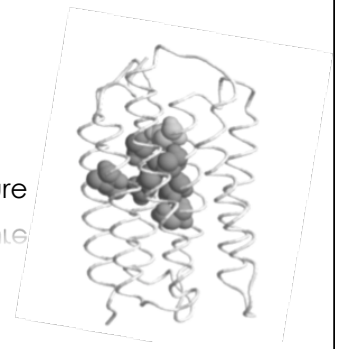
part 3: analysis of natural selection pressure



types of codon models

"OMEGA MODELS"

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

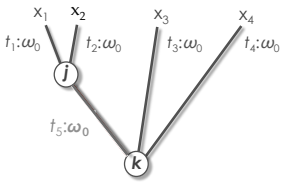
Goldman and Yang (1994)
Muse and Gaut (1994)

this codon model "**M0**"

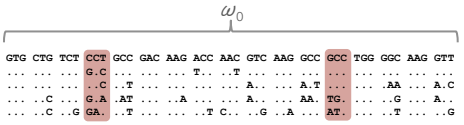
"OMEGA MODELS"

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

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

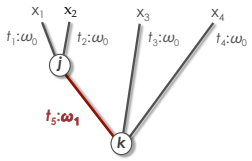


same ω
for all branches

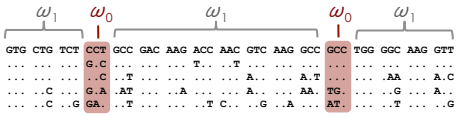


same ω
for all sites

two basic types of models

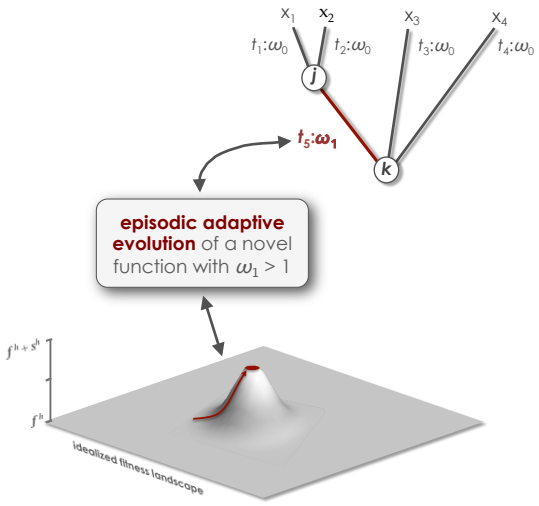


branch models
(ω varies among branches)

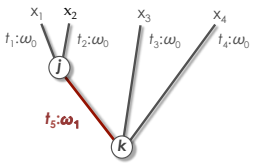


site models
(ω varies among sites)

interpretation of a branch model



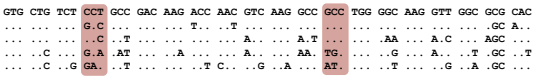
branch models*



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***

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)

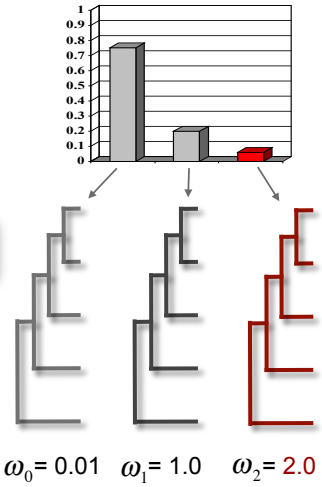
- *useful when at some sites evolve under **diversifying selection** pressure over long periods of time*
- *this is not a comprehensive list*

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

MIXTURE-MODEL LIKELIHOOD

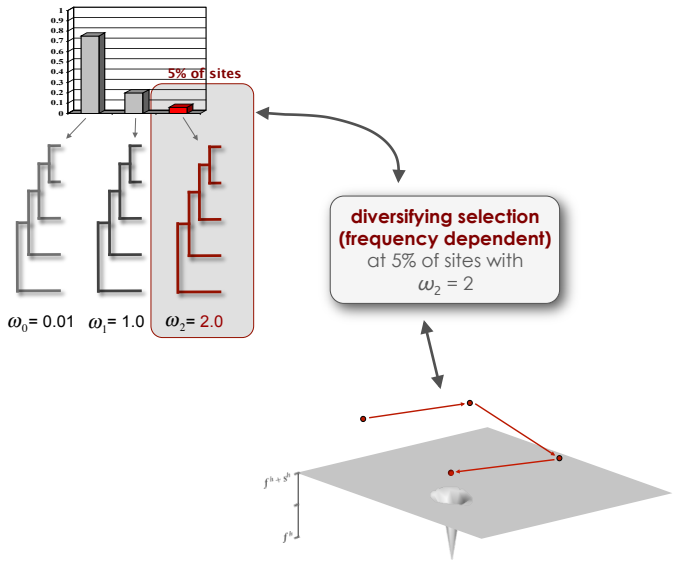
$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$

conditional likelihood calculation (see part 1)

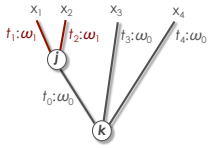


$\omega_0 = 0.01$ $\omega_1 = 1.0$ $\omega_2 = 2.0$

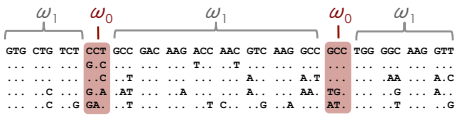
interpretation of a sites-model



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

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	covarion-like (ML)

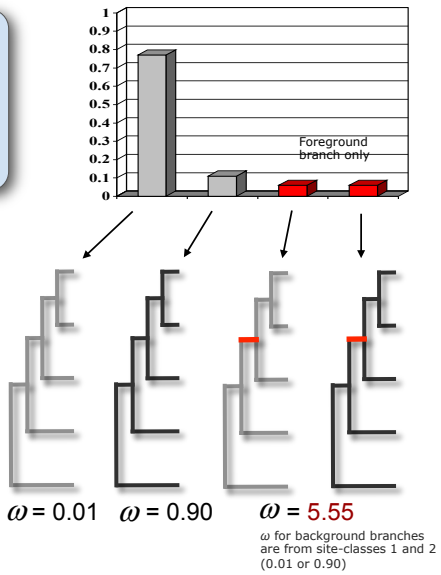
* *these methods can be useful when selection* ***pressures change over time at just a fraction of sites***

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

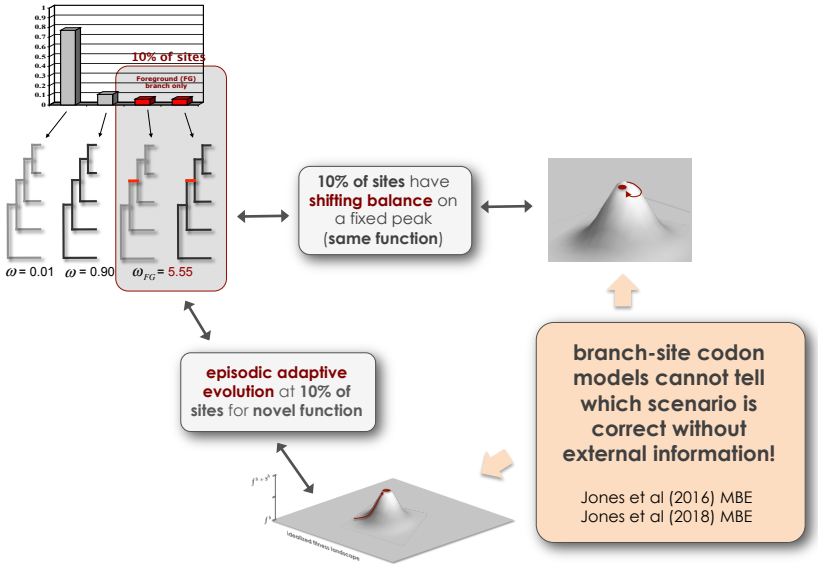
branch-site "Model B"

MIXTURE-MODEL LIKELIHOOD

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



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



model-based inference

"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)

model based inference

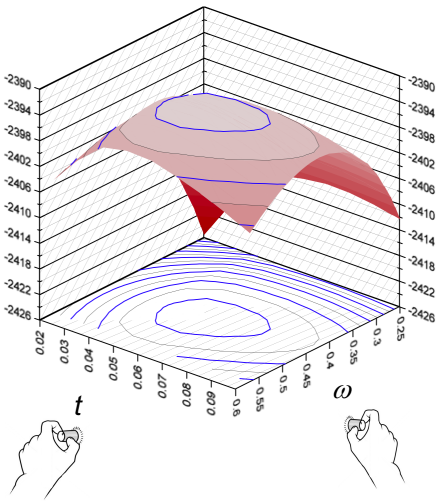
3 analytical tasks

task 1. parameter estimation (e.g., ω)

task 2. hypothesis testing

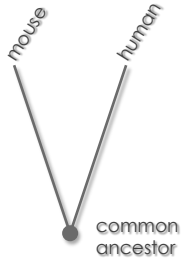
task 3. make predictions (e.g., sites having $\omega > 1$)

task 1: parameter estimation



Parameters: t and ω

Gene: acetylcholine α receptor



lnL = -2399

task 2: statistical significance

task 1. parameter estimation (e.g., ω) ✔

task 2. hypothesis testing ⬅ LRT

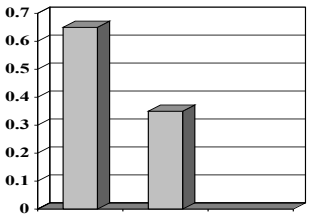
task 3. prediction / site identification

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)

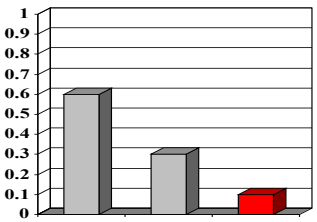
Compare **$2\Delta l$** = $2(l_1 - l_0)$ with a χ^2 distribution

Model 1a (**M1a**)



$\hat{\omega}$ = 0.5 (ω = 1)

Model 2a (**M2a**)



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

task 3: identify the selected sites

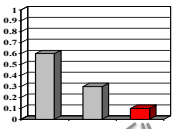
task 1. parameter estimation (e.g., ω) ✔

task 2. hypothesis testing ✔

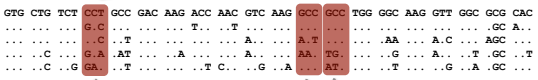
task 3. prediction / site identification ← **Bayes' rule**

task 3: which sites have *dN*/*dS* > 1

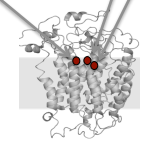
model:
10% have $\omega > 1$



Bayes' rule:
site 4, 12 & 13



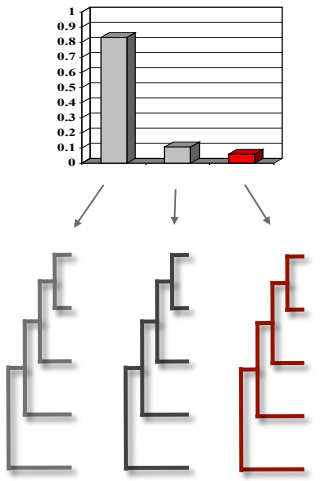
structure:
sites are in contact



review the mixture likelihood (model **M3**)

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

Total probability | Prior | Likelihood



ω_0= 0.03 ω_1= 0.40 ω_2= 14.1

p_0= 0.85 p_1= 0.10 p_2= 0.05

Bayes' rule for identifying selected sites

Site class 0: ω_0 = .03, 85% of codon sites

Site class 1: ω_1 = .40, 10% of codon sites

? → **Site class 2: ω_2 = 14, 05% of codon sites**

Prior probability of hypothesis (ω_2)

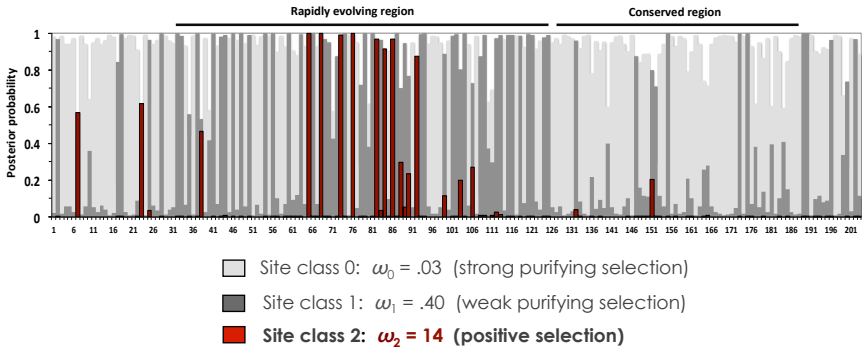
Likelihood of hypothesis (ω_2)

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

Posterior probability of hypothesis (ω_2)

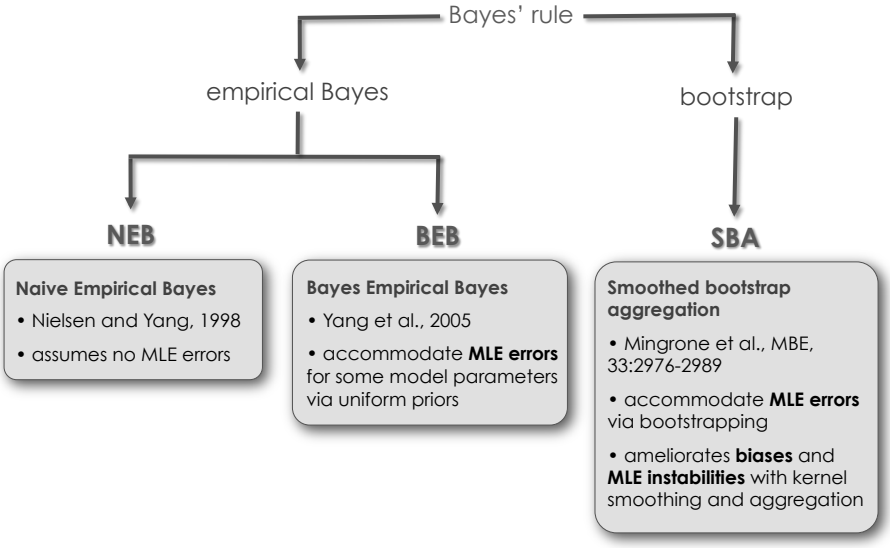
Marginal probability (Total probability) of the data

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

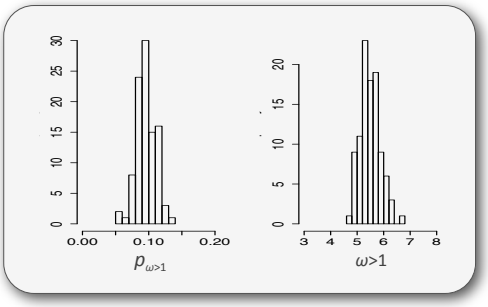


critical question:

Have the requirements for maximum likelihood inference been met?

(rarely addressed in real data analyses)

regularity conditions have been met

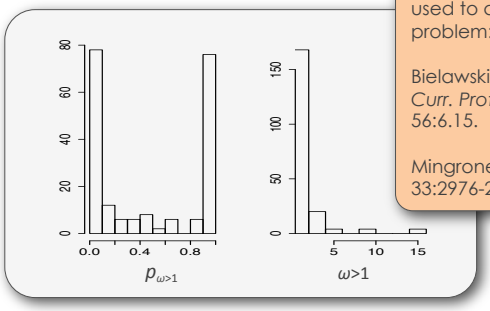


Normal MLE uncertainty (M2a)

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

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

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*, 33:2976-2989

MLE instabilities (M2a)

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

software for codon models in the ML framework

PAML: a package of programs for process modeling

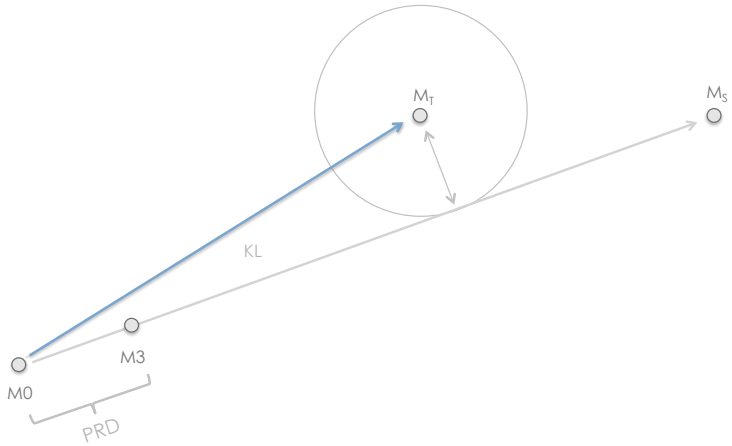
HyPhy: comparative sequence analysis using stochastic evolutionary models; http://www.hyphy.org/

DataMonkey: a server that supports a variety of HYPHY tools at no cost; http://www.datamonkey.org/

COLD: a program that implements a general-purpose parametric (GPP) codon model. Most codon models are special cases of the GPP codon model. https://github.com/tjk23/COLD

codeml_SBA: a program that implements smoothed Bootstrap Aggregation for Assessing Selection Pressure at Amino Acid Sites. https://github.com/Jehops/codeml_sba.

ModL: a program for restoring regularity when testing for positive selection using codon models https://github.com/jehops/codeml_modl



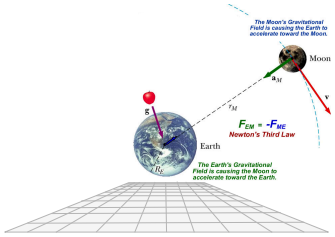
part 4: phenomenological load and biological inference

biological inference
part 4: phenomenological load and

phenomenological load

review types of models

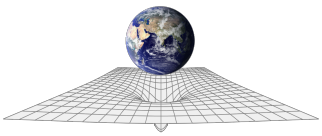
phenomenological



Newton

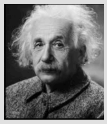
$$F = -\frac{Gm_1m_2}{r^2}$$

mechanistic



Einstein

$$G_{\alpha\beta} = 8\pi T_{\alpha\beta}$$



phenomenological load

molecular evolution is **process** and **pattern**

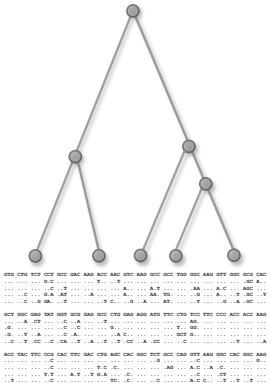
process ⇨ **pattern**

"MUTSEL MODELS"

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

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

Halpern and Bruno (1998)



phenomenological load

Maximum phenomenological model for sequence data: explains all variation in a particular dataset

- so-called "**saturated model**" (multinomial model)
- does not generalize to other datasets
- no information about process
- highest lnL score (useless?)

site pattern

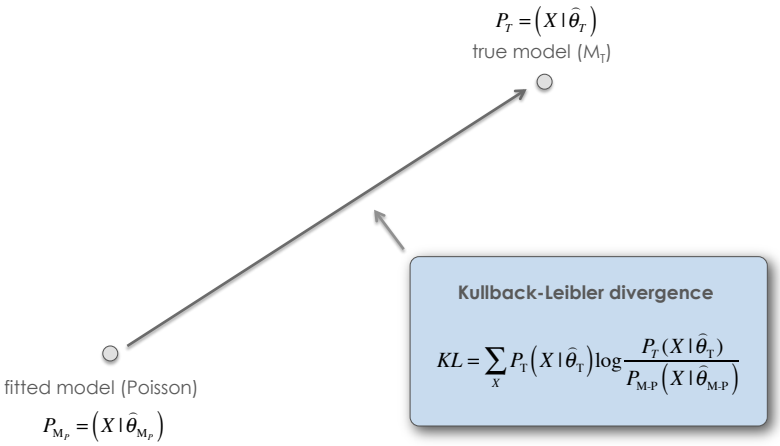
4

```
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 ...
```

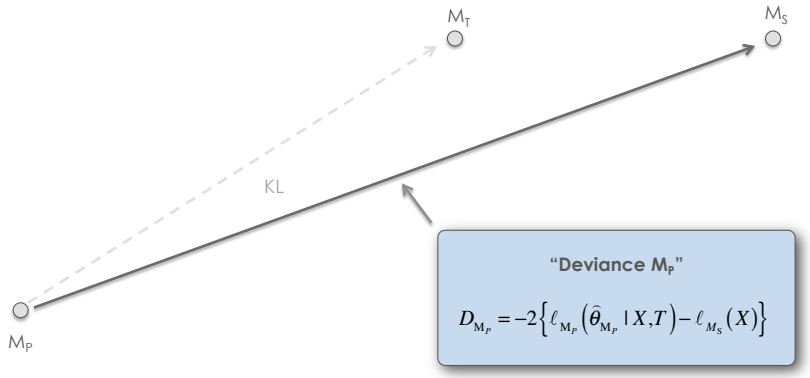
Question: Does anyone really care, at all, that **site pattern No.4** occurs 33 times in *my sample* of 5 mammalian mt genomes?

phenomenological load

a different look at the issue …



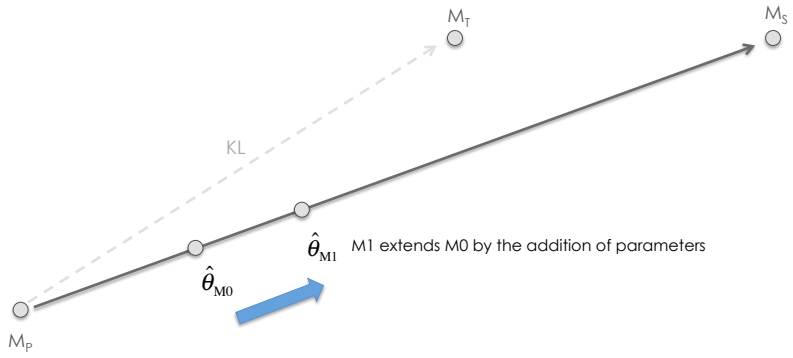
Not to scale!



Poisson model (M_P):
single rate parameter

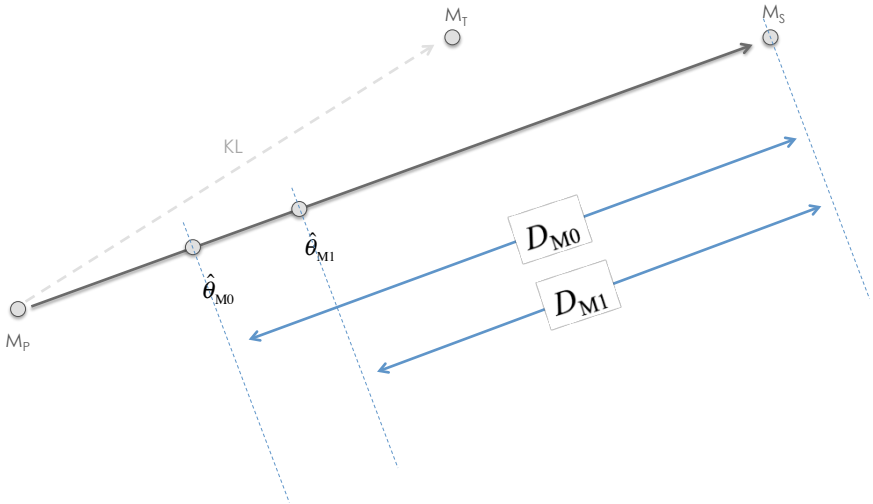
Line:
subspace

Saturated model (M_S):
as many parameters as unique site patterns



KEY POINT: addition of any parameter will reduce the deviance

$$\text{LLR} = D_{M0} - D_{M1} = -2\left\{\ell_{M1}\left(\hat{\theta}_{M1}\middle|X,T\right) - \ell_{M2}\left(\hat{\theta}_{M2}\middle|X,T\right)\right\}$$



The **Likelihood Ratio Test (LRT)** "manages" phenomenological variability (not mechanistic variability)

let's do a simulation study

and

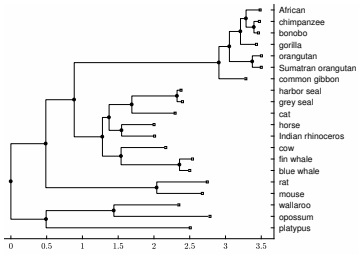
let's use "double mutations" and "triple mutations" as an example

example double (D): ATG (Met) ➔ AAA (Lys)

example triple (T): AAA (Lys) ➔ GGG (GLY)

the simulation and the outcomes…

process ($\mathbf{M_T}$):

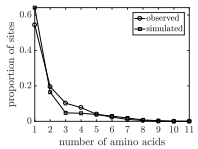


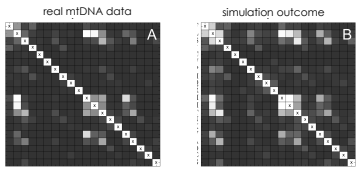
simulation

- MutSel
- f^h differ for each site
- **NO** DT-mutations
- 12 mt proteins (3331 codons)
- 20 mammals

outcome ($\boldsymbol{X}$):

we need outcomes to match up





heat maps: proportion of sites having a given pair of AAs

Our simulated data LOOKS LIKE the REAL DATA!

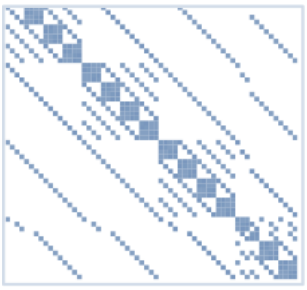
DT: Double and Triple mutations

Example double: ATG (Met) → AAA (Lys) [α parameter]

Example triple: AAA (Lys) → GGG (GLY) [β parameter]

M0 Q matrix

- **2** parameters (κ and ω)
- DT not allowed

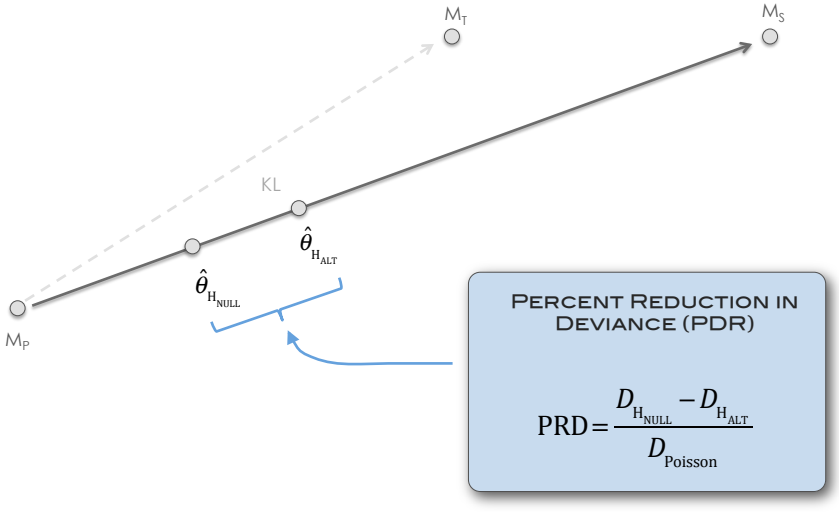


white: probability = 0

New Q matrix: M0 + DT

- **4** parameters (κ, ω, α, β)
- DT allowed (via α and β)

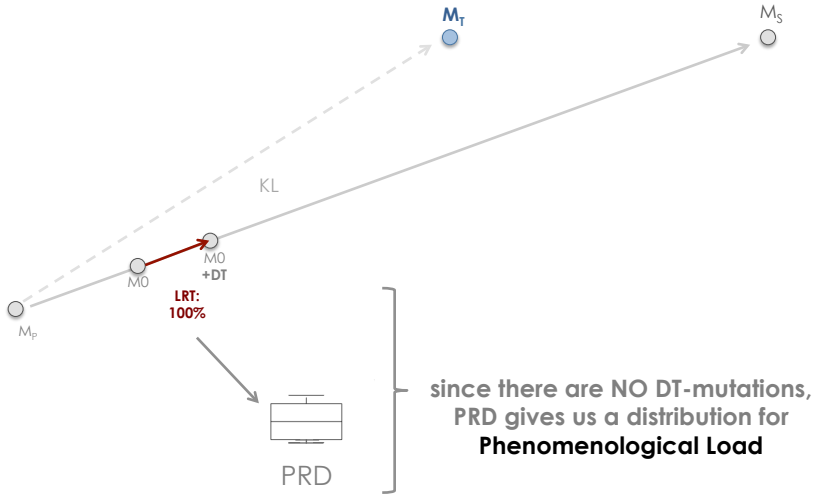
How to test such a model?



Percent Reduction in Deviance (PDR)

$$\text{PRD} = \frac{D_{H_{NULL}} - D_{H_{ALT}}}{D_{\text{Poisson}}}$$

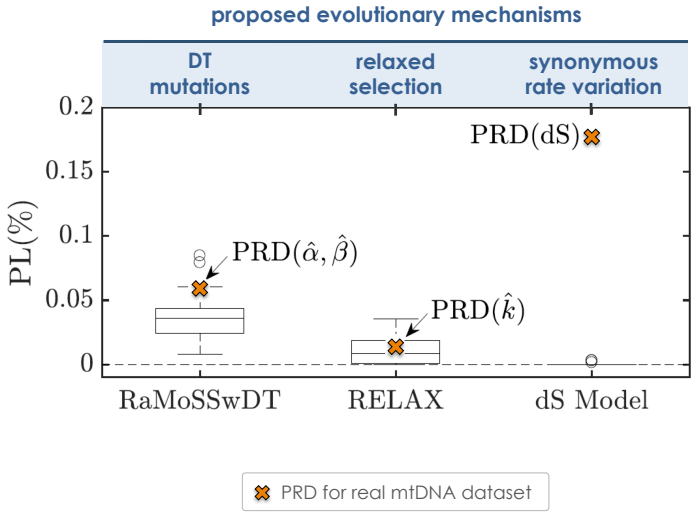
simulation for M_T:
MutSel with NO DT-mutations



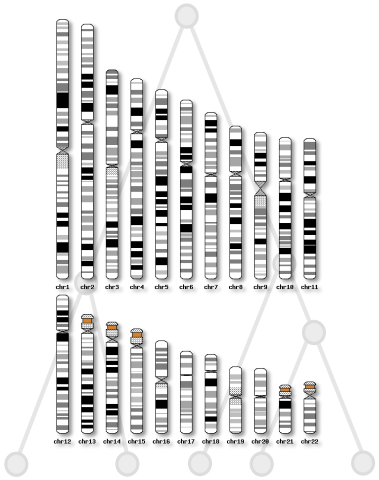
since there are NO DT-mutations,
PRD gives us a distribution for
Phenomenological Load

phenomenological load

testing PL on three proposed mechanisms for mtDNA

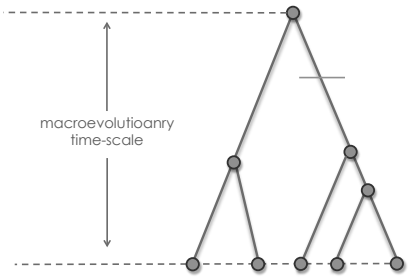


part 5: re-assessing long-held paradigms for evidence of adaptive evolution



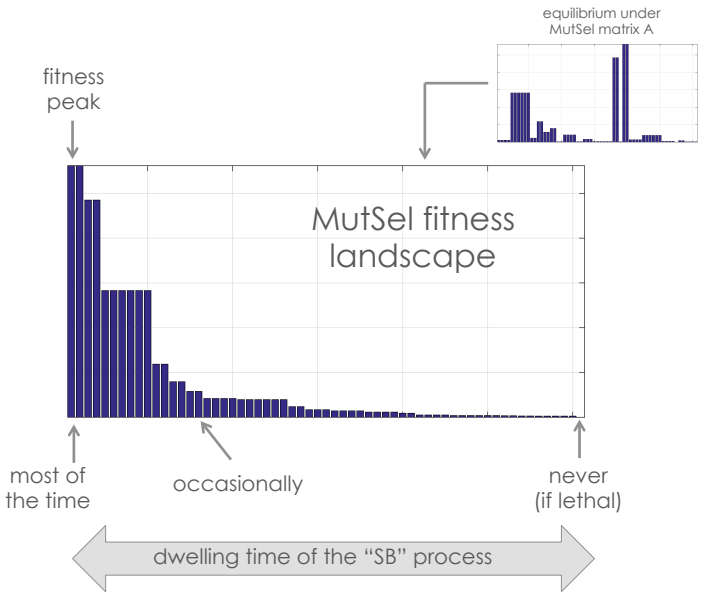
Three paradigms for "this" side:

1. codon substitution model: $d_N/d_S > 1$ is evidence of adaptive evolution
2. "mechanistic" substitution models are better
3. It's easy to test and predict model performance via simulation

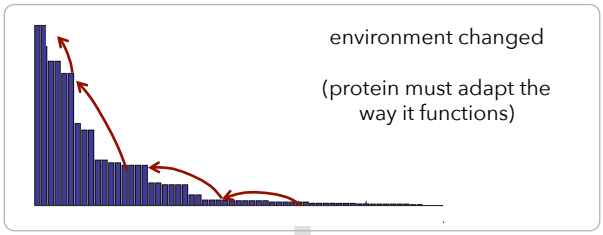


Paradigm 1: $d_N/d_S > 1$ is evidence of adaptive evolution of function

the MutSel fitness landscape



the MutSel fitness landscape: **adaptive evolution**



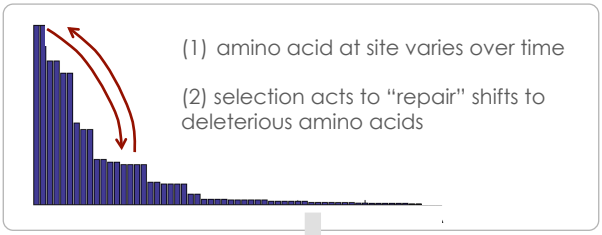
key result 1:

adaptive evolution: $p_+ > p_-$

("peak shift")

$d_N/d_S > 1$ (transient)

the MutSel fitness landscape: **non-adaptive shifting balance**



key result 2:

purifying selection: $p_+ = p_-$

(static landscape)

$d_N/d_S > 1$ (transient)

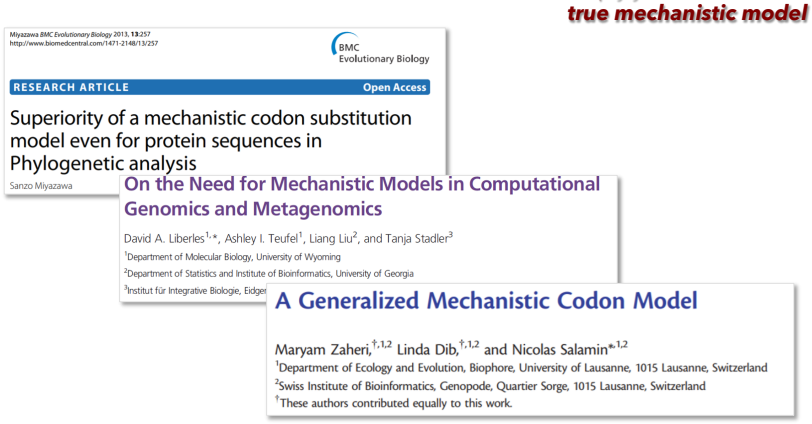
Paradigm 1 [illegible] adaptive [illegible]nction

Reality: $d_N/d_S > 1$ on a fixed landscape with **no change in function**

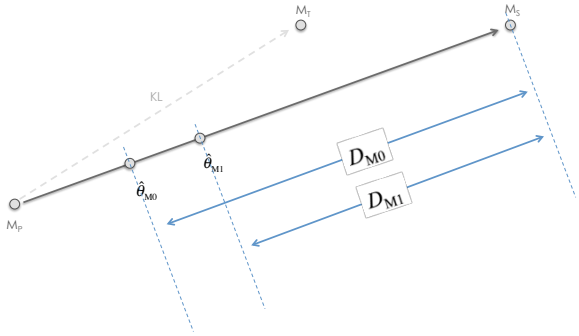
Proposal: develop new frameworks that do NOT depend on the $d_N/d_S > 1$ paradigm

Paradigm 2: "mechanistic" substitution models should be better

All imply you move closer to a
true mechanistic model



$$\mathrm{LRR} = D_{M0} - D_{M1} = -2\left\{\ell_{M1}\left(\hat{\theta}_{M1}|X,T\right) - \ell_{M2}\left(\hat{\theta}_{M2}|X,T\right)\right\}$$



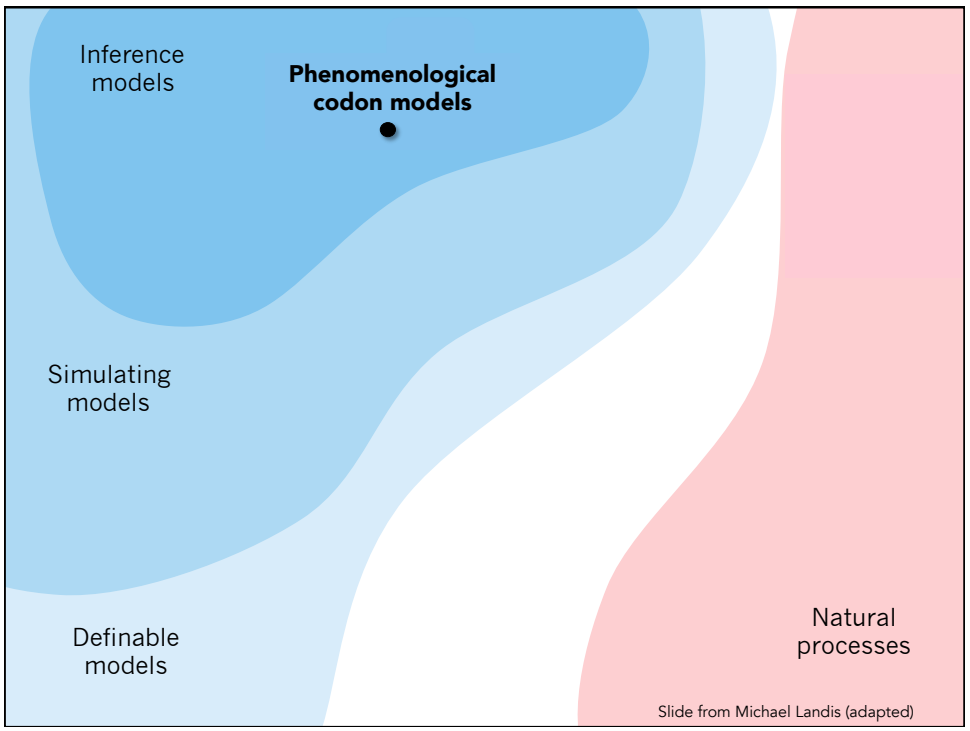
The **Likelihood Ratio Test (LRT)** "manages" phenomenological variability
(not mechanistic variability)

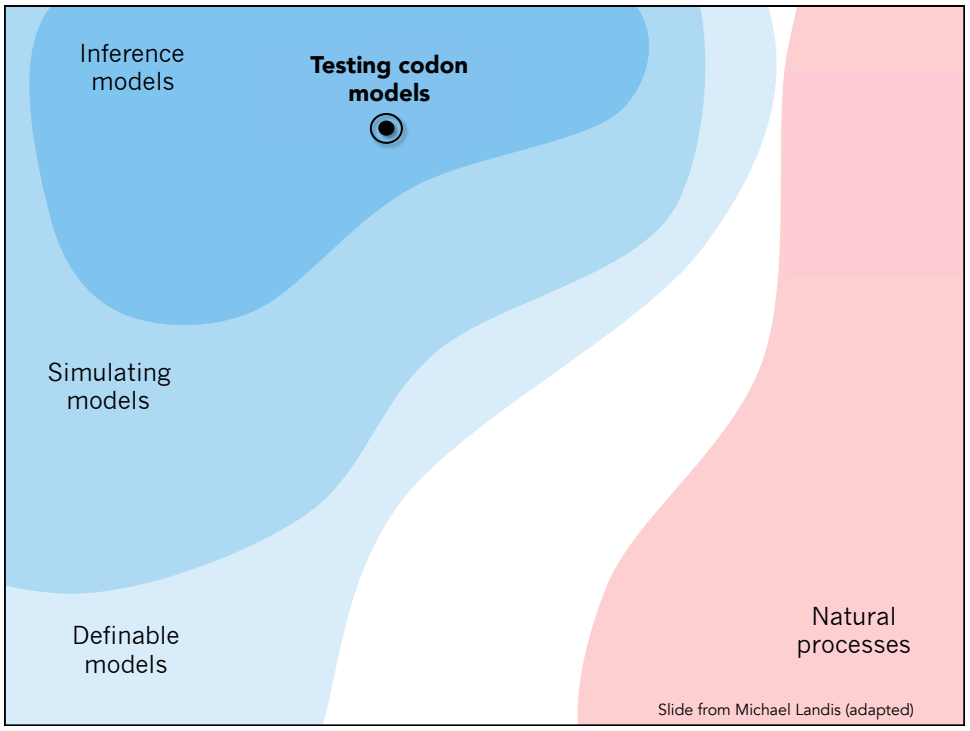
Paradigm 2: ...stitution models ...be better

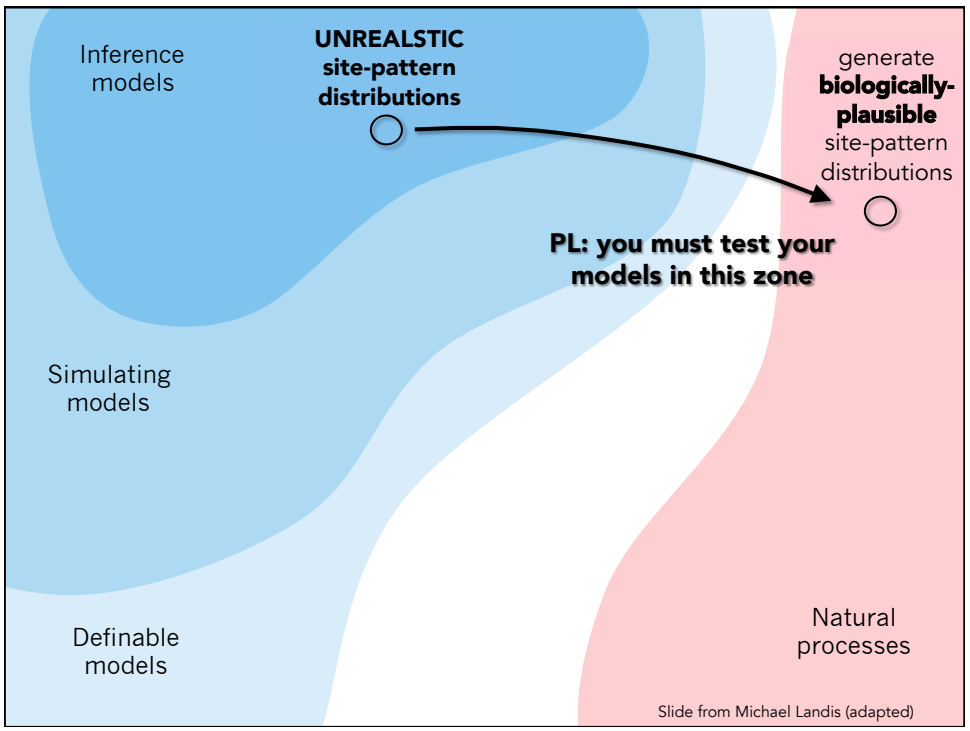
For real data: mechanistic parameters within models are expected to carry some **Phenomenological Load**

Proposal: intentionally add phenomenological parameters that improve inferences (e.g., covarion δ)

Paradigm 3: It's easy to test and predict model performance via simulation


Inference models
Phenomenological codon models
Simulating models
Definable models
Natural processes
Slide from Michael Landis (adapted)


Inference models
Testing codon models
Simulating models
Definable models
Natural processes
Slide from Michael Landis (adapted)



Paradigm 3: It's ...ct model performa... MYTH BUSTED

In reality it's hard to (1) compare complex site pattern distributions, and (2) identify models that produce biologically plausible distributions

Proposal: we need to do more work on how to generate "realistic" site pattern distributions and change the way we think about testing model performance

How can you really tell if you have learned anything relevant to the function of your protein?

- combine computational and **experimental approaches** (B. Chang, next lecture; "**Gold Standard**")
- informal cross-validation via comparison with **external phenotypic information** (B. Chang, next lecture)
- formally **include phenotypic information within the likelihood inference framework** (we have this working; the paper is in revision… "*stay tuned*")

THE END.