Codon Models

The Genetic Code

First 12 nucleotides at the 5' end of the *rbc*L gene in corn:

```
5'-ATG|TCA|CCA|CAA-3'   coding strand
3'-TAC|AGT|GGT|GTT-5'   template strand
```

DNA double helix

transcription

```
5'-AUG|UCA|CCA|CAA-3'   mRNA
```

translation

```
N-Met|Ser|Pro|Gln-C     polypeptide
```

Codon models treat codons as the independent units, not individual nucleotide sites.

Genetic Code

	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

http://www.langara.bc.ca/biology/mario/Assets/Geneticode.jpg

Table 1 from: Lewis (2001b)

Table I. Part of Muse and Gaut's 61 × 61 instantaneous rate matrix[a]

Codon before substitution (the 'from' state)	Codon after substitution (the 'to' state)							
	TTT (Phe)	TTC (Phe)	TTA (Leu)	TTG (Leu)	CTT (Leu)	CTC (Leu)	…	GGG (Gly)
TTT (Phe)	– – –	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$	$\beta\pi_C$	0	…	0
TTC (Phe)	$\alpha\pi_T$	– – –	$\beta\pi_A$	$\beta\pi_G$	0	$\beta\pi_C$	…	0
TTA (Leu)	$\beta\pi_T$	$\beta\pi_C$	– – –	$\alpha\pi_G$	0	0	…	0
TTG (Leu)	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	– – –	0	0	…	0
CTT (Leu)	$\beta\pi_T$	0	0	0	– – –	$\alpha\pi_C$	…	0
CTC (Leu)	0	$\beta\pi_T$	0	0	$\alpha\pi_T$	– – –	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
GGG (Gly)	0	0	0	0	0	0	…	– – –

Compare to HKY85 model:

Muse and Gaut (1994) Goldman and Yang (1994)

Interpreting codon model results

$\omega = \beta/\alpha$ is the nonsynonymous/synonymous rate ratio

omega	mode of selection	example
$\omega < 1$		
$\omega = 1$		
$\omega > 1$		

Amino Acid Models

Amino Acid Models

AA models began with Margaret Dayhoff's 1978 PAM (Point Accepted Mutation) matrices.

PAM: a *point mutation* that results in replacement of an amino acid at a particular site in a protein sequence *and which has been accepted* by natural selection

PAM1 is a matrix recording the probability of each possible amino acid substitution for proteins that differ in 1% of their amino acid positions.

1572 subst., 71 families

Dayhoff et al. (1978)

Simulation to show why PAM1 approximates rate matrix

1 million sites simulated for two taxa under HKY:

	A	C	G	T
A	246563	306	2366	605
C	166	246110	429	3184
G	801	317	248759	639
T	157	1577	463	247558

kappa = 5

pi = .1 .2 .3 .4

edge length = 0.01

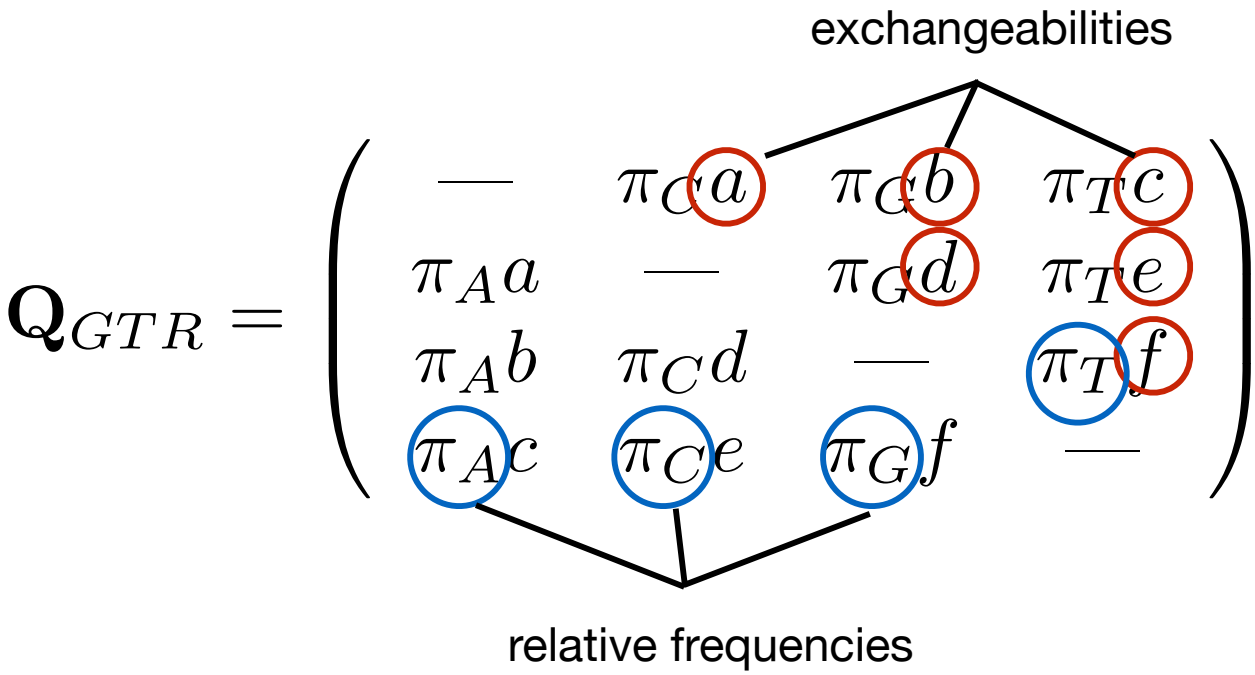
---	0.02779	0.21490	0.05495
0.01508	---	0.03896	0.28919
0.07275	0.02879	---	0.05804
0.01426	0.14323	0.04205	---

counts normalized so that sum of off-diagonal elements is 1.0

---	0.02857	0.21429	0.05714
0.01429	---	0.04286	0.28571
0.07143	0.02857	---	0.05714
0.01429	0.14286	0.04286	---

true rate matrix

PAM1 matrix comparable to the **instantaneous rate matrix** of nucleotide models:



PAM30, PAM60, etc., are comparable to **transition matrices**: they are obtained by extrapolating the rate matrix (PAM1) to larger amounts of time

Amino acid models

WAG (Whelan And Goldman) exchangeabilities:

0.55157																			
0.50985	0.63535																		
0.73900	0.14730	5.42942																	
1.02704	0.52819	0.26526	0.03029																
0.90860	3.03550	1.54364	0.61678	0.09882															
1.58285	0.43916	0.94720	6.17416	0.02135	5.46947														
1.41672	0.58467	1.12556	0.86558	0.30667	0.33005	0.56772													
0.31695	2.13715	3.95629	0.93068	0.24897	4.29411	0.57003	0.24941												
0.19334	0.18698	0.55424	0.03944	0.17014	0.11392	0.12740	0.03045	0.13819											
0.39792	0.49767	0.13153	0.08480	0.38429	0.86949	0.15426	0.06130	0.49946	3.17097										
0.90627	5.35142	3.01201	0.47986	0.07403	3.89490	2.58443	0.37356	0.89043	0.32383	0.25756									
0.89350	0.68316	0.19822	0.10375	0.39048	1.54526	0.31512	0.17410	0.40414	4.25746	4.85402	0.93428								
0.21049	0.10271	0.09616	0.04673	0.39802	0.09992	0.08113	0.04993	0.67937	1.05947	2.11517	0.08884	1.19063							
1.43855	0.67949	0.19508	0.42398	0.10940	0.93337	0.68236	0.24357	0.69620	0.09993	0.41584	0.55690	0.17133	0.16144						
3.37079	1.22419	3.97423	1.07176	1.40766	1.02887	0.70494	1.34182	0.74017	0.31944	0.34474	0.96713	0.49391	0.54593	1.61328					
2.12111	0.55441	2.03006	0.37487	0.51298	0.85793	0.82277	0.22583	0.47331	1.45816	0.32662	1.38698	1.51612	0.17190	0.79538	4.37802				
0.11313	1.16392	0.07192	0.12977	0.71707	0.21574	0.15656	0.33698	0.26257	0.21248	0.66531	0.13751	0.51571	1.52964	0.13941	0.52374	0.11086			
0.24074	0.38153	1.08600	0.32571	0.54383	0.22771	0.19630	0.10360	3.87344	0.42017	0.39862	0.13326	0.42844	6.45428	0.21605	0.78699	0.29115	2.48539		
2.00601	0.25185	0.19625	0.15234	1.00214	0.30128	0.58873	0.18725	0.11836	7.82130	1.80034	0.30543	2.05845	0.64989	0.31489	0.23274	1.38823	0.36537	0.31473	

Determined from comparing 3905 closely-related protein sequences.

http://www.ebi.ac.uk/goldman/WAG/wag.dat

Relative amino acid frequencies

0.08663	0.04397	0.03909	0.05705	0.01931	0.03673	0.05806	0.08325	0.02443	0.04847	0.08621	0.06203	0.01950	0.03843	0.04576	0.06952	0.06101	0.01439	0.03527	0.07090
Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Whelan and Goldman (2001)

Amino acid model evolution

PAM ➢ Dayhoff et al. (1978)

- used closely-related amino-acid sequences

JTT ➢ Jones et al. (1992)

- based on much larger database

WAG ➢ Whelan and Goldman (2001)

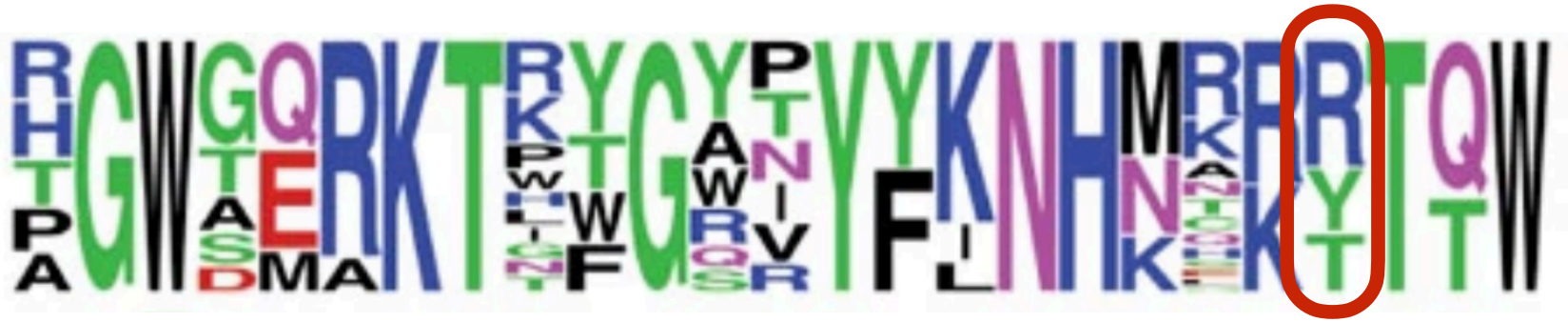
- **Q** MLE allowed more divergent sequences

LG ➢ Le and Gascuel (2008)

- added rate heterogeneity to **Q** MLE
- based on 50,000 sequences

IQ-TREE implements 26 additional amino acid models!

Protein mixture models



Which amino acid equilibrium relative frequencies would provide a better fit for the site indicated?

$$\boldsymbol{\pi} = \left\{ \underset{\texttt{A}}{\frac{1}{20}}, \underset{\texttt{R}}{\frac{1}{20}}, \underset{\texttt{N}}{\frac{1}{20}}, \underset{\texttt{D}}{\frac{1}{20}}, \underset{\texttt{C}}{\frac{1}{20}}, \underset{\texttt{Q}}{\frac{1}{20}}, \underset{\texttt{E}}{\frac{1}{20}}, \underset{\texttt{G}}{\frac{1}{20}}, \underset{\texttt{H}}{\frac{1}{20}}, \underset{\texttt{I}}{\frac{1}{20}}, \underset{\texttt{L}}{\frac{1}{20}}, \underset{\texttt{K}}{\frac{1}{20}}, \underset{\texttt{M}}{\frac{1}{20}}, \underset{\texttt{F}}{\frac{1}{20}}, \underset{\texttt{P}}{\frac{1}{20}}, \underset{\texttt{S}}{\frac{1}{20}}, \underset{\texttt{T}}{\frac{1}{20}}, \underset{\texttt{W}}{\frac{1}{20}}, \underset{\texttt{Y}}{\frac{1}{20}}, \underset{\texttt{V}}{\frac{1}{20}} \right\}$$

$$\boldsymbol{\pi} = \left\{ \underset{\texttt{A}}{0}, \underset{\texttt{R}}{\frac{1}{2}}, \underset{\texttt{N}}{0}, \underset{\texttt{D}}{0}, \underset{\texttt{C}}{0}, \underset{\texttt{Q}}{0}, \underset{\texttt{E}}{0}, \underset{\texttt{G}}{0}, \underset{\texttt{H}}{0}, \underset{\texttt{I}}{0}, \underset{\texttt{L}}{0}, \underset{\texttt{K}}{0}, \underset{\texttt{M}}{0}, \underset{\texttt{F}}{0}, \underset{\texttt{P}}{0}, \underset{\texttt{S}}{0}, \underset{\texttt{T}}{\frac{1}{4}}, \underset{\texttt{W}}{0}, \underset{\texttt{Y}}{\frac{1}{4}}, \underset{\texttt{V}}{0} \right\}$$

Fig. 2c from Fowler et al. (2010)

C20 protein mixture model

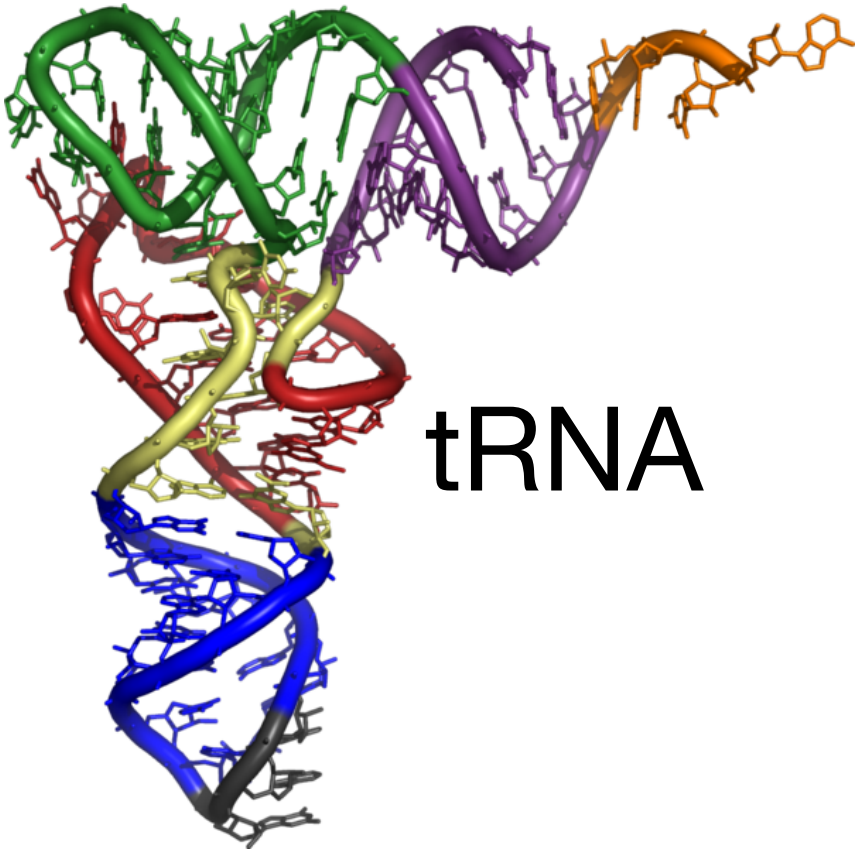
$$L(D_k) = p(\boldsymbol{\pi}_1)p(D_k|\boldsymbol{\pi}_1) + p(\boldsymbol{\pi}_2)p(D_k|\boldsymbol{\pi}_2) + \cdots + p(\boldsymbol{\pi}_{20})p(D_k|\boldsymbol{\pi}_{20})\text{-}$$

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	3	1	0	0	1	0	0	38	0	9	2	0	0	0	0	0	2	42	0	0
2	17	3	0	1	1	2	0	12	1	9	3	1	2	1	1	5	12	27	0	1
3	4	1	1	4	0	2	3	1	28	3	1	3	1	9	28	4	4	2	0	1
4	1	0	0	0	3	0	0	31	0	36	5	0	0	0	0	0	1	20	0	0
5	7	0	27	34	0	3	1	0	4	1	0	4	1	6	2	4	3	1	0	0
6	9	1	3	2	1	4	1	1	3	1	1	8	2	2	2	30	26	2	0	1
7	3	0	29	8	0	8	3	0	4	0	0	25	1	3	2	8	4	0	0	1
8	9	0	8	20	0	2	2	1	14	2	1	4	1	15	9	4	5	2	0	0
9	44	4	0	1	1	11	0	1	0	1	1	1	2	1	1	21	7	5	0	0
10	1	1	0	0	39	0	3	1	0	5	1	0	0	0	0	1	0	1	5	41
11	2	1	0	0	8	0	0	12	0	49	16	0	0	1	0	1	1	6	1	1
12	5	1	1	4	2	1	3	9	8	18	6	2	1	7	9	3	6	9	1	2
13	16	0	2	4	1	3	1	1	3	3	1	2	38	3	3	9	5	3	0	1
14	5	2	2	3	10	2	16	2	3	6	2	5	1	5	5	5	3	3	3	18
15	10	1	5	11	1	2	2	6	7	5	2	3	2	7	6	7	12	11	0	1
16	4	2	0	0	20	1	2	12	1	22	4	1	1	1	1	1	2	12	3	10
17	11	1	6	3	0	47	1	0	3	1	0	8	1	2	3	8	2	0	0	0
18	18	0	8	13	0	6	2	1	8	1	1	5	2	8	6	13	7	2	0	0
19	5	1	6	5	1	4	12	1	9	2	1	16	1	10	9	8	5	1	0	2
20	12	0	13	17	0	3	1	0	7	1	0	4	20	5	3	7	4	1	0	0

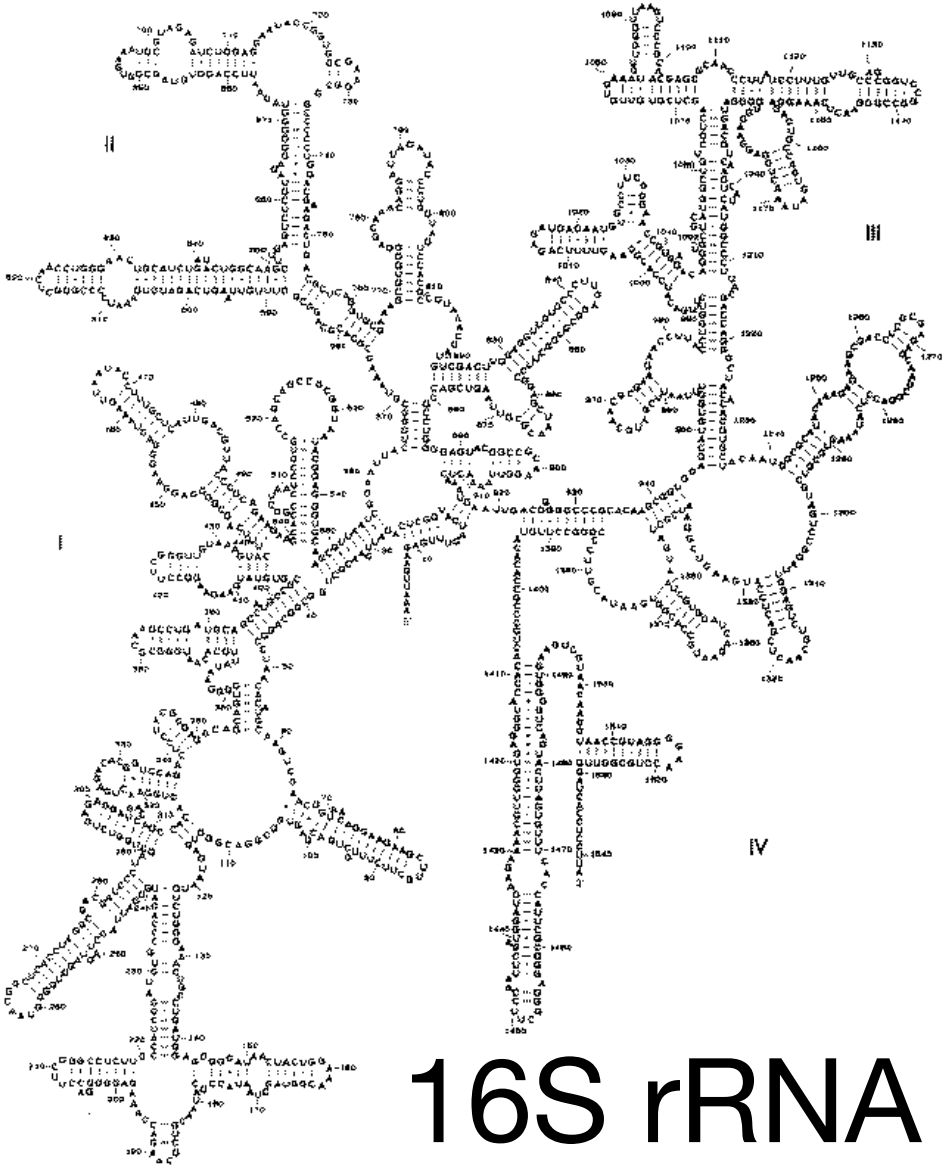
Si Quang et al. (2008)

Stems and loops

http://en.wikipedia.org/wiki/Image:3d_tRNA.png



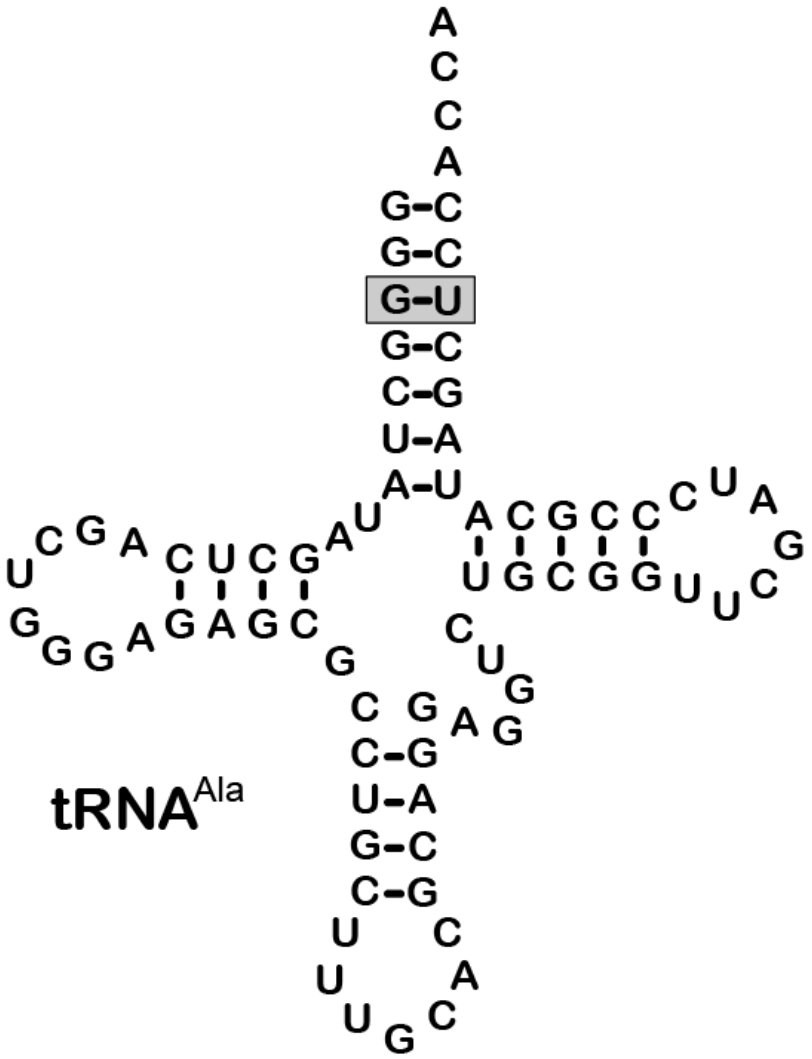
tRNA



16S rRNA

http://bass.bio.uci.edu/~hudel/bs99a/lecture22/lecture3_4.html

Compensatory substitutions



tRNA image adapted from http://web.mit.edu/beuning/www/formerresearch.html

Muse (1995) stem model

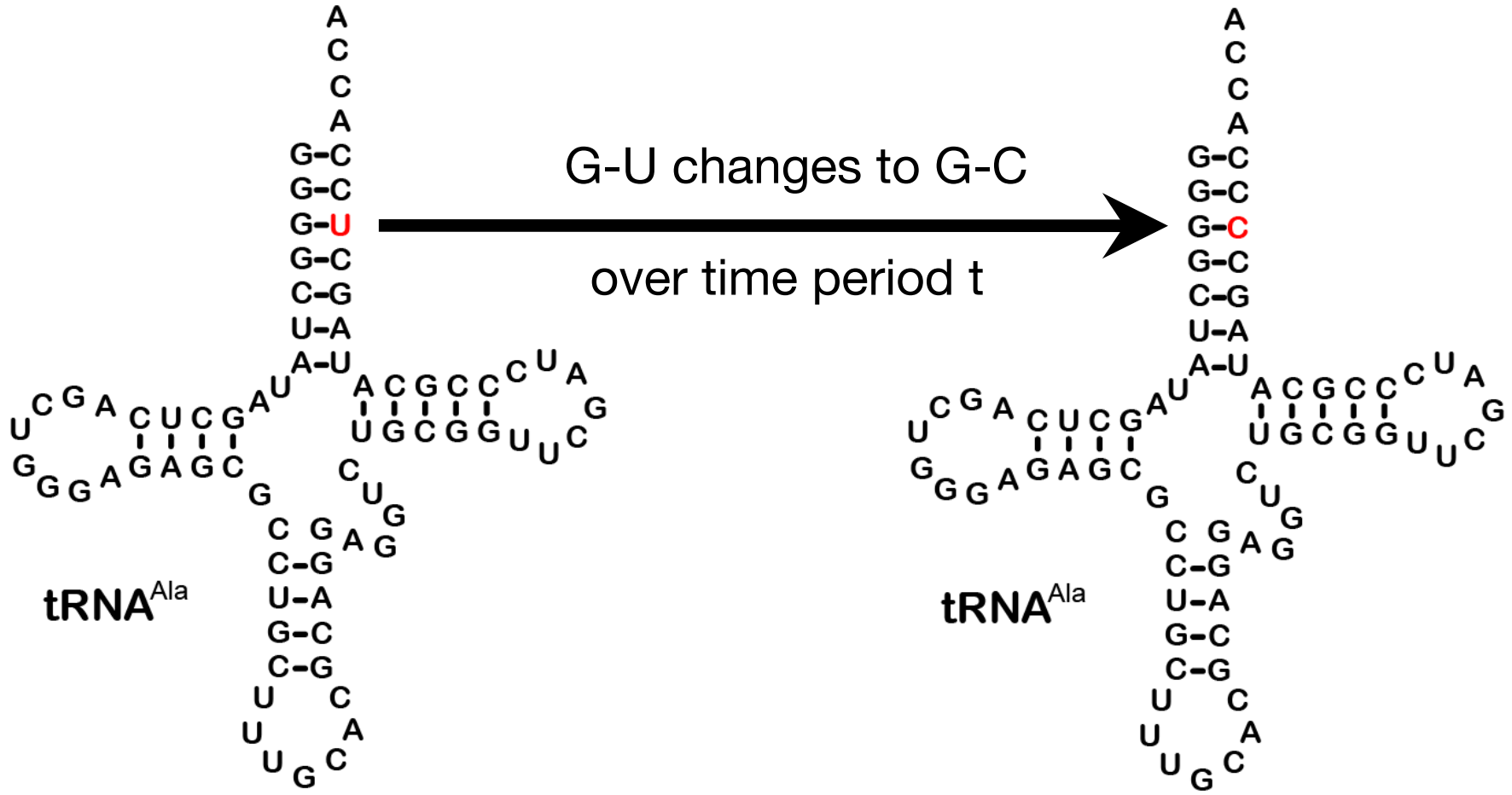
- Apply normal substitution model (e.g. JC69, K80, HKY85, etc.) to loop regions
- Apply a special "stem model" to the stem regions
- The stem model has 16 states:
  AA, AC, AG, AT, CA, CC, ..., TT
  where each state consists of one possible pairing of two nucleotides across a stem

Muse (1995) stem model

- Idea is to compare null (independence) model to the alternative (dependence) model using a likelihood ratio test
- Independence model assumes each site evolves independently
- Dependence model allows rate of evolution to be *higher* for changes that *improve stem stability* and lower for changes that destroy stability

Independence model

(applied to one stem site)



$$L_{\text{stem ``site''}} = \left[\left(\frac{1}{4}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-4\beta t}\right)\right]\left[\left(\frac{1}{4}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4\beta t}\right)\right]$$

G → G (no change) U → C (change)

Independence model rate matrix*

	AA	AC	AG	AU	CA	CC	CG	CU	GA	…
AA	-	β	β	β	β	0	0	0	β	…
AC	β	-	β	β	0	β	0	0	0	…
AG	β	β	-	β	0	0	β	0	0	…
AU	β	β	β	-	0	0	0	β	0	…
CA	β	0	0	0	-	β	β	β	0	…
CC	0	β	0	0	β	-	β	β	0	…
CG	0	0	β	0	β	β	-	β	0	…
CU	0	0	0	β	β	β	β	-	0	…
GA	β	0	0	0	0	0	0	0	-	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

*I have used a JC model here; Muse used HKY85

Dependence model rate matrix

	AA	AC	AG	AU	CA	CC	CG	CU	GA	…
AA	-	β	β	$\lambda\beta$	β	0	0	0	β	…
AC	β	-	β	$\lambda\beta$	0	β	0	0	0	…
AG	β	β	-	$\lambda\beta$	0	0	$\lambda\beta$	0	0	…
AU	β/λ	β/λ	β/λ	-	0	0	0	β/λ	0	…
CA	β	0	0	0	-	β	$\lambda\beta$	β	β	…
CC	0	β	0	0	β	-	$\lambda\beta$	β	0	…
CG	0	0	β/λ	0	0	β/λ	-	β/λ	β/λ	…
CU	0	0	0	$\lambda\beta$	0	β	$\lambda\beta$	-	β	…
GA	β	0	0	0	0	0	0	0	-	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Muse (1995) dependence model

- The instantaneous rate of those changes that increase pairing are multiplied by parameter λ
- The instantaneous rate of those changes that decrease pairing are divided by parameter λ
- The parameter λ thus measures the degree to which pairing is favored in evolution
- Setting $\lambda = 1$ equals independence
- LRT statistic = -2 (log L_{indep} - log L_{dep})
  1 d.f. (λ is the only extra parameter)
- Muse (1995) found that $\lambda = 5.29$ for ribonuclease P (LRT = 224.30, P=0.0)