University of Oxford, Department of Zoology
Evolutionary Biology Group

Department of Zoology
University of Oxford
South Parks Road
Oxford OX1 3PS, U.K.
Fax: +44 1865 271249

Evolutionary Analysis of Viral Genomes

Lecture 4: Natural Selection & Viral Adaptation

Oliver G. Pybus

Department of Zoology, University of Oxford

MOLECULAR SEQUENCES

Alignment Methods

BIOINFORMATICS

ALIGNMENT

Sequence Evolution Models

MOLECULAR EVOLUTION

GENETIC DISTANCES

Phylogenetic Inference

PHYLOGENETICS

EVOLUTIONARY TREE

(time scale = genetic distance)

Molecular Clock Models

PHYLOGENETICS

EVOLUTIONARY TREE

(time scale = years)

??????????

POPULATION GENETICS

POPULATION PROCESSES

(e.g. adaptation, migration, population size change)

Natural Selection

- "This preservation of favourable variations and the rejection of injurious variations, I call natural selection. Variations neither useful nor injurious would not be affected by natural selection, and would be a left a fluctuating element" (Darwin, On The Origin of Species).

- The **selection coefficient** (s) of a mutation measures the degree by which it increases or decreases the fitness of an organism.

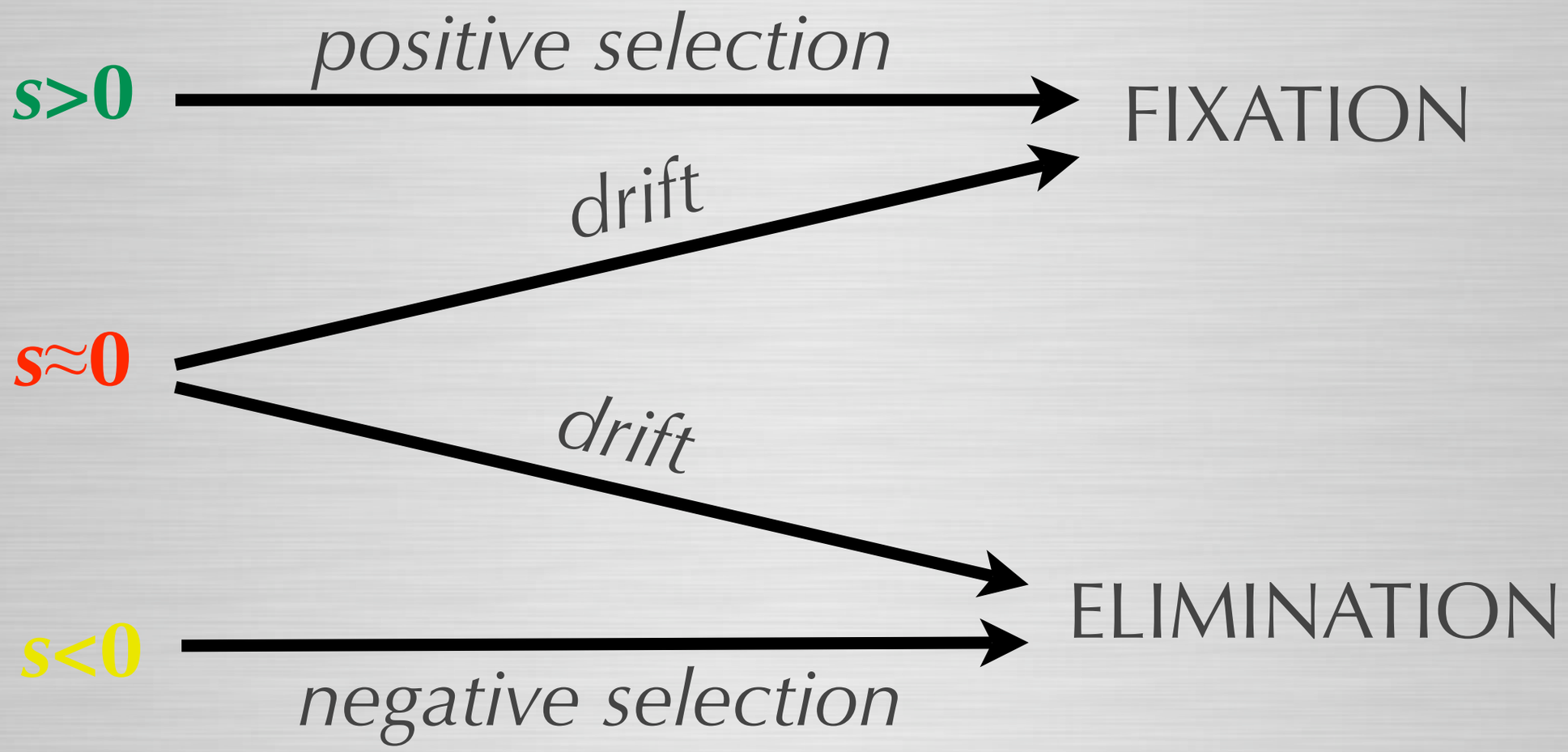
- Mutations are classified as:

  advantageous / benficial ($s>0$)

  disadvantageous / deleterious ($s<0$)

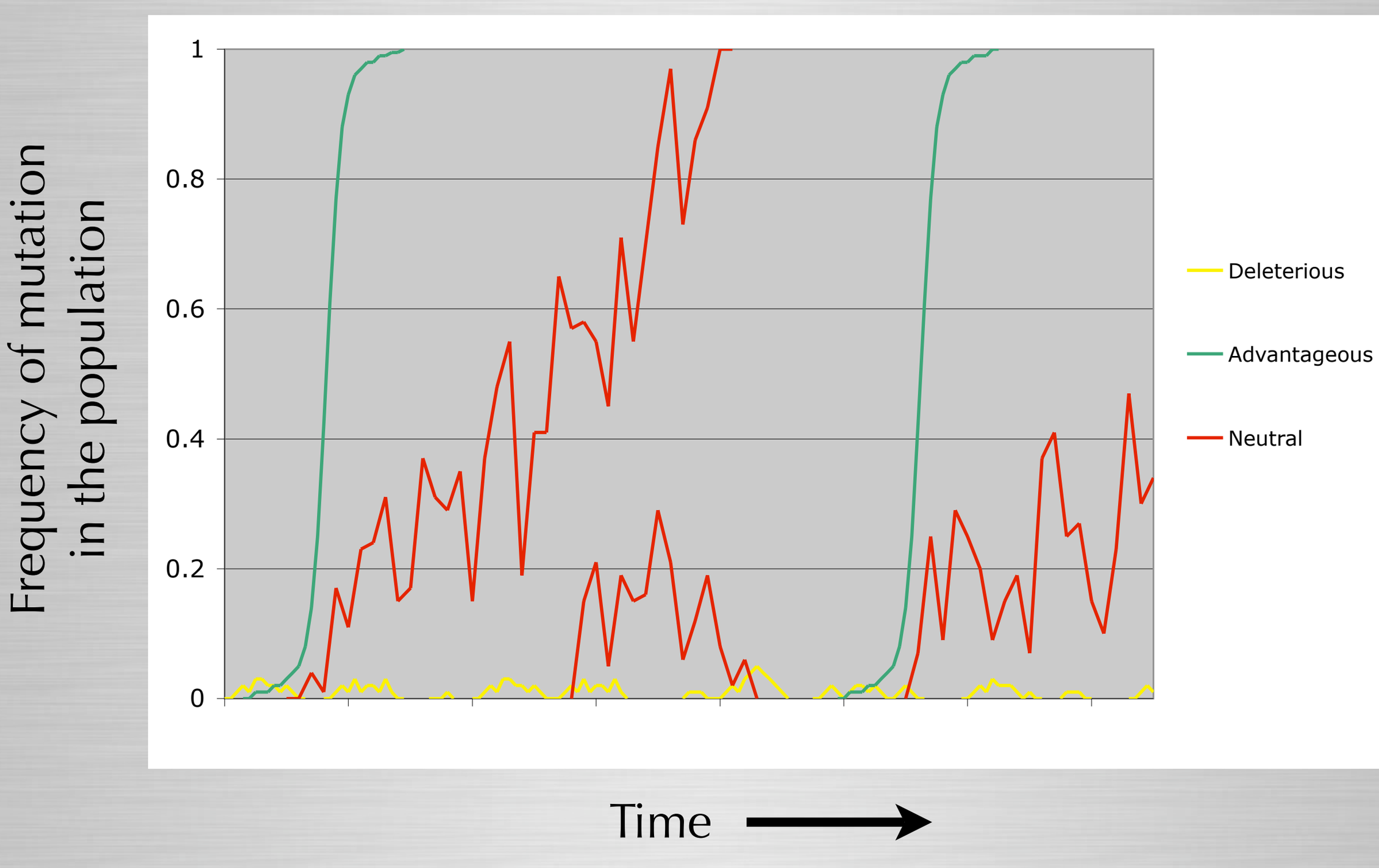
  neutral ($s=0$)

The Fate of Mutations (Simplified)



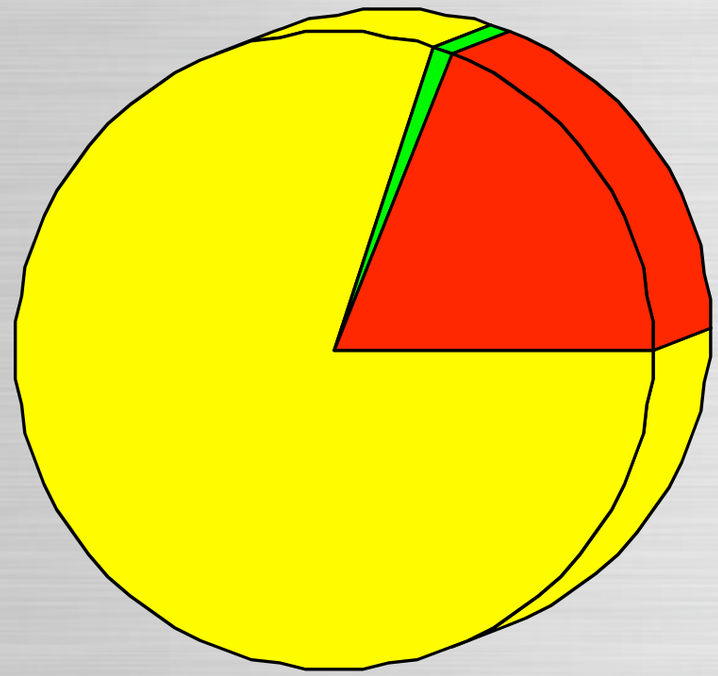
Fixation = Nucleotide Substitution

The Fate of Mutations (Simplified)



Distribution of Selection Coefficients

Neutralist Model



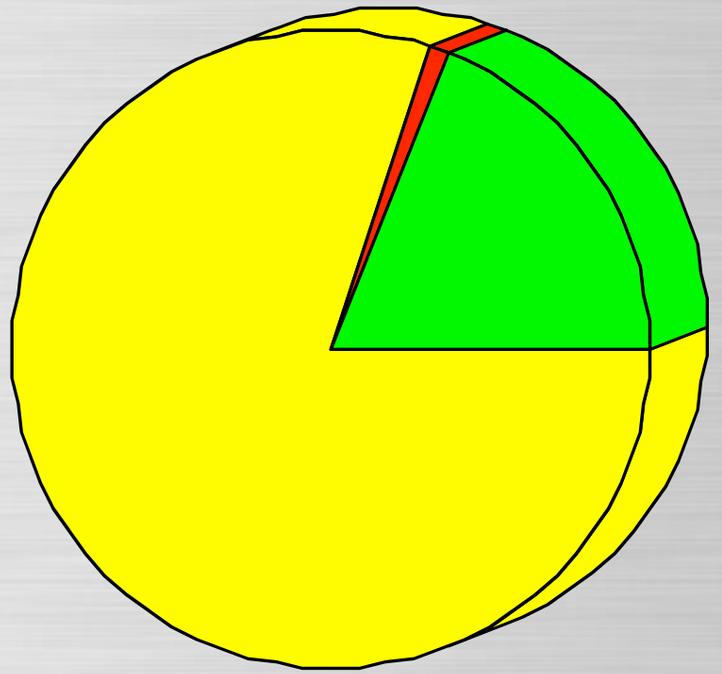
Most *fixed* mutations are neutral

s<0

s>0

s≈0

Selectionist Model



Most *fixed* mutations are advantageous

A Model Of Substitution Rate

- m = mutation rate per site per individual

  N = population size

  $f(s)$ = distribution of selection coefficients

  = probability that a given mutation has selection coefficient s

  $p(s)$ = probability of fixation of mutations with selection coefficient s

  $\sim (1-\exp(-2s)) / (1-\exp(-2Ns))$ in a constant-size population

  $r(s)$ = rate at which mutations with selection coefficient s are fixed

  $= N m f(s) p(s)$

- μ = total substitution rate

  $\mu = \int r(s).ds$

  $= Nm \int f(s)\ p(s).ds$

A Model Of Substitution Rate

- **For neutral mutations, *s*=0**

$$p(0) = 1/N \quad , \quad r(0) \approx mNf(0)/N = mf(0)$$

- **For advantageous mutations, *s*>0**

$$p(s>0) \approx 2s \quad , \quad r(s>0) \approx 2Nmsf(s>0)$$

- **For disadvantageous mutations, *s*<0**

$$p(s<0) \approx 0 \text{ hence } r \approx 0$$

dN/dS

- Are fixed mutations neutral or advantageous?
- Silent mutations are likely to be neutral.
- Replacement changes could be neutral, beneficial or deleterious.
- *dN/dS* = ω = ratio of replacement and silent substitution rates.

  dN/dS = 1 if all replacement changes are neutral

  dN/dS < 1 if some replacement changes are deleterious

  dN/dS = 0 if all replacement changes are deleterious

  dN/dS > 1 **only if** some replacement changes are beneficial

When are silent changes not neutral?

- Overlapping genes and reading frames.

  Very common in viruses (e.g. Hepatitis B Virus)

- Secondary RNA / DNA structure

  e.g. stem-loop structures

- Codons for the same amino acid differ in fitness

  May result from different translational efficiencies. Gives rise to codon bias.

- Untranslated or regulatory regions

Detecting Selection Using *dN/dS*

- **Pairwise method:** Calculates *dN/dS* from each pair of sequences, then averages *dN/dS* across all pairs.

  e.g. Nei & Gojobori 1986 (uses Jukes-Cantor model)

- **Parsimony method**: Start with a phylogeny. A parsimony algorithm then estimates the ancestral codons at the internal nodes. dN/dS is then calculated from the tree with specified ancestral codons.

  e.g. ADAPTSITE (Suzuki & Gojobori 1999) http://mep.bio.psu.edu

- **Substitution model method:** Starts with a phylogeny. Uses a codon-based substitution model that contains *dN/dS* as a parameter. Allows *dN/dS* to vary among codons and can identify individual codons under selection. Uses realistic models of substitution.

  e.g. CODEML (Yang et al. 2000) http://abacus.gene.ucl.ac.uk/

Codon Substitution Model

- Define a 61×61 substitution matrix (**Q**). Each element in the matrix $q_{i,j}$ defines the substitution rate of codon *i* to codon *j*.

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position} \\ \pi_j, & \text{for synonymous transversion} \\ \kappa\pi_j, & \text{for synonymous transition} \\ \omega\pi_j, & \text{for nonsynonymous transversion} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition} \end{cases}$$

- κ is the transition/transversion rate ratio
- ω is the silent/replacement rate ratio
- π_j is the frequency of codon *j*

Codon Substitution Models

- Codon models are calculated like nucleotide models.
- The genetic distance between codon *i* and codon *j* is given by the distribution $P_{i,j}(t) = \exp(\mathbf{Q}t)$.
- This probability is multiplied across all codons and across all branches in the tree. As before, it is then integrated over all possible ancestral codons at internal nodes.
- This gives the final phylogenetic probability: $P(seqs|\mathbf{T},\mathbf{B},\mathbf{Q})$
- If ω varies among codons, then the final probability is obtained by integrating across the specified distribution of ω among sites.

Disadvantages of *dN/dS* methods

- Assumes all changes are the result of past fixation events. Caution is required if genetic variation represents non-fixed polymorphism (e.g. within-patient viral data sets).

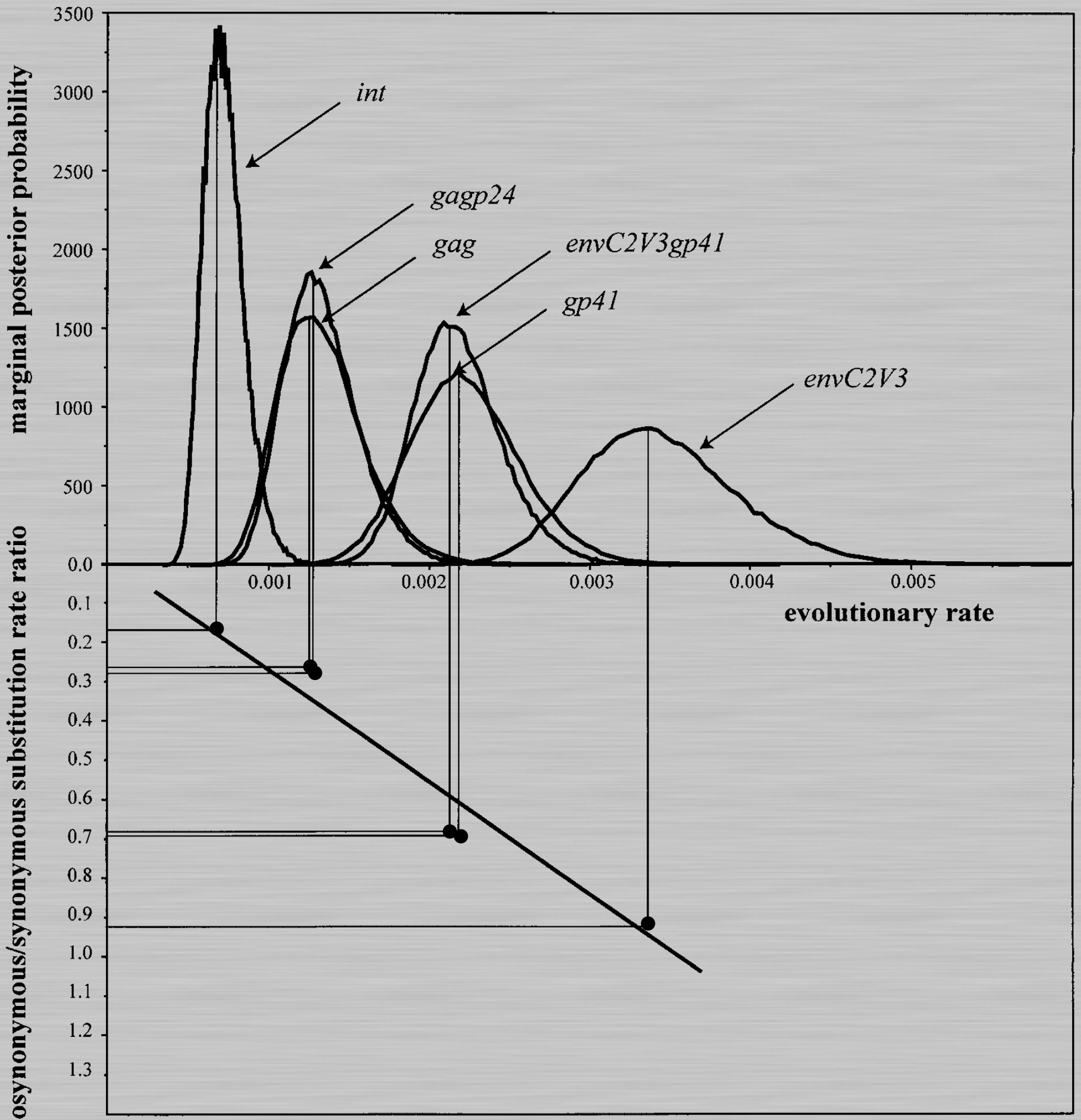
- Conservative. Will fail to detect many cases of selection, especially if a site has only undergone a single selective sweep.

- Such methods therefore tend to identify sites that are repeatedly positively selected.

- Assumes substitutions are independent at each codon (no epistasis or clonal interference).

Table 1. Selected examples of protein-coding genes in which positive selection was detected by using the d_N/d_S ratio

Gene	Organism	Refs
Genes involved in defensive systems or immunity		
Class I chitinase gene	*Arabis* and *Arabidopsis*	41
Colicin genes	*Escherichia coli*	45
Defensin genes	Rodents	46
Fv1	*Mus*	47
Immunoglobulin V_H genes	Mammals	48
MHC genes	Mammals	49
Polygalacturonase inhibitor genes	Legume and dicots	50
RH blood group and RH50 genes	Primates and rodents	51
Ribonuclease genes	Primates	52
Transferrin gene	Salmonid fishes	53
Type I interferon- v gene	Mammals	54
a_1-Proteinase inhibitor genes	Rodents	55
Genes involved in evading defensive systems or immunity		
Capsid gene	FMD virus	42
CSP, *TRAP*, *MSA-2* and *PF83*	*Plasmodium falciparum*	56
Delta-antigen coding region	Hepatitis D virus	57
E gene	Phages *G4*, *f X174*, and *S13*	3
Envelope gene	HIV	40
gH glycoprotein gene	Pseudorabies virus	3
Hemagglutinin gene	Human influenza A virus	33
Invasion plasmid antigen genes	*Shigella*	3
Merozoite surface antigen-1 gene	*Plasmodium falciparum*	58
msp 1 a	*Anaplasma marginale*	3
nef	HIV	38
Outer membrane protein gene	*Chlamydia*	3
Polygalacturonase genes	Fungal pathogens	50
Porin protein 1 gene	*Neisseria*	59
S and *HE* glycoprotein genes	Murine coronavirus	60
Sigma-1 protein gene	Reovirus	3
Virulence determinant gene	*Yersinia*	3
Genes involved in reproduction		
18-kDa fertilization protein gene	Abalone (*Haliotis)*	61
Acp26Aa	*Drosophila*	62
Androgen-binding protein gene	Rodents	63
Bindin gene	*Echinometra*	64
Egg-laying hormone genes	*Aplysia californica*	3
Ods homeobox gene	*Drosophila*	65
Pem homeobox gene	Rodents	66
Protamine P1 gene	Primates	67
Sperm lysin gene	Abalone (*Haliotis)*	61
S-Rnase gene	Rosaceae	68
Sry gene	Primates	69
Genes involved in digestion		
k-casein gene	Bovids	70
Lysozyme gene	Primates	23
Toxin protein genes		
Conotoxin genes	*Conus* gastropods	71
Phospholipase A_2 gene	Crotalinae snakes	72
Genes related to electron transport and/or ATP synthesis		
ATP synthase F_0 subunit gene	*Escherichia coli*	3
COX7A isoform genes	Primates	73
COX4 gene	Primates	74
Cytokine genes		
Granulocyte-macrophage SF gene	Rodents	75
Interleukin-3 gene	Primates	75
Interleukin-4 gene	Rodents	75
Miscellaneous		
CDC6	*Saccharomyces cerevisiae*	3
Growth hormone gene	Vertebrates	76
Hemoglobin *b*-chain gene	Antarctic fishes	77
Jingwei	*Drosophila*	78
Prostatein peptide C3 gene	Rat	3

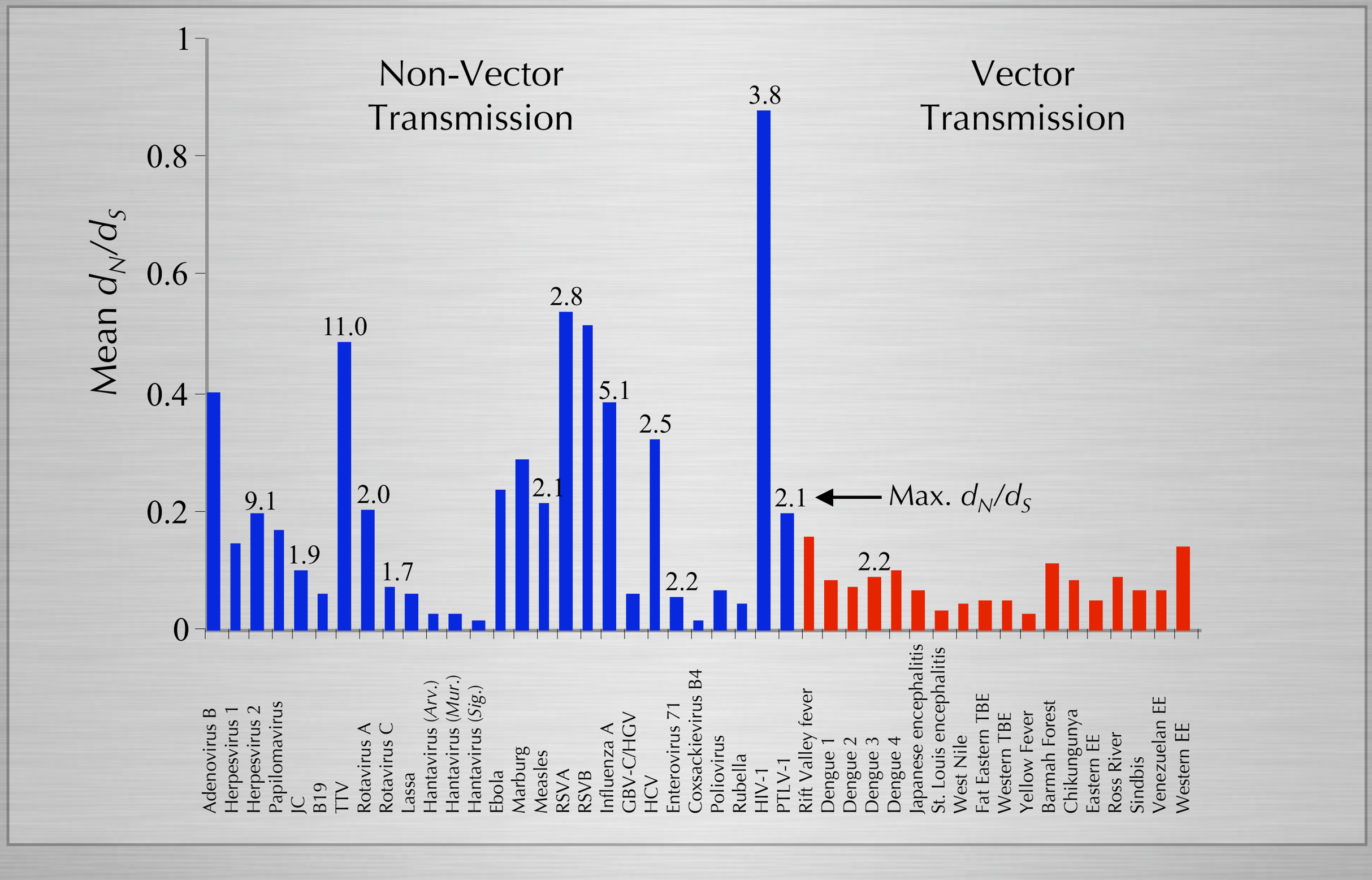


Example:

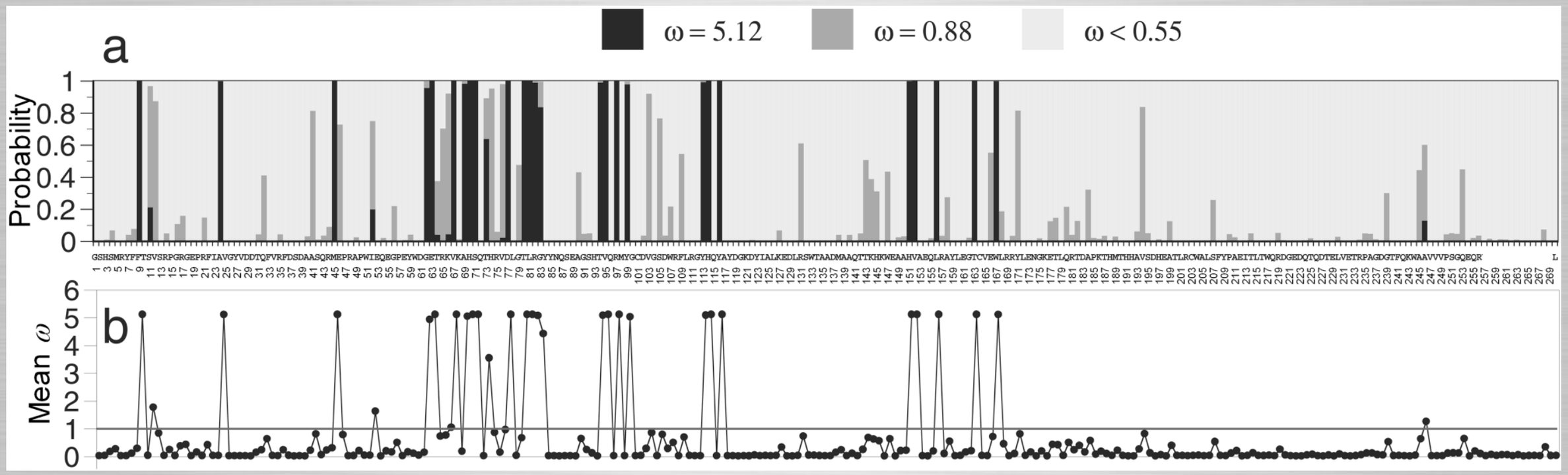
dN/dS varies among genes

(HIV-1 group O)

Example: *dN/dS* varies among species



dN/dS varies among codons

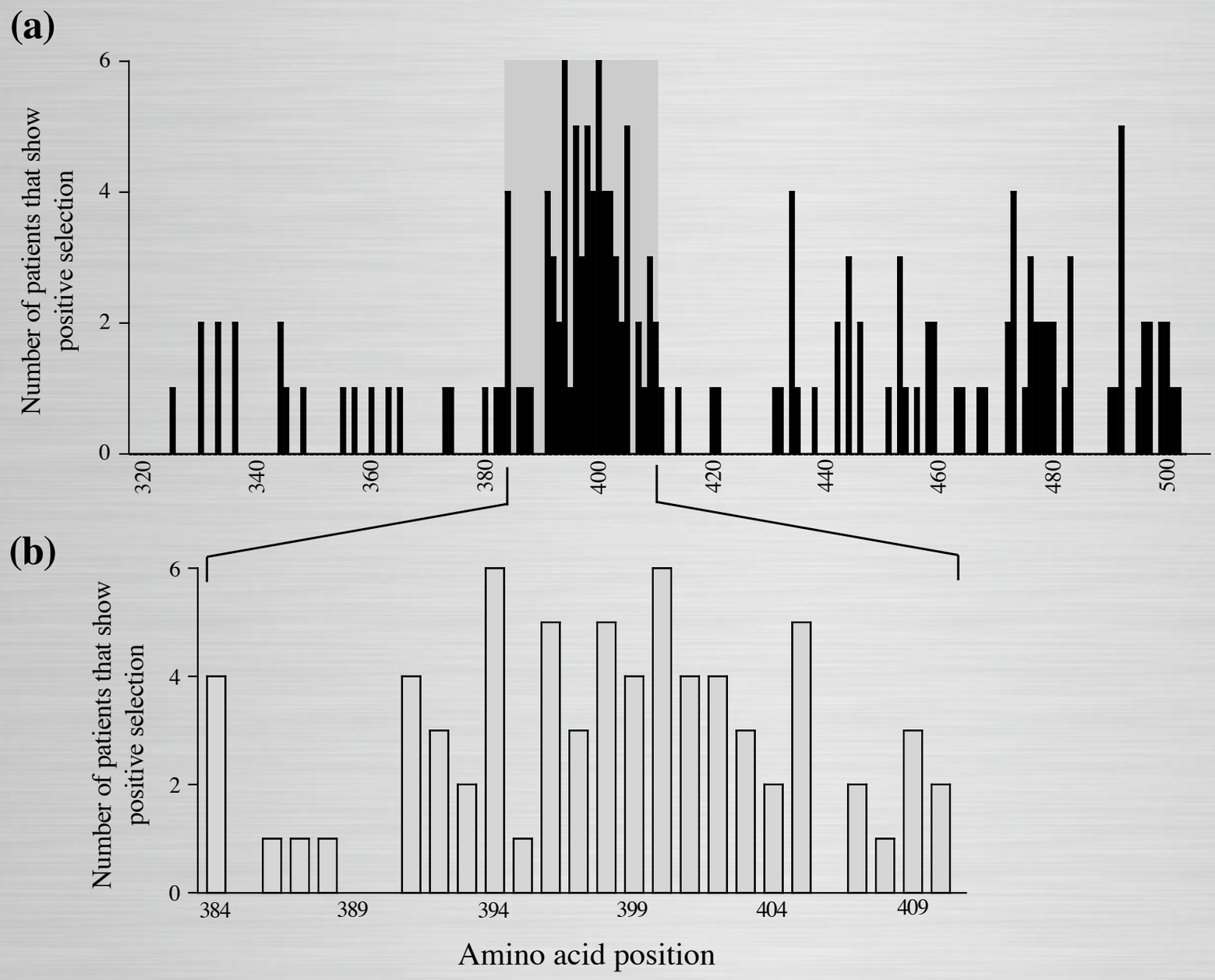


MHC class I gene

High *dN/dS* found in codons that form the antigen-recognition site

(Yang & Swanson. 2002. Mol. Biol. Evol. 19:49-57)

dN/dS varies among codons

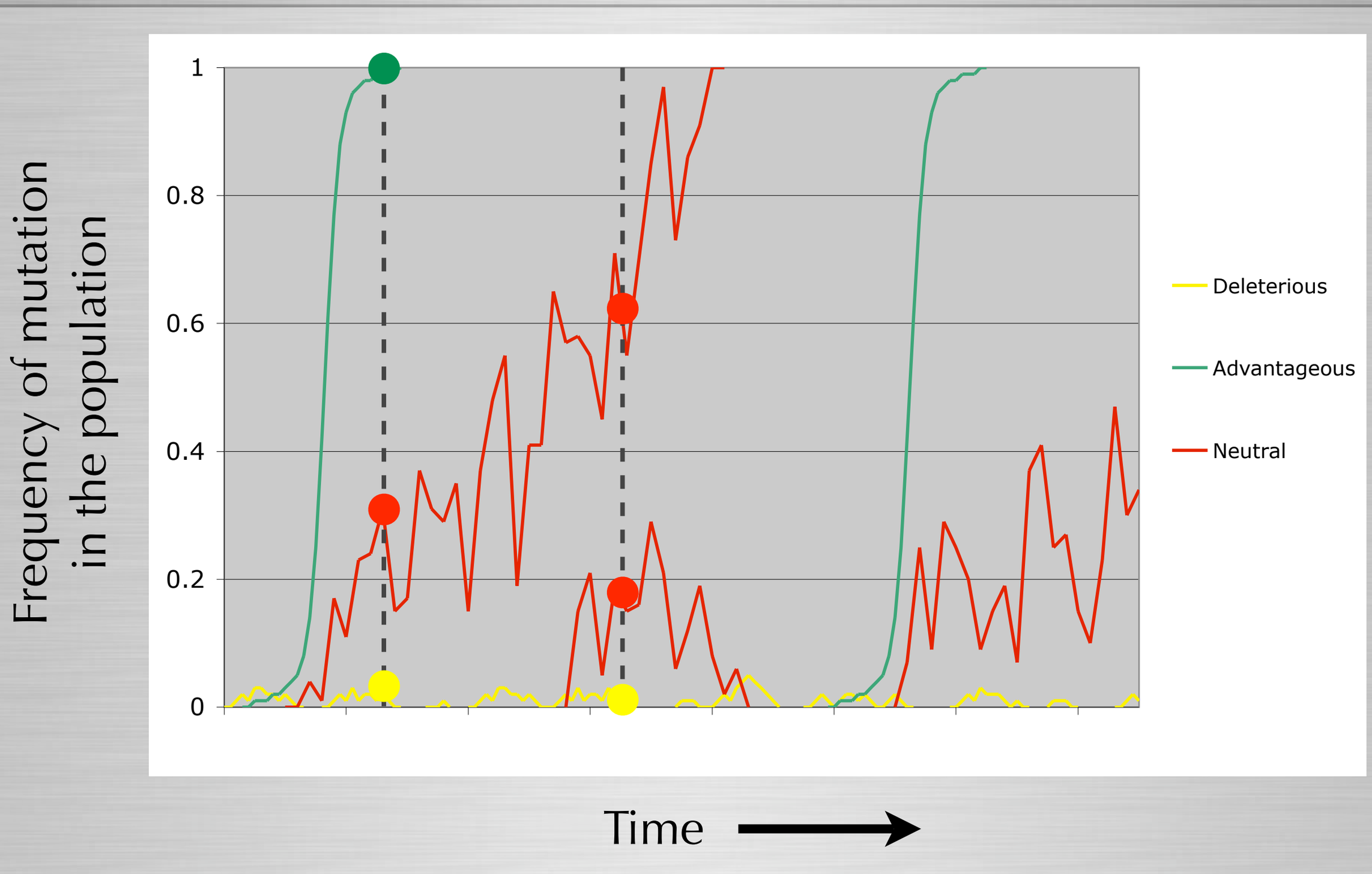


Positively selected codons in the Hepatitis C Virus envelope genes

Hypervariable region

(Sheridan, Pybus, Holmes & Klenerman. 2004. J. Virol. 78:3447-3454)

Mutation Site Frequencies



Mutation Site Frequencies

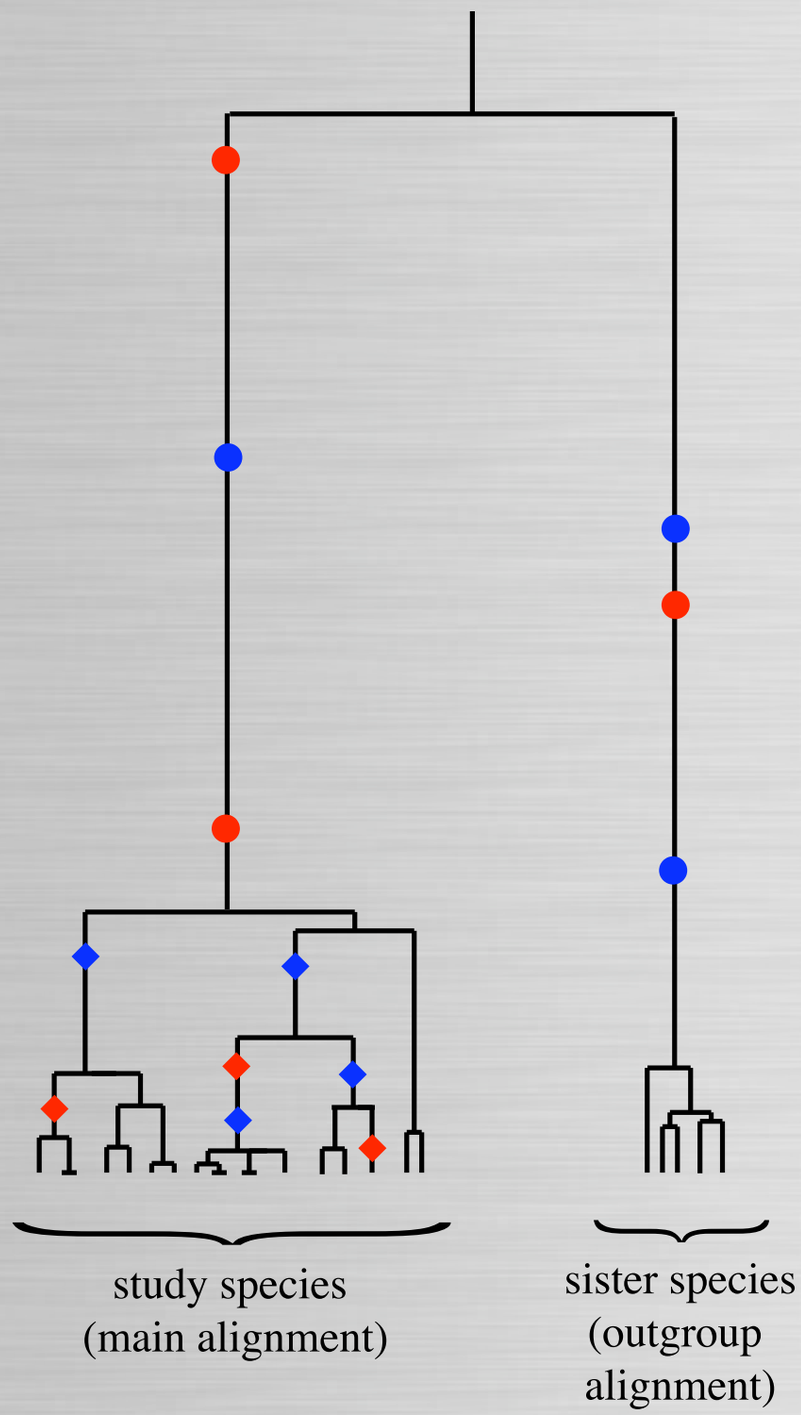
- The distribution of site frequencies (the site-frequency spectrum) in a sample of sequences contains information about natural selection.

- Can be analyzed using Poisson Random Field (PRF) models.

- PRF models assume that polymorphic sites are unlinked (effectively on different chromosomes).

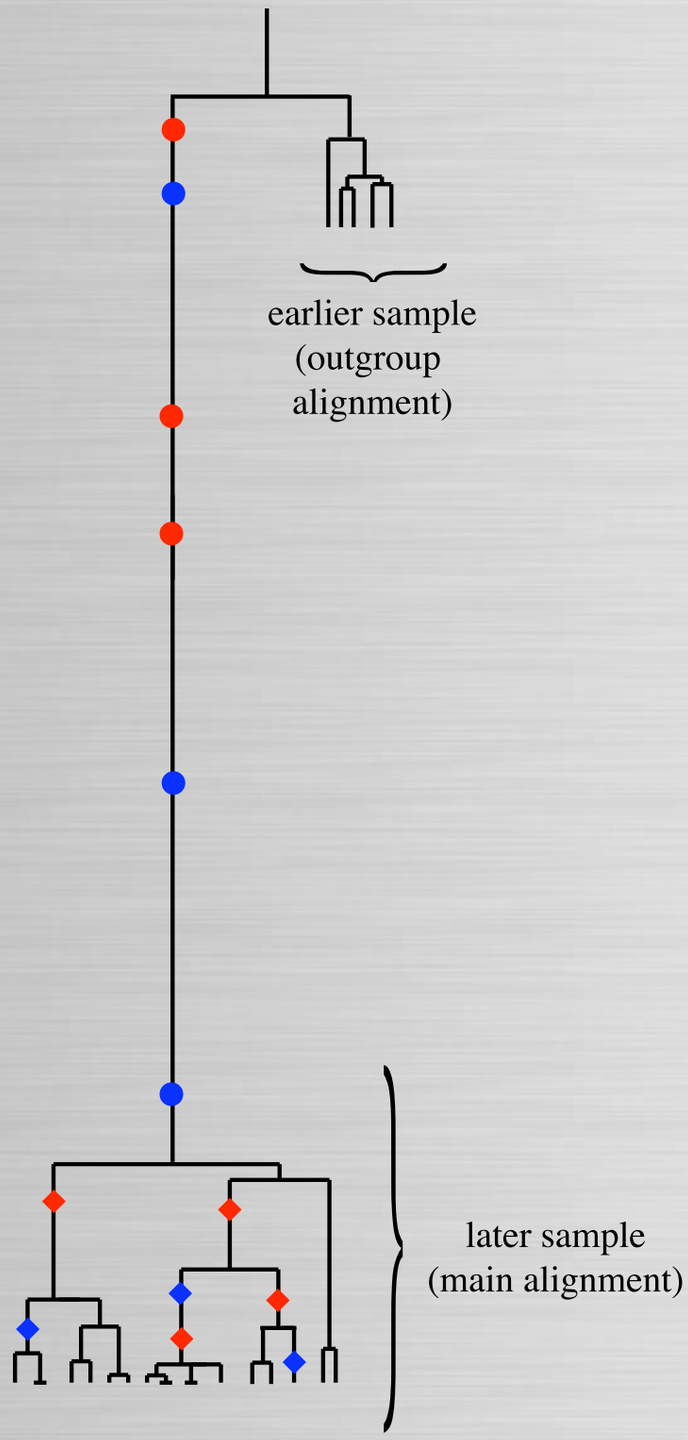
- PRF models are more suitable for human genome analyses than for pathogen genome analyses.

McDonald-Kreitman Method



- Circles represent *fixation* between species. Diamonds represent *polymorphism* within species.
- If there is no selection, then both fixation and polymorphism are driven by genetic drift.
- The ratio of silent/replacement fixations should equal the ratio of silent/non-replacement polymorphisms.
- A statistical deviation from equality suggests natural selection.
- **Assumes all polymorphism is neutral.**

Estimating Viral Adaptation Rate



- Can also be applied to pathogen sequences sampled through time.
  - $s_{<1}$ = # silent polymorphisms (◆)
  - $r_{<1}$ = # replacement polymorphisms (◆)
  - s_1 = # silent fixations (●)
  - r_1 = # replacement fixations (●)
  - a_1 = # positively selected fixations
- If there is no selection, then, on average:

$$(r_1 - a_1)/r_{<1} = s_1 / s_{<1}$$

- Rearranging, we can estimate a_1 using:

$$a_1 = r_1\left[1 - \left(\frac{s_1}{r_1}\right)z\right], \text{ where } z = \left(\frac{r_{<1}}{s_{<1}}\right).$$

Estimating Viral Adaptation Rate

- In virus populations, mutations that are highly-frequent but not yet fixed may also be adaptations.

  $s_{<0.5}$ = # silent polymorphisms with a frequency < 0.5

  $r_{<0.5}$ = # replacement polymorphisms with a frequency < 0.5

  $s_{>0.5}$ = # silent polymorphisms with a frequency > 0.5

  $r_{>0.5}$ = # replacement polymorphisms with a frequency > 0.5

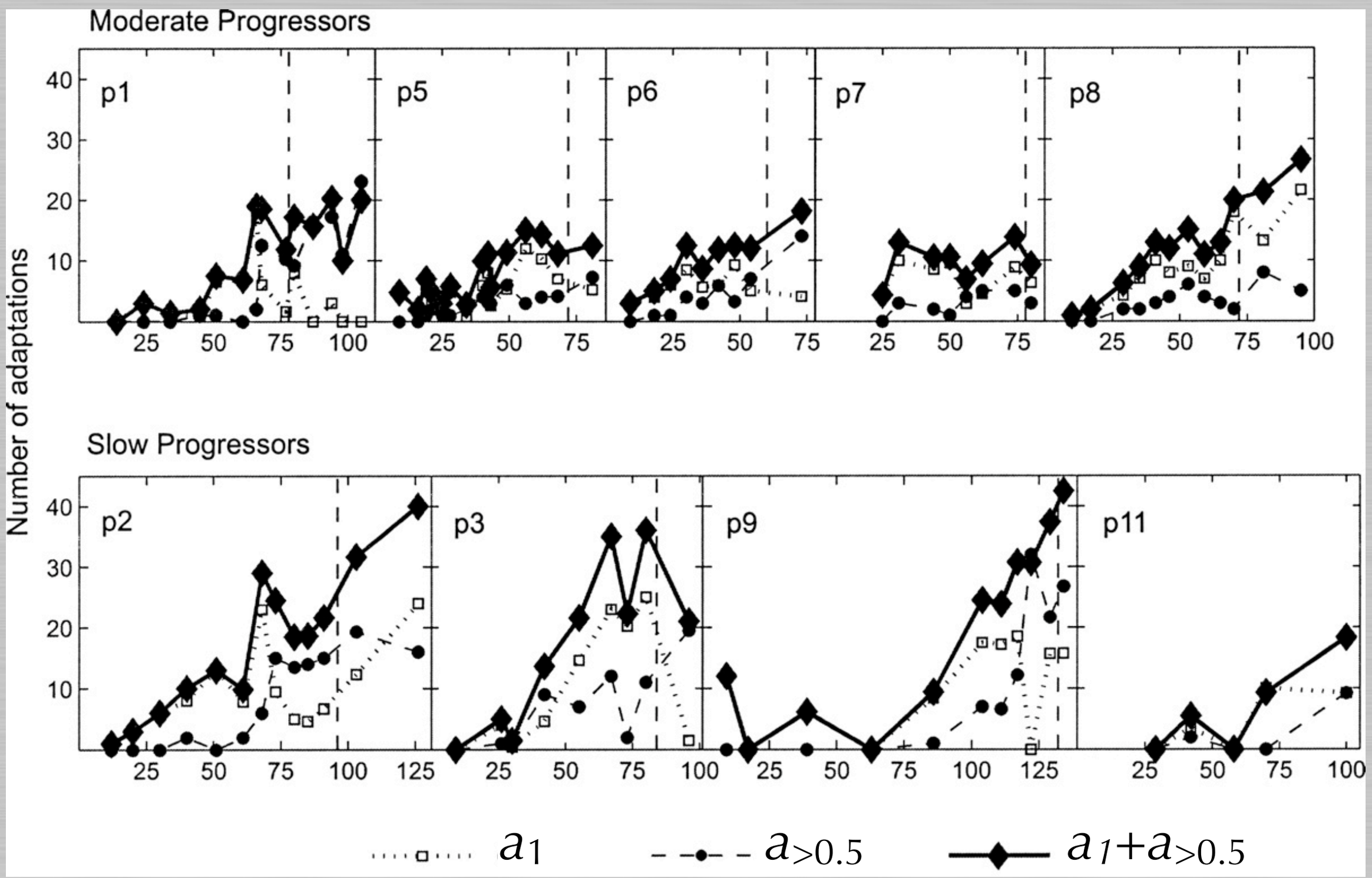
  $a_{>0.5}$ = # positively selected polymorphisms with a frequency > 0.5

- We can now estimate both a_1 and $a>_{0.5}$ using:

$$a_1 = r_1\left[1-\left(\frac{s_1}{r_1}\right)z\right], \text{ where } z=\left(\frac{r_{<0.5}}{s_{<0.5}}\right) \qquad a_{>0.5} = r_{>0.5}\left[1-\left(\frac{s_{>0.5}}{r_{>0.5}}\right)z\right], \text{ where } z=\left(\frac{r_{<0.5}}{s_{<0.5}}\right).$$

- This approach assumes that only polymorphisms with a frequency<0.5 are neutral.

Accumulation of adaptations during HIV infection



(Williamson. 2003. Mol. Biol. Evol. 20:1318-1325)