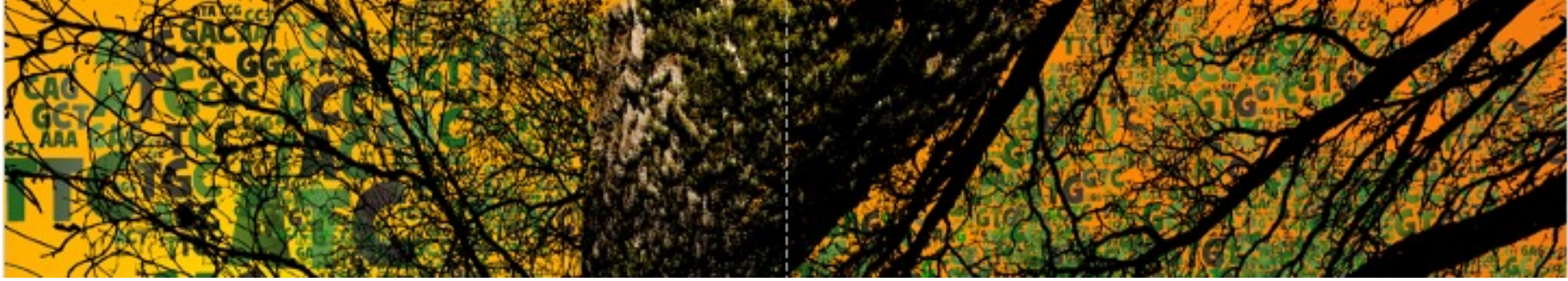


Detecting positive selection with branch-site codon models (3)

Maria Anisimova

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

Life Sciences and Facility Management

Institute of Computational Life Sciences

12.05.2022, Lund

Swiss Institute of Bioinformatics

Modeling selection variability

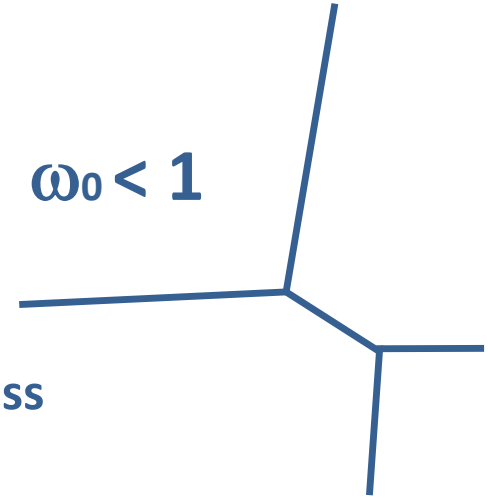
By modeling variable ω over time and across sites we can study:

WHEN (in which lineages) did positive selection occur?

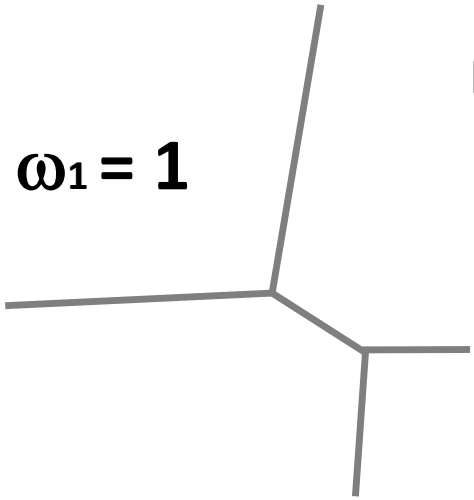
AND

WHERE in the sequence did positive selection occur?

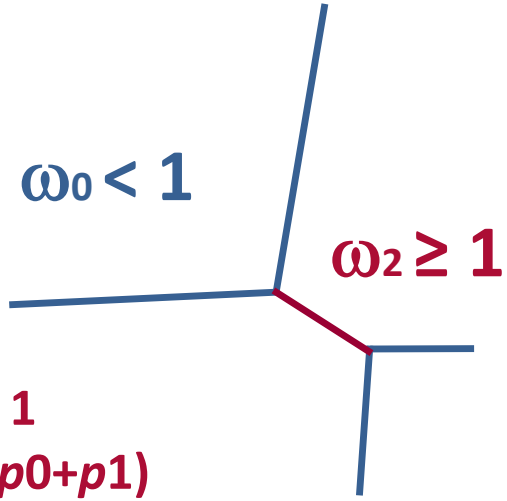
Branch-site codon model A (Yang et al 2005)



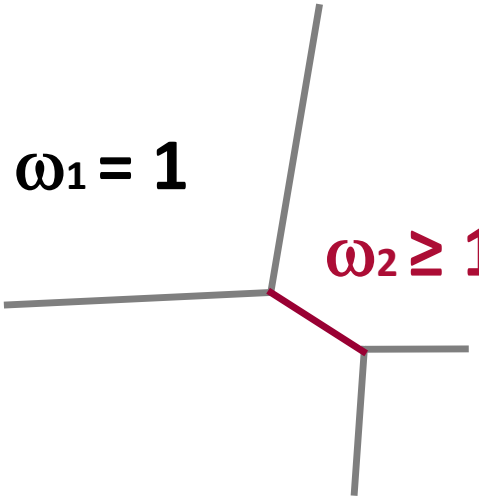
Conserved class
*p*0



Neutral class
*p*1



Variable class 1
*p*0(1-*p*0-*p*1)/(*p*0+*p*1)



Variable class 2
*p*1(1-*p*0-*p*1)/(*p*0+*p*1)

LRT for positive selection based on branch-site codon model

Null:
Model A
$\omega_2 = 1$ fixed

Alternative:
Model A
$\omega_2 \geq 1$ estimated

ℓ_0 ℓ_1

LRT statistic $2(\ell_0 - \ell_1) \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$

Foreground branches (with ω_2) are defined *a priori*

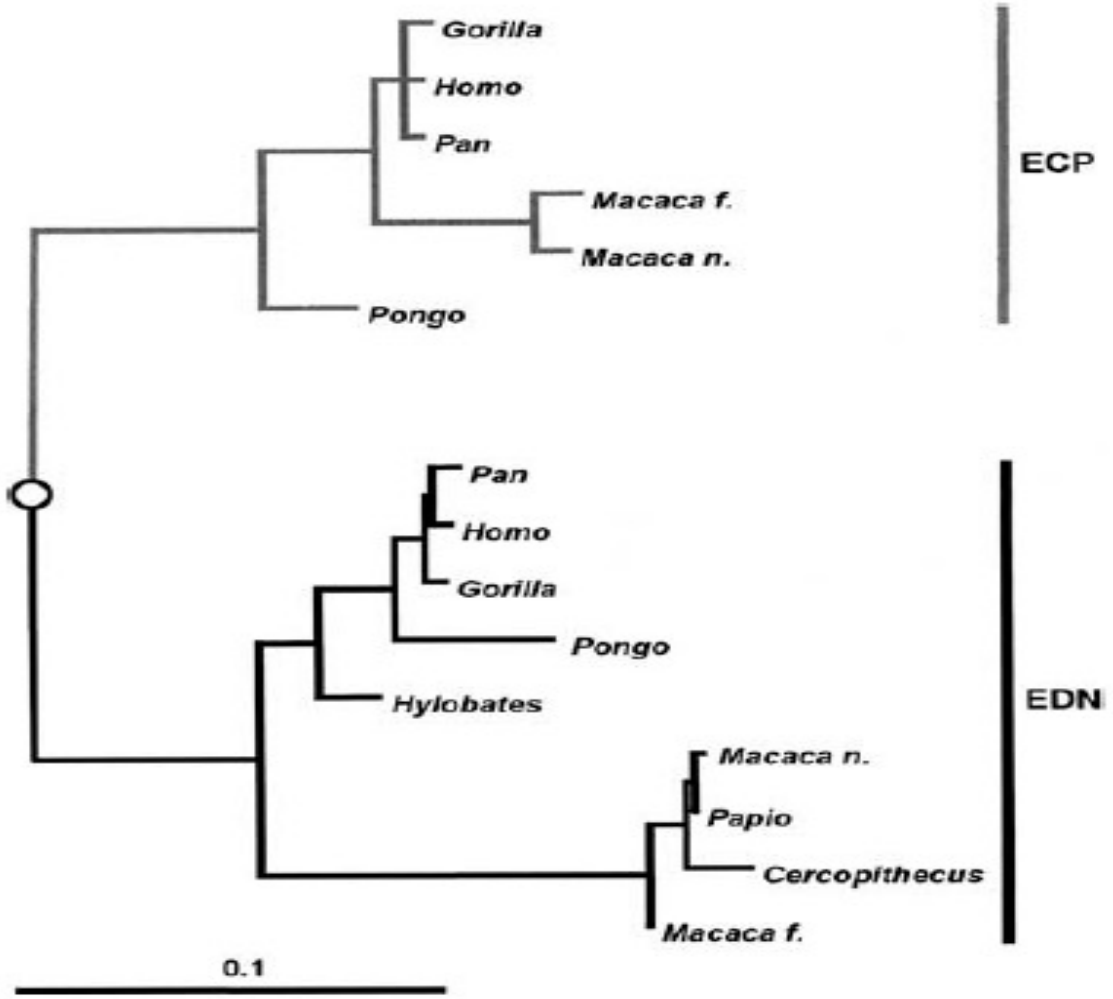


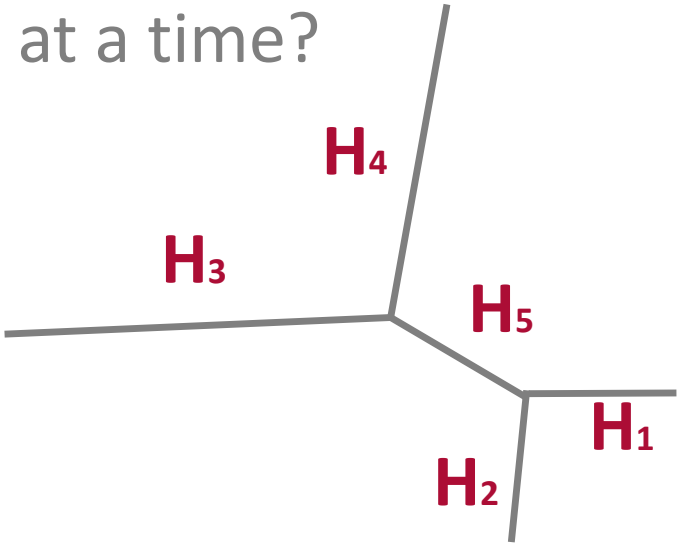
Fig. 3. Gene tree for 15 sequences from the ECP–EDN gene family. The topology was obtained by using maximum likelihood analysis under the HKY85 substitution matrix combined with a correction for among-site rate variation (discrete gamma model). The scale bar indicates the mean number of substitutions per nucleotide site. The open circle indicates the duplication event that gave rise to the ECP and EDN genes. Under Model D, a fraction of sites was allowed to evolve under divergent selection pressure, with ω_{1A} and ω_{1B} for the two paralogous clades, respectively.

Figure from Bielawski and Yang (2004)

To test for selection **after gene duplication:** branches of one clade following the duplication event are set as foreground

Testing multiple hypotheses

Test one branch at a time?



Anisimova & Yang (2007), MBE

Are p_1, p_2, p_3, p_4, p_5 significant at an overall threshold α?

Adjust individual thresholds α_1, α_2, α_3, α_4, α_5

so overall type I error rate $\leq \alpha$

Multiple testing correction: FWER or FDR?

Family-Wise Error Rate (FWER): overall type I error (FP rate)

FWER = Pr (reject at least one null when it's true)
For *n* independent true null hypotheses tested at α:
$$\text{FWER} = 1 - (1 - \alpha)^n$$

***e.g.* testing 10 hypotheses at 5% each we may get FWER=40%!**

If in some cases the null hypotheses is expected to be wrong, small percentage of false rejections is tolerable

FDR = False Discovery Rate
FDR = *E*(# false rejections/# all rejections)

Example: how do FWER and FDR compare

100 simulated datasets with first 6 null hypotheses true

For each sample, test 10 hypotheses, making 1 error per sample

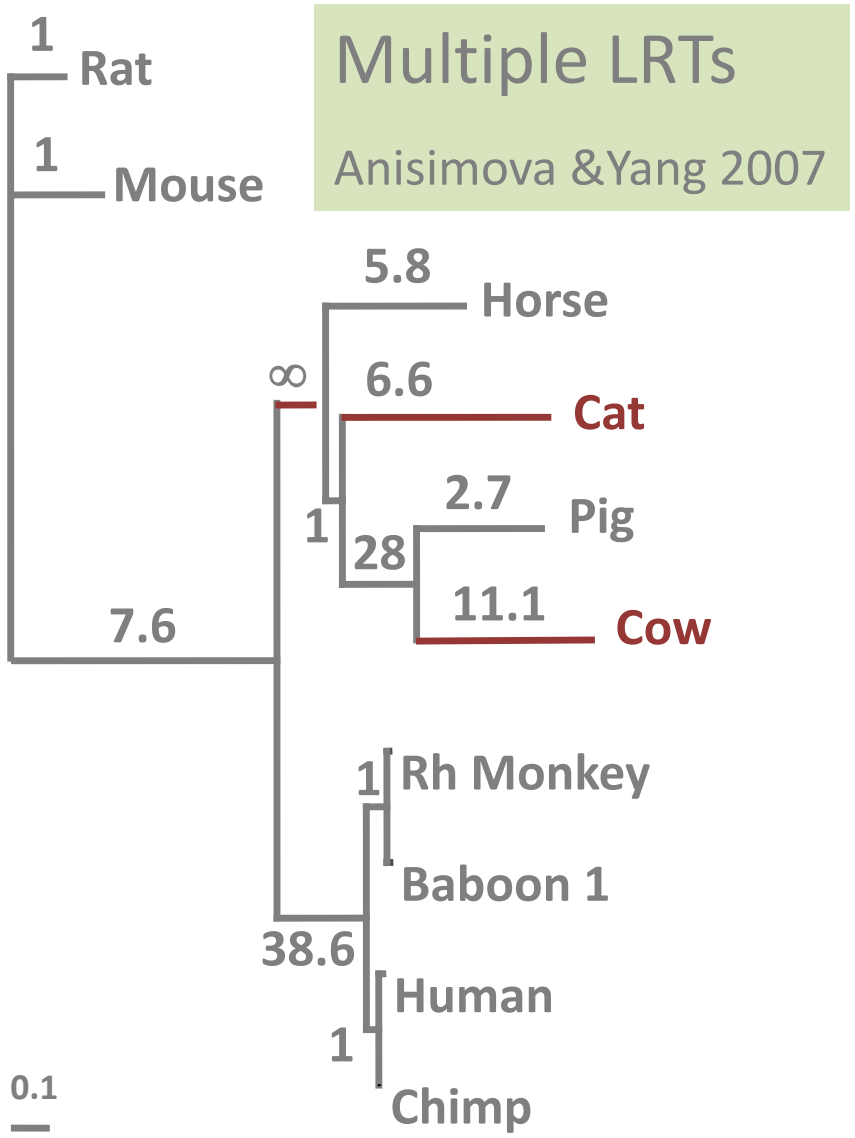
Test results: 1=sign / 0=not sign

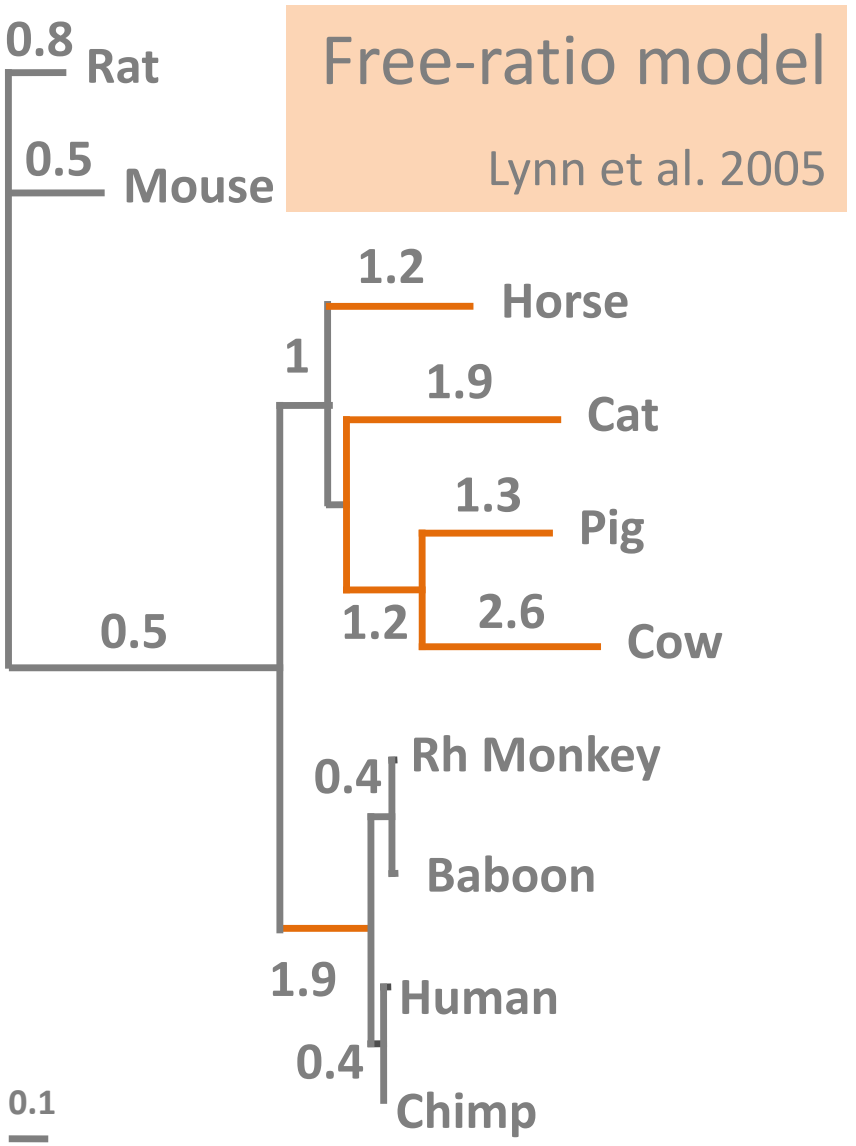
1	0	0	0	0	0	**1**	1	1	1	
2	0	**1**	0	0	0	0	1	1	1	
3	0	0	0	0	**1**	0	1	1	1	
...										
100	0	**1**	0	0	0	0	1	1	1	
	T	T	T	T	T	T	F	F	F	F

FDR = 20%

FWER = 100%

Multiple branch-site LRTs example: CD2 extra-cellular domain





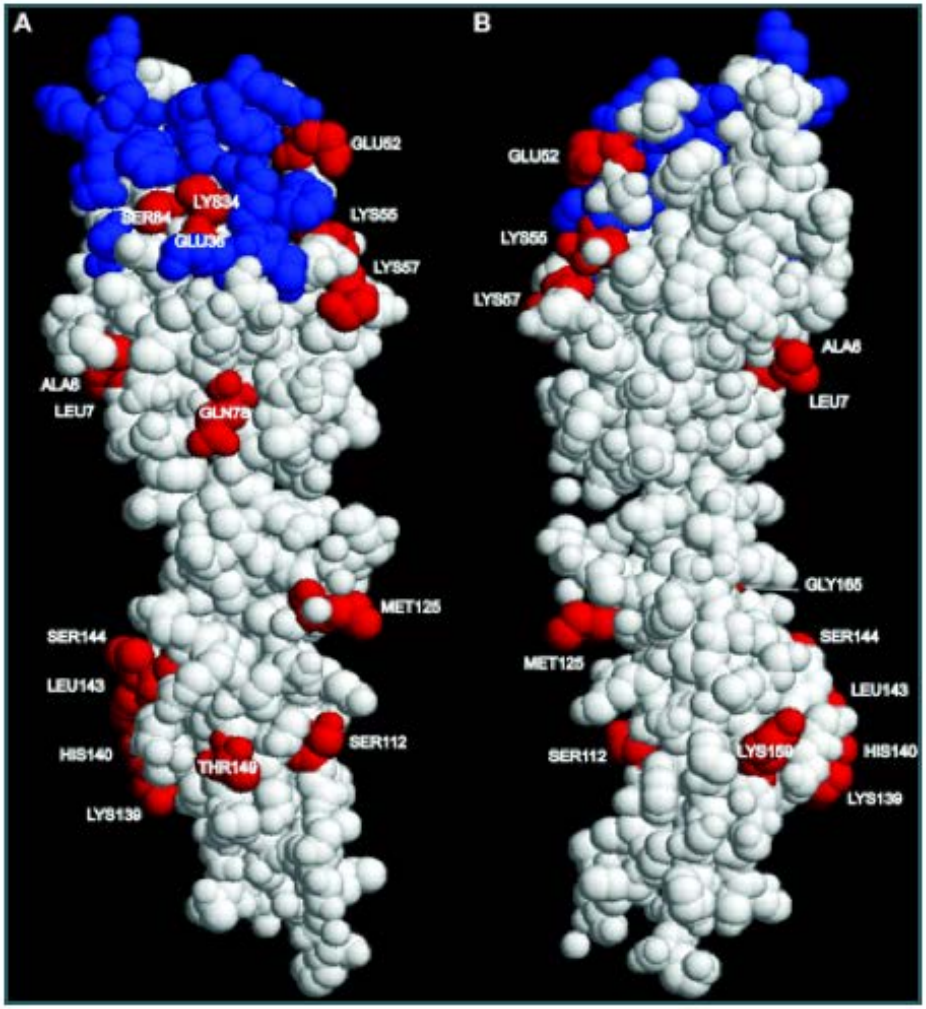
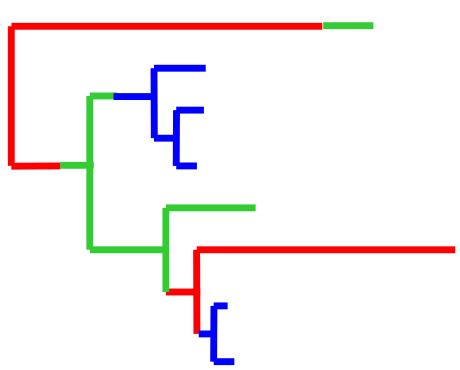


FIGURE 3.—

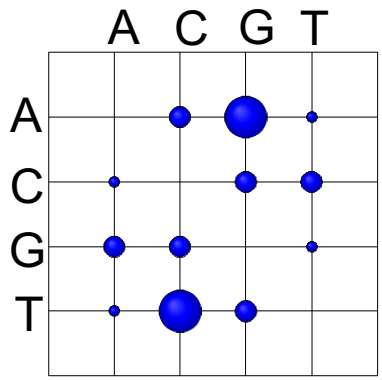
The three-dimensional structure of human CD2 extracellular domain [Protein Data Bank (PDB) http://www.rcsb.org/pdb/entry=1HNF]. Sites shown in red are those sites predicted to be under positive selection (model 8). The sites are labeled according to the numbering scheme used in the PDB file (ALA6 corresponds to site 14 in Table 1). Sites known to be involved in CD58 binding are shown in blue. A and B show two opposite faces of the CD2 molecule. The structure was displayed using RasMol V2.7.2.1.1 (http://www.openrasmol.org/software/rasmol/).

All but two sites under positive selection are found in the extra-cellular domain of CD2

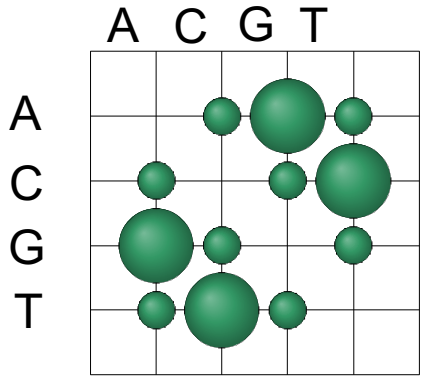
Alternatively, use covarion models



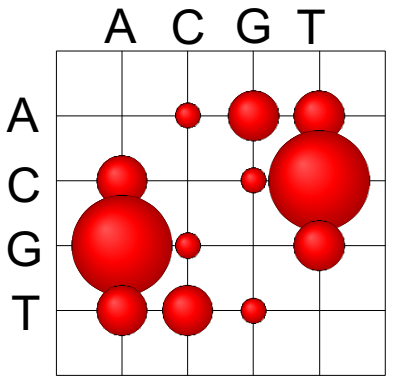
```
Seq1   TCTTTATTGACGTGTATGGACAATTC
Seq2   TCTTTGTTAACGTGCATGGACAATTC
Seq3   TCCTTGCTAACATGCATGGACAATTC
Seq4   TCTTTGCTAACGTGCATGGATAATTC
Seq5   TCTT---TAACGTGCATAGATAACTC
Seq6   TCAC---TAACATGTATAGATAACTC
Seq7   TCTCTTCTAACGTGCATTGTGAAGTC
Seq8   TCTCTTTTGACATGTATTGAAAAATC
```



Rate = 0.5

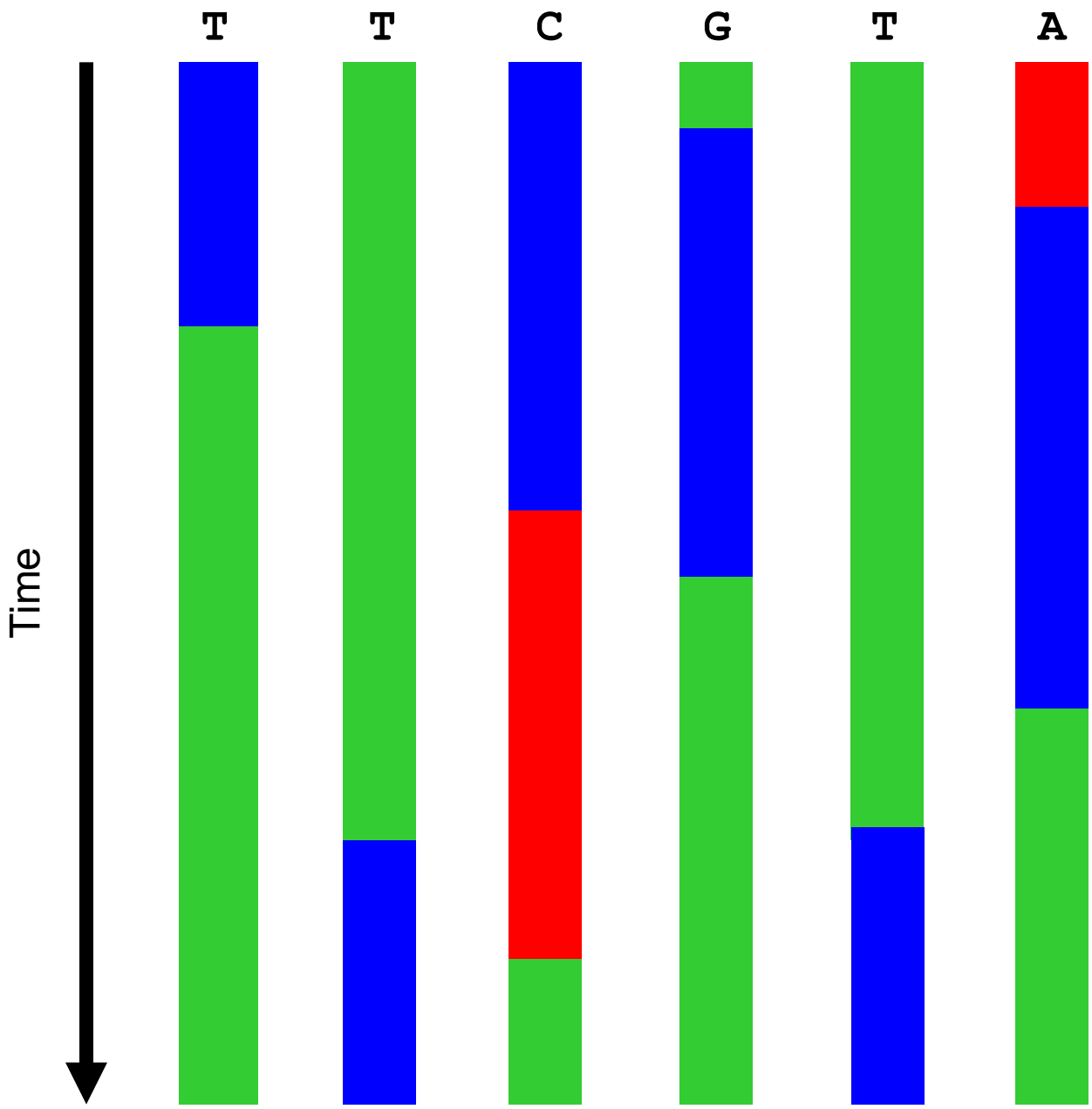


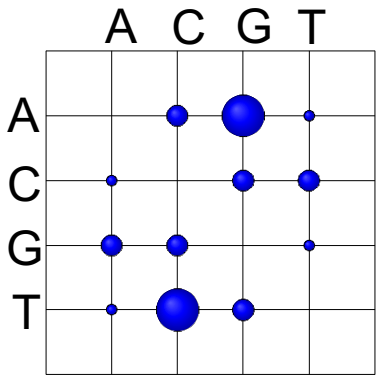
Rate = 1.0



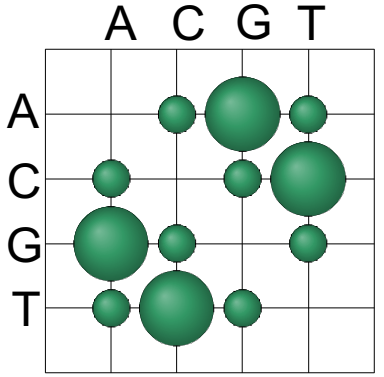
Rate = 2.0

Figure by Simon Whelan

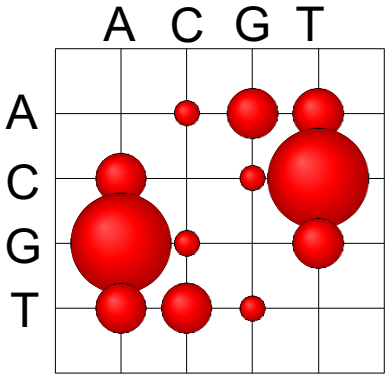




Rate = 0.5



Rate = 1.0



Rate = 2.0

Slide by Simon Whelan

Markov Modulated Codon Model

$$Q_x(ij) = \begin{cases} 0: \text{ if codons } i \text{ and } j \text{ differ at more than one nucleotide position} \\ \omega_x \pi_j: \text{ nonsynonymous transversion} \\ \pi_j: \text{ synonymous transversion} \\ \kappa\omega_x \pi_j: \text{ nonsynonymous transition} \\ \kappa\pi_j: \text{ synonymous transition} \end{cases}$$

Q_x describes instanteneous rates for sites from selection regime *x*
Codon models M2 and M3 are considered (each has 3 classes of sites)

Guindon et al. 2004 PNAS
https://github.com/stephaneguindon/fitmodel

$$\mathbf{R} = \delta \begin{pmatrix} -(p_2 + p_3\alpha) & p_2 & p_3\alpha \\ p_1 & -(p_1 + p_3\beta) & p_3\beta \\ p_1\alpha & p_2\beta & -(p_1\alpha + p_2\beta) \end{pmatrix}$$

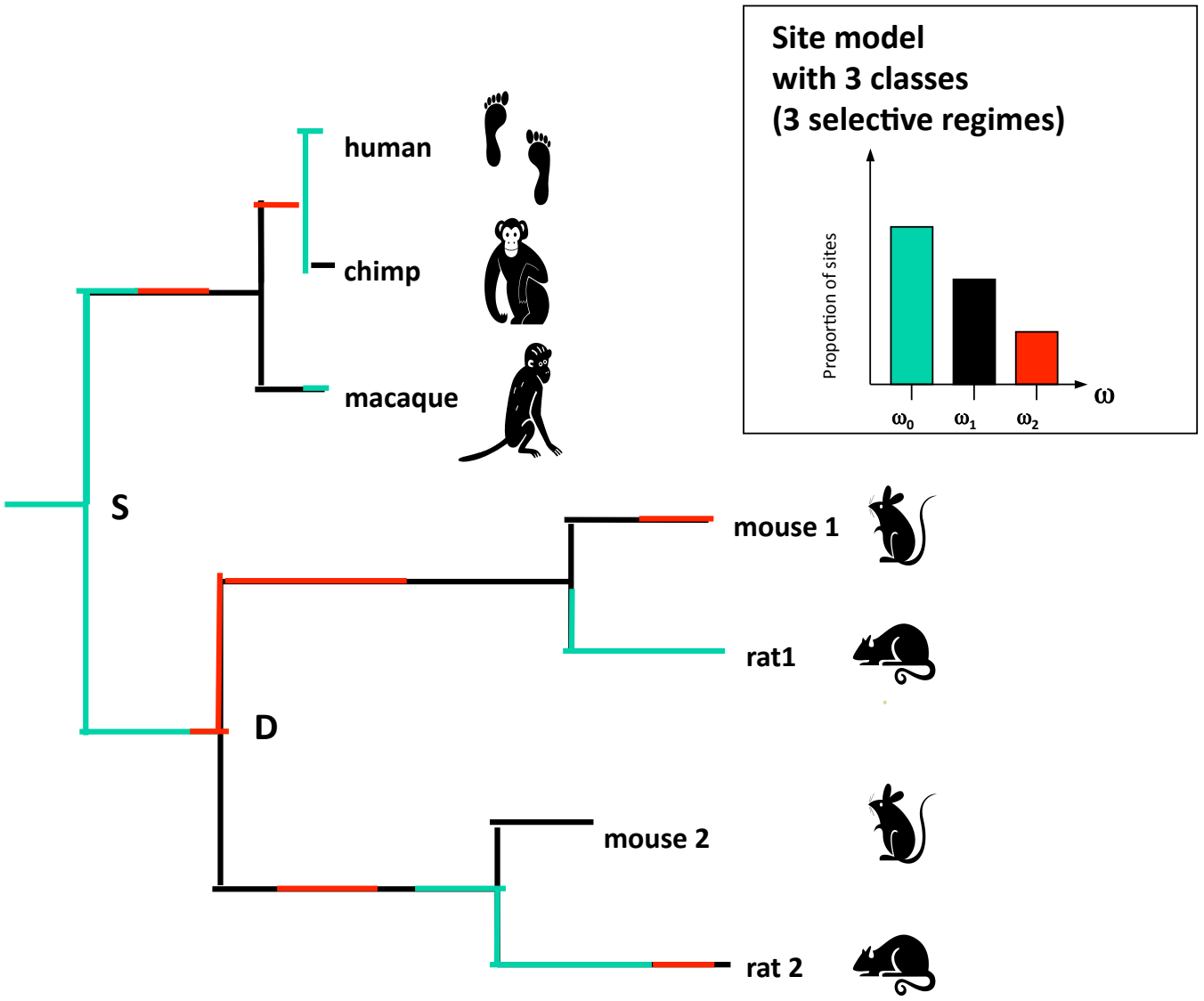
R describes rate switches between selection regimes 1, 2 and 3 ($\omega_1 < \omega_2 < \omega_3$)
p_1, p_2, p_3 are equilibrium frequencies of sites in each selection regime (add up to 1)
α is relative rate of changes between 1 and 3
β is relative rate of changes between 2 and 3

Combined process:

$$\mathbf{S} = \begin{pmatrix} \mathbf{Q}_1 & 0 & 0 \\ 0 & \mathbf{Q}_2 & 0 \\ 0 & 0 & \mathbf{Q}_3 \end{pmatrix} + \delta \begin{pmatrix} -(p_2 + p_3\alpha)\mathbf{I} & p_2\mathbf{I} & p_3\alpha\mathbf{I} \\ p_1\mathbf{I} & -(p_1 + p_3\beta)\mathbf{I} & p_3\beta\mathbf{I} \\ p_1\alpha\mathbf{I} & p_2\beta\mathbf{I} & -(p_1\alpha + p_2\beta)\mathbf{I} \end{pmatrix}$$

δ is the rate of switch between selection regimes

Markov Modulated Codon Model



LRTs of temporal variation in selection

H_0: $\delta = 0$ (no switches btw regimes or M3)
H_1: $\delta \neq 0$

H_0: $\delta = 0$ (no switches btw regimes)
H_1: $\beta = \alpha = 1$ (switching but **no bias in switching pattern)**

H_0: $\beta = \alpha = 1$ (no bias in switching pattern)
H_1: $\beta \neq \alpha$

Model notations: +S1 ($\beta = \alpha = 1$)
+S2 ($\beta = \alpha$ are free)

LRTs of temporal variation in selection

Guindon et el. 2004 PNAS

Significant at 5%

Table 1. Likelihood analysis of eight HIV-1 *env* gene sequence data sets

	M2	M2+S1	M2+S2	M3	M3+S1	M3+S2
P1						
ln*L*	−3,050.46	−3,021.78	−3,019.93	−3,036.87	−3,021.15	−3,019.13
$\omega_1\ \omega_2\ \omega_3$	0.00 1.00 8.31	0.00 1.00 9.40	0.00 1.00 10.01	0.15 1.22 7.50	0.04 0.91 8.62	0.04 0.71 9.43
$p_1\ p_2\ p_3$	0.39 0.56 0.04	0.67 0.29 0.05	0.64 0.32 0.05	0.70 0.26 0.03	0.69 0.26 0.05	0.60 0.35 0.05
P2						
ln*L*	−3,672.49	−3,652.61	−3,651.67	−3,658.85	−3,652.30	−3,651.23
$\omega_1\ \omega_2\ \omega_3$	0.00 1.00 4.39	0.00 1.00 3.86	0.00 1.00 4.47	0.15 1.14 3.85	0.06 1.36 4.23	0.03 0.49 3.98
$p_1\ p_2\ p_3$	0.30 0.62 0.07	0.57 0.33 0.10	0.55 0.38 0.08	0.58 0.37 0.06	0.65 0.28 0.07	0.46 0.42 0.13
P3						
ln*L*	−3,205.90	−3,171.99	−3,169.07	−3,184.05	−3,165.13	−3,162.90
$\omega_1\ \omega_2\ \omega_3$	0.00 1.00 5.20	0.00 1.00 5.07	0.00 1.00 14.17	0.19 2.10 5.95	0.00 2.92 9.99	0.00 2.83 13.82
$p_1\ p_2\ p_3$	0.36 0.49 0.15	0.71 0.15 0.14	0.75 0.20 0.05	0.73 0.22 0.05	0.78 0.18 0.03	0.79 0.19 0.02
P5						
ln*L*	−3,889.82	−3,819.30	−3,817.56	−3,838.40	−3,816.79	−3,815.98
$\omega_1\ \omega_2\ \omega_3$	0.00 1.00 11.88	0.00 1.00 10.01	0.00 1.00 10.44	0.14 1.04 7.34	0.05 1.71 11.51	0.05 1.39 10.80
$p_1\ p_2\ p_3$	0.35 0.62 0.04	0.73 0.23 0.03	0.71 0.26 0.03	0.77 0.20 0.04	0.84 0.14 0.02	0.79 0.18 0.03
P7						
ln*L*	−4,121.97	−4,060.46	−4,057.37	−4,084.47	−4,050.26	−4,049.37
$\omega_1\ \omega_2\ \omega_3$	0.00 1.00 8.40	0.00 1.00 11.61	0.00 1.00 11.81	0.32 2.70 11.84	0.19 3.29 14.56	0.17 3.07 15.09
$p_1\ p_2\ p_3$	0.25 0.63 0.12	0.61 0.32 0.07	0.58 0.35 0.07	0.79 0.17 0.04	0.83 0.13 0.04	0.81 0.14 0.05
P8						
ln*L*	−4,174.14	−4,098.80	−4,092.67	−4,136.79	−4,095.89	−4,090.22
$\omega_1\ \omega_2\ \omega_3$	0.00 1.00 5.34	0.00 1.00 9.20	0.00 1.00 15.05	0.10 1.03 4.17	0.03 1.41 9.93	0.05 1.06 14.85
$p_1\ p_2\ p_3$	0.38 0.53 0.09	0.68 0.27 0.05	0.68 0.29 0.03	0.64 0.28 0.07	0.74 0.22 0.04	0.71 0.26 0.03

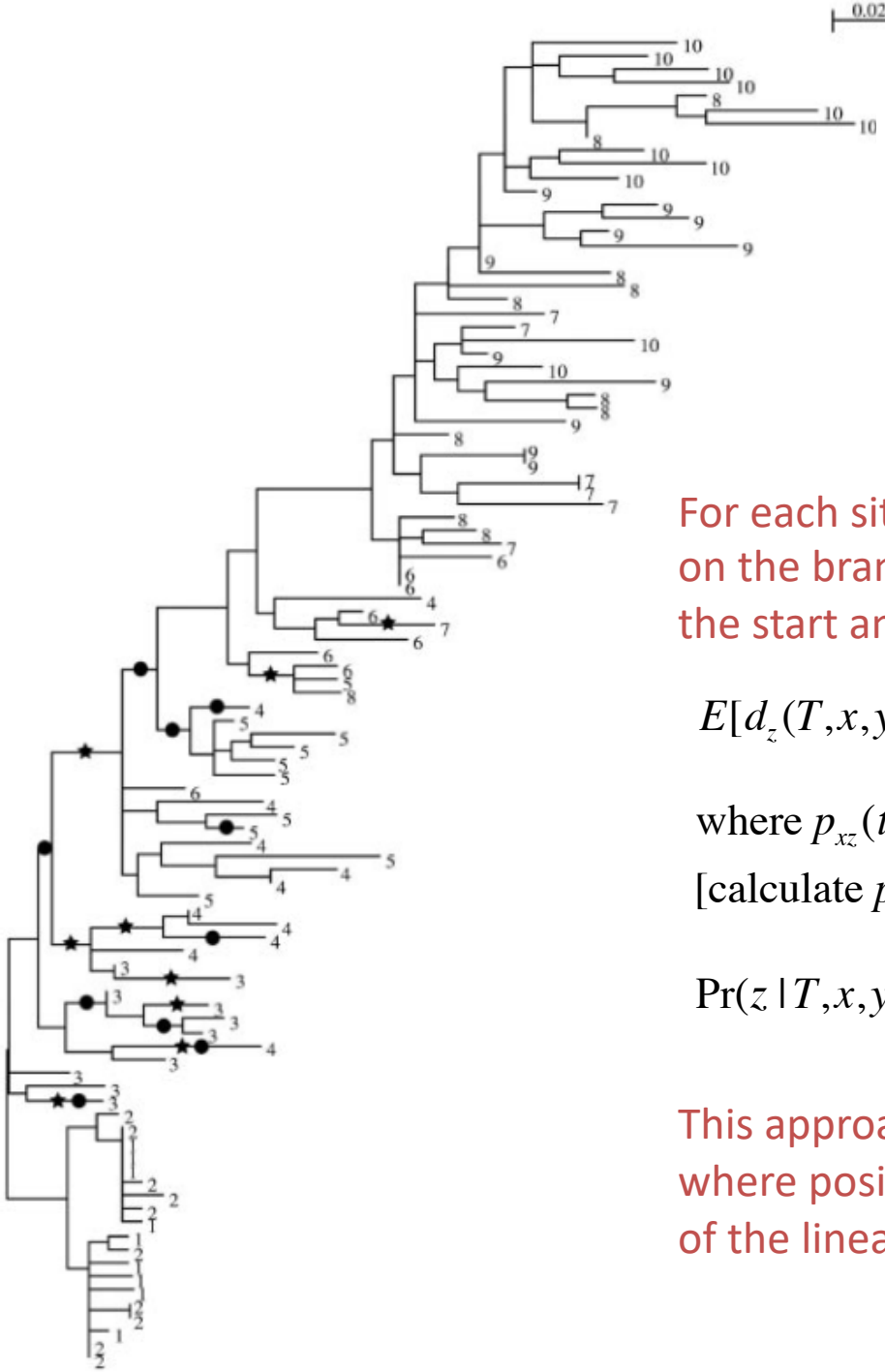


Fig. 1. Phylogenetic positions of substitutions inferred at two amino acid sites of patient 6 data set. M3 strongly supports the hypothesis that sequences evolved under positive selection at these sites, whereas the statistical support given by M3+S1 to the same hypothesis is less important. ★ and ● correspond to the substitutions inferred at sites 41 and 180, respectively. All of these substitutions are likely to be nonsynonymous. The leaves of the tree are labeled with the rank of the corresponding sample time (1 is the earliest sample and 10 is the latest). The position of the root was determined by using outgroup sequences collected during the earliest stages of the infection.

For each site, the expected time spent in selection class *z* on the branch of length *T*, which had selection regime *x* at the start and *y* at the end:

$$E[d_z(T,x,y)] = \int_0^T \frac{p_{xz}(t)\,p_{zy}(T-t)}{p_{xy}(T)}\,dt$$

where $p_{xz}(t)$ is the probability of change $x \rightarrow y$ over time t

[calculate $p_{xz}(t)$ from $P_R(t) = \exp(tR)$]

$$\Pr(z \mid T,x,y) = E[d_z(T,x,y)]/T$$

This approach is used to detect sites in the alignment where positive selection is likely to have occurred in most of the lineages

Guindon et el. 2004 PNAS

Exercises with PAML (codeml)

Focus of exercise #4:

1. ML estimation with branch-site models
2. Optional: Try out with codon tree (CodonPhyML)