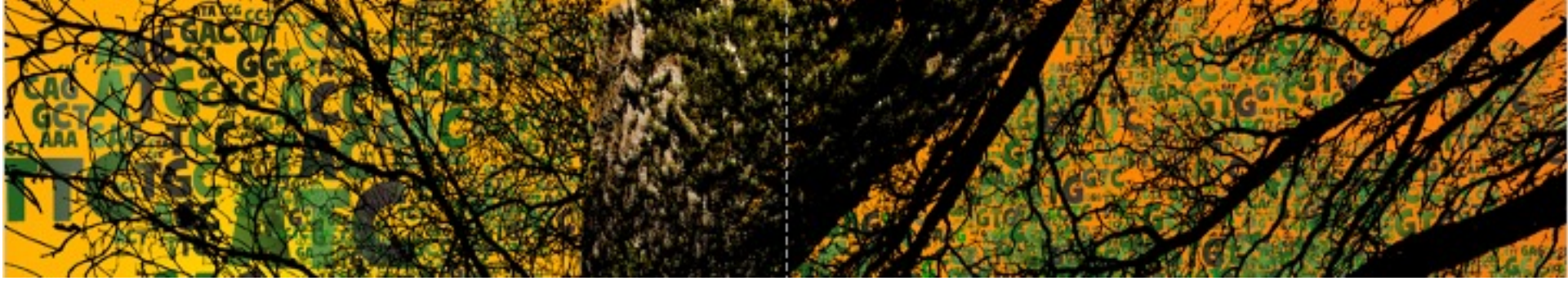


Detecting positive selection with codon models (2)

Maria Anisimova

Institute of Computational Life Sciences
Zurich University of Applied Sciences - ZHAW

Life Sciences and Facility Management

Institute of Computational Life Sciences

11.05.2022, Lund

Swiss Institute of Bioinformatics

Likelihood ratio test for positive selection

Two nested models:

Model 0 no positive selection

(H0: ω is always ≤1)

Model 1 allows positive selection

(H1: ω >1 for some sites or in certain lineages)

LRT statistic: $2\Delta\ell = 2(\ell_1 - \ell_0) \sim \chi^2_{d.f.}$

d.f. = difference in numbers of parameters

Modeling selection variability

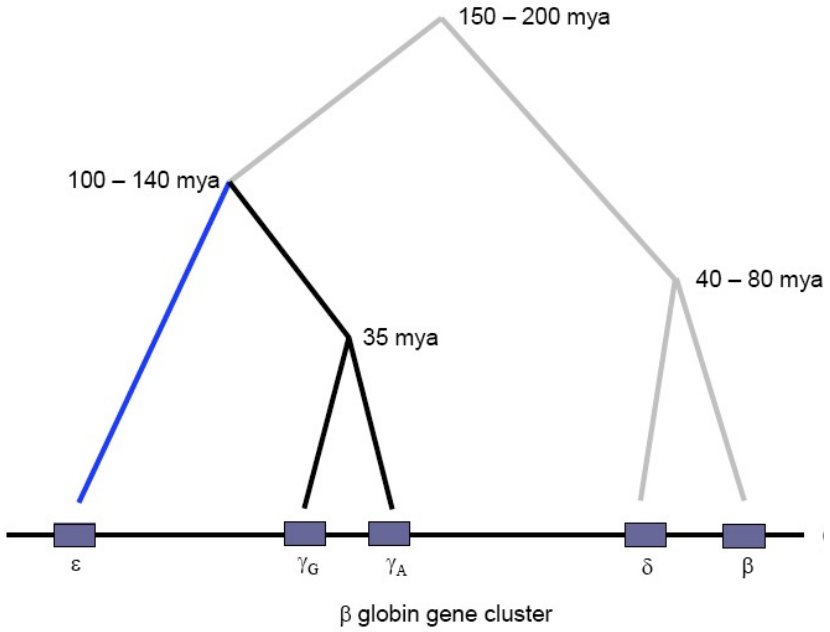
Assuming *constant selective pressure* across the whole sequence and over the whole phylogeny renders the *power of the test low*

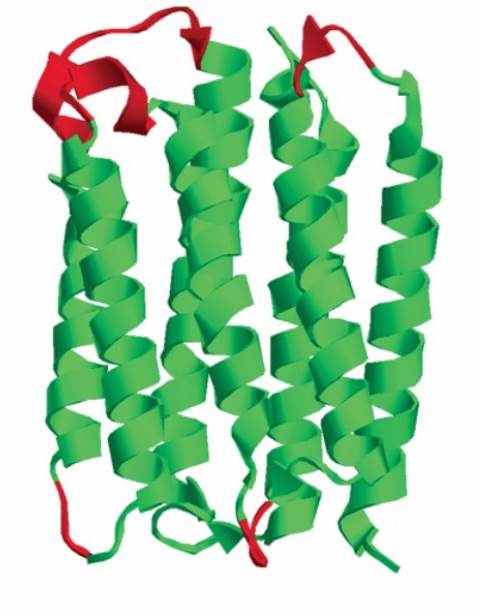
e.g., Endo et al (1996) detected only 17 out of 3595 analyzed genes to be under selection

Positive selection usually affects:

only in a few lineages/branches

only few codon sites





Modeling selection variability

By modeling variable ω over time and across sites we can study:

WHEN (in which lineages) did positive selection occur?

WHERE in the sequence did positive selection occur?

Modeling ω variability across sites

M-series models vary only by distributions used to model ω

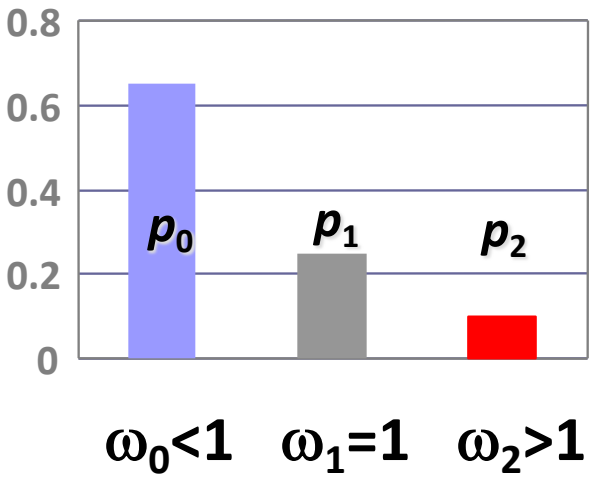
Yang et al. (2000), MBE

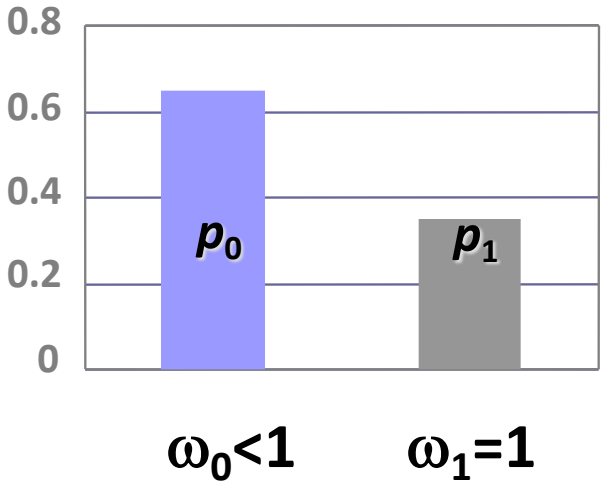
Model	Code	NP	Parameters
One-ratio	M0	1	ω
Neutral	M1a	2	$p_0, \omega_0,$
Selection	M2a	4	$p_0, p_1, \omega_0, \omega_2$
Discrete	M3	2K-1	$p_0, p_1, ..., p_{K-2}$ $\omega_0, \omega_1, \ldots, \omega_{K-1}$
Frequency	M4	5	$p_0, p_1, ..., p_4$
Gamma	M5	2	α, β
2Gamma	M6	4	$p_0, \alpha_0, \beta_0, \alpha_1$
Beta	M7	2	p, q
Beta&ω	M8	4	p_0, p, q, ω
Beta&gamma	M9	5	p_0, p, q, α, β
Beta&normal+1	M10	5	p_0, p, q, α, β
Beta&normal>1	M11	5	p_0, p, q, μ, σ
0&2normal>1	M12	5	$p_0, p_1, \mu_2, \sigma_1, \sigma_2$
3normal>0	M13	6	$p_0, p_1, \mu_2, \sigma_0, \sigma_1, \sigma_2$

It is hard to say what distribution shapes better reflects the data

Examples of nested site models

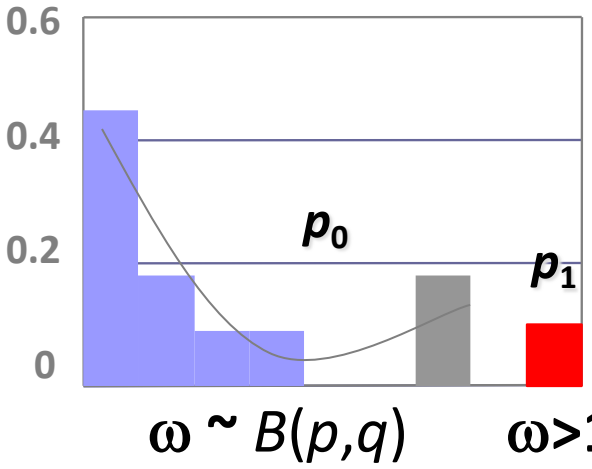
M2

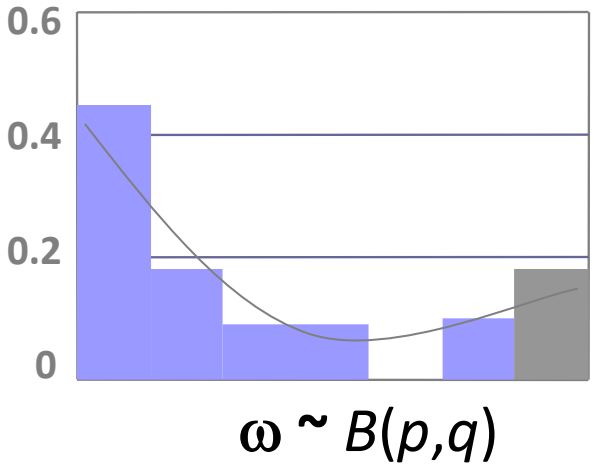




M1

M8

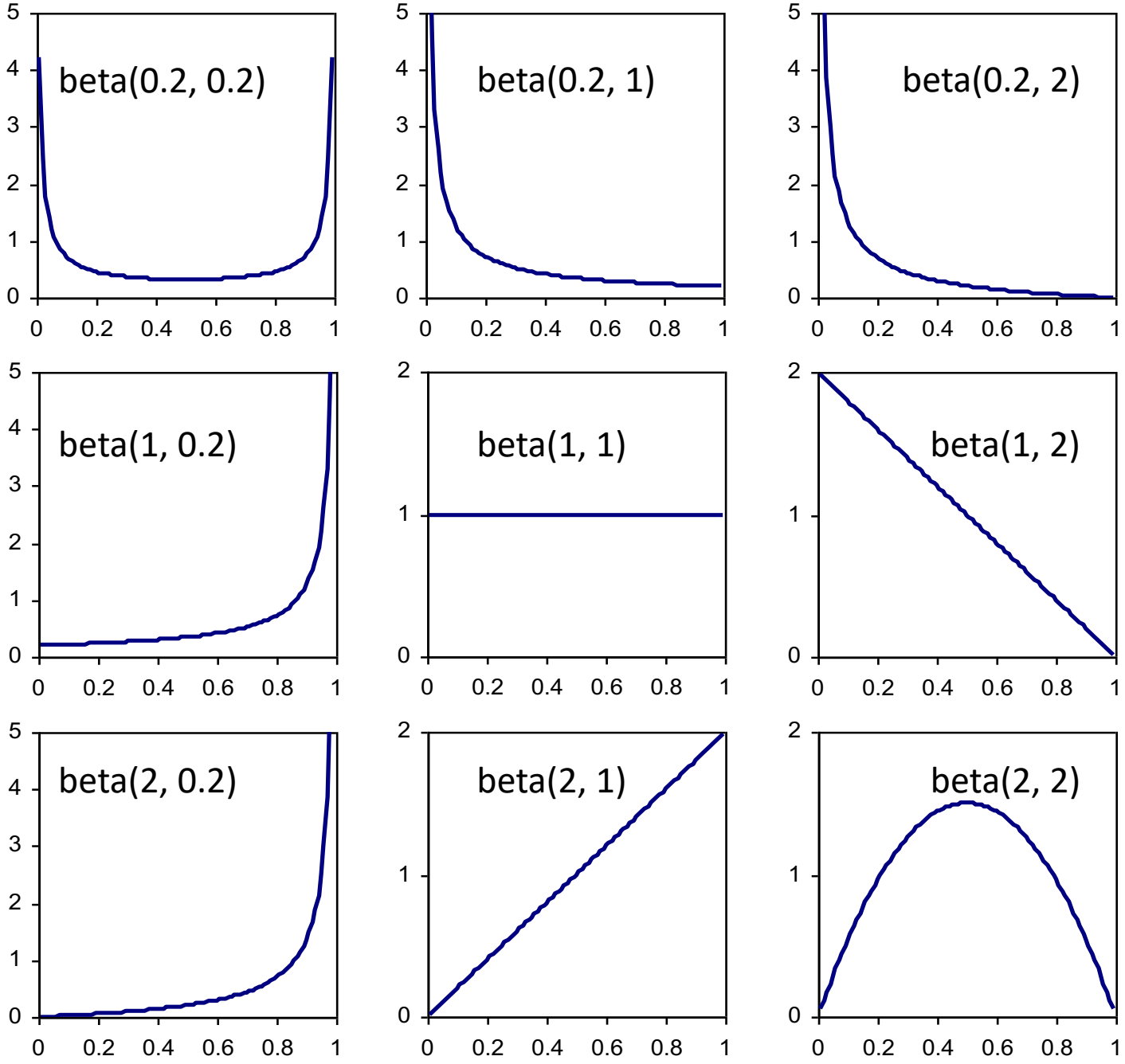




M7

Alternative **Null**

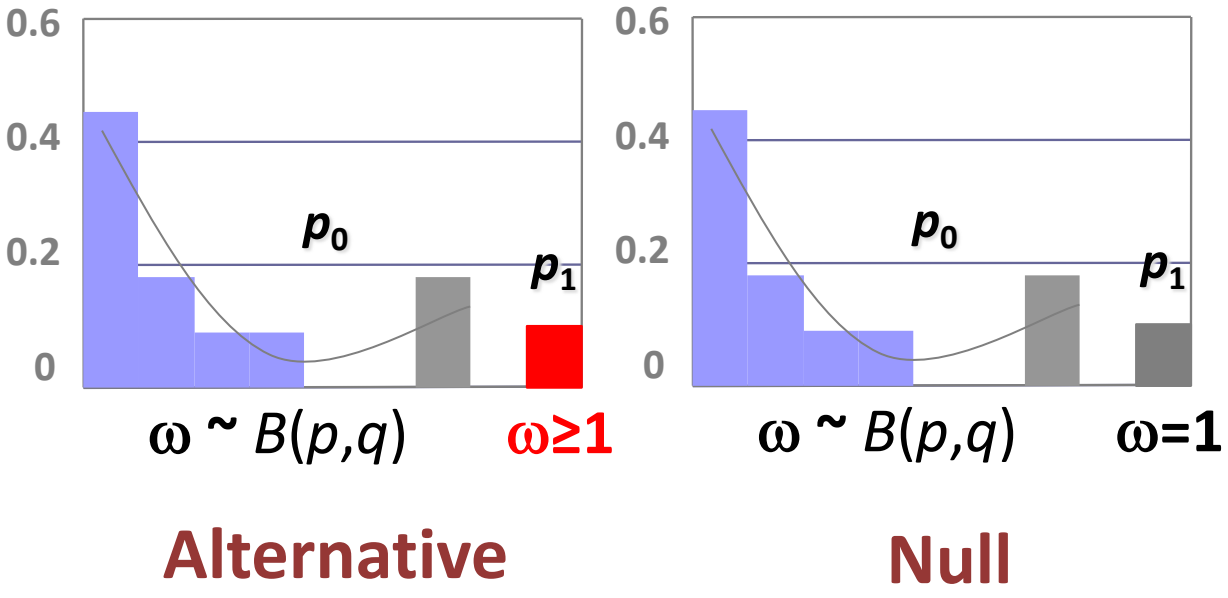
$0 \le B(p,q) \le 1$


beta(0.2, 0.2)
beta(0.2, 1)
beta(0.2, 2)
beta(1, 0.2)
beta(1, 1)
beta(1, 2)
beta(2, 0.2)
beta(2, 1)
beta(2, 2)

Examples of nested site models

A better defined LRT:
The null is 50:50 X^2mixture (with d.f. = 1 and 0)

M8



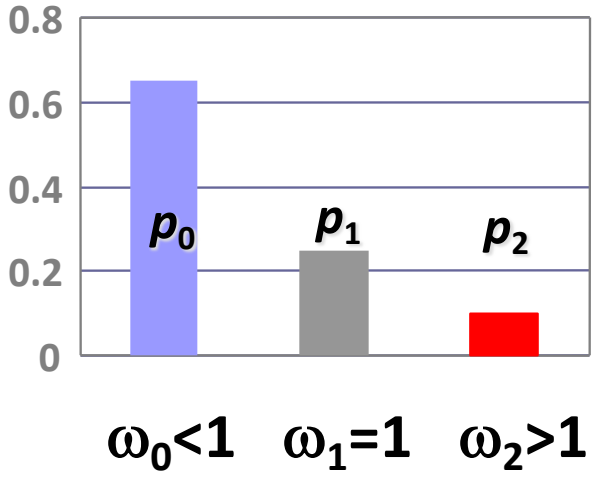
M8a

Alternative

Null

Examples of nested site-specific models

M2



Likelihood calculation should take into account
that a site may come from a number of different classes:

$$L_h = \Pr(\text{data}_{\text{site}}) = \sum_{class=1}^{K} \Pr(data_{site} \mid \omega_{site} = \omega_{class}) p_{class}$$

Example: Human MHC Class I data
192 alleles, 270 codons

Model	ℓ	Parameter estimates
M1a (neutral)	−7,490.99	$p_0 = 0.830$, $\omega_0 = 0.041$ $p_1 = 0.170$, $\omega_1 = 1$
M2a (selection)	−7,231.15	$p_0 = 0.776$, $\omega_0 = 0.058$ $p_1 = 0.140$, $\omega_1 = 1$ $p_2 = 0.084$, $\omega_2 = 5.389$

LRT of positive selection:

$2\Delta\ell = 2 \times 259.84 = 519.68$, $P < 0.000$ (d.f. = 2)

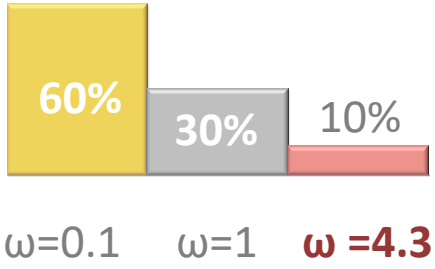
Yang and Swanson (2002) MBE

So far we used
models with variable selection
to test if selection affected the data

If LRT for positive selection is ***significant***
we can proceed inferring WHEN and WHERE…
(but this is more difficult)

Prediction of sites with Bayesian approach

ω site classes (GDD or M3):



For each site compute posterior probability:

$$P(\square \mid \begin{matrix}\text{CTC}\\ \text{TTA}\\ \text{TTG}\\ \text{TTA}\\ \text{CTG}\end{matrix}) = \frac{P(\begin{matrix}\text{CTC}\\ \text{TTA}\\ \text{TTG}\\ \text{TTA}\\ \text{CTG}\end{matrix} \mid \square)P(\square)}{P(\begin{matrix}\text{CTC}\\ \text{TTA}\\ \text{TTG}\\ \text{TTA}\\ \text{CTG}\end{matrix} \mid \square)P(\square) + P(\begin{matrix}\text{CTC}\\ \text{TTA}\\ \text{TTG}\\ \text{TTA}\\ \text{CTG}\end{matrix} \mid \square)P(\square) + P(\begin{matrix}\text{CTC}\\ \text{TTA}\\ \text{TTG}\\ \text{TTA}\\ \text{CTG}\end{matrix} \mid \square)P(\square)}$$

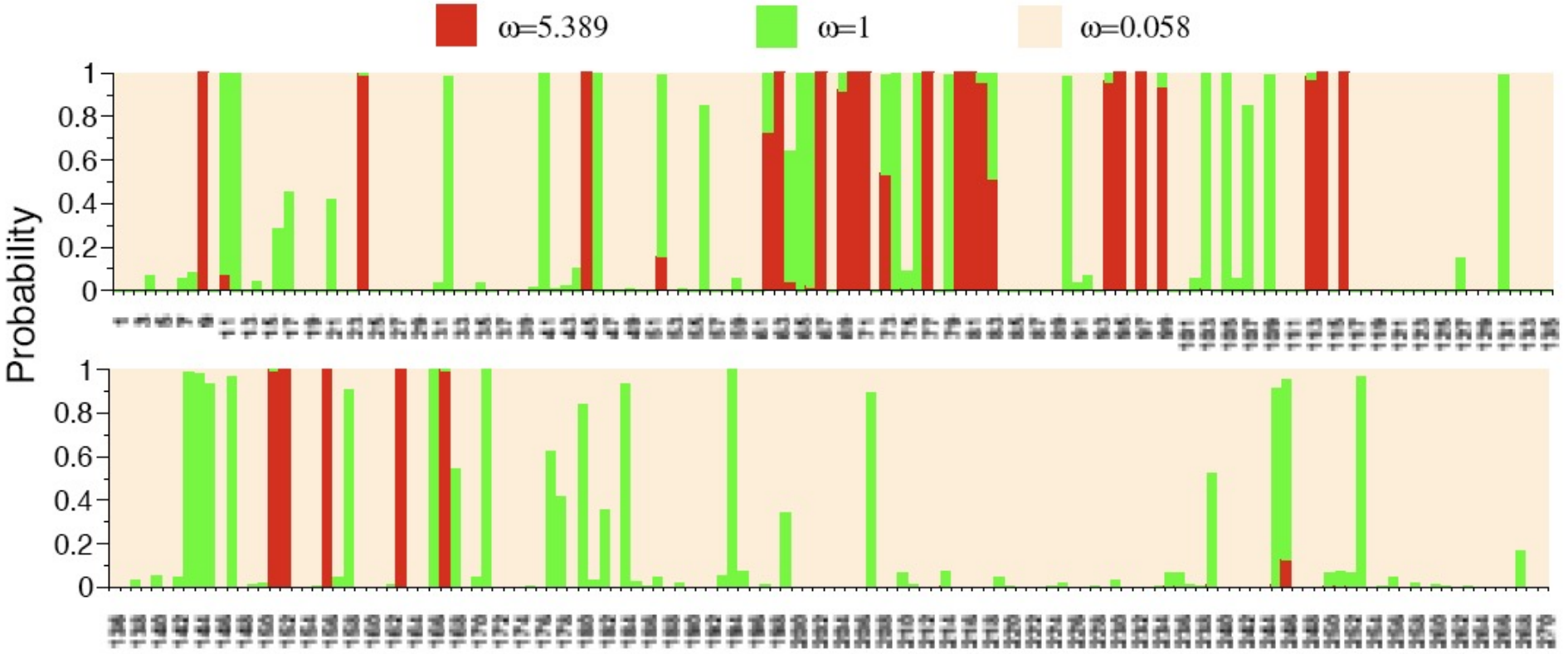
Sites with high posteriors (≥0.95)
may be inferred to be under positive selection

Empirical Bayesian calculation of posterior probabilities that a site is under positive selection with $\omega > 1$.

- Naïve Empirical Bayes (NEB) ignores sampling errors in parameter estimates.
- Bayes Empirical Bayes (BEB) accounts for sampling errors by integrating over a prior.

Nielsen & Yang. 1998 *Genetics* **148**
Yang, Wong & Nielsen 2005 *Mol Biol Evol* **22**

Posterior probabilities of ? for MHC (M2a)



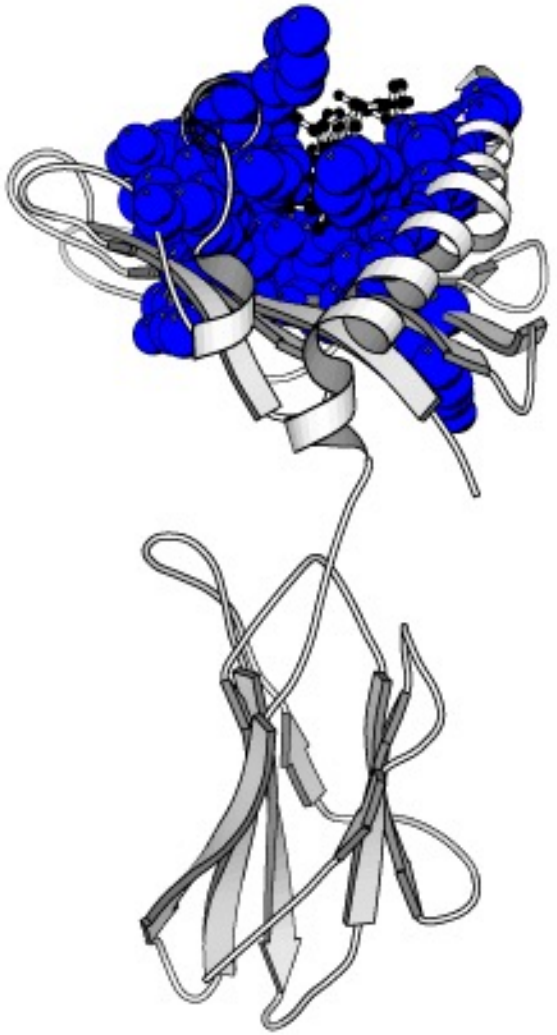
$$p(\omega_{\text{site}} = \omega_{\text{class}} \mid \text{data}_{\text{site}}) = \frac{p(\text{data}_{\text{site}} \mid \omega_{\text{class}})\, p_{\text{class}}}{\sum_{j=\text{site class}} p(\text{data}_{\text{site}} \mid \omega_j)\, p_j}$$

Human MHC Class I: 3D structure

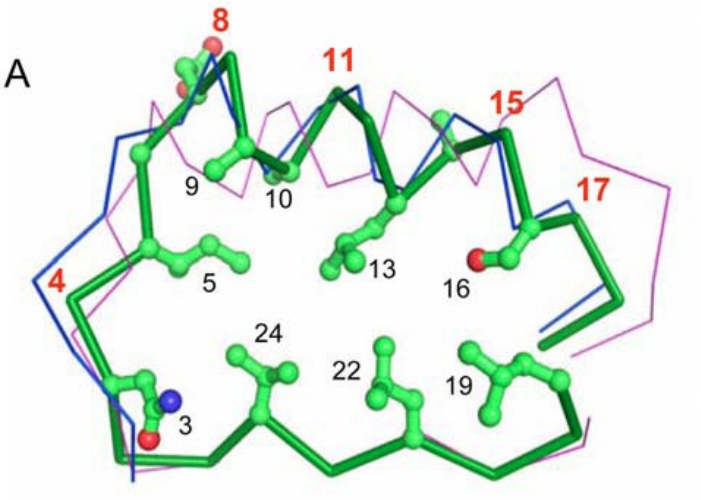
25 sites identified under M2a

All sites cluster together in the antigen recognition domain (blue)

Yang and Swanson (2002)



Positive selection in bacterial GALA



Bacterial GALA (type III effectors) acquired from host plants by LGT: residues under positive selection are found on the convex side of horse-shoe & involved in binding

Data from Kajava, Anisimova, Peeters (2008)

Figure 2. Structural model of GALA-LRR. (A) Cα-trace superposition of a modeled GALA-LRR and the known CC-LRR from human Skp2 protein [10] and RI-LRR from porcine ribonuclease inhibitor [46]. GALA-LRR model is shown in a ball-and-stick representation, CC-LRR is shown by a blue trace and RI-LRR by a magenta trace. Numbering of the conserved GALA-LRR residues is taken from Figure 1. Numbers in red point to positions inferred to be under positive selection. The carbon atoms are in green, oxygen in red, nitrogen in blue. (B) A ribbon diagram of a structural model of the C-terminal LRR domain of GALA4 type III effector protein from *R. solanacearum* (strain MolK2, region 170 to 460, accession code ZP_00946474). The figure was generated with Pymol [47]. The atomic coordinates of the model are available on request.

With more genomes sequenced, the approach of evolutionary comparison becomes more powerful.

It provides a way of generating interesting biological hypotheses, which can be validated by experimentation.

Ivarsson, Mackey, Edalat, Pearson, and Mannervik (2002) Identification of residues in glutathione transferase capable of driving functional diversification in evolution: **a novel approach to protein design**. *J. Biol. Chem.* 278:8733-8738.

Bielawski, Dunn, Sabehi, and Beja (2004) Darwinian **adaptation of proteorhodopsin to different light intensities** in the marine environment. *Proc. Natl. Acad. Sci. U.S.A.* 101:14824-14829.

Positive selection of primate *TRIM5α* identifies a critical species-specific retroviral restriction domain

Sara L. Sawyer*, Lily I. Wu†, Michael Emerman*†, and Harmit S. Malik*‡

Divisions of *Basic Sciences and †Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Communicated by Mark T. Groudine, Fred Hutchinson Cancer Research Center, Seattle, WA, December 29, 2004 (received for review December 8, 2004)

Primate genomes encode a variety of innate immune strategies to defend themselves against retroviruses. One of these, TRIM5α, can restrict diverse retroviruses in a species-specific manner. Thus, whereas rhesus TRIM5α can strongly restrict HIV-1, human TRIM5α only has weak HIV-1 restriction. The biology of TRIM5α restriction

genome defense predates the origin of primate lentiviruses (11, 12) and that many other *APOBEC* cytidine deaminase genes likely participate in defending the primate genome against retroviruses.

Here, we show that the *TRIM5α* restriction factor has

Rhesus TRIM5α restricts HIV-1 while human TRIM5α has only weak restriction.

Phylogenetic analysis detected a 13-aa patch with many positive-selected sites.

Functional studies of chimeric TRIM5α genes demonstrated that the patch was largely responsible for the difference in function. (Sawyer et al 2005)

Exercises with codeml

Focus of exercise #3:

ML estimation with site models