codon substitution models and the analysis of natural selection pressure

Joseph P. Bielawski

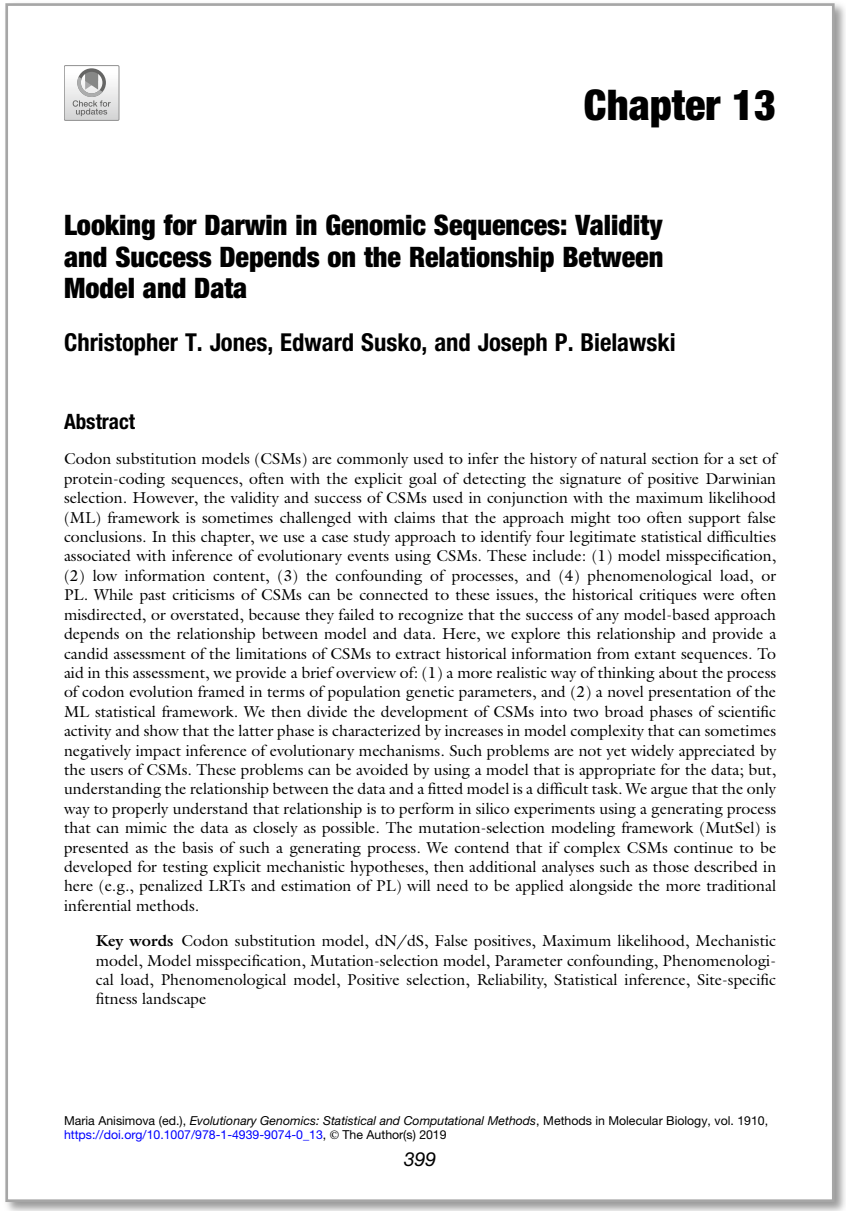
Department of Biology
Department of Mathematic and Statistics
Institute of Comparative genomics
Dalhousie University, Halifax, Nova Scotia, Canada

Details matter!

Codon models: waaaaay **too much** to cover in this talk

Macro-evolutionary inference of selection intensity:

- Very complex and diverse modelling strategies
- Deep statistical issues
- Model testing and interpretation
- Strong opinions about “the right thing to do”



Chapter 13

Looking for Darwin in Genomic Sequences: Validity and Success Depends on the Relationship Between Model and Data

Christopher T. Jones, Edward Susko, and Joseph P. Bielawski

Abstract

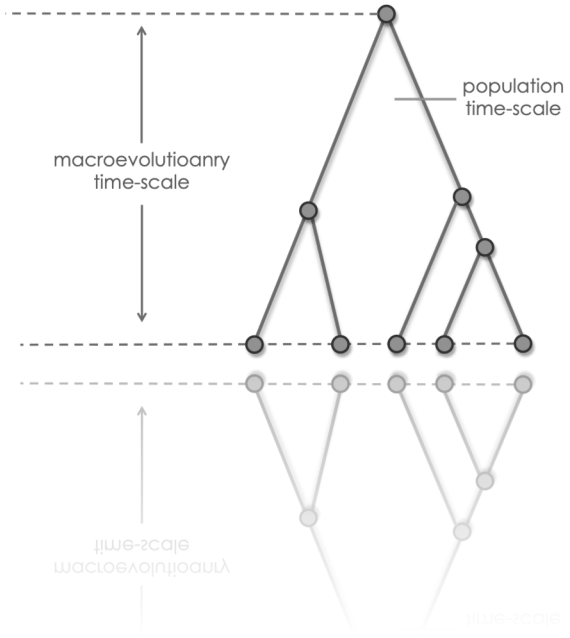
Codon substitution models (CSMs) are commonly used to infer the history of natural section for a set of protein-coding sequences, often with the explicit goal of detecting the signature of positive Darwinian selection. However, the validity and success of CSMs used in conjunction with the maximum likelihood (ML) framework is sometimes challenged with claims that the approach might too often support false conclusions. In this chapter, we use a case study approach to identify four legitimate statistical difficulties associated with inference of evolutionary events using CSMs. These include: (1) model misspecification, (2) low information content, (3) the confounding of processes, and (4) phenomenological load, or PL. While past criticisms of CSMs can be connected to these issues, the historical critiques were often misdirected, or overstated, because they failed to recognize that the success of any model-based approach depends on the relationship between model and data. Here, we explore this relationship and provide a candid assessment of the limitations of CSMs to extract historical information from extant sequences. To aid in this assessment, we provide a brief overview of: (1) a more realistic way of thinking about the process of codon evolution framed in terms of population genetic parameters, and (2) a novel presentation of the ML statistical framework. We then divide the development of CSMs into two broad phases of scientific activity and show that the latter phase is characterized by increases in model complexity that can sometimes negatively impact inference of evolutionary mechanisms. Such problems are not yet widely appreciated by the users of CSMs. These problems can be avoided by using a model that is appropriate for the data; but, understanding the relationship between the data and a fitted model is a difficult task. We argue that the only way to properly understand that relationship is to perform in silico experiments using a generating process that can mimic the data as closely as possible. The mutation-selection modeling framework (MutSel) is presented as the basis of such a generating process. We contend that if complex CSMs continue to be developed for testing explicit mechanistic hypotheses, then additional analyses such as those described in here (e.g., penalized LRTs and estimation of PL) will need to be applied alongside the more traditional inferential methods.

Key words Codon substitution model, dN/dS, False positives, Maximum likelihood, Mechanistic model, Model misspecification, Mutation-selection model, Parameter confounding, Phenomenological load, Phenomenological model, Positive selection, Reliability, Statistical inference, Site-specific fitness landscape

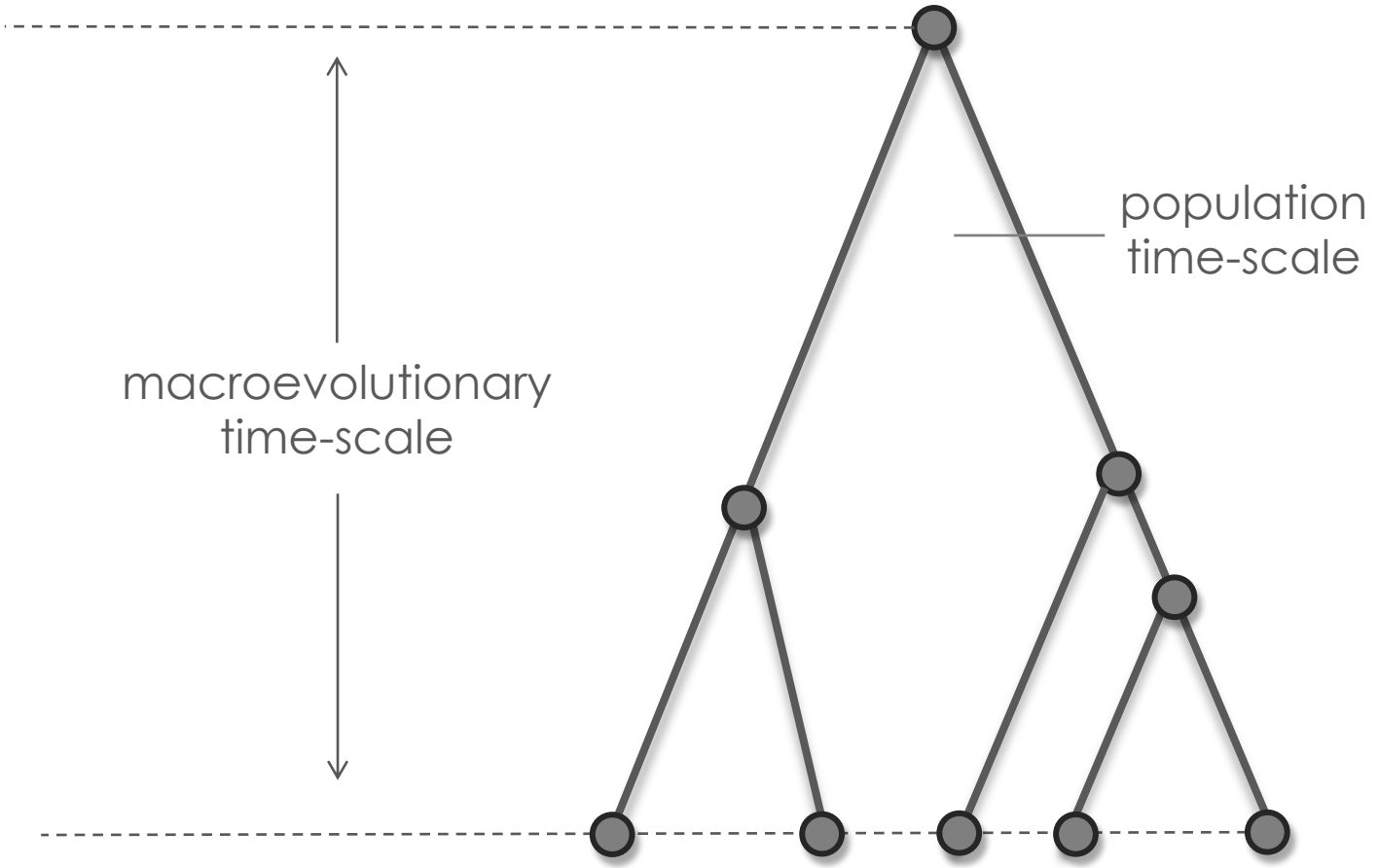
Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_13,

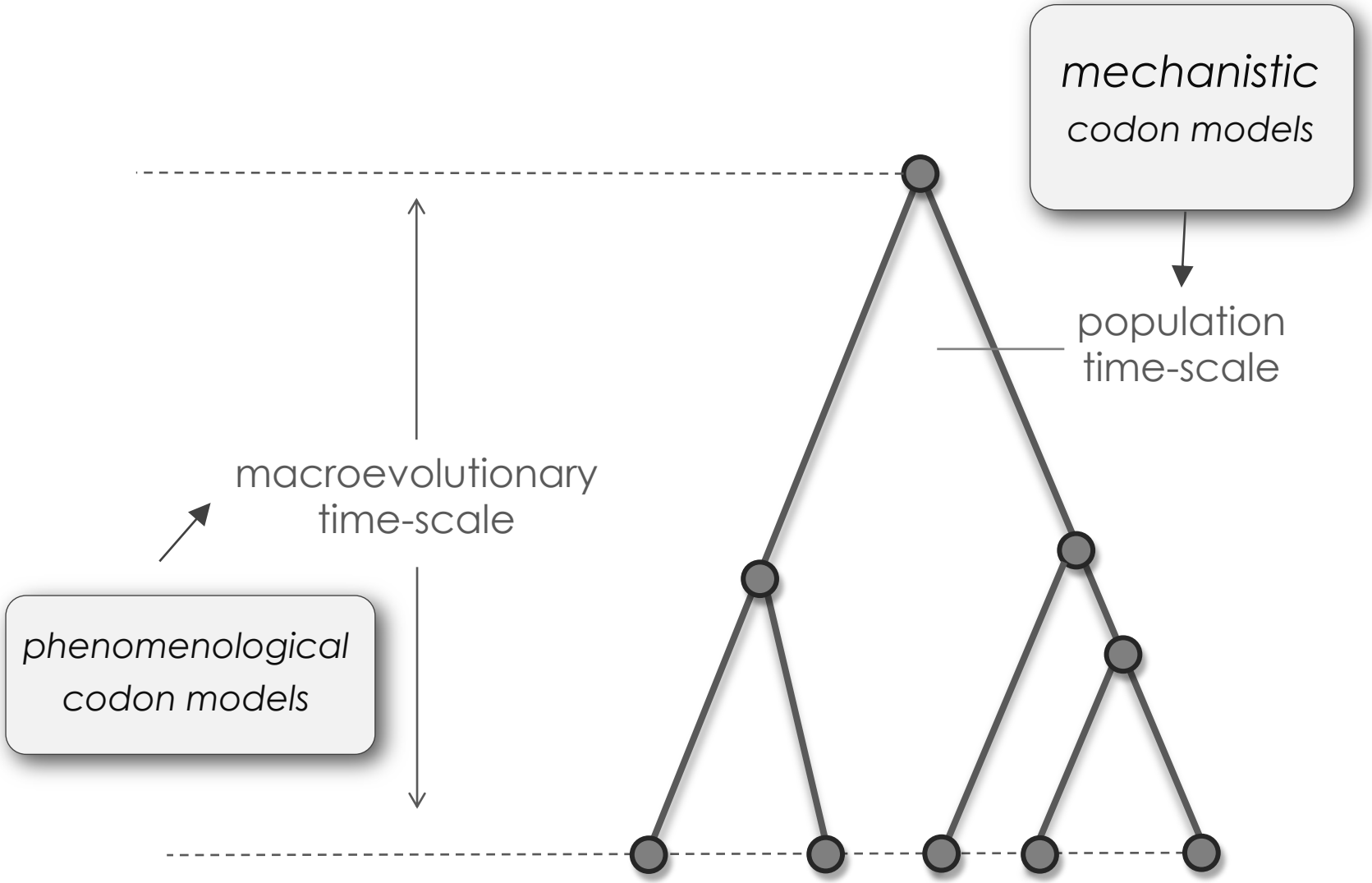
Details: **book chapter** (PDF) on course website

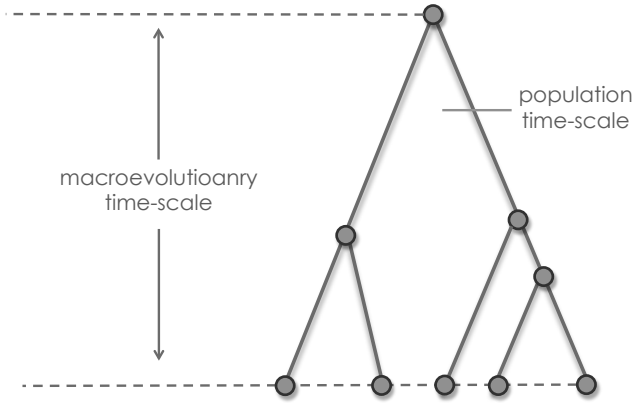
1. mechanistic codon models



reconciling evolutionary time scales




mechanistic
codon models
population
time-scale
macroevolutionary
time-scale
phenomenological
codon models



mechanistic codon models

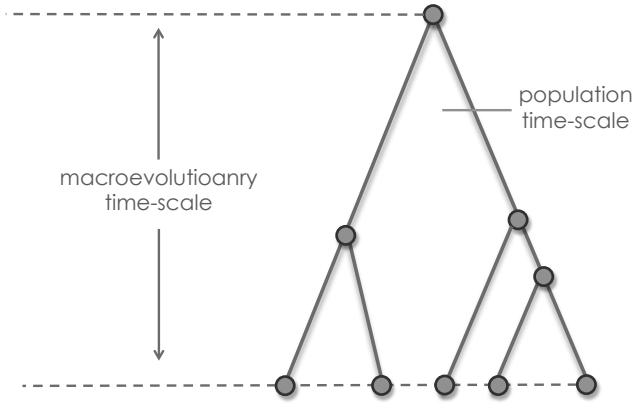
"MUTSEL MODELS"

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

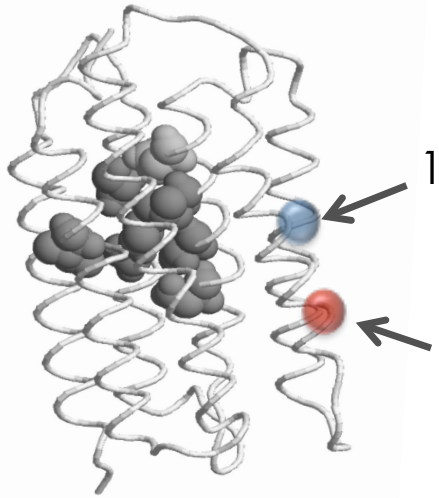
$$s_{ij} = \Delta f_{ij}$$

Halpern and Bruno (1998)

- Wright-Fisher population
- drift: ***N***
- mutation: $\boldsymbol{\mu}$
- selection: $\boldsymbol{s_{ij}}$
- s_{ij} vary among sites AND amino acids



mechanistic codon models

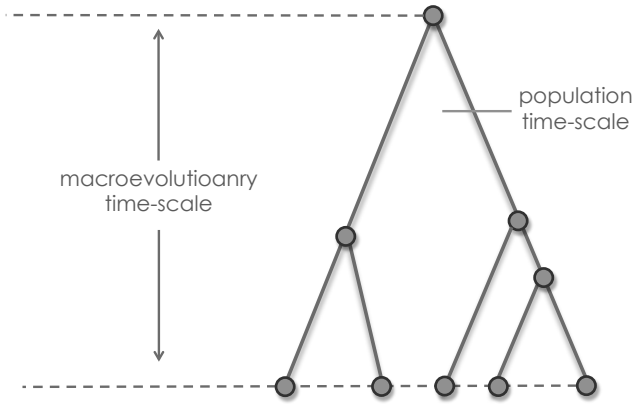


1. **A**TA (**Ile**) ➔ **T**TA (**Leu**): $\Delta f^{h}_{Ile \rightarrow Leu}$ (**conservative**)

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

- **realism**: fixation probability depends on fitness of ancestral and derived amino acids in the context of the protein.

- **the cost of realism**: **usually too complex to fit such a model to real data** (caveat: some versions will allow new ways to analyze big datasets)



mechanistic codon models

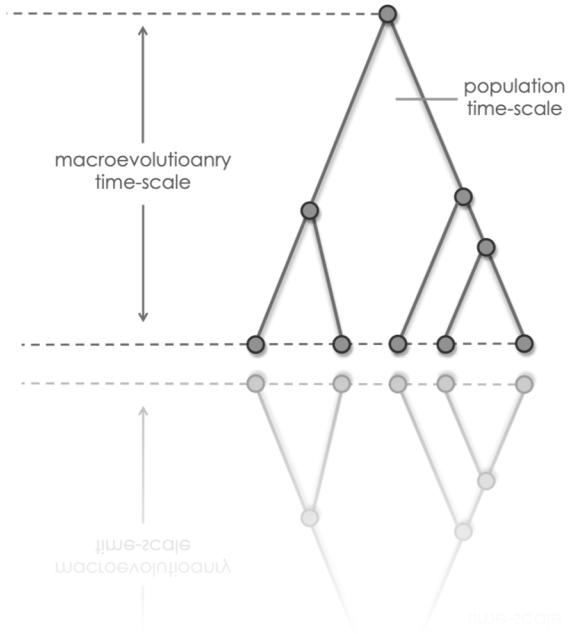
population genetics at a single codon site (*h*)

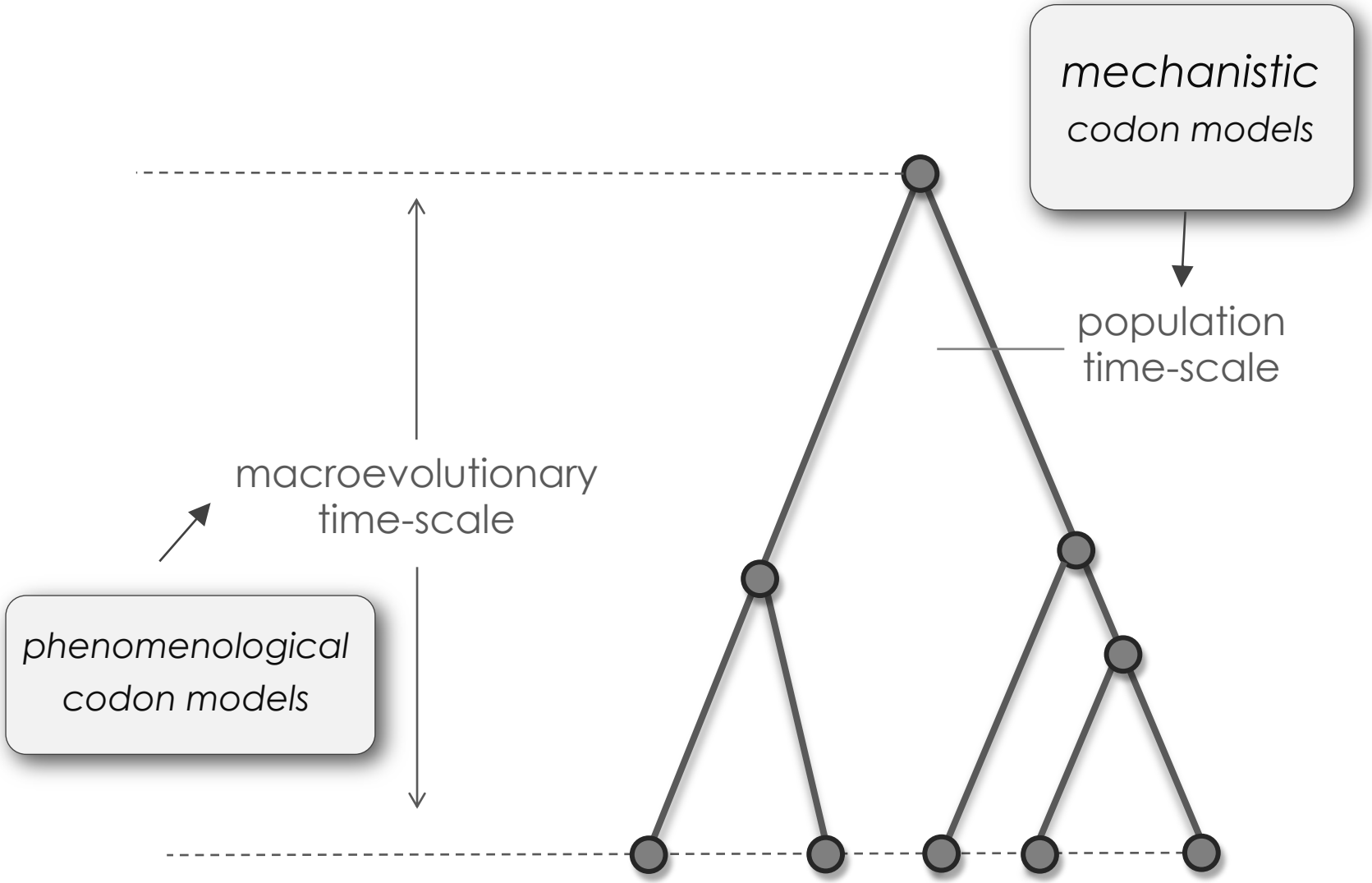
fitness coefficients $f^h = \langle f_1 , \ldots, f_{61} \rangle$

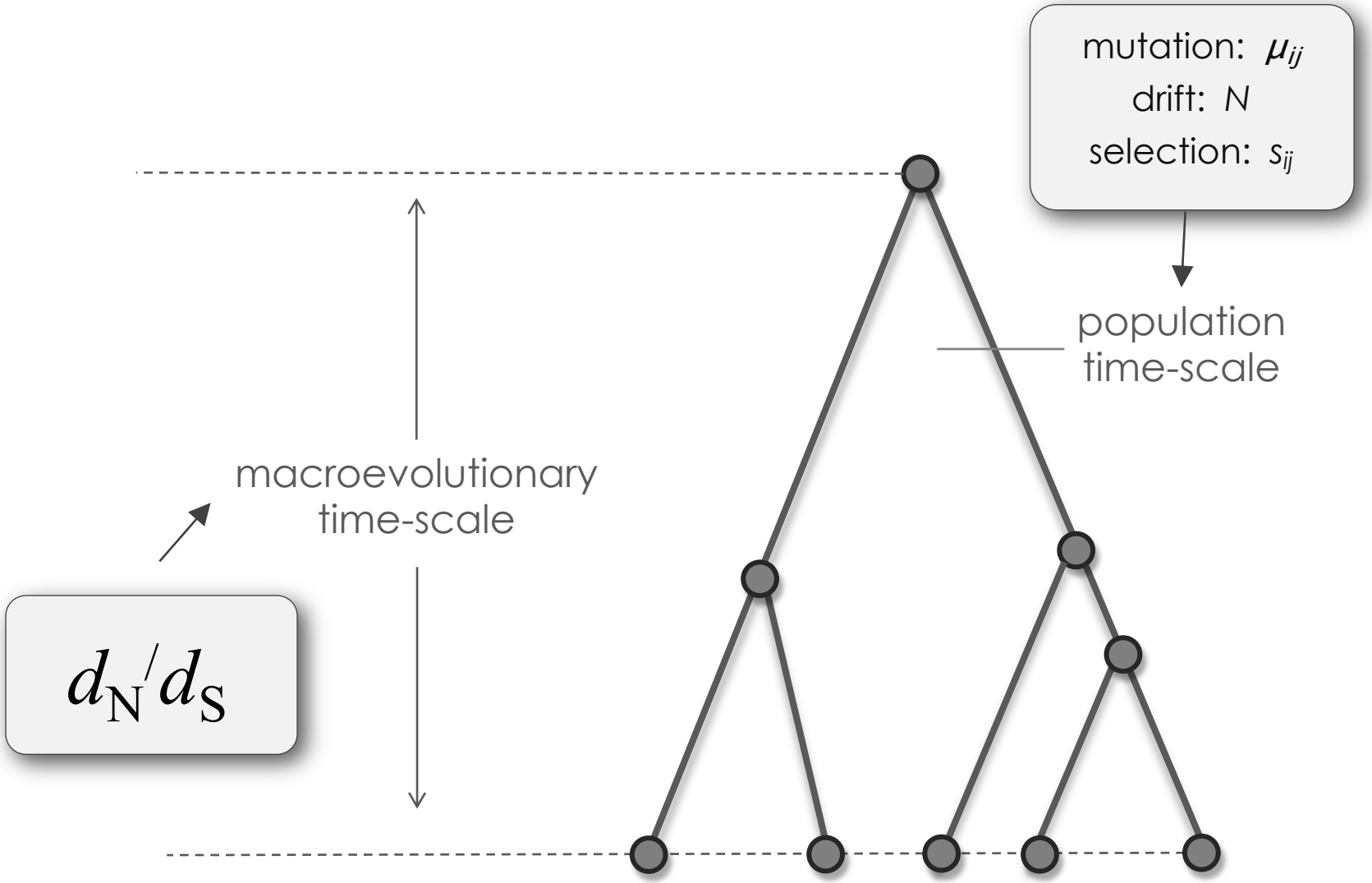
selection coefficients $s_{ij}^h = f_j^h - f_i^h$

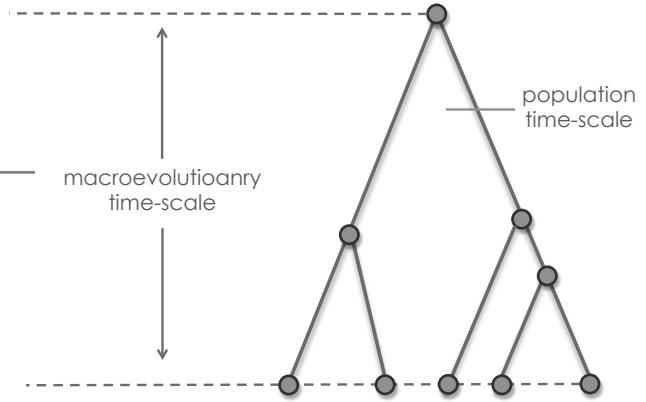
fixation probability (Kimura, 1962) $\Pr(s_{ij}^h) = \frac{2 s_{ij}^h}{1 - e^{-2Ns_{ij}^h}}$

2. phenomenological codon models




mechanistic
codon models
population
time-scale
macroevolutionary
time-scale
phenomenological
codon models


mutation: μ_{ij}
drift: N
selection: s_{ij}
population time-scale
macroevolutionary time-scale
d_N/d_S



phenomenological models

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

"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)

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

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

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

→ Phenomenological codon models: just a few parameters can cover the 3721 changes 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 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).

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

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

Phenomenological codon models: just a few parameters can cover the 3721 changes between codons!

context matters…

	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	—

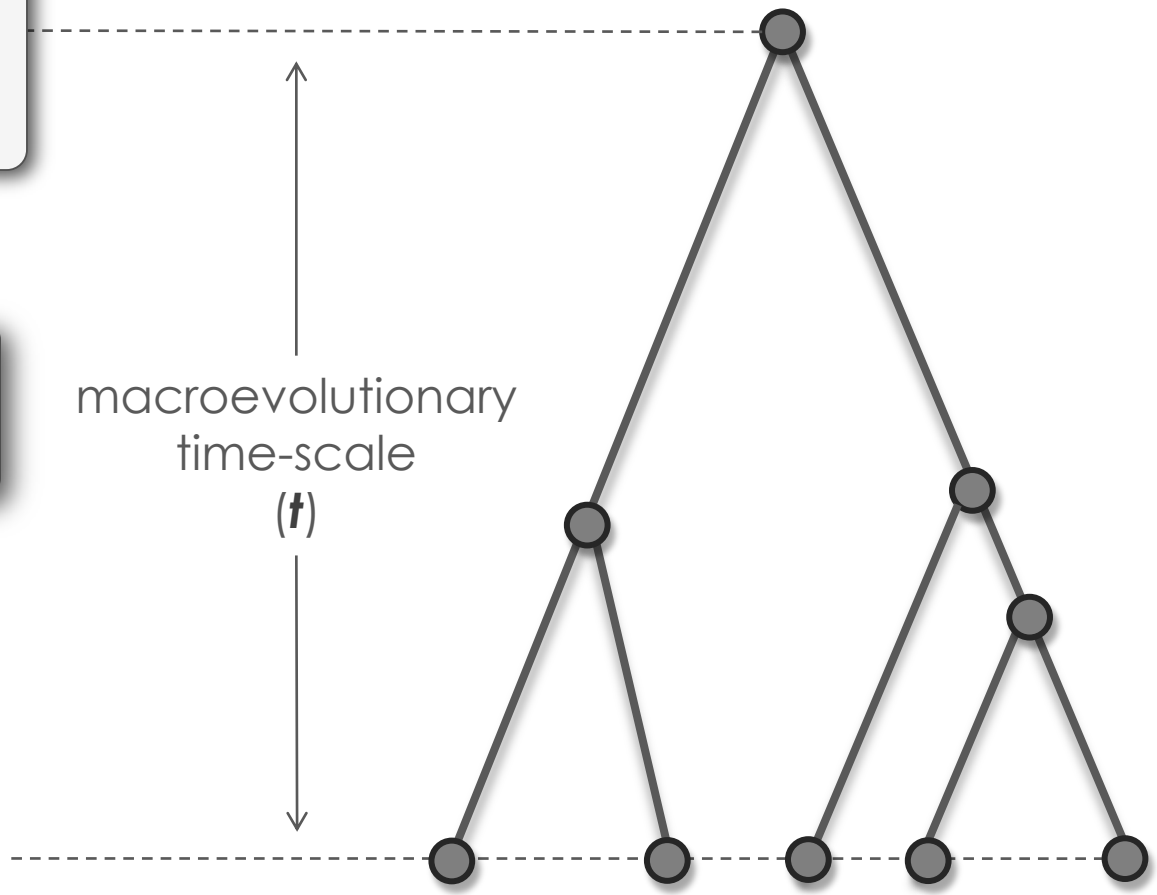
* This is equivalent to the codon model of Goldman and Yang (1994). Parameter ω is the ratio d_N/d_S, κ is the transition/transversion rate ratio, and π_i is the equilibrium frequency of the target codon (i).

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

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

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

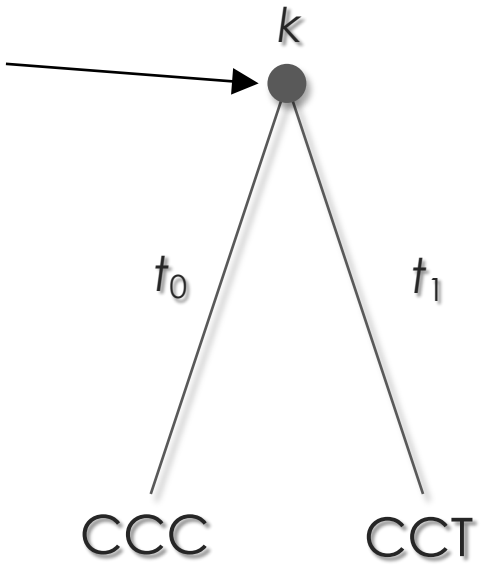
recall that **Paul Lewis** introduced **Q matrices** and how to obtain **transition probabilities**



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.

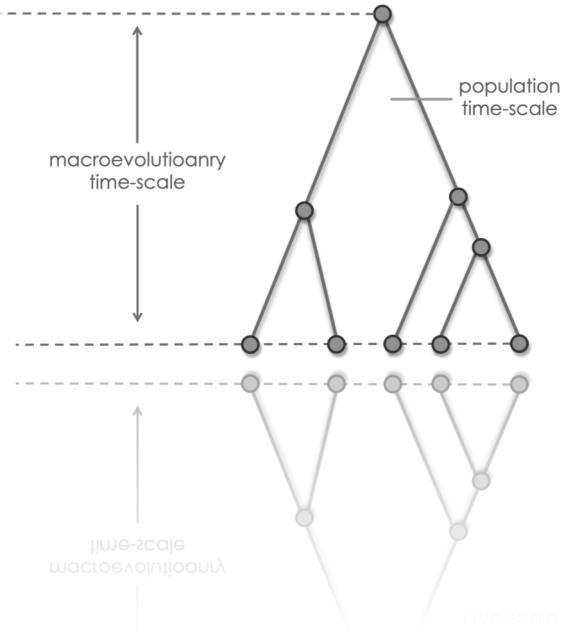
$$L = L_1 \times L_2 \times L_3 \times \ldots \times L_N = \prod_{h=1}^{N} L_h$$

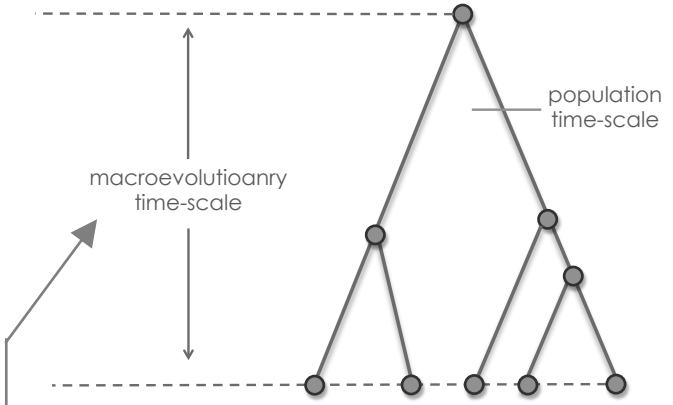
see **Paul Lewis's** lecture slides for more about likelihoods vs. log-likelihoods

The log likelihood is a sum over all sites.

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

3. bridging selection between time-scales





mechanistic models

"MutSel framework"

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

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

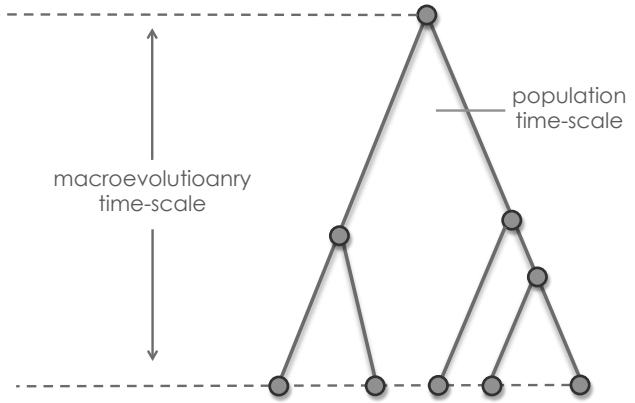
Halpern and Bruno (1998)
Jones et al. (2016)

phenomenological models

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

1. map fitness to equilibrium frequencies

2. expected index of selection intensity

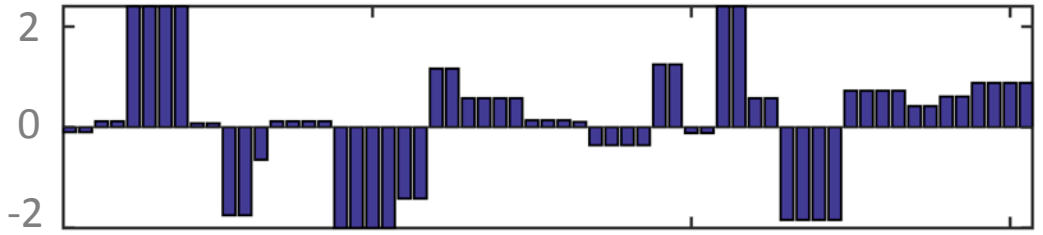


(1) Sella and Hirsh 2005; (2)Jones et al. 2016

1. fitness coefficients map to stationary codon frequencies

fitness coefficients

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



codon frequencies

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

(Sella and Hirsh 2005)

2. from fitness coefficients to espected *dN/dS*

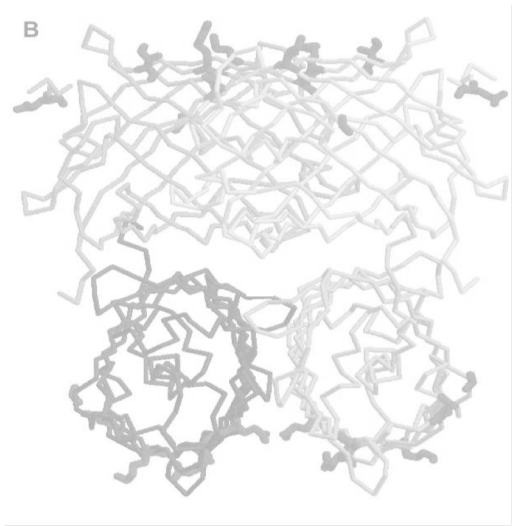
MutSel rate matrix

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

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

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

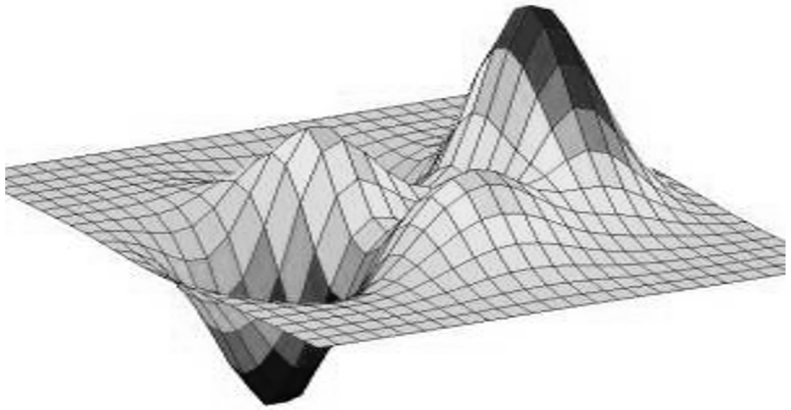
4. three positive selection scenarios



1932: adaptive landscapes and “shifting balance”

Sewall Wright

- introduces “ADAPTIVE LANDSCAPE” as a metaphor



- introduces “SHIFTING BALANCE” as a model

  (SBT **more complex** than I will present)

positive selection: 3 evolutionary scenarios

1. frequency dependent selection
2. episodic adaptation

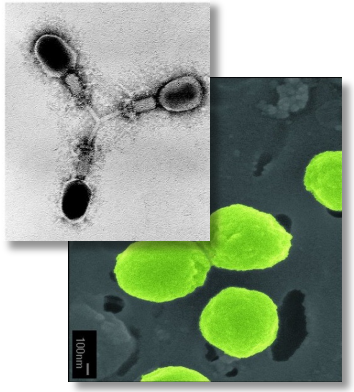
dynamic fitness landscape

3. non-adaptive shifting balance

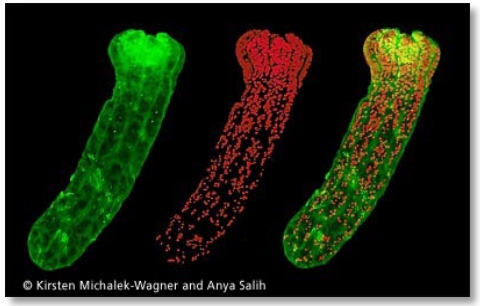
static fitness landscape

1. antagonistic evolutionary interaction

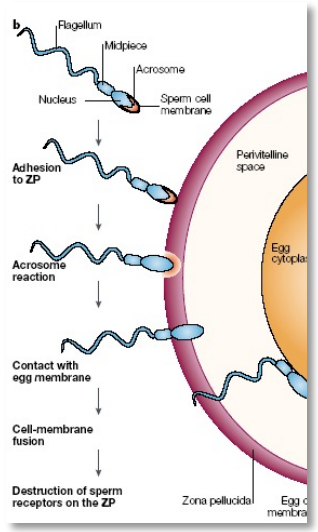
host-pathogen



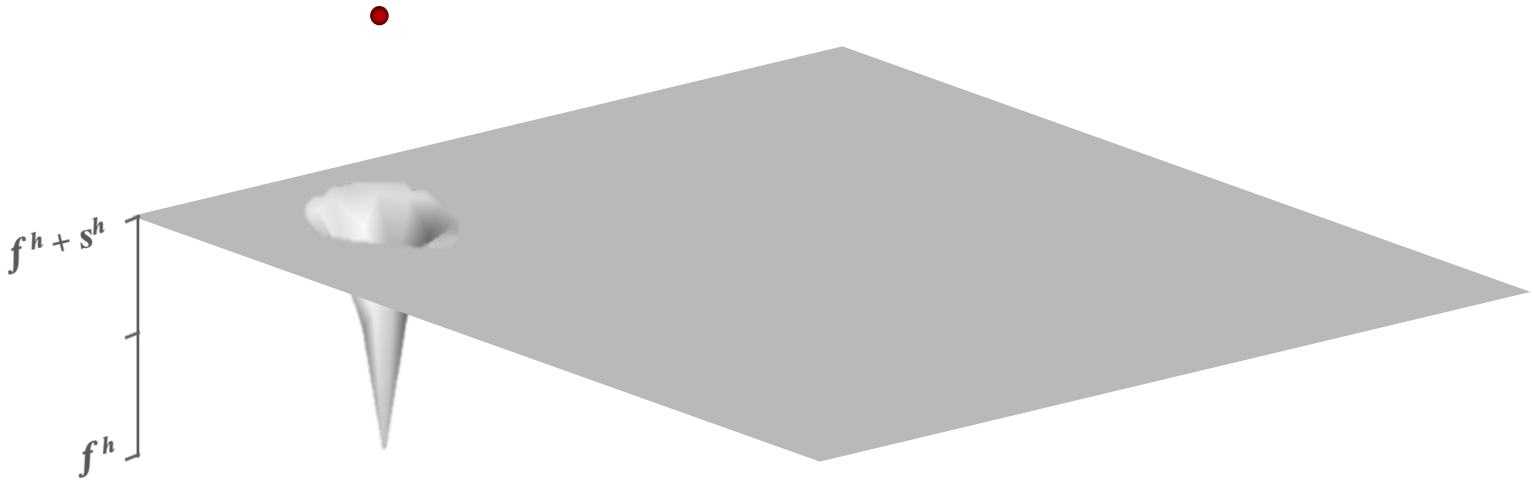
molecular-interactions



sexual-conflict

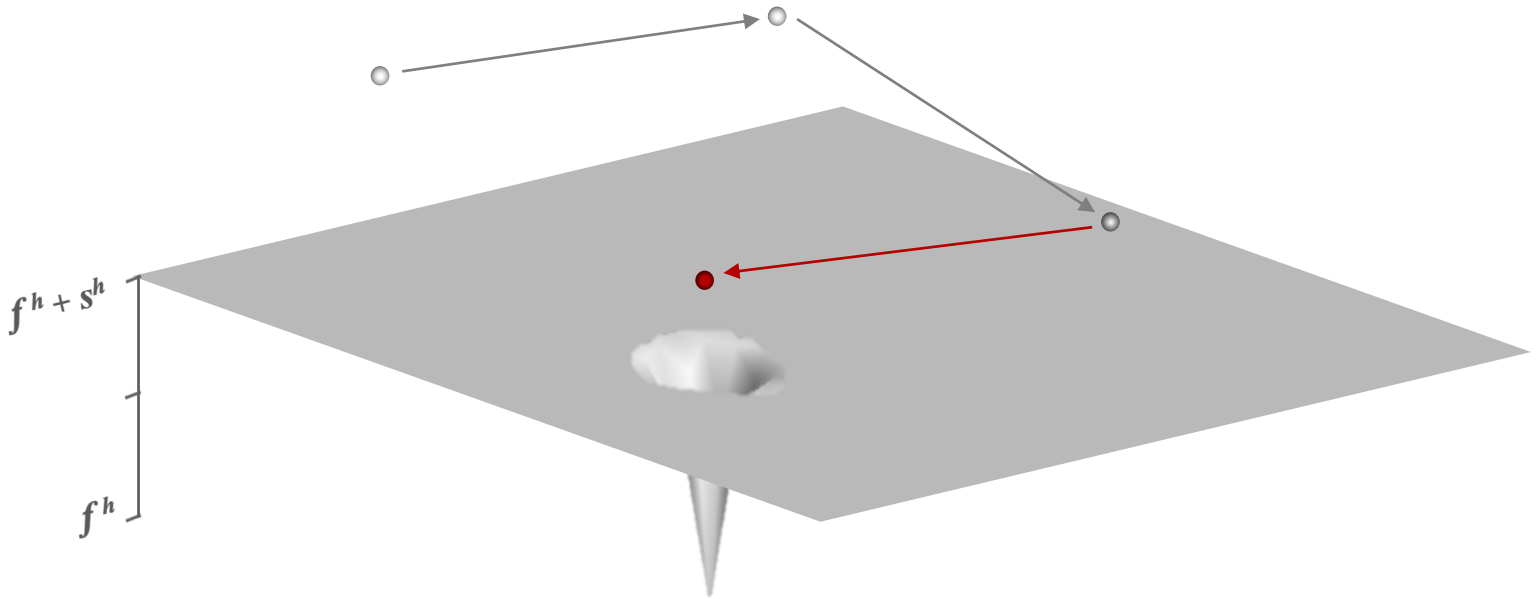


1 frequency-dependent adaptive landscape (weird)



1 frequency-dependent adaptive landscape (weird)

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



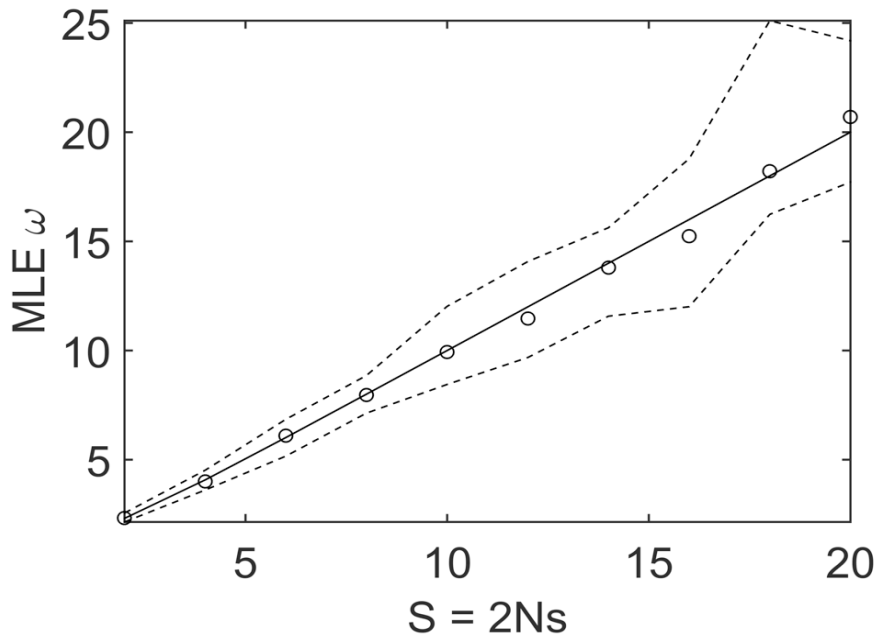
MutSelM0: (1) and (2) above imply Markov chain properties with the same rate matrix Q as **codon model M0**

"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)

1 frequency-dependent adaptive landscape (weird)



generating process:

MutSelM0

expectation = dN^h/dS^h

symbol = ——

fitted model:

model M0

inference = MLE ω

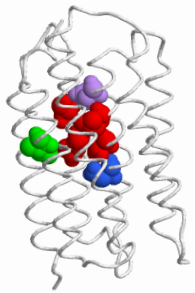
symbol = O

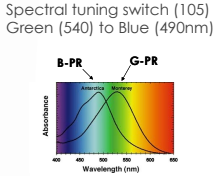
conclusion: phenomemologcial codon models assume frequency-dependent selection

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

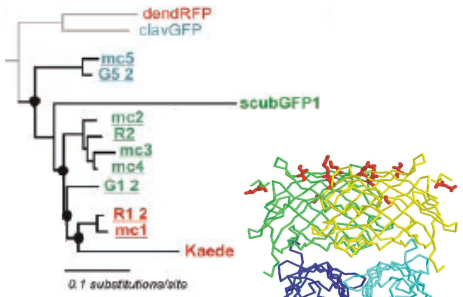
2. episodic Darwinian adaptation

exploitation of a new niche

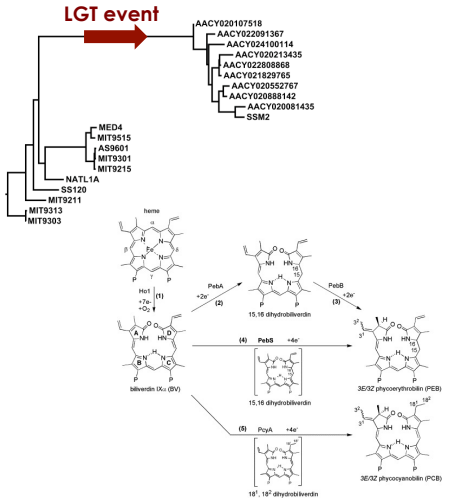




gene duplication

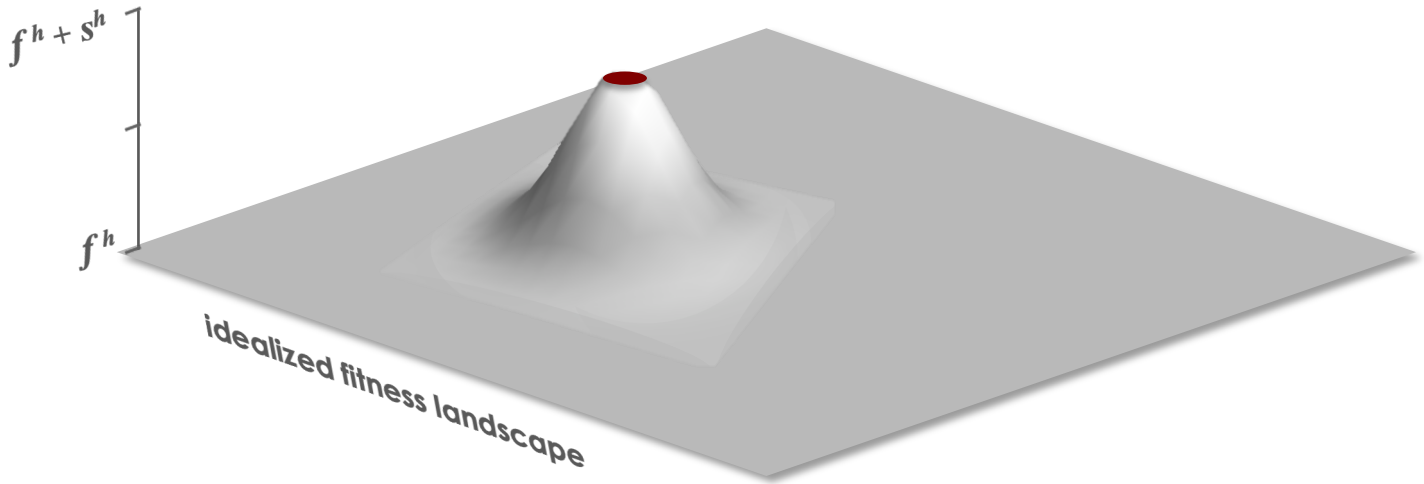


lateral gene transfer (LGT)



2 adaptive peak shift: evolution of novel function

optimal function in a stable environment



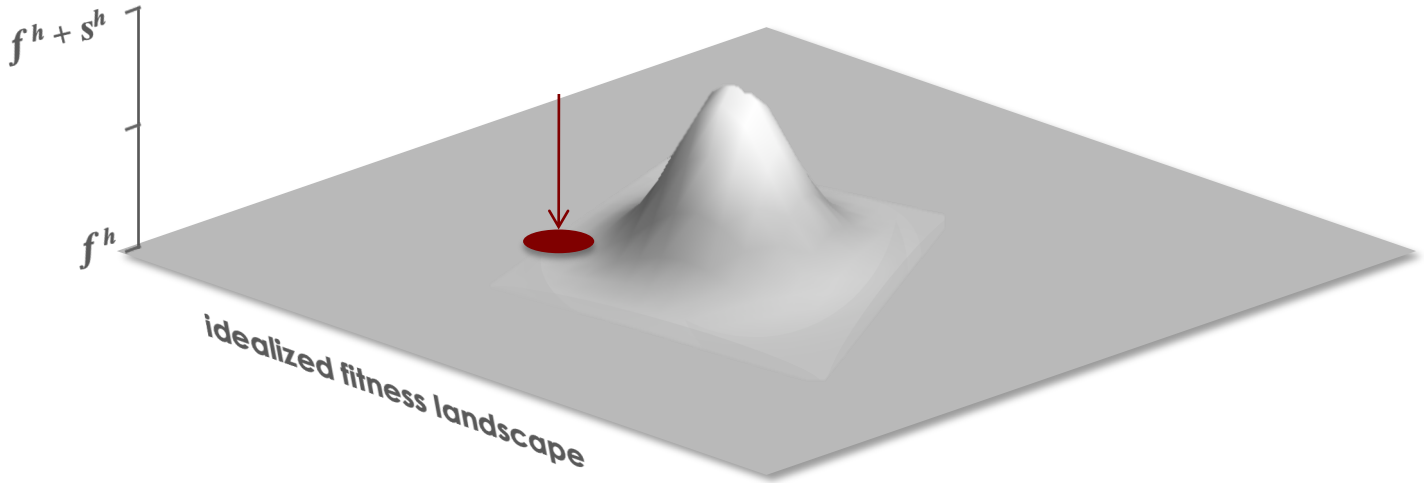
population: at fitness peak

fitness peak: stationary

FFTNS: keeps population at peak

2 adaptive peak shift: evolution of novel function

sub-optimal function in a novel environment



population: lower fitness

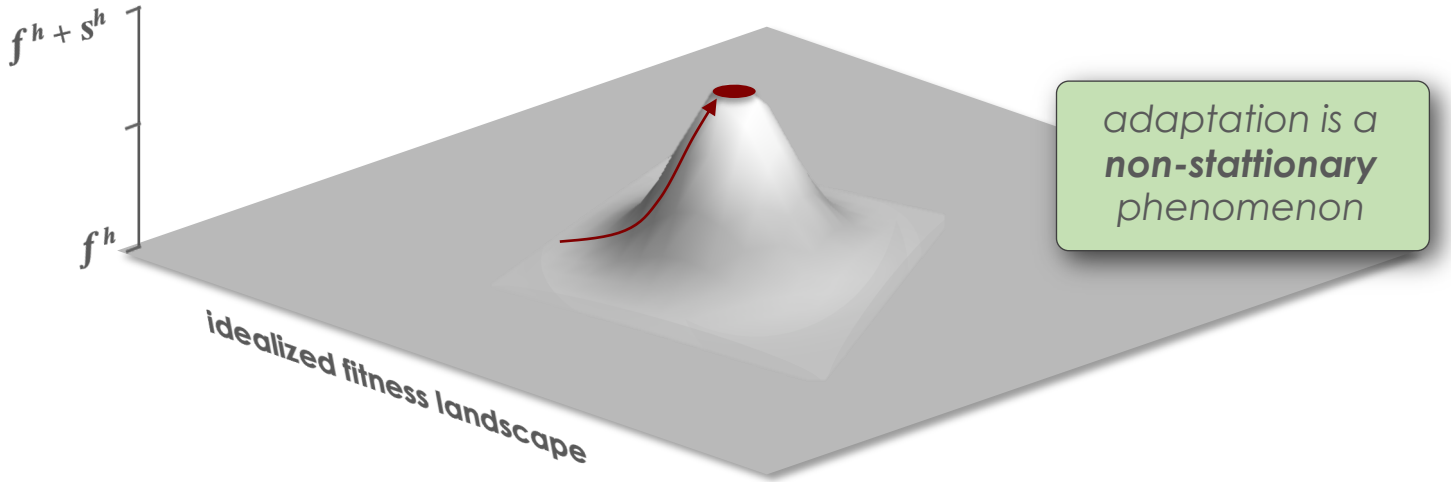
fitness peak: moving

FFTNS: increase population mean fitness

(non-stationary process)

2 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

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

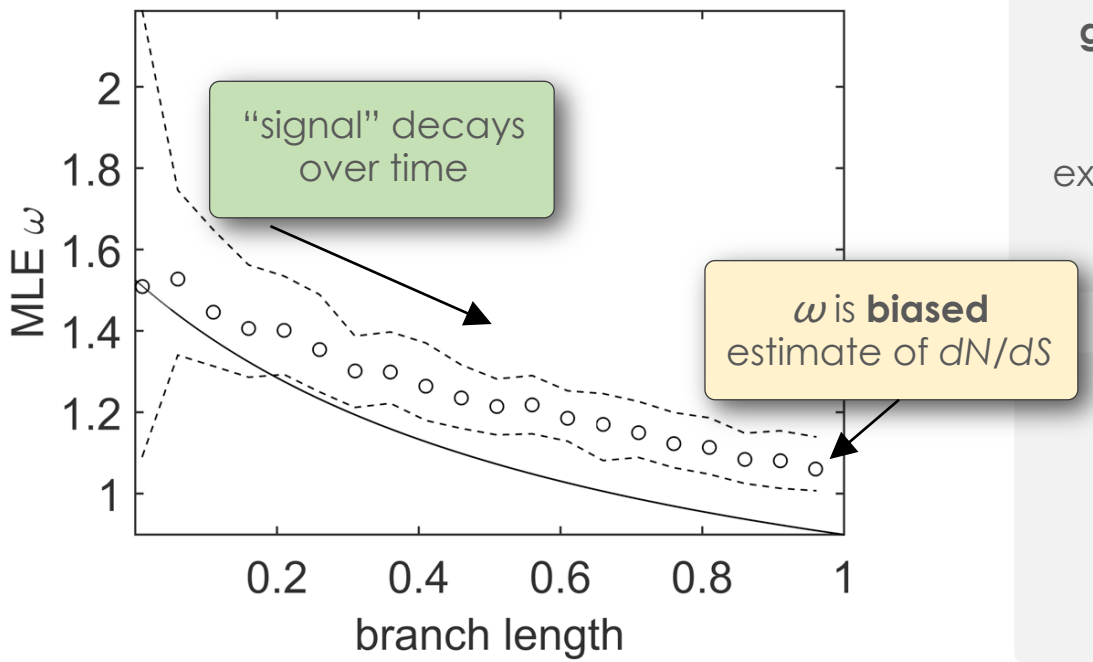
Mario dos Reis

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

2 adaptive peak shift: MutSel-ES



generating process:

MutSel-ES

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.

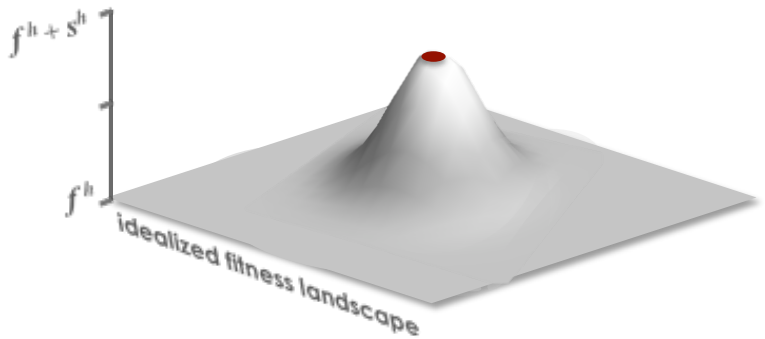
conclusion : episodic models "work" because they are sensitive to non-stationary behavior

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

3. fitness coefficients are constant (fixed-peak)

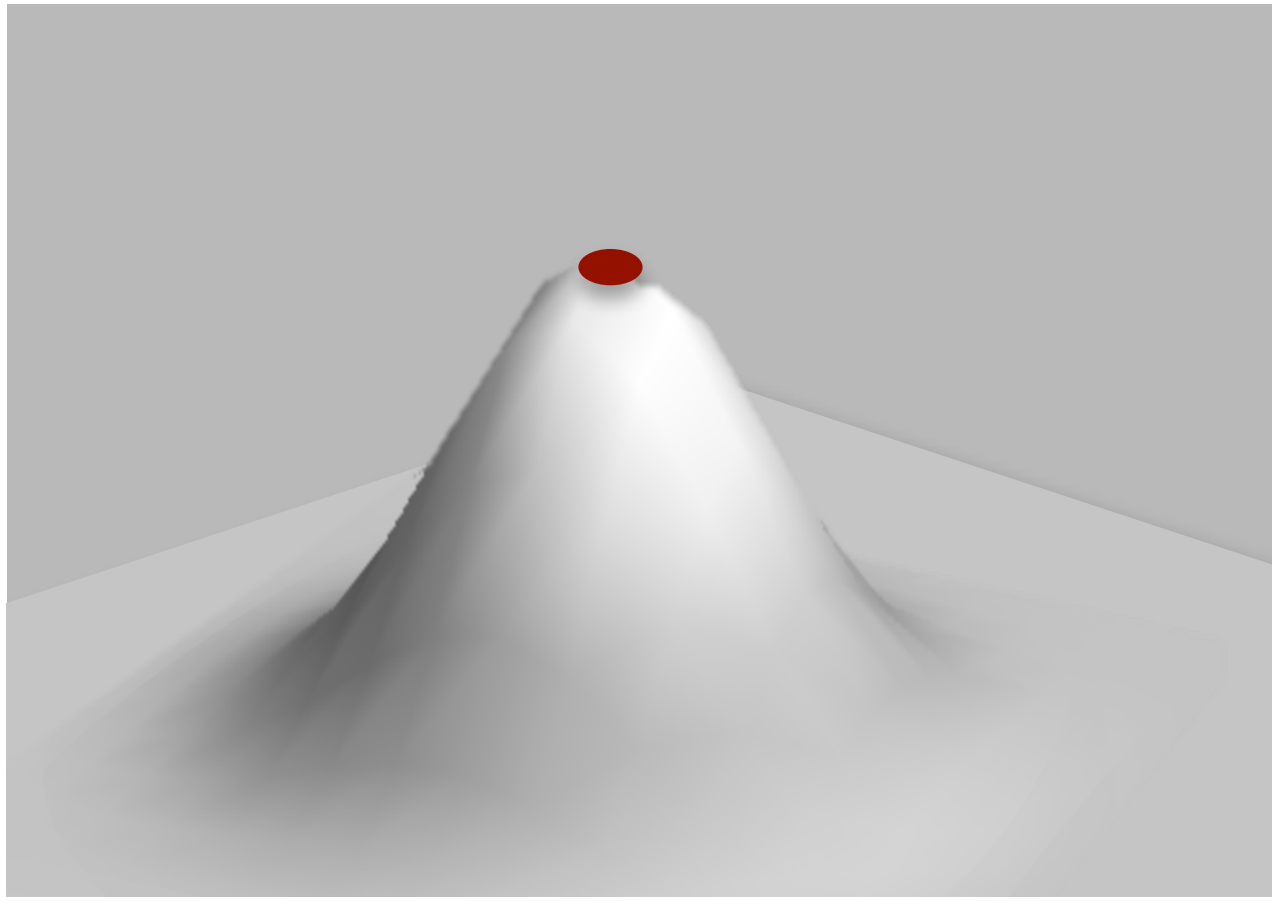
Spielman and Wilke (2015)

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



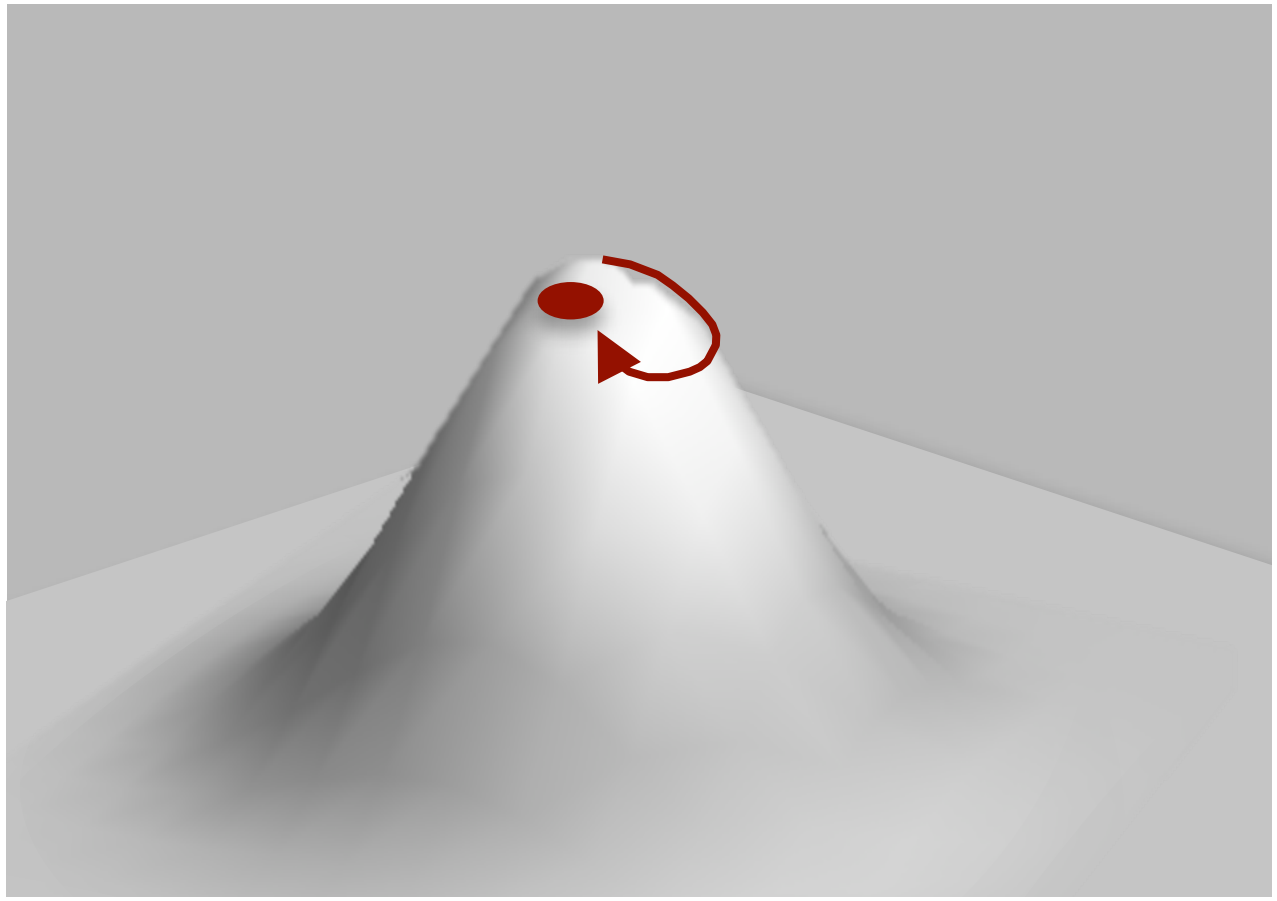
[Spielman and Wilke, (2015); Jones at al., (2016)]

3 shifting balance: movement around stationary peak (non-adaptive)



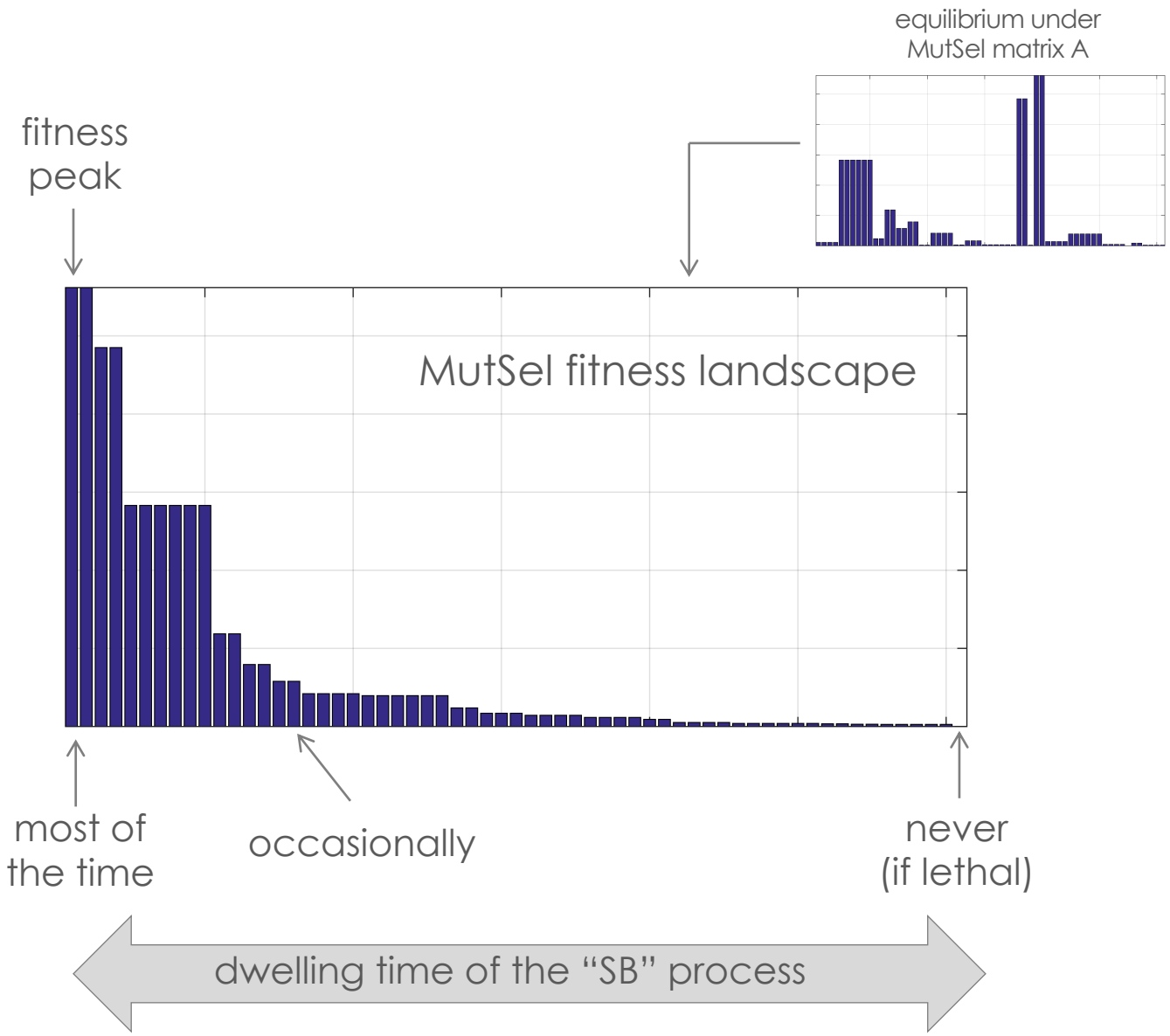
mutation and **drift** can move a pop. off a fitness peak

3 shifting balance: movement around stationary peak (non-adaptive)

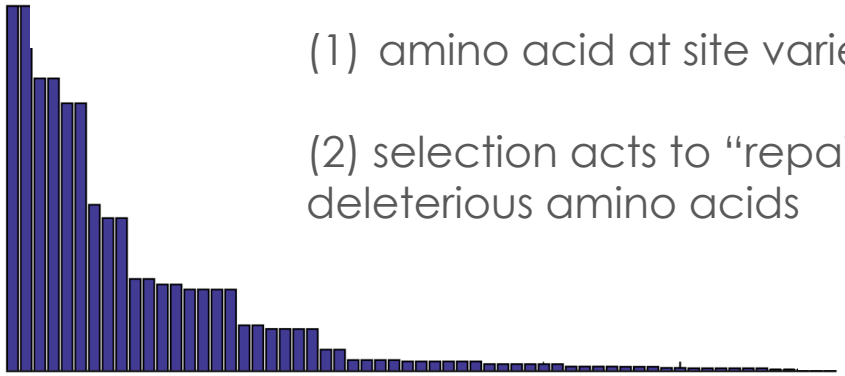


mutation and **drift** can move a pop. off a fitness peak

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



3 shifting balance: positive selection on a MutSel landscape



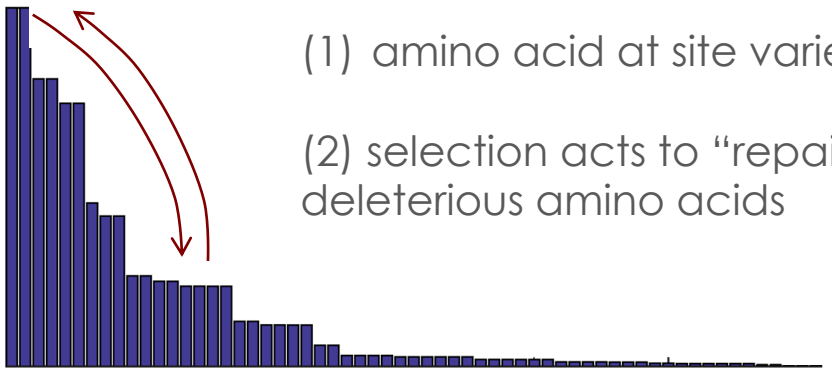
(1) amino acid at site varies over time

(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



(1) amino acid at site varies over time

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

key result:

purifying selection: $p_+ = p_-$

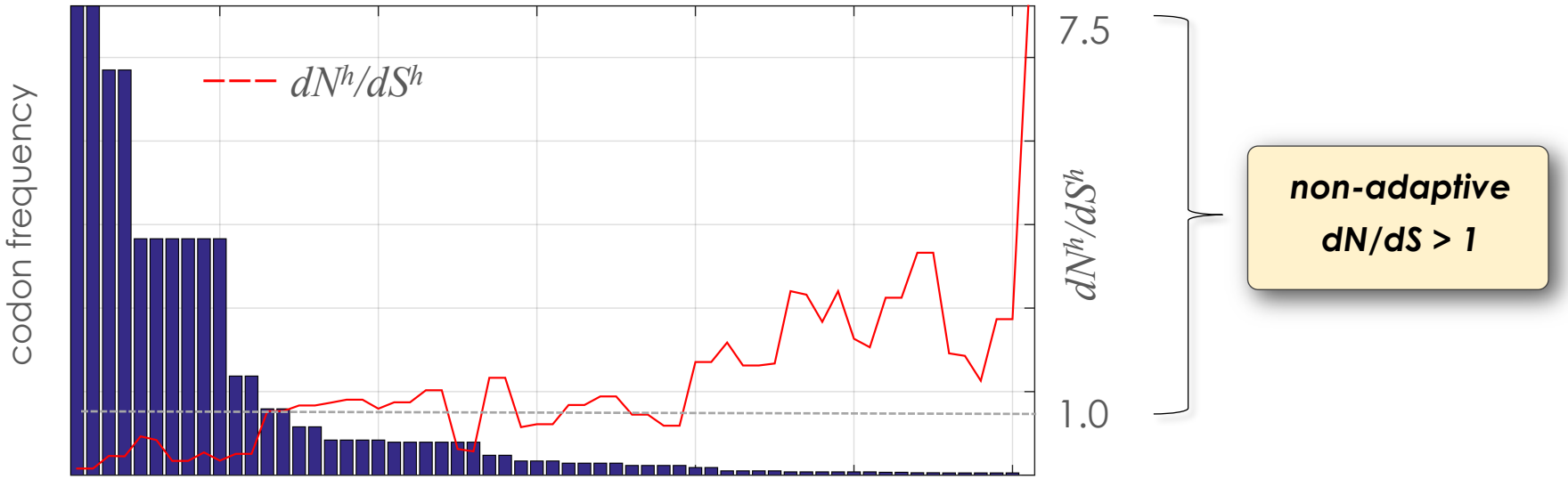
(static landscape)

p+ = positive selection without adaptation (***maintenance***!)

p- = related to "fixed drift load"

3 shifting balance: the MutSel landscape

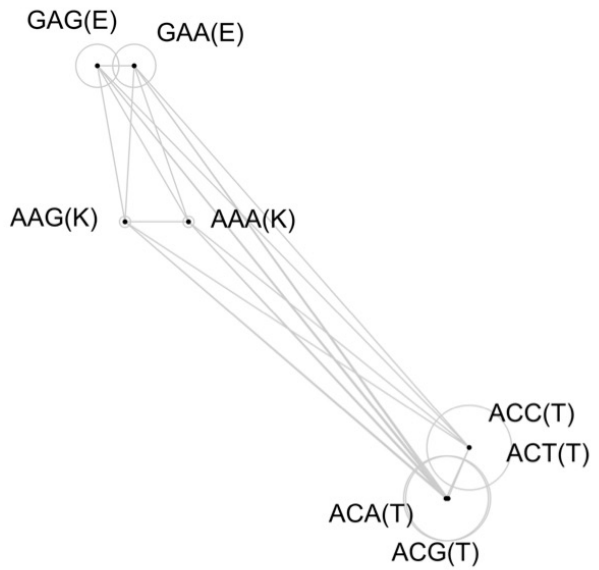
dN^h/dS^h depends on the current amino acid



temporal average dN^h/dS^h = 0.61

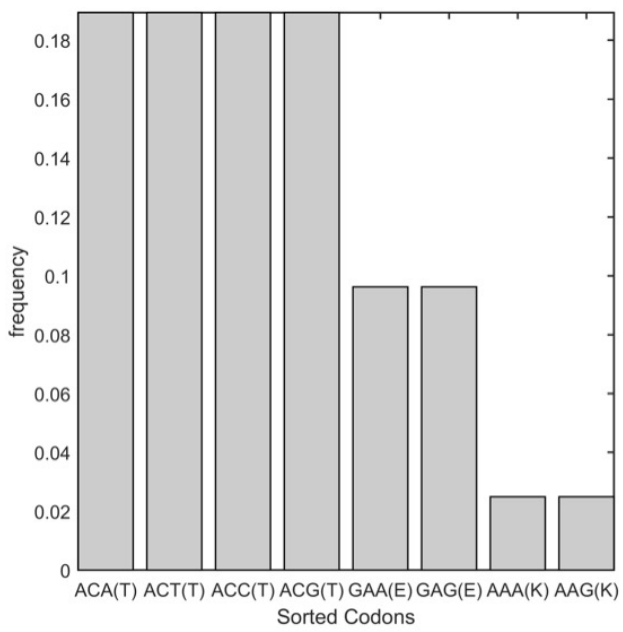
conclusion: positive selection operates on a stationary fitness peak in the same way as when there is an adaptive peak shift

5. nearly-neutral theory and “heterotachy”

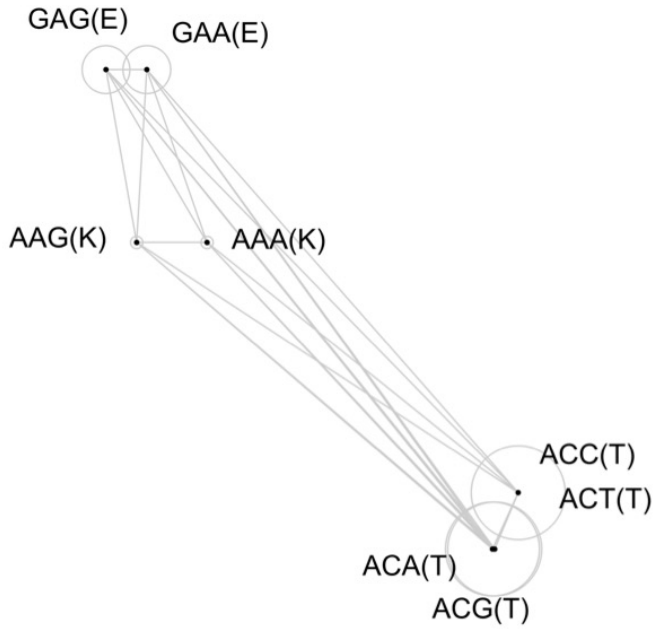


landscapes have unique structures

MutSel landscape



McCandlish landscape



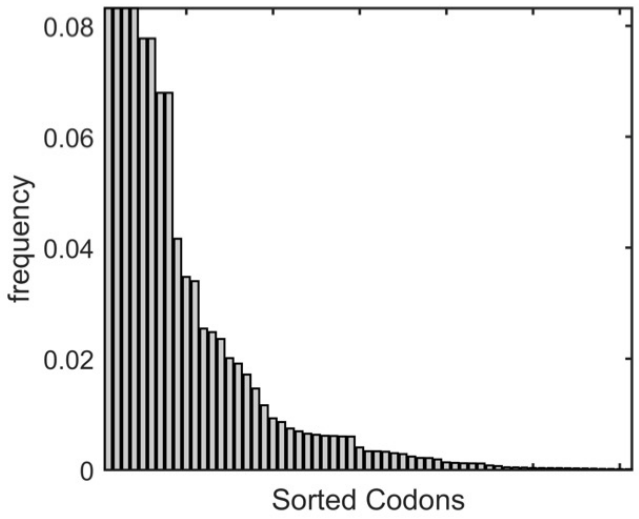
conclusion: A population can get to a sub-optimal codon (E) by drift and reside there for some time (b/c moving between T and E requires changes ≥ 2 codons).

landscapes have unique structures

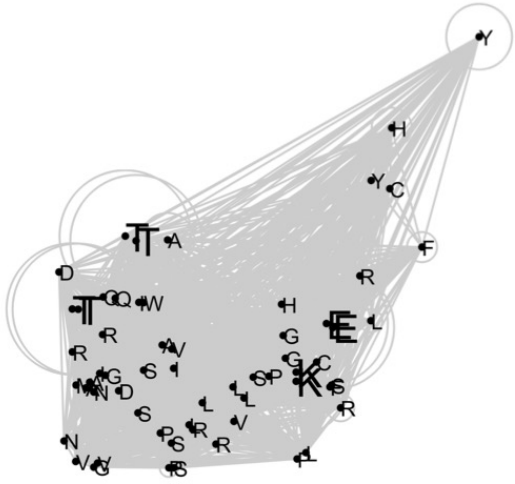
same site... 10x decrease in *N* (f^h have not changed!)

Nearly-neutral theory:
selection and drift interact!

MutSel landscape



McCandlish landscape



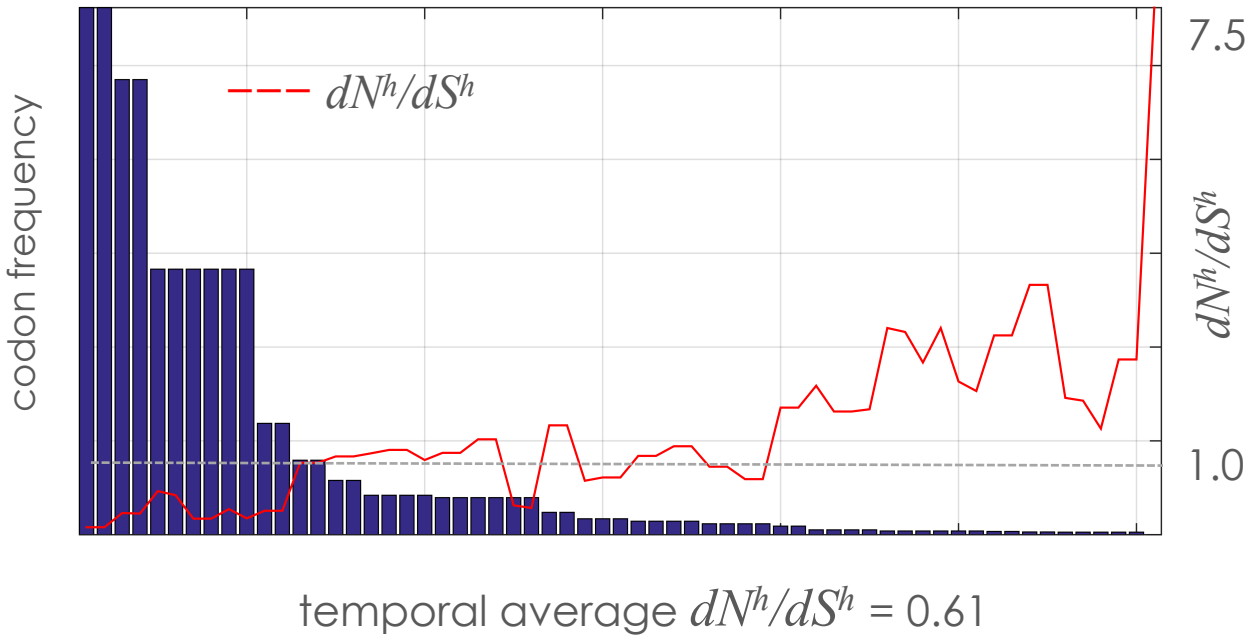
Rate of evolution depends on **population size**

conclusion: decreasing N changes:

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

shifting balance: the MutSel landscape

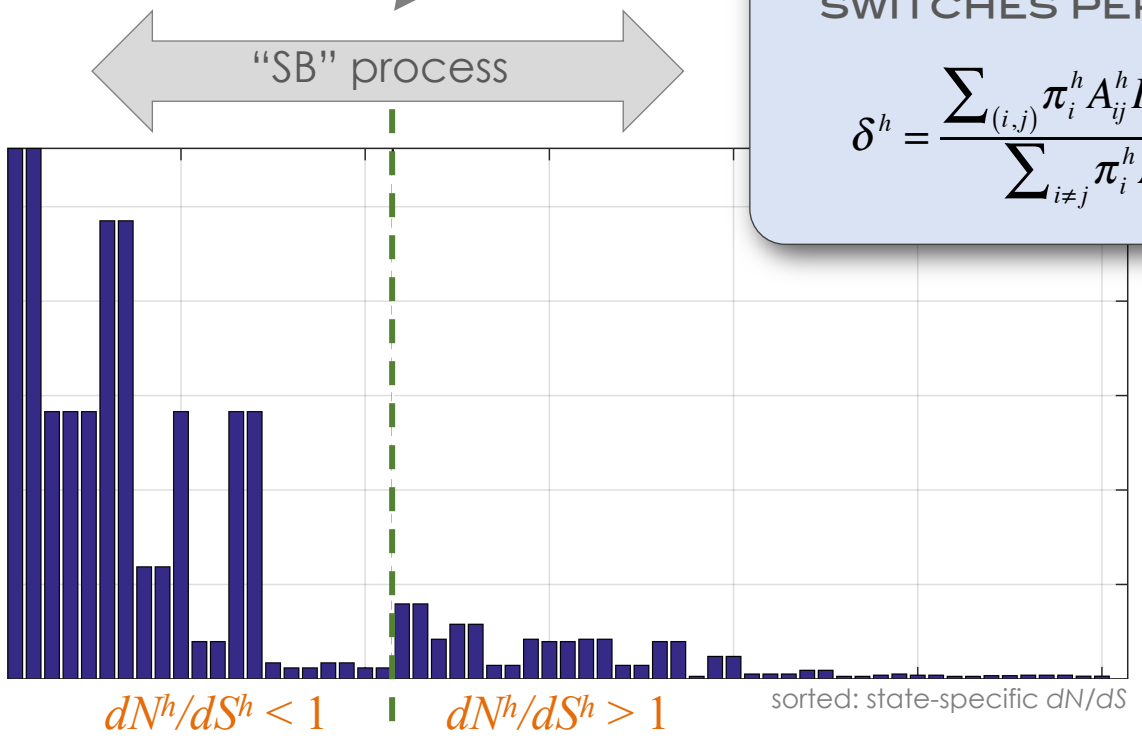
dN^h/dS^h depends on the current amino acid



temporal average dN^h/dS^h = 0.61

shifting balance: a mechanistic model

NOTE: Rate switching like this is called "**Heterotachy**"



EXPECTED NO. OF SWITCHES PER SUB.

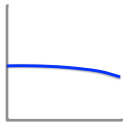
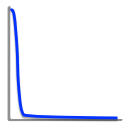
$$\delta^h = \frac{\sum_{(i,j)} \pi_i^h A_{ij}^h I_{SWITCH}}{\sum_{i \neq j} \pi_i^h A_{ij}^h}$$

$$\omega^h < 1 = \frac{\sum_{i \in I_p^h} \frac{\pi_i^h}{p_1} A_{ij}^h I_N}{\sum_{i \in I_p^h} \frac{\pi_i^h}{p_1} \mu_{ij} I_N}$$

$$\omega^h > 1 = \frac{\sum_{i \in I_t^h} \frac{\pi_i^h}{p_2} A_{ij}^h I_N}{\sum_{i \in I_t^h} \frac{\pi_i^h}{p_2} \mu_{ij} I_N}$$

What does heterotachy "look" like on a tree?

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^h

σ: {0.0001, 0.001, 0.01}

N = 1000

We can model heterotachy with a covarion-like model

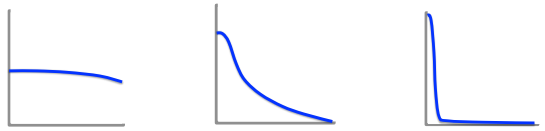
$$Q = \left[\begin{array}{c|c} \textbf{evolutionary regime 1}: \omega_1 = \text{low} \quad (\text{“near the peak”}) & \textbf{switching process:} \quad \omega_1 \rightarrow \omega_2 \\ \hline \textbf{switching process:} \quad \omega_1 \leftarrow \omega_2 & \textbf{evolutionary regime 2}: \omega_2 = \text{high}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

[Guindon et al., (2004); Jones et al. (2016); Jones et al. (2018); Jones et al. 2019]

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

shifting balance: a mechanistic model

shifting balance over landscape



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 codon models!

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

6. some common types of codon models

"OMEGA MODELS"

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

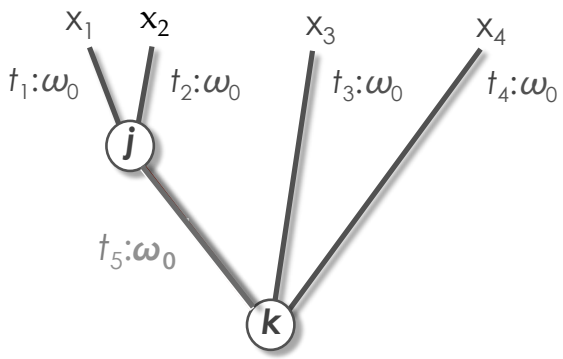
Goldman and Yang (1994)
Muse and Gaut (1994)

this codon model "**M0**"

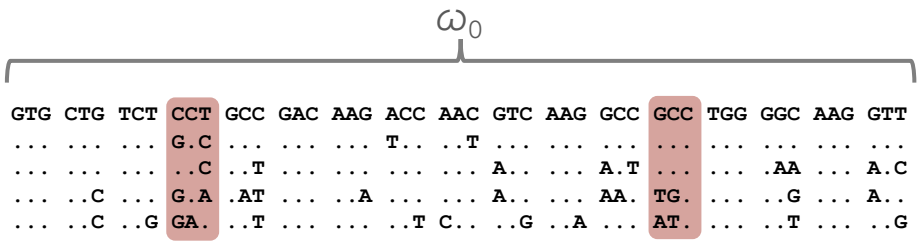
"OMEGA MODELS"

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

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

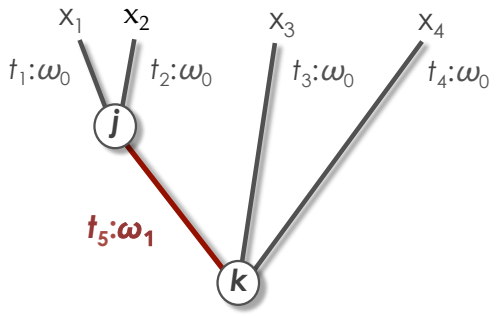


same ω
for all branches

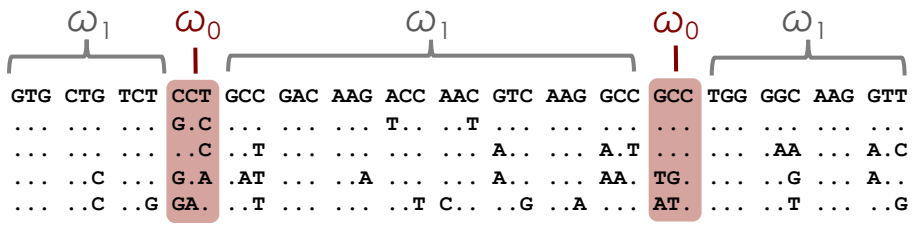


same ω
for all sites

two basic types of models…

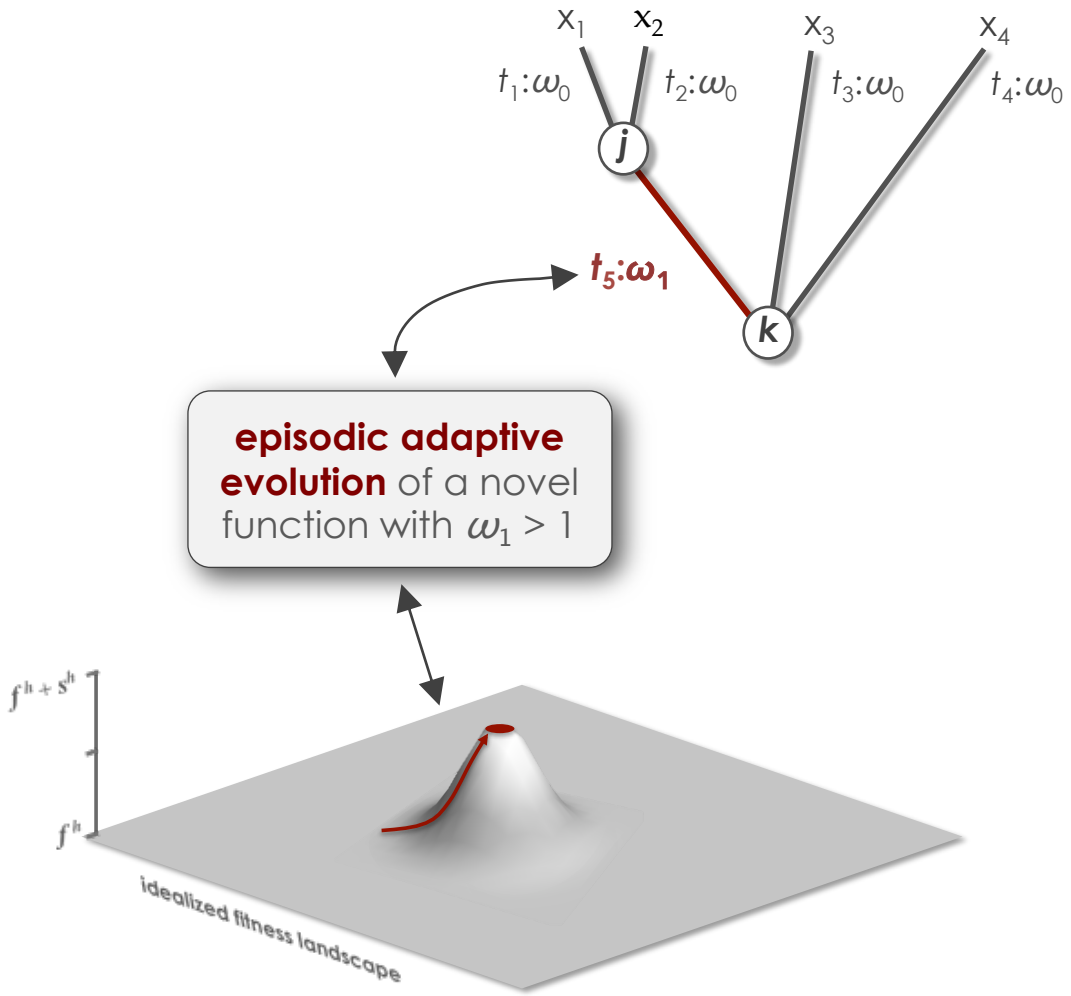


branch models
(*ω* varies among branches)

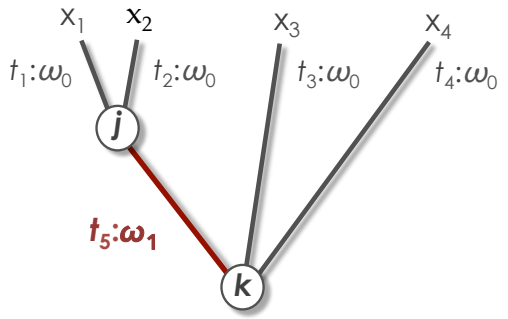


site models
(*ω* varies among sites)

interpretation of a branch model



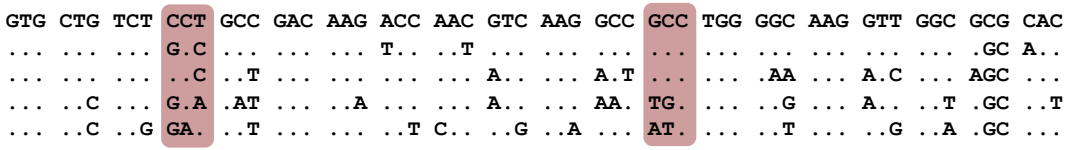
branch models*



variation (ω) among branches:	approach
Yang, 1998	fixed effects
Bielawski and Yang, 2003	fixed effects
Seo et al. 2004	auto-correlated rates
Kosakovsky Pond and Frost, 2005	genetic algorithm
Dutheil et al. 2012	clustering algorithm

* *these methods can be useful when selection pressure is strongly* ***episodic and functional change is substantial***

site models*



variation (ω) among sites:	approach
Yang and Swanson, 2002	fixed effects (ML)
Bao, Gu and Bielawski, 2006	fixed effects (ML)
Massingham and Goldman, 2005	site wise (LRT)
Kosakovsky Pond and Frost, 2005	site wise (LRT)
Nielsen and Yang, 1998	mixture model (ML)
Kosakovsky Pond, Frost and Muse, 2005	mixture model (ML)
Huelsenbeck and Dyer, 2004; Huelsenbeck et al. 2006	mixture (Bayesian)
Rubenstein et al. 2011	mixture model (ML)
Bao, Gu, Dunn and Bielawski 2008 & 2011	mixture (LiBaC/MBC)
Murell et al. 2013	mixture (Bayesian)

This is **NOT** a comprehensive list!

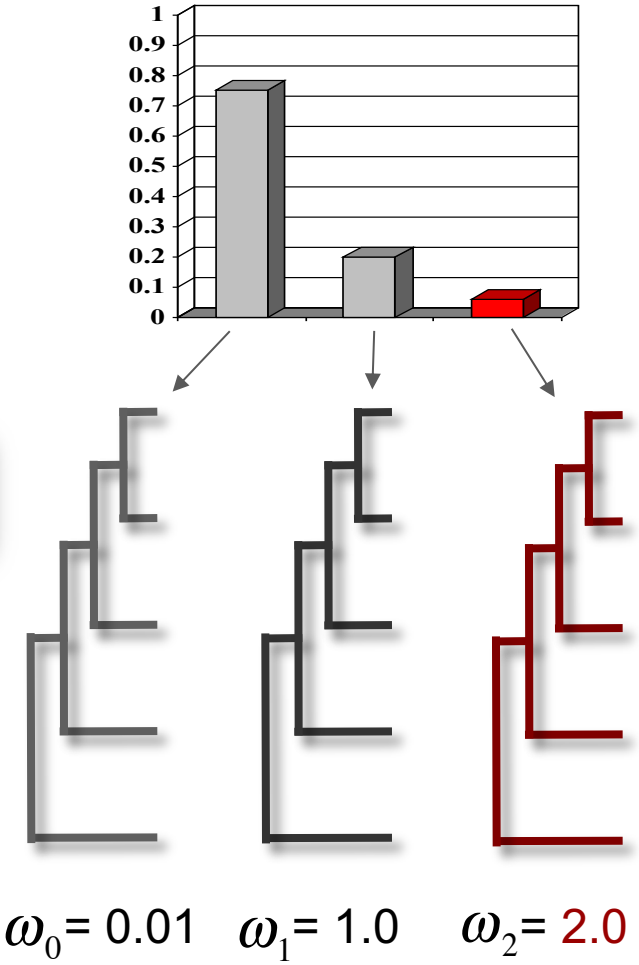
- *useful when at some sites evolve under **diversifying selection** pressure over long periods of time*

site models: discrete mixture model (**M3**)

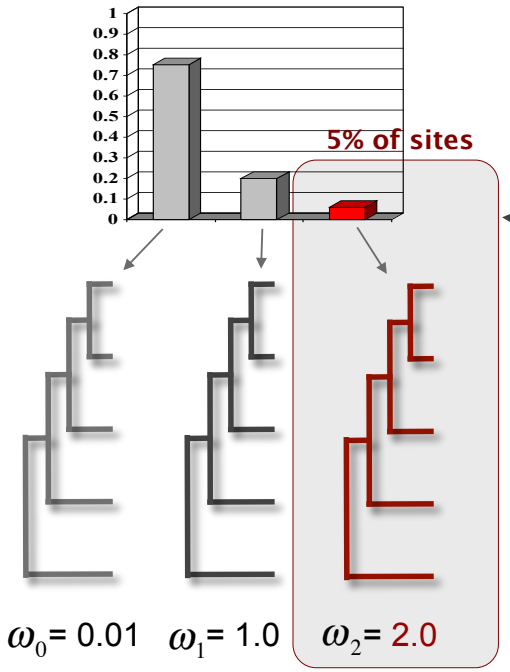
MIXTURE-MODEL LIKELIHOOD

$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$

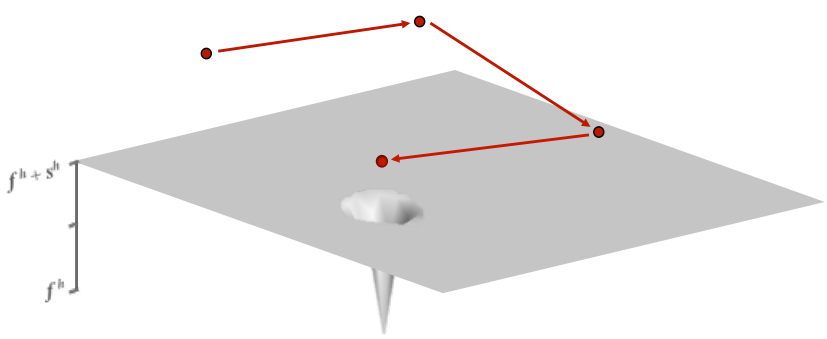
conditional likelihood calculation (see part 1)



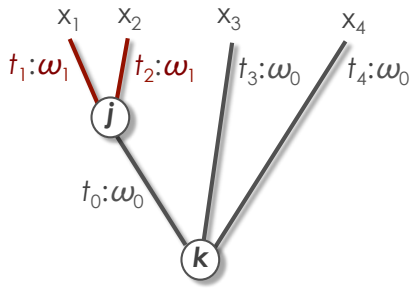
interpretation of a sites-model

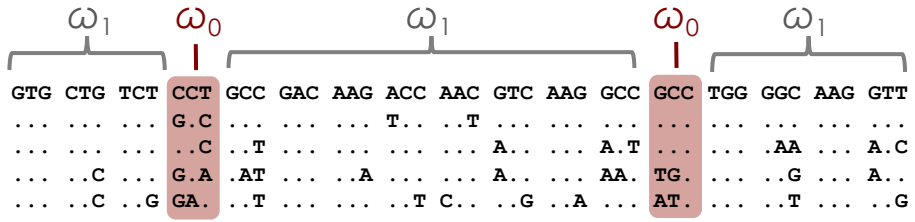


diversifying selection (frequency dependent)
at 5% of sites with
$\omega_2 = 2$



models for variation among branches & sites





branch models
(ω varies among branches)

site models
(ω varies among sites)

branch-site models
(combines the features of above models)

models for variation among branches & sites

This is **NOT** a comprehensive list!

variation (ω) among branches & sites:	**approach**
Yang and Nielsen, 2002	fixed+mixture (ML)
Forsberg and Christiansen, 2003	fixed+mixture (ML)
Bielawski and Yang, 2004	fixed+mixture (ML)
Giundon et al., 2004	covarion-like (ML)
Zhang et al. 2005	fixed+mixture (ML)
Kosakovsky Pond et al. 2011, 2012	full mixture (ML)
Jones et al., 2016, 2018, 2020	mix-covarion-like (ML)

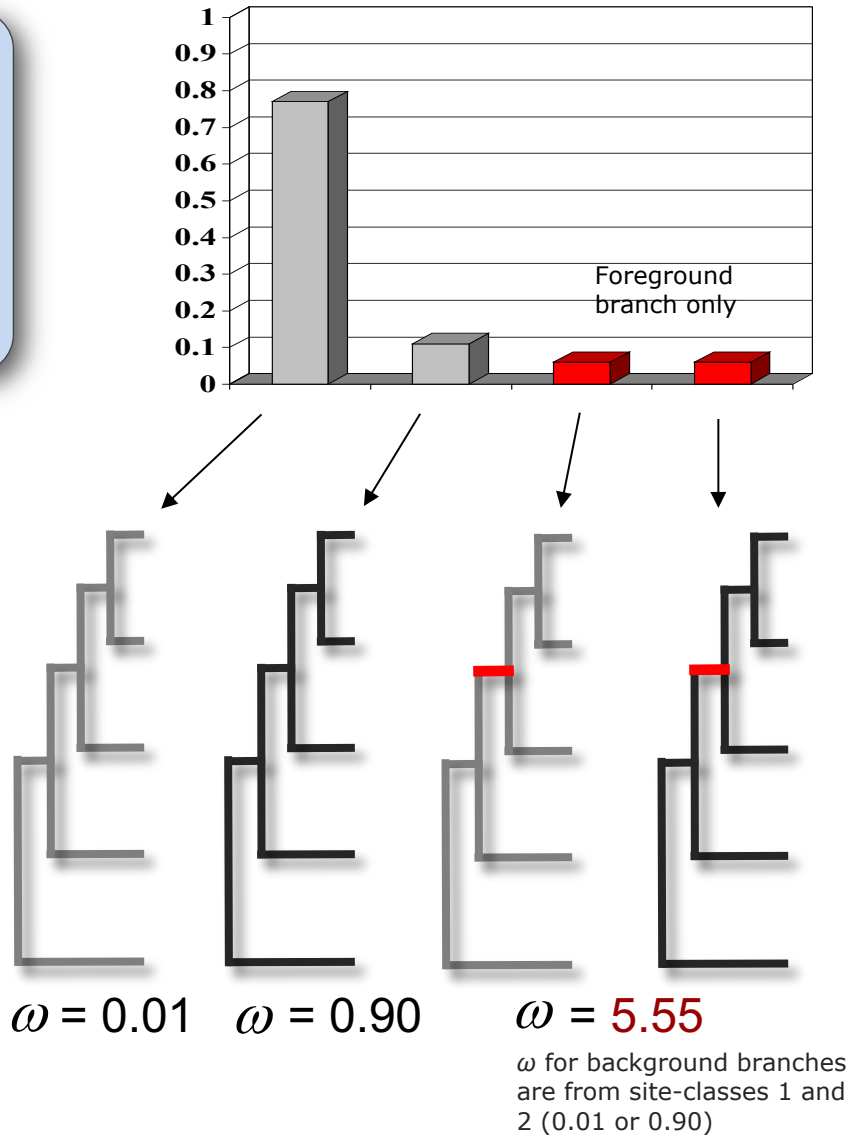
* *these methods can be useful when selection* ***pressures change over time at just a fraction of sites***

* *it can be a challenge to apply these methods properly*

branch-site "Model B"

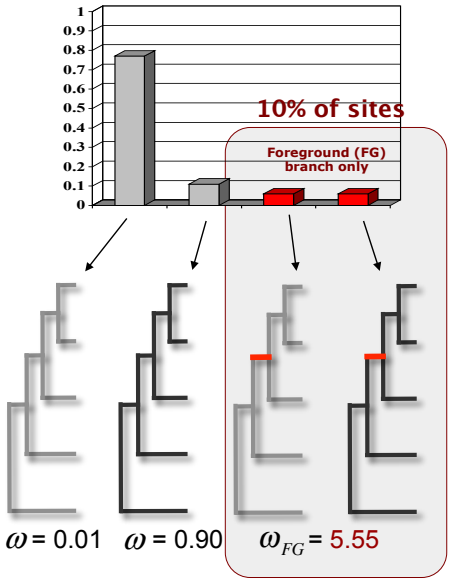
MIXTURE-MODEL LIKELIHOOD

$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$

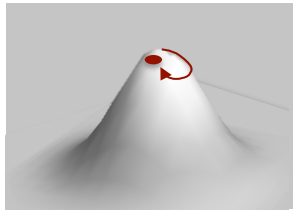


ω for background branches are from site-classes 1 and 2 (0.01 or 0.90)

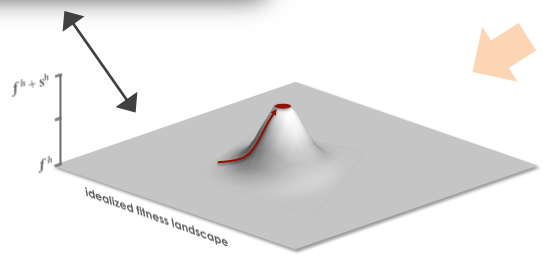
two scenarios can yield branch-sites with $d_N/d_S > 1$



10% of sites have **shifting balance** on a fixed peak (**same function**)



episodic adaptive evolution at **10% of sites** for **novel function**

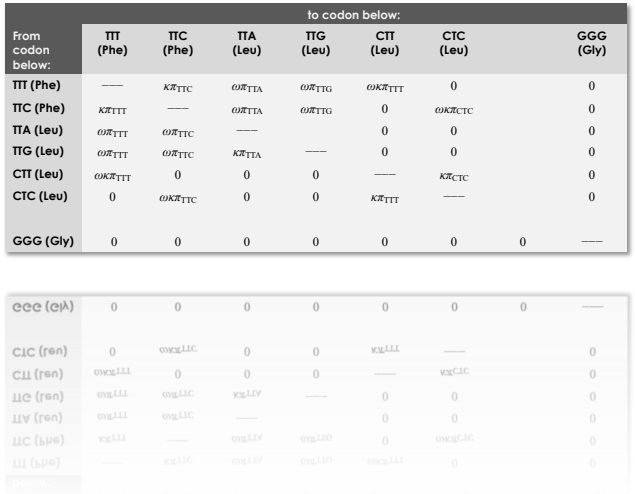


branch-site codon models cannot tell which scenario is correct without external information!

Jones et al (2016) MBE
Jones et al (2018) MBE

7. "bells –n– whistles"…

codon models + "other processes"



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	—

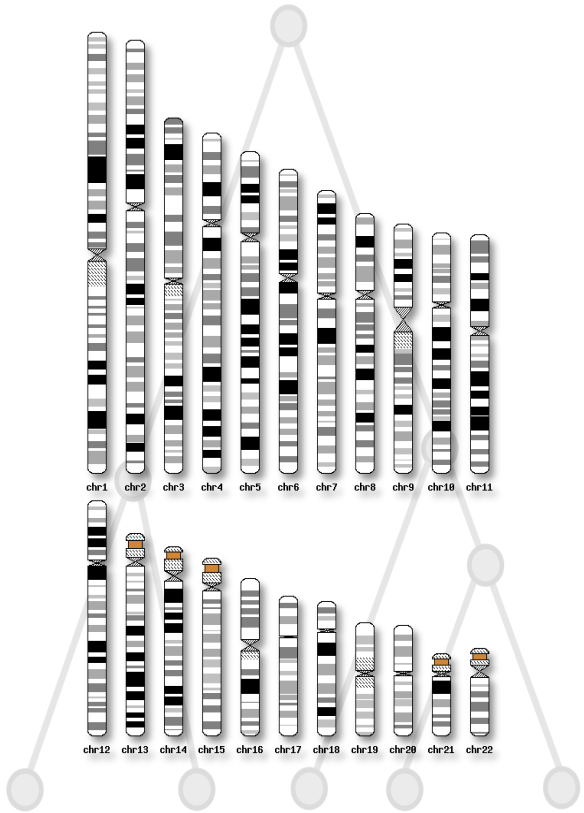
"*bells –n– whistles*"… some general categories

1. alternative models of codon frequencies (*the $\boldsymbol{\pi}$'s parameters are important*)
2. GTR process at DNA-level (*this is NOT a mutational process*)
3. among-site synonymous rate (d_S) variation (*phenomenologically important*)
4. double & triple nucleotide changes (*confounded with heterotachy*)
5. amino acid exchangeabilities (*confounded with codon frequencies via fitness*)
6. multi-process variation among sites (*do we really want this much complexity?)*
7. multi-pattern (tree) variation among sites (*this can be important*)

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	——

Is there some way that we can become <u>less-dependent</u> on all this phenomenological complexity which can obscure evolutionary implications?

8. Phenotype-Genotype codon models



Phenotype-Genotype codon models?

another justification…

Trends in Ecology & Evolution

CellPress
REVIEWS

Opinion

Phylogenetics is the New Genetics (for Most of Biodiversity)

Stacey D. Smith,[1,6,*,@] Matthew W. Pennell,[2] Casey W. Dunn,[3] and Scott V. Edwards[4,5]

Despite substantial progress in understanding the genetic basis for differences in morphology, physiology, and behavior, many phenotypes of interest are difficult to study with traditional genetic approaches because their origin traces to deep nodes in the tree of life. Moreover, many species are not amenable to either large-scale sampling or laboratory crosses. We argue that phylogenetic methods and theory provide tremendous power to identify the functional genetic variation underlying trait evolution. We anticipate that existing statistical comparative approaches will be more commonly applied to studying the genetic basis for phenotypic evolution as whole genomes continue to populate the tree of life. Nevertheless, new methods and approaches will be needed to fully capitalize on the power of clade-scale genomic datasets.

Highlights

Genome sequencing is rapidly spreading beyond model organisms, opening the door to comparative studies that can reveal the genetic basis for phenotypic variation across species. Nevertheless, statistical comparative methods have not been frequently applied to these data.

New phylogenetic methods have been developed with the explicit goal of linking genes and even specific mutations to species differences ('PhyloG2P'). Applications of these methods show great promise for uncovering new sources of functional variation and tackling traits beyond the reach of traditional genetic approaches.

Parallel advances in statistical comparative methods present new avenues for expanding the phylogenetic toolkit and creating tailored approaches for mapping genotype to phenotype.

Most of Biodiversity Is Beyond the Reach of Classical Genetics

One of the fundamental goals of biology is to connect variation across genomes to differences in phenotypes. With advances in sequencing and molecular genetic techniques, this area of biology has blossomed in recent years, revealing the genetic basis for traits ranging from floral scent [1] to sociality [2] to herbivory [3]. At the same time, statistical methods for analyzing these data have also proliferated [4–6]. At their core, however, all classical and population genetic methods for **genotype-to-phenotype mapping** (see Glossary) work by associating genetic variation with differences in the trait of interest. Thus, they require a population with segregating phenotypic variation, which could be produced artificially through crosses or mutagenesis or could occur naturally, such as in polymorphic species or hybrid zones between species. As with any statistical approach, association methods [e.g., **genome-wide association studies (GWASs)**] have significant challenges and pitfalls [6,7]. Still, the loci uncovered by association mapping and similar methods have often been validated in subsequent functional studies [8,9], confirming their ability to identify regions of the genome that contribute to phenotypic differences.

Despite the success of this population genetic program for genotype–phenotype mapping, it presents significant limitations for understanding the genetic basis of phenotypes for most of biodiversity. First, many species cannot be propagated artificially or sampled in the wild at the scale needed for association mapping (usually hundreds of individuals, depending on the trait of interest). Second, and more importantly, many traits of interest are not found segregating in nature nor can different species with contrasting phenotypes be crossed. For example, mammals with and without pouches cannot be crossed, precluding the creation of a mapping population segregating for pouches. As a consequence, our understanding of the genetic basis for phenotypic diversity is concentrated around a narrow range of species and traits – often those that vary in model organisms amenable to genetic studies. Although loci discovered through genetic studies of model species often later help to explain variation at deeper phylogenetic levels (i.e., across species [10,11]), we wonder what we might discover if this research program were inverted (Figure 1). We suggest, and recent studies confirm, that beginning from a phylogenetic

[1]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309, USA
[2]Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[3]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA
[4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[5]Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA
[6]http://www.colorado.edu/smithlab

*Correspondence: Stacey.D.Smith@colorado.edu (S.D. Smith).
@Twitter: @iochromaland (S.D. Smith).

Phenotype only models:

Cornwell, W. and Nakagawa, S. 2017. Phylogenetic comparative methods. Curr.Biol., 27: 327–338. — phenotype models

Phenotype + Genotype models:

Mayrose, I. and Otto, S. P. (**2011**). *A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution*. Mol. Biol. Evol., 28: 759–770. — + DNA model

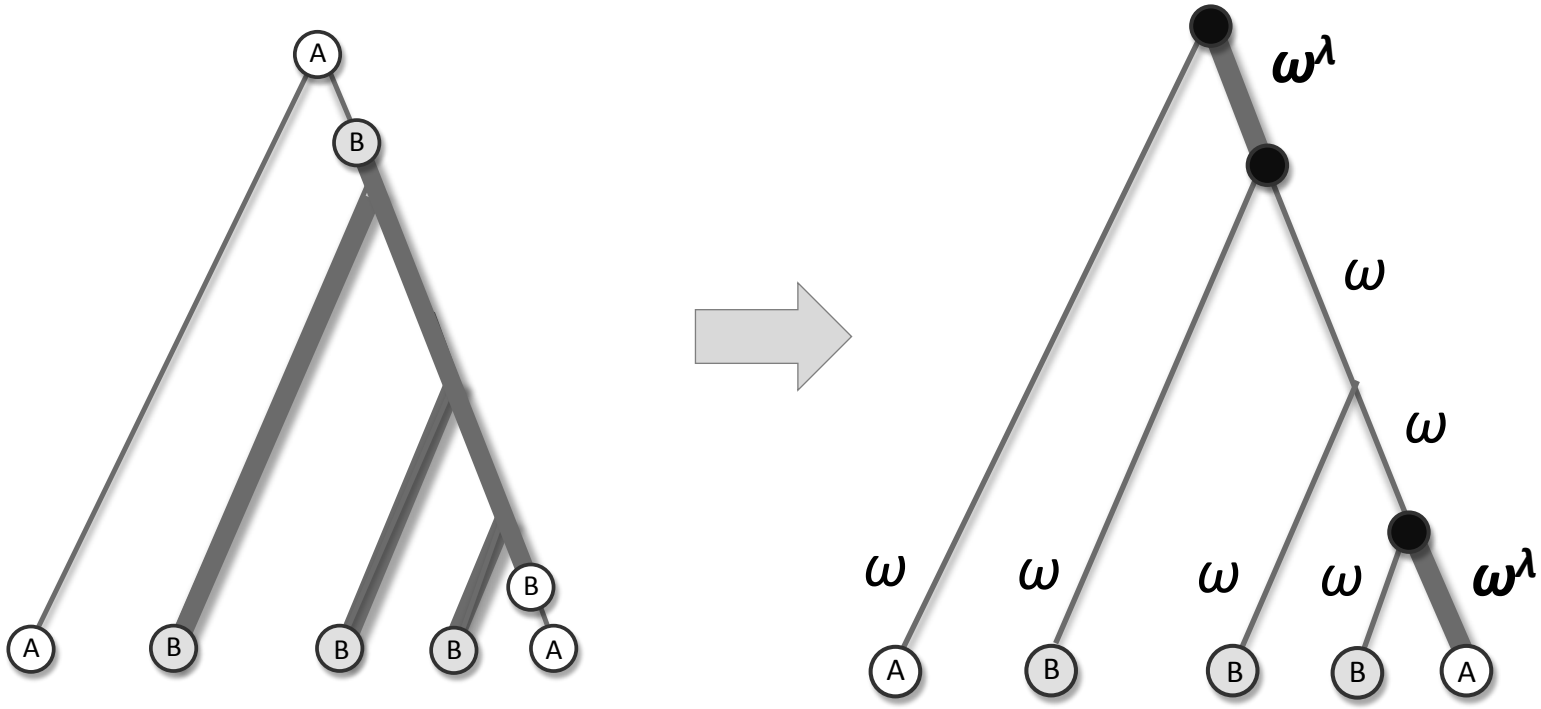
Lartillot, N. and Poujol, R. (**2011**). *A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters*. Mol. Biol. Evol., 28: 729–744. — + codon model

O'Connor, T. D. and Mundy, N. I. (**2013**). *Evolutionary modeling of geneotype-phenotype association and application to the primate coding and non-coding mtdna rate variation*. Evolutionary Bioinformatics, 9: 301–316. — + DNA model

Karin, E. L., Wicke, S., Pupko, T., and Mayrose, I. (**2017**). *An integrated model of phenotypic trait changes and site-specific sequence evolution*. Syst. Biol., 66: 917–933. — + DNA model

Jones, C. T., Youssef, N., Susko, E., & Bielawski, J. P. (**2020**). A Phenotype–Genotype Codon Model for Detecting Adaptive Evolution. Systematic Biology, 69(4), 722-738. — + codon model

Halabi, K., Karin, E. L., Guéguen, L., & Mayrose, I. (**2021**). *A codon model for associating phenotypic traits with altered selective patterns of sequence evolution.* Systematic Biology, 70(3), 608-622. — + codon model



phenotype mapping 1 (of many)

gene evolution

Phenotype only models:

Cornwell, W. and Nakagawa, S. 2017. Phylogenetic comparative methods. Curr.Biol., 27: 327–338. — phenotype models

Phenotype + Genotype models:

Mayrose, I. and Otto, S. P. (**2011**). *A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution*. Mol. Biol. Evol., 28: 759–770. — + DNA model

Lartillot, N. an
substitution ra — + codon model

Detect adaptive molecular evolution

(possibly without $d_N/d_S > 1$)

O'Connor, T. D
association an
Evolutionary B — + DNA model

Karin, E. L., Wicke, S., Pupko, T., and Mayrose, I. (**2017**). *An integrated model of phenotypic trait changes and site-specific sequence evolution*. Syst. Biol., 66: 917–933. — + DNA model

Jones, C. T., Youssef, N., Susko, E., & Bielawski, J. P. (**2020**). A Phenotype–Genotype Codon Model for Detecting Adaptive Evolution. Systematic biology, 69(4), 722-738. — + codon model

Halabi, K., Karin, E. L., Guéguen, L., & Mayrose, I. (**2021**). *A codon model for associating phenotypic traits with altered selective patterns of sequence evolution.* Systematic Biology, 70(3), 608-622. — + codon model

9. model-based inference

You will get "*the basics*"
in the evening
PAML-Lab