Mon, Nov 23

Lecture 11A:

Detecting selection in proteins

Asclepias syriaca
© Matilda Adams/
Missouri Botanical Garden

Practical Bioinformatics (Biol 4220)
Instructor: Michael Landis
Email: michael.landis@wustl.edu

Lecture 11A outline

1. Protein evolution
2. *dN/dS* and hypotheses
3. Counting method
4. Phylogenetic method
5. Lab 11A overview

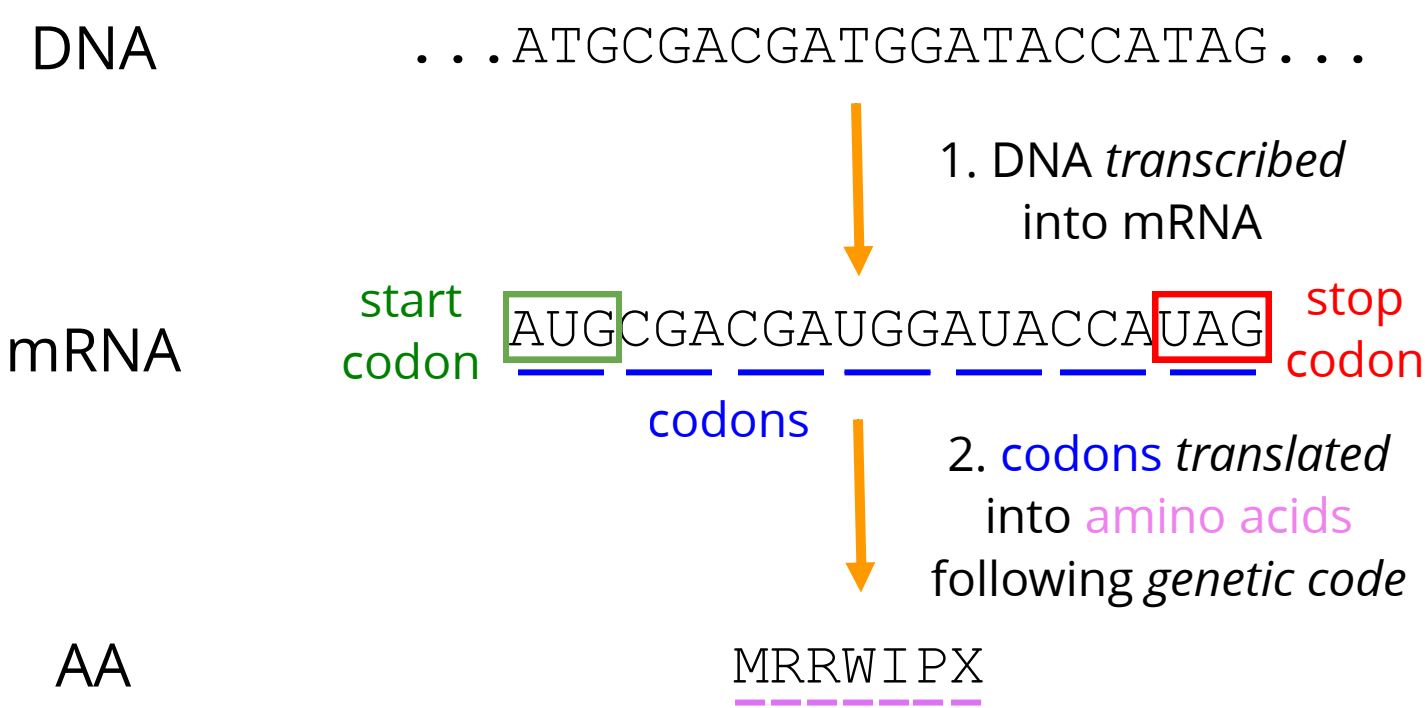
How do proteins evolve?

Selection may act upon DNA mutations
in protein-coding genes

- DNA sequences ***mutate*** and are ***heritable***
- DNA from protein-coding genes is ***transcribed*** into RNA, then ***translated*** into AA through the ***genetic code***
- AA sequences determine ***protein structure***
- Protein structure (largely) determines ***protein function***
- Protein function may influence ***organismal fitness***

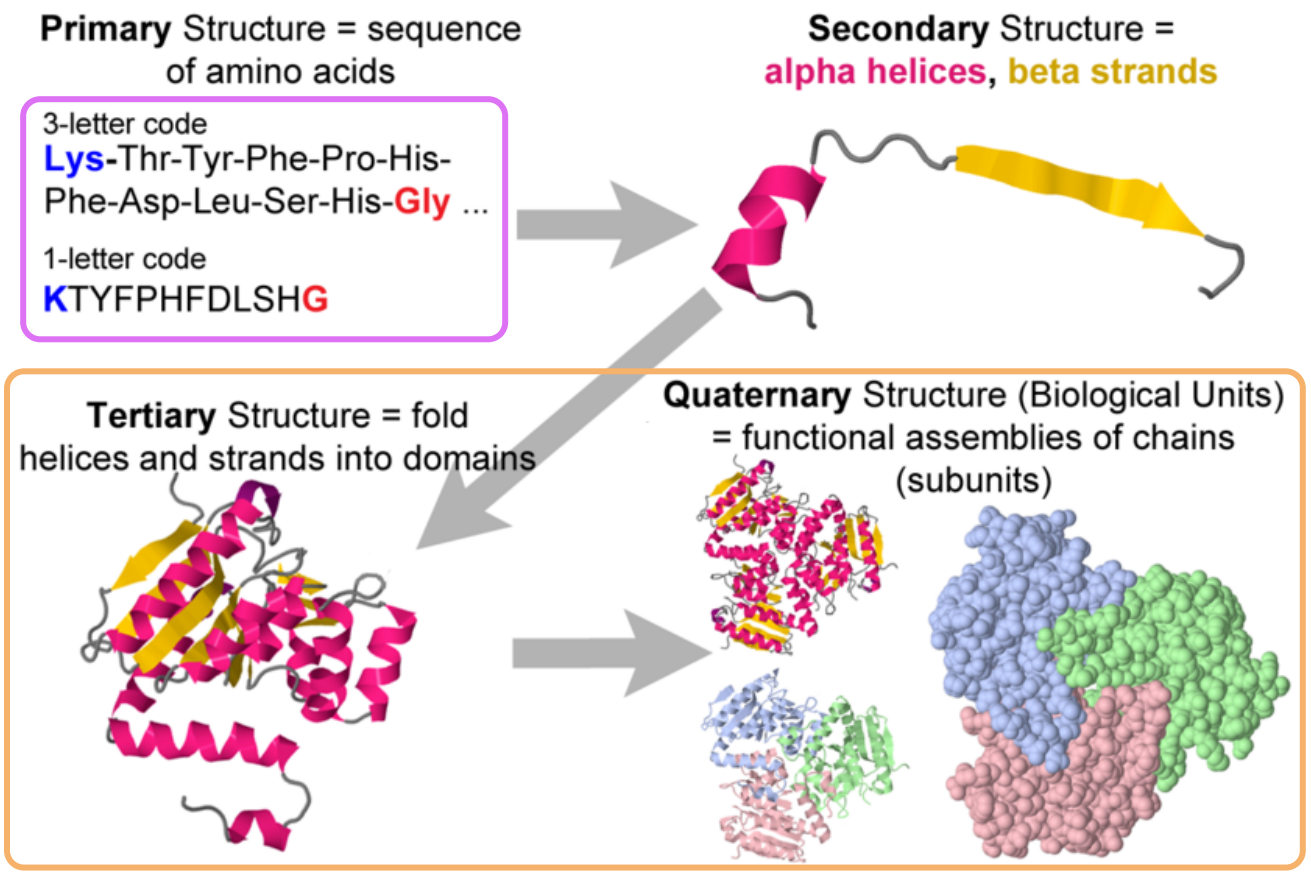
How can we detect if a protein's structure evolves faster or slower than expected?

How do proteins evolve?



How do proteins evolve?

Amino acid sequences

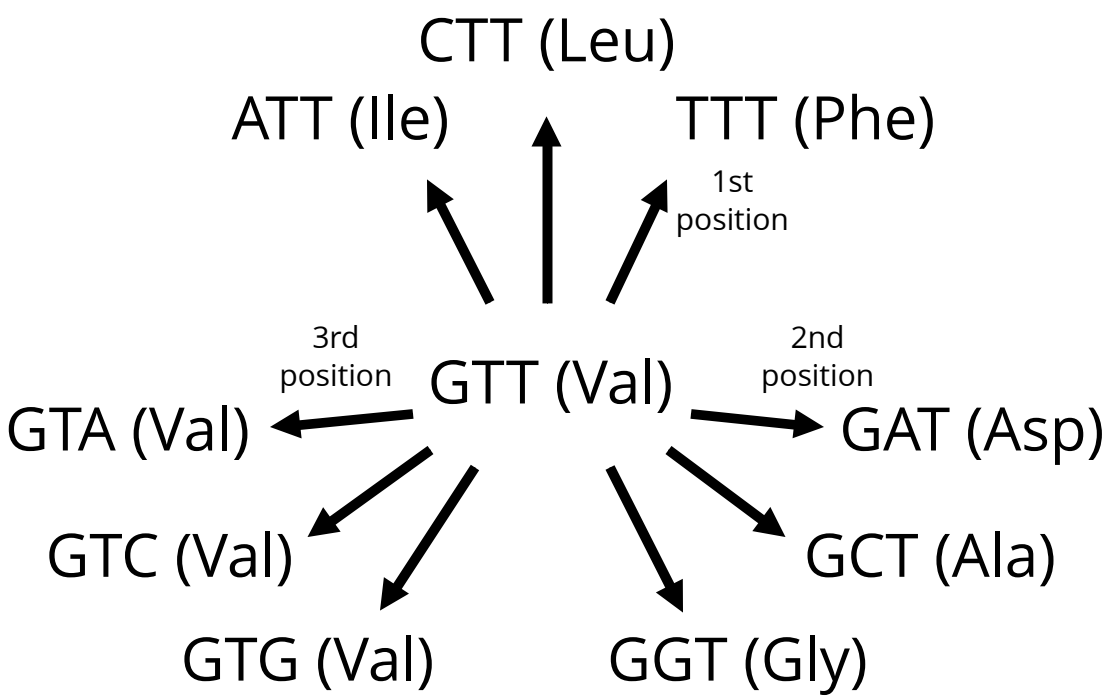


higher order protein structure increasingly determines protein function

proteopedia.org/wiki/index.php/Protein_primary%2C_secondary%2C_tertiary_and_quaternary_structure

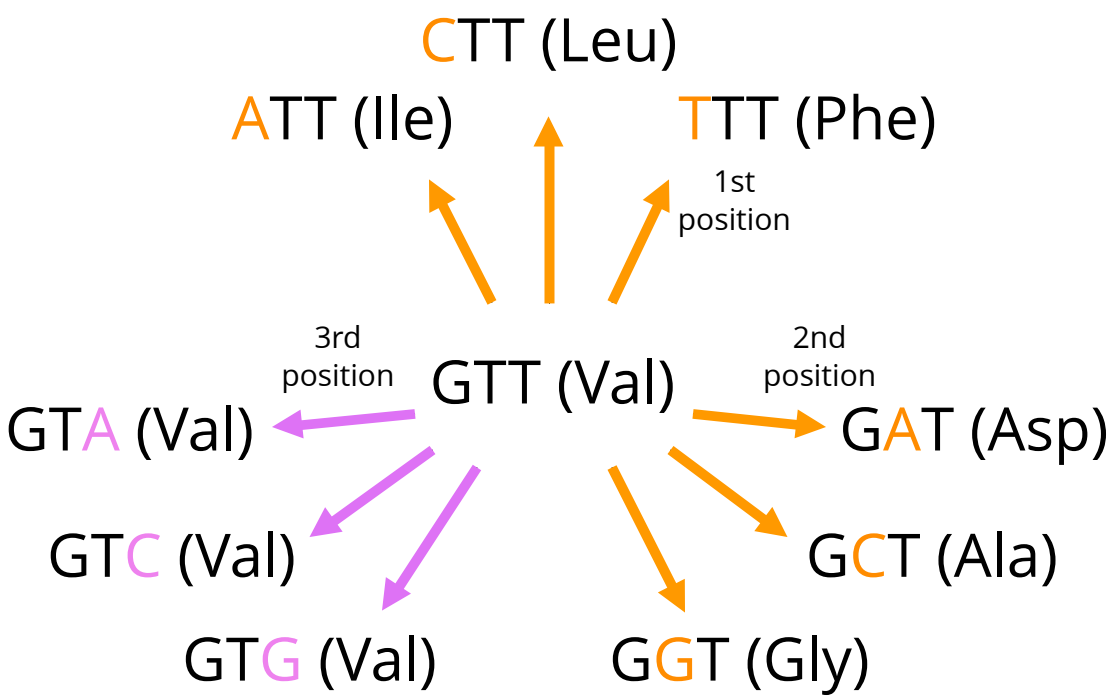
What mutations might induce changes in protein structure?

A point mutation could change a codon into any of nine "adjacent" codons



Synonymous substitutions are silent, and do not cause the amino acid to change;

Nonsynonymous substituions are visible, and chance the encoded amino acid



Scenarios for protein evolution

The relative rate of nonsynonymous vs. synonymous substitution events (called ***dN/dS***) can help us infer what type of selection pressures a protein encountered

If *dN/dS* < 1, then DNA mutations that change AA tend to be discarded (consistent with ***purifying selection***)

If *dN/dS* > 1, then DNA mutations that change AA tend to be kept (consistent with ***positive selection***)

If *dN/dS* = 1, then DNA mutations are kept regardless of effect on AA (consistent with ***neutrality***)

Purifying selection

Genes that encode proteins reponsible for core molecular functions, called ***housekeeping genes***, often have highly conserved protein sequences, structures, and functions

Table 2. Averaged ω in Branches of Phylogenetic Tree of Mammalian H1.1–H1.5 Gene Family.

Hypothesis	lnL	Branches	Omega (ω)
H0	−22708.4	All the branches	0.14116
H1	−22692.1	H1.1	0.18982
		Rest of the branches	0.12314
H2	−22690.59	H1.5	0.08853
		Rest of the branches	0.15575
H3	−22694.48	H1.2	0.1097
		H1.3	0.1741
		H1.4	0.0952
		Rest of the branches	0.1497

dN/dS < 1

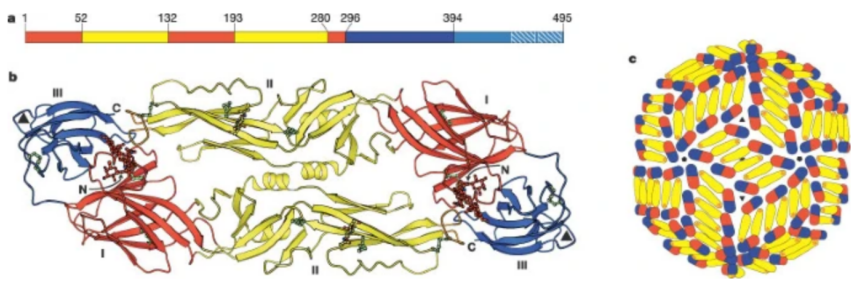
Histone, H1.1

Ponte *et al.* (2017, *Mol. Biol. Evol.*)

Positive selection

Protein-coding genes that are adapting to changing environmental conditions, e.g. genes that participate in ***host-pathogen arms races***, may be enriched for amino acid changes due to positive selection

Figure 1: Structure of the dimer of dengue E soluble fragment (sE) in the mature virus particle.



a, The three domains of dengue sE. Domain I is red, domain II is yellow, domain III is blue. A 53-residue 'stem' segment links the stably folded sE fragment with the C-terminal transmembrane anchor. **b**, The sE dimer[10]. This is the conformation of E in the mature virus particle and in solution above the fusion pH. **c**, Packing of E on the surface of the virus. Electron cryomicroscopy image reconstructions show that 90 E dimers pack in an icosahedral lattice[13].

Modis *et al.* (2004, *Nature*)

Table 2
Maximum Ratio of Nonsynonymous to Synonymous Substitutions for Each DEN-4 Gene Region Examined in This Study

	d_N/d_S[a]		
Gene	Max. d_N/d_S[b]	Proportion of Codons[c]	P[d]
Capsid / membrane	0.822	0.167	0.997
Envelope / NS1	2.110	0.017	0.157
NS2A	4.574	0.009	0.725
NS4B	1.851	0.014	0.937

[a] Values given for the M3 model of codon evolution that allows three classes of d_N/d_S per gene sequence alignment, all of which are estimated from the data.
[b] Highest d_N/d_S for a set of codons estimated under the M3 model.
[c] Proportion of codons with the maximum d_N/d_S value.
[d] Significance value obtained from a likelihood ratio test involving M3 and the neutral codon model M1 (which allows two classes of d_N/d_S, 0 and 1).

dN/dS > 1

Bennett *et al.* (2003, *Mol. Biol. Evol.*)

Neutral theory

For many genes, it is difficult to show that amino acid substitutions occur more or less often than expected under neutrality.

Neutrality is useful as a ***null hypothesis***: how would proteins evolve in the absence of selection?

Some biologists argue that most molecular variation is not adaptive, but rather due to "neutral" processes like genetic drift. This is called ***neutral theory***.

Counting method

Are nonsynonymous substitutions
relatively rare or common?

A simple count-based test for a pair of sequences:

1. Compute # of nonsyn. changes per nonsyn. site (*dN*)
2. Compute # of syn. changes per syn. site (dS)
3. Compute ratio, *dN/dS*, to test whether nonsyn. substitutions occur more or less often than expected by chance
4. Interpret *dN/dS* in terms of purifying, positive, or neutral selection

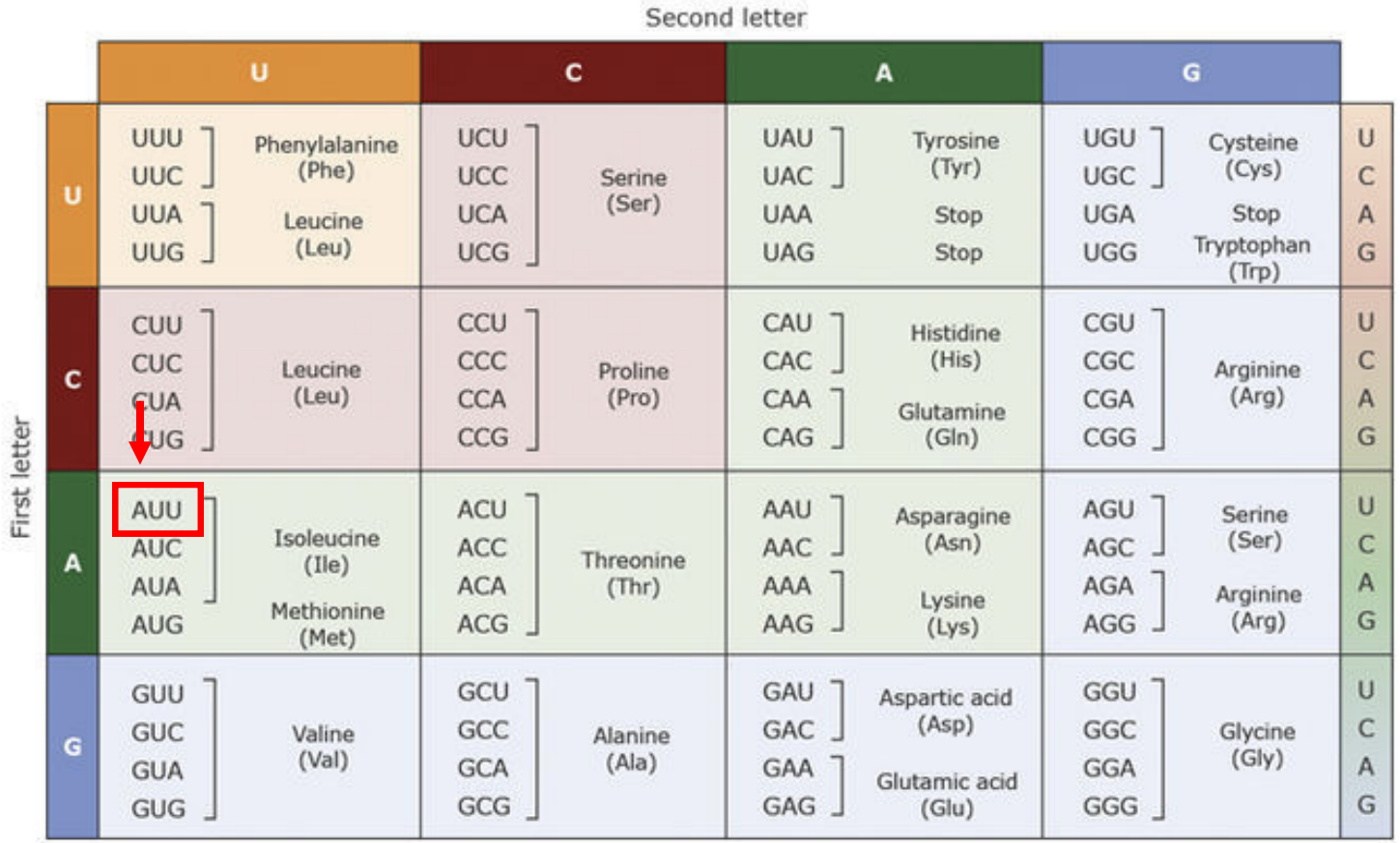
Nei and Gojobori (1986, *Mol Biol Evol*)

Counting method

Compute the number of synonymous sites (*S*)
and the number of nonsynonymous sites (*N*=*L*-*S*)

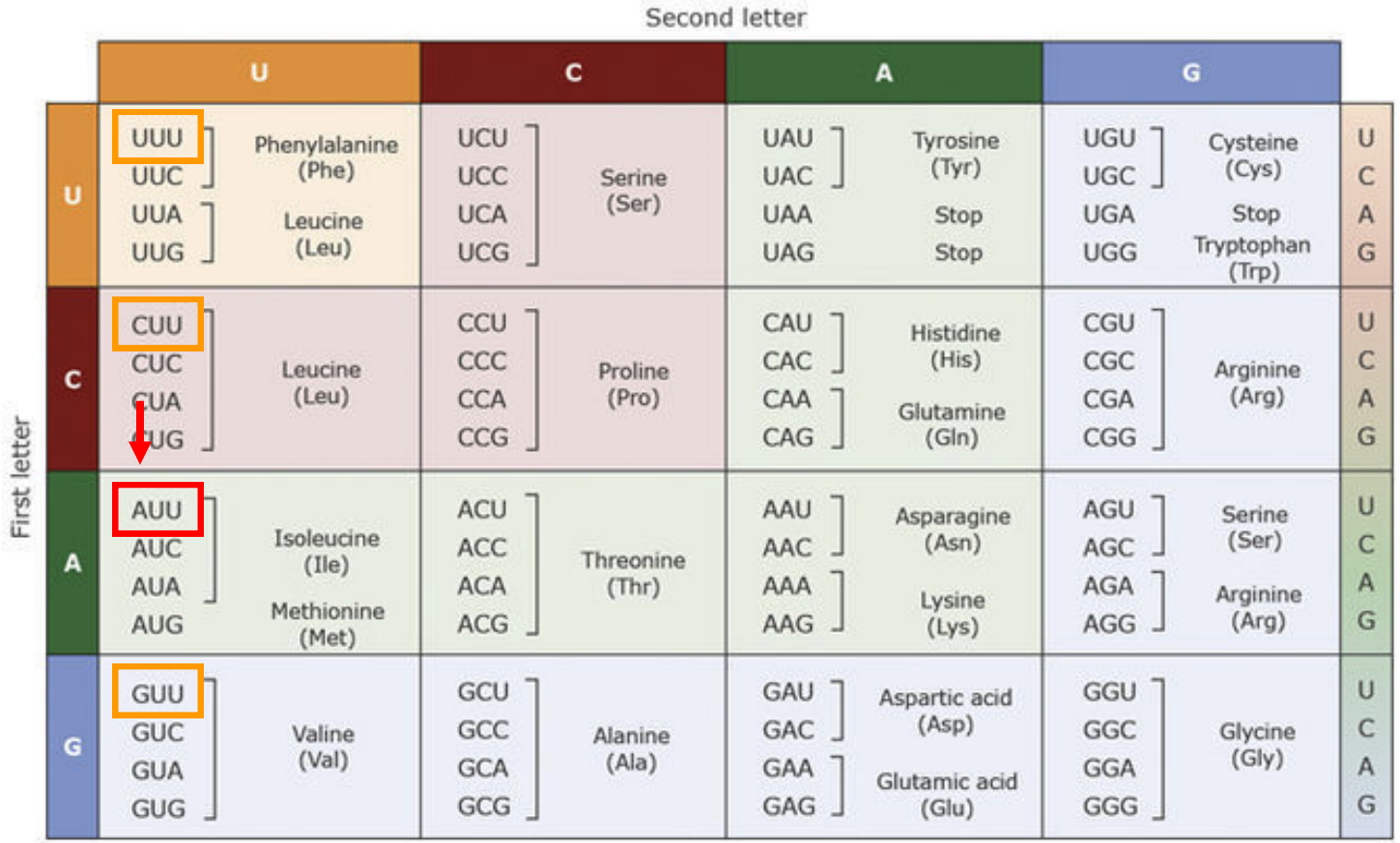
```
Sp_1   GTT ATT GAT GCT TCA GTC
Sp_2   GTT ACT GAC GCA CCA GTC
```

Nei and Gojobori (1986, *Mol Biol Evol*)



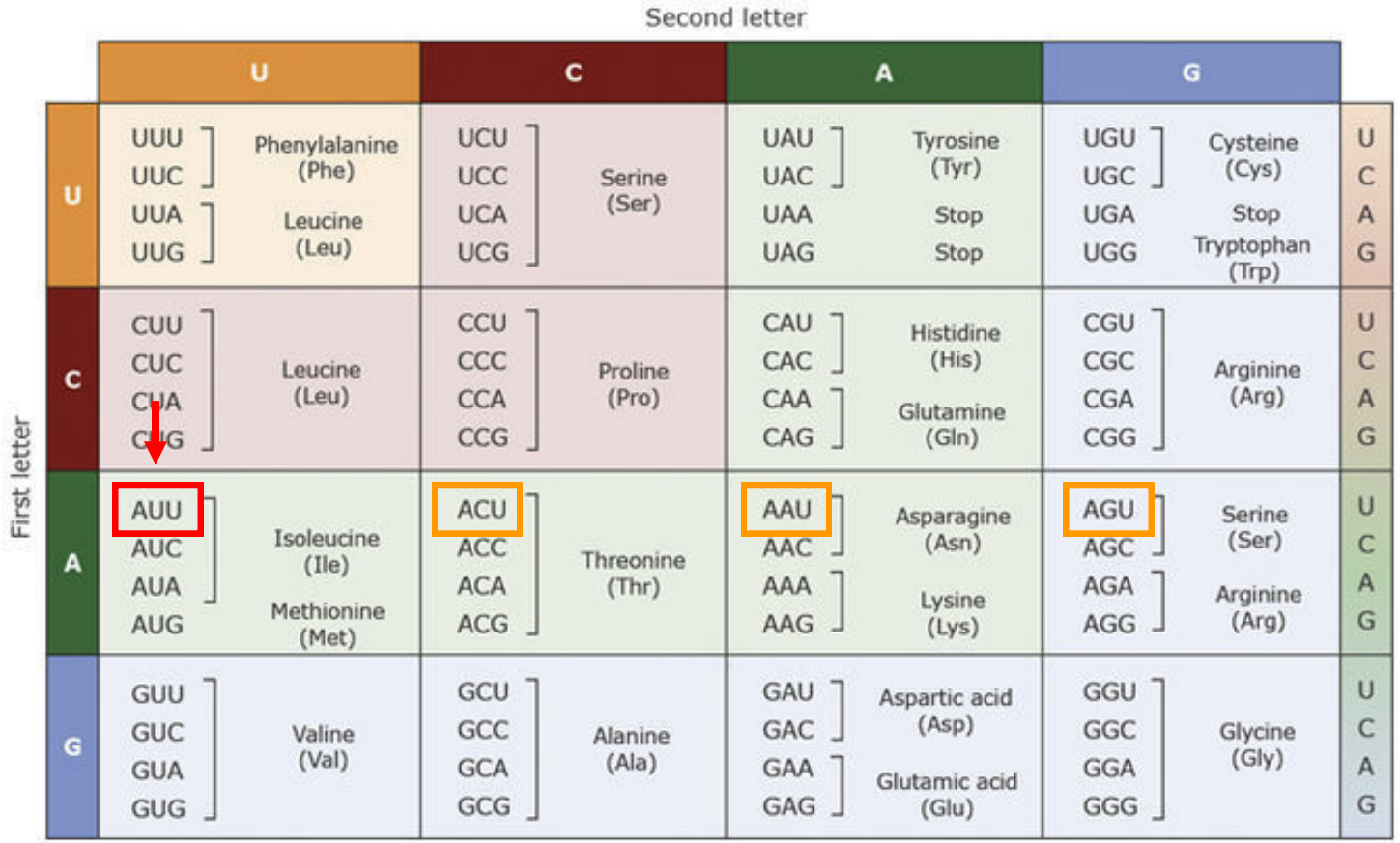
What % of ATT mutations in the ***first codon position*** result in ***synonymous*** changes?

f[1] = ?
f[2] = ?
f[3] = ?



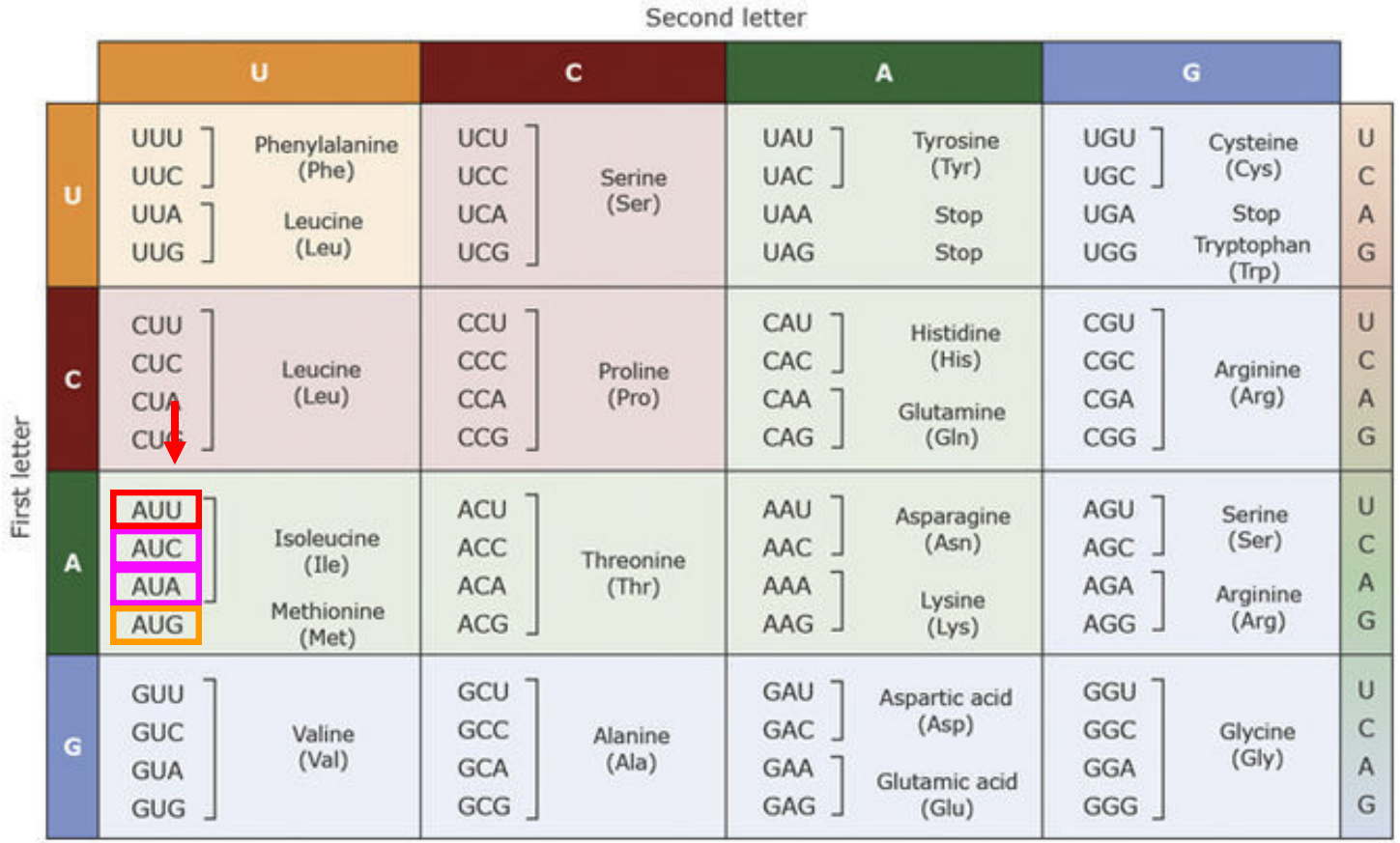
What % of ATT mutations in the ***first codon position*** result in ***synonymous*** changes?

f[1] = 0
f[2] = ?
f[3] = ?



What % of ATT mutations in the ***second codon position*** result in ***synonymous*** changes?

f[1] = 0
f[2] = 0
f[3] = ?



What % of ATT mutations in the ***third codon position*** result in ***synonymous*** changes?

f[1] = 0
f[2] = 0
f[3] = 2

Counting method

Compute the number of synonymous sites (*S*)
and the number of nonsynonymous sites (*N*=*L*-*S*)

0+0+2

```
Sp_1   GTT ATT GAT GCT TCA GTC
Sp_2   GTT ACT GAC GCA CCA GTC
```

Nei and Gojobori (1986, *Mol Biol Evol*)

Counting method

Compute the number of synonymous sites (*S*)
and the number of nonsynonymous sites (*N*=*L*-*S*)

		0+0+2	0+0+1	0+0+3	0+0+3	
Sp_1	GTT	ATT	GAT	GCT	TCA	GTC
Sp_2	GTT	ACT	GAC	GCA	CCA	GTC
		0+0+3	0+0+1	0+0+3	0+0+3	

(can ignore codons w/ no changes)

Nei and Gojobori (1986, *Mol Biol Evol*)

Counting method

L = **12**

S = (1/2) * (2+1+3+3+3+1+3+3)/3 = 19/6 = **3.16**

N = L-S = 12-3.16 = **8.83**

		0+0+2	0+0+1	0+0+3	0+0+3	
Sp_1	GTT	ATT	GAT	GCT	TCA	GTC
Sp_2	GTT	ACT	GAC	GCA	CCA	GTC
		0+0+3	0+0+1	0+0+3	0+0+3	

(can ignore codons w/ no changes)

Nei and Gojobori (1986, *Mol Biol Evol*)

Counting method

Compute the number of synonymous changes (*Sd*)
and the number of nonsynonymous changes (*Nd*)

```
Sp_1  Val Ile Asp Ala Ser Val
Sp_2  Val Thr Asp Ala Pro Val
```

Sd = 2
Nd = 2

Nei and Gojobori (1986, *Mol Biol Evol*)

Counting method

Finally, compute the number of synonymous and nonsynonymous changes per site

dN = Nd / N = 2 / 8.83

dS = Ns / S = 2 / 3.16

dN/dS = 0.36 < 1

This estimate is consistent with purifying selection.

Nei and Gojobori (1986, *Mol Biol Evol*)

Counting method limitations

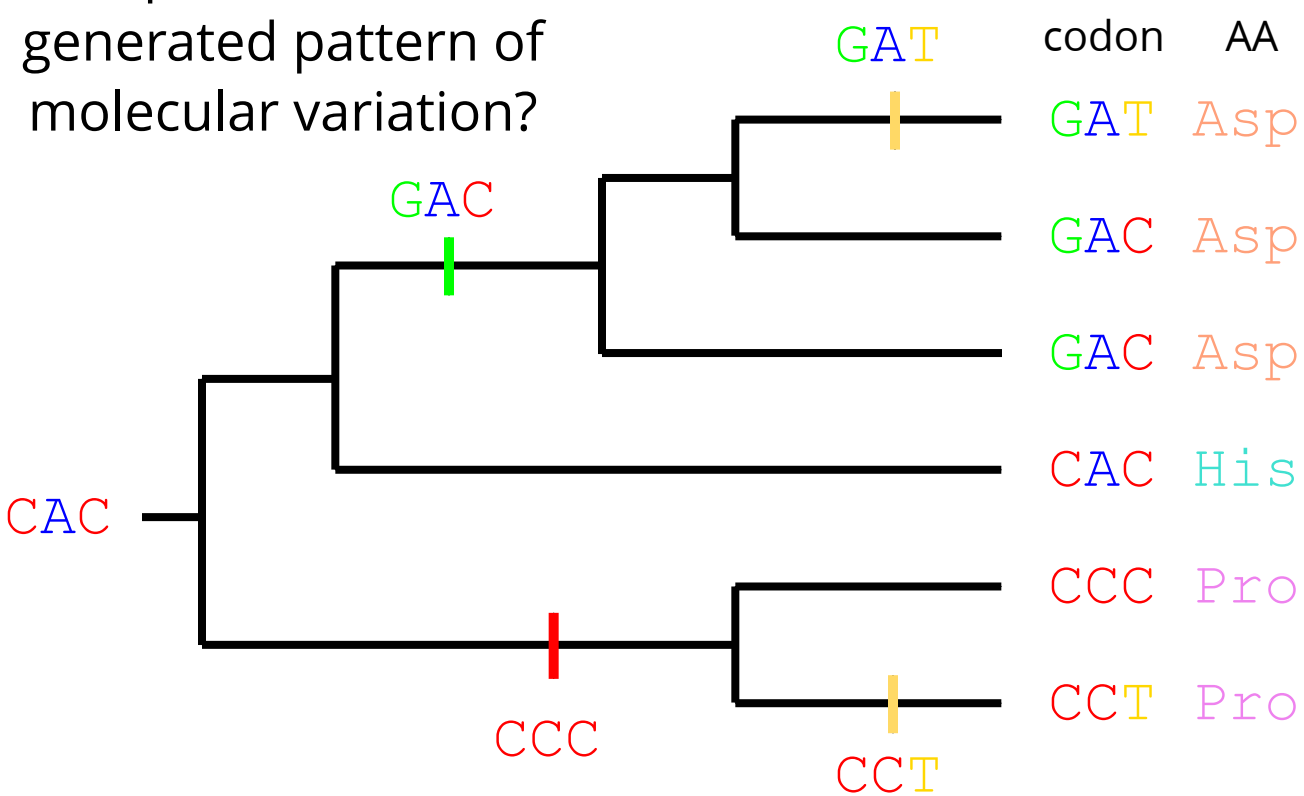
Not designed for multiple sequence tests

Assumes slow mutation rate, shallow timescales

Describes a pattern instead of modeling process

Phylogenetic models of codon evolution

What sequence of events led generated pattern of molecular variation?



Muse & Gaut (1994)
Goldman & Yang (1994)

Phylogenetic models of codon evolution

Define probabilities of codon change for each increment of time using a ***rate matrix*** model

$$P_{ij}(dt) = \begin{cases} \pi_j dt & \text{synonymous change} \\ \pi_j \omega dt & \text{nonsynonymous change} \\ 0 & 2+ \text{ changes needed} \end{cases}$$

Muse & Gaut (1994)
Goldman & Yang (1994)

Phylogenetic models of codon evolution

transition probability of codon *i* to codon *j*, *P[i,j]*

$$P_{ij}(dt) = \begin{cases} \pi_j dt & \text{synonymous change} \\ \pi_j \omega dt & \text{nonsynonymous change} \\ 0 & 2+ \text{ changes needed} \end{cases}$$

instant of time, *dt*

Muse & Gaut (1994)
Goldman & Yang (1994)

Phylogenetic models of codon evolution

transition probability of codon *i* to codon *j*, *P[i,j]*

equilibrium probability of having codon *j*, *π[j]*

$$P_{ij}(dt) = \begin{cases} \pi_j dt & \text{synonymous change} \\ \pi_j \omega dt & \text{nonsynonymous change} \\ 0 & 2+ \text{ changes needed} \end{cases}$$

instant of time, *dt*

the relative rate of nonsynonymous vs. synonymous change, ω

Muse & Gaut (1994)
Goldman & Yang (1994)

Phylogenetic models of codon evolution

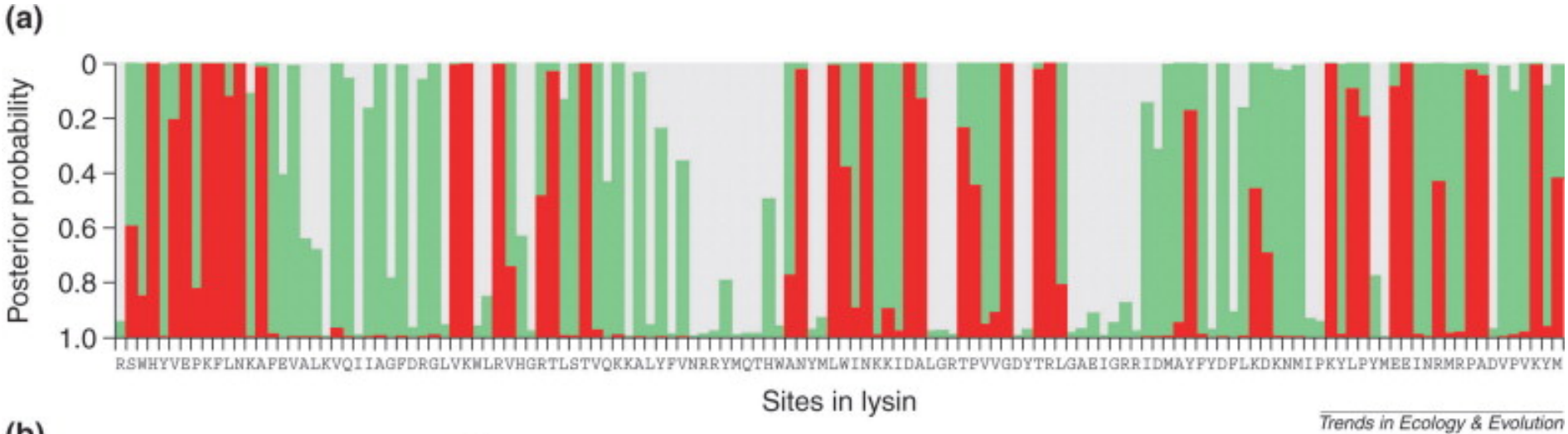
The relative rate of nonsynonymous vs. synonymous change, ω, is the dN/dS ratio!

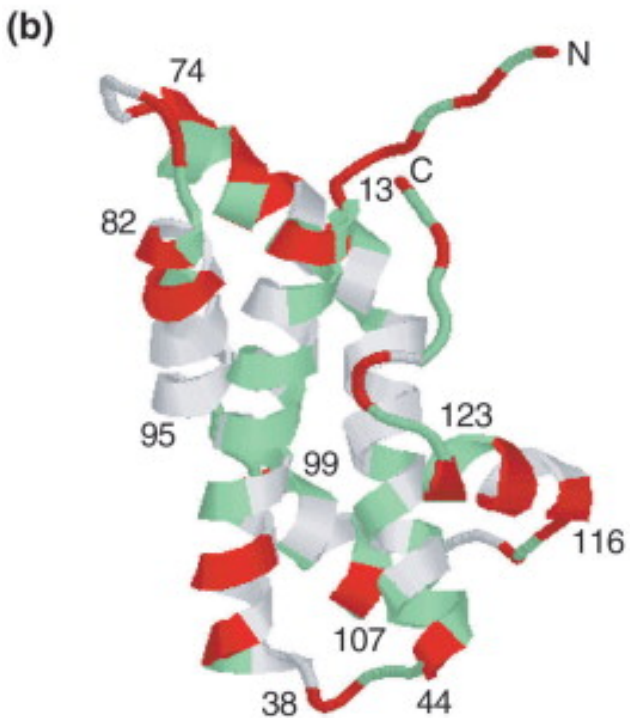
Similar to before, parameter estimates for ω may be interpreted as follows:

ω = 1, neutrality

ω < 1, purifying selection

ω < 1, positive selection





Selection estimates per site
(sperm lysin from 25 abalone spp.)

purifying selection (dN/dS = 0.085)

nearly neutral (dN/dS = 0.911)

positive selection (dN/dS = 3.065)

Yang & Bielawski (2000, *Trends Ecol. Evol.*)

Phylogenetic Analysis by Maximum Likelihood (PAML)

Ziheng Yang

PAML is a powerful program for analyzing molecular data with phylogenetic models.

PAML provides a suite of sophisticated modeling tools in the *codeml* module to estimate *dN/dS* ratios

- global, ω
- branchwise, $\omega[i]$
- sitewise, $\omega[j]$
- branch-and-sitewise, $\omega[i,j]$

Yang (1997, *Bioinfo.*)

Lab 11A

github.com/WUSTL-Biol4220/home/labs/lab_11A.md