

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

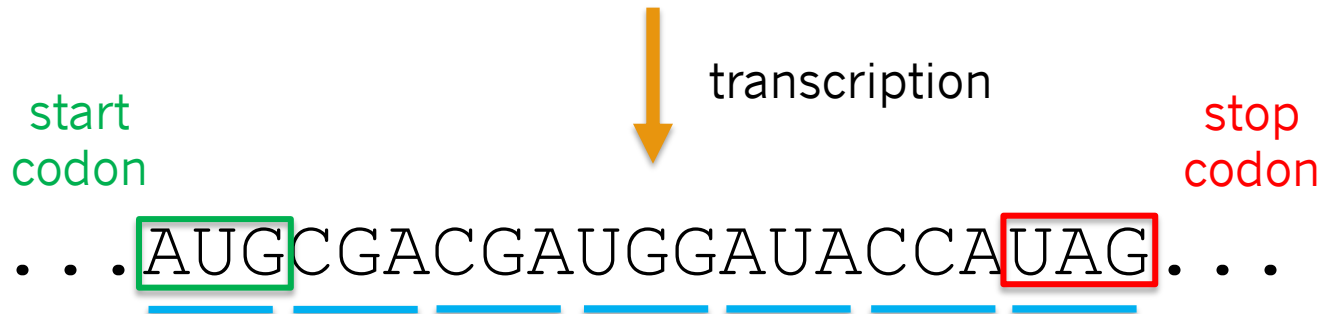
Selection may act upon DNA mutations
in protein-coding genes

- 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

. . . ATGCGACGATGGATACCATAG . . .

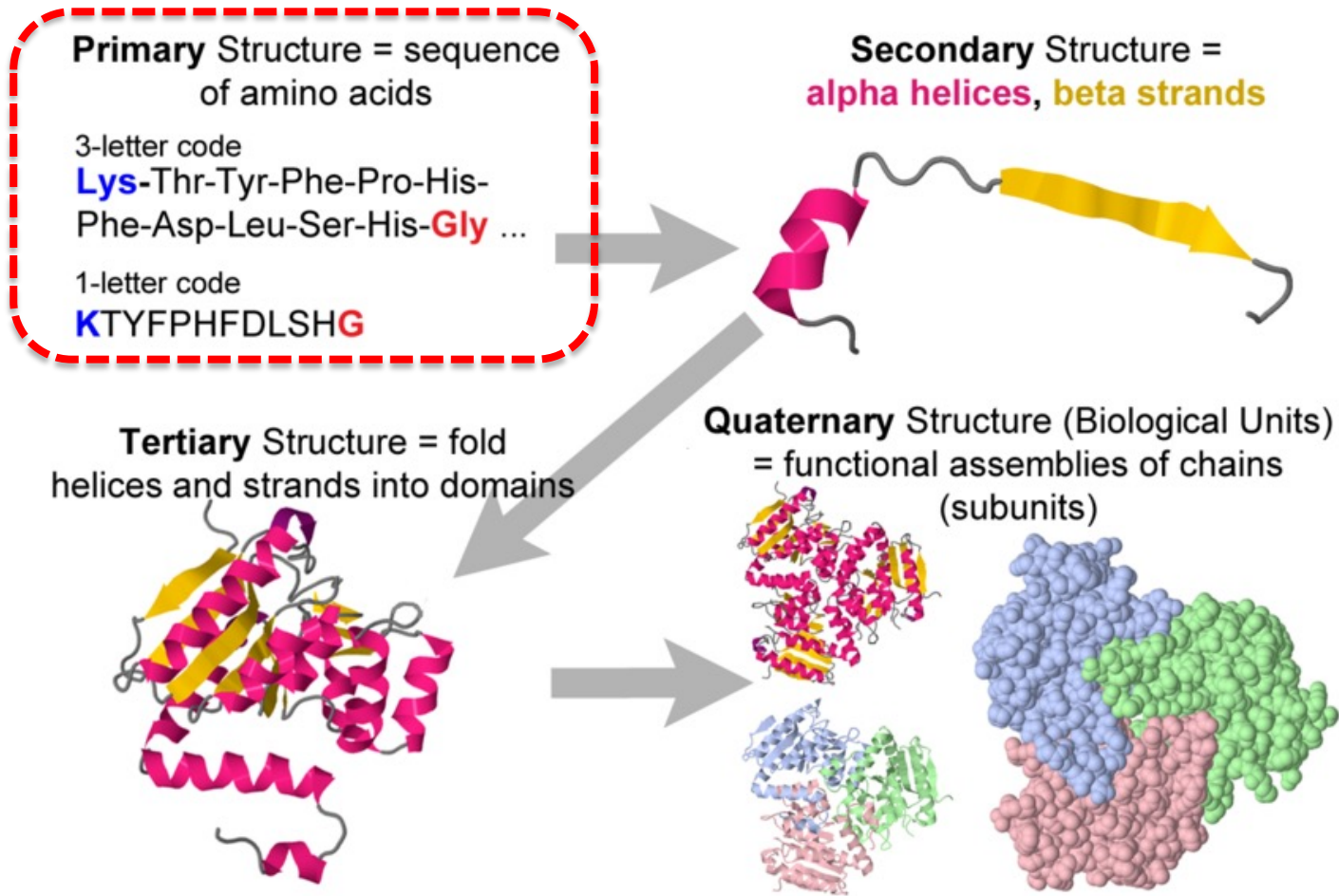


. . . MRRWIPX . . .

amino acids

AA sequences influence function

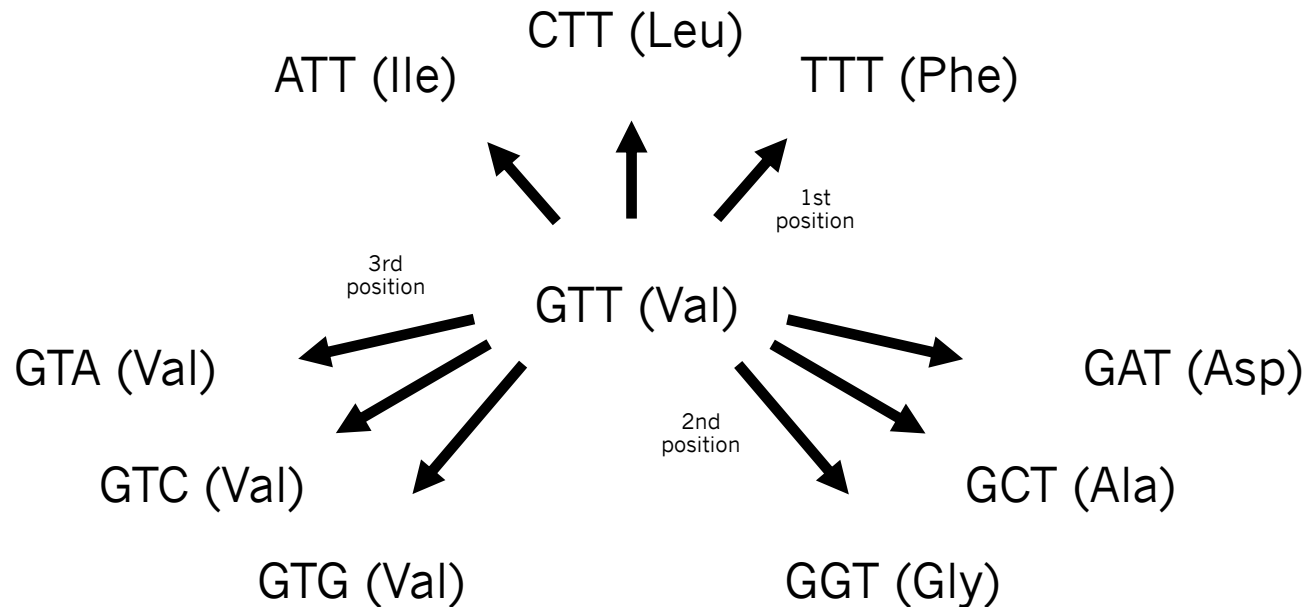
AA sequence



AA sequence ultimately shapes
higher-order structure and protein function

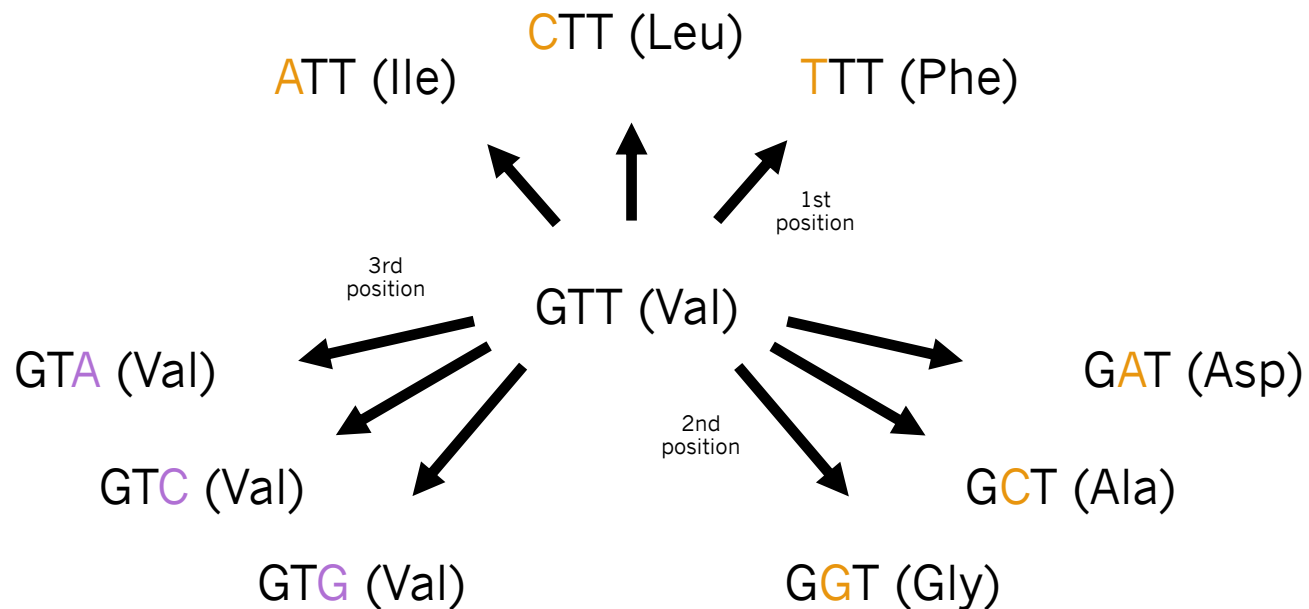
What mutations might induce changes in protein function?

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***)

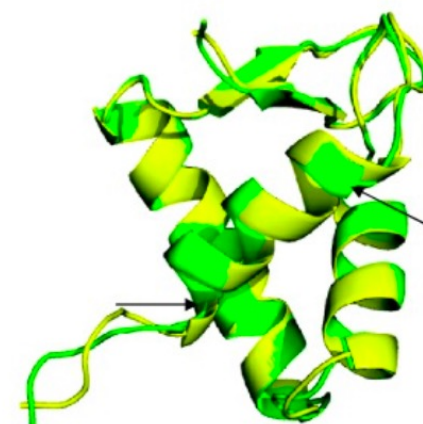
Purifying selection

Genes that encode proteins responsible 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

$$dN/dS < 1$$

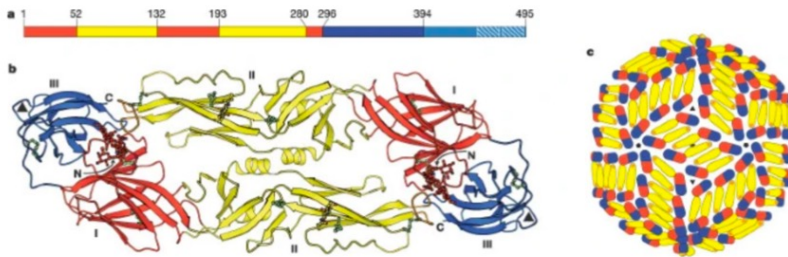


Histone, H1.1

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 dimer¹⁰. 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 lattice¹³.

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

Gene	d_N/d_S^a		P^d
	Max. d_N/d_S^b	Proportion of Codons ^c	
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).

$$dN/dS > 1$$

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

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

		Second letter				
		U	C	A	G	
First letter	U	UUU] Phenylalanine (Phe) UUC] UUA] Leucine (Leu) UUG]	UCU] Serine (Ser) UCC] UCA] UCG]	UAU] Tyrosine (Tyr) UAC] UAA] Stop UAG] Stop	UGU] Cysteine (Cys) UGC] UGA] Stop UGG] Tryptophan (Trp)	U C A G
	C	CUU] Leucine (Leu) CUC] CUA] CUG]	CCU] Proline (Pro) CCC] CCA] CCG]	CAU] Histidine (His) CAC] CAA] Glutamine (Gln) CAG]	CGU] Arginine (Arg) CGC] CGA] CGG]	U C A G
	A	AUU] Isoleucine (Ile) AUC] AUA] Methionine (Met) AUG]	ACU] Threonine (Thr) ACC] ACA] ACG]	AAU] Asparagine (Asn) AAC] AAA] Lysine (Lys) AAG]	AGU] Serine (Ser) AGC] AGA] Arginine (Arg) AGG]	U C A G
	G	GUU] Valine (Val) GUC] GUA] GUG]	GCU] Alanine (Ala) GCC] GCA] GCG]	GAU] Aspartic acid (Asp) GAC] GAA] Glutamic acid (Glu) GAG]	GGU] Glycine (Gly) GGC] GGA] GGG]	U C A G

© Copyright, 2014, University of Waikato. All rights reserved.
www.biotechlearn.org.nz

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

$f[1] = ?$

$f[2] = ?$

$f[3] = ?$

		Second letter					
		U	C	A	G		
First letter	U	UUU] Phenylalanine (Phe) UUC] UUA] Leucine (Leu) UUG]	UCU] Serine (Ser) UCC] UCA] UCG]	UAU] Tyrosine (Tyr) UAC] UAA] Stop UAG] Stop	UGU] Cysteine (Cys) UGC] UGA] Stop UGG] Tryptophan (Trp)	U	C
	C	CUU] Leucine (Leu) CUC] CUA] CUG]	CCU] Proline (Pro) CCC] CCA] CCG]	CAU] Histidine (His) CAC] CAA] Glutamine (Gln) CAG]	CGU] Arginine (Arg) CGC] CGA] CGG]	U	C
	A	AUU] Isoleucine (Ile) AUC] AUA] Methionine (Met) AUG]	ACU] Threonine (Thr) ACC] ACA] ACG]	AAU] Asparagine (Asn) AAC] AAA] Lysine (Lys) AAG]	AGU] Serine (Ser) AGC] AGA] Arginine (Arg) AGG]	U	C
	G	GUU] Valine (Val) GUC] GUA] GUG]	GCU] Alanine (Ala) GCC] GCA] GCG]	GAU] Aspartic acid (Asp) GAC] GAA] Glutamic acid (Glu) GAG]	GGU] Glycine (Gly) GGC] GGA] GGG]	U	C

© Copyright, 2014. University of Waikato. All rights reserved.
www.biotechlearn.org.nz

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

$f[1] = 0$

$f[2] = ?$

$f[3] = ?$

		Second letter				
		U	C	A	G	
First letter	U	UUU] Phenylalanine (Phe) UUC] UUA] Leucine (Leu) UUG]	UCU] Serine (Ser) UCC] UCA] UCG]	UAU] Tyrosine (Tyr) UAC] UAA] Stop UAG] Stop	UGU] Cysteine (Cys) UGC] UGA] Stop UGG] Tryptophan (Trp)	U C A G
	C	CUU] Leucine (Leu) CUC] CUA] CUG]	CCU] Proline (Pro) CCC] CCA] CCG]	CAU] Histidine (His) CAC] CUA] Glutamine (Gln) CUG]	CGU] Arginine (Arg) CGC] CGA] CGG]	U C A G
	A	AUU] Isoleucine (Ile) AUC] AUA] Methionine (Met) AUG]	ACU] Threonine (Thr) ACC] ACA] ACG]	AAU] Asparagine (Asn) AAC] AAA] Lysine (Lys) AAG]	AGU] Serine (Ser) AGC] AGA] Arginine (Arg) AGG]	U C A G
	G	GUU] Valine (Val) GUC] GUA] GUG]	GCU] Alanine (Ala) GCC] GCA] GCG]	GAU] Aspartic acid (Asp) GAC] GAA] Glutamic acid (Glu) GAG]	GGU] Glycine (Gly) GGC] GGA] GGG]	U C A G

© Copyright, 2014, University of Waikato. All rights reserved.
www.biotechlearn.org.nz

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

$$f[1] = 0$$

$$f[2] = 0$$

$$f[3] = ?$$

		Second letter					
		U	C	A	G		
First letter	U	UUU] Phenylalanine (Phe) UUC] UUA] Leucine (Leu) UUG]	UCU] Serine (Ser) UCC] UCA] UCG]	UAU] Tyrosine (Tyr) UAC] UAA Stop UAG Stop	UGU] Cysteine (Cys) UGC] UGA Stop UGG Tryptophan (Trp)	U C A G	
	C	CUU] Leucine (Leu) CUC] CUA] CUG]	CCU] Proline (Pro) CCC] CCA] CCG]	CAU] Histidine (His) CAC] CAA] Glutamine (Gln) CAG]	CGU] Arginine (Arg) CGC] CGA] CGG]	U C A G	
	A	AUU] AUC] Isoleucine (Ile) AUA] AUG] Methionine (Met)	ACU] Threonine (Thr) ACC] ACA] ACG]	AAU] Asparagine (Asn) AAC] AAA] Lysine (Lys) AAG]	AGU] Serine (Ser) AGC] AGA] Arginine (Arg) AGG]	U C A G	
	G	GUU] Valine (Val) GUC] GUA] GUG]	GCU] Alanine (Ala) GCC] GCA] GCG]	GAU] Aspartic acid (Asp) GAC] GAA] Glutamic acid (Glu) GAG]	GGU] Glycine (Gly) GGC] GGA] GGG]	U C A G	

© Copyright, 2014. University of Waikato. All rights reserved.
www.biotechlearn.org.nz

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	A T T	GA T	G C T	T CA	GTC
Sp_2	GTT	A C T	GA C	GC A	C CA	GTC
		0+0+3	0+0+1	0+0+3	0+0+3	

(can ignore codons w/ no changes)

Counting method

divide by three codon site positions

number of analyzed variable sites

$L = 12$

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

divide by two sequences

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

		0+0+2	0+0+1	0+0+3	0+0+3	
Sp_1	GTT	A ^T T	GA ^T	GCT ^T	^T CA	GTC
Sp_2	GTT	A ^C T	GA ^C	GCA ^A	^C CA	GTC
		0+0+3	0+0+1	0+0+3	0+0+3	

(can ignore codons w/ no changes)

Counting method

Compute the number of **synonymous changes** (S_d)
and the number of **nonsynonymous changes** (N_d)

$$S_d = 2$$

$$N_d = 2$$

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

(can ignore codons w/ no changes)

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

$$\begin{aligned}dN &= Nd / N \\ &= 2 / 8.83 \\ &= 0.227\end{aligned}$$

$$\begin{aligned}dS &= Sd / S \\ &= 2 / 3.16 \\ &= 0.633\end{aligned}$$

$$\begin{aligned}dN/dS &= 0.227 / 0.633 \\ &= 0.36 < 1\end{aligned}$$

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

Counting method limitations

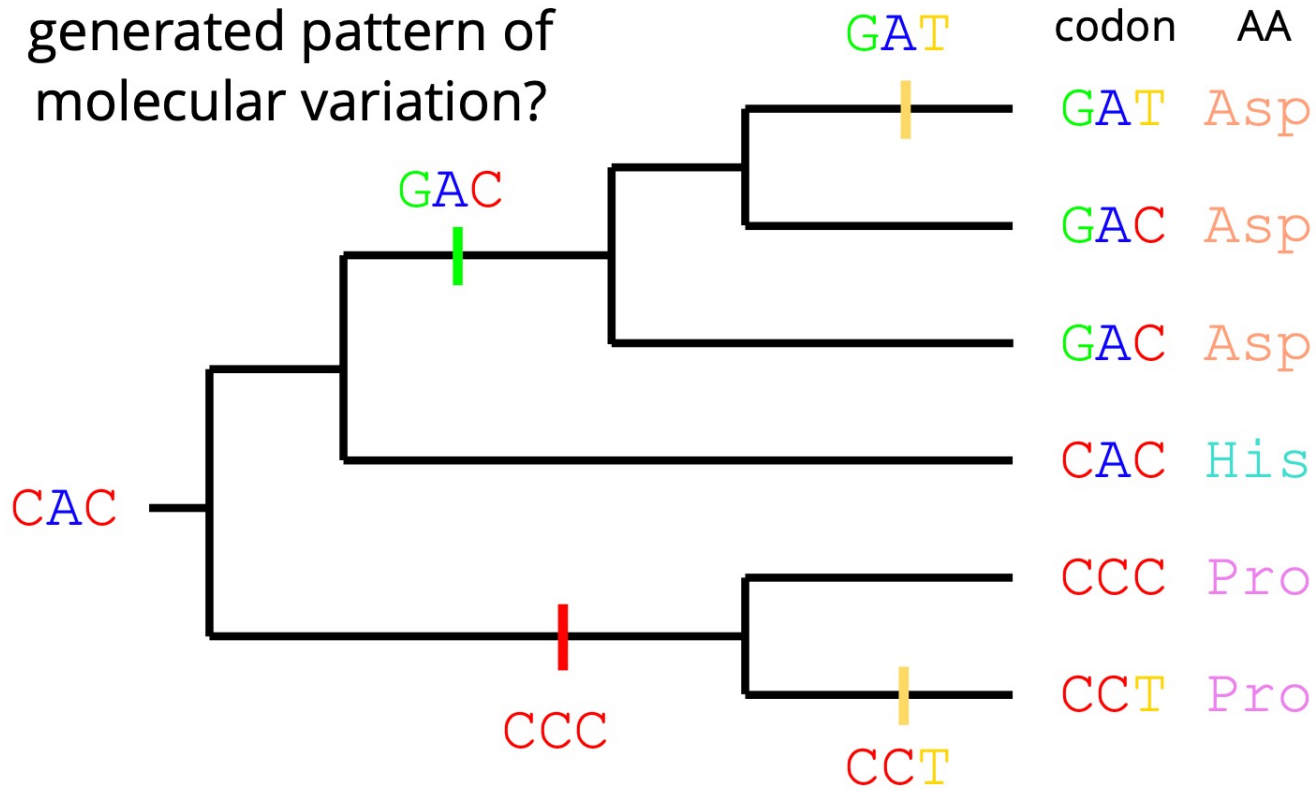
Not designed for multiple sequence tests

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?



Phylogenetic models of codon evolution

Define rates
(instantaneous probabilities)
of codon change for each
moment of time using
a **Markov model**



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

Phylogenetic models of codon evolution

probability of changing
from codon i
into codon j



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



instant of time

Phylogenetic models of codon evolution

probability of changing from codon i into codon j

equilibrium probability of having codon j

instant of time

relative rate of nonsynonymous vs. synonymous change

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

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

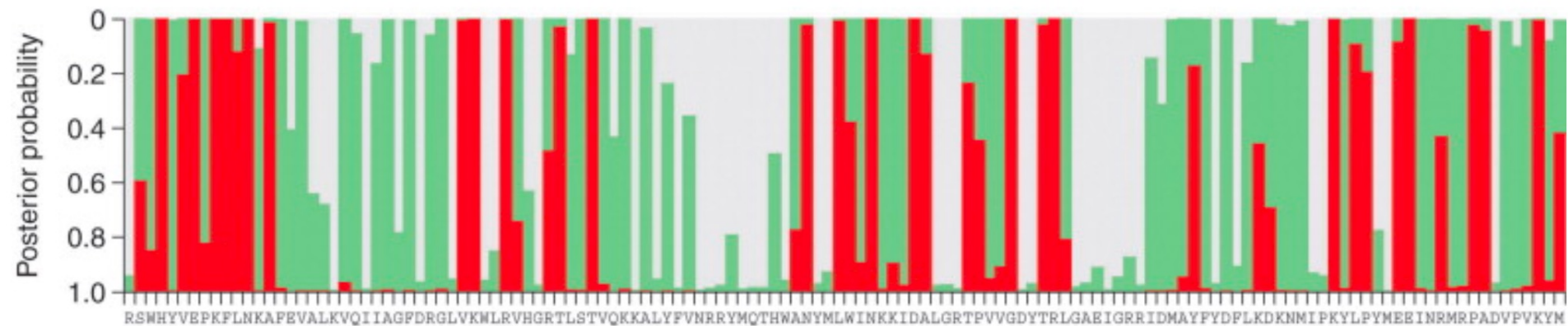
Example transition
series:

AAA -> AAG -> AGG

	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

(a)

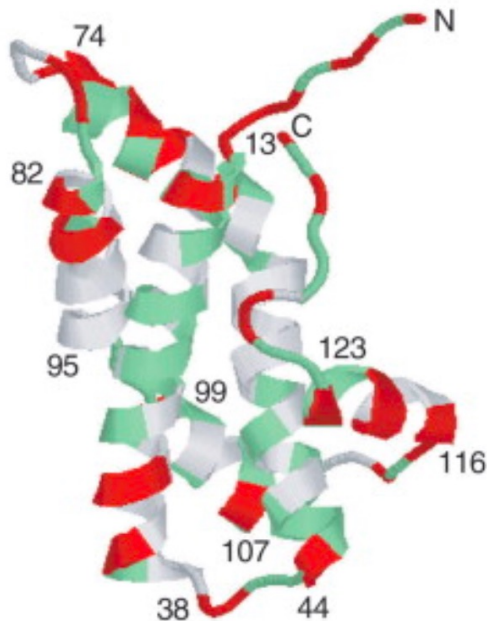


Selection estimates per site
(sperm lysin from 25 abalone *spp.*)

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 ($dN/dS = 3.065$)

(b)



Overview for Lab 18