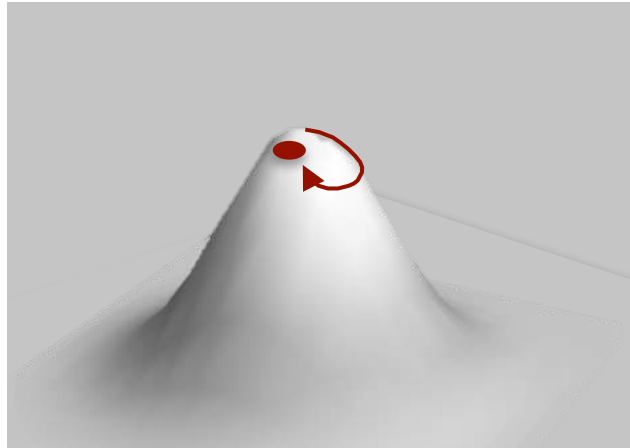


shifting balance: movement around peak



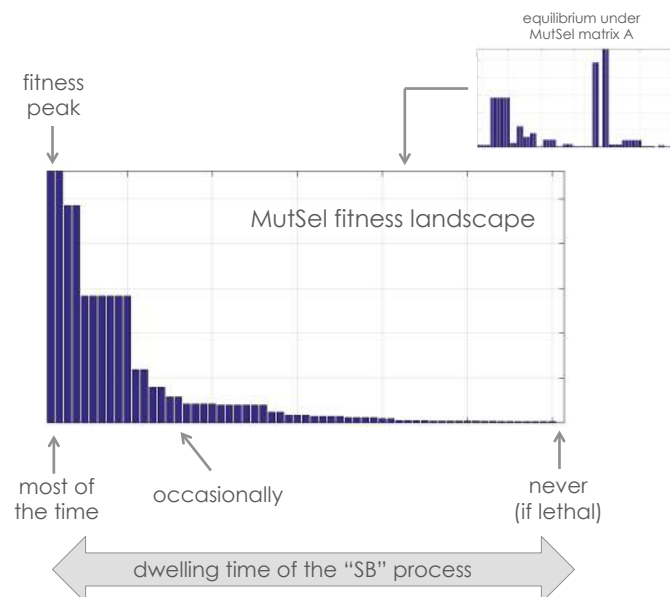
mutation and **drift** can move a pop₂ off a fitness peak

shifting balance: movement around peak

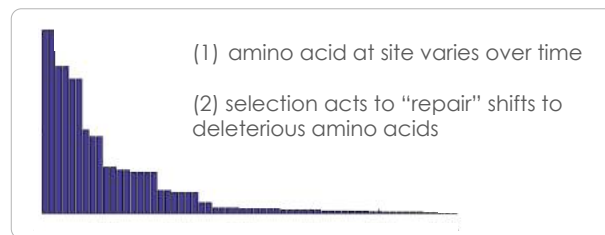


mutation and **drift** can move a pop₂ off a fitness peak

shifting balance: the MutSel landscape



shifting balance: positive selection on a MutSel landscape



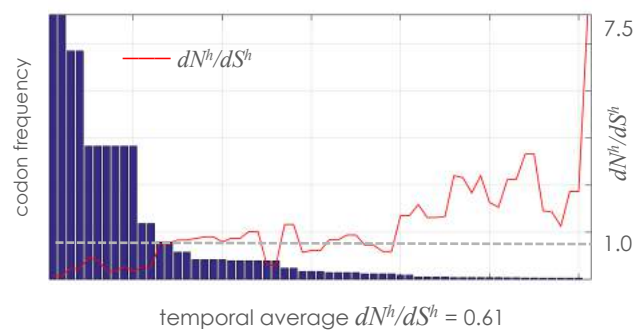
EXPECTED PROPORTION OF
MUTATIONS FIXED BY SELECTION

$$p_+^h = \frac{\sum_{(i,j)} \pi_i^h (A_{ij}^h - \mu_i) I_+}{\sum_{i \neq j} \pi_i^h A_{ij}^h}$$

conclusion: $p_+ > 0$ as long as number of viable amino acids > 1 at a site

shifting balance: the MutSel landscape

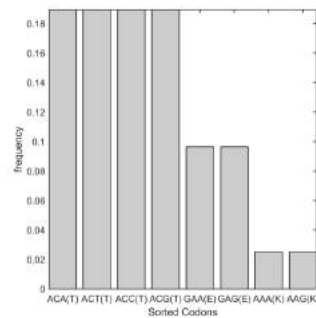
dN^h/dS^h depends on the current amino acid



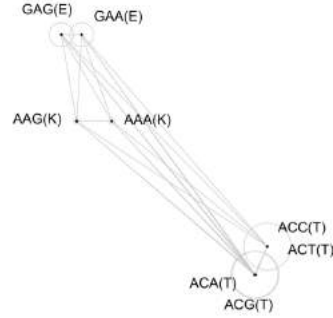
conclusion: positive selection operates on a stationary fitness peak in the same way as when there is an adaptive peak shift

landscapes have unique structures

MutSel landscape



McCandlish landscape

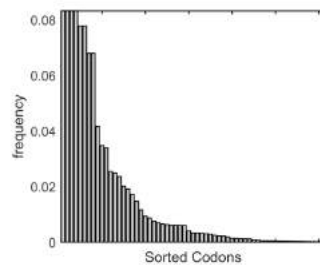


conclusion: A population can get to a sub-optimal codon (E) by drift and reside there for some time (b/c moving between T and E requires changes ≥ 2 codons).

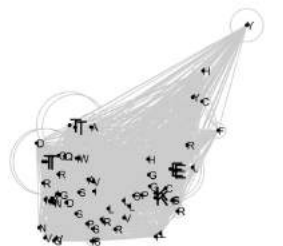
landscape structure depends on N

same site... 10x decrease in N (f^h have not changed!)

MutSel landscape



McCandlish landscape

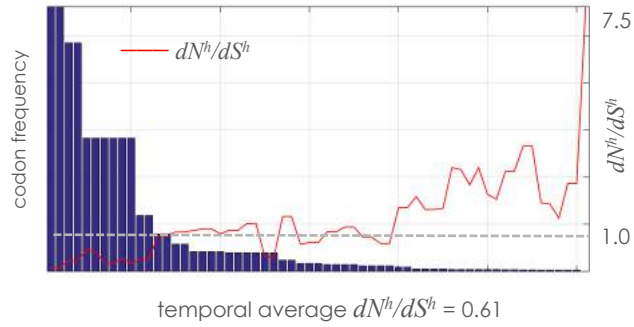


conclusion: decreasing N changes:

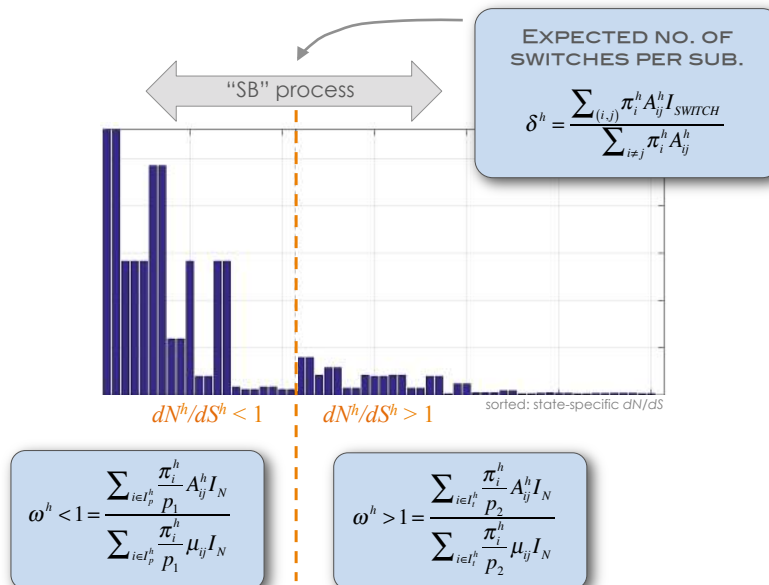
- the "space" for shifting balance
- mean dN/dS
- equilibrium frequencies

shifting balance: the MutSel landscape




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



shifting balance: a mechanistic model






shifting balance: a mechanistic model

	shifting balance over landscape			
	high	moderate	low	
				
median switching rate (δ)	0.45		<0.01	
expected probability of a site being in the "tail" of the landscape ($p_{w>1}$)	high (>20%)		very low (<0.1%)	
Expected dN/dS in the "tail" of the landscape	~ 1.1		$\gg 1$	
Expected dN/dS near the "peak" of the landscape	~ 0.95		<0.01	
rate of evolution (i.e., "type of site")	"fast"		"conserved"	

landscapes:
 $250 \text{ } f^h$
 $\sigma: \{0.0001, 0.001, 0.01\}$
 $N = 1000$

shifting balance: a mechanistic model

	shifting balance over landscape			
	high	moderate	low	
				
median switching rate (δ)	0.45	0.25	<0.01	
expected probability of a site being in the "tail" of the landscape ($p_{w>1}$)	high (>20%)	moderate (1%-25%)	very low (<0.1%)	
Expected dN/dS in the "tail" of the landscape	~ 1.1	1-3	$\gg 1$	
Expected dN/dS near the "peak" of the landscape	~ 0.95	<0.4	<0.01	
rate of evolution (i.e., "type of site")	"fast"	"informative"	"conserved"	

landscapes:
 $250 \text{ } f^h$
 $\sigma: \{0.0001, 0.001, 0.01\}$
 $N = 1000$

gene sequences

human
cow
rabbit
rat
opossum

```

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

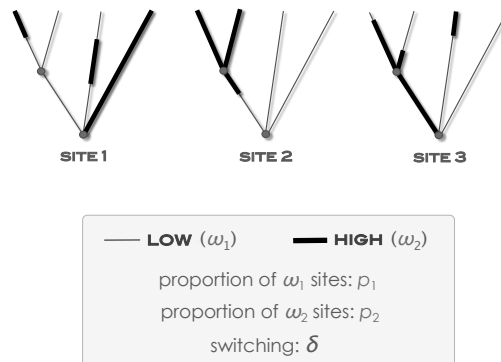
GCT GGC GAG TAT GGT GCG GAG GCC CTG GAG AGG ATG TTC CTG TCC TTC CCC ACC ACC AAG
... ..A ..CT ... ..C ..A ... ..T ... .. .. .. .. AG. ... .. .. ..
..G. ... .. ..C ..C ... ..G.. ... .. ..T.. GG. ... .. .. ..
..G. ..T ..A ... ..C ..A. ... ..A C.. ... ..GCT G.. ... .. ..
..C ..T ..CC ..C ..CA ..T ..A ..T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

```

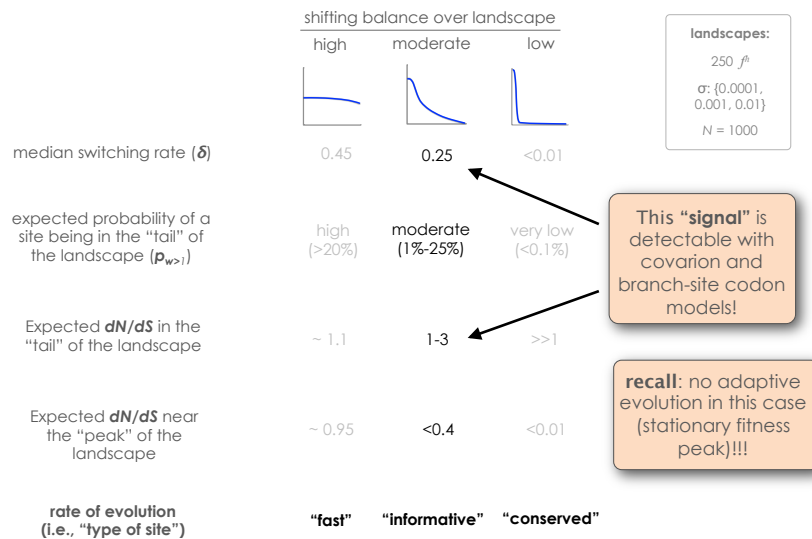
covarion-like model of evolution

2 selective regimes (low & high): sites **CAN** switch regime



the covarion-like codon model can be **fit to real data** (Guindon et al., 2004)

shifting balance: a mechanistic model



summary

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