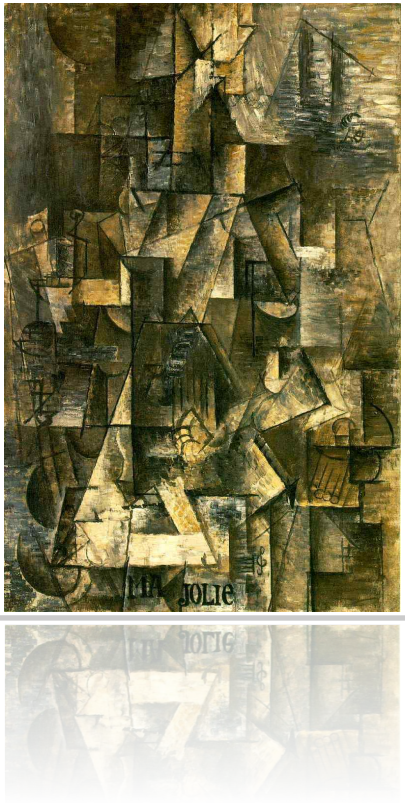
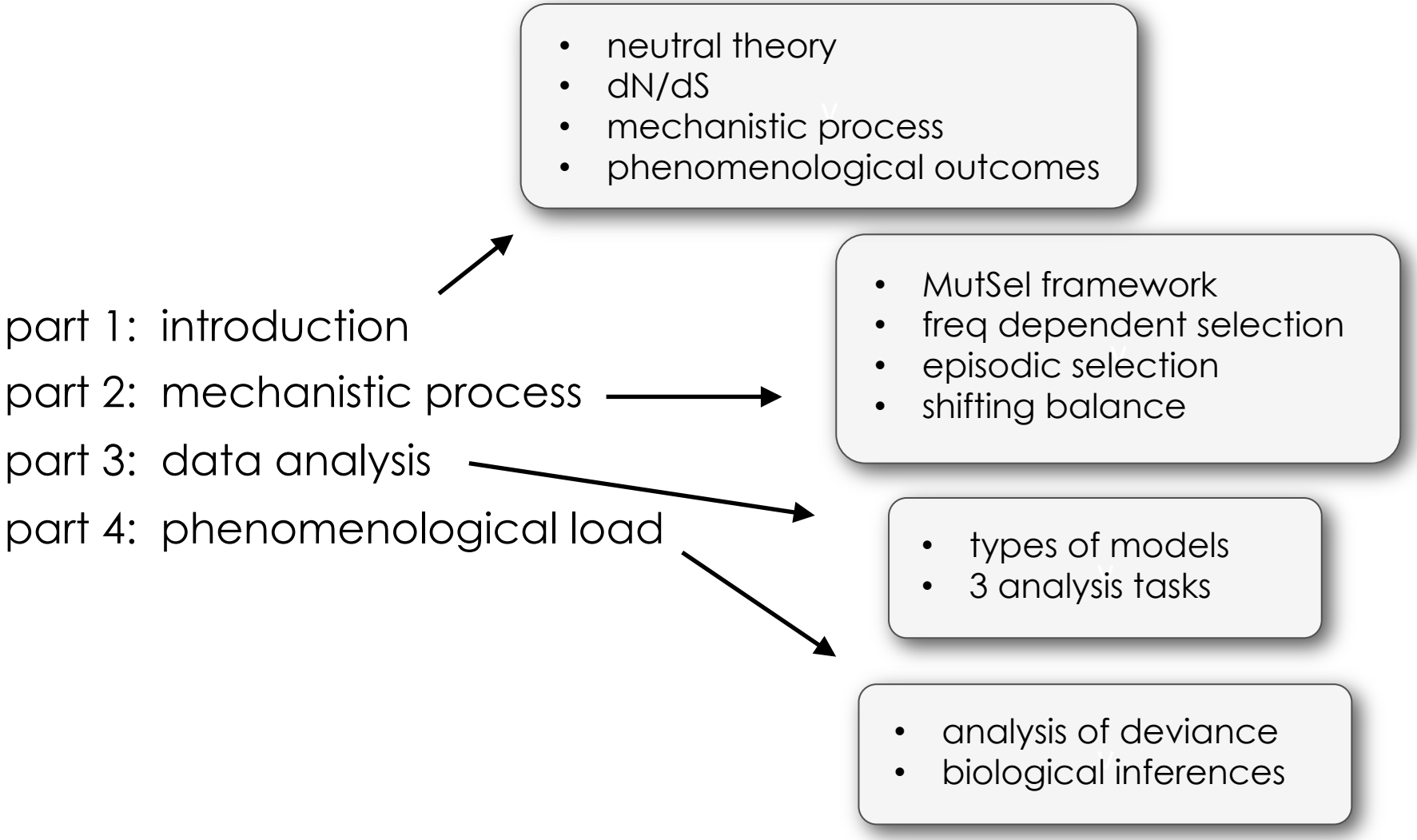
codon substitution models and the analysis of natural selection pressure

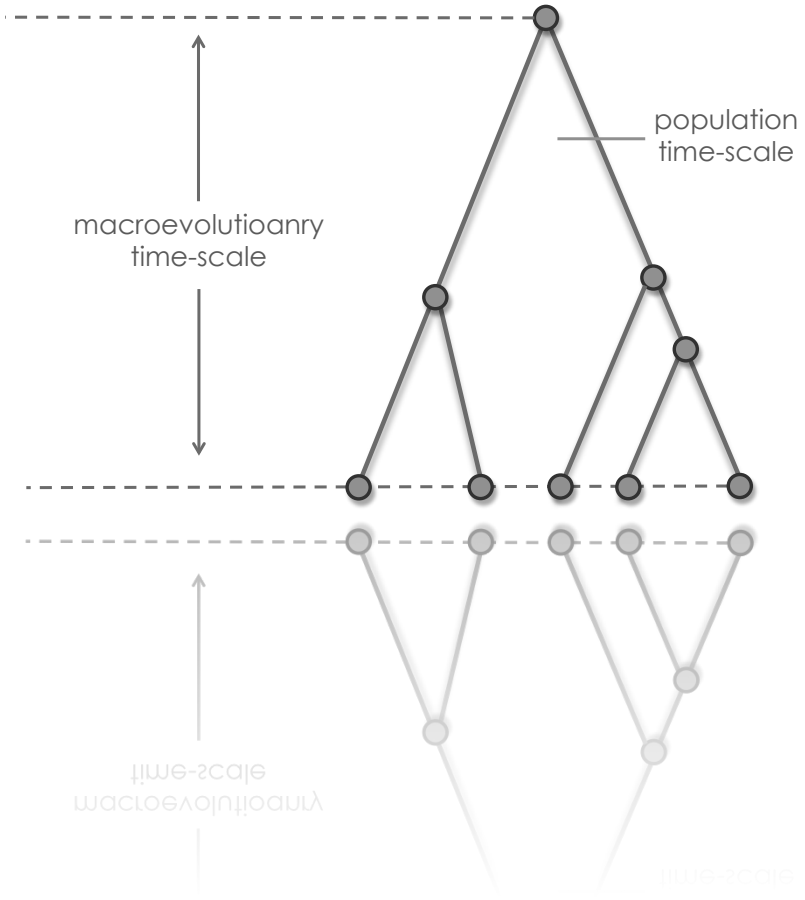


Joseph P. Bielawski
Department of Biology
Department of Mathematics & Statistics
Dalhousie University

The goals and the plan



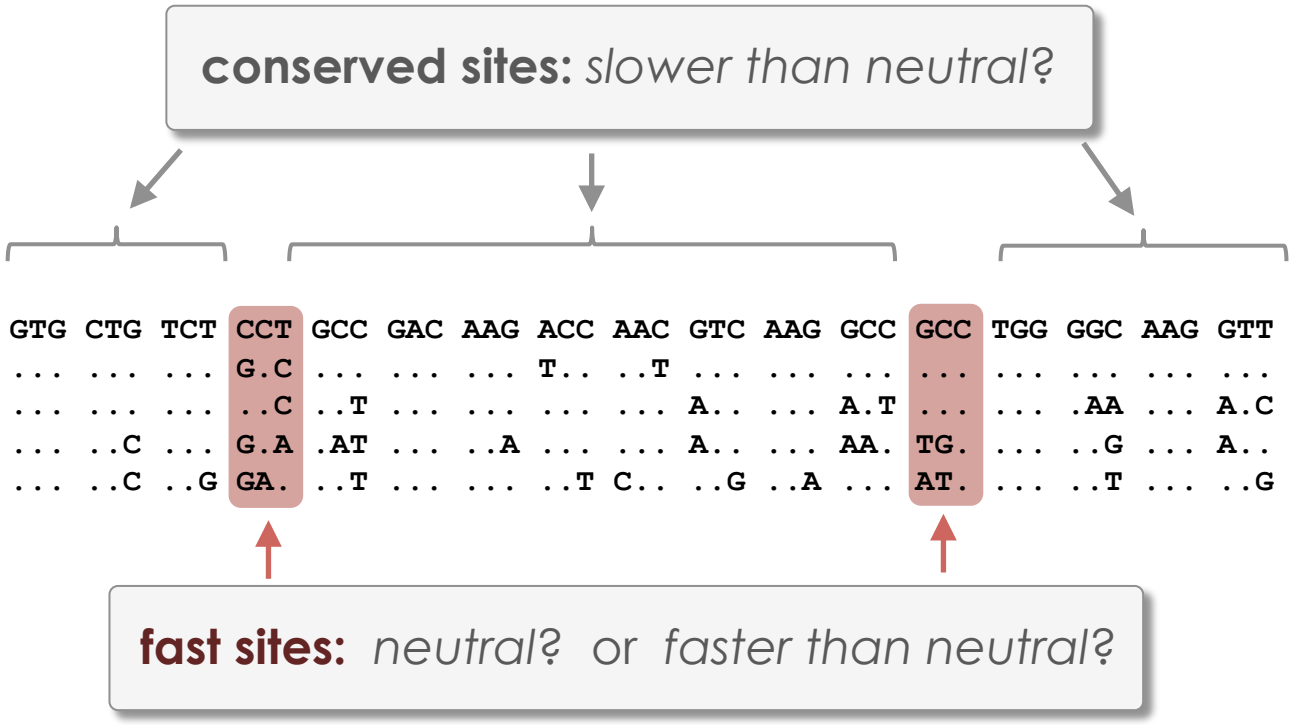
part 1: introduction



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?

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

http://www.langara.bc.ca/biology/mario/Assets/Geneticode.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 (1983)

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

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	**1st positions**	**2nd positions**	**3rd positions**
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
Nonsyonymous	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%

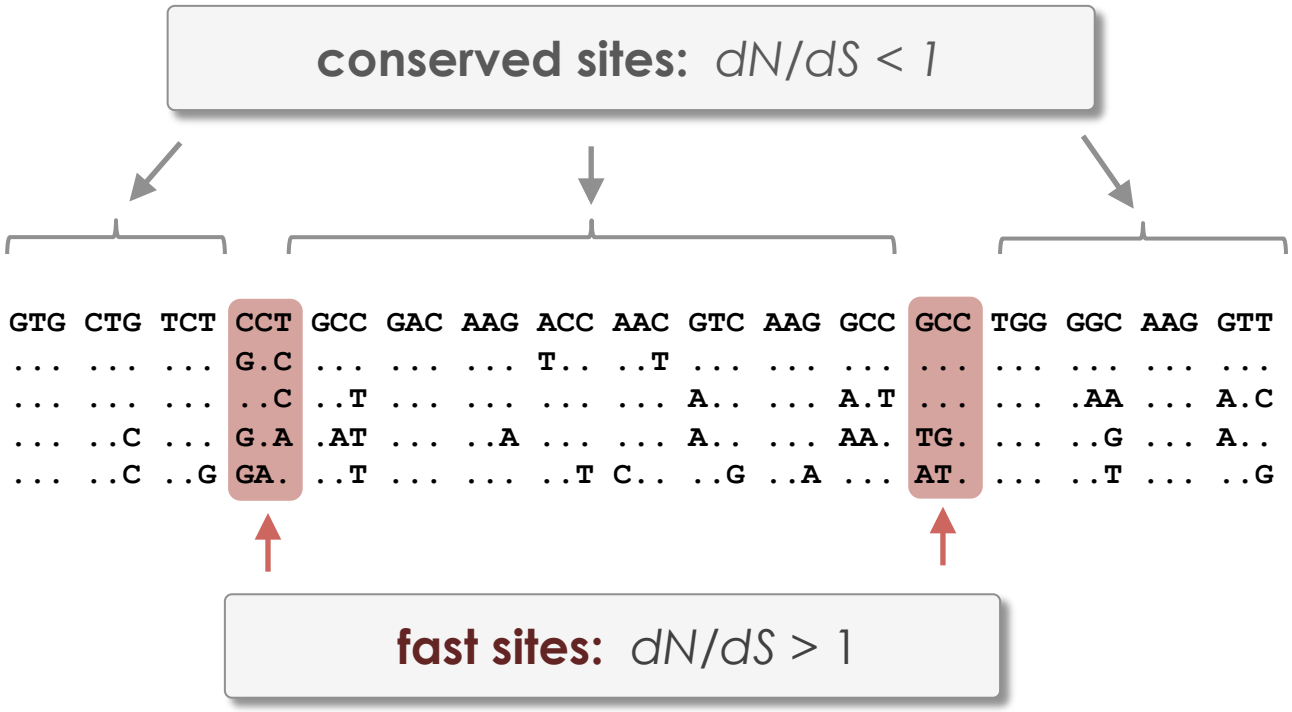
$N = 300 \times 3 \times 74.5\% = 670.5$

So, $d_S = 5/229.5 = 0.0218$

$d_N = 5/670.5 = 0.0075$

d_N/d_S (ω) = 0.34, **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

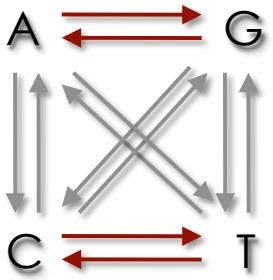
Relative proportion of different types of mutations in hypothetical protein coding sequence.				
	Expected number of changes (proportion)			
Type	**All 3 Positions**	**1st positions**	**2nd positions**	**3rd positions**
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)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



ts/tv = 2.71

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

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

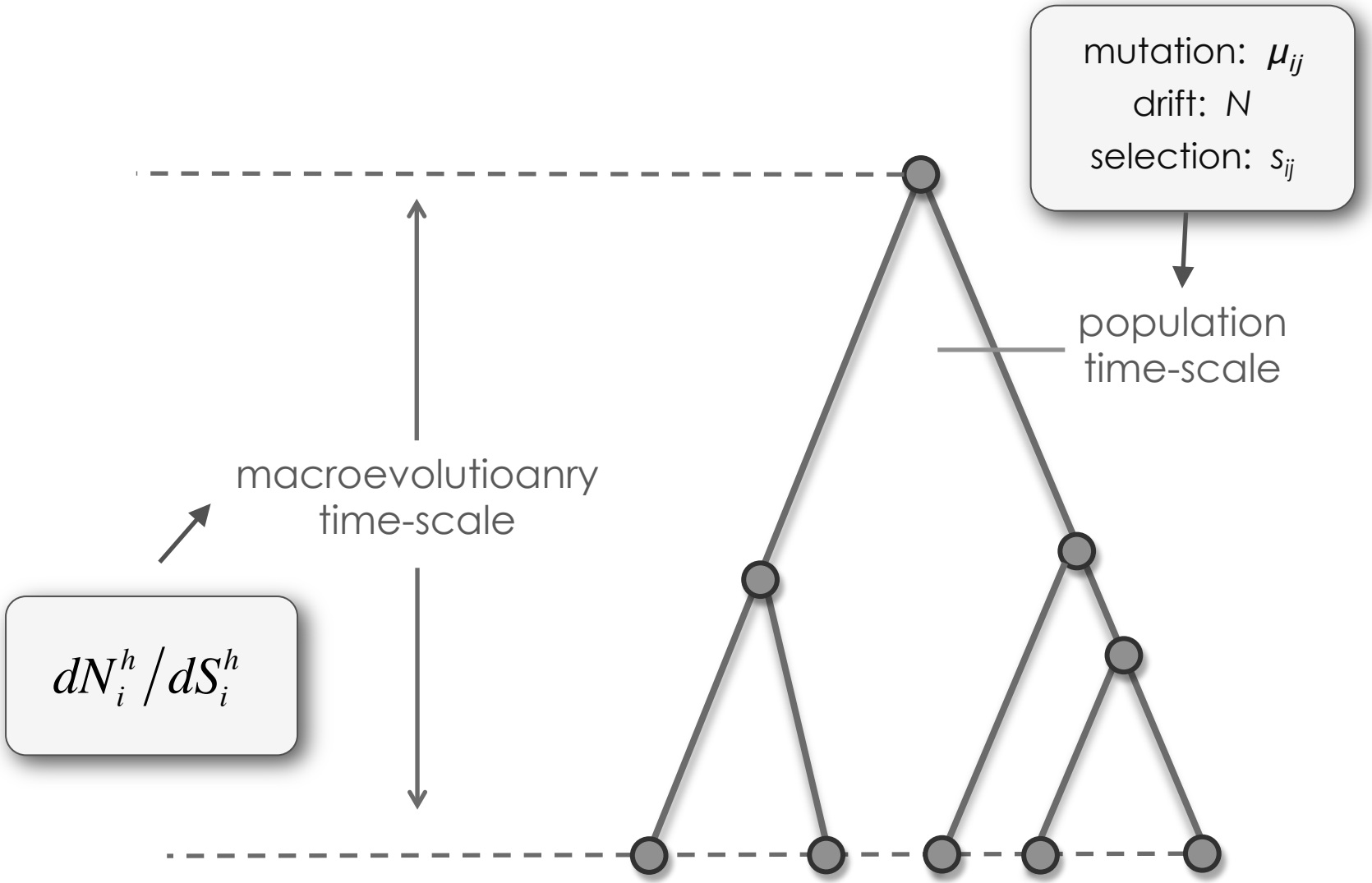
Phe F	TTT	0	Ser S	TCT	0	Tyr Y	TAT	1	Cys C	TGT	0
	TTC	27		TCC	15		TAC	22		TGC	6
Leu L	TTA	0		TCA	0	*** *	TAA	0	*** *	TGA	0
	TTG	1		TCG	1		TAG	0	Trp W	TGG	8
Leu L	CTT	2	Pro P	CCT	1	His H	CAT	0	Arg R	CGT	1
	CTC	2		CCC	15		CAC	4		CGC	7
	CTA	0		CCA	3	Gln Q	CAA	0		CGA	0
	CTG	29		CCG	1		CAG	14		CGG	0

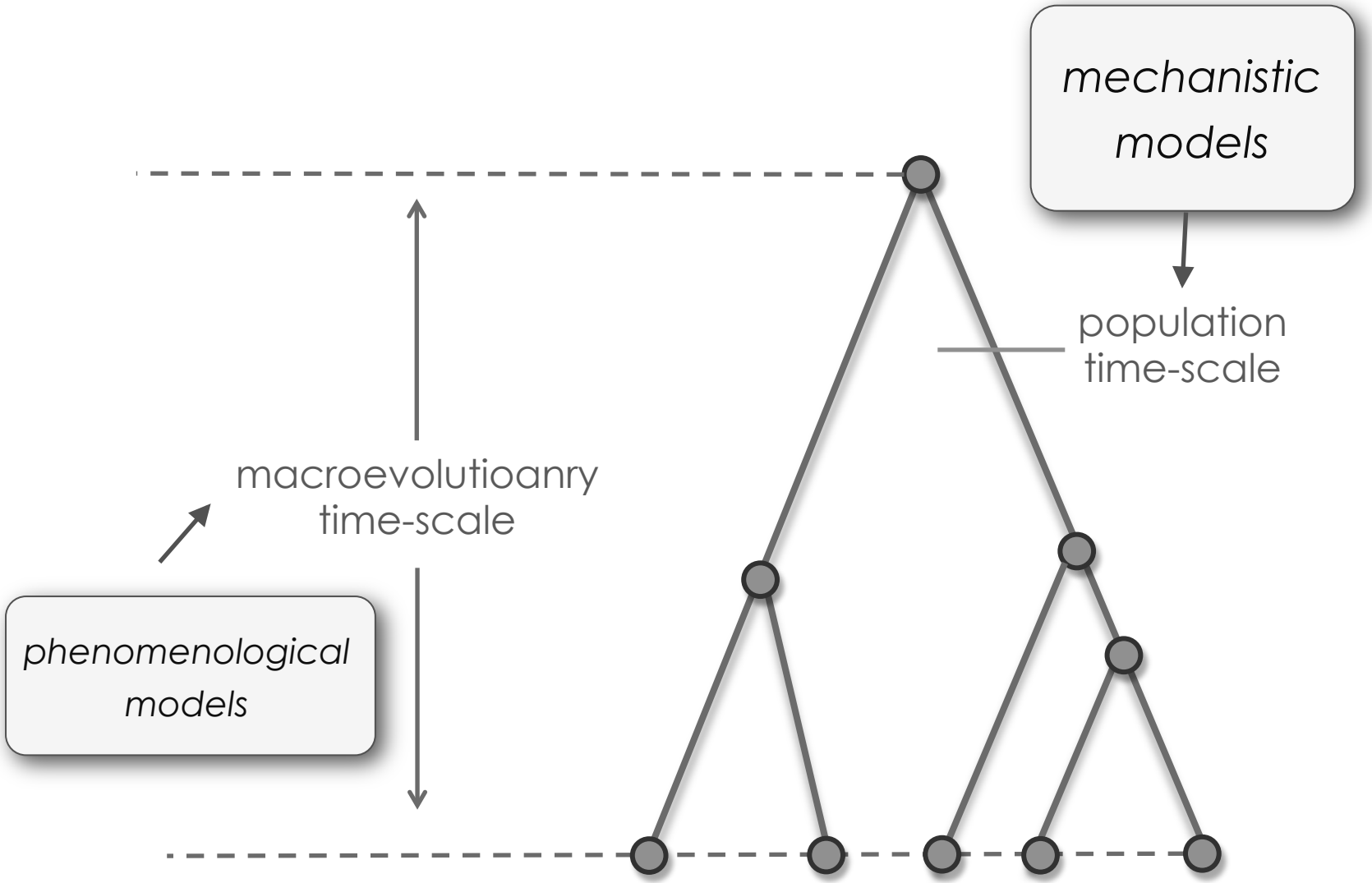
an index of selection pressure acting on the protein

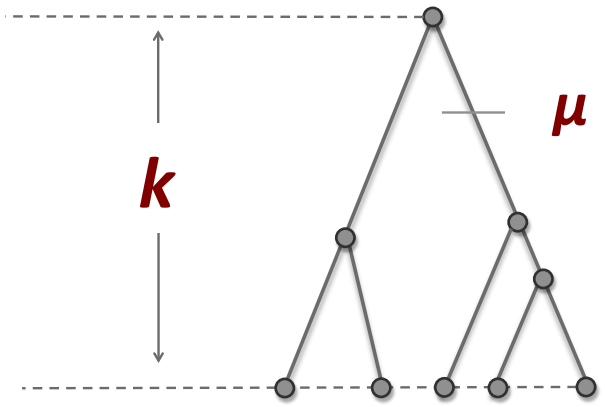
$$\omega = \frac{dN}{dS}$$

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**!


mutation: μ_{ij}
drift: N
selection: s_{ij}
population time-scale
macroevolutioanry time-scale
dN_i^h / dS_i^h


mechanistic models
population time-scale
macroevolutioanry time-scale
phenomenological models



mechanistic models

"MutSel models"

$$\Pr = \begin{cases} \mu_{ij} & \text{if neutral} \\ \mu_{ij} N \times \dfrac{2 s_{ij}}{1 - e^{-2 N s_{ij}}} & \text{if selected} \end{cases}$$

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

Halpern and Bruno (1998)

- Wright-Fisher population
- drift: ***N***
- mutation: ***μ***
- selection: s_{ij}
- s_{ij} vary among sites AND amino acids
- expected dN^h/dS^h

fixation probability with selection

population genetics at a single codon site (*h*)

fitness coefficients

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

selection coefficients

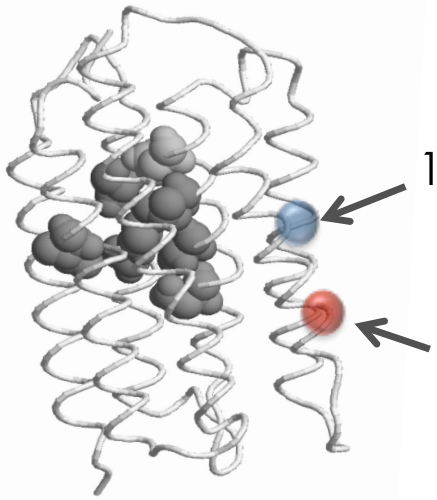
$$s_{ij}^h = f_j^h - f_i^h$$

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)

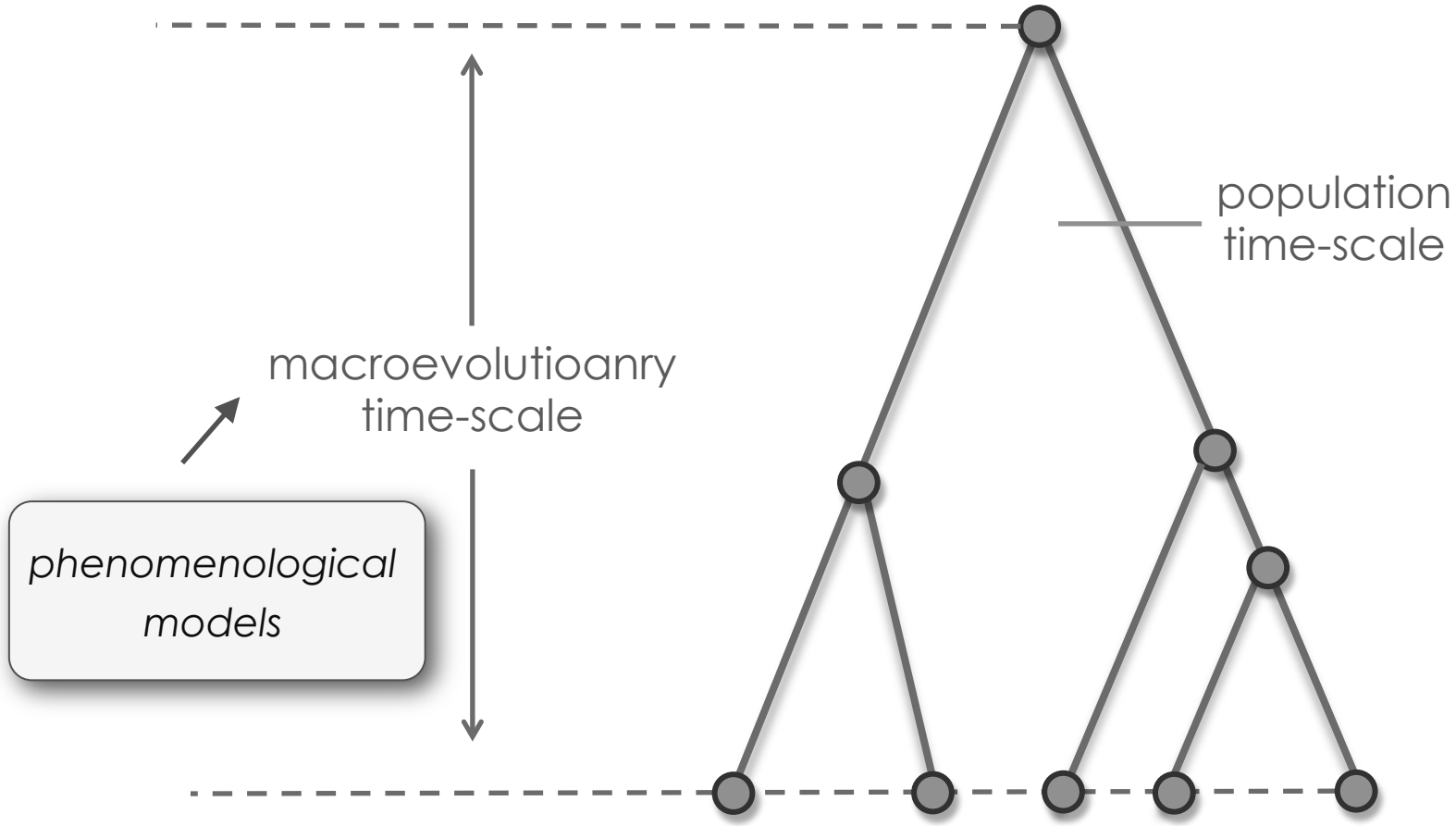


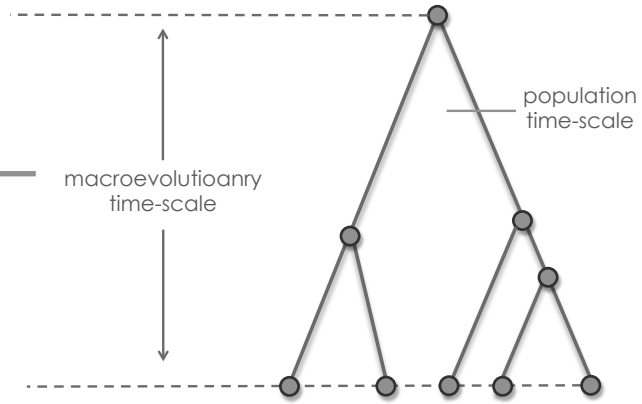
1. **A**TA (**Ile**) ➔ **T**TA (**Leu**): $\Delta f^{h}_{Ile \to Leu}$ (**conservative**)

2. A**T**A (**Ile**) ➔ A**A**A (**Lys**): $\Delta f^{h}_{Ile \to Lys}$ (**radical**)

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)


population
time-scale
macroevolutioanry
time-scale
phenomenological
models



phenomenological models

"OMEGA MODELS"

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by} > 1 \\ \pi_j & \text{for synonymous tv.} \\ \kappa\pi_j & \text{for synonymous ts.} \\ \omega\pi_j & \text{for non-synonymous tv.} \\ \omega\kappa\pi_j & \text{for non-synonymous ts.} \end{cases}$$

Goldman and Yang (1994)
Muse and Gaut (1994)

- phenomenological parameters
- ts/tv ratio: κ
- codon frequencies: π_j
- $\omega = dN/dS$
- parameter estimation via ML
- *stationary process*

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

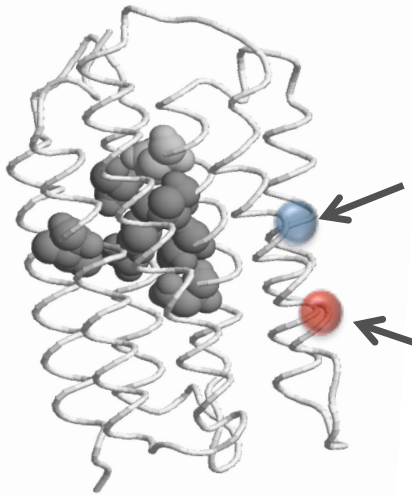
phenomenological codon models: just a few parameters are needed to cover the 3721 transitions between codons!

	to codon below:						⋯▸	
From 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 $\boldsymbol{\omega}$ is the ratio d_N/d_S, $\boldsymbol{\kappa}$ is the transition/transversion rate ratio, and $\boldsymbol{\pi_i}$ is the equilibrium frequency of the target codon (i).

substitution probability with selection

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



1. **A**TA (**Ile**) ➔ **T**TA (**Leu**): $\omega_{ile \rightarrow leu}$ (**conservative**)

2. ATA (**Ile**) ➔ A**A**A (**Lys**): $\omega_{ile \rightarrow lys}$ (**radical**)

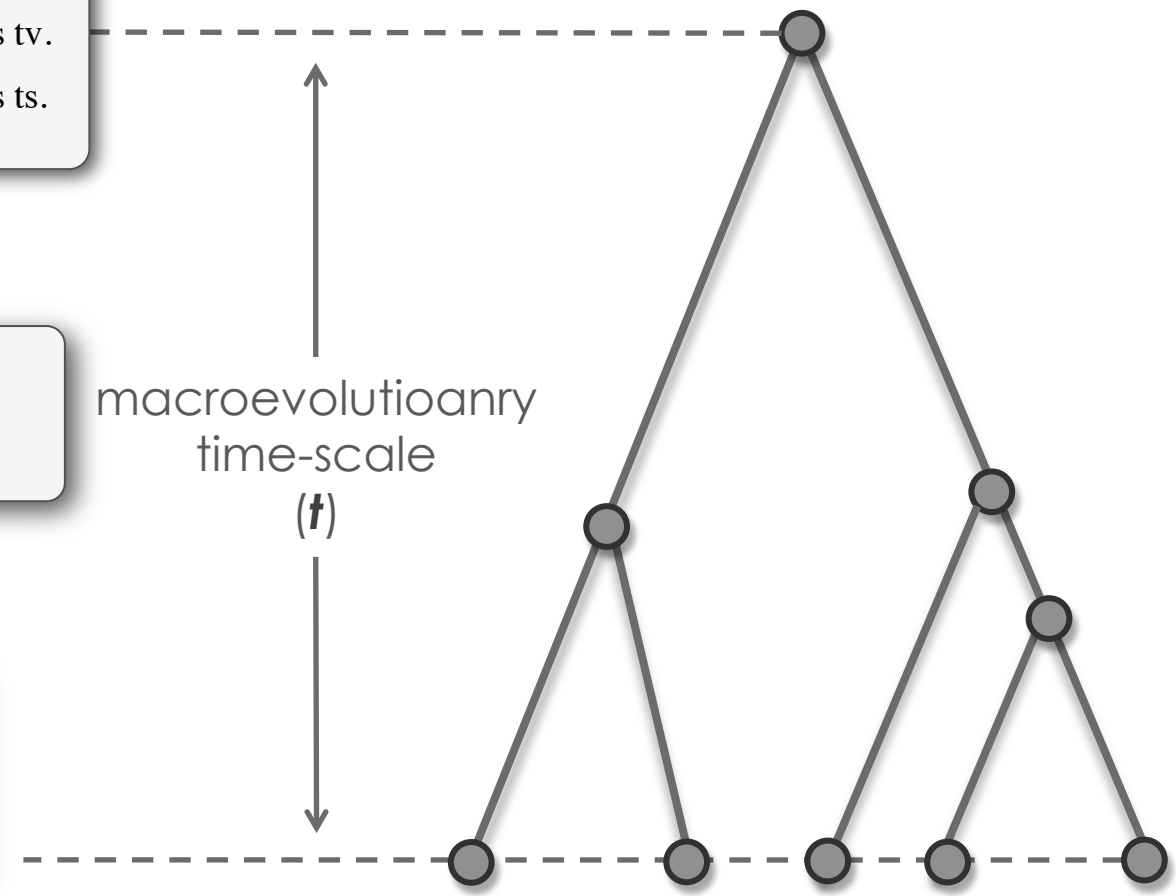
contradiction? selection should favour amino acids with higher fitness.

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

$$Q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by} > 1 \\ \pi_j & \text{for synonymous tv.} \\ \kappa\pi_j & \text{for synonymous ts.} \\ \omega\pi_j & \text{for non-synonymous tv.} \\ \omega\kappa\pi_j & \text{for non-synonymous ts.} \end{cases}$$

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

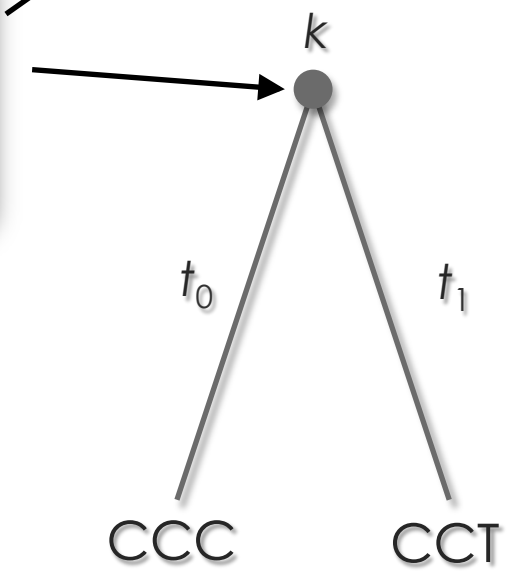
recall that **Paul Lewis** introduced **Q matrices** and how to obtain **transition probabilities**



likelihood of the data at a site

$$L_h(CCC,CCT) = \sum_k \pi_k p_{kCCC}(t_0) p_{kCCT}(t_1)$$

the likelihood is a **sum over all possible ancestral codon states** that could have been observed at node k



recall that **Paul Lewis** described how to compute the likelihood of the data at a site for a DNA model. The only difference here is that the states are codons rather than nucleotides

note: analysis is typically done by using an unrooted tree

likelihood of the data at all sites

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

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

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

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

The log likelihood is a sum over all sites.

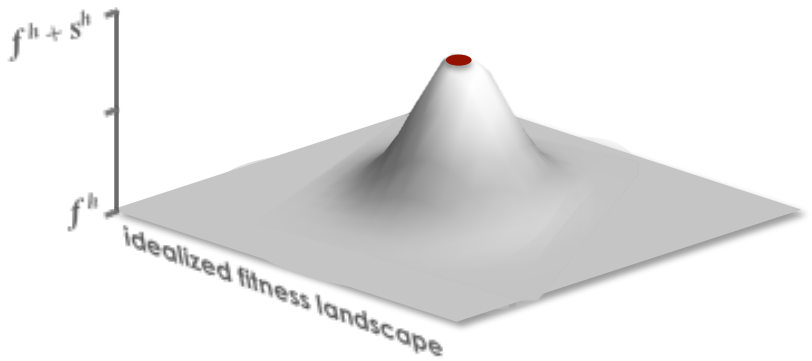
$$\ell = \ln\{L\} = \ln\{L_1\} + \ln\{L_2\} + \ln\{L_3\} + \ldots + \ln\{L_N\} = \sum_{h=1}^{N} \ln\{L_h\}$$

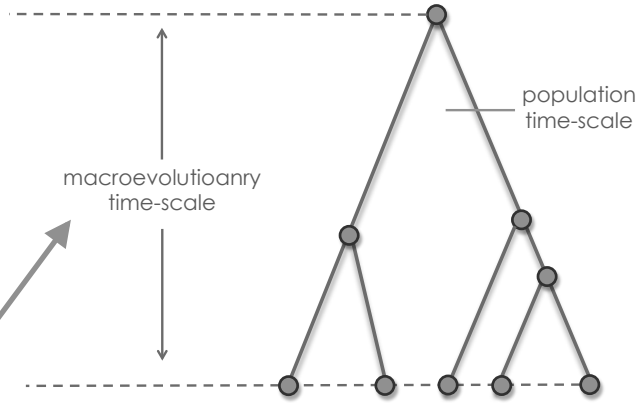
we made some progress…

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!

part 2: mechanistic processes of codon evolution





mechanistic models

"MutSel framework"

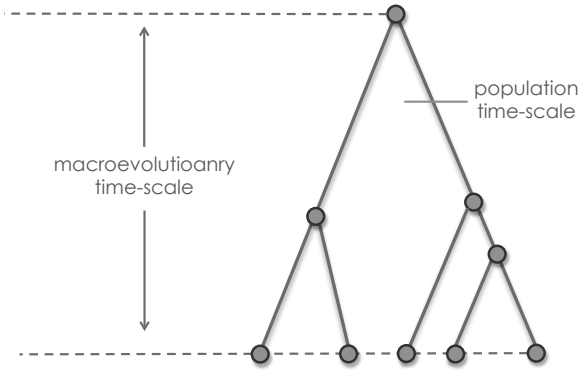
$$A_{ij}^{h} = \begin{cases} \mu_{ij} & \text{if } s_{ij}^{h} = 0 \\ \mu_{ij} N \times \dfrac{2 s_{ij}^{h}}{1 - \mathrm{e}^{-2 N s_{ij}^{h}}} & \text{otherwise} \end{cases}$$

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

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

phenomenological models

site-specific MutSel rate matrix



MutSel rate matrix

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

- 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)

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

1. map fitness to equilibrium frequencies

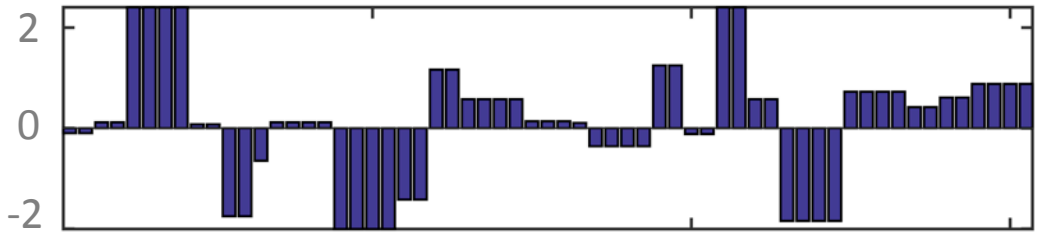
2. macroevolution index of selection intensity

(1) Sella and Hirsh 2005; (2) Jones et al. 2016

1. fitness coefficients map to stationary codon frequencies

fitness coefficients

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



codon frequencies

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

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* = ω 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

1. frequency dependent selection
2. episodic adaptation

dynamic fitness landscape

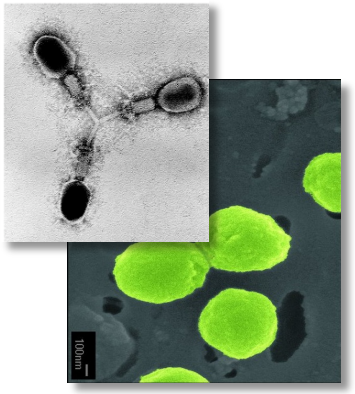
3. shifting balance

static fitness landscape

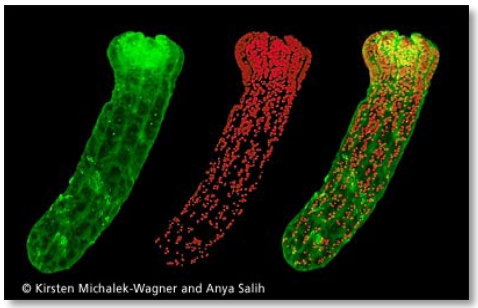
scenario 1: frequency dependent selection

1. antagonistic evolutionary interaction

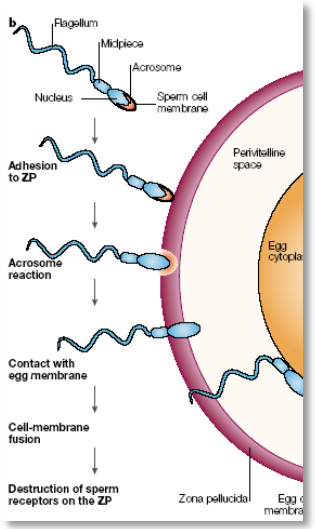
host-pathogen



molecular-interactions

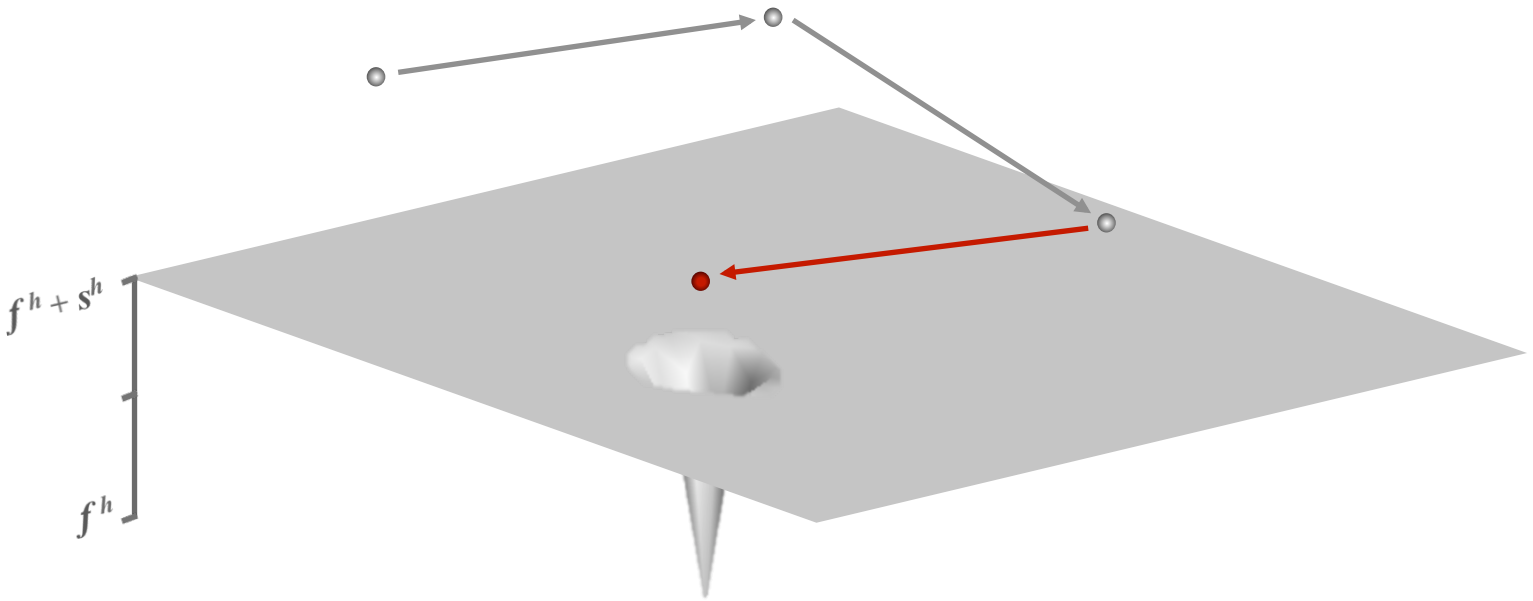


sexual-conflict



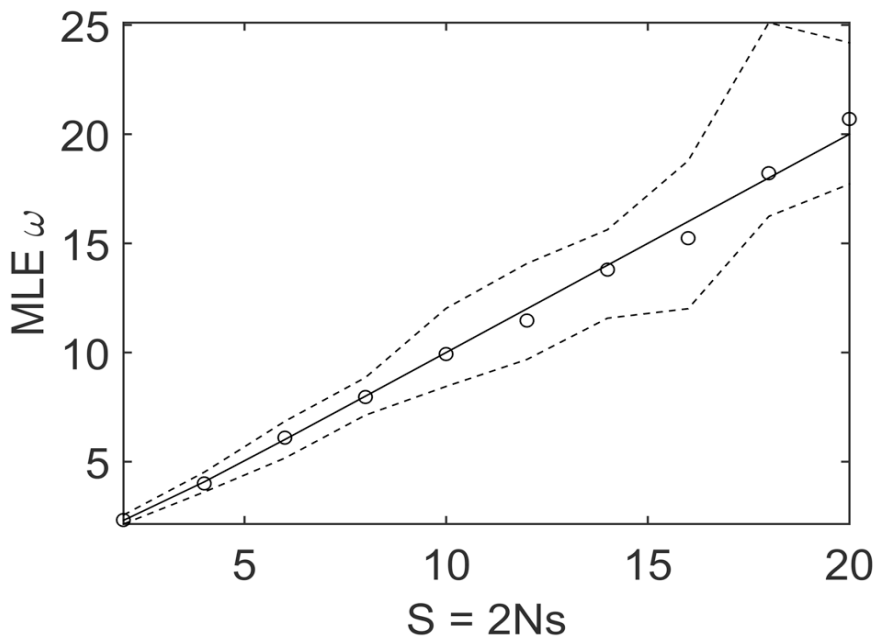
frequency-dependent selection: MutSelM0

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 = O

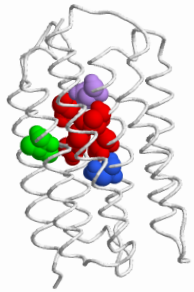
conclusion: phenomemologcial codon models assume frequency-dependent selection

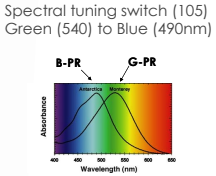
[dos Reis (2015); Jones et al. (2016)]

scenario 2: adaptive peak shift

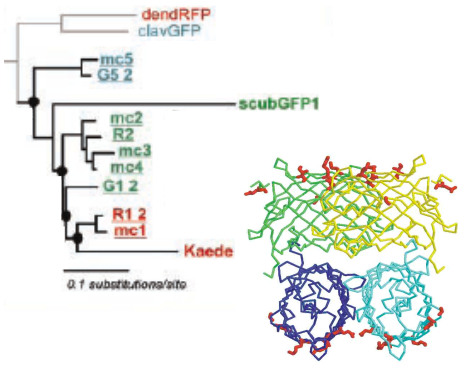
2. episodic Darwinian adaptation

exploitation of a new niche

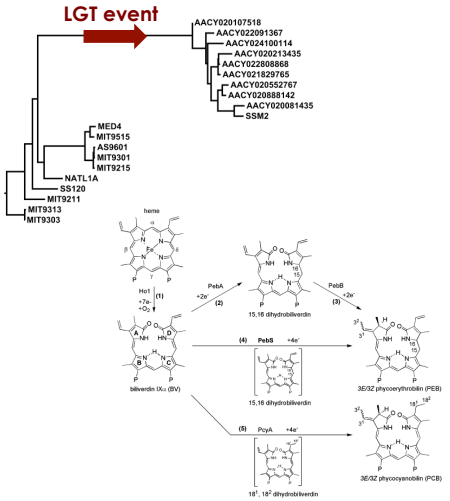




gene duplication

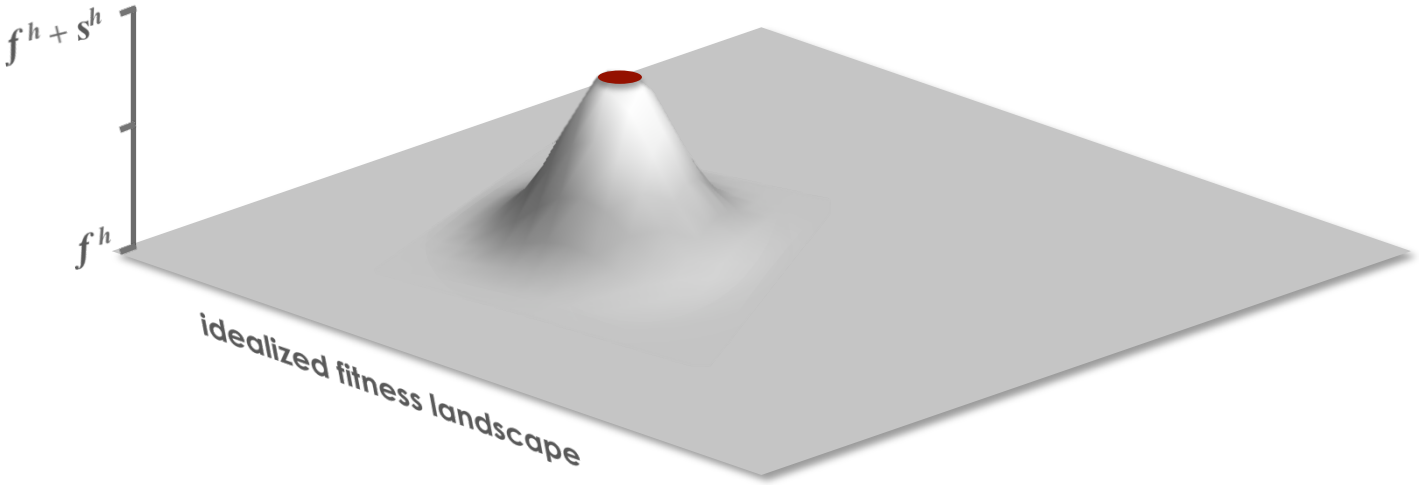


lateral gene transfer (LGT)



adaptive peak shift: evolution of novel function

optimal function in a stable environment



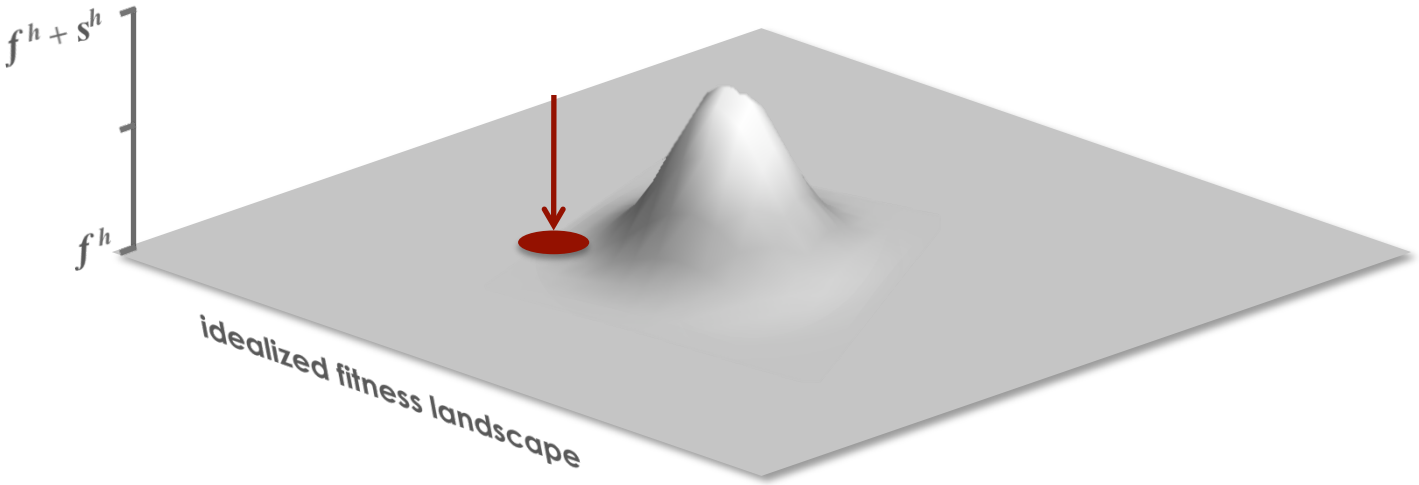
population: at fitness peak

fitness peak: stationary

FFTNS: keeps population at peak

adaptive peak shift: evolution of novel function

sub-optimal function in a novel environment



population: lower fitness

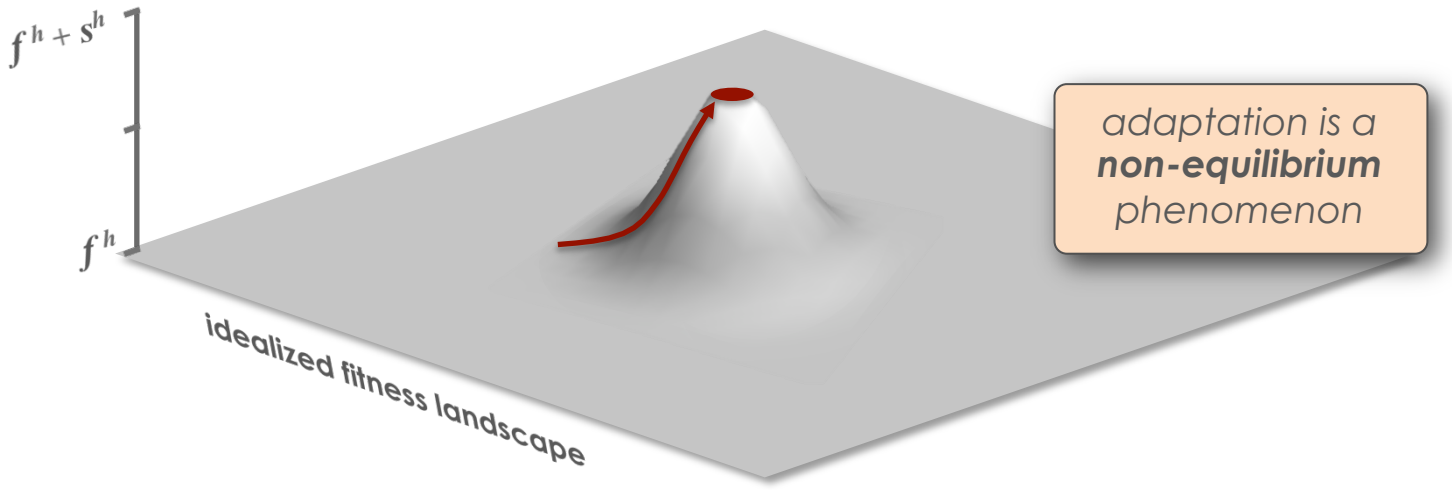
fitness peak: moving

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

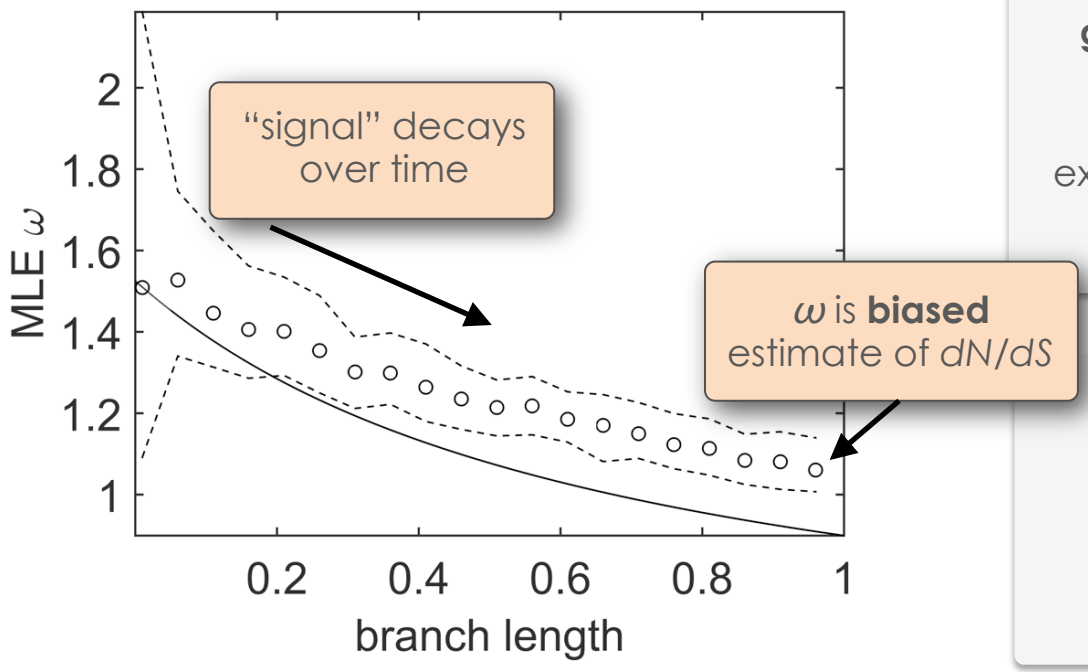
Mario dos Reis

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



generating process:

MutSelES

expectation = dN^h/dS^h

symbol = ——

fitted model:

model M0

inference = MLE ω

symbol = O

conclusion : episodic models "work" because w>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

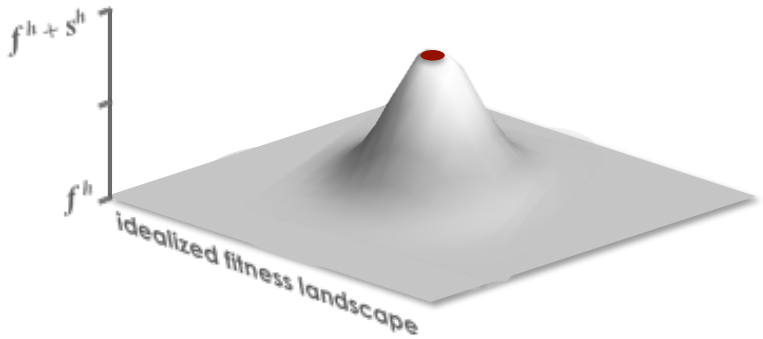
[dos Reis (2015); Jones et al. (2016)]

Scenario 3: non-adaptive evolution

3. fitness coefficients are constant (fixed-peak)

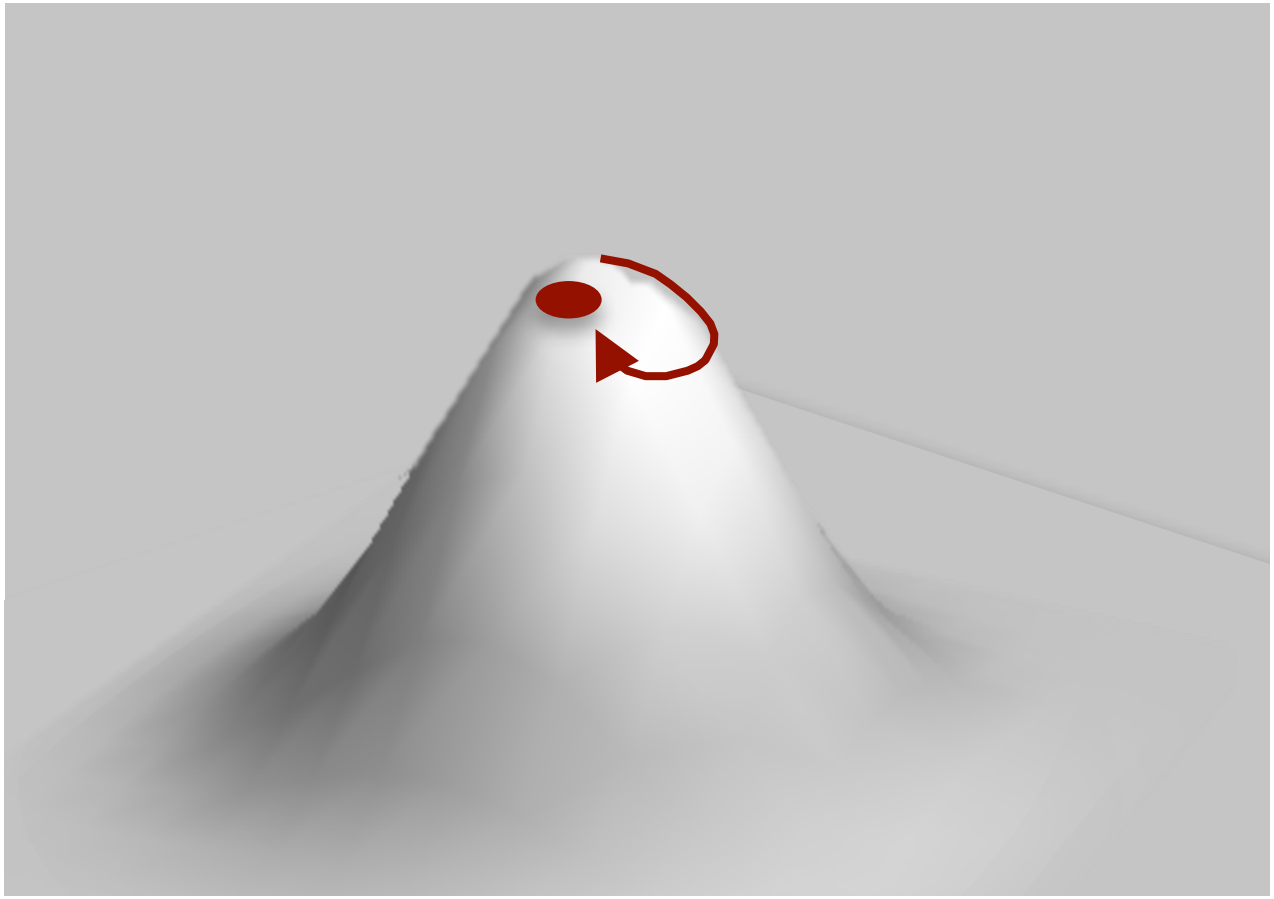
Spielman and Wilke (2015)

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



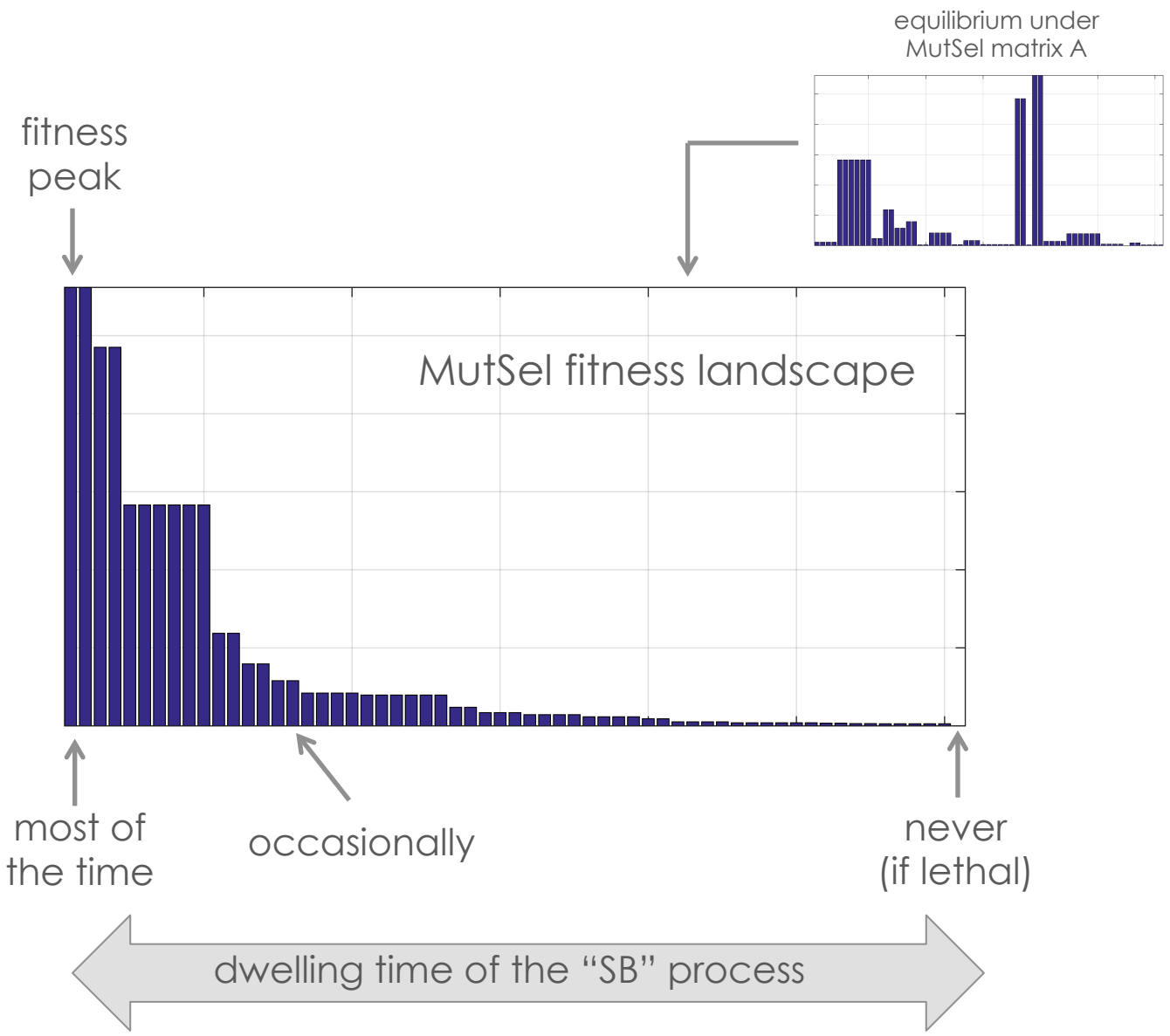
[Spielman and Wilke, (2015); Jones at al., (2016)]

shifting balance: movement around peak

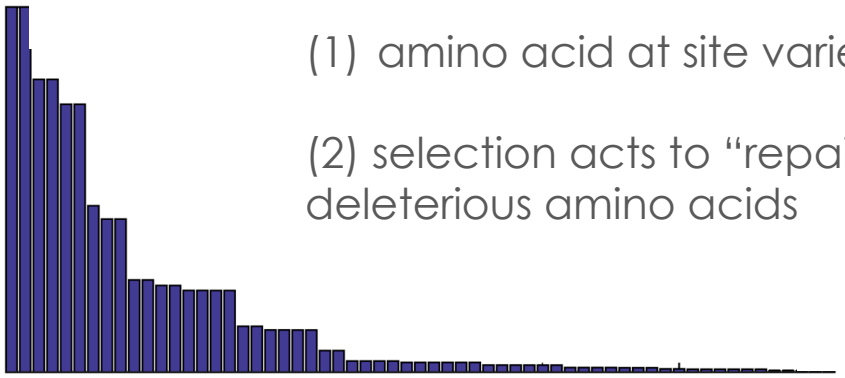


mutation and **drift** can move a pop. off a fitness peak

shifting balance: the MutSel landscape (Jones at el. 2016)



shifting balance: positive selection on a MutSel landscape



(1) amino acid at site varies over time

(2) selection acts to "repair" shifts to deleterious amino acids

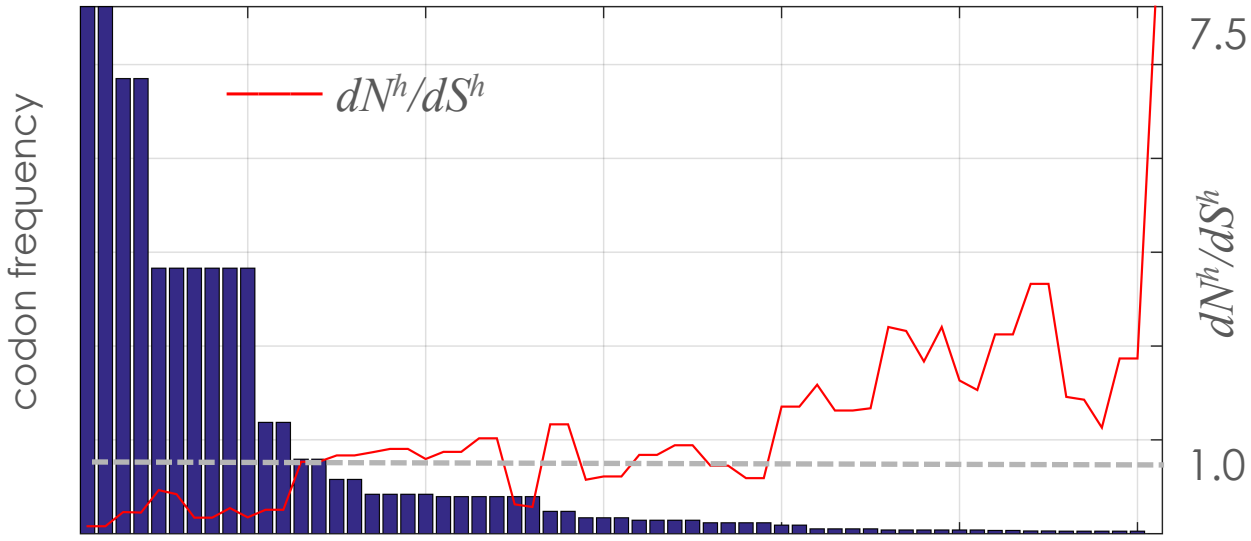
EXPECTED PROPORTION OF MUTATIONS FIXED BY SELECTION

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

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

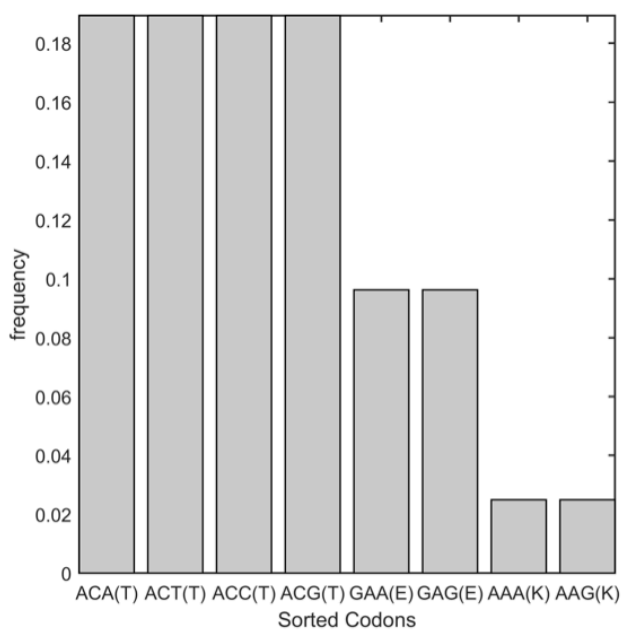


temporal average dN^h/dS^h = 0.61

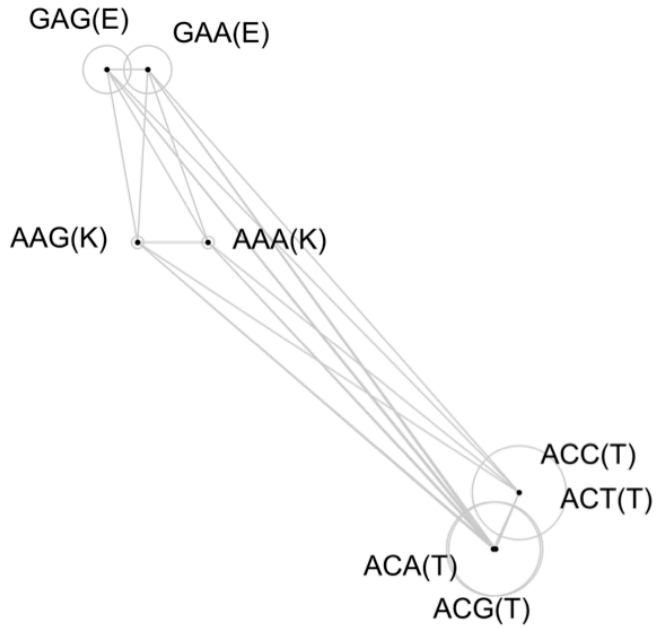
conclusion: positive selection operates on a stationary fitness peak in the same way as when there is an adaptive peak shift

landscapes have unique structures

MutSel landscape



McCandlish landscape

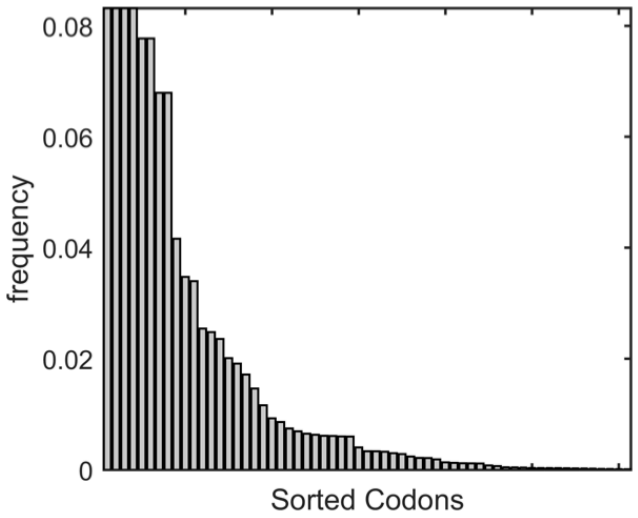


conclusion: A population can get to a sub-optimal codon (E) by drift and reside there for some time (b/c moving between T and E requires changes ≥ 2 codons).

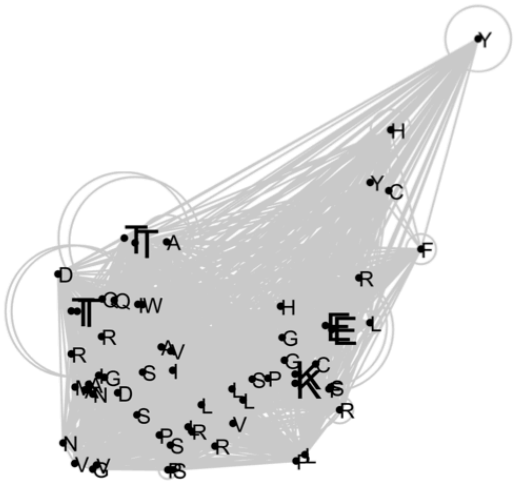
landscape structure depends on N

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

MutSel landscape



McCandlish landscape

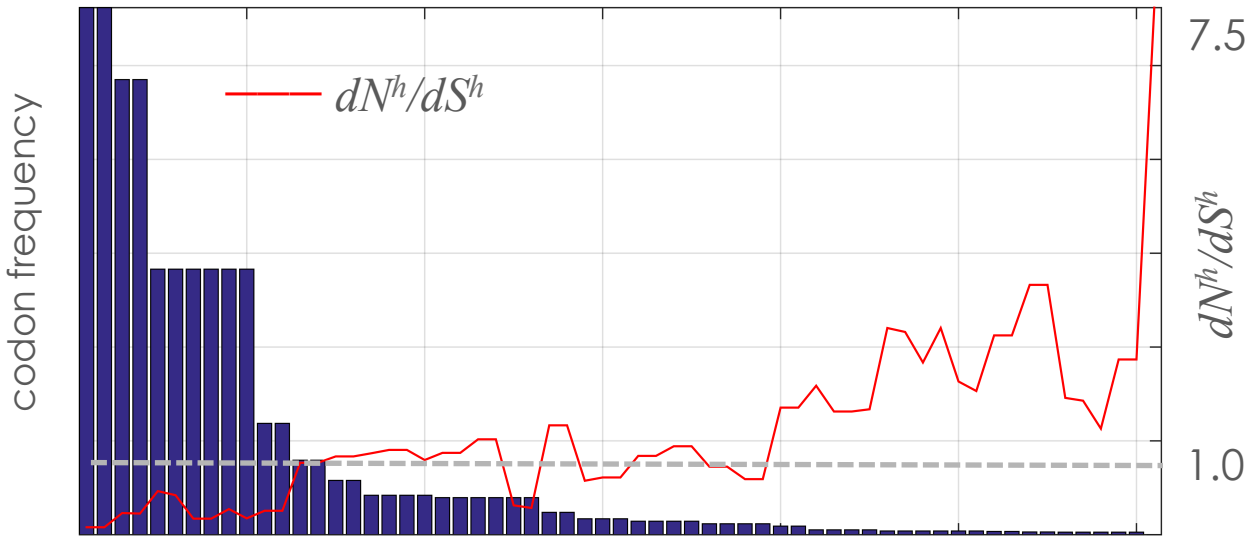


conclusion: decreasing N changes:

i. the "space" for shifting balance
ii. mean dN/dS
iii. equilibrium frequencies

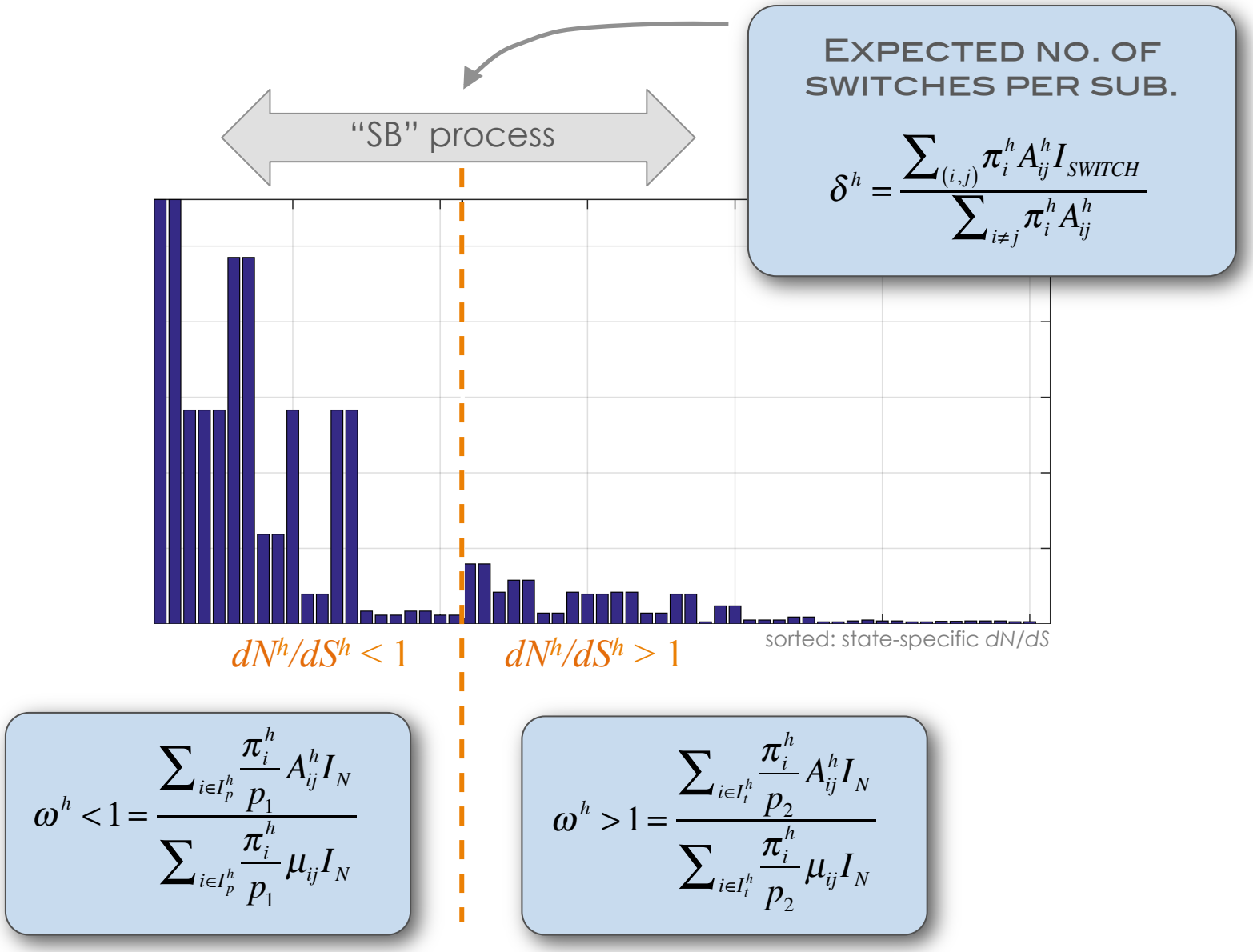
shifting balance: the MutSel landscape

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

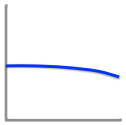
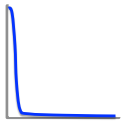


temporal average dN^h/dS^h = 0.61

shifting balance: a mechanistic model



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	>>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 f^{n}

σ: {0.0001, 0.001, 0.01}

N = 1000

gene sequences

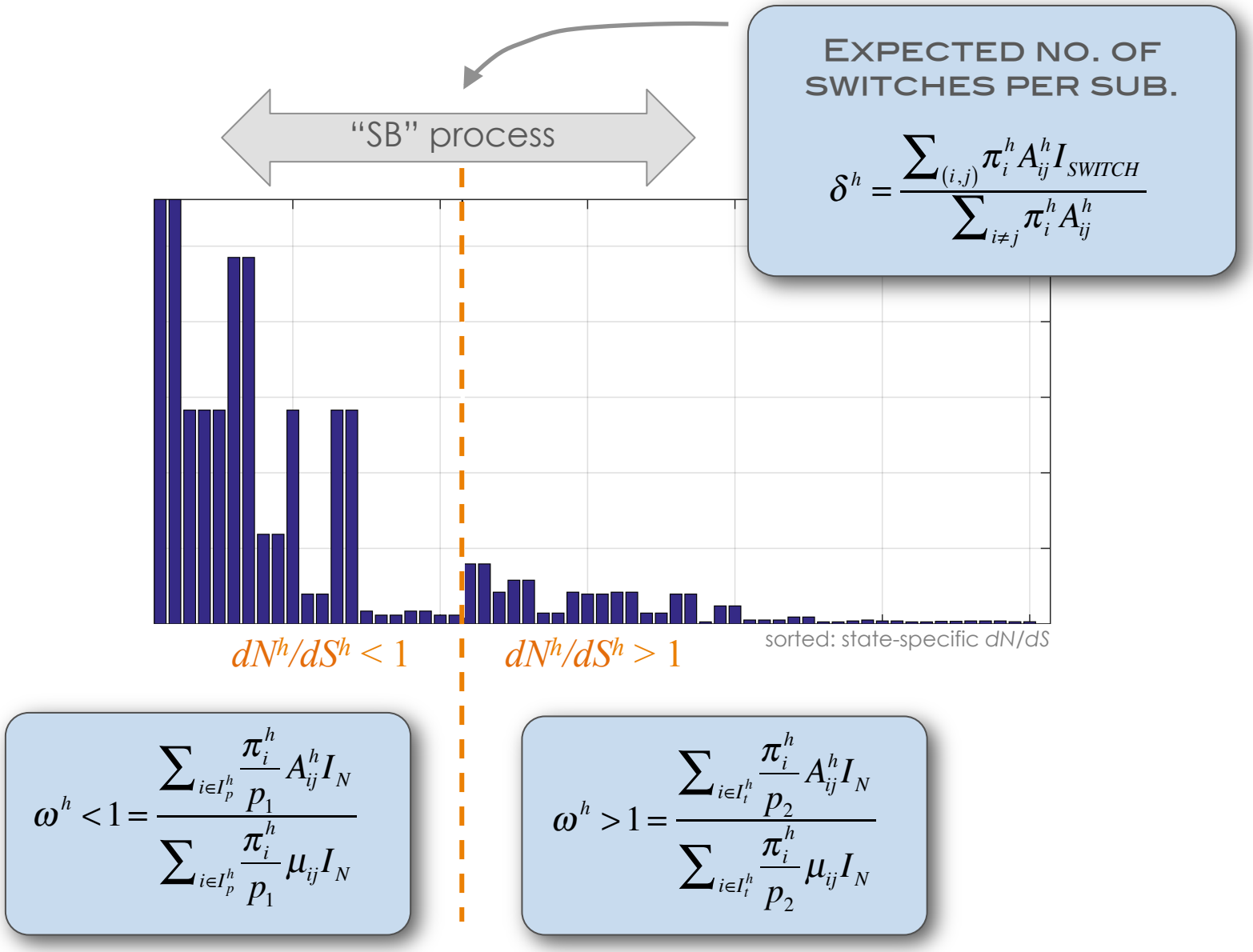
```
human
cow
rabbit
rat
opossum

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

shifting balance: a mechanistic model



covarion-like model of evolution

$$Q = \begin{pmatrix} \textbf{evolutionary regime 1}: \omega_1 = \text{low} \ (\text{“near the peak”}) & \textbf{switching process:} \ \omega_1 \rightarrow \omega_2 \\ \textbf{switching process:} \ \omega_2 \rightarrow \omega_1 & \textbf{evolutionary regime 2}: \omega_1 = \text{high} \ (\text{“in the tail”}) \end{pmatrix}$$

[Guindon et al., (2004) ; Jones at al. (2016)]

covarion-like model of evolution
(phenomenological)

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

LOW (ω_1) HIGH (ω_2)

p_1: proportion of time sites are in ω_1

p_2: proportion of time sites are in ω_2

switching: δ

the covarion-like codon model can be **fit to real data**

shifting balance: a mechanistic model

landscapes:

250 f^h

σ: {0.0001, 0.001, 0.01}

N = 1000

	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	>>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"**

This **"signal"** is detectable with covarion and branch-site codon models!

recall: no adaptive evolution in this case (stationary fitness peak)!!!

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)

REFERNCES

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. Biology letters, 11(4), 20141031.

Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular biology and evolution, 11(5), 725-736.

Guindon, S., Rodrigo, A. G., Dyer, K. A., & Huelsenbeck, J. P. (2004). Modeling the site-specific variation of selection patterns along lineages. Proceedings of the National Academy of Sciences, 101(35), 12957-12962.

Halpern, A. L., & Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Molecular biology and evolution, 15(7), 910-917.

Kimura, Motoo (1968). "Evolutionary rate at the molecular level". Nature. 217: 624–626

Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge University Press.

Jones, C. T., Youssef, N., Susko, E., & Bielawski, J. P. (2016). Shifting Balance on a Static Mutation–Selection Landscape: A Novel Scenario of Positive Selection. Molecular biology and evolution, 34(2), 391-407.

Muse, S. V., & Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular biology and evolution, 11(5), 715-724.

Sella, G., & Hirsh, A. E. (2005). The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences, 102(27), 9541-9546.

Spielman, S. J., & Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. Molecular biology and evolution, 32(4), 1097-1108.