

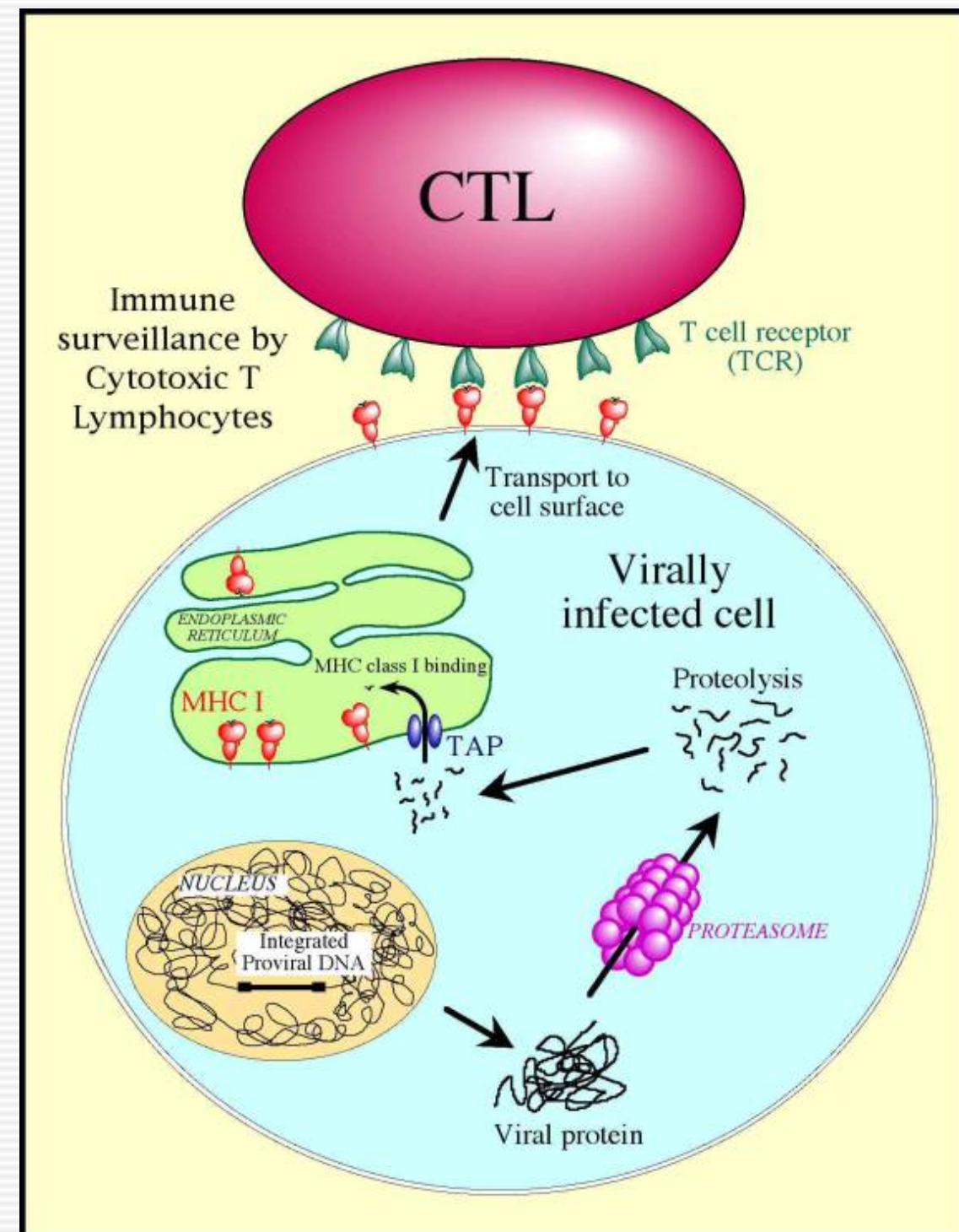
MOLECULAR EVOLUTION: NATURAL SELECTION

NATURAL SELECTION

- Mutation, recombination and other processes introduce variation into genomes of organisms
- The fitness of an organism describes how well it can survive/grow/function/replicate in a given environment, or how well it can pass on its genetic material to future generations
- Any particular mutation can be
 - Neutral: no or little change in fitness
 - Deleterious: reduced fitness
 - Adaptive: increased fitness

EXAMPLE IN HIV: MHC-RESTRICTED CTL KILLING

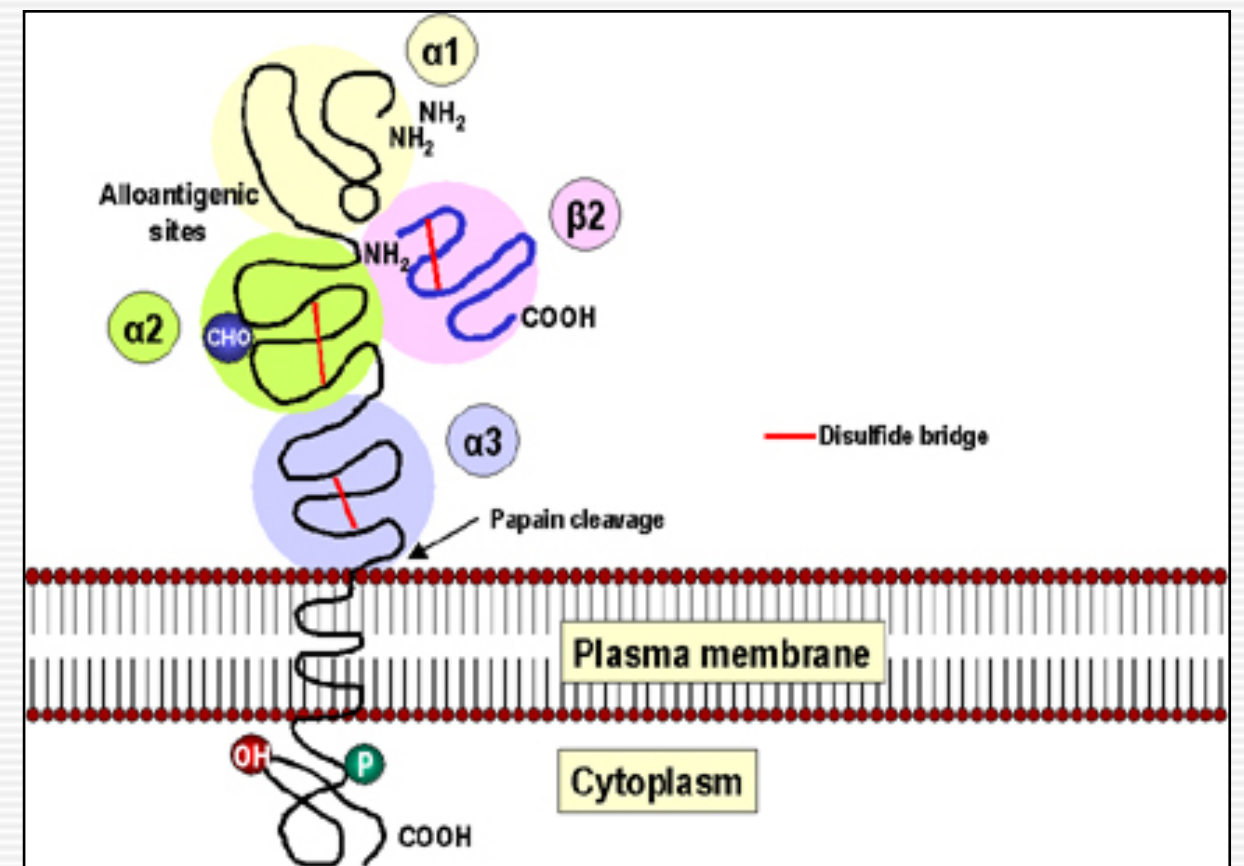
- Cytotoxic T-Lymphocytes effect cell-mediated immune response
- Viral proteins are cleaved by the proteasome, transported by TAP and loaded onto the MHC Class I molecule.
- MHC Class I presents a restricted polypeptide (epitope) on the surface of the cell.
- A CD8+ cell binds to presented foreign peptides via a T cell receptor (TCR) and initiates cell apoptosis.



MHC CLASS I MOLECULES

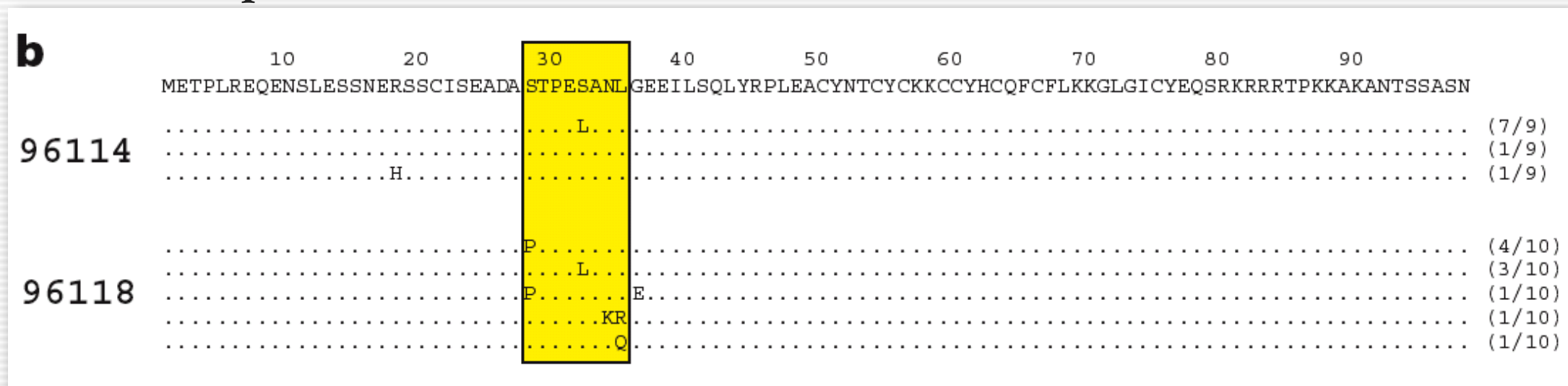
- Presents foreign peptides which are 9-10 amino-acid long
- Anchor sites (2 and 9) are usually important for binding and recognition
- Mutations which alter the peptide can hinder or prevent CTL response activation

Antigen Binding Site



RAPID SIV SEQUENCE EVOLUTION IN MACAQUES

- SIV: only animal model of HIV (rhesus macaques)
- Experimental infection with MHC-matched strain of SIV
- Virus sequenced from 2 weeks post infection
- Only variation was in an epitope recognized by the MHC
- “CTL escape”



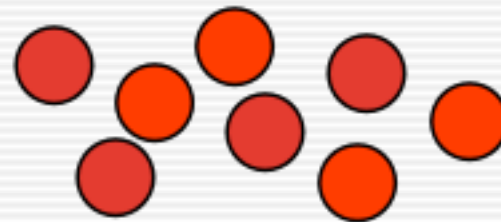
Before selection



After selection



Final population



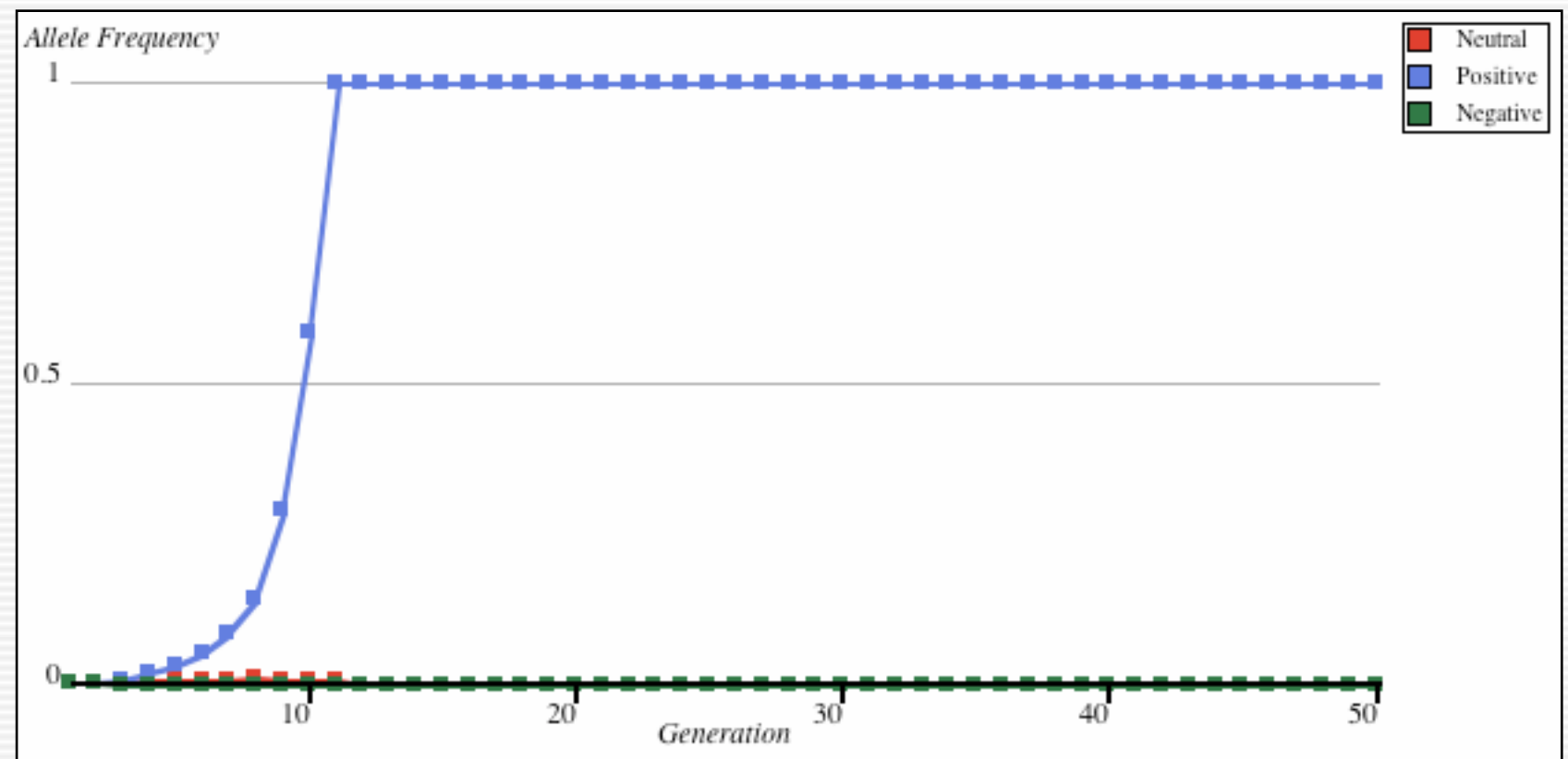
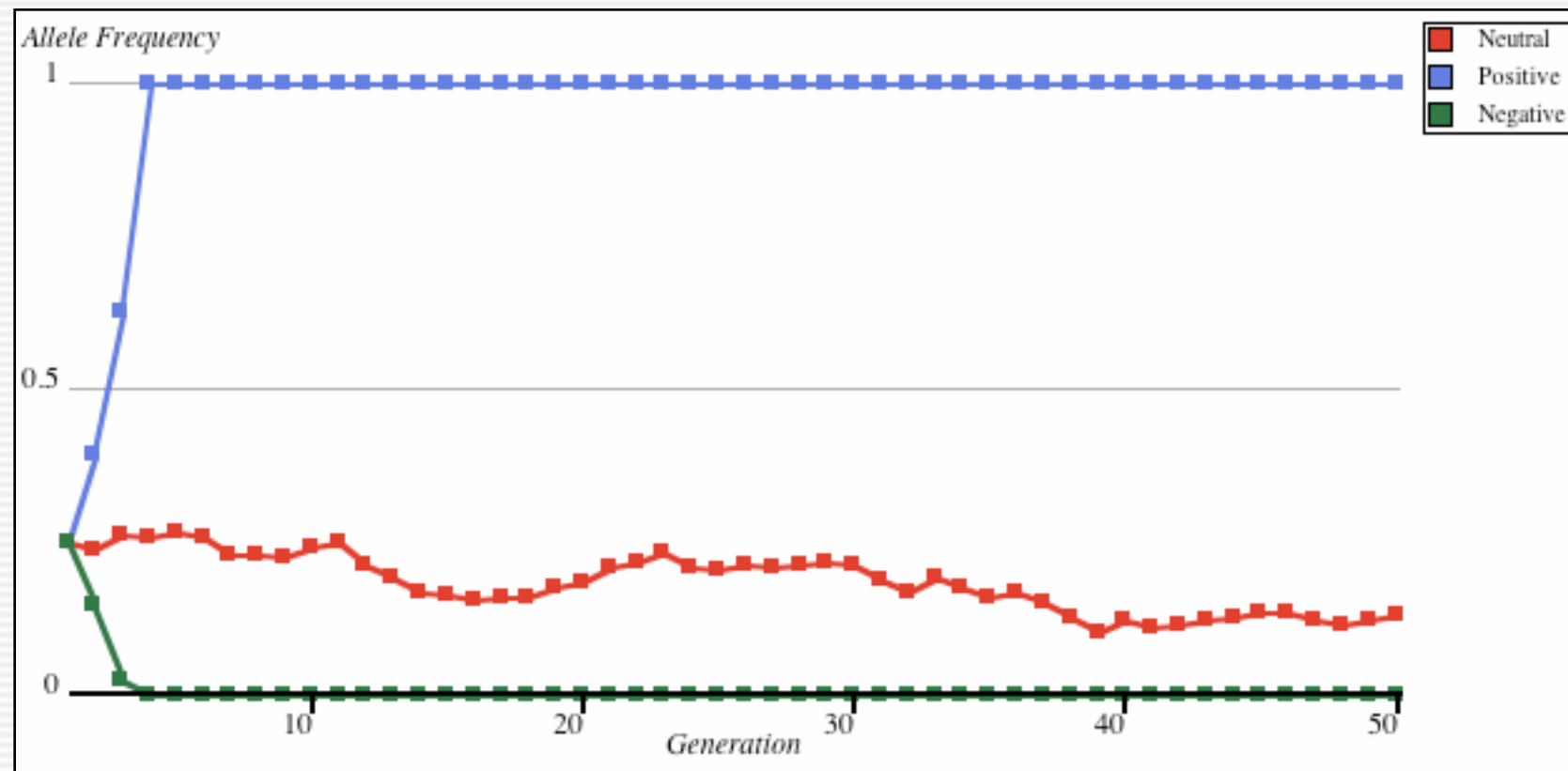
Resistance level



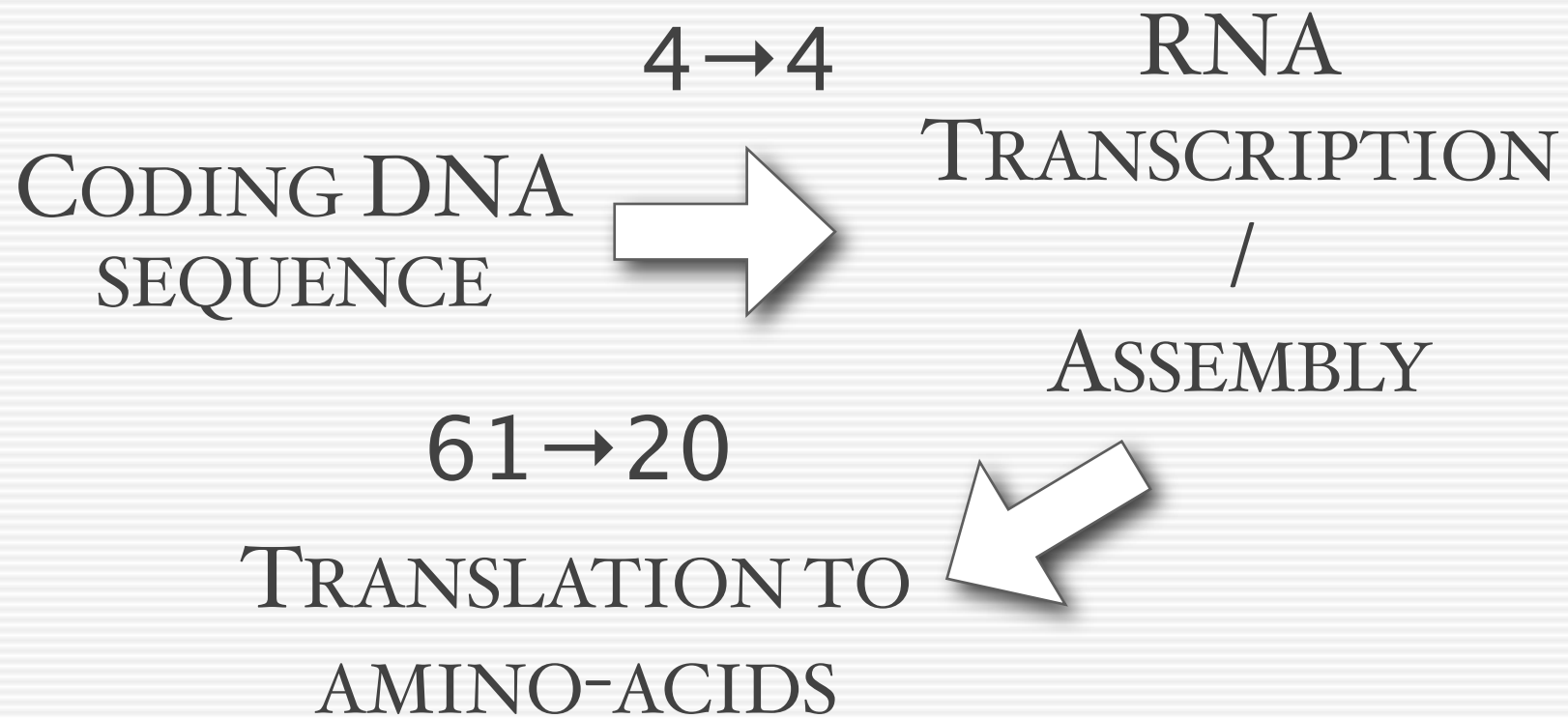
[HTTP://EN.WIKIPEDIA.ORG/WIKI/FILE:ANTIBIOTIC_RESISTANCE.SVG](http://en.wikipedia.org/wiki/File:Antibiotic_resistance.svg)

EVOLUTIONARY MODES

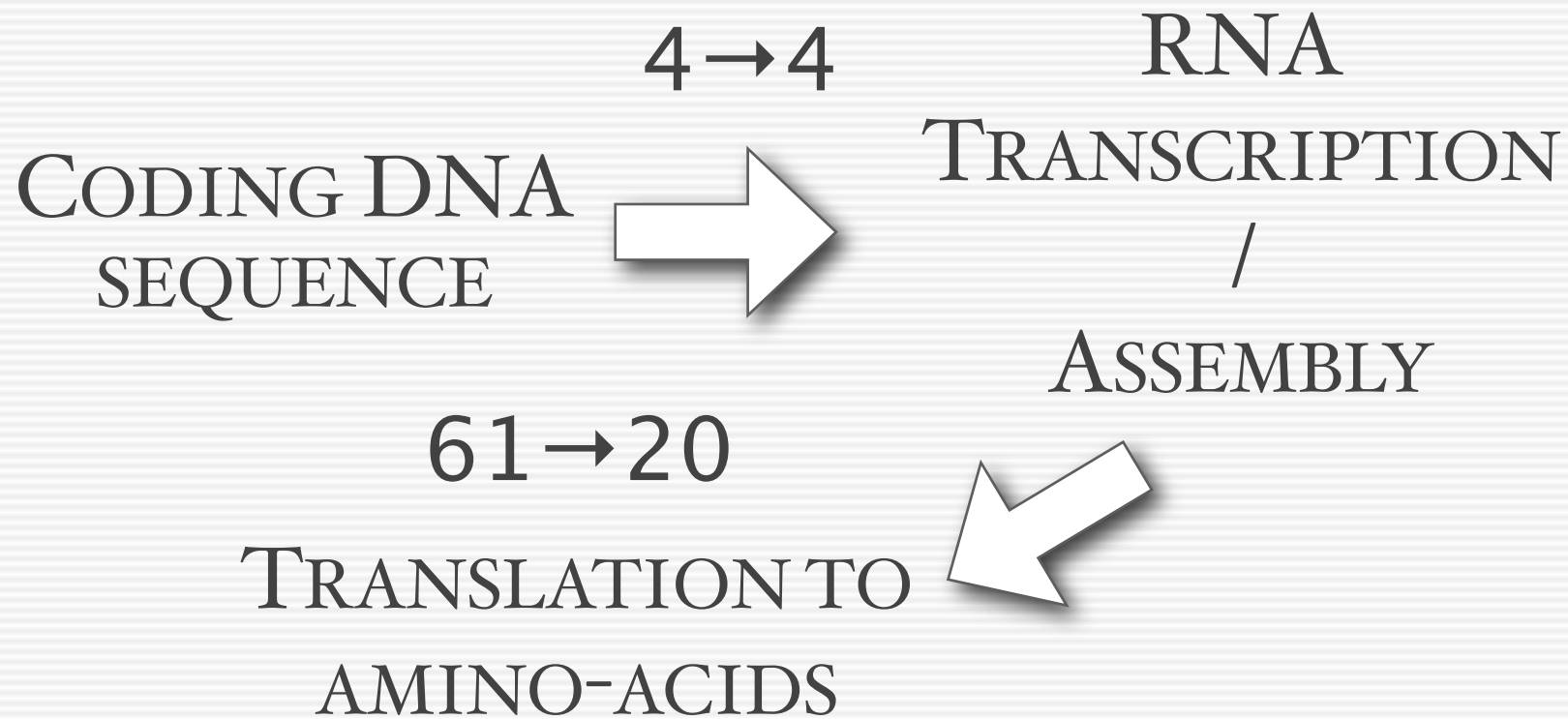
- Evolutionary selective processes can be classified into
 - **Neutral evolution:** mutations, randomly introduced into the gene increase or decrease in frequency stochastically (genetic drift)
 - **Negative selection:** the frequency of deleterious mutations is reduced by natural selection more rapidly than just by genetic drift
 - **Positive selection:** the frequency of adaptive mutations is increased by natural selection more rapidly than just by genetic drift



CODING SEQUENCES.

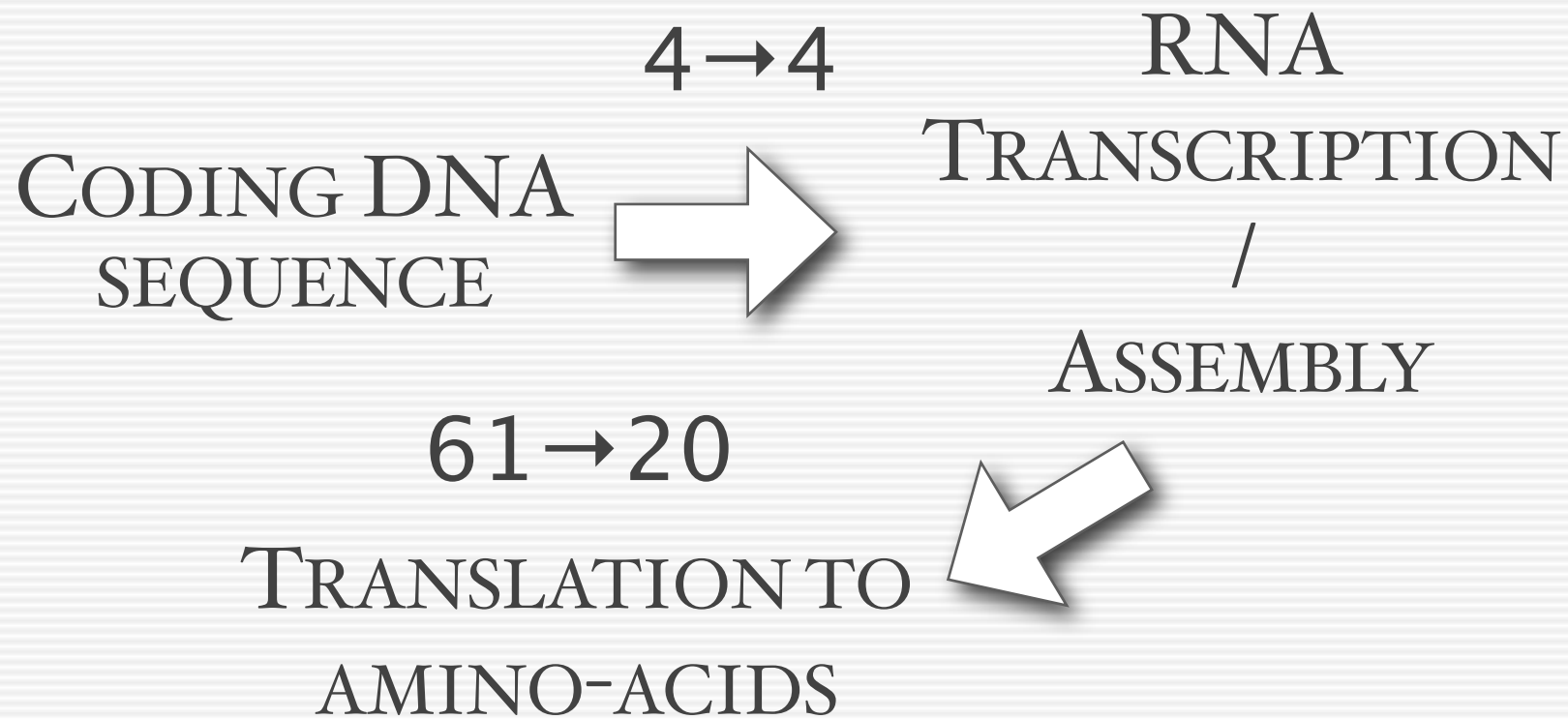


CODING SEQUENCES.



AMINOACID	CODONS	REDUNDANCY
ALANINE	GC*	4
CYSTEINE	TGC,TGT	2
ASPARTIC ACID	GAC,GAT	2
GLUTAMINE ACID	GAA,GAG	2
PHENYLALANINE	TTC,TTT	2
GLYCIN	GG*	4
HISTIDINE	CAC,CAT	2
ISOLEUCINE	ATA,ATC,ATT	3
LYSINE	AAA,AAG	2
LEUCINE	CT*,TTA,TTG	6
METHIONINE	ATG	1
ASPARGINE	AAC,AAT	2
PROLINE	CC*	4
GLUTAMINE	CAA,CAG	2
ARGININE	AGA,AGG,CG*	6
SERINE	AGC,AGT,TC*	6
THREONINE	AC*	4
VALINE	GT*	4
TRYPTOPHAN	TGG	1
TYROSINE	TAC,TAT	2
STOP	TAA,TAG,TGA	3

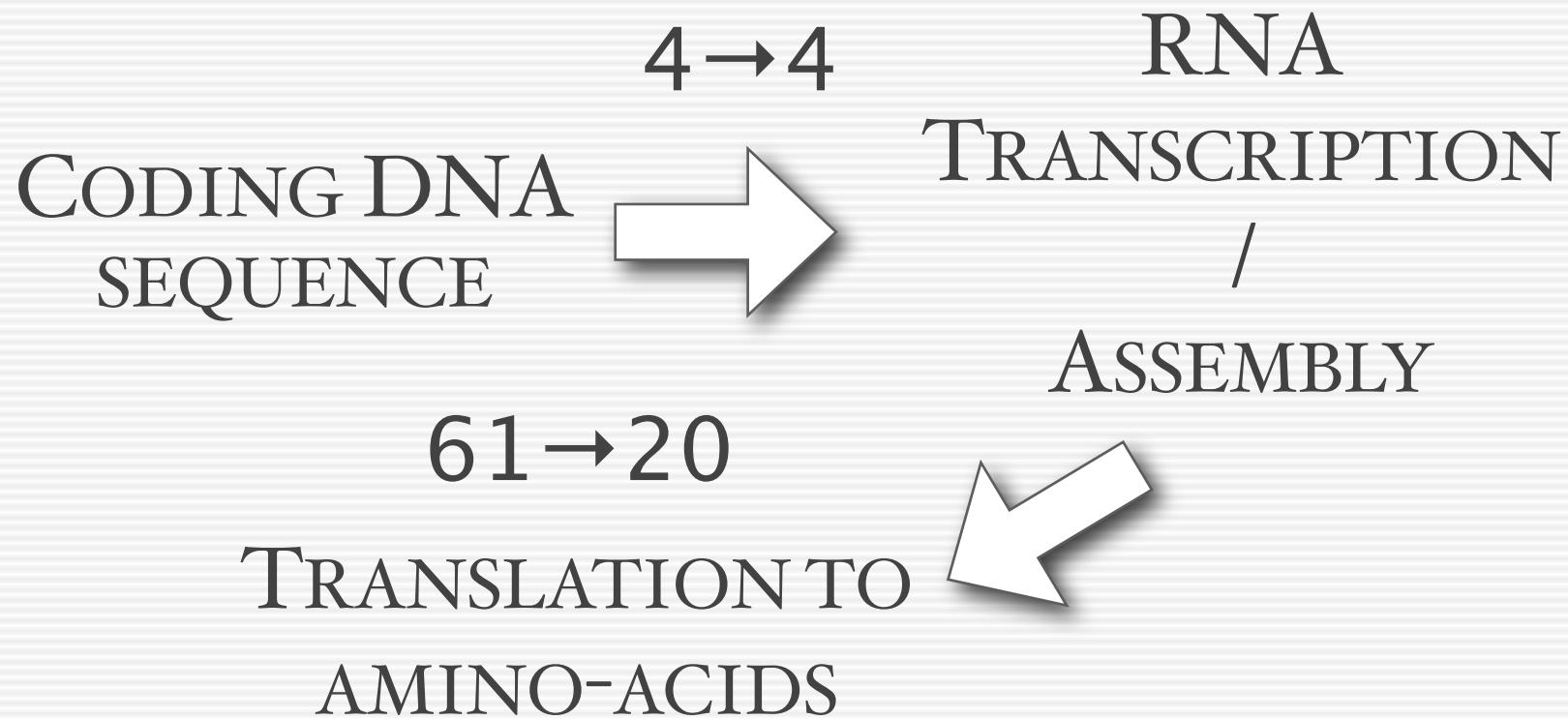
CODING SEQUENCES.



- Proper unit of evolution is a triplet of nucleotides - a codon

AMINOACID	CODONS	REDUNDANCY
ALANINE	GC*	4
CYSTEINE	TGC,TGT	2
ASPARTIC ACID	GAC,GAT	2
GLUTAMINE ACID	GAA,GAG	2
PHENYLALANINE	TTC,TTT	2
GLYCIN	GG*	4
HISTIDINE	CAC,CAT	2
ISOLEUCINE	ATA,ATC,ATT	3
LYSINE	AAA,AAG	2
LEUCINE	CT*,TTA,TTG	6
METHIONINE	ATG	1
ASPARGINE	AAC,AAT	2
PROLINE	CC*	4
GLUTAMINE	CAA,CAG	2
ARGININE	AGA,AGG,CG*	6
SERINE	AGC,AGT,TC*	6
THREONINE	AC*	4
VALINE	GT*	4
TRYPTOPHAN	TGG	1
TYROSINE	TAC,TAT	2
STOP	TAA,TAG,TGA	3

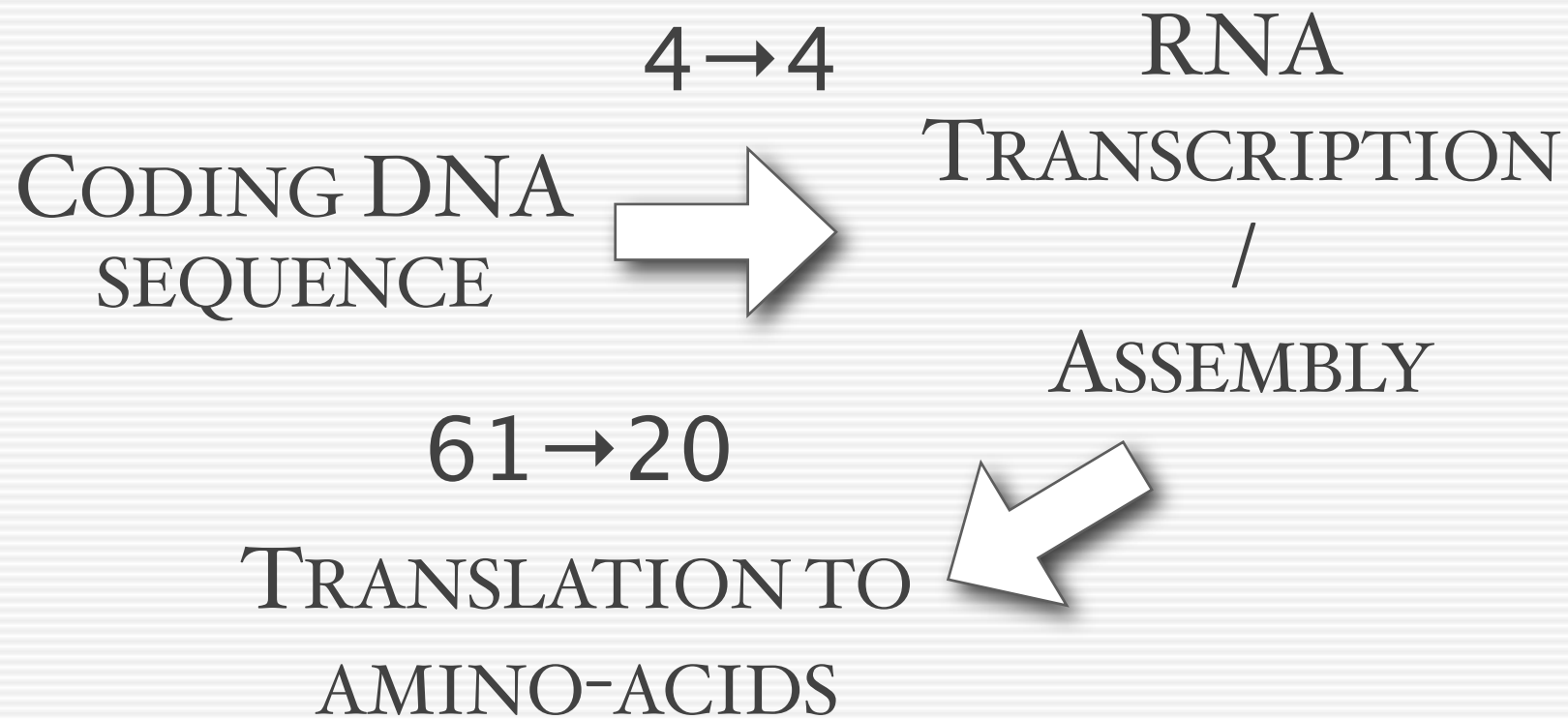
CODING SEQUENCES.



- Proper unit of evolution is a triplet of nucleotides - a codon
- Multiple and differing redundancies in the genetic code

AMINOACID	CODONS	REDUNDANCY
ALANINE	GC*	4
CYSTEINE	TGC,TGT	2
ASPARTIC ACID	GAC,GAT	2
GLUTAMINE ACID	GAA,GAG	2
PHENYLALANINE	TTC,TTT	2
GLYCIN	GG*	4
HISTIDINE	CAC,CAT	2
ISOLEUCINE	ATA,ATC,ATT	3
LYSINE	AAA,AAG	2
LEUCINE	CT*,TTA,TTG	6
METHIONINE	ATG	1
ASPARGINE	AAC,AAT	2
PROLINE	CC*	4
GLUTAMINE	CAA,CAG	2
ARGININE	AGA,AGG,CG*	6
SERINE	AGC,AGT,TC*	6
THREONINE	AC*	4
VALINE	GT*	4
TRYPTOPHAN	TGG	1
TYROSINE	TAC,TAT	2
STOP	TAA,TAG,TGA	3

CODING SEQUENCES.



- Proper unit of evolution is a triplet of nucleotides - a codon
- Multiple and differing redundancies in the genetic code
- Synonymous and non-synonymous substitutions are fundamentally different

AMINOACID	CODONS	REDUNDANCY
ALANINE	GC*	4
CYSTEINE	TGC,TGT	2
ASPARTIC ACID	GAC,GAT	2
GLUTAMINE ACID	GAA,GAG	2
PHENYLALANINE	TTC,TTT	2
GLYCIN	GG*	4
HISTIDINE	CAC,CAT	2
ISOLEUCINE	ATA,ATC,ATT	3
LYSINE	AAA,AAG	2
LEUCINE	CT*,TTA,TTG	6
METHIONINE	ATG	1
ASPARGINE	AAC,AAT	2
PROLINE	CC*	4
GLUTAMINE	CAA,CAG	2
ARGININE	AGA,AGG,CG*	6
SERINE	AGC,AGT,TC*	6
THREONINE	AC*	4
VALINE	GT*	4
TRYPTOPHAN	TGG	1
TYROSINE	TAC,TAT	2
STOP	TAA,TAG,TGA	3

MOLECULAR SIGNATURES OF SELECTION

- Because synonymous mutations do not alter the protein, we may assume that they are neutral
- The rate of accumulation of synonymous mutations (**dS**) gives the neutral background
- We can compare the rate of accumulation of non-synonymous mutations (**dN**), which alter the protein sequence, to classify the nature of the evolutionary process

SELECTIVE PRESSURE

dS = “RATE” OF SYNONYMOUS SUBSTITUTIONS

dN = “RATE” OF NON-SYNONYMOUS SUBSTITUTIONS

SELECTIVE PRESSURE

dS = “RATE” OF SYNONYMOUS SUBSTITUTIONS

dN = “RATE” OF NON-SYNONYMOUS SUBSTITUTIONS

Positive Selection

$$dS < dN$$

SELECTIVE PRESSURE

dS = “RATE” OF SYNONYMOUS SUBSTITUTIONS

dN = “RATE” OF NON-SYNONYMOUS SUBSTITUTIONS

Positive Selection

$$dS < dN$$

Negative Selection

$$dS > dN$$

SELECTIVE PRESSURE

dS = “RATE” OF SYNONYMOUS SUBSTITUTIONS

dN = “RATE” OF NON-SYNONYMOUS SUBSTITUTIONS

Positive Selection

$$dS < dN$$

Negative Selection

$$dS > dN$$

Neutral Evolution

$$dS = dN$$

ESTIMATING DS AND DN

- Consider two aligned sequences

ACA	ATA	AT C	TTT	AAT T	CAA
T	I	I	F	N	Q
ACA	ATA	AC C	TTT	AAC C	CAA
T	I	T	F	N	Q

ESTIMATING dS AND dN

- Consider two aligned sequences

ACA	ATA	AT C	TTT	AAT T	CAA
T	I	I	F	N	Q
ACA	ATA	AC C	TTT	AAC C	CAA
T	I	T	F	N	Q

- Can we say that $dN/dS = 1$, because there is one synonymous and one non-synonymous substitution?

This genetic code has 61 sense (non-termination) codons

Substitution types

	Synonymous				Non-synonymous				To a stop codon
	Transitions	Transversions	Total		Transitions	Transversions	Total		Total
1st position:	8	0	8		140	26	166		9
2nd position:	0	0	0		148	28	176		7
3rd position:	58	68	126		2	48	50		7

Total	66	68	134		290	102	392		23

NEUTRAL EXPECTATION

- A random mutation is more likely to be non-synonymous than synonymous (~3 times more likely, depending on the variety of factors, such as codon composition, transition/transversion ratios etc)
- We need to estimate the proportion of random mutations that will be synonymous, and use that as the reference to compute dN and dS.
- Introduce the concept of synonymous and non-synonymous sites.

GAA (GLUTAMINE ACID)

SYNONYMOUS

GAG (GLUTAMINE ACID)

NON-SYNONYMOUS

AAA (LYSINE)

GCA (ALANINE)

CAA (GLUTAMINE)

GGA (GLYCINE)

TAA (STOP)

GTA (VALINE)

GAC (ASPARTIC ACID)

GAT (ASPARTIC ACID)

8/9 NON-SYNONYMOUS SITES

1/9 SYNONYMOUS SITES

GAA (GLUTAMINE ACID)

SYNONYMOUS

GAG (GLUTAMINE ACID)

NON-SYNONYMOUS

AAA (LYSINE)

GCA (ALANINE)

CAA (GLUTAMINE)

GGA (GLYCIN)

TAA (STOP)

GTA (VALINE)

GAC (ASPARTIC ACID)

GAT (ASPARTIC ACID)

8/9 NON-SYNONYMOUS SITES

1/9 SYNONYMOUS SITES

AMINOACID	CODONS	REDUNDANCY
ALANINE	GC*	4
CYSTEINE	TGC,TGT	2
ASPARTIC ACID	GAC,GAT	2
GLUTAMINE ACID	GAA,GAG	2
PHENYLALANINE	TTC,TTT	2
GLYCIN	GG*	4
HISTIDINE	CAC,CAT	2
ISOLEUCINE	ATA,ATC,ATT	3
LYSINE	AAA,AAG	2
LEUCINE	CT*,TTA,TTG	6
METHIONINE	ATG	1
ASPARGINE	AAC,AAT	2
PROLINE	CC*	4
GLUTAMINE	CAA,CAG	2
ARGININE	AGA,AGG,CG*	6
SERINE	AGC,AGT,TC*	6
THREONINE	AC*	4
VALINE	GT*	4
TRYPTOPHAN	TGG	1
TYROSINE	TAC,TAT	2
STOP	TAA,TAG,TGA	3

NEI-GOJOBORI DN/DS ESTIMATE

- For each codon **C**, define $ES(C)$ and $EN(C)$ - the fractions of synonymous and non-synonymous sites in a codon
 - e.g. $ES(GAA) = 1/9$, $EN(GAA) = 8/9$
 - Can also define them as fractions of substitutions that do not lead to stop codons, e.g. $ES(GAA) = 1/9$, $EN(GAA) = 7/9$
- The sum of ES and EN over all codons in a sequence gives an estimate of expected synonymous and non-synonymous sites in a sequence. For two sequences, average $ES(C)$ and $EN(C)$ at each site.
- ES/EN is the expected ratio of synonymous to non-synonymous substitutions under neutral evolution

ESTIMATING DS AND DN

- Consider two aligned sequences

ACA	ATA	AT C	TTT	AAT T	CAA
T	I	I	F	N	Q
ACA	ATA	AC C	TTT	AAC C	CAA
T	I	T	F	N	Q

- ES = ?, NS = ?

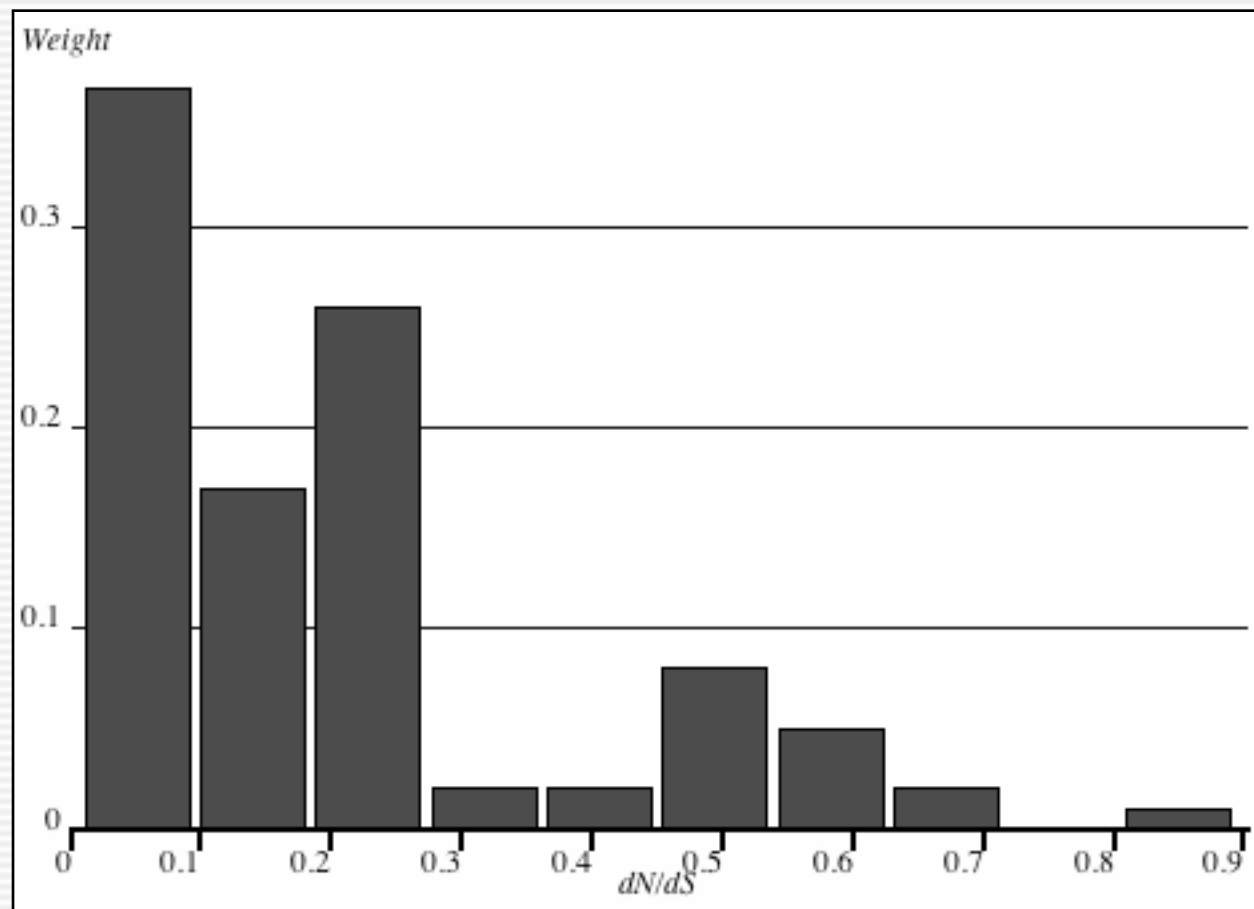
ESTIMATING dS AND dN

- Consider two aligned sequences

ACA	ATA	AT C	TTT	AAT T	CAA
T	I	I	F	N	Q
ACA	ATA	AC C	TTT	AAC C	CAA
T	I	T	F	N	Q

- $ES = ?$, $NS = ?$
- $ES = 3.5$, $NS = 14.1667$, $dN = 1/14.1667$, $dS = 1/3.5$, $dN/dS \sim 0.25$

BOOTSTRAPPED DISTRIBUTION OF dN/dS



Count = 100
Mean = 0.207385
Median = 0.166687
Variance = 0.0490168
Std.Dev = 0.221397
COV = 1.06757
Sum = 20.7385
Sq. sum = 9.15351
Skewness = 0.266313
Kurtosis = 33.381
Min = 0
2.5% = 0
97.5% = 0.741176
Max = 1

ESTIMATE dN/dS

- Compute ES, EN for the two sequences
- Count the number of synonymous substitutions S, and non-synonymous substitutions NS
 - How to deal with multiple substitutions?
- Define $dS = S/ES$, $dN = NS/EN$
- Compute dN/dS and compare to 1
- How to assess significance?

Substitutions = 6

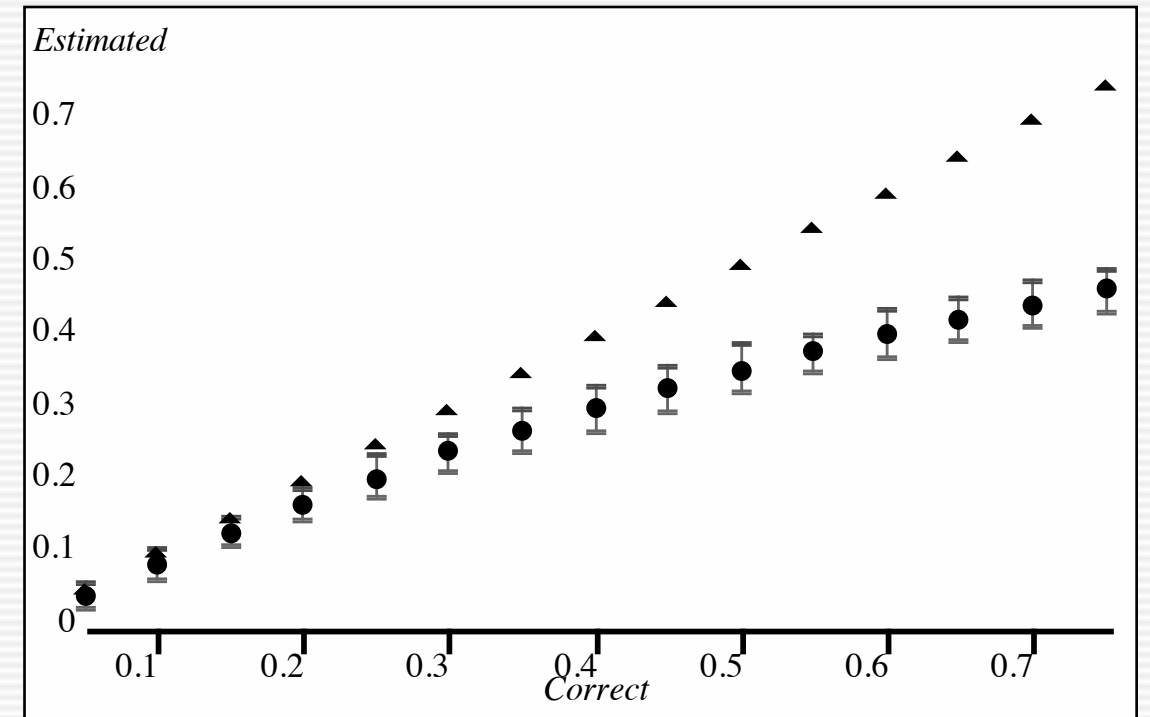
$$p = 0.4$$

HIGH DIVERGENCE

A	T	G	A	A	A	G	C	G	A
				T	C				
A	G	T	A	G	A	G	T	G	A

Multiple hits

Reversion



- Analogous to how p-distance underestimates true divergence due to multiple hits.
- Simulated 100 replicates of 1000 nucleotide long sequences for various divergence levels (substitutions/site)
- Plotted 'true' divergence vs that estimated by p-distance.
- Even for divergence of 0.25 (1/4 sites have mutation on average), p distance already significantly underestimates the true level: 0.2125 (0.19-0.241 95% range)
- Underestimation becomes progressively worse for larger divergence levels.

THE EFFECT OF PHYLOGENIES

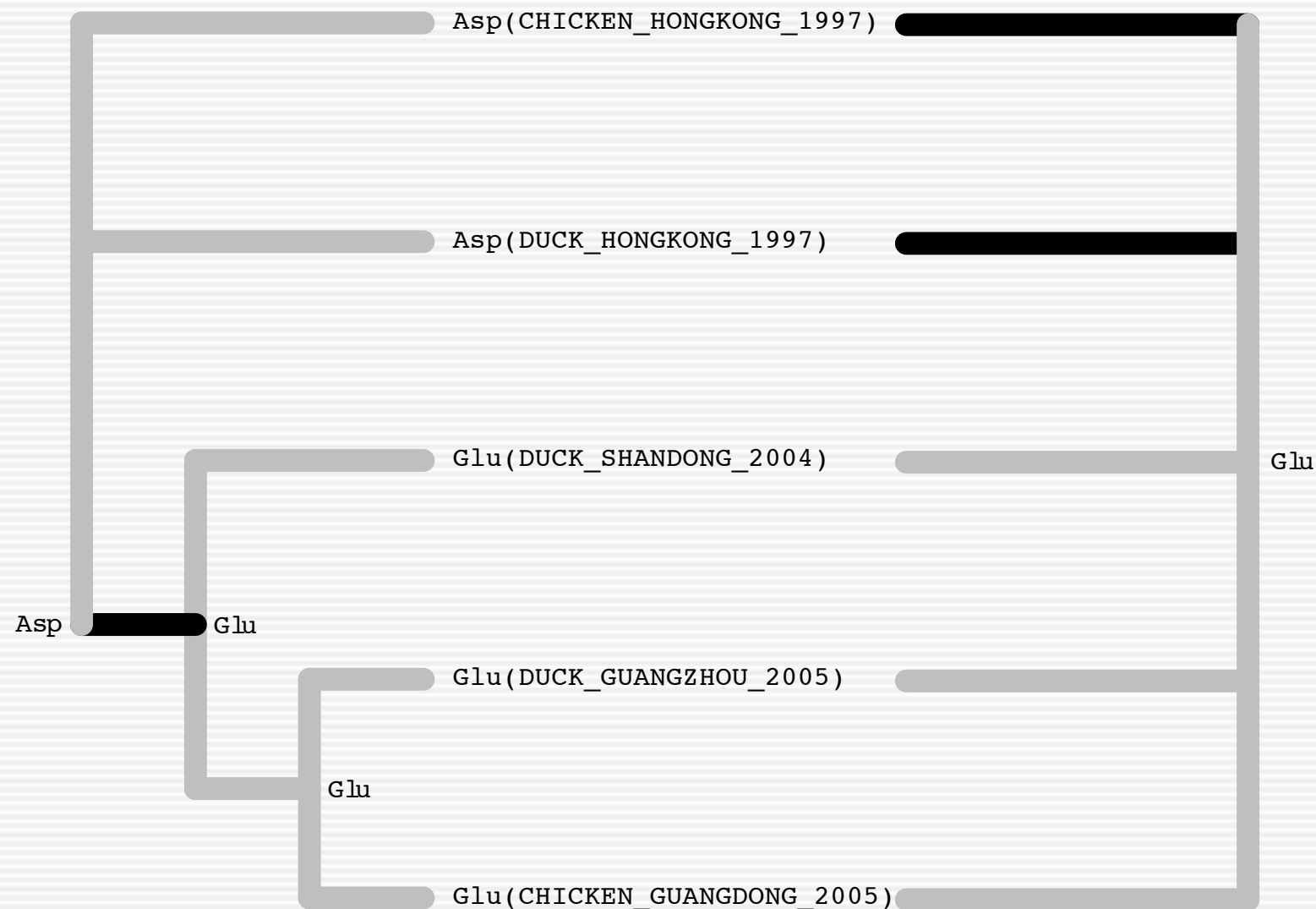
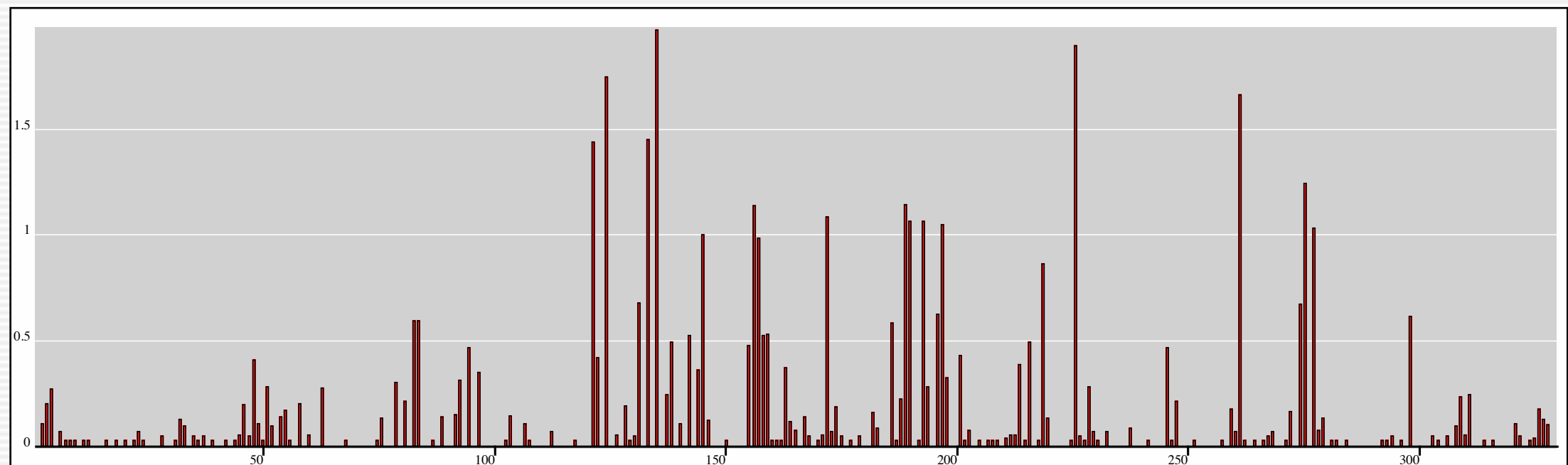


Fig. 1.1. Effect of phylogeny on estimating synonymous and nonsynonymous substitution counts in a dataset of Influenza A/H5N1 haemagglutinin sequences. Using the maximum likelihood tree on the left, the observed variation can be parsimoniously explained with one nonsynonymous substitution along the darker branch, whereas the star tree on the right involves at least two.

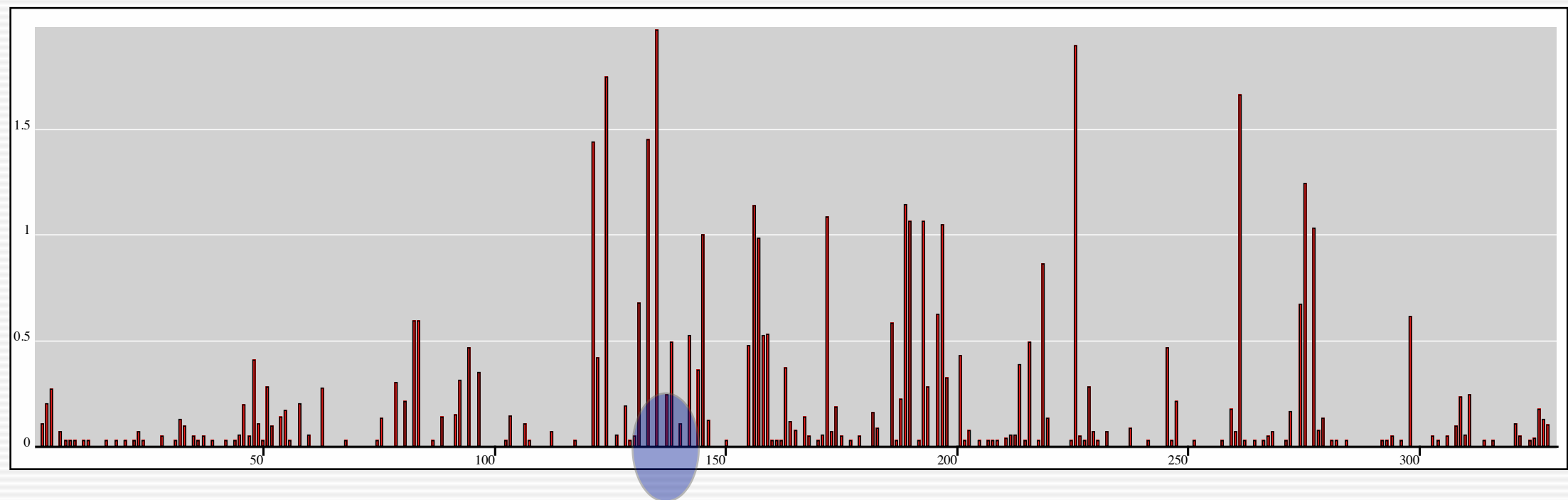
SELECTION IS VARIABLE ACROSS A GENE

- Different sites in a gene will be subject to different selective forces
- A “gene-wide” measure of selection is going to “average” these effects
- Most sites in most genes are negatively selected
- Positively selected sites are of great biological interest, because they point to **how** a particular gene can respond to selective pressures
- Must develop methods that are able to disentangle the contributions of individual sites

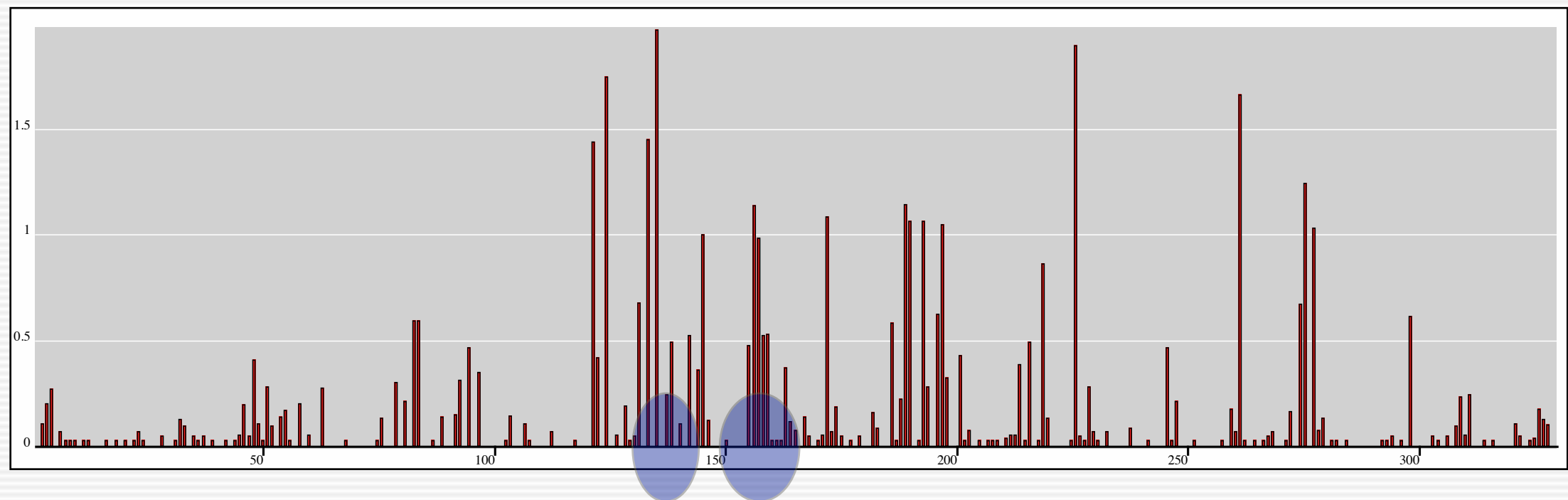
Influenza A hemagglutinin



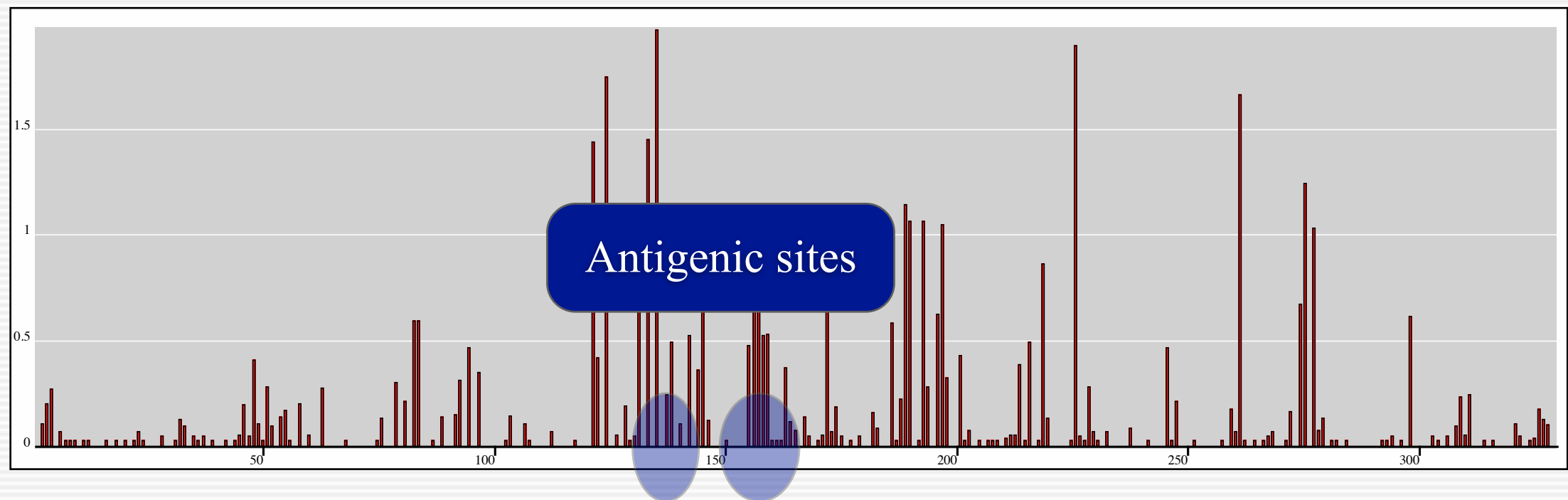
Influenza A hemagglutinin



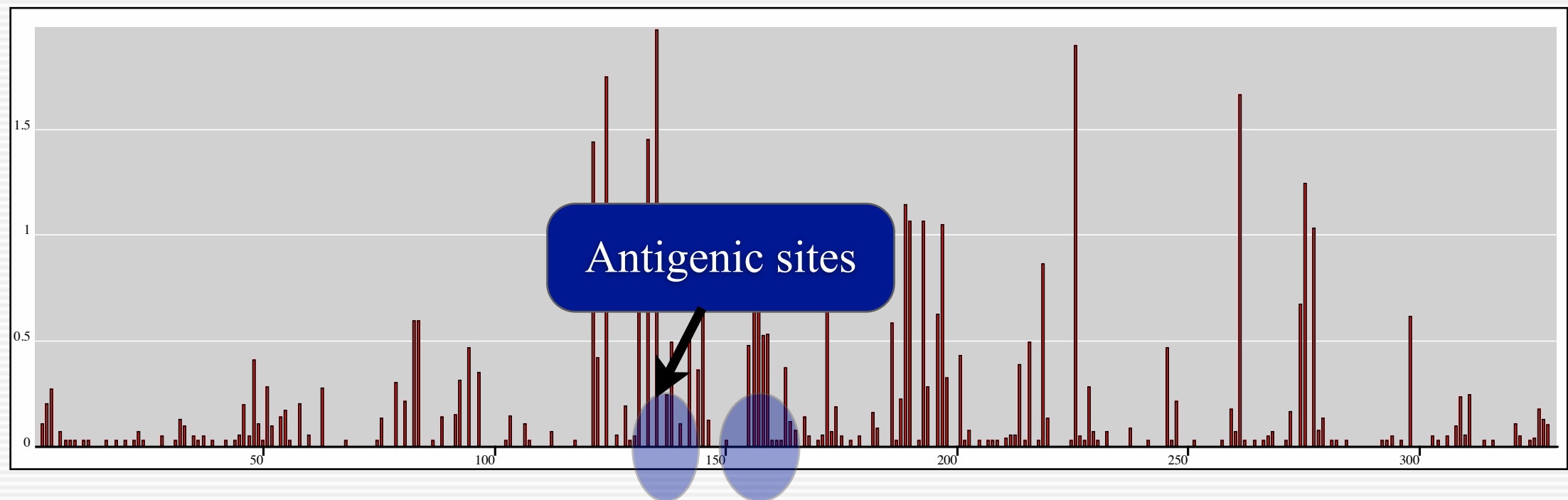
Influenza A hemagglutinin



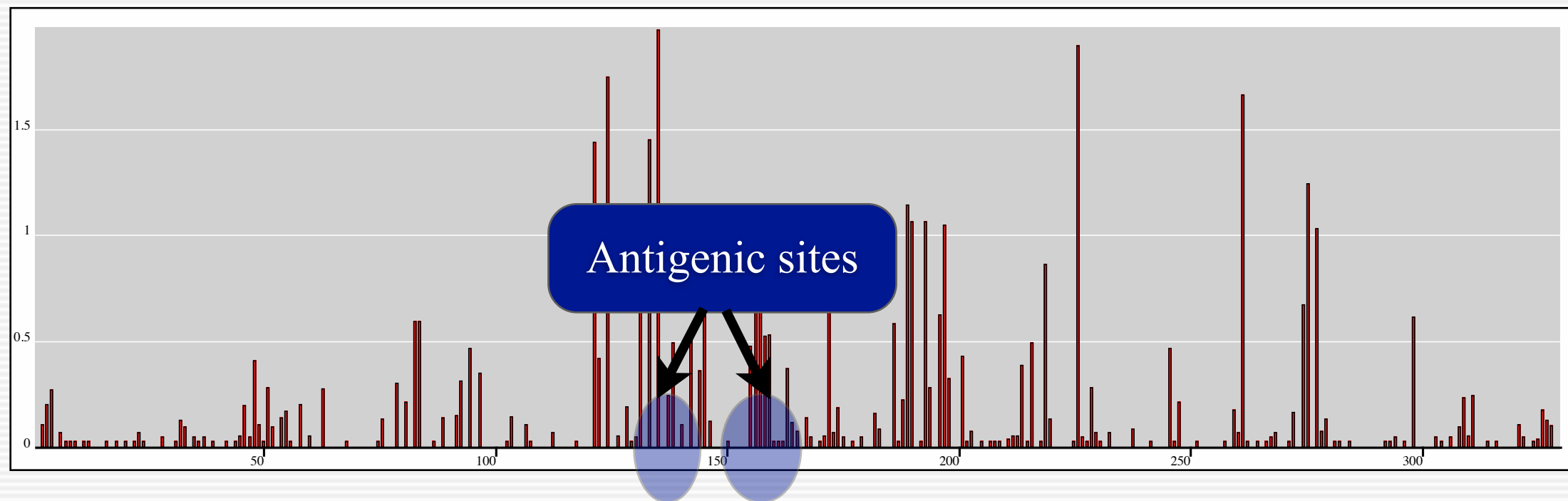
Influenza A hemagglutinin



Influenza A hemagglutinin



Influenza A hemagglutinin



SUZUKI GOJOBORI (SG)

- Uses the tree compute dN/dS at a given site
- First, reconstruct ancestral sequences by parsimony
- Second, compute EN and ES using labeled branches; define $p = ES/EN$
- Third, compute S and NS for each site
- Fourth, estimate the probability that the number S is unusually low (positive selection) or unusually high (negative selection).

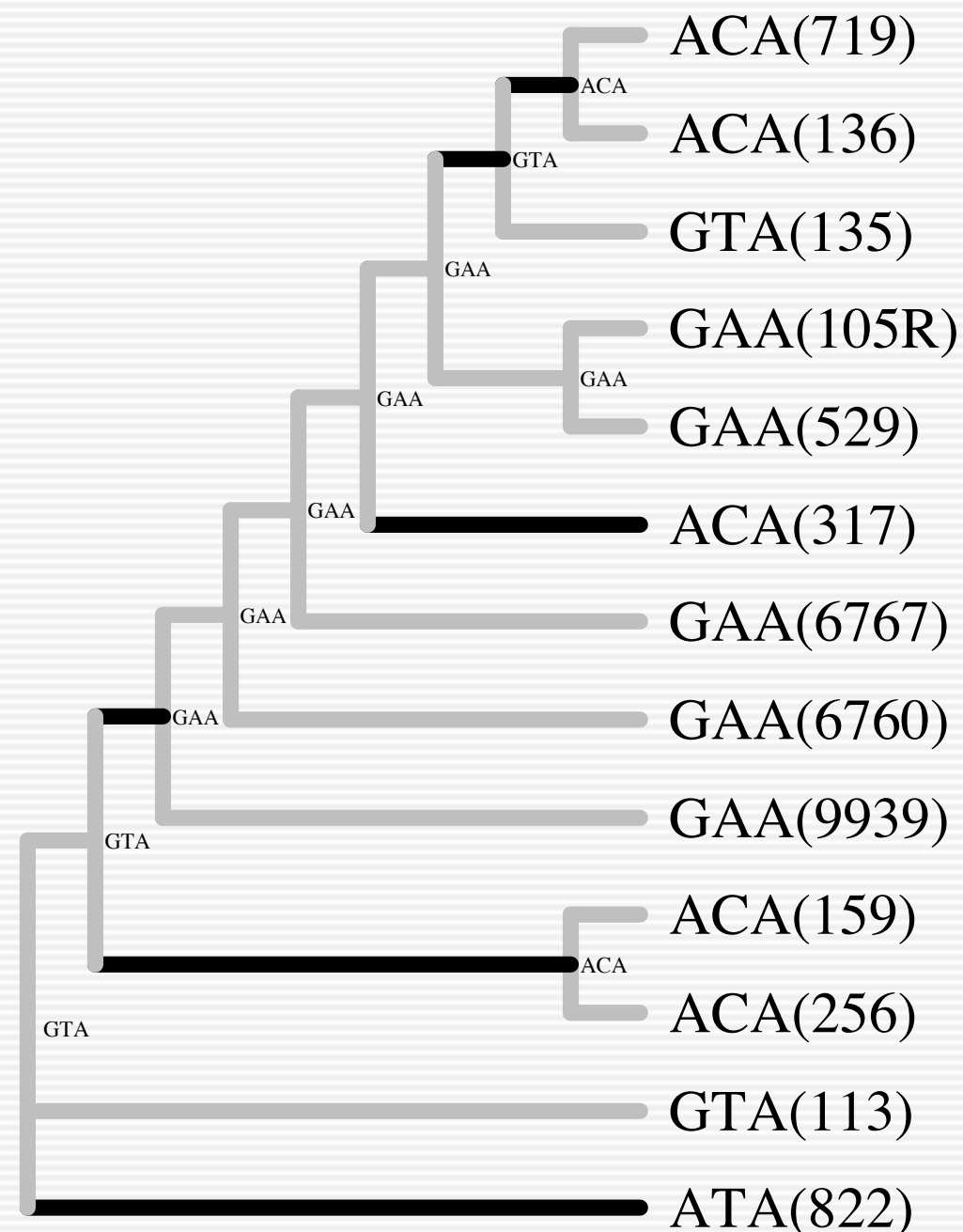


Fig. 1.6. An illustration of SLAC method, applied to a small HIV-1 envelope V3 loop alignment. Sequence names are shown in parentheses. Likelihood state ancestral reconstruction is shown at internal nodes. The parsimonious count yields 0 synonymous and 9 non-synonymous substitutions (highlighted with a dark shade) at that site. Based on the codon composition of the site and branch lengths (not shown), the expected proportion of synonymous substitutions is $p_e = 0.25$. An extended binomial distribution on 9 substitutions with the probability of success of 0.25, the probability of observing 0 synonymous substitutions is 0.07, hence the site is borderline significant for positive selection.

MODELS OF CODON EVOLUTION

- In 1994, first tractable evolutionary models for codon sequences were proposed.

$$(\text{Rate})_{X,Y}(dt) = \begin{cases} \alpha R_{xy} \pi_t dt & , \text{ one-step, synonymous substitution,} \\ \beta R_{xy} \pi_t dt & , \text{ one-step, non-synonymous substitution,} \\ 0 & , \text{ multi-step.} \end{cases}$$

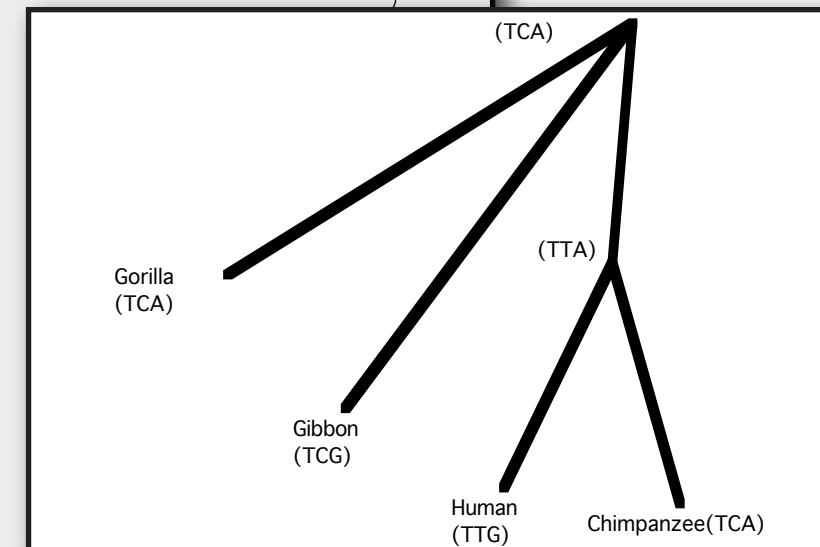
$X, Y = \text{AAA} \dots \text{TTT}$ (excluding stop codons),
 π_t - frequency of the target nucleotide.

Example substitutions:

$\text{AAC} \rightarrow \text{AAT}$ (one step, synonymous - Asparagine)

$\text{CAC} \rightarrow \text{GAC}$ (one step, non-synonymous - Histidine to Aspartic Acid)

$\text{AAC} \rightarrow \text{GTC}$ (multi-step).



αR_{CT}

βR_{CG}

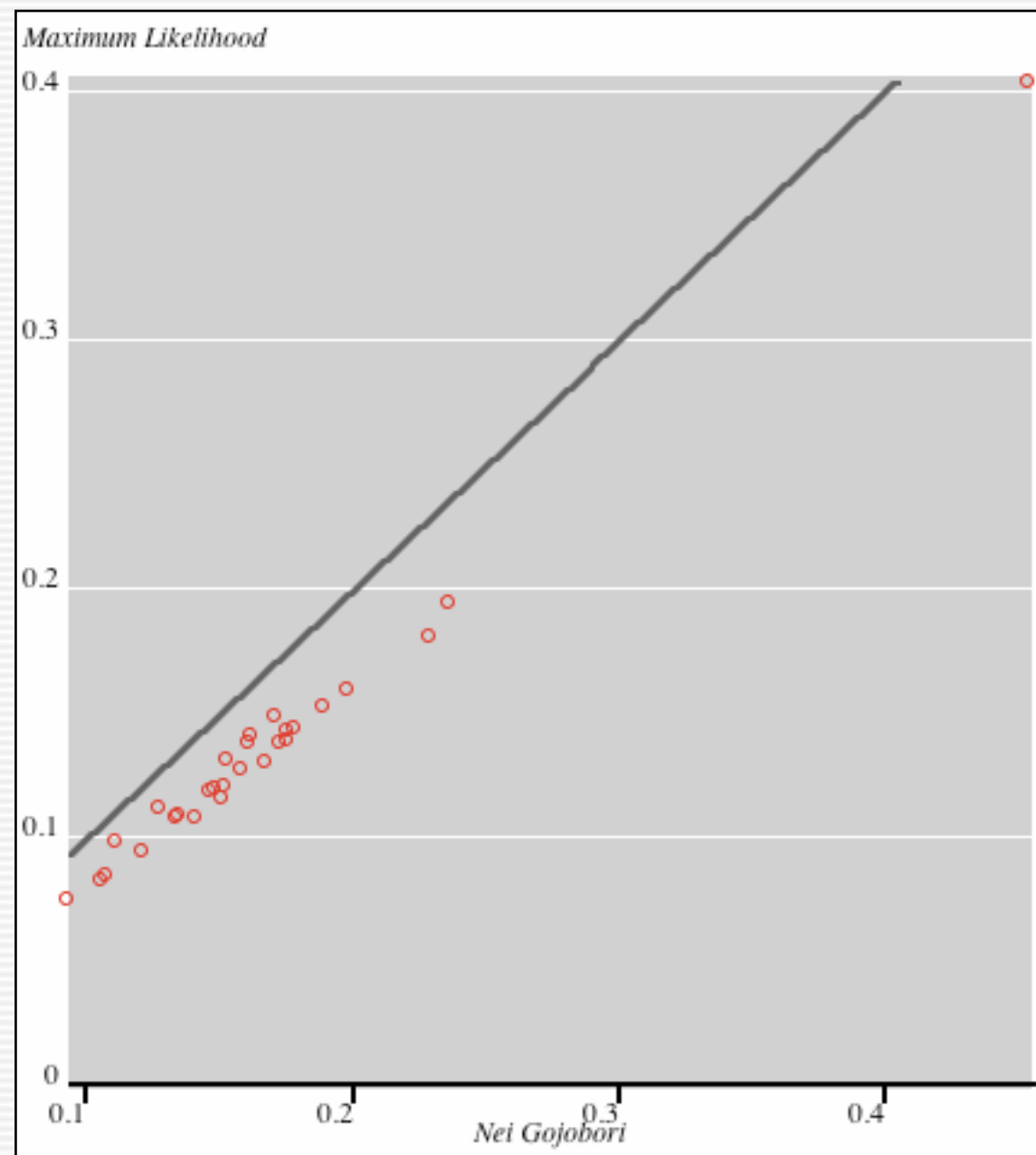
COMPUTING THE TRANSITION PROBABILITIES

- In order to recover transition probabilities $T(t)$ from the rate matrix Q , one computes the matrix exponential $T(t) = \exp Qt$
- Because the computational complexity of matrix exponentiation scales as the cube of the matrix dimension, codon based models require roughly $(61/4)^3 \approx 3500$ more operations than nucleotide models
- This explains why codon probabilistic models were not introduced until the 1990s.

MULTIPLE SUBSTITUTIONS

- The model assumes that point mutations alter one nucleotide at a time, hence most of the instantaneous rates ($3134/3761$ or 84.2% in the case of the universal genetic code) are **0**.
- This restriction, however, does not mean that the model disallows any substitutions that involve multiple nucleotides (e.g. $ACT \rightarrow AGG$).
- Such substitutions must simply be realized via several single nucleotide steps, e.g. $ACT \rightarrow AGT \rightarrow AGG$
- In fact the (i,j) element of $T(t) = \exp(Qt)$ sums the probabilities of all such possible pathways of length **t**, including reversions
- Compare this to the rather naive NG parsimony approach

NG VS A CODON MODEL

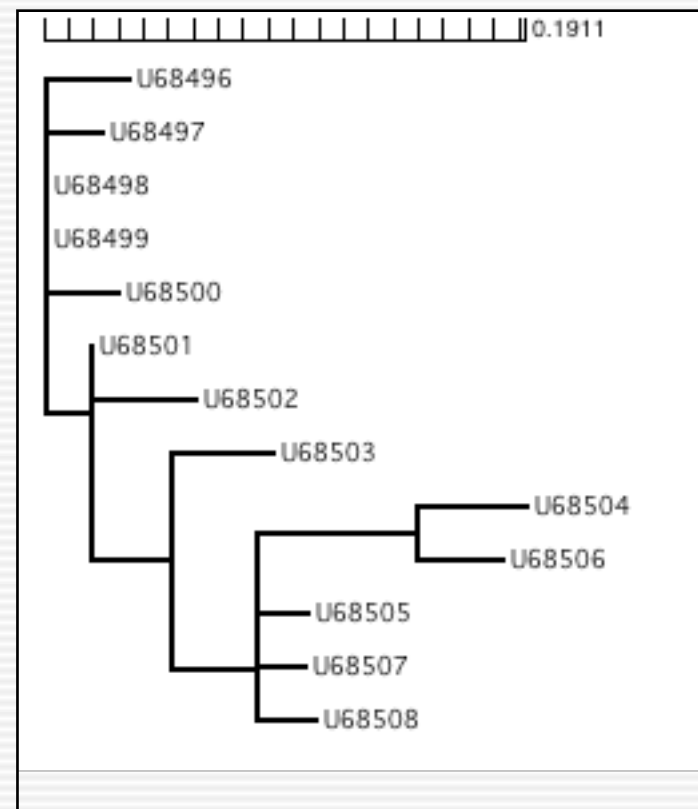
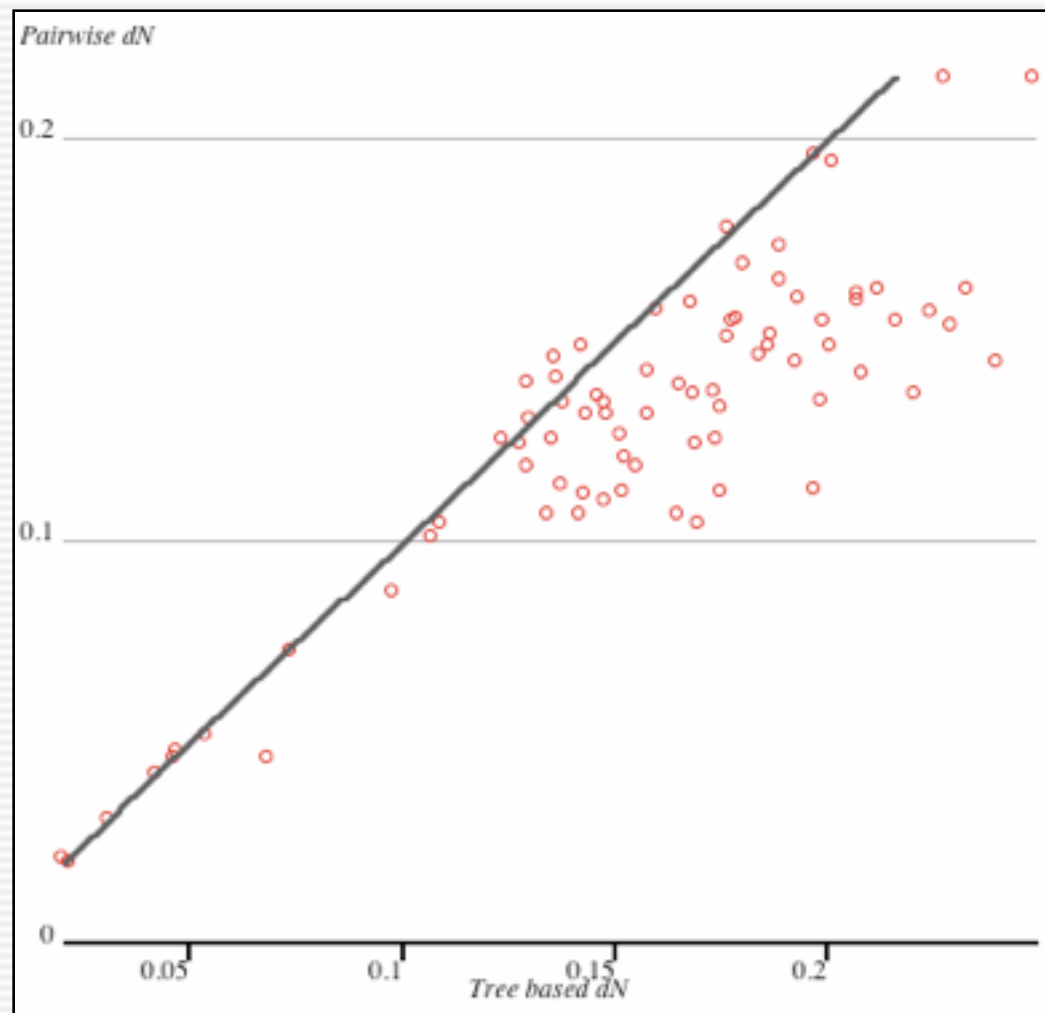


PAIRWISE dN/dS ESTIMATES ON HIV-1 P51 (RT) SEQUENCES

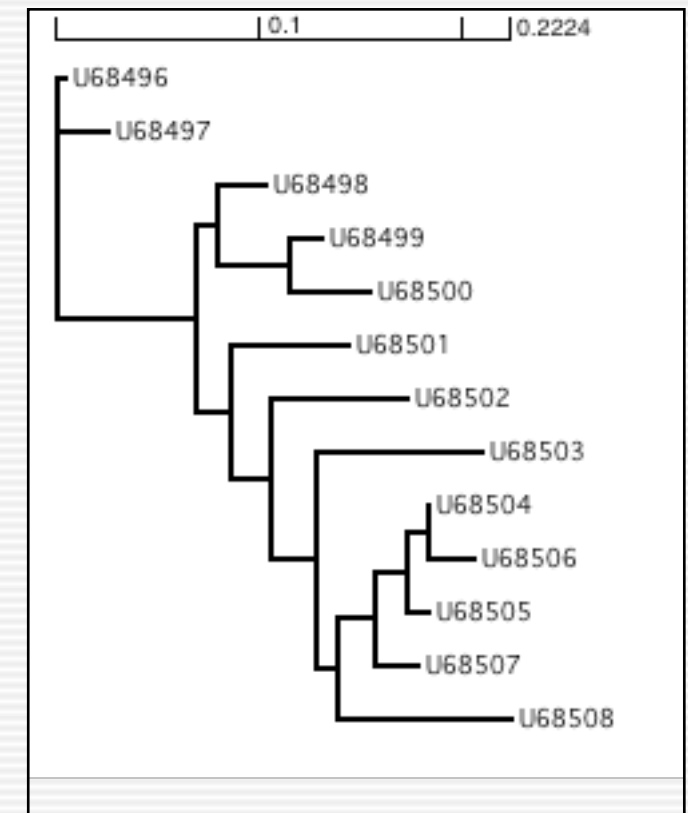
ALIGNMENT-WIDE ESTIMATES

- Using standard MLE approaches it is straightforward to obtain point estimates of $dN/dS := \beta/\alpha$
- Can also easily test whether or not $dN/dS > 1$, or < 1 using the likelihood ratio test
- Codon models also support the concepts of synonymous and non-synonymous distances between sequences using standard properties of Markov processes (exponentially distributed waiting times)
- $$E[subs] = - \sum_i \pi_i \hat{q}_{ii}, \quad E[subs] = E[syn] + E[nonsyn] = - \sum_i \pi_i \hat{q}_{ii}^s - \sum_i \pi_i \hat{q}_{ii}^{ns}.$$

TOPOLOGY-BASED VS PAIRWISE



DS



DN

- Pairwise distance calculations do not account for shared ancestry and can bias dN and dS estimates (usually UNDERestimate)

SITE TO SITE RATE VARIATION

- Global α and β models are the simplest computationally and contain the fewest number of estimable parameters, hence they are suitable for
 - coarse data characterization (exploratory analysis),
 - analyses of small samples (a few sequences or very short alignments)
 - or when substitution rates are a nuisance parameter (i.e. they are used as a means to estimate something else, e.g. phylogeny or ancestral sequences)

SITE TO SITE RATE VARIATION

- When α and β are the primary object of evolutionary analysis, the global model is nearly always a poor choice for inference.
- Selective regimes may differ from site to site in a gene due to varying functional and structural constraints, selective pressures and other biological processes.
- Since the global model is only able to estimate the mean, it reveals remarkably little about the unknown distribution of rates.
- Two genes with mean $\omega = 0.5$ may, for example, have dramatically different distributions of ω across sites, hence it might be erroneous to state, based solely on the equality of the means, that two genes are evolving under similar selective pressures.

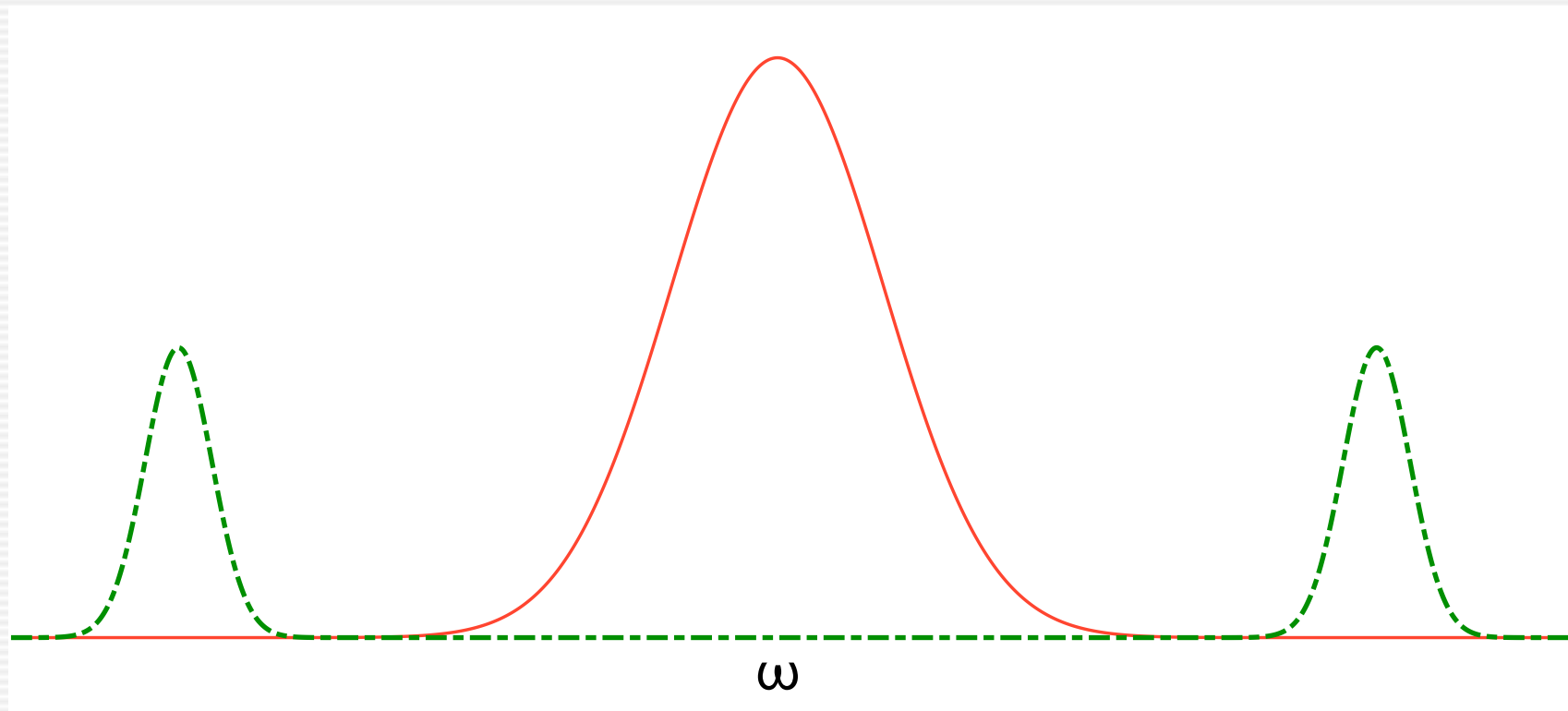


Fig. 1.2. Two different rate distributions (solid and dashed) which have the same mean ω .

STEP 1. FITTING GLOBAL MG94xHKY85 TO OBSERVED DATA.

DONE. ESTIMATED MEAN dN/dS = 0.23 AND TS/TV RATIO = 6.69416

STEP 2A. SIMULATING UNDER A POINT MASS DISTRIBUTION (GLOBAL MODEL) WITH dN/dS = 0.230856

STEP 2B. FITTING A GLOBAL MODEL...

DONE. ESTIMATED MEAN dN/dS = 0.26 AND TS/TV RATIO = 7.18668

STEP 3A. SIMULATING UNDER A BIMODAL DISTRIBUTION WITH 50% AT 0.129985 AND 50% AT 0.389954

STEP 3B. FITTING A GLOBAL MODEL...

DONE. ESTIMATED MEAN dN/dS = 0.26 AND TS/TV RATIO = 6.9184

STEP 4A. SIMULATING UNDER A TRIMODAL DISTRIBUTION WITH 60% AT 0.0262829, 30% AT 0.262829 AND 10% AT 1.68211

STEP 4B. FITTING A GLOBAL MODEL...

DONE. ESTIMATED MEAN dN/dS = 0.25 AND TS/TV RATIO = 5.52756

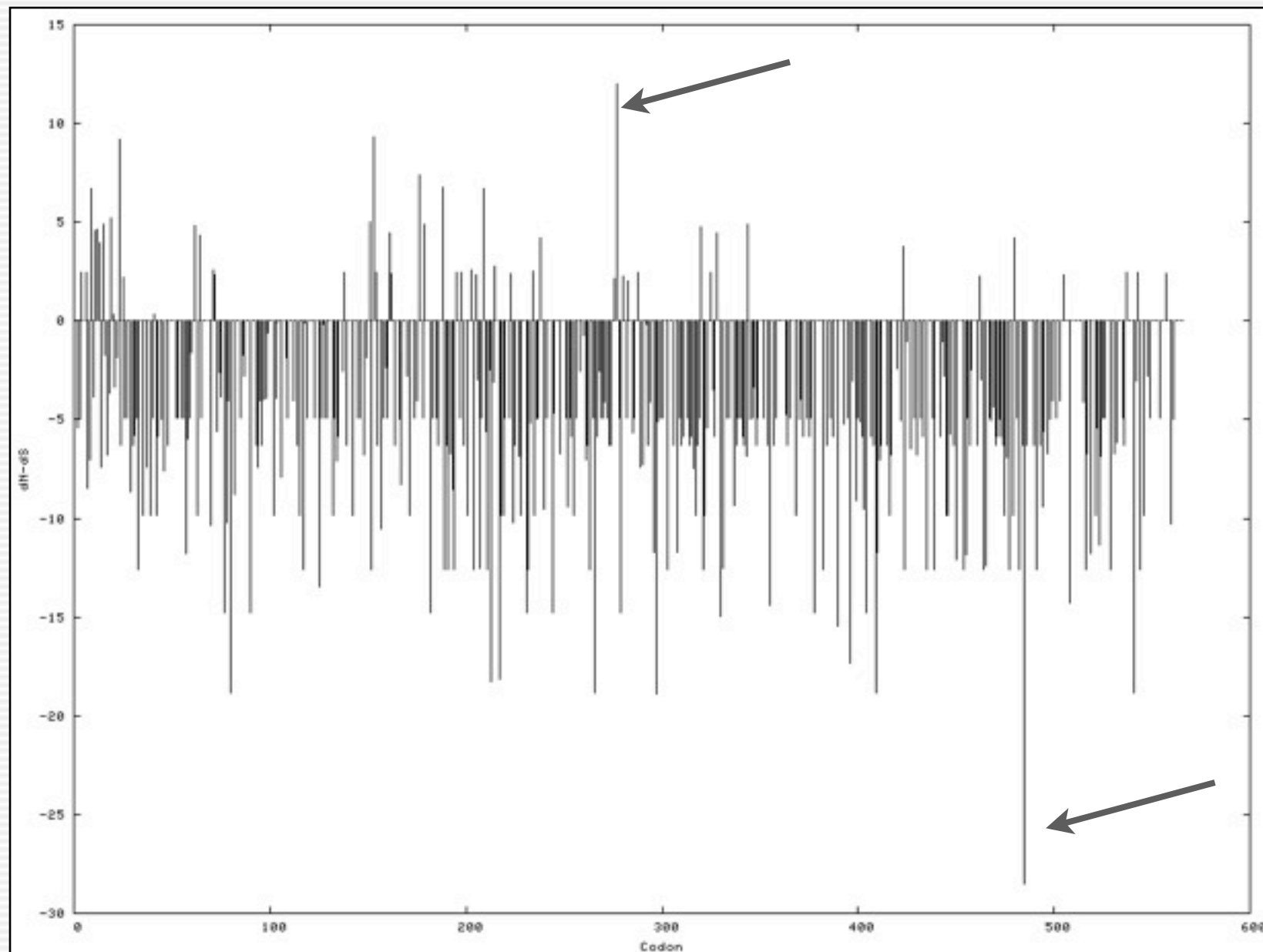
STEP 5A. SIMULATING UNDER A GAMMA DISTRIBUTION WITH VARIANCE 0.600982

STEP 5B. FITTING A GLOBAL MODEL...

DONE. ESTIMATED MEAN dN/dS = 0.17 AND TS/TV RATIO = 6.90859

SELECTION: WHERE?

- Locate codon sites which evolve non-neutrally.
- Provide confidence estimates (p-values, etc)



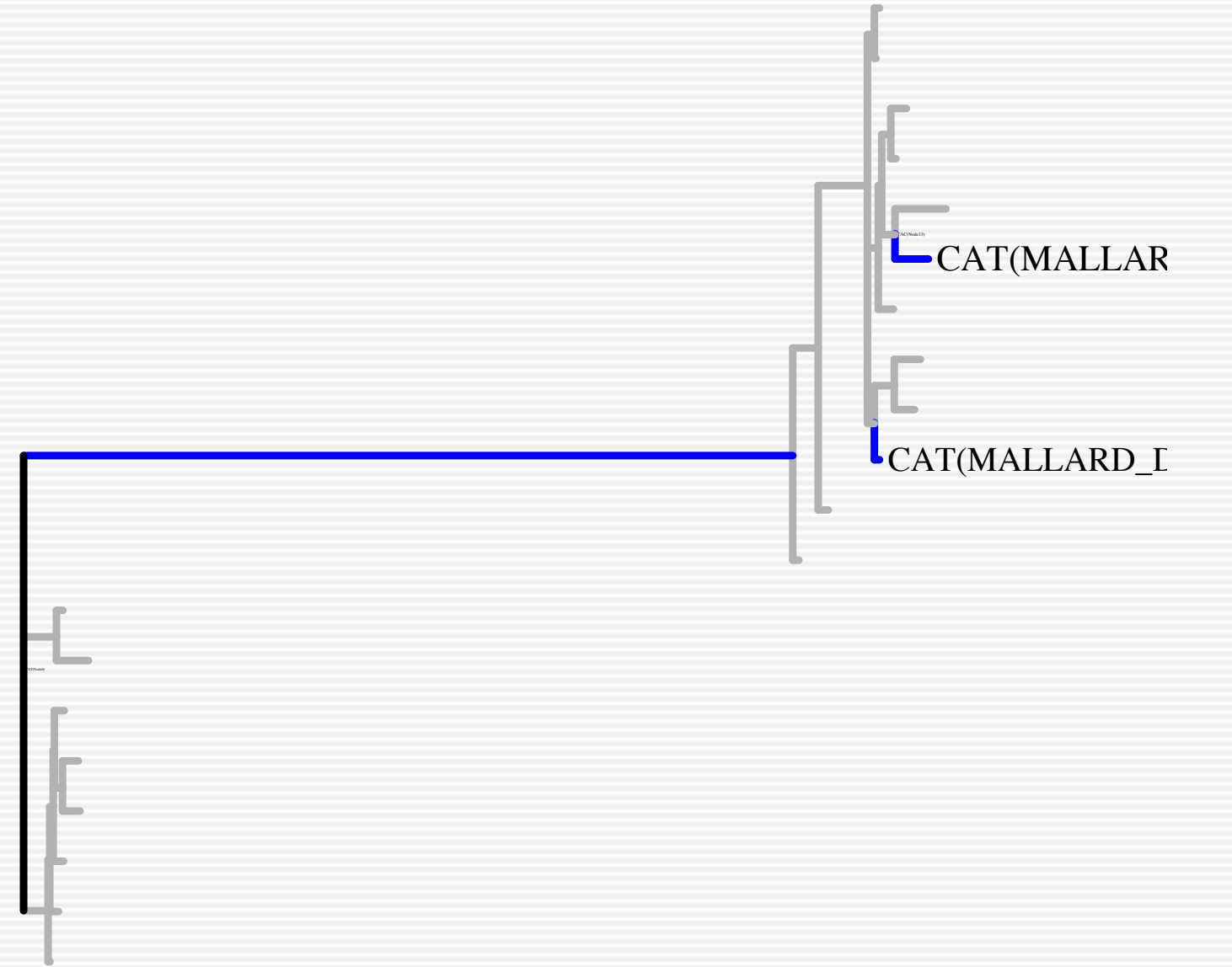
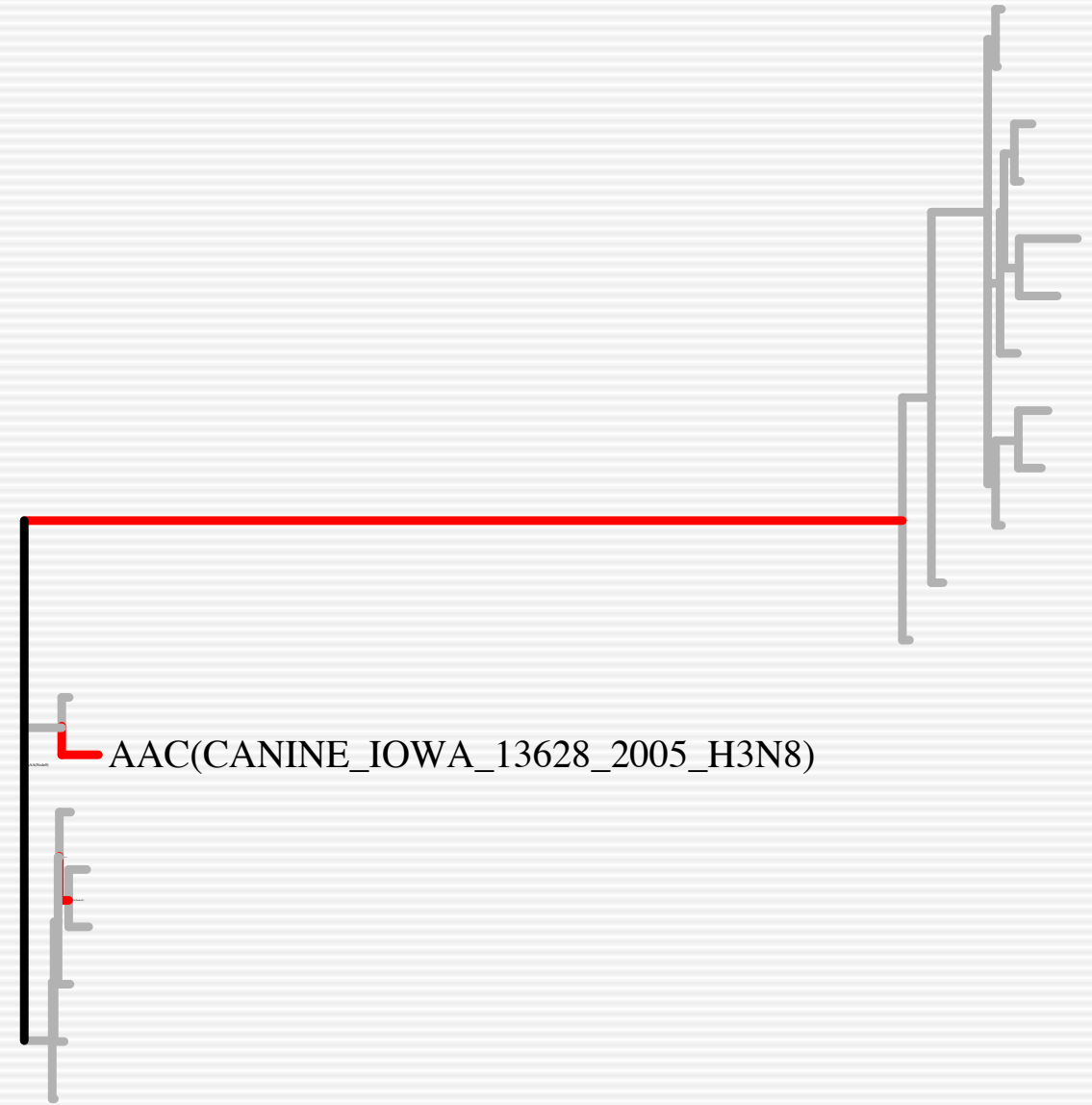
INFLUENZA H₃N*
MEAN DN/DS = 0.101841
TREMENDOUS VARIATION
ACROSS SITES

SINGLE LIKELIHOOD ANCESTOR COUNTING (SLAC)

- ☐ Use a simple codon model and maximum likelihood to reconstruct the evolutionary history of the sample
- ☐ Compute dN and dS based on ML ancestral reconstructions (best, mean, sampled) and compare to “neutral” expectation
- ☐ Very fast (100s sequences, in minutes on a desktop)
- ☐ Least powerful

0.1

0.1



P ~ 0.02

P ~ 0.15

Branch	From		To		Substitutions	
	Codon	AA	Codon	AA	Synonymous	Non-synonym
Node1	CAT	His	CAC	His	1.0	0.0
MALLARD_ALB_394_1988_H3N3	CAC	His	CAT	His	1.0	0.0
MALLARD_DUCK_ALB_525_1985_H3N8	CAC	His	CAT	His	1.0	0.0

Branch	From		To		Substitutions	
	Codon	AA	Codon	AA	Synonymous	Non-synonymous
Node1	AAA	Lys	CGC	Arg	0.0	3.0
Node32	AAA	Lys	AGA	Arg	0.0	1.0
CANINE_IOWA_13628_2005_H3N8	AAA	Lys	AAC	Asn	0.0	1.0

FIXED EFFECTS LIKELIHOOD (FEL)

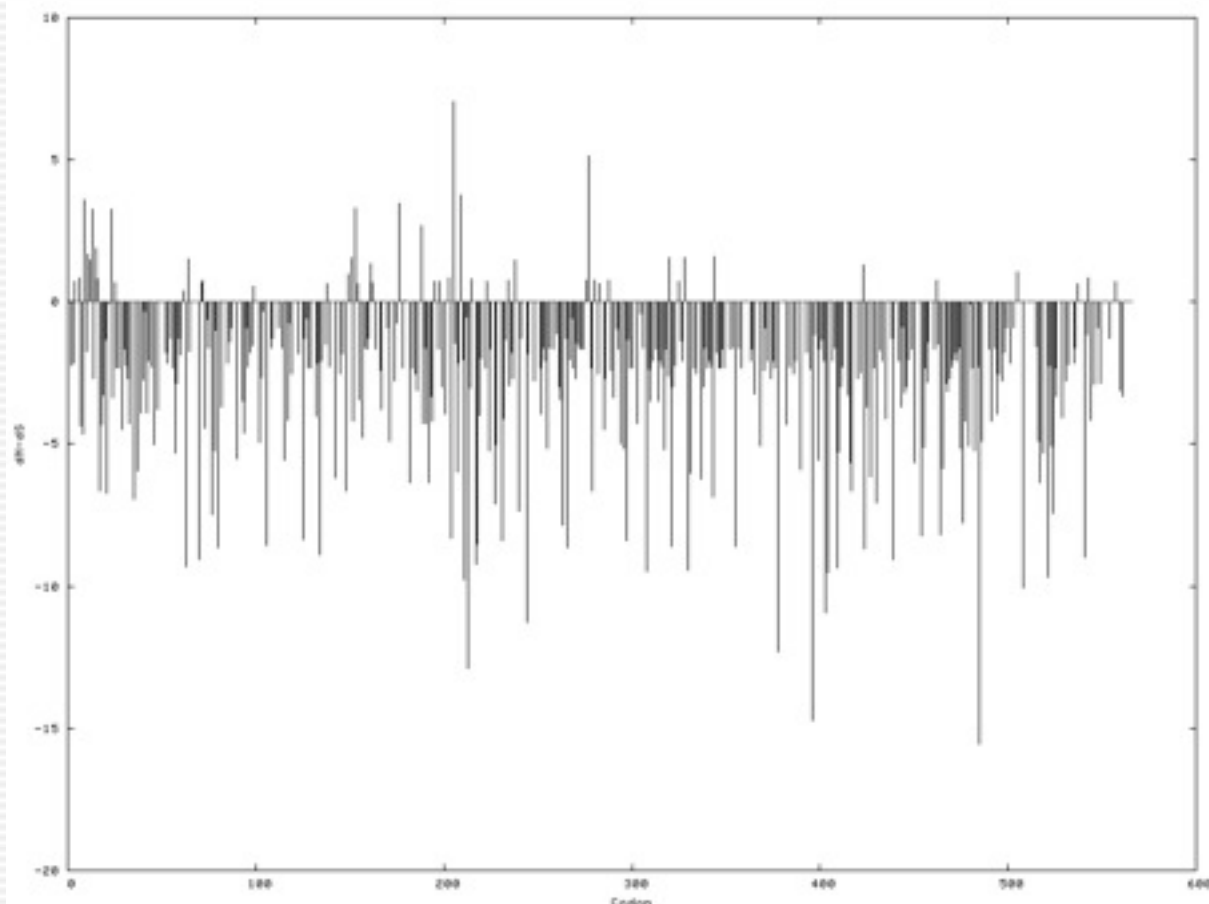
- Estimate common parameters (branch lengths, nucleotide biases) from the entire alignment
- Fit dN and dS individually to every site using maximum likelihood
- Examine the hypothesis of neutrality by a likelihood ratio test (dN=dS vs dN≠dS)
- Can be easily parallelized. 100s sequences in a few minutes on an MPI cluster

Found **2** positively selected sites (significance level)

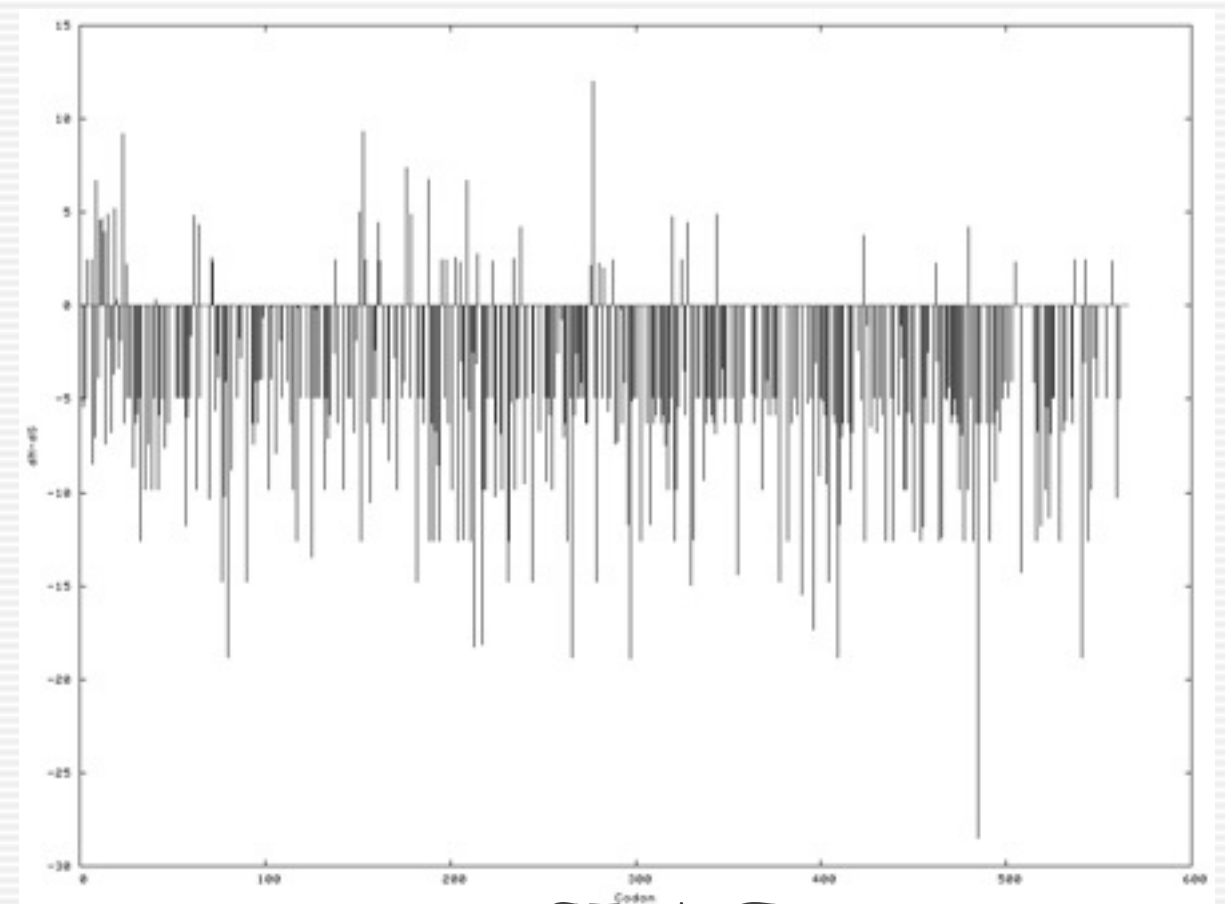
Codon	dS	dN	dN/dS	Normalized dN-dS	p-value	Additional Information
23	0	1.72616	Infinite	3.2581	0.0888338	[Codons][AA][Counts]
176	0	1.84806	Infinite	3.48818	0.0602959	[Codons][AA][Counts]

Found **197** negatively selected sites (0.1 significance level)

Codon	dS	dN	dN/dS	Normalized dN-dS	p-value	Additional Information
3	1.13671	0	0.000	-2.14552	0.0899295	[Codons][AA][Counts]
7	2.31472	1.11022e-16	0.000	-4.36899	0.0419049	[Codons][AA][Counts]
17	3.51811	0	0.000	-6.64037	0.0614126	[Codons][AA][Counts]



FEL



SLAC

RANDOM EFFECTS LIKELIHOOD (REL)

- When the objective of an analysis is to estimate, for each codon $c = 1 \dots S$ in the alignment, a pair of **unobserved** rates α_s and β_s , the random effects approach posits that there exists a distribution of rates (assumed discrete with D categories for computational feasibility), with values (α_d, β_d) , and the probability of drawing each pair of values is (p_d)
- To compute the likelihood function at codon site c , one now has to compute an expectation over the distribution of rates

$$L(\text{site } c) = \sum_{d=1}^D L(\text{site } c | \alpha_c = \alpha_d, \beta_c = \beta_d) p^d,$$

Synonymous rates α	Non-synonymous rates β	Likelihood Expression
Constant	Constant	$L_{\text{data}} \left(\mathcal{D} \mathcal{T}, \vec{\theta}, \alpha, \beta \right) = \prod_{i=1}^N L_{\text{site}} \left(\mathcal{D}_i \mathcal{T}, \vec{\theta}, \alpha, \beta \right)$
Constant	Drawn from a univariate density $F(r; B)$	$L_{\text{data}} \left(\mathcal{D} \mathcal{T}, \vec{\theta}, \alpha, B \right) = \prod_{i=1}^N \int L_{\text{site}} \left(\mathcal{D}_i \mathcal{T}, \vec{\theta}, \alpha, \beta = r \right) F(r; B) dr$
Drawn from a bivariate density $F(r, s; B)$		$L_{\text{data}} \left(\mathcal{D} \mathcal{T}, \vec{\theta}, B \right) = \prod_{i=1}^N \int \int L_{\text{site}} \left(\mathcal{D}_i \mathcal{T}, \vec{\theta}, \alpha = s, \beta = r \right) F(r, s; B) dr ds$
Independently estimated for each site		$L_{\text{data}} \left(\mathcal{D} \mathcal{T}, \vec{\theta}, \vec{\alpha}, \vec{\beta} \right) = \prod_{i=1}^N L_{\text{site}} \left(\mathcal{D}_i \mathcal{T}, \vec{\theta}, \alpha_i, \beta_i \right)$

SITES UNDER SELECTION

- To find individual codon sites under positive selective pressure, REL methods can use an empirical Bayes approach, whereby the posterior probability of a rate class at every site is computed
- A simple application of Bayes rule, treating the inferred distribution of rates as a prior shows that

$$Pr(\alpha_s = \alpha_d, \beta_s = \beta_d | \text{site } c) = \frac{L(\text{site } c | \alpha_s = \alpha_d, \beta_s = \beta_d) p^d}{L(\text{site } c)}.$$

- A site can be classified as selected if the posterior probabilities for all those rate classes with $\beta > \alpha$ exceed a fixed threshold (e.g. 0.95)
- The Bayesian analog of a p-value is the Bayes factor for some event E, defined as

$$BF(E) = \frac{\text{posterior odds } E}{\text{prior odds } E} = \frac{Pr_{\text{posterior}}(E)/(1 - Pr_{\text{posterior}}(E))}{Pr_{\text{prior}}(E)/(1 - Pr_{\text{prior}}(E))}.$$

- If $BF(\beta > \alpha)$ at site c exceeds some predefined value (e.g. 20), a site can be called selected.

Inferred rate distribution

Rate class has $dN > dS$ Rate class has $dN < dS$

Rate Class	1	2	3	4	5	6	7	8	9	Summary
dS	1.08	1.08	1.08	0.00	0.00	0.00	0.00	0.00	0.00	
dN	0.01	0.03	0.49	0.01	0.01	0.03	0.03	0.49	0.49	
dN-dS	-1.06	-1.04	-0.59	0.01	0.01	0.03	0.03	0.49	0.49	Mean dN-dS: -0.893
Prob.	0.521	0.234	0.173	0.000	0.040	0.000	0.018	0.000	0.013	Std.Dev : 0.333

