

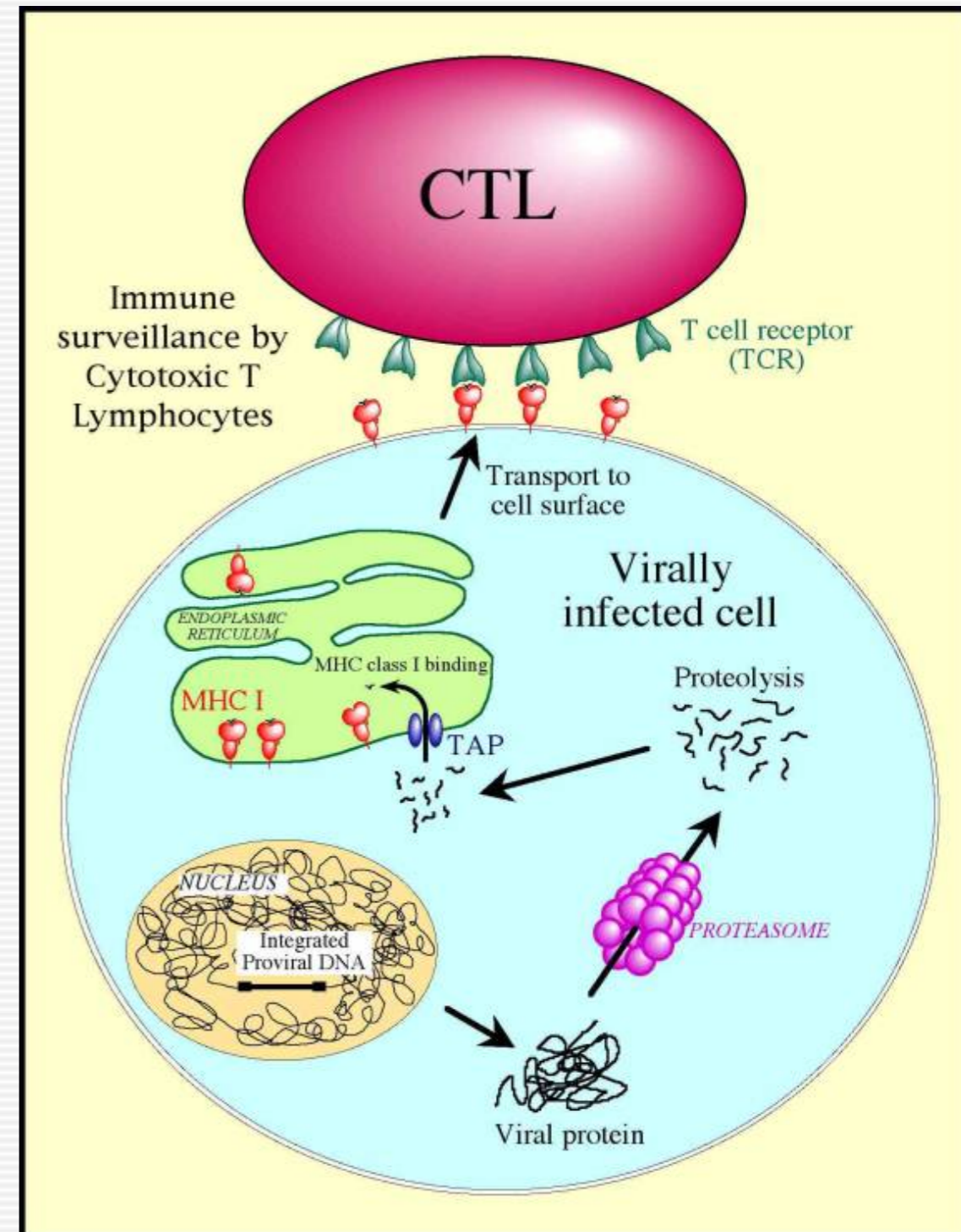
MOLECULAR EVOLUTION: NATURAL SELECTION

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

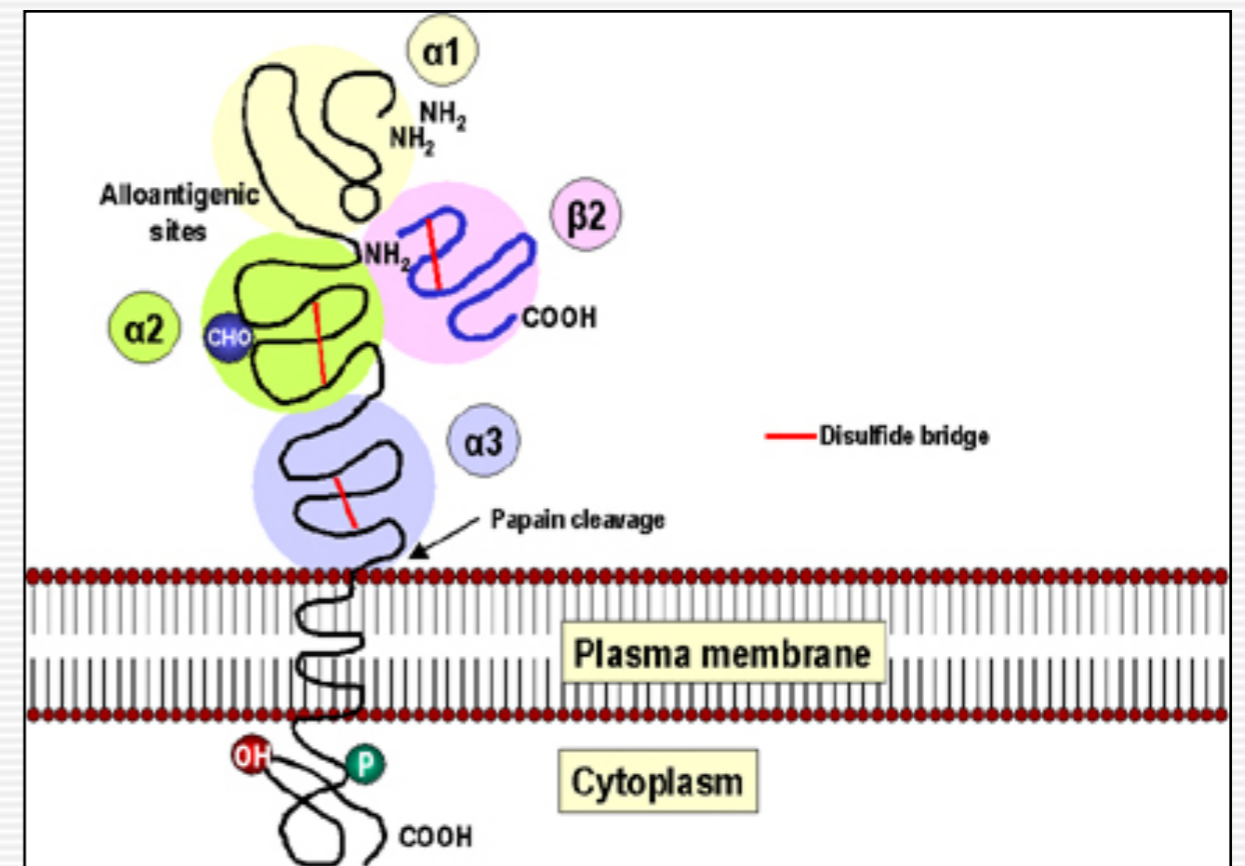
- Cytotoxic T-Lymphocytes effect cell-mediated immune response
- Viral proteins are cleaved by the proteasome, transported by TAP and loaded onto the MHC Class I molecule.
- MHC Class I presents a restricted polypeptide (epitope) on the surface of the cell.
- A CD8+ cell binds to presented foreign peptides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



MHC CLASS I MOLECULES

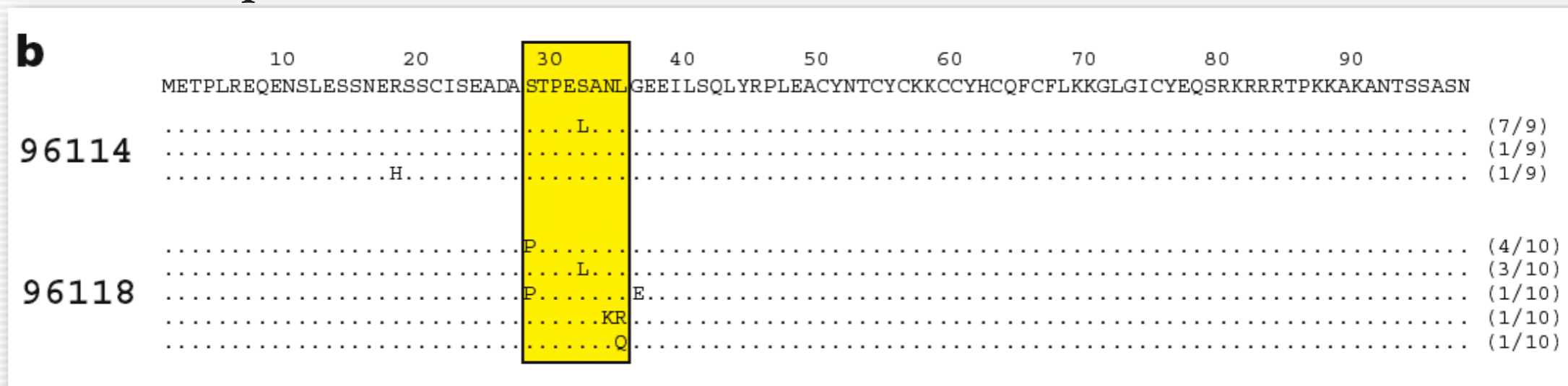
- Presents foreign peptides which are 9-10 amino-acid long
- Anchor sites (2 and 9) are usually important for binding and recognition
- Mutations which alter the peptide can hinder or prevent CTL response activation

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”



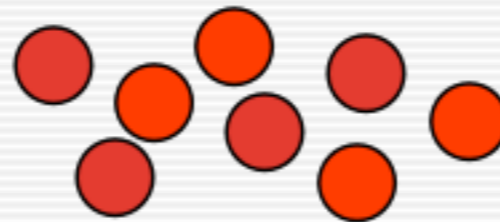
Before selection



After selection



Final population



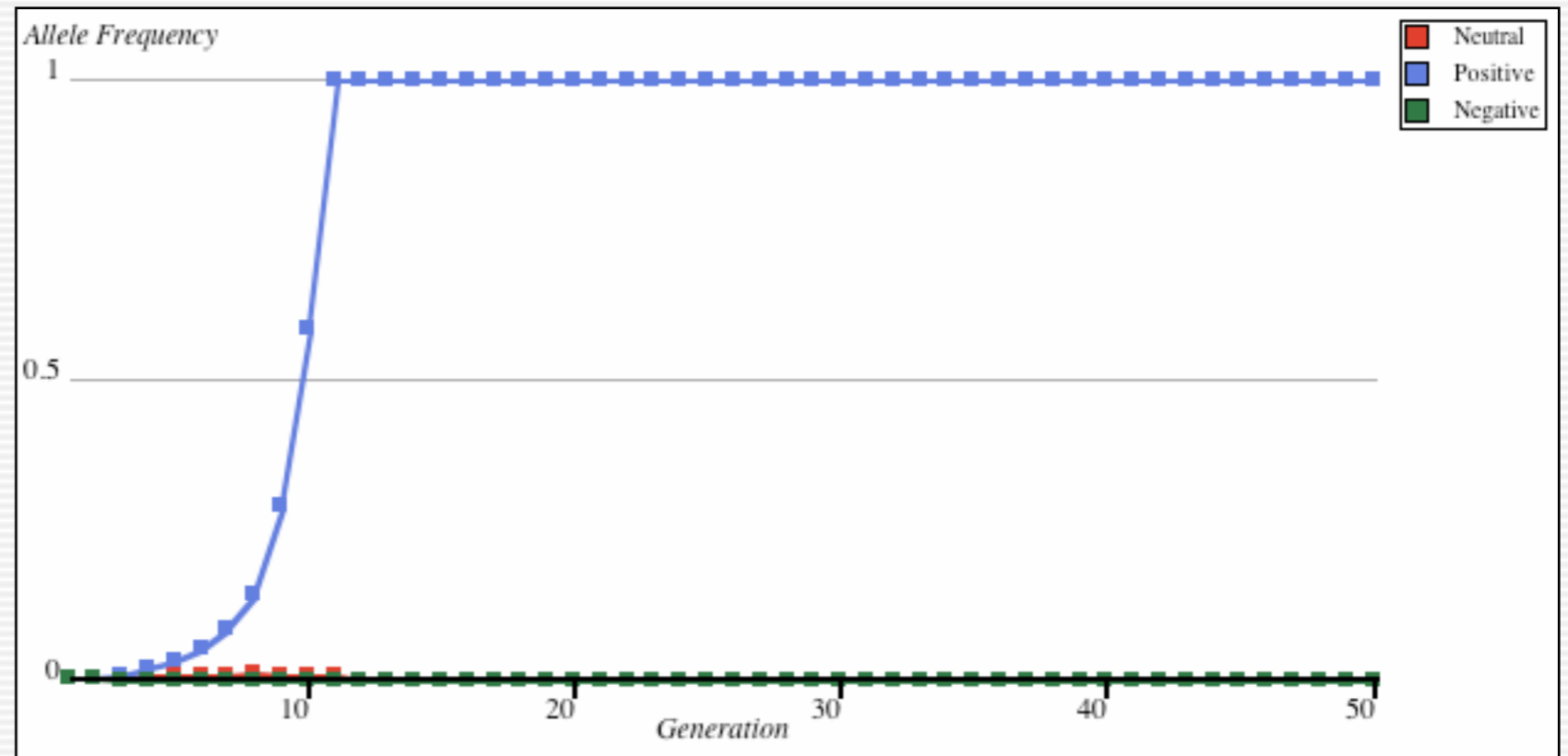
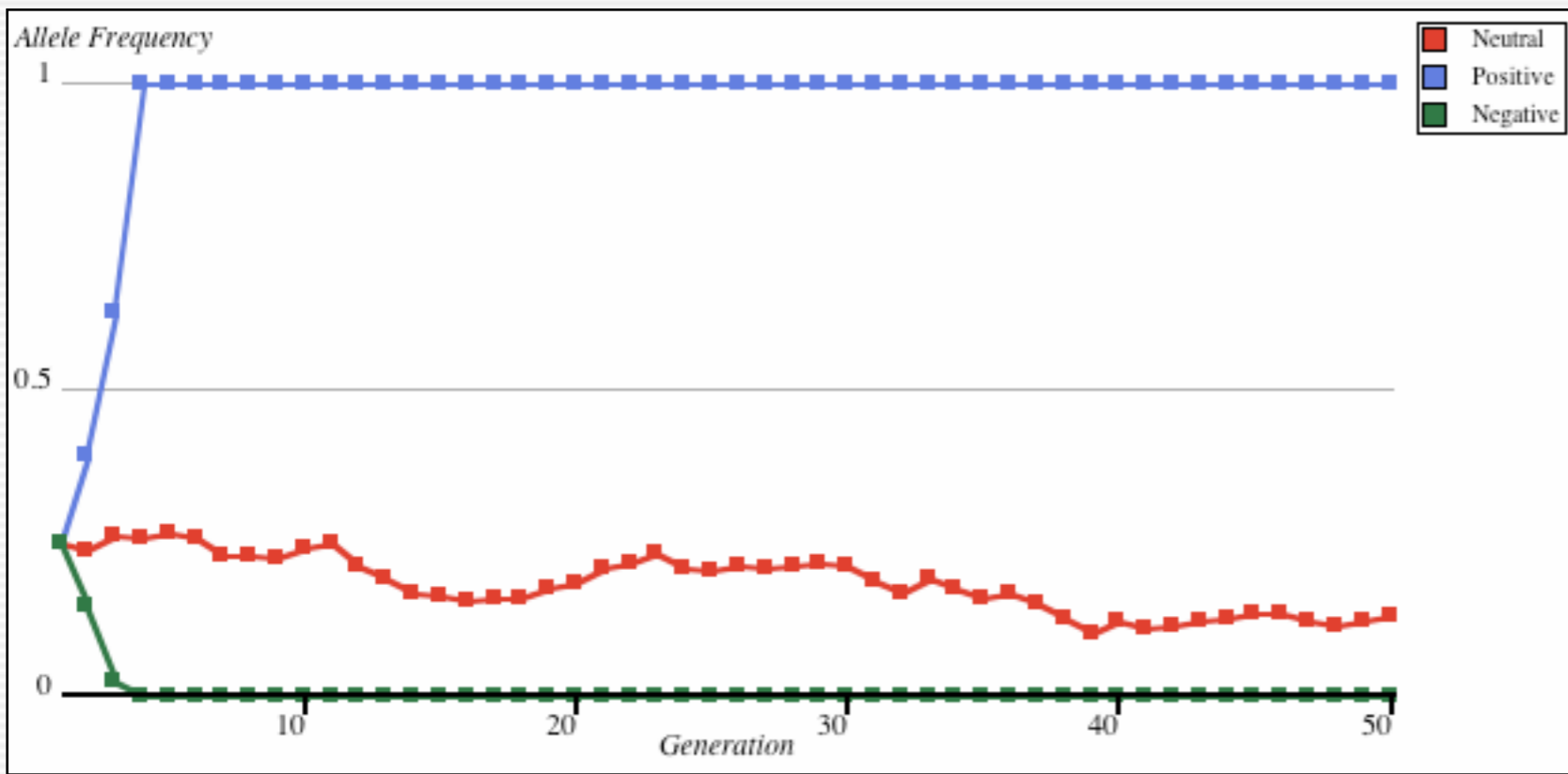
Resistance level



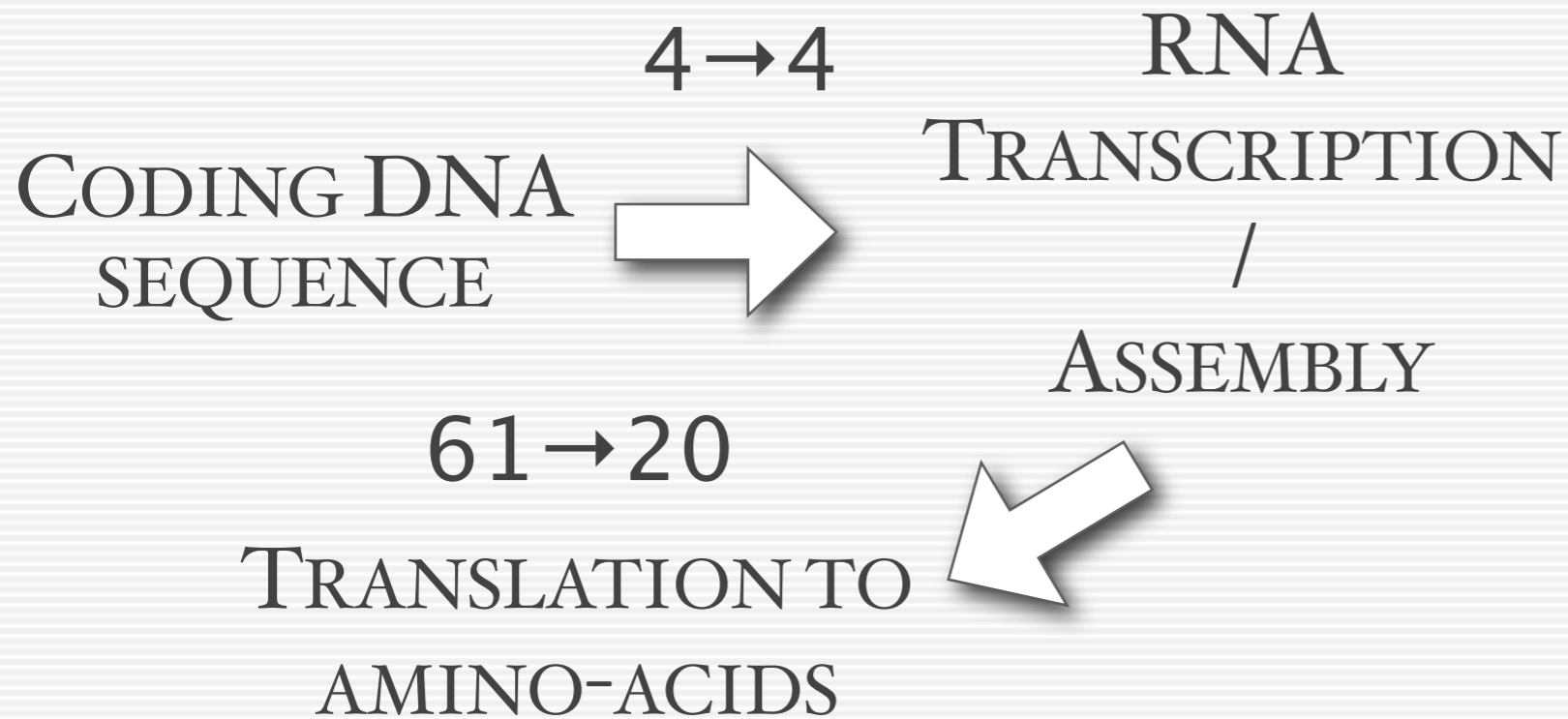
[HTTP://EN.WIKIPEDIA.ORG/WIKI/FILE:ANTIBIOTIC_RESISTANCE.SVG](http://en.wikipedia.org/wiki/File:Antibiotic_resistance.svg)

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.



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

- 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

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

Positive Selection

$$dS < dN$$

Negative Selection

$$dS > dN$$

Neutral Evolution

$$dS = dN$$

ESTIMATING dS AND dN

- Consider two aligned sequences

ACA	ATA	ATC	TTT	AAT	CAA
T	I	I	F	N	Q
ACA	ATA	ACC	TTT	AAC	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 (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

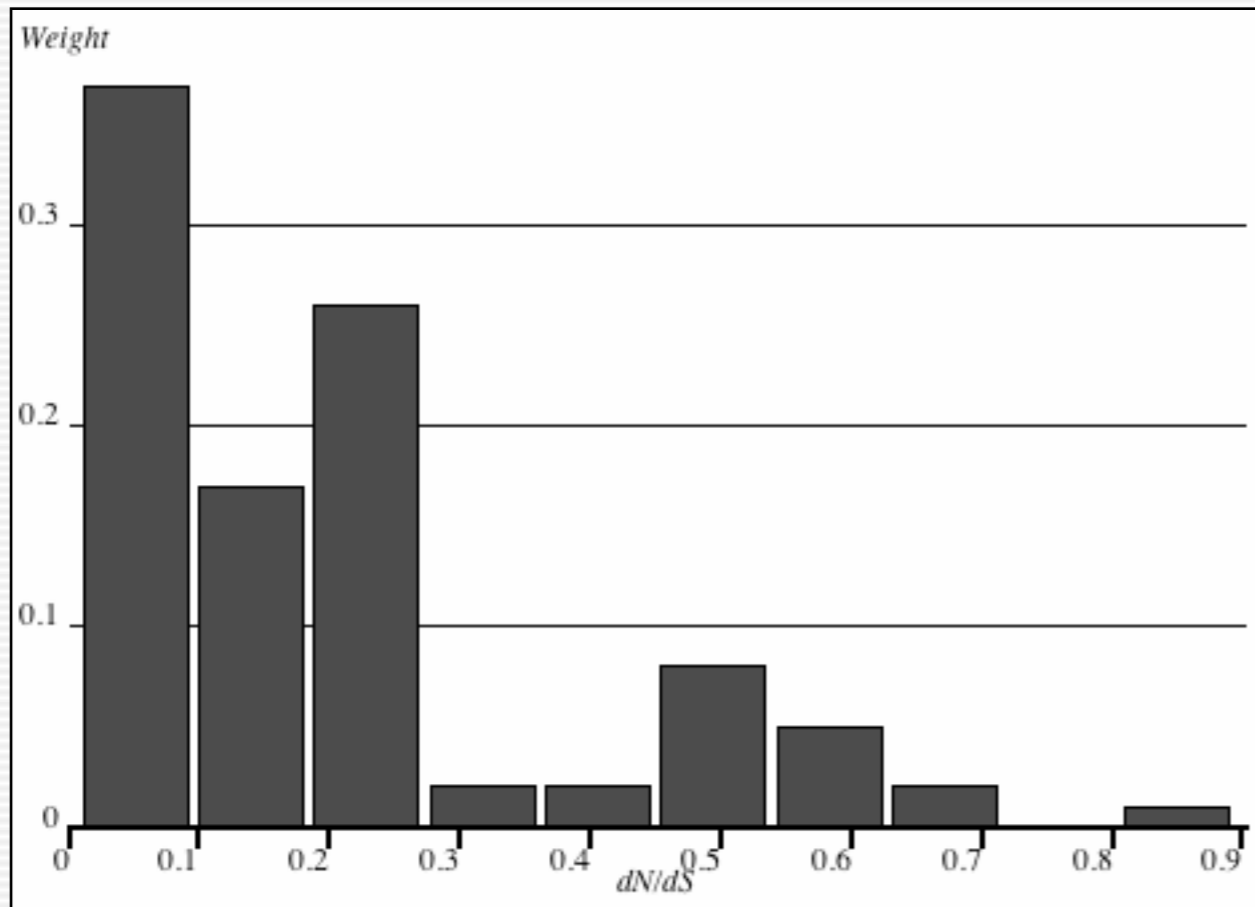
- Consider two aligned sequences

ACA	ATA	ATC	TTT	AAT	CAA
T	I	I	F	N	Q
ACA	ATA	ACC	TTT	AAC	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 E_S , E_N for the two sequences
- Count the number of synonymous substitutions S , and non-synonymous substitutions NS
 - How to deal with multiple substitutions?
- Define $d_S = S/E_S$, $d_N = NS/E_N$
- Compute d_N/d_S and compare to 1
- How to assess significance?

PROBLEMS WITH NG DN/DS

Substitutions = 6

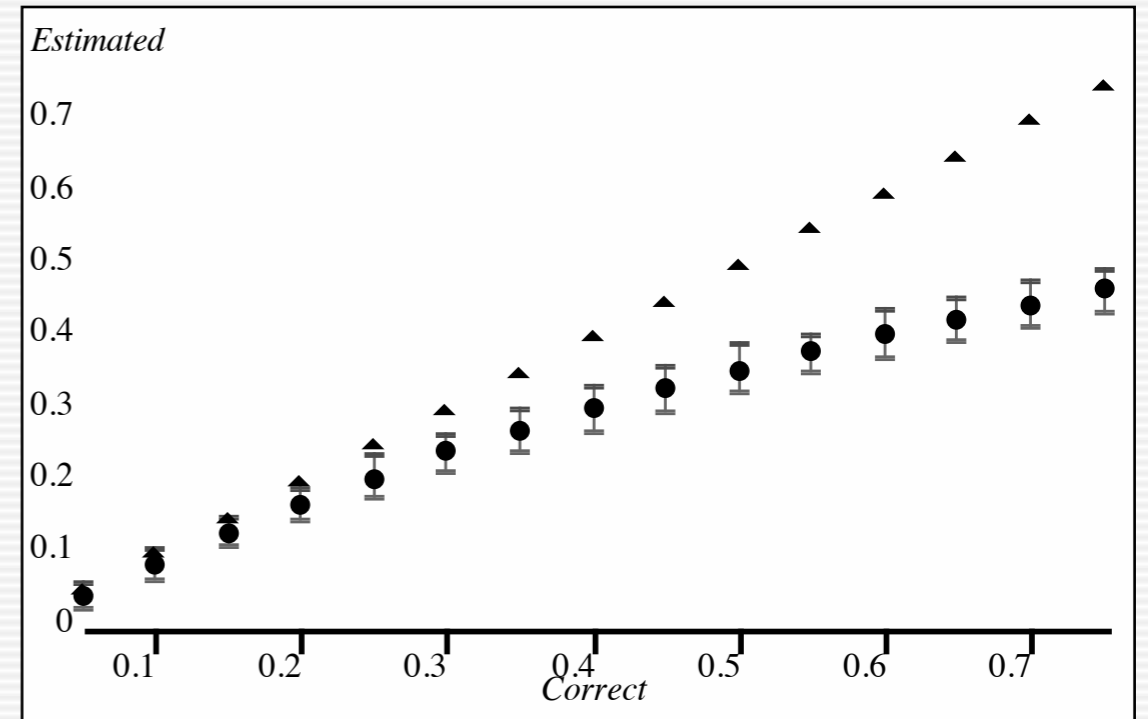
$p = 0.4$

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

†

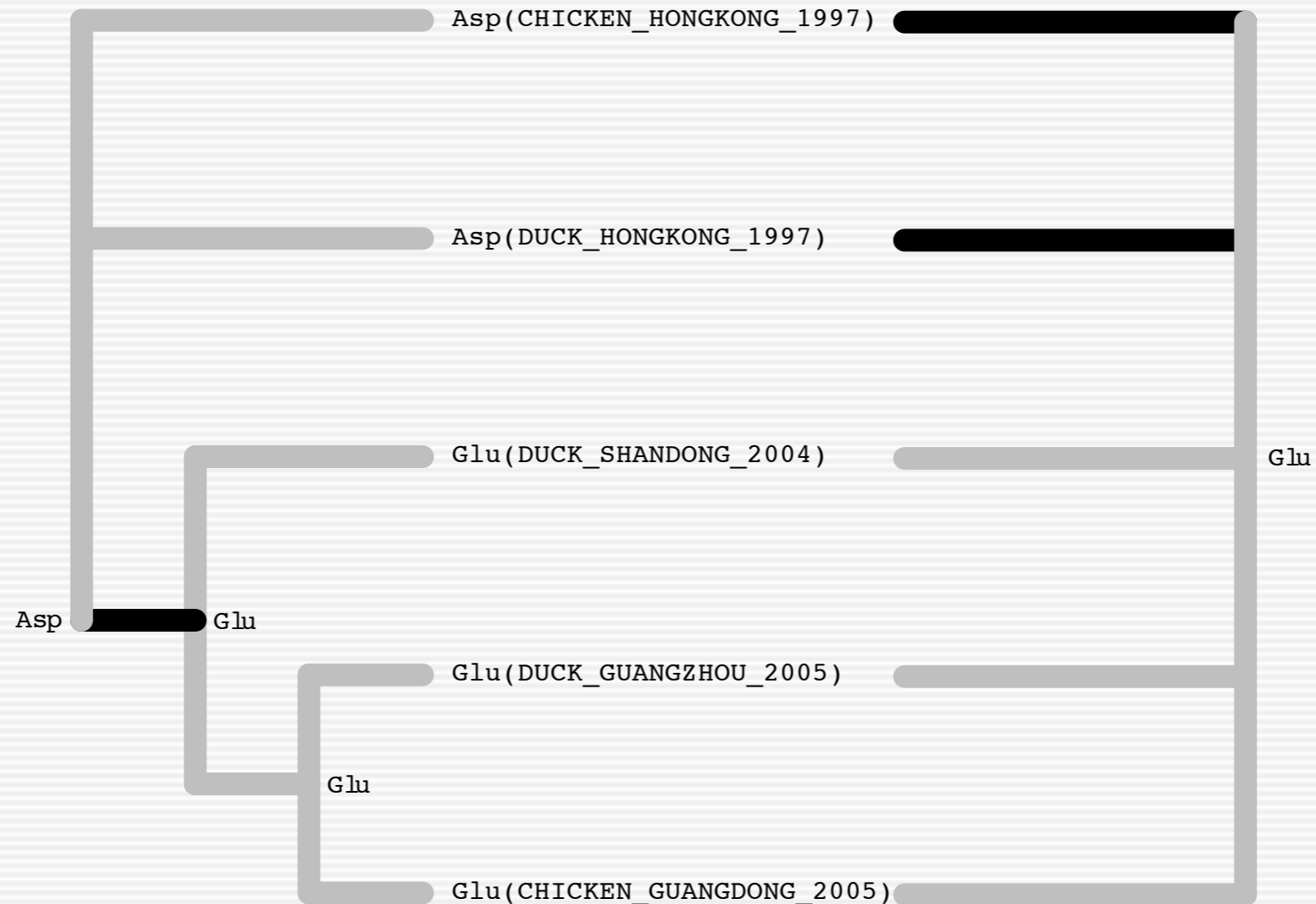
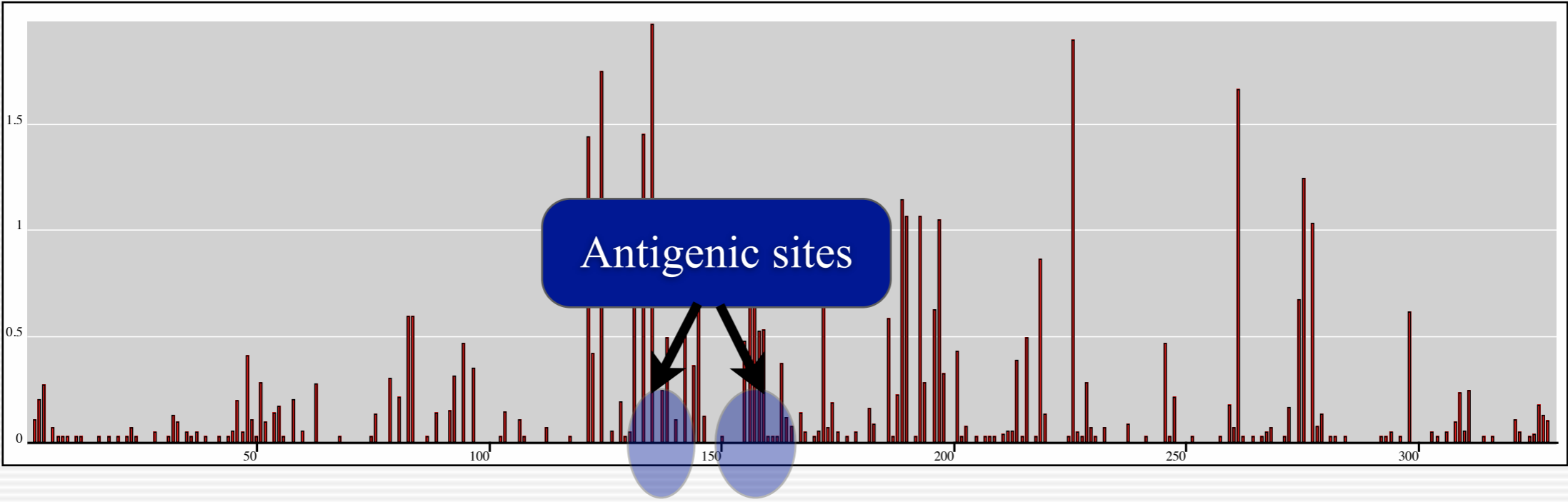


Fig. 1.1. Effect of phylogeny on estimating synonymous and nonsynonymous substitution counts in a dataset of Influenza A/H5N1 haemagglutinin sequences. Using the maximum likelihood tree on the left, the observed variation can be parsimoniously explained with one nonsynonymous substitution along the darker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



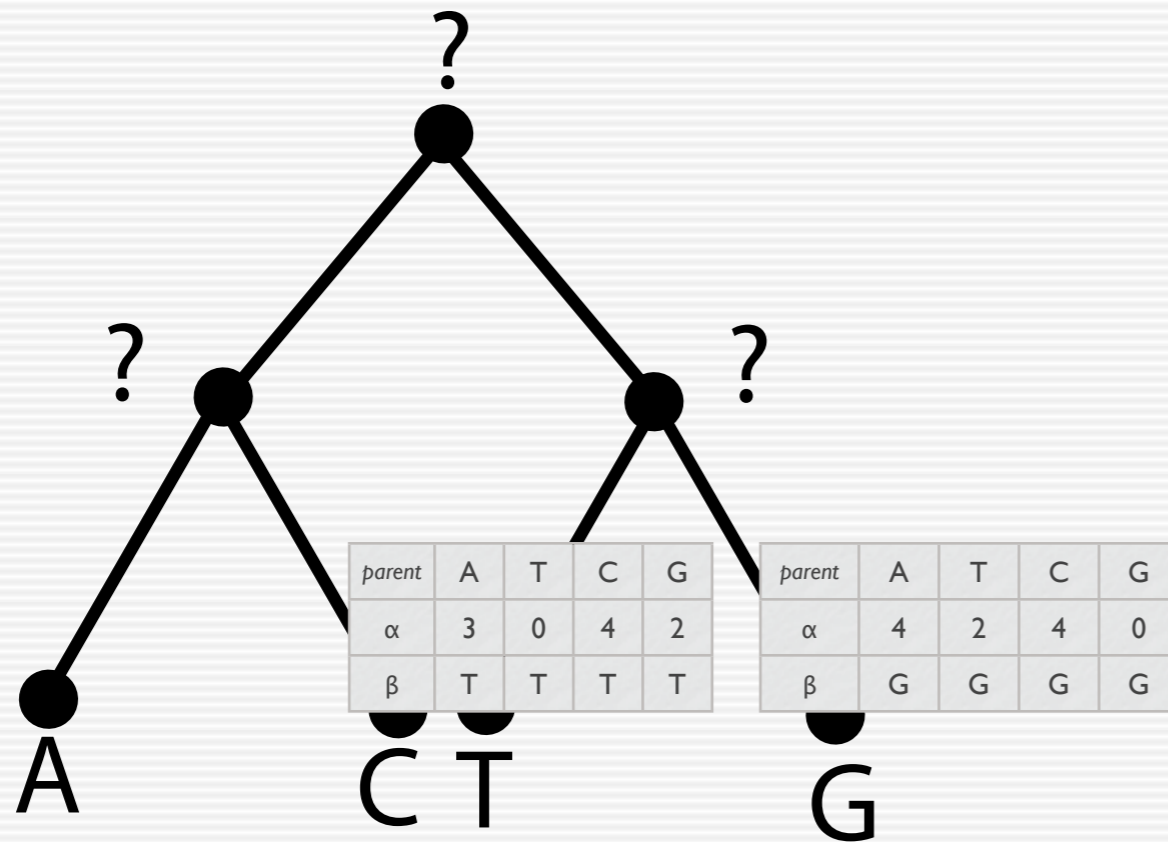
SUZUKI GOJOBORI (SG)

- Uses the tree compute dN/dS at a given site
- First, reconstruct ancestral sequences by parsimony
- Second, compute E_N and E_S using labeled branches; define $p = E_S/E_N$
- 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).

SANKOFF'S ALGORITHM

- Permits, for a fixed topology, to compute the optimal interior node label assignment and the parsimony score for a user specified cost function $\mathbf{c}(\mathbf{x}, \mathbf{y})$
- Uses the fact that the score of the subtree rooted at some interior node \mathbf{n} is independent of the rest of the tree given the label of \mathbf{n} 's parent.
- For each node \mathbf{n} in the tree, the algorithm populates two arrays (of dimension equal to the size of the alphabet) for each node – leaf and internal – in the topology (except the arbitrarily chosen interior node designated as root):
 - $\alpha_{\mathbf{n}}(\mathbf{i})$ - the optimal score of the subtree rooted at \mathbf{n} , given that the label of \mathbf{n} 's parent is \mathbf{i} .
 - $\beta_{\mathbf{n}}(\mathbf{i})$ - the label at \mathbf{n} that achieves score $\alpha_{\mathbf{n}}(\mathbf{i})$
- The arrays can be computed recursively from the leaves up to the tree root
- The second pass from the root down to the leaves assigns the optimal labels

STEP 1: Traverse the tree from the leaves up (postorder) and populate cost/label arrays



parent	A	T	C	G
α	0	3	9	4
β	A	A	A	A

parent	A	T	C	G
α	9	4	4	∞
β	C	C	C	C

Substitution costs

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

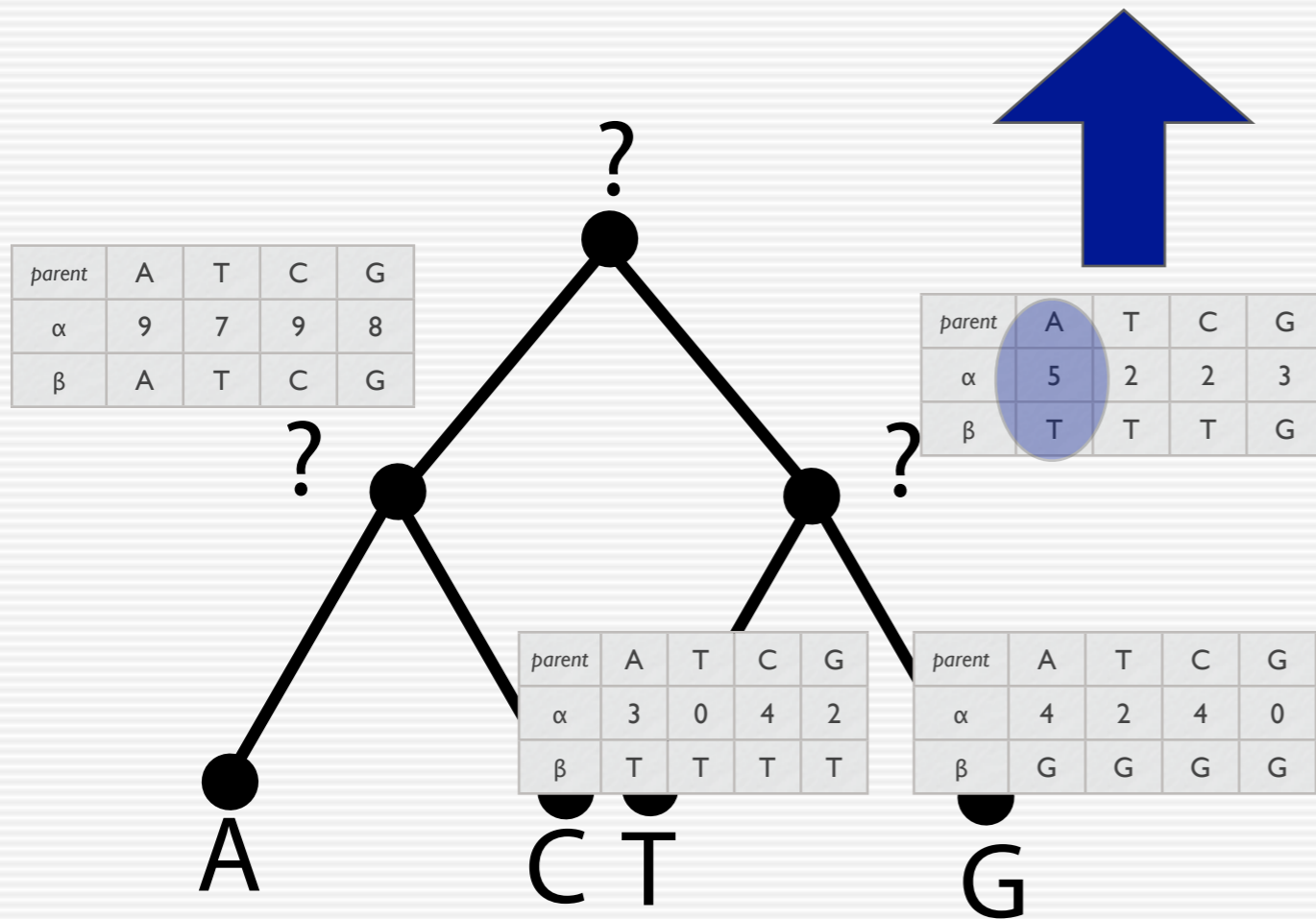
STEP I...

TRY A: $c(A,A) + 3+4=7$

TRY T: $c(A,T)+0+2=5$

TRY C: $c(A,C)+4+4=17$

TRY G: $c(A,G)+2+0=6$



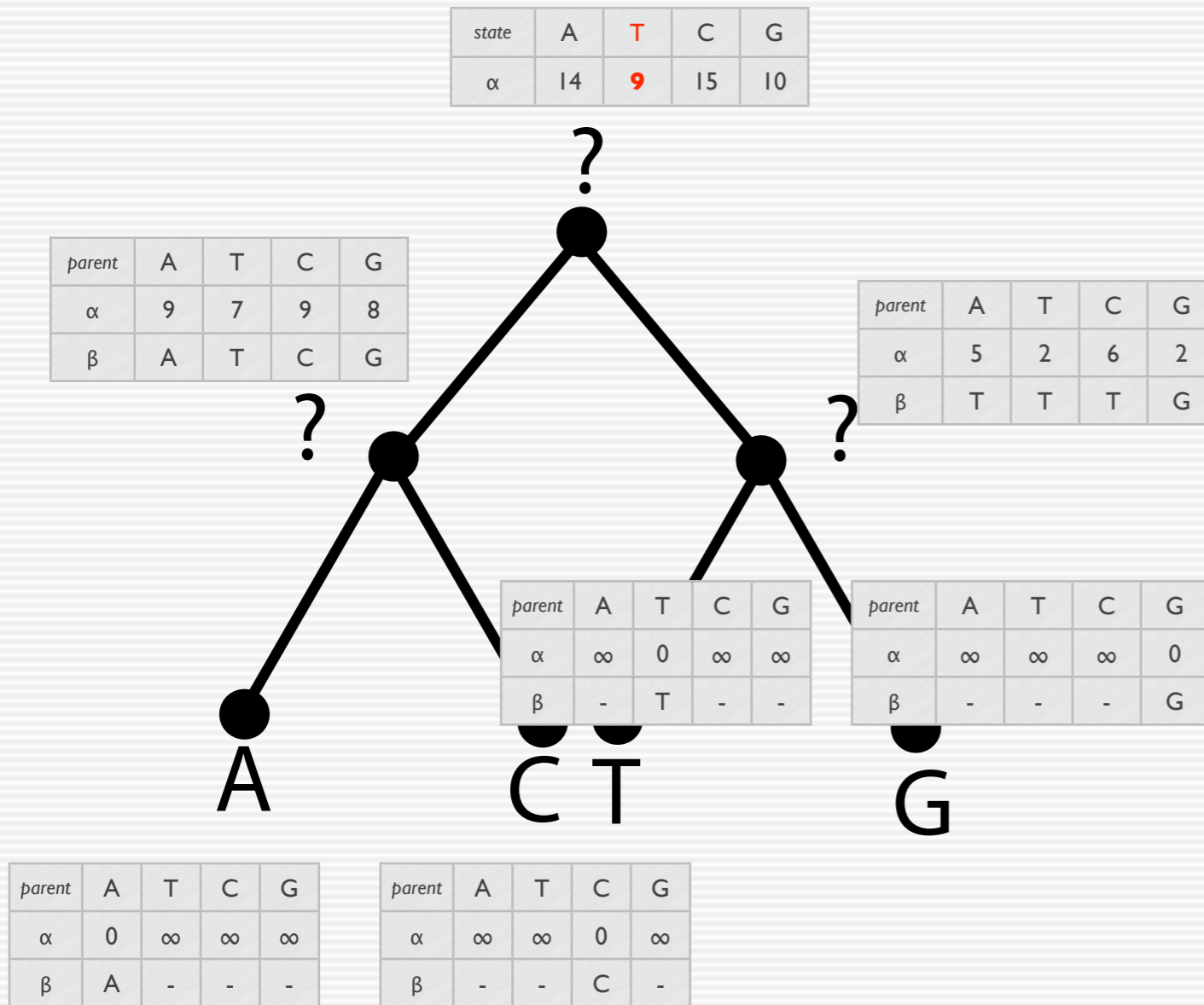
Substitution costs

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

parent	A	T	C	G
α	0	3	9	4
β	A	A	A	A

parent	A	T	C	G
α	9	4	4	∞
β	C	C	C	C

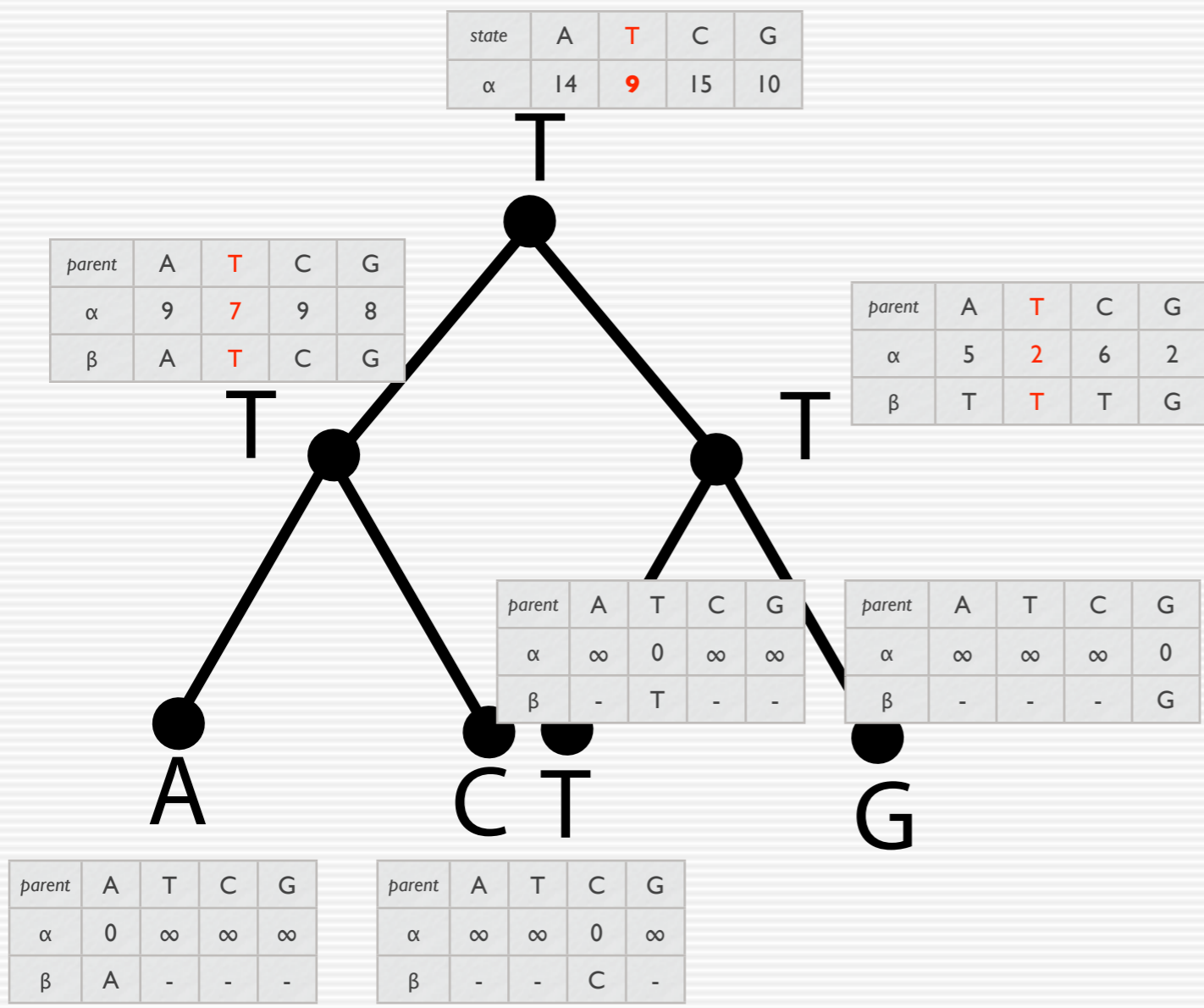
STEP 2: label the root



Substitution costs

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

STEP 3: label the rest of the tree



Substitution costs

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

Optimal cost: 9. Run time?

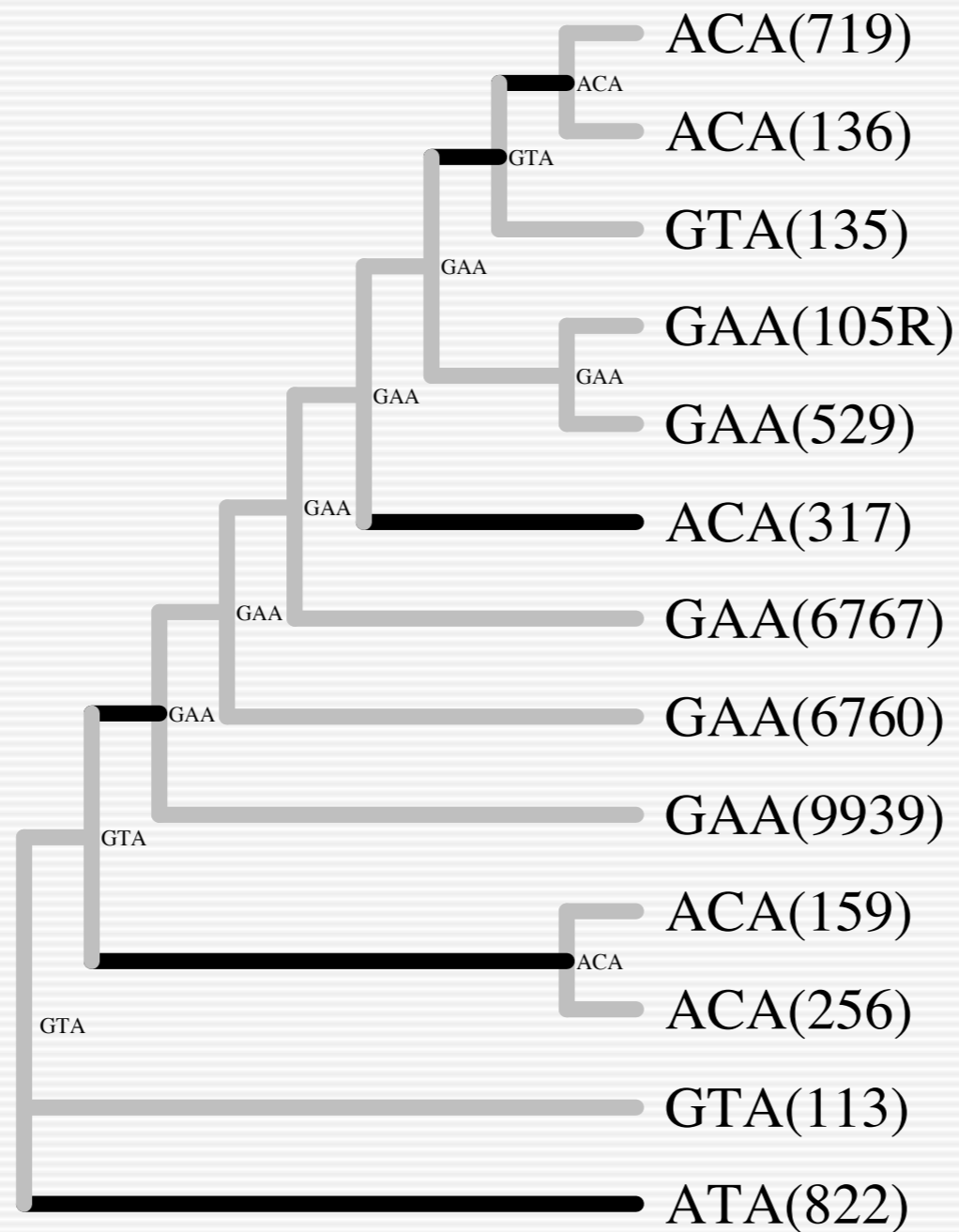


Fig. 1.6. An illustration of SLAC method, applied to a small HIV-1 envelope V3 loop alignment. Sequence names are shown in parentheses. Likelihood state ancestral reconstruction is shown at internal nodes. The parsimonious count yields 0 synonymous and 9 non-synonymous substitutions (highlighted with 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.