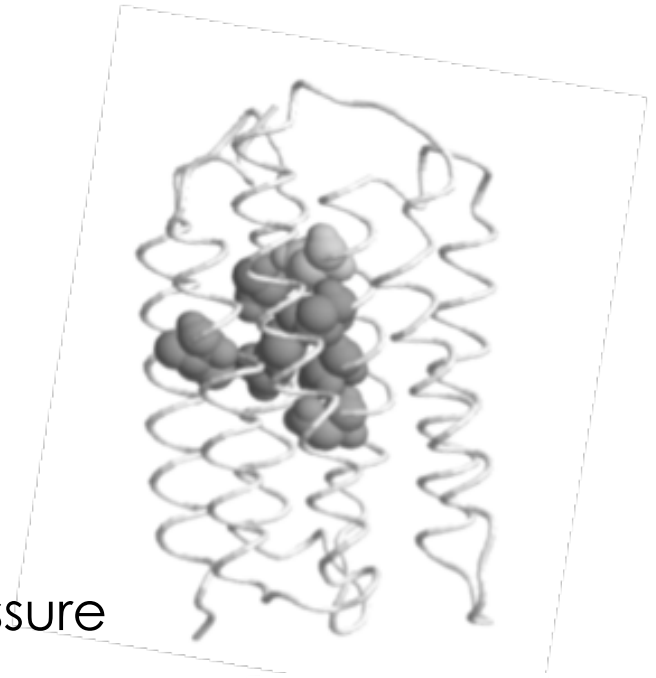


part 3: analysis of natural selection pressure

part 3: analysis of natural selection pressure



markov models are good

---

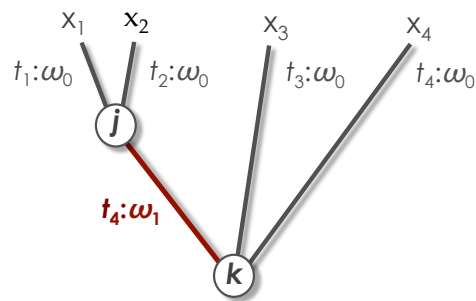
phenomenological codon models do have many benefits:

- principled framework for statistical inference
- avoiding *ad hoc* corrections of “counting” methods
- computation of transition probabilities \*
- explicit use of phylogeny
- model  $\omega$  variation among sites
- model  $\omega$  variation among branches
- many other kinds of models for  $\omega$

\* Computation of transition probabilities accomplishes, in just one step, (1) a proper correction for multiple substitutions, (2) weighting for alternative pathways between codons and (3) is the basis for estimating the values of the model parameters from the data in hand.

## two basic types of models

---



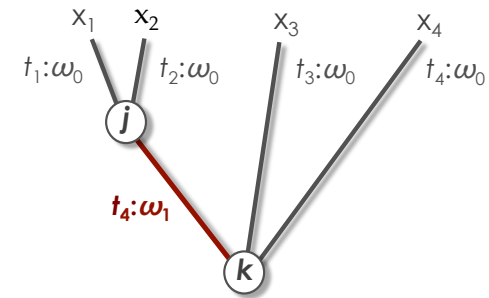
**branch models**  
( $\omega$  varies among  
branches)

$\omega_1$			$\omega_0$	$\omega_1$								$\omega_0$	$\omega_1$			
GTG	CTG	TCT	CCT	GCC	GAC	AAG	ACC	AAC	GTC	AAG	GCC	GCC	TGG	GGC	AAG	GTT
...	...	...	G.C	...	...	...	T..	..T	...	...	...	...	...	...	...	...
...	...	...	..C	..T	...	...	...	...	A..	...	A.T	...	...	..AA	...	A.C
...	..C	...	G.A	..AT	...	..A	...	...	A..	...	AA.	TG.	...	..G	...	A..
...	..C	..G	GA.	..T	...	...	..T	C..	..G	..A	...	AT.	...	..T	...	..G

**site models**  
( $\omega$  varies among sites)

## branch models\*

---

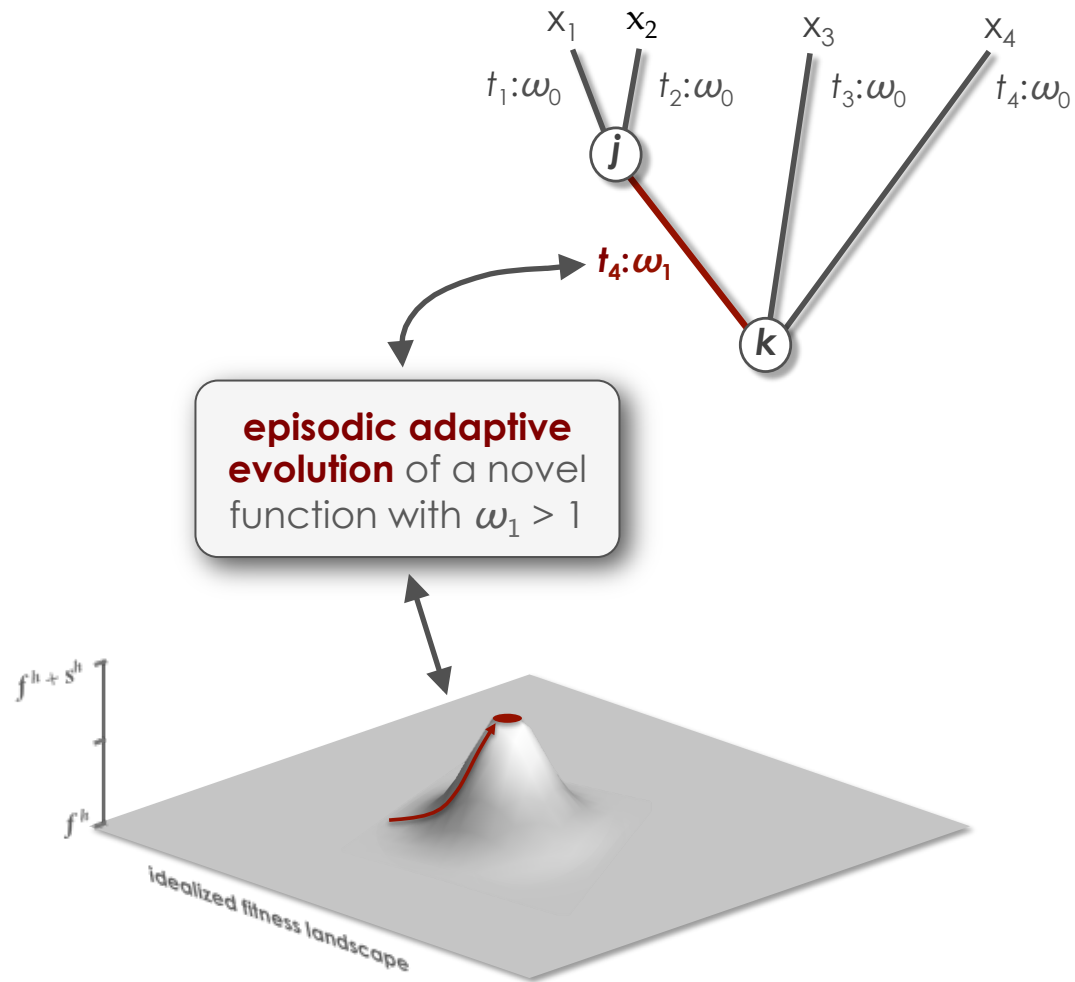


variation ( $\omega$ ) among branches:	approach
Yang, 1998	fixed effects
Bielawski and Yang, 2003	fixed effects
Seo et al. 2004	auto-correlated rates
Kosakovsky Pond and Frost, 2005	genetic algorithm
Dutheil et al. 2012	clustering algorithm

\* these methods can be useful when selection pressure is strongly **episodic**

## interpretation of a branch model

---



## site models\*

```

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 ...

```

variation ( $\omega$ ) 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)

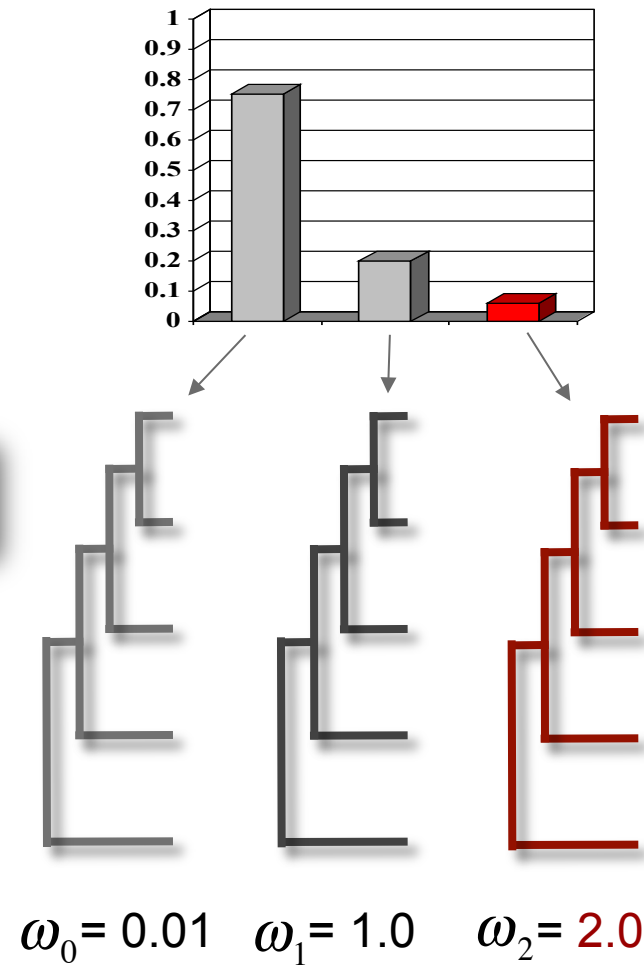
- useful when at some sites evolve under **diversifying selection** pressure over long periods of time
- this is not a comprehensive list

## site models: discrete model (M3)

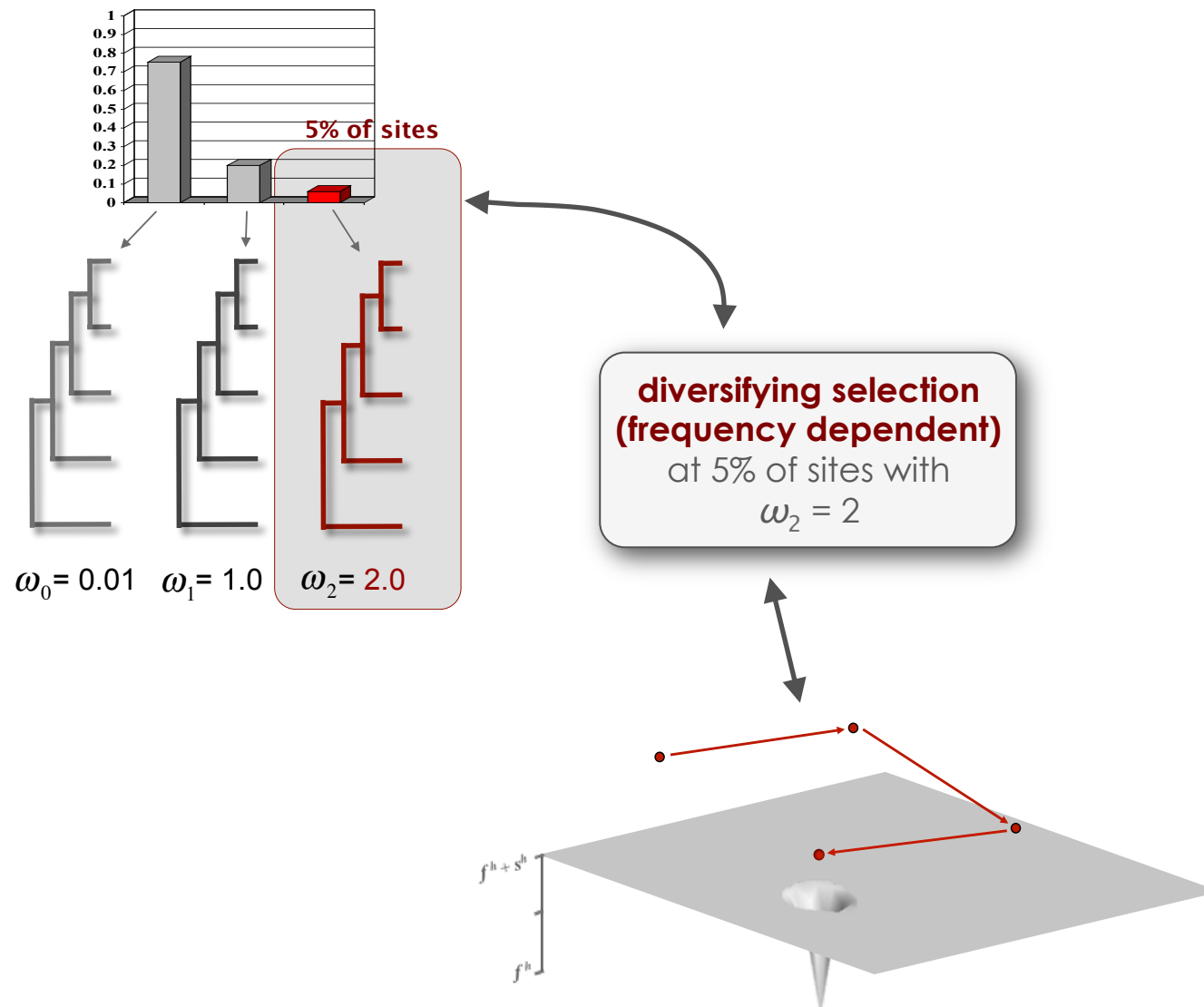
MIXTURE-MODEL LIKELIHOOD

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

conditional likelihood  
calculation (see part 1)



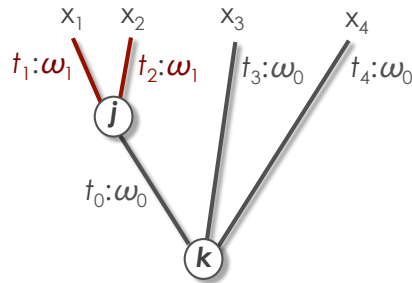
# interpretation of a sites-model





## models for variation among branches & sites

---



$\omega_1$			$\omega_0$	$\omega_1$							$\omega_0$	$\omega_1$				
GTG	CTG	TCT	CCT	GCC	GAC	AAG	ACC	AAC	GTC	AAG	GCC	GCC	TGG	GGC	AAG	GTT
...	...	...	G.C	...	...	...	T..	..T	...	...	...	...	...	...	...	...
...	...	...	..C	..T	...	...	...	...	A..	...	A.T	...	...	..AA	...	A.C
...	..C	...	G.A	.AT	...	..A	...	...	A..	...	AA.	TG.	...	..G	...	A..
...	..C	..G	GA.	..T	...	...	..T	C..	..G	..A	...	AT.	...	..T	...	..G

**branch models**  
( $\omega$  varies among  
branches)

**site models**  
( $\omega$  varies among sites)



## models for variation among branches & sites

---

variation ( $\omega$ ) 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	switching (ML)
Zhang et al. 2005	fixed+mixture (ML)
Kosakovsky Pond et al. 2011, 2012	full mixture (ML)

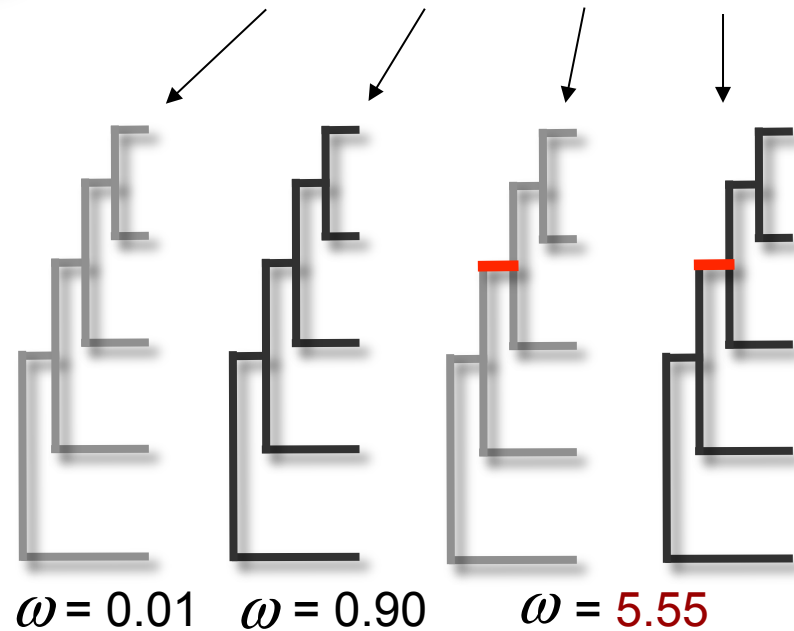
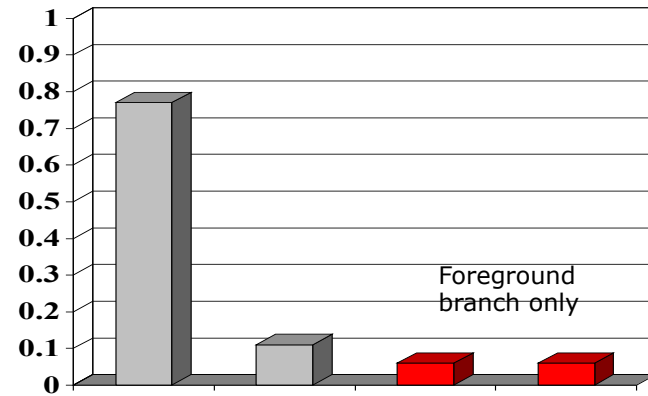
\* *these methods can be useful when selection **pressures change over time at just a fraction of sites***

\* *it can be a challenge to apply these methods properly (**more about this later**)*

## branch-site “Model B”

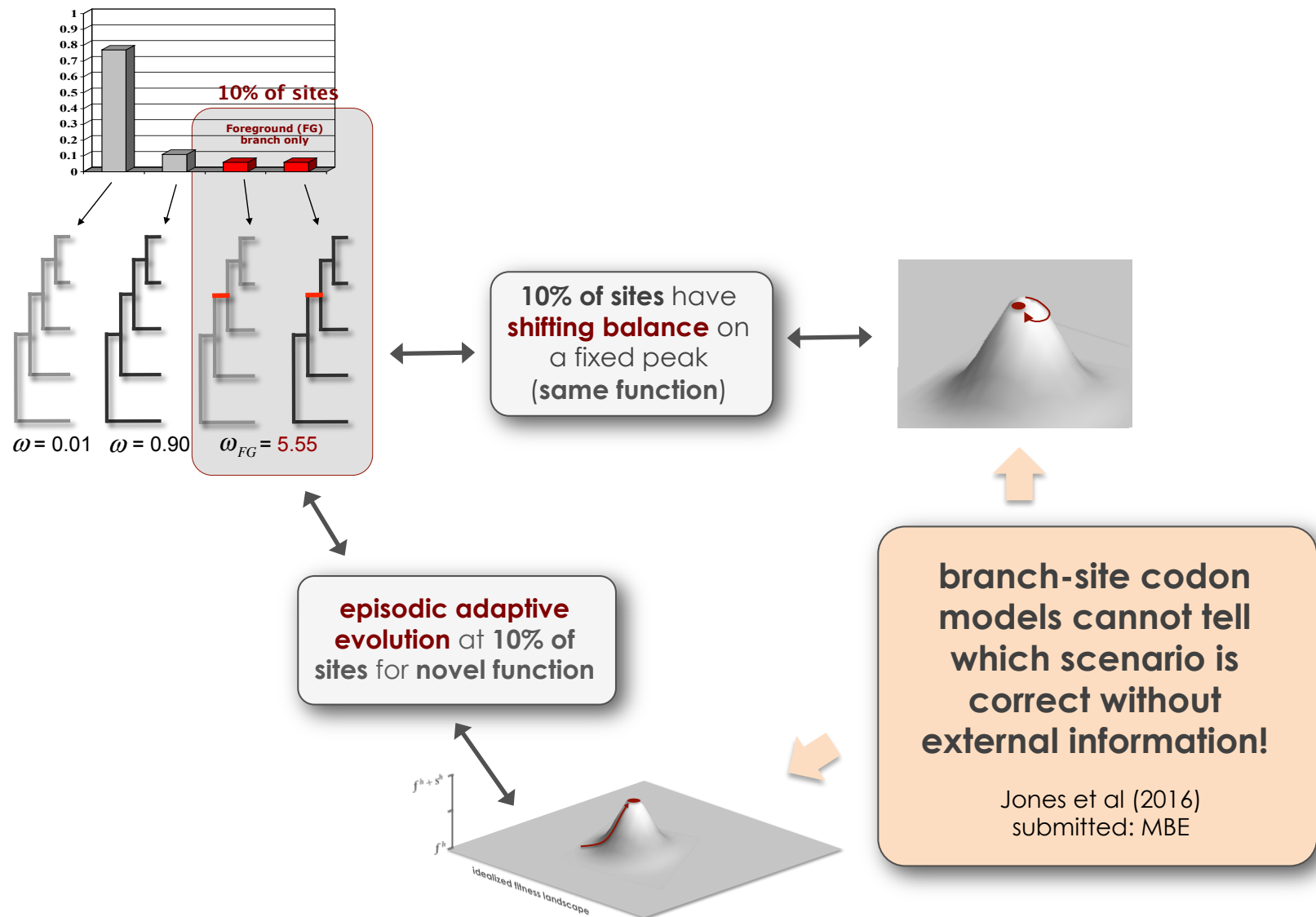
MIXTURE-MODEL LIKELIHOOD

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



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

two scenarios can yield branch-sites with  $dN/dS > 1$



model-based inference

### 3 analytical tasks

**task 1.** parameter estimation (e.g.,  $\omega$ ) 

**task 2.** hypothesis testing

**task 3.** make predictions (e.g., sites having  $\omega > 1$  )

## task 1: parameter estimation

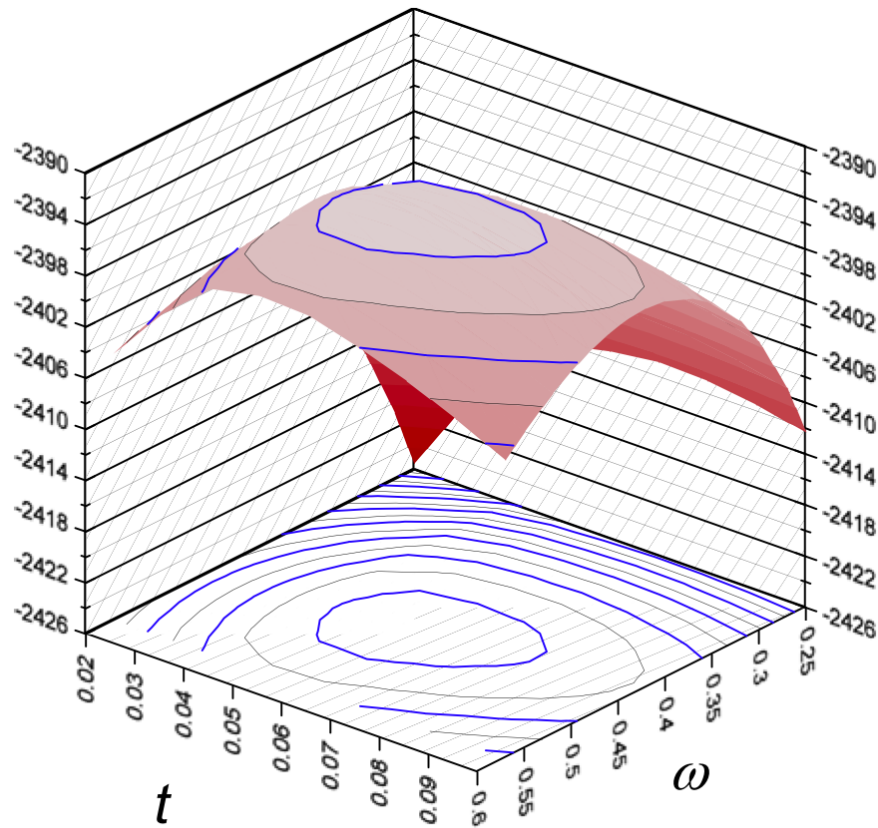
---

$t, \kappa, \omega$  = unknown constants estimated by ML

$\pi$ 's = empirical [GY: F3×4 or F61 in Lab]

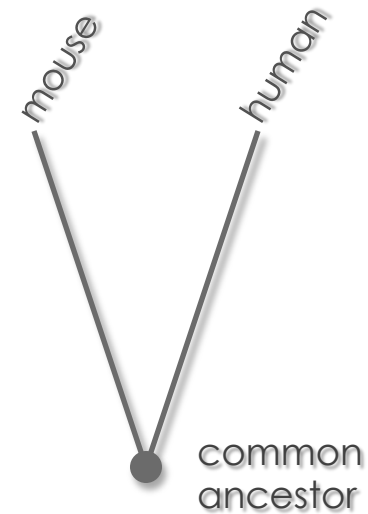
use a numerical hill-climbing algorithm to maximize the likelihood function

## task 1: parameter estimation



**Parameters:**  $t$  and  $\omega$

**Gene:** acetylcholine  $\alpha$  receptor



$\ln L = -2399$



## task 2: statistical significance

---

task 1. parameter estimation (e.g.,  $\omega$ ) ✓

task 2. hypothesis testing ← **LRT**

task 3. prediction / site identification

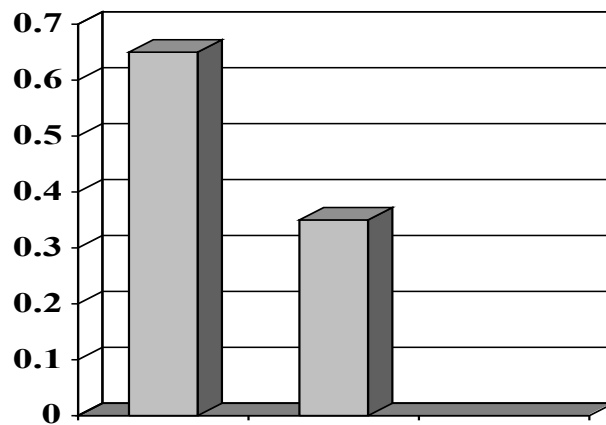
## task 2: likelihood ratio test for positive selection

$H_0$ : variable selective pressure but NO positive selection (M1)

$H_1$ : variable selective pressure with positive selection (M2)

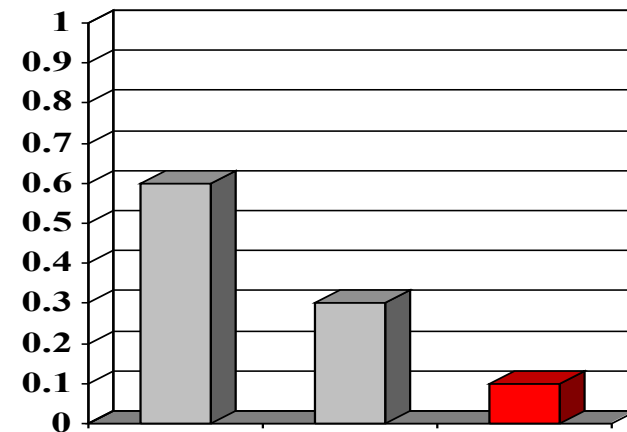
Compare  $2\Delta l = 2(l_1 - l_0)$  with a  $\chi^2$  distribution

Model 1a



$\hat{\omega} = 0.5$  ( $\omega = 1$ )

Model 2a



$\hat{\omega} = 0.5$  ( $\omega = 1$ )  $\hat{\omega} = 3.25$

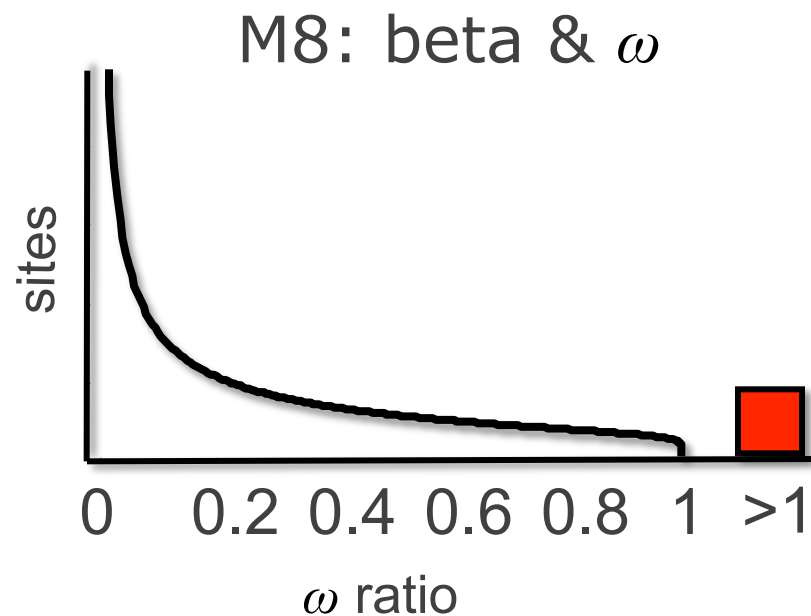
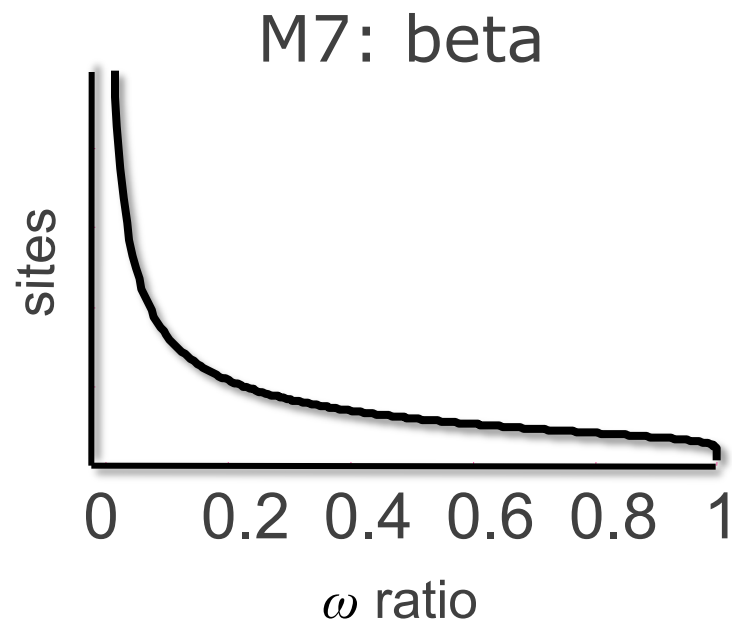
## task 2: likelihood ratio test for positive selection

---

**H<sub>0</sub>**: Beta distributed variable selective pressure (M7)

**H<sub>1</sub>**: Beta plus positive selection (M8)

Compare  $2\Delta l = 2(l_1 - l_0)$  with a  $\chi^2$  distribution



task 3: identify the selected sites

---

task 1. parameter estimation (e.g.,  $\omega$ ) ✓

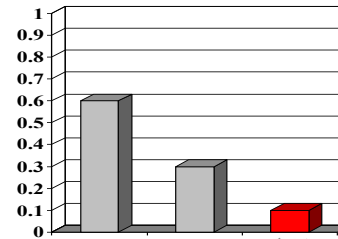
task 2. hypothesis testing ✓

task 3. prediction / site identification ← **Bayes' rule**

## task 3: which sites have $dN/dS > 1$

---

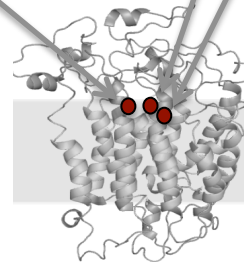
**model:**  
9% have  $\omega > 1$



**Bayes' rule:**  
site 4, 12 & 13

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	...

**structure:**  
sites are in contact



## Bayes' rule: yet another (silly) example

---

**Suppose that a population consists of 60% males and 40% females, and a disease occurs at the rate 1% in males and 0.1% in females.**

$Q_1$ : What is the probability that any individual carries the disease?

$$A_1: 0.6 \times 0.01 + 0.4 \times 0.001 = 0.0064$$

$$P(D) = P(M)P(D|M) + P(F)P(D|F)$$

## Bayes' rule: yet another (silly) example

---

Q<sub>2</sub>: Given that an individual carries the disease, what is the probability that it is a male?

$$A_2: 0.6 \times 0.01 / 0.0064 = 0.94$$

$$P(M|D) = \frac{P(M) P(D|M)}{P(D)}$$

See Yang and Bielawski (2000) TREE 15:496-503 for a detailed presentation of this example

from Paul Lewis' lecture ....

## Bayes' rule in statistics

The diagram illustrates Bayes' rule with the following components and labels:

- Likelihood of hypothesis  $\theta$** : Points to the term  $\Pr(D|\theta)$  in the numerator.
- Prior probability of hypothesis  $\theta$** : Points to the term  $\Pr(\theta)$  in the numerator.
- Posterior probability of hypothesis  $\theta$** : Points to the term  $\Pr(\theta|D)$  on the left side of the equation.
- Marginal probability of the data (marginalizing over hypotheses)**: Points to the denominator  $\sum_{\theta} \Pr(D|\theta) \Pr(\theta)$ .

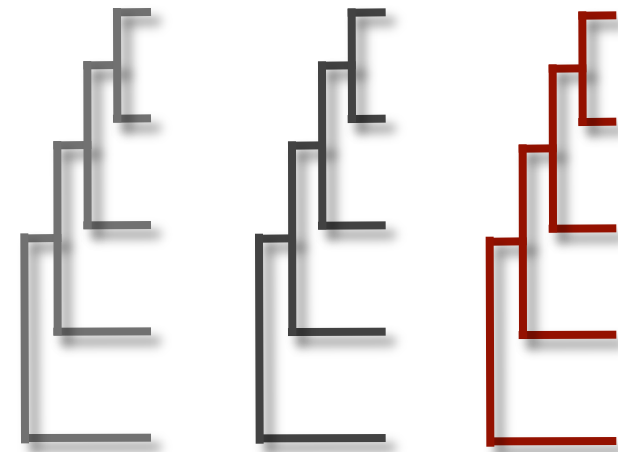
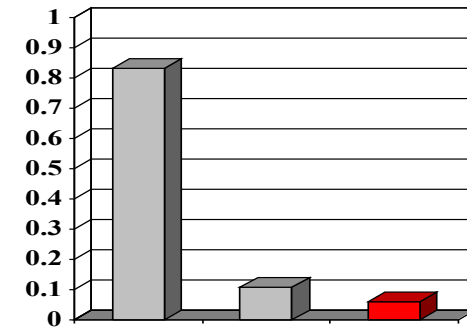
$$\Pr(\theta|D) = \frac{\Pr(D|\theta) \Pr(\theta)}{\sum_{\theta} \Pr(D|\theta) \Pr(\theta)}$$



identifying selected sites under a codon model

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

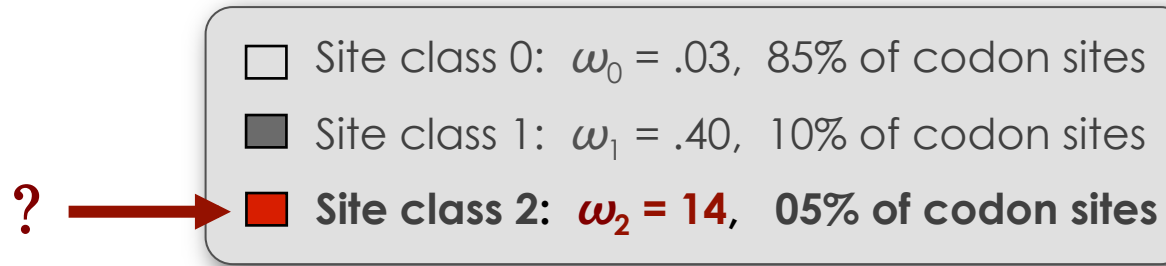
↑                      ↑                      ↑  
**Total**                      **Prior**                      **Likelihood**  
**probability**



$$\begin{array}{lll} \omega_0 = 0.03 & \omega_1 = 0.40 & \omega_2 = 14.1 \\ p_0 = 0.85 & p_1 = 0.10 & p_2 = 0.05 \end{array}$$

## Bayes' rule for identifying selected sites

---



Prior probability of hypothesis ( $\omega_2$ )

Likelihood of hypothesis ( $\omega_2$ )

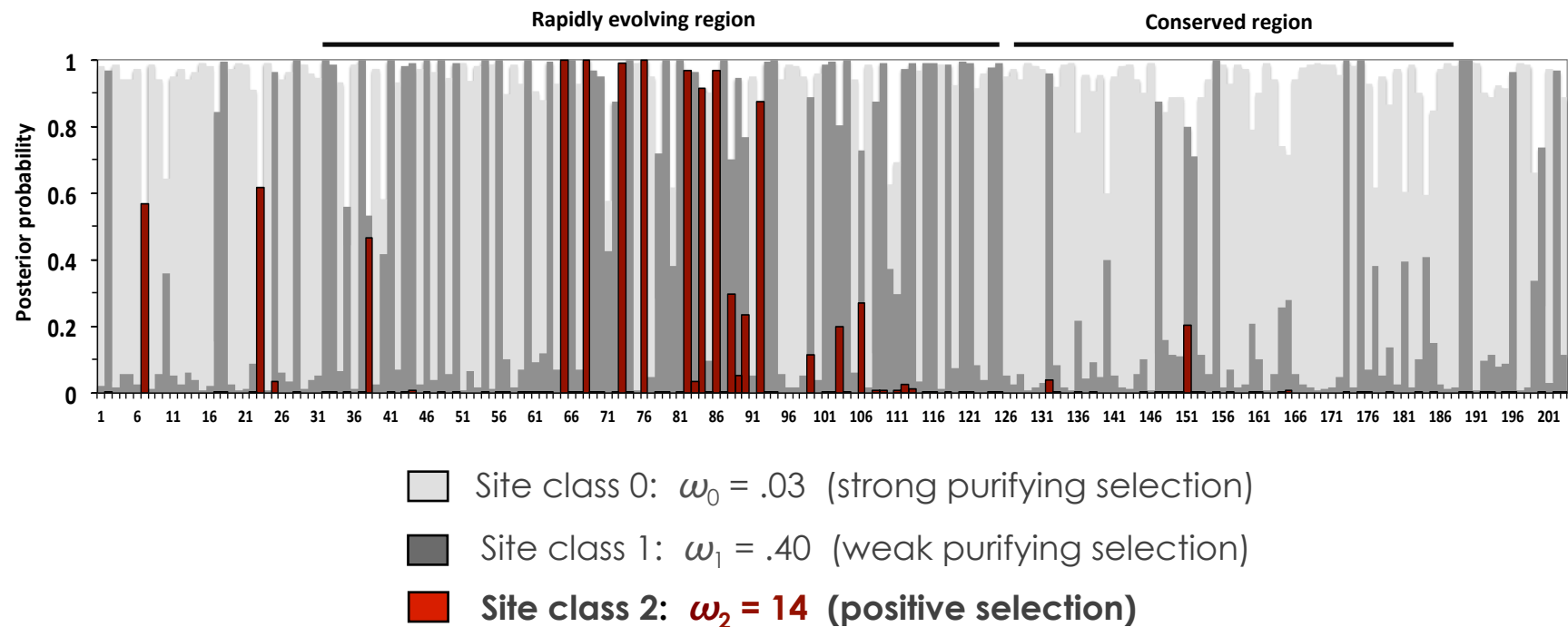
$$P(\omega_2 | x_h) = \frac{P(\omega_2)P(x_h | \omega_2)}{\sum_{i=0}^{K-1} P(\omega_i)P(x_h | \omega_i)}$$

Posterior probability of hypothesis ( $\omega_2$ )

Marginal probability (Total probability) of the data

### task 3: Bayes rule for which sites have $dN/dS > 1$

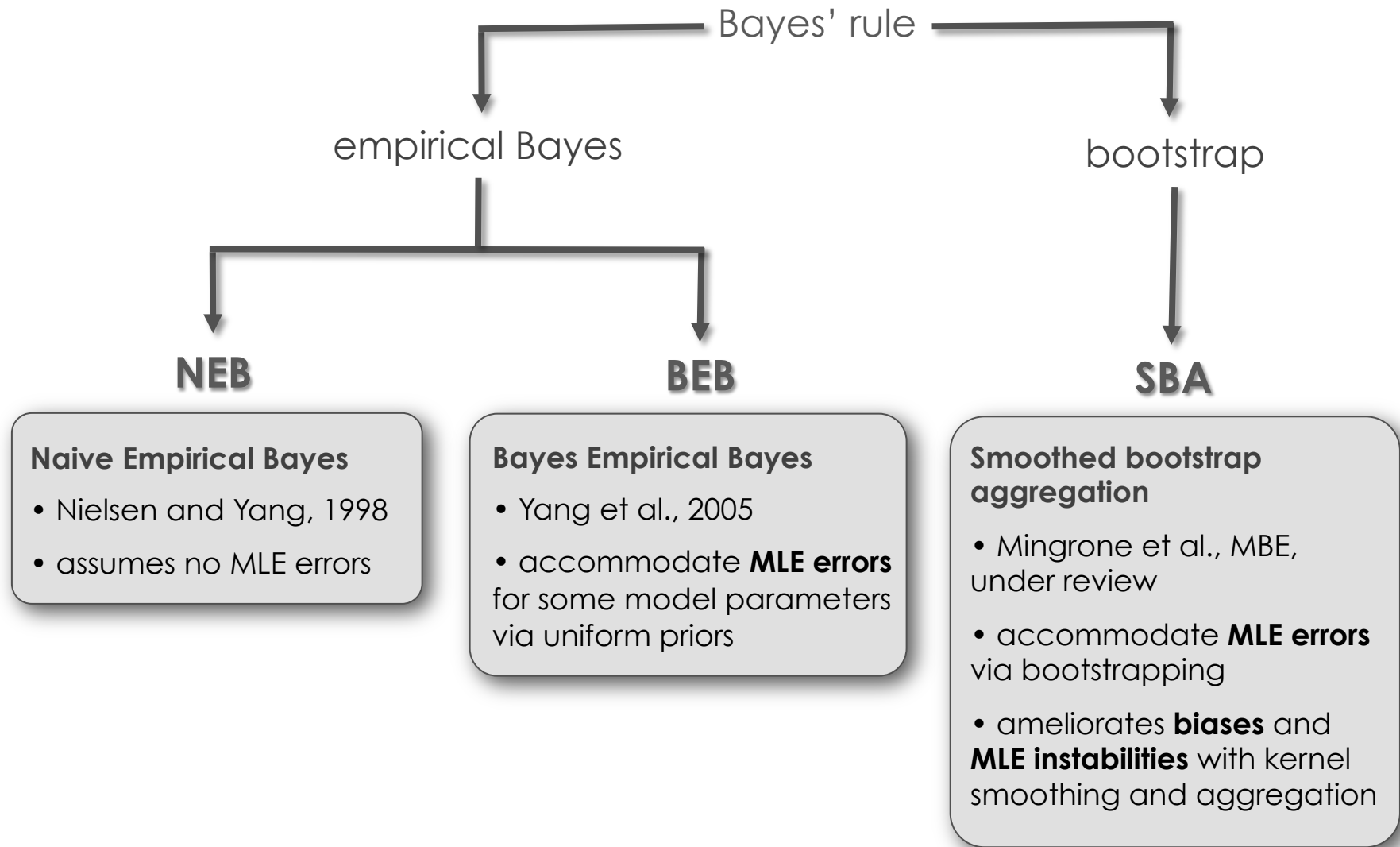
---



**NOTE:** The posterior probability should NOT be interpreted as a “P-value”; it can be interpreted as a measure of relative support, although there is rarely any attempt at “calibration”.

### task 3: Bayes rule for which sites have $dN/dS > 1$

---



## model based inference

---

task 1. parameter estimation (e.g.,  $\omega$ )

task 2. hypothesis testing

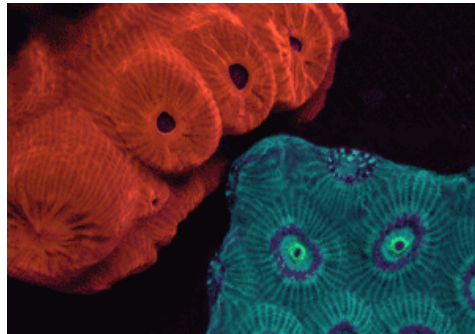
task 3. prediction / site identification

**let's put this into practice ...**

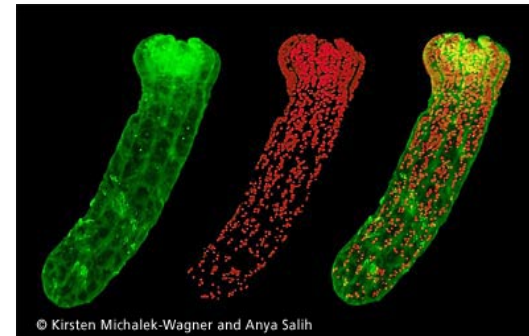
example analysis (& *experimental validation*)

## colour diversity of coral pigments (GFPs)

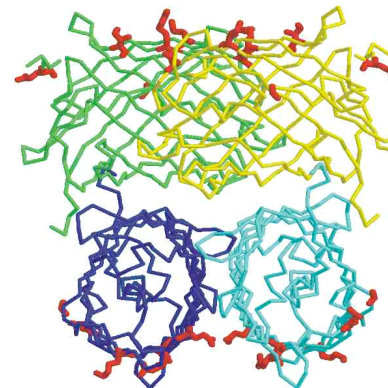
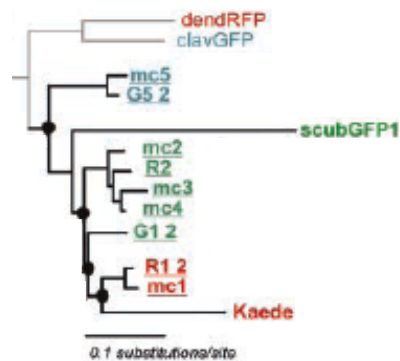
---



Red/blue colour morphs of the great star coral *Montastraea cavernosa*



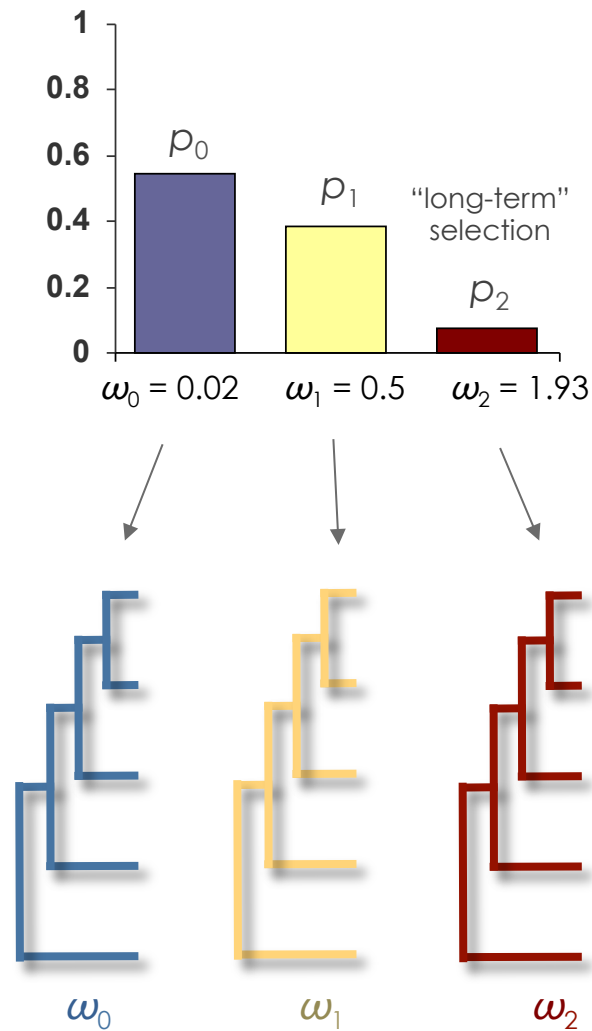
© Kirsten Michalek-Wagner and Anya Salih



- Is color diversity tuned by natural selection?
- Is there a relationship between colour and endosymbiotic algae?

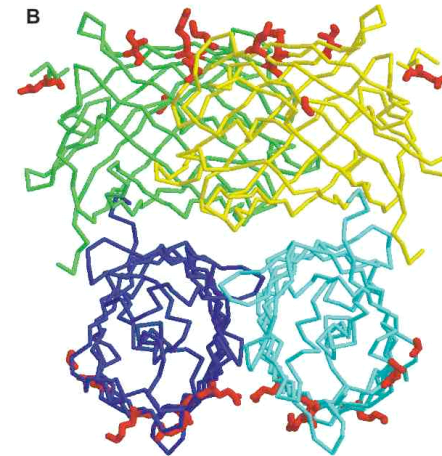
See Field et al. 2006 J. Mol. Evol. 62(3):332-9 for details.

## signal 1: long term (diversifying) selection



### Bayes' rule:

$$P(\omega_2 | \mathbf{x}_h) = \frac{p_2 P(x_h | \omega_2)}{P(x_h)} = \frac{p_2 P(x_h | \omega_2)}{\sum_{i=0}^{K-1} p_i P(x_h | \omega_i)}$$

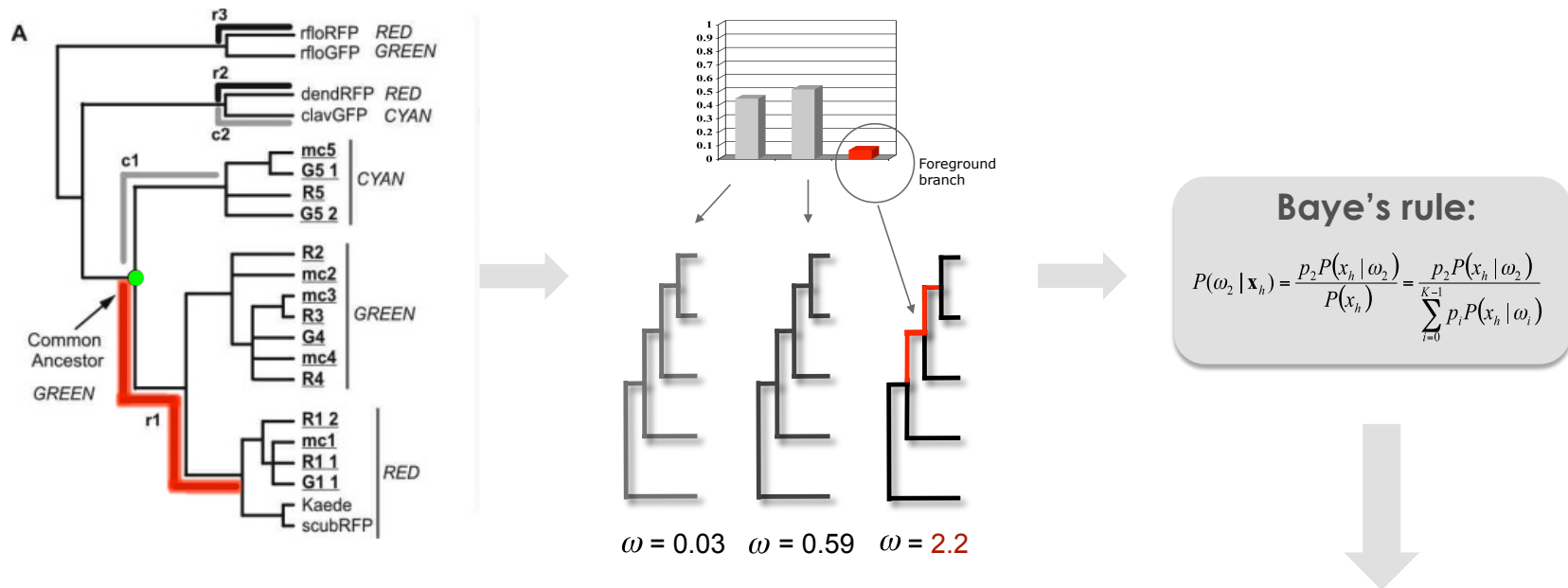


sites in red correspond to the protein-binding region of **non-colored** homologs of these GFP proteins

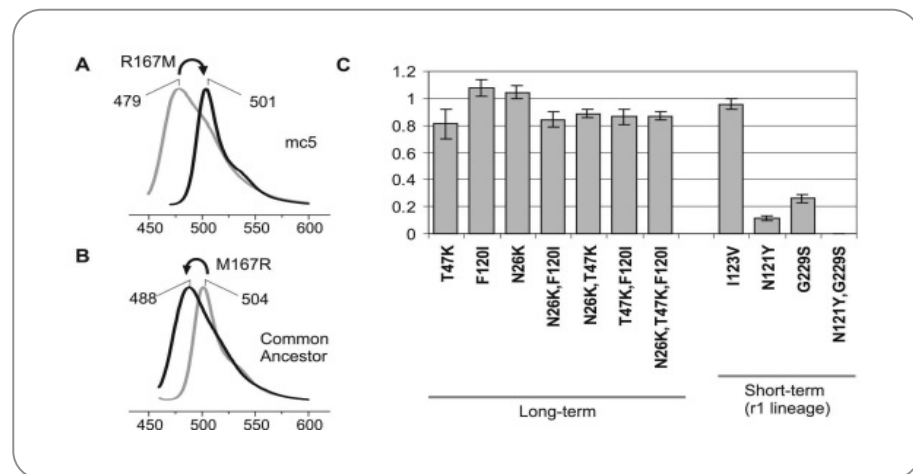
See Field et al. 2006 J. Mol. Evol. 62(3):332-9 for details.



## signal 2: episodic selection

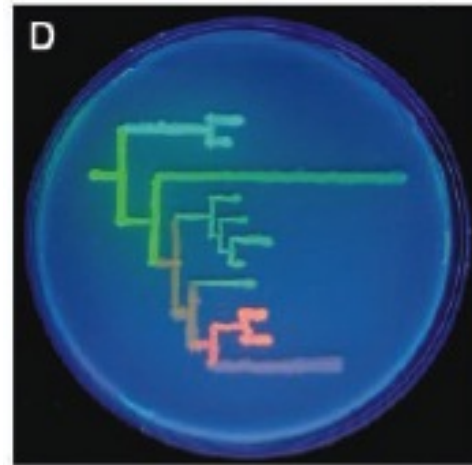
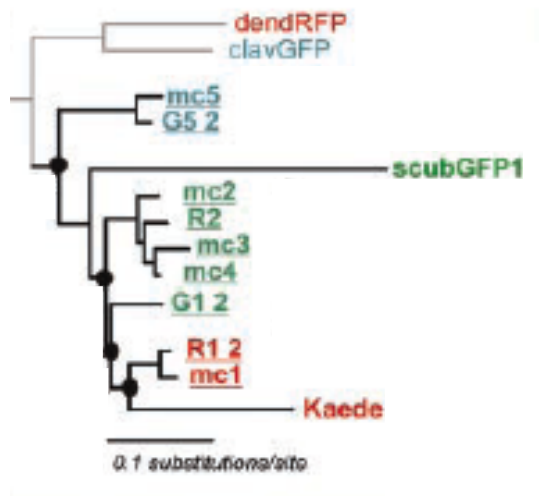


See Field et al. 2006 J. Mol. Evol. 62(3):332-9 for details.



just for fun ....

---




Bacteria were engineered to express the extant and ancestral GFP-like proteins. These bacteria were then cultured in a pattern that corresponded to the GFP-LIKE gene tree.

Ugalde JA, Chang BS, Matz MV.  
Evolution of coral pigments recreated.  
Science. (2003). 305:1433.

false biological conclusions

## false biological conclusions

---

1. codon usage 
2. process variation among sites
3. process variation over time
4. recombination
5. regularity conditions not met

# how to model codon frequencies?

---

Sums of codon usage counts 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
Ile I ATT	4	Thr T ACT	2	Asn N AAT	5	Ser S AGT	1
ATC	12	ACC	11	AAC	17	AGC	4
ATA	0	ACA	2	Lys K AAA	1	Arg R AGA	0
Met M ATG	4	ACG	4	AAG	37	AGG	1
Val V GTT	0	Ala A GCT	0	Asp D GAT	2	Gly G GGT	4
GTC	2	GCC	38	GAC	11	GGC	6
GTA	1	GCA	2	Glu E GAA	0	GGA	11
GTG	25	GCG	3	GAG	30	GGG	0

## how to model codon frequencies?

---

	substitution rates are proportional to empirical frequency of:
Goldman and Yang 1994 (GY):	target codon
Muse and Gaut 1994 (MG):	target nucleotide

See Rodrigue et al. (2008) for a comparison of GY and MG style codon models that suggests the MG style, combined with parameters for codon preferences, might be the most desirable core-model for future development.

The MutSel process (part 1) is inherently a process whereby the transition probability depends on the target nucleotide (MG).

## how to model codon frequencies?

---

example: A  $\rightarrow$  C

AAA  $\rightarrow$  CAA

AAA  $\rightarrow$  ACA

AAA  $\rightarrow$  AAC

depending on the gene/  
genome, the method could  
yield **biased estimates of  $dN/dS$** , See the following for cases:

- Aris-Brosou & Bielawski (2006) Gene 378: 58-64.
- Yap et al. (2010) MBE 27: 726-734.
- Spielman & Wilke (2015) MBE 32: 1097- 1108.

---

$\Delta$  at codon position

1<sup>st</sup>

2<sup>nd</sup>

3<sup>rd</sup>

GY

$\pi_{CAA}$

$\pi_{ACA}$

$\pi_{AAC}$

MG

$\pi_c^1$


$\pi_c^2$

$\pi_c^3$

---

## false biological conclusions

---

1. codon usage
2. process variation among sites 
3. process variation over time
4. recombination
5. regularity conditions not met



## sequence evolution is complex

---



loop structures extend into  
extra-cellular space: **Hydrophilic  
amino acids here**

$\omega_0 \pi_0 \kappa_0 c_0$

cell membrane in grey; helix  
structures span the membrane:  
**Hydrophobic amino acids here**

$\omega_1 \pi_1 \kappa_1 c_1$

loop structures extend into  
cytoplasm: **Hydrophilic amino  
acids here**


$\omega_2 \pi_2 \kappa_2 c_2$

**codon models:** biological interpretation of differences among sites in  $\omega$  requires that such differences are due to selection pressure alone

GY-type codon models: variable  $\omega$ 's + c's among sites = variable  $d_N$  &  $d_S$  among sites

## modeling process variation among sites


---

process variation among sites	software & references
<ul style="list-style-type: none"><li>• synonymous rate</li><li>• nonsynonymous rate</li></ul>	several methods in: <b>HyPhy</b> : Kosakovsky Pond et al. (2005) <b>Datamonkey</b> : Delpont et al. (2010)
<ul style="list-style-type: none"><li>• baseline DNA/RNA substitution rate</li><li>• nonsynonymous rate</li></ul>	<b>MultiLayer</b> : Rubinstein et al. (2011)
<ul style="list-style-type: none"><li>• baseline DNA/RNA substitution rate</li><li>• transition/transversion ratio</li><li>• codon frequencies</li><li>• nonsynonymous rate</li></ul>	<b>LiBaC</b> : Bao et al. (2008) 

several studies show **false signal** for  $dN/dS > 1$  is possible when process variation among sites in inadequately modeled

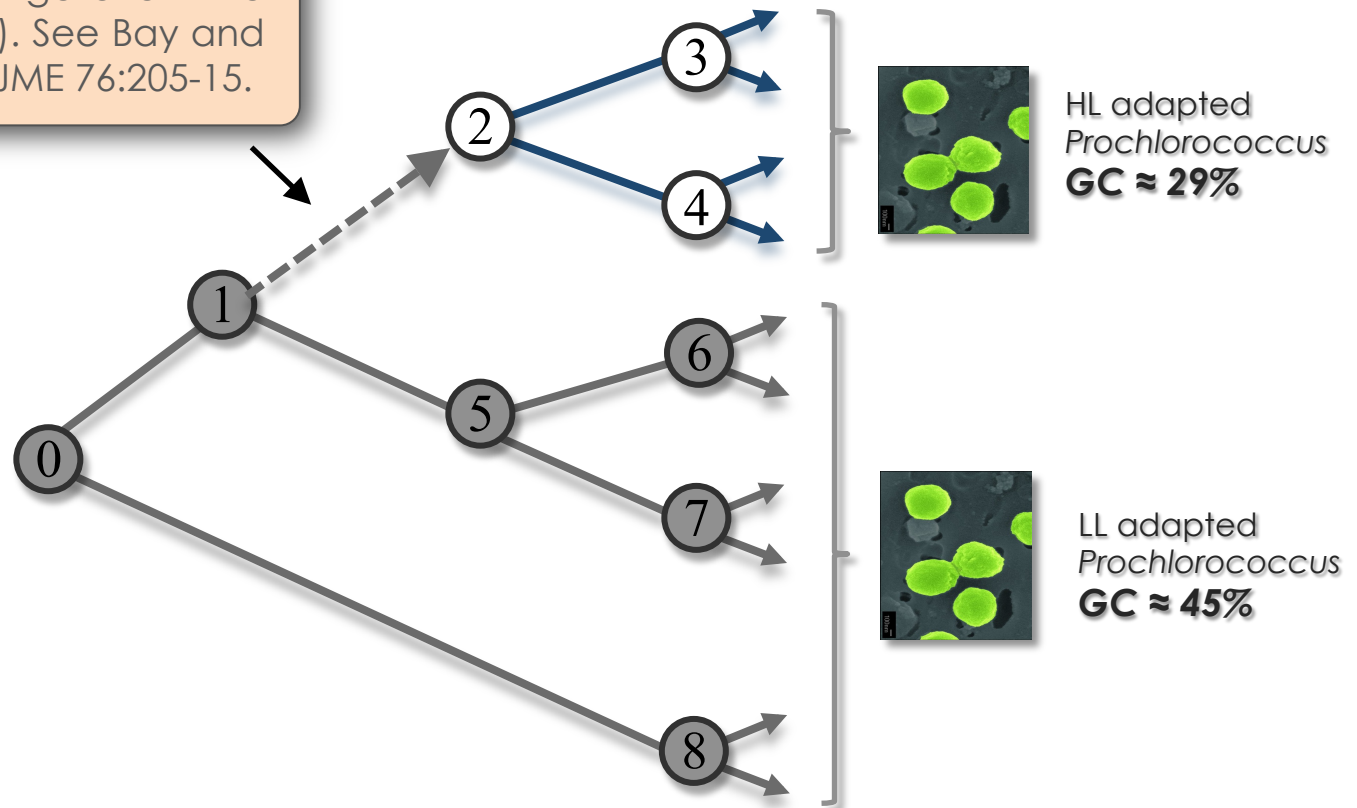
## false biological conclusions

---

1. codon usage
2. process varies among sites
3. process varies over time 
4. recombination
5. regularity conditions not met


# non-stationary codon frequencies

false signal for  $dN/dS > 1$  is possible when codon frequencies change over time (non-stationarity). See Bay and Bielawski (2013) JME 76:205-15.



## false biological conclusions

---

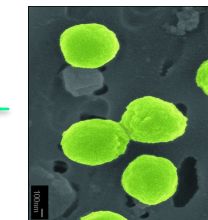
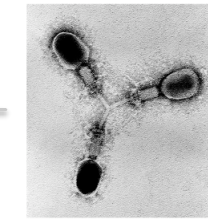
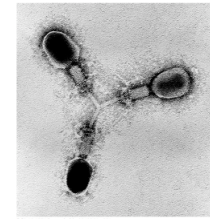
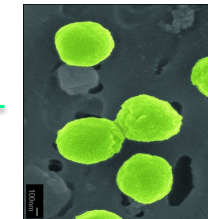
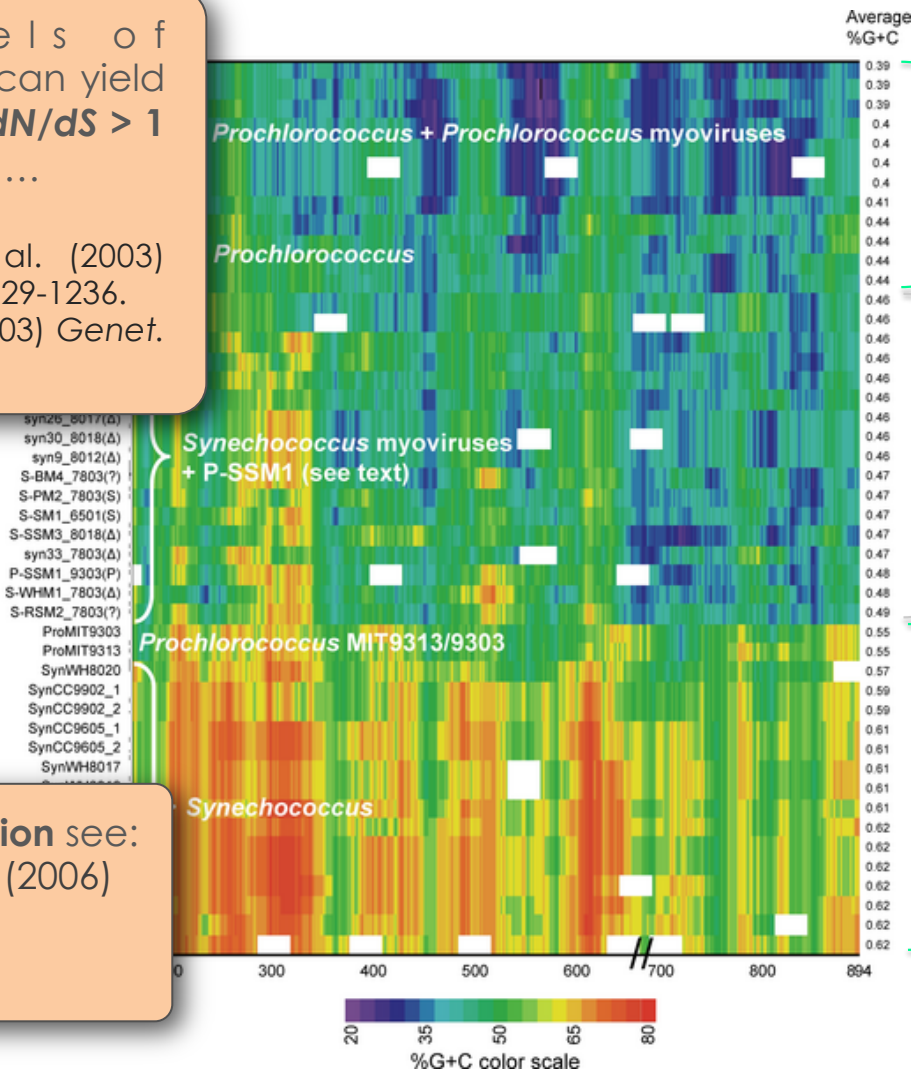
1. codon usage
2. variation among sites
3. variation over time
4. recombination 
5. regularity conditions not met

# recombination

high levels of recombination can yield **false signal for  $dN/dS > 1$**  via the LRT. see ...

- Anisimova, et al. (2003) *Genetics*, 164:1229-1236.
- Shriner et al. (2003) *Genet. Res.* 81:115-121


**for a nice solution** see:  
Scheffler et al. (2006)  
*Bioinformatics*,  
22:2493-2499.



Note: Recombination adds among site variation relative to both process and phylogeny! See Sullivan et al. 2006 PLoS Biology 4: e234 for details.

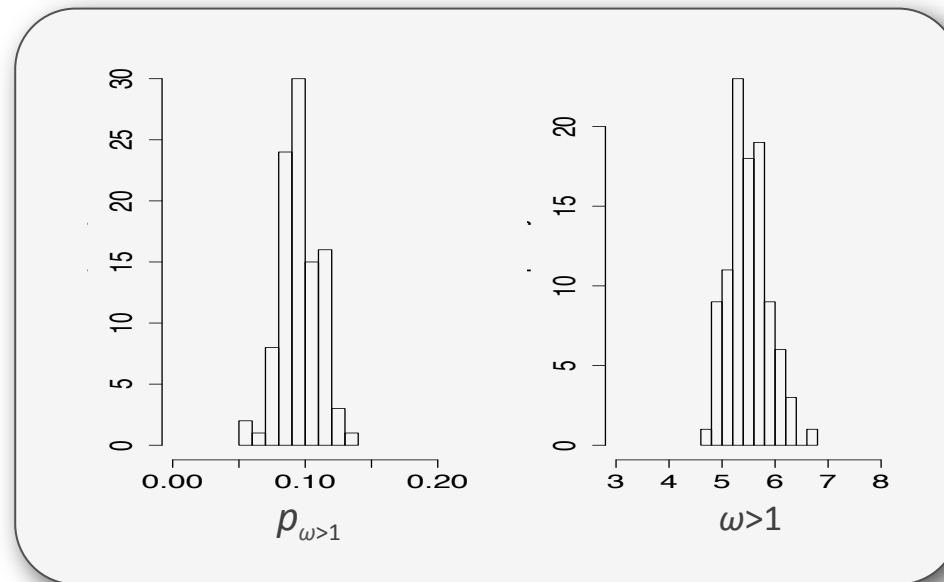
## false biological conclusions

---

1. codon usage
2. variation among sites
3. variation over time
4. recombination
5. regularity conditions not met 

regularity conditions have been met

---



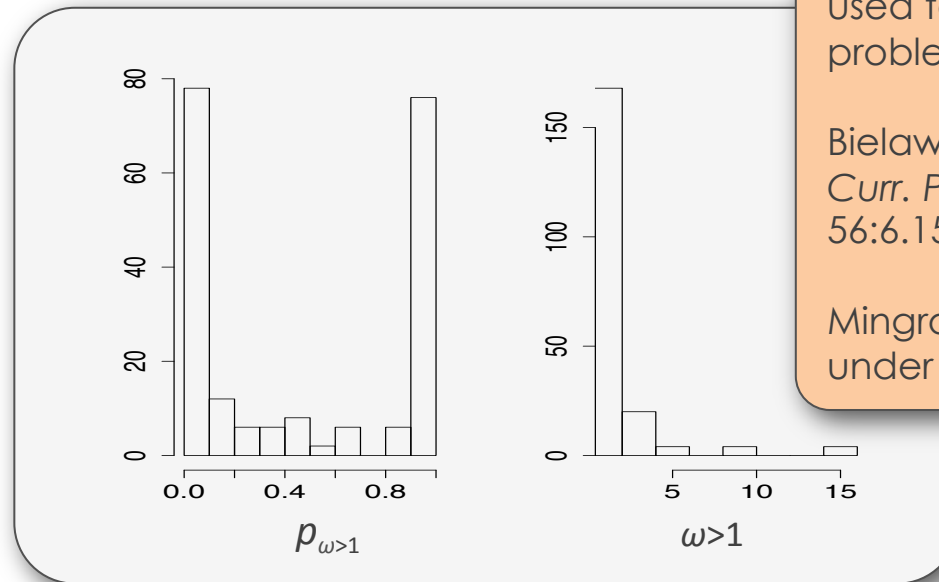
### Normal MLE uncertainty (M2a)

- large sample size with regularity conditions
- MLEs approximately unbiased and minimum variance

$$\hat{\theta} \sim N\left(\theta, I(\hat{\theta})^{-1}\right)$$



regularity conditions have **NOT** been met



**bootstrapping** can be used to diagnose this problem:

Bielawski et al. (2016)  
*Curr. Protoc. Bioinf.*  
56:6.15.

Mingrone et al., *MBE*,  
under review

### MLE instabilities (M2a)

- small sample sizes and  $\theta$  on boundary
- continuous  $\theta$  has been discretized (e.g., M2a)
- non-Gaussian, over-dispersed, divergence among datasets

best practices

## best practices in evolutionary surveys

---

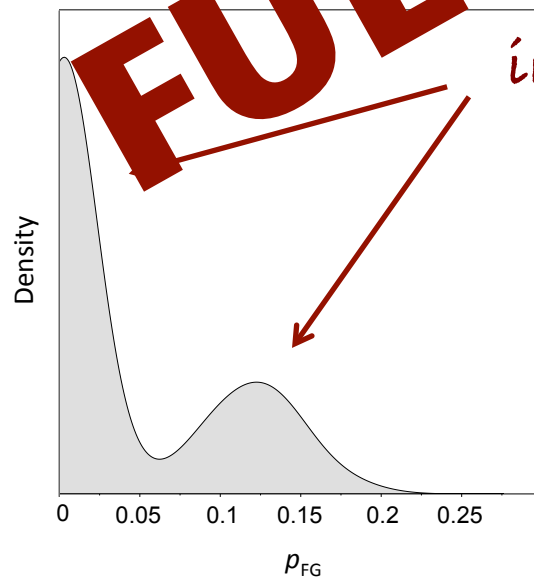
1. processing and Q.C. (in large scale surveys)
2. alignment (independent evaluations)
3. recombination
4. robustness: MG vs GY style codon model
5. robustness: alternative tree topologies
6. robustness: variation in baseline DNA/RNA rates
7. bootstrapping

for discussion of best practices in large scale gene surveys see:

- Baker et al. (2016) *Genetics*, 203:905-22
- Bielawski et al. (2016) *Curr. Protoc. Bioinf.*, 56: Unit 6.15

## nuclear receptor *NR1D1*: positive selection along human lineage ?

1. alignment (independent evaluations) ✓
2. recombination ✓
3. robustness: MG vs GY ✓
4. robustness: tree topologies ✓
5. robustness: baseline DNA/RNA rates ✓
6. bootstrapping



instabilities in  
the MLEs

### KEY

$\omega_M$  :  $\omega_{\text{Mammal}}$

$\omega_{GA}$  :  $\omega_{\text{GreatApe}}$


$\omega_{HC}$  :  $\omega_{\text{Human-Chimpanzee}}$

$\omega_H$  :  $\omega_{\text{Human}}$

What are the next steps in codon models?

# What are the next steps in codon models?

1. applications of the MutSel framework

- 
- Tamuri AU et al. (2014) *Genetics* 197:257
  - Tamuri et al. (2012) *Genetics* 190:1101
  - Yang Z & Nielsen R. (2008) *Mol Biol Evol.* 25:568
  - Nielsen & Yang Z. (2003) *Mol Biol Evol* 20:1231

2. joint modeling of genotype & phenotype

- 
- Nabholz et al. (2013) *Genome Biol Evol* 5:1273
  - Lartillot & Delsuc (2012) *Evolution* 66:1773
  - Lartillot & Poujol (2011) *Mol Biol Evol.* 28:729

**THE END.**