Lecture 18
protein evolution

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

Lecture 18 outline

Last time: Biopython

This time: protein evolution

- *dN/dS* and hypotheses
- counting method
- phylogenetic method

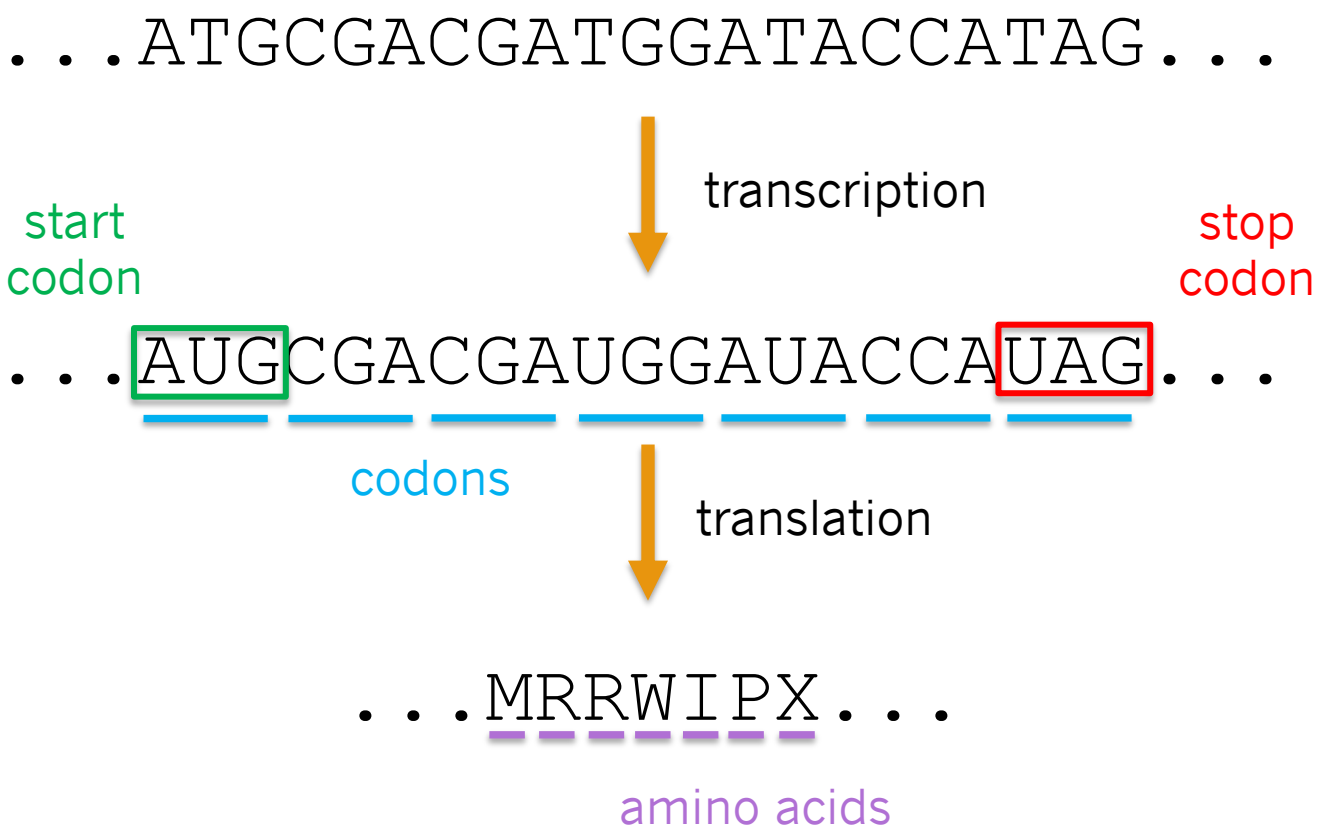
Protein evolution

Selection may act upon DNA mutations in protein-coding genes

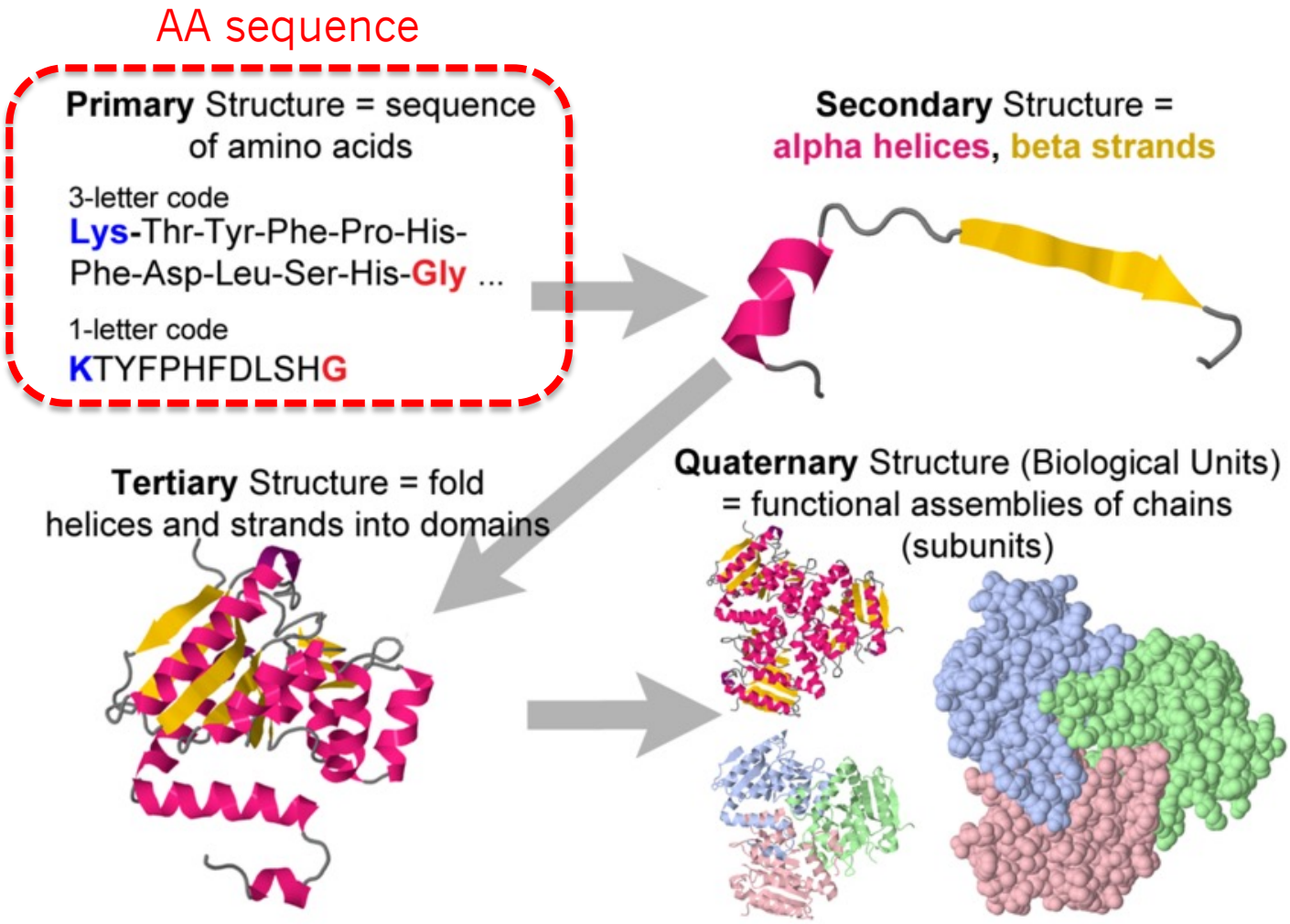
- DNA sequences ***mutate*** and are ***inherited***
- DNA from protein-coding genes is ***transcribed*** into RNA then ***translated*** into AA through the ***genetic code***
- AA sequences determine ***protein structure***
- Protein structure (largely) determines ***protein function***
- Protein function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

Can we detect if a protein evolves faster or slower than it would in the absence of selection?

DNA to RNA to AA



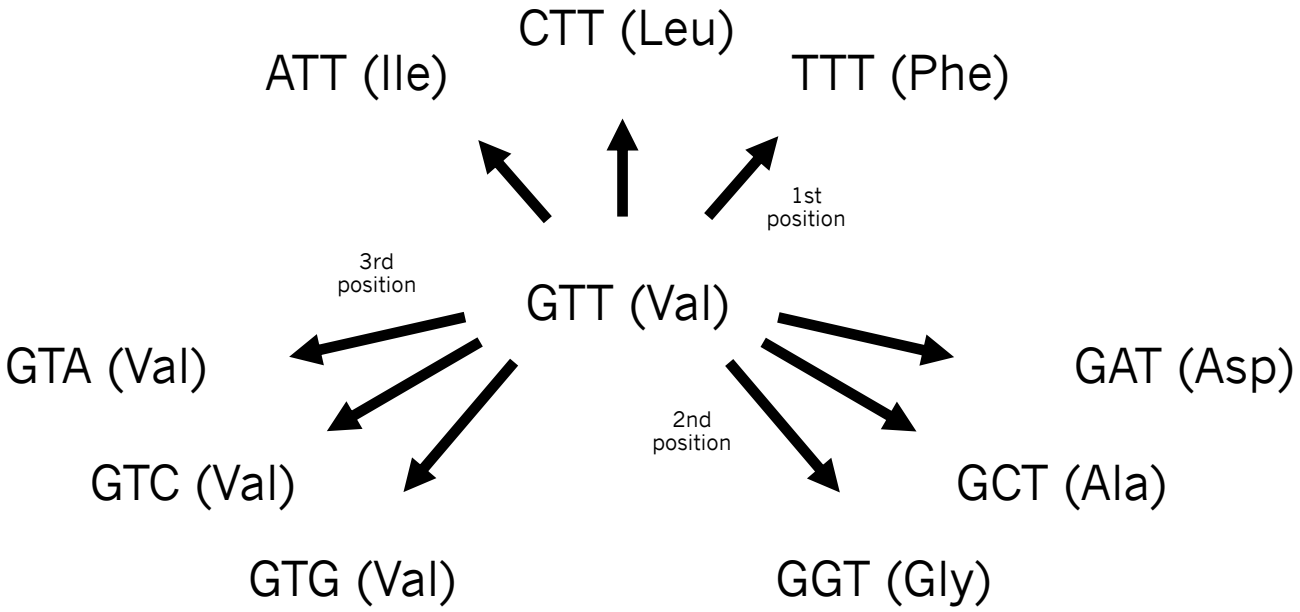
AA sequences influence function



AA sequence ultimately shapes higher-order structure and protein function

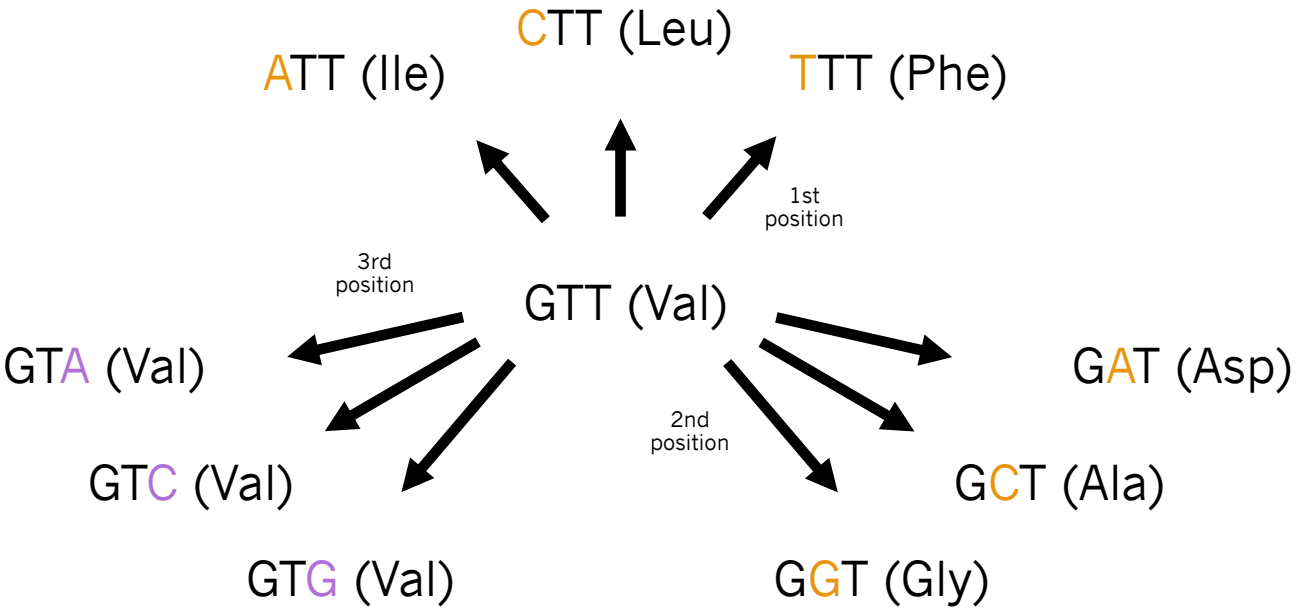
What mutations might induce changes in protein function?

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



Synonymous substitutions are silent and do not induce AA change;

Nonsynonymous substitutions are visible and do cause AA change



Common 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

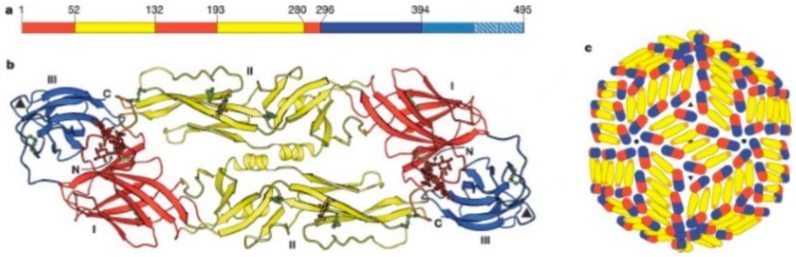
Histone, H1.1

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

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].

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

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

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

Neutral theory

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

The ***neutral theory of molecular evolution*** argues that most alleles evolve according to *neutral processes*

Neutral processes include mutation, migration, recombination, and genetic drift – but not selection!

Neutral processes adequately explain many patterns of molecular variation, both within and between species.

Counting method

Are nonsynonymous substitutions relatively rare or common?

A simple count-based test for pairs 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*)
4. Interpret *dN/dS* in terms of purifying, positive, or neutral selection

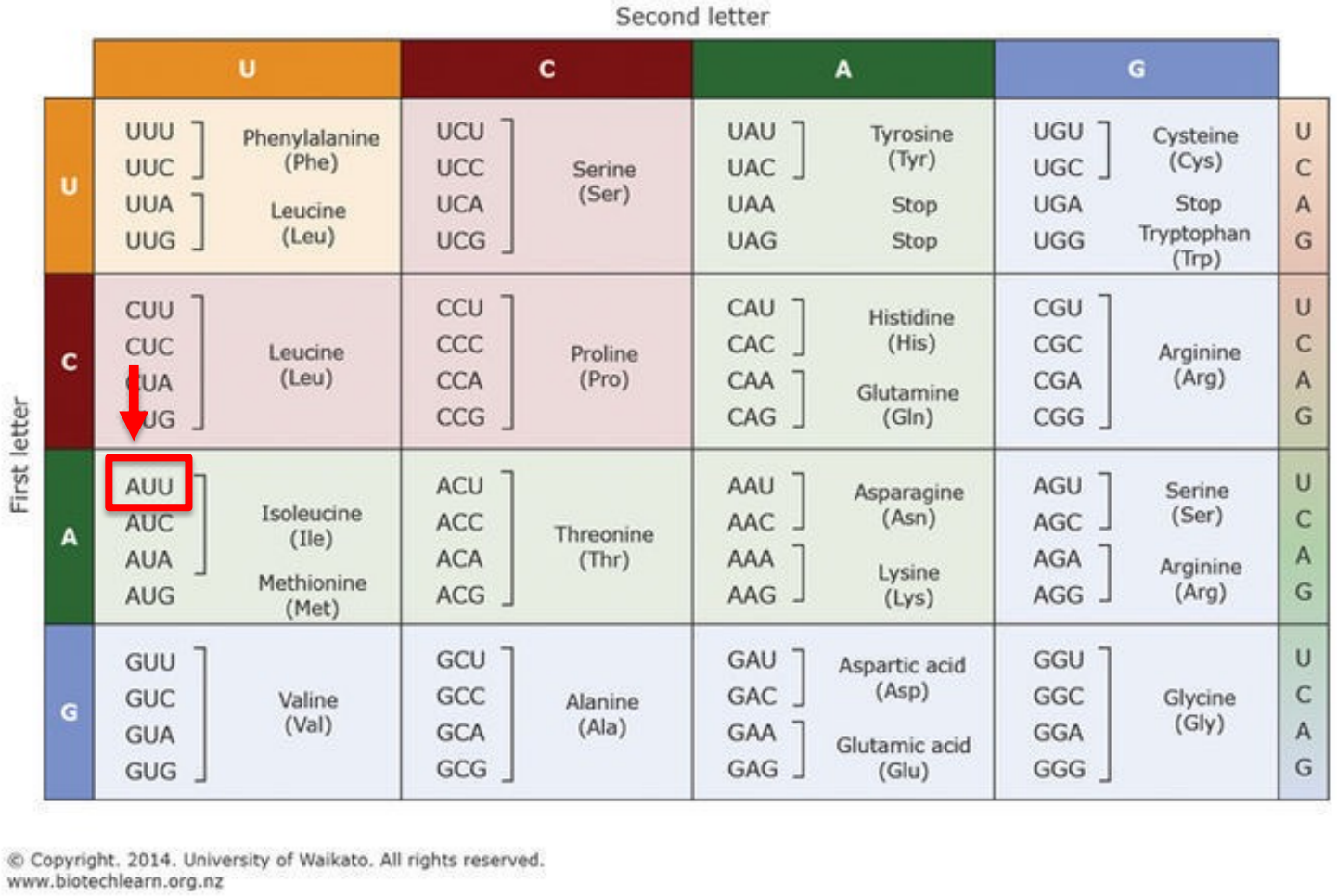
Nei and Gojobori (*Mol. Biol. Evol.* 1986)

Counting method

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

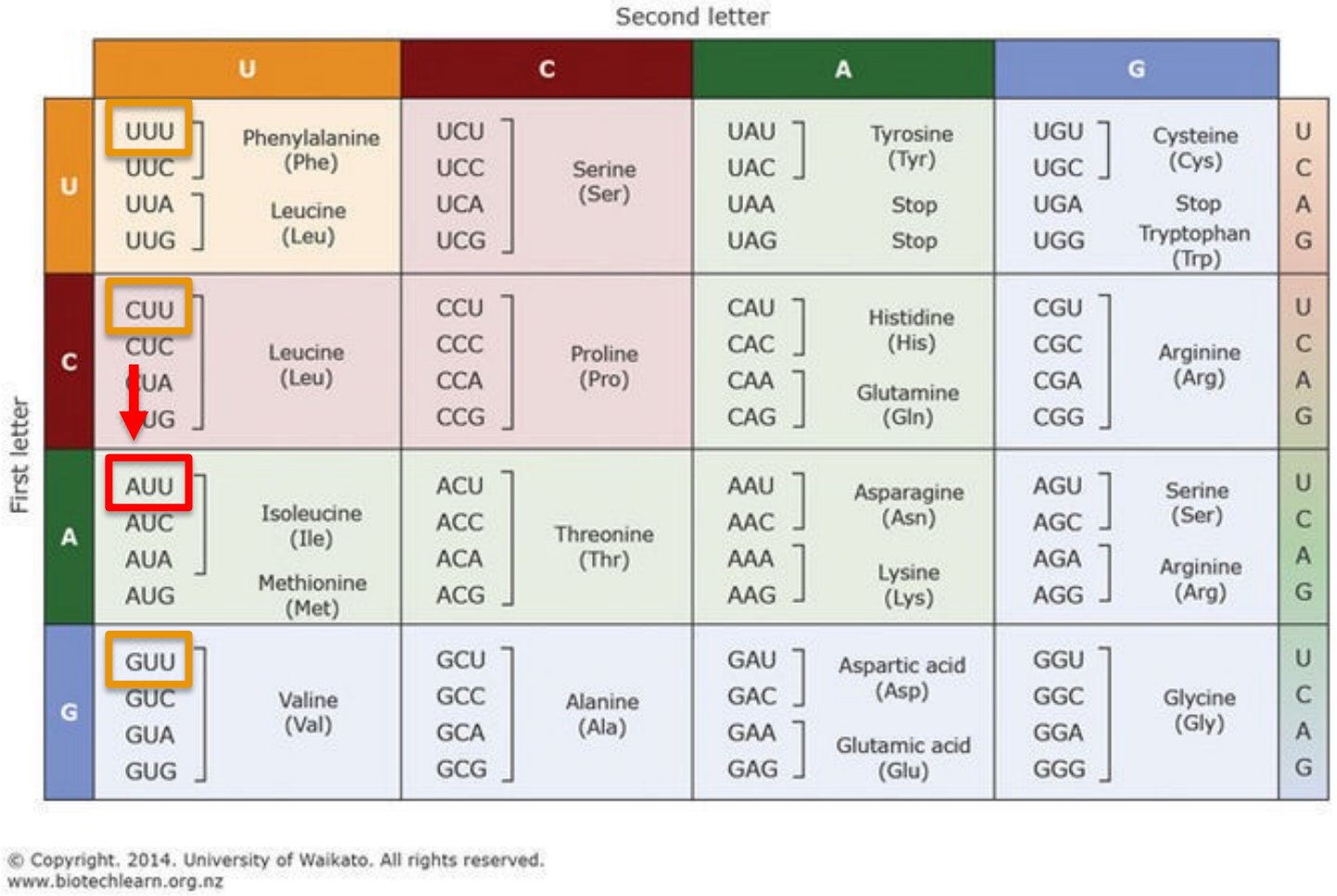
```
Sp_1 GTT ATT GAT GCT TCA GTC
Sp_2 GTT ACT GAC GCA CCA GTC
```

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



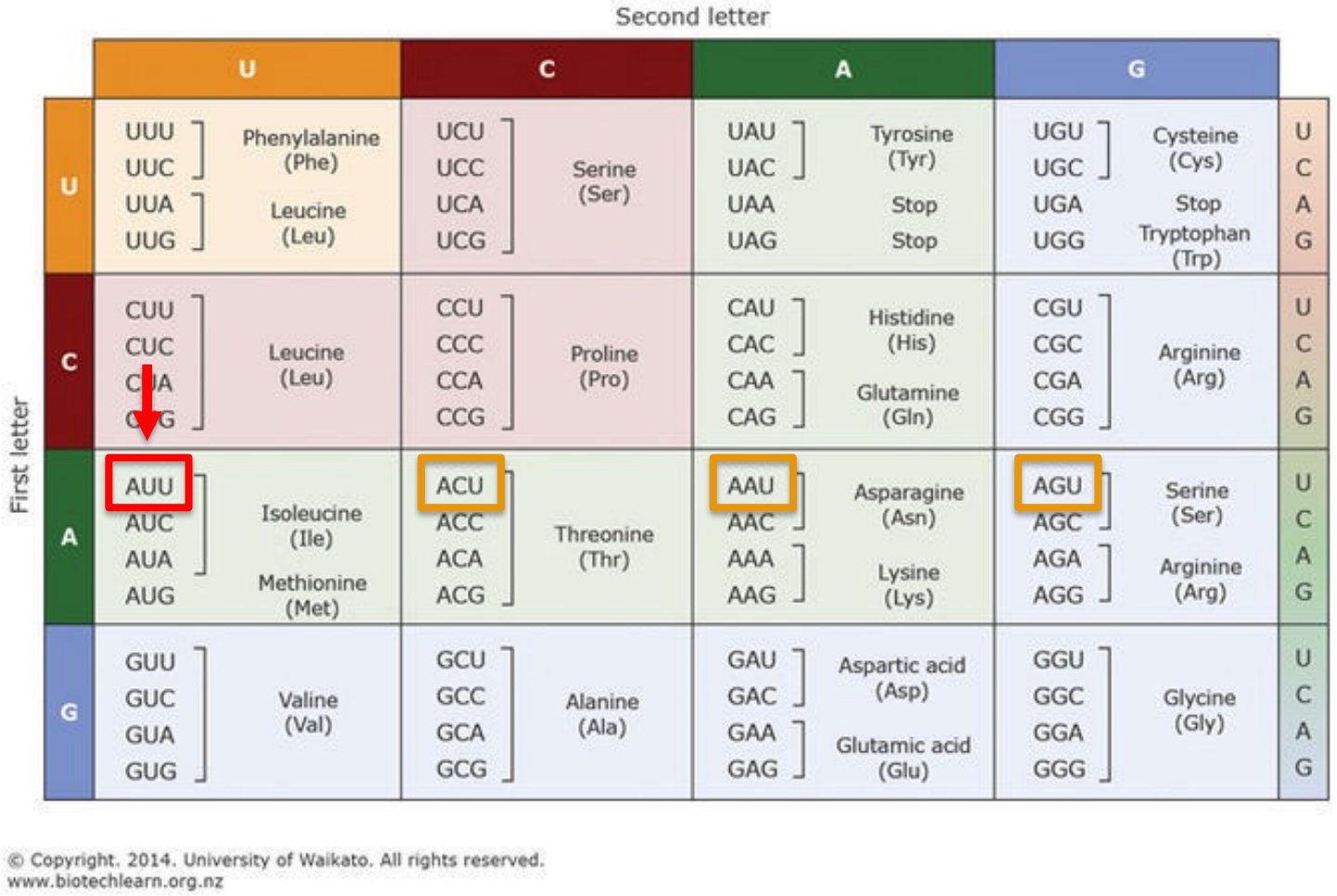
What % of ATT mutations in the ***first codon position*** result in ***synonymous*** changes?

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



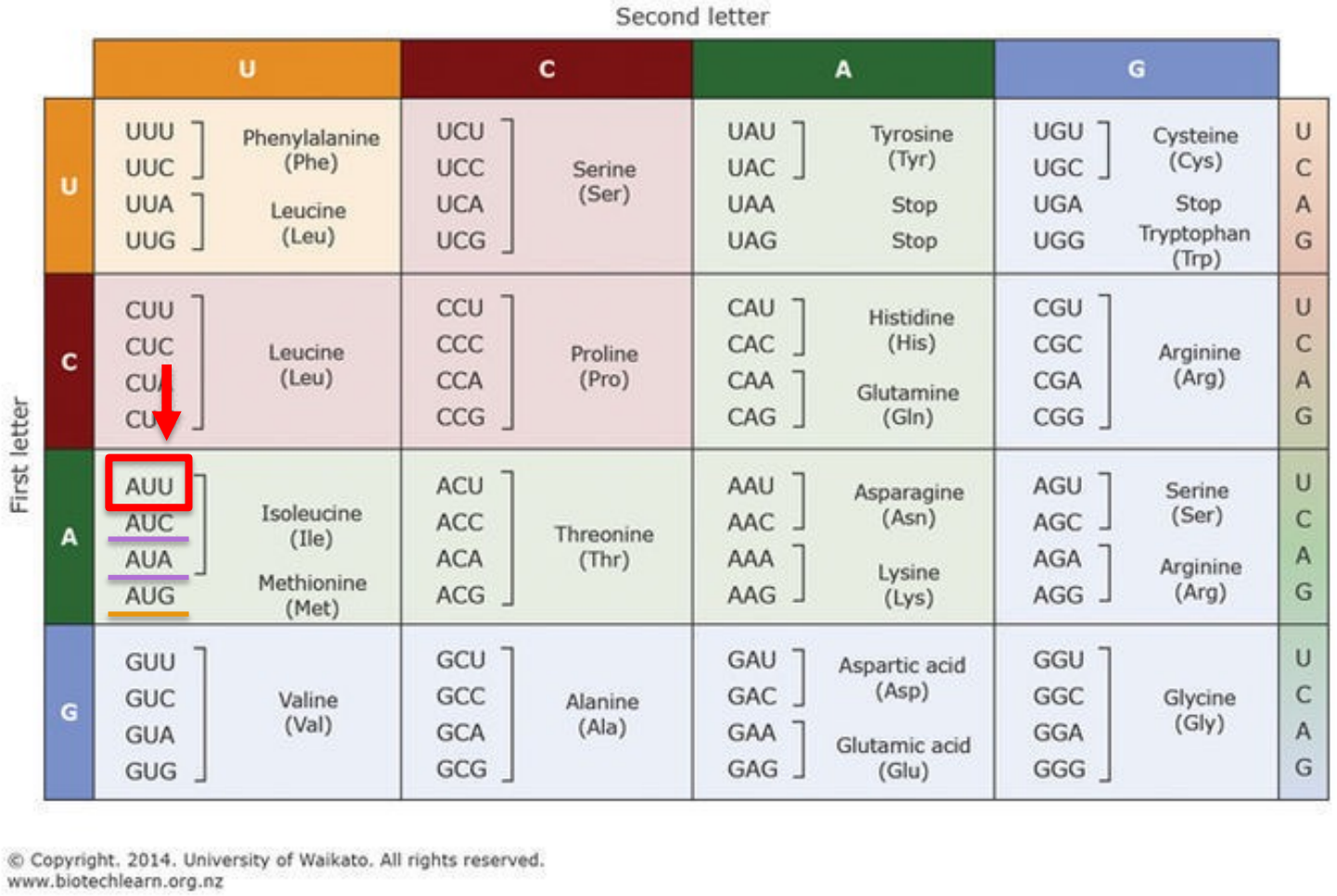
What % of ATT mutations in the ***first codon position*** result in ***synonymous*** changes?

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*) in a sequence of length *L*

```
                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 (*Mol. Biol. Evol.* 1986)

Counting method

divide by three codon site positions

number of analyzed variable sites

L = **12**

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

N = L – S = 12 - 3.16 = **8.83**

divide by two sequences

		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 (*Mol. Biol. Evol.* 1986)

Counting method

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

Sd = 2
Nd = 2

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

(can ignore codons w/ no changes)

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

Counting method

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

$$dN = Nd / N = 2 / 8.83$$
$$dS = Sd / S = 2 / 3.16$$
$$\mathbf{dN/dS = 0.36 < 1}$$

The estimate of $dN/dS < 1$ is consistent with purifying selection.

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

Counting method limitations

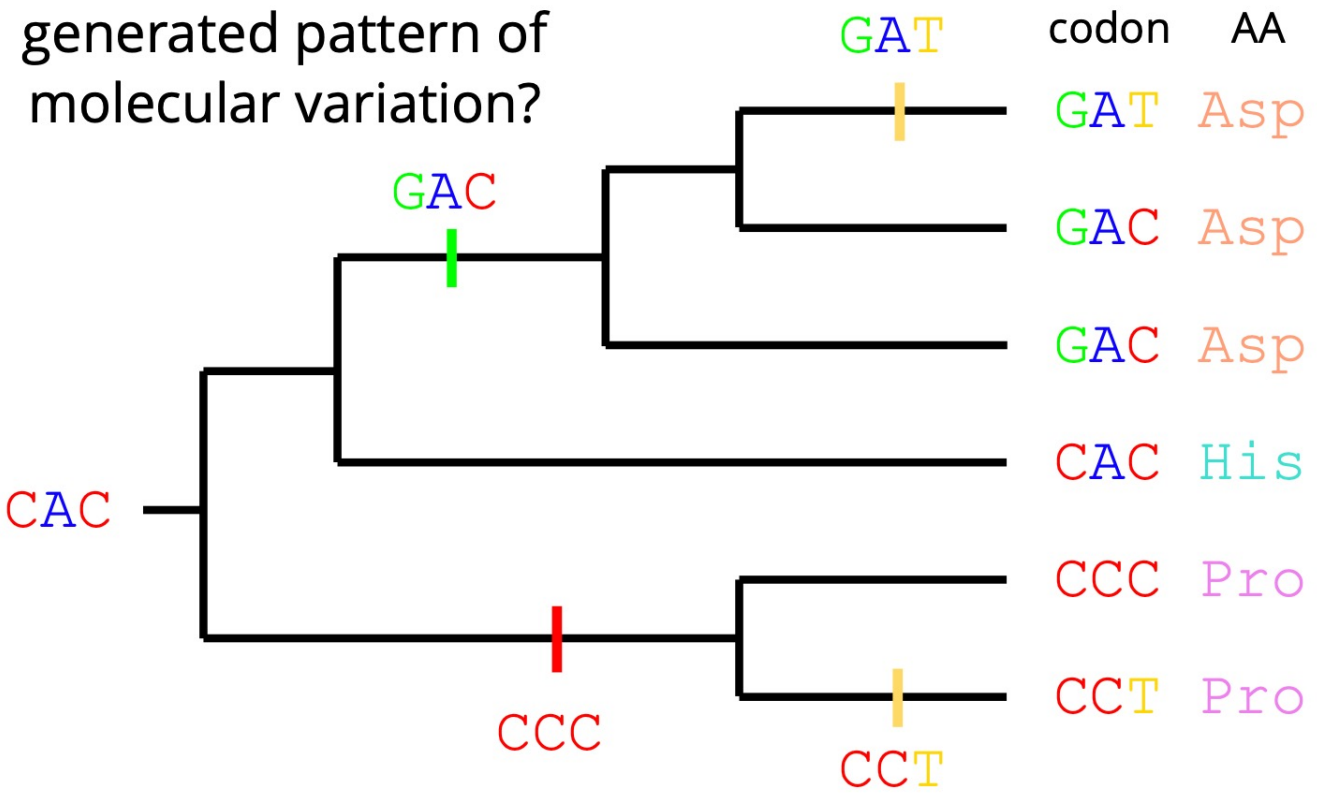
Not designed for multiple sequence tests

Assumes slow mutation rate, shallow timescales

Describes pattern instead of modelling the 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 rates (instantaneous probabilities) of codon change for each moment of time using a ***Markov 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

probability of changing from codon *i* into codon *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

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

Phylogenetic models of codon evolution

probability of changing from codon *i* into codon *j*

equilibrium probability of having codon *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

relative rate of nonsynonymous vs. synonymous change

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

Codon rate matrix structure (61 x 61)

Estimate the *dN/dS* ratio of nonsynonymous versus synonymous substitutions with the parameter, ω, using a phylogenetic framework

Numbers give codon site position change

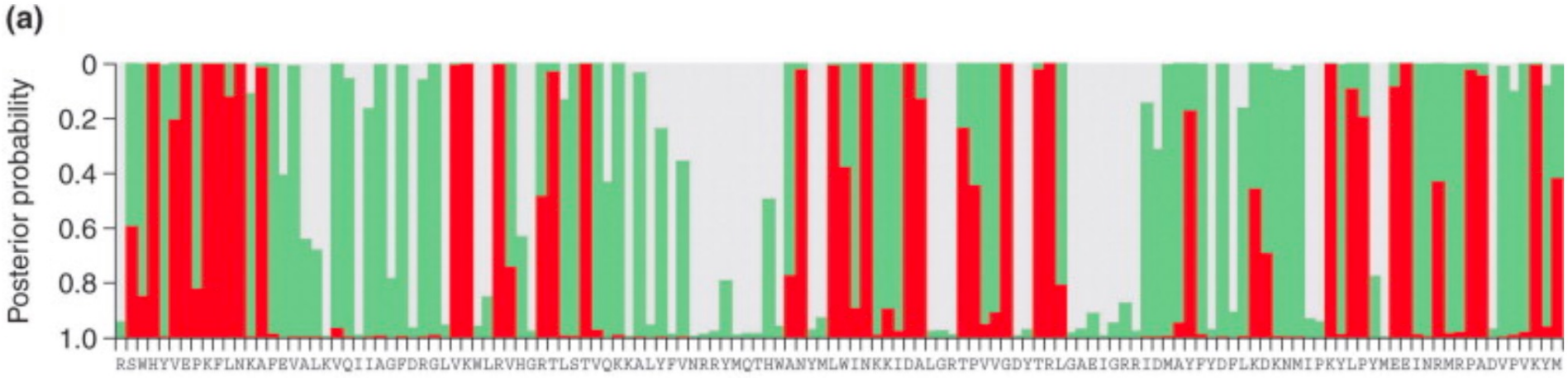
Empty cells indicate impossible transitions

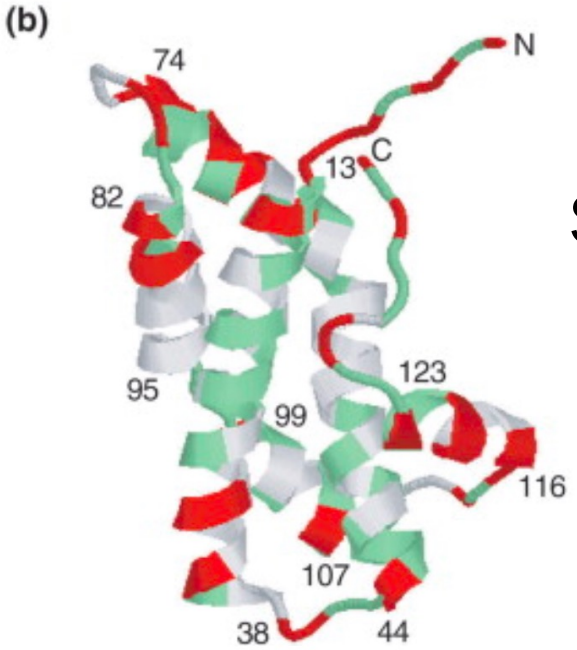
Example transition series:

AAA -> AAG -> AGA

	AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT	AGA	AGC	AGG	AGT	ATA	ATC	ATG	ATT	CAA	CAC	...	TTG	TTT
AAA	–	3	3	3	2				2				2				1				
AAC	3	–	3	3		2				2				2				1			
AAG	3	3	–	3			2				2				2						
AAT	3	3	3	–				2				2				2					
ACA	2				–	3	3	3	2				2								
ACC		2			3	–	3	3		2				2							
ACG			2		3	3	–	3			2				2						
ACT				2	3	3	3	–				2				2					
AGA	2				2				–	3	3	3	2								
AGC		2				2			3	–	3	3		2							
AGG			2				2		3	3	–	3			2						
AGT				2				2	3	3	3	–				2					
ATA	2				2				2				–	3	3	3					
ATC		2				2				2			3	–	3	3					
ATG			2				2				2		3	3	–	3				1	
ATT				2				2				2	3	3	3	–					1
CAA	1																–	3			
CAC		1															3	–			
...																			–		
TTG															1					–	3
TTT																1				3	–

structure of a codon rate matrix





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

Shows probability of each site belonging to any of three selection regimes

purifying selection (*dN/dS = 0.085*)
nearly neutral (*dN/dS = 0.911*)
positive selection (*dN/dS = 3.065*)

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

Overview for Lab 18