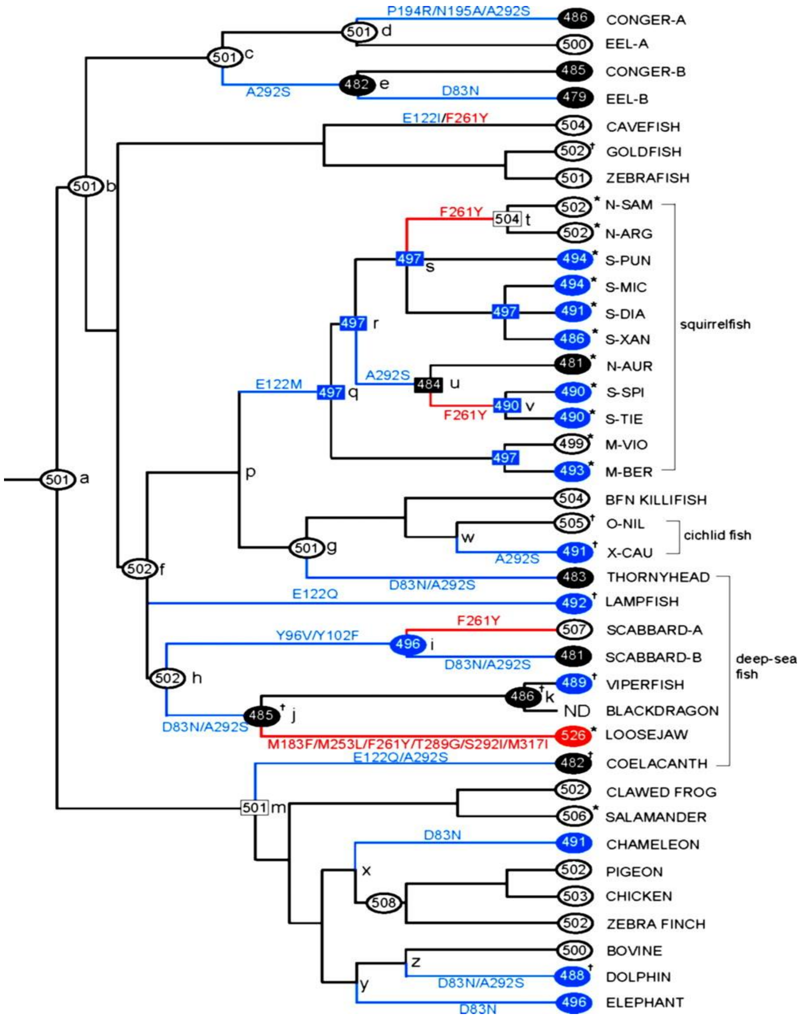
Quick Lesson on dN/dS

- Neutral Selection
- Codon Degeneracy
- Synonymous vs. Non-synonymous
- dN/dS ratios
- Why Selection?
- The Problem

What does selection “look” like?



When moving into new dim-light environments, vertebrate ancestors adjusted their dim-light vision by modifying their rhodopsins

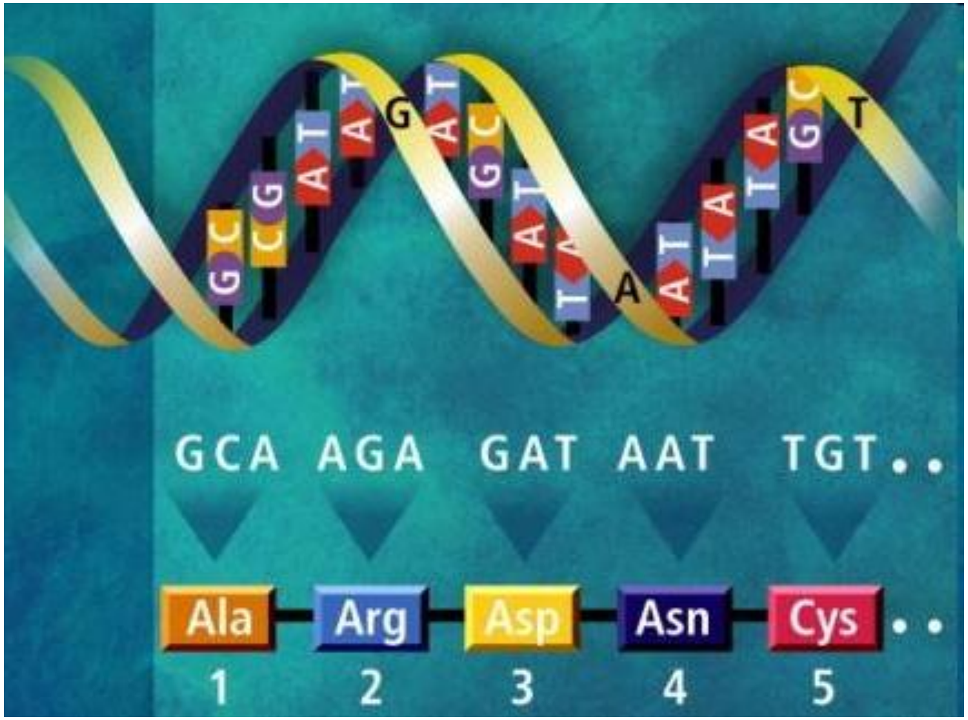
- Functional changes have occurred
- Biologically significant shifts have occurred multiple times
- How do we know whether these shifts are adaptive or random?

Yokoyama S et al. PNAS 2008;105:13480-13485

PNAS

Neutral Selection

Mutations will occur evenly throughout the genome.



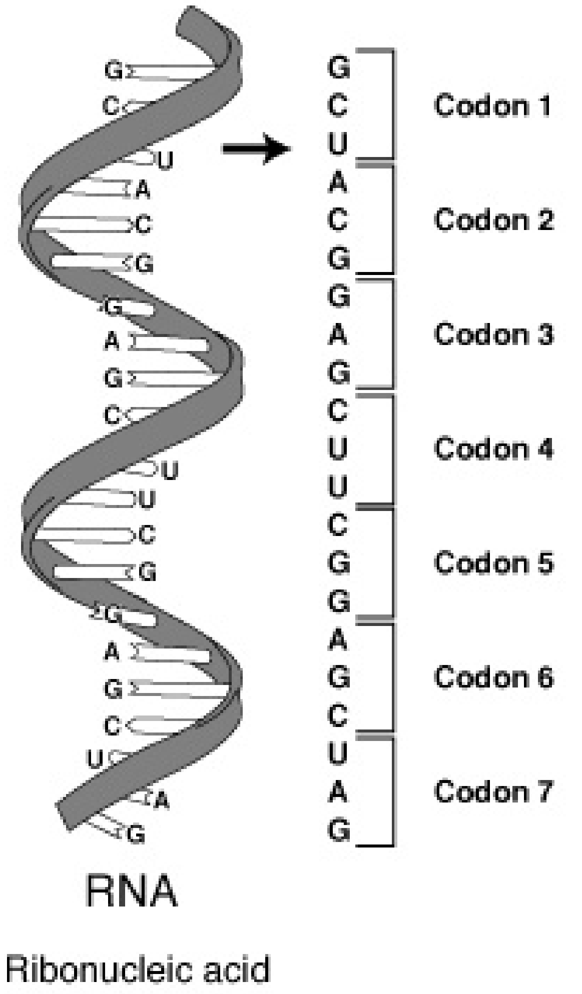
Pseudogenes?

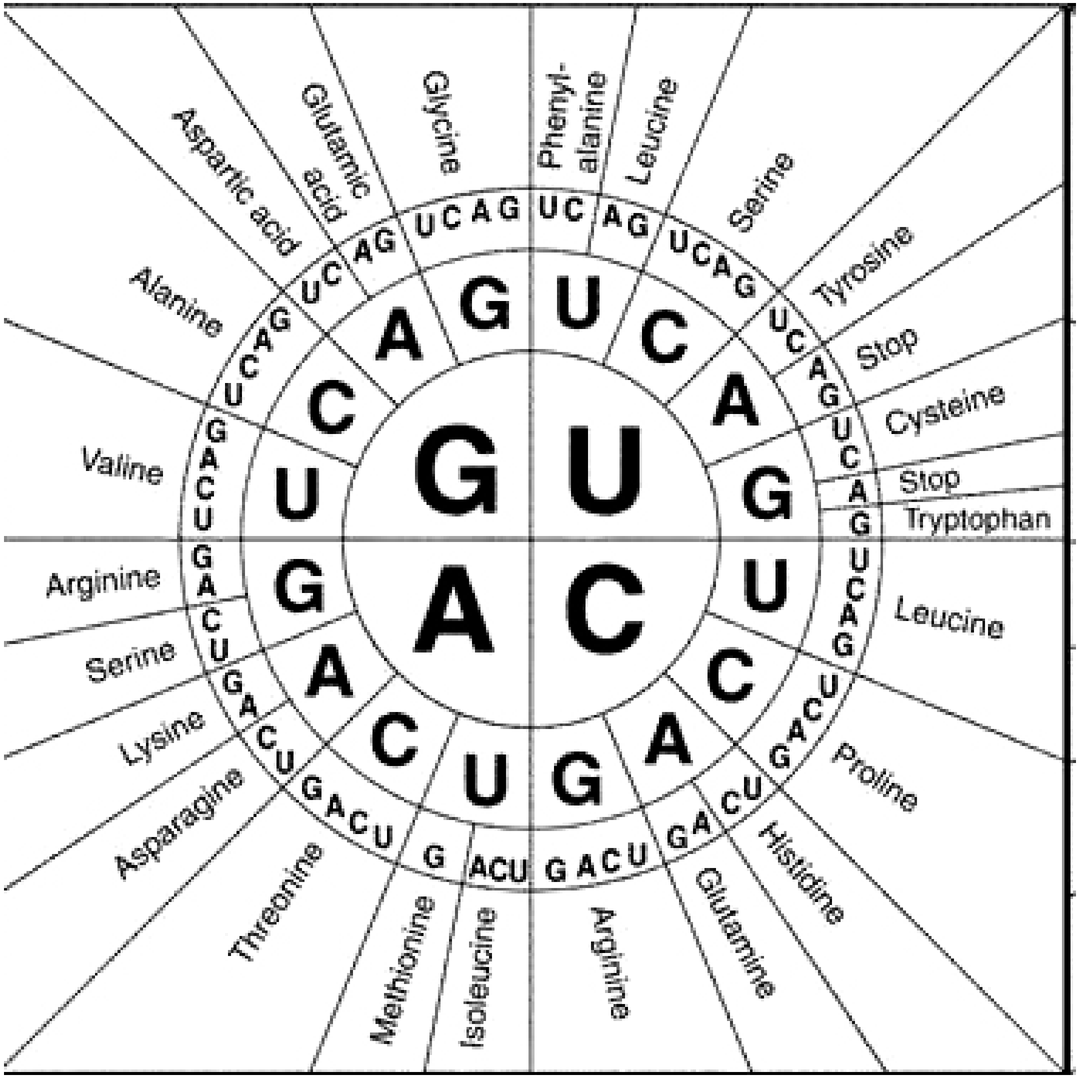
Introns?

Promoters?

Coding Regions?

Codon Degeneracy





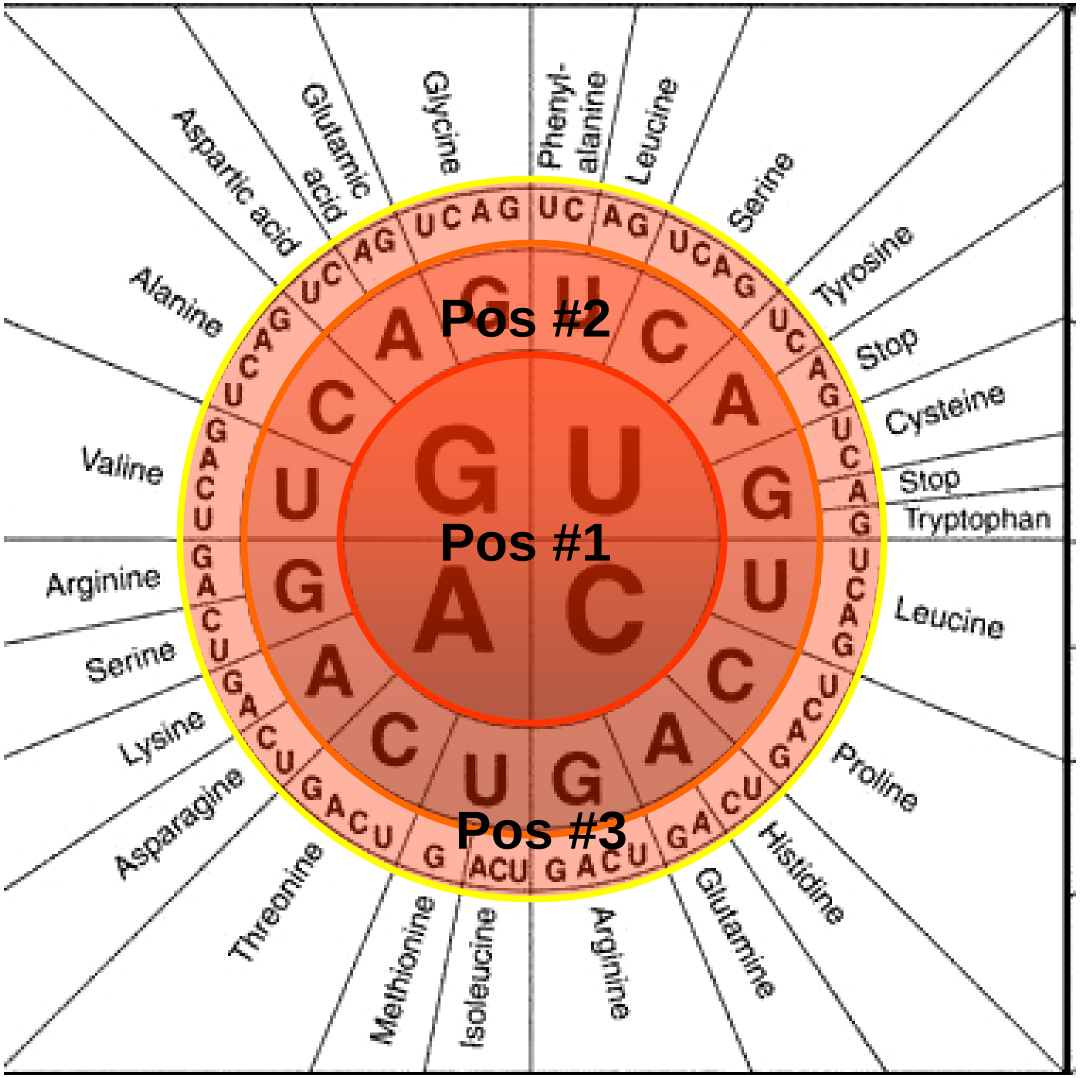
Codon Degeneracy

1st position = strongly conserved

2nd position = conserved

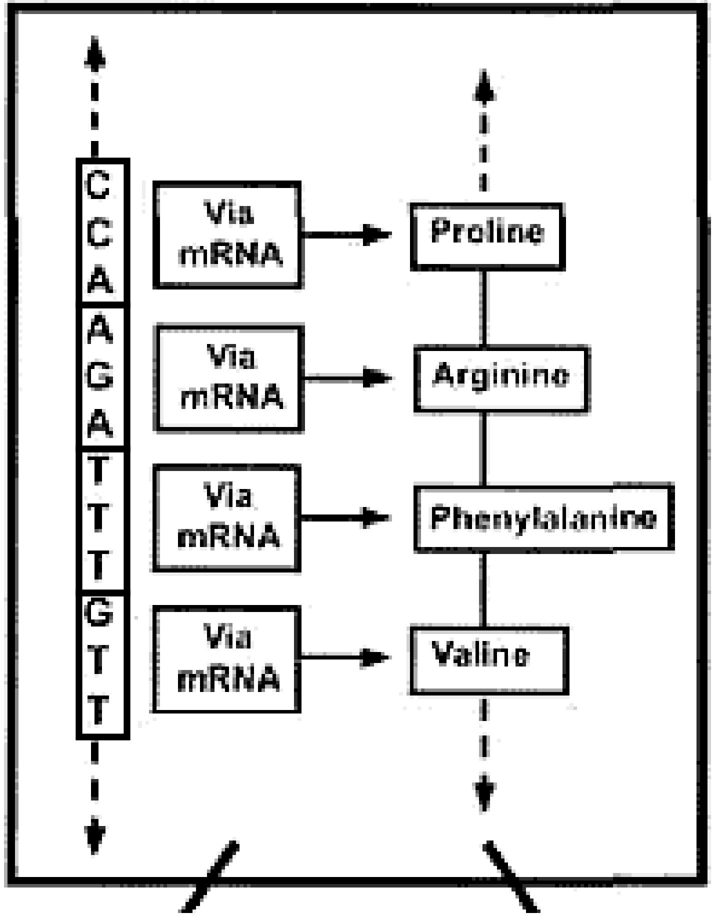
3rd position = “wobbly”

Wobble effect – an AA coded for by more than one codon



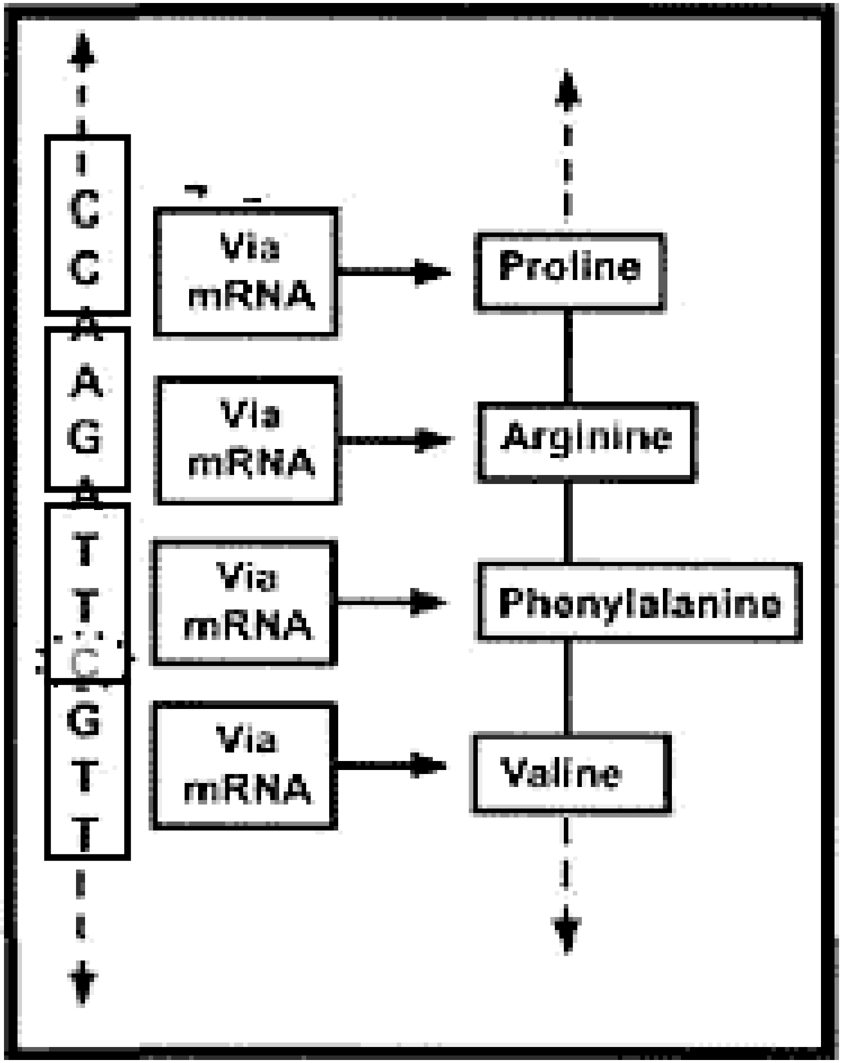
Synonymous vs Non-synonymous

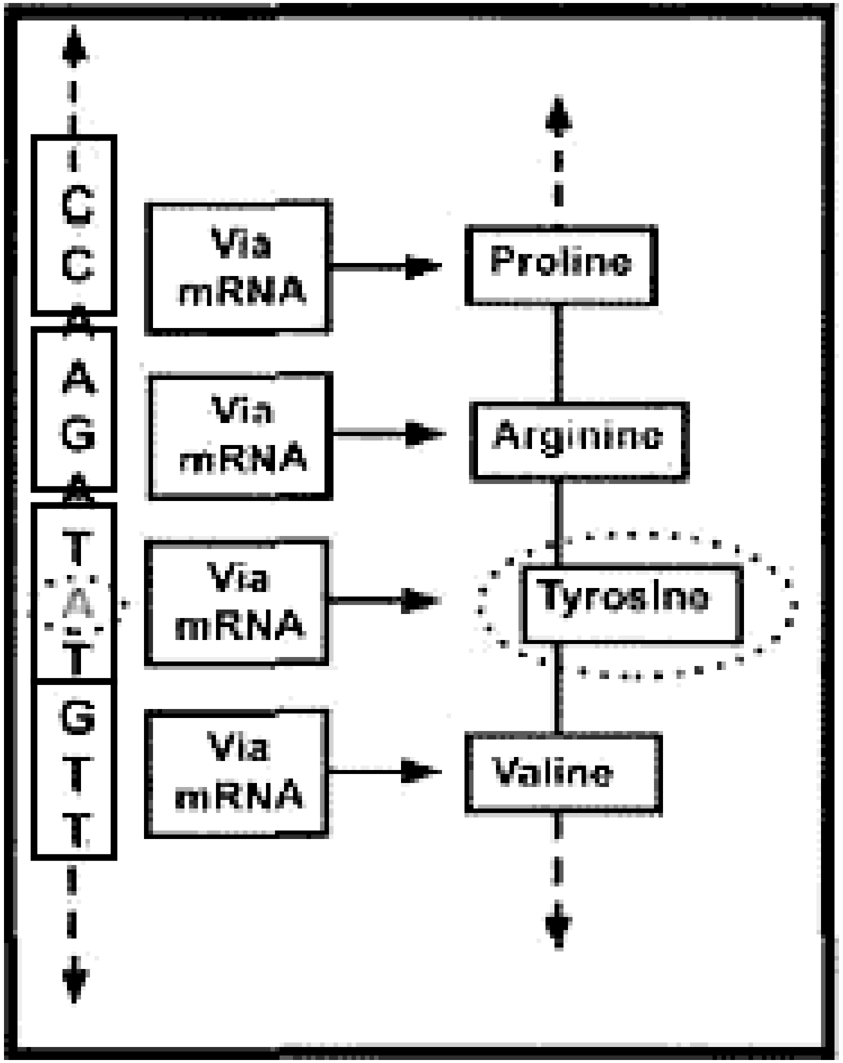
Synonymous:
no AA change



Non-synonymous:
AA change

Synonymous vs Non-synonymous





dN/dS ratios

N = Non-synonymous change

S = Synonymous change

dN = rate of Non-synonymous changes

dS = rate of Synonymous changes

dN / dS = the rate of Non-synonymous changes over the rate of Synonymous changes

Selection and dN/dS

dN / dS == 1 => neutral selection

No selective pressure

dN / dS <= 1 => negative selection

Selective pressure to stay the same

dN / dS >= 1 => positive selection

Selective pressure to change

Why Selection?

Identify important gene regions

Find drug resistance

Locate thrift genes or mutations

dN/dS Problem

Analyzes whole gene or large segments

But, selection occurs at amino acid level

This method lacks statistical power

Thus the purpose of this paper

SLAC

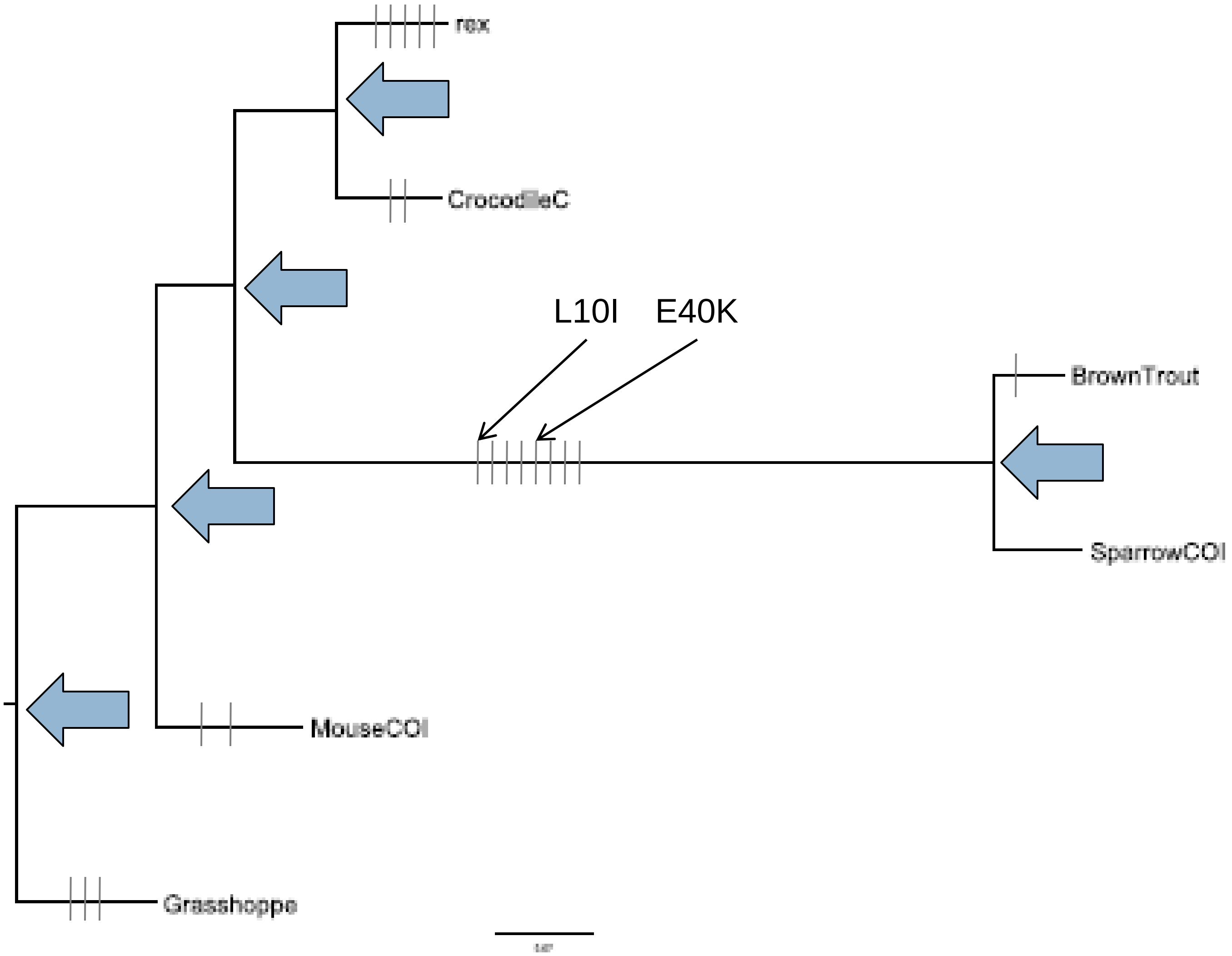
single likelihood ancestor counting

- **The basic idea:**
  Count the number of synonymous and nonsynonymous changes at each codon over the evolutionary history of the sample

$$NN[D_s \mid T, A]$$

$$NS[D_s \mid T, A]$$

SLAC



L10I E40K

SLAC

Strengths:

- Computationally inexpensive
- More powerful than other counting methods in simulation studies

Weaknesses:

- We are assuming that the reconstructed states are correct
- Adding the number of substitutions over all the branches may hide significant events
- Simulation studies shows that SLAC underestimates substitution rate

Runtime estimates

- Less than a minute for 200-300 sequence datasets

FEL

fixed effects likelihood

- **The basic idea:**
  Use the principles of maximum likelihood to estimate the ratio of nonsynonymous to synonymous rates at each site

FEL

$$MG94^{*} \times REV_{x,y}(\mathrm{d}t) = \begin{cases} 0, & x \rightarrow y \text{ requires} \geq 2 \text{ nucleotide substitutions,} \\ \underbrace{\alpha_s \hat{R}_{ij} \pi_{n_y}}_{\text{fixed}} \mathrm{d}t, & x \rightarrow y \text{ is a synonymous substitution of nucleotide } i \text{ with nucleotide } j, \\ \underbrace{\beta_s \hat{R}_{ij} \pi_{n_y}}_{\text{fixed}} \mathrm{d}t, & x \rightarrow y \text{ is a nonsynonymous substitution of nucleotide } i \text{ with nucleotide } j. \end{cases}$$

Likelihood Ratio Test

$$H_o: \alpha = \beta$$

$$H_a: \alpha \neq \beta$$

FEL

Strengths:

- In simulation studies, substitution rates estimated by FEL closely approximate the actual values
- Models variation in both the synonymous and nonsynonymous substitution rates
- Easily parallelized, computational cost grows linearly

Weaknesses:

- To avoid estimating too many parameters, we fix the tree topology, branch lengths and rate parameters

Runtime Estimates:

- A few hours on a small cluster for several hundred sequences

REL

random effects likelihood

- **The basic idea:**
  Estimate the full likelihood nucleotide substitution model and the synonymous and nonsynonymous rates simultaneously.
- Compromise: Use discrete categories for the rate distributions

REL

$$MG94^{*}\times REV_{x,y}(\mathrm{d}t) = \begin{cases} 0, & x\rightarrow y \text{ requires } \geq 2 \text{ nucleotide substitutions,} \\ \alpha_s \hat{R}_{ij}\pi_{n_y}\mathrm{d}t, & x\rightarrow y \text{ is a synonymous substitution of nucleotide } i \text{ with nucleotide } j, \\ \beta_s \hat{R}_{ij}\pi_{n_y}\mathrm{d}t, & x\rightarrow y \text{ is a nonsynonymous substitution of nucleotide } i \text{ with nucleotide } j. \end{cases}$$

1. Posterior Probability
2. Ratio of the posterior and prior odds having $\omega > 1$

REL

Strengths:

- Estimates synonymous, nonsynonymous and nucleotide rates simultaneously
- Most powerful of the three methods for large numbers sequences

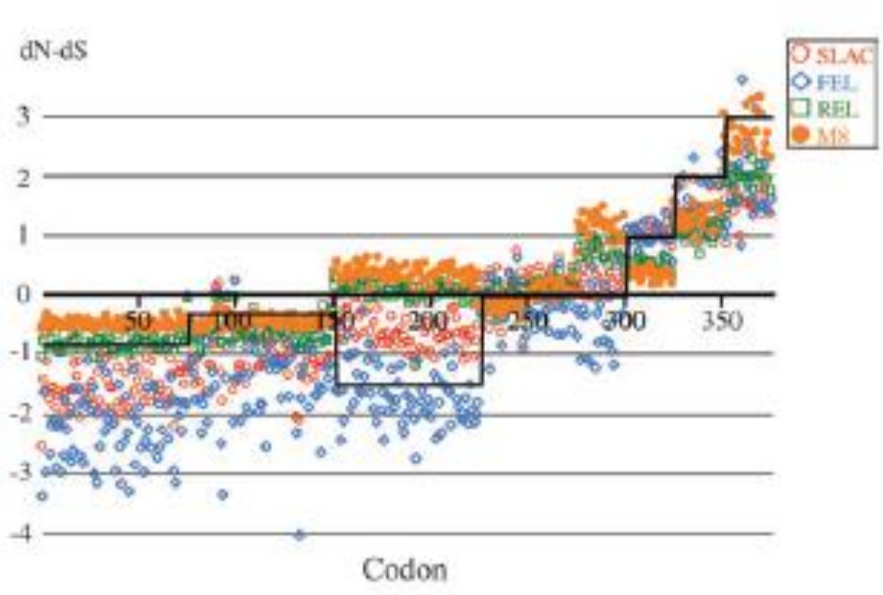
Weaknesses:

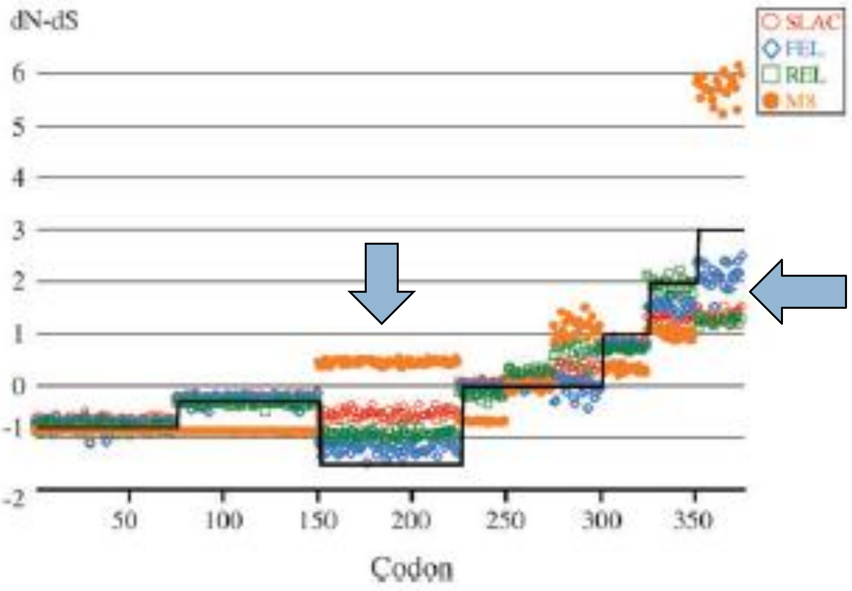
- Performs poorly with small numbers of sequences
- Computationally demanding

Runtime Estimates:

- Not mentioned

Simulation Performance





8 sequences

64 sequences

Selection and dN/dS

dN / dS == 1 => neutral selection

No selective pressure

dN / dS <= 1 => negative selection

Selective pressure to stay the same

dN / dS >= 1 => positive selection

Selective pressure to change