Variation in rate among sites

The previous distances all assume that the rate of nucleotide substitution is the same for all sites.

What if it is different? Some sites changing rapidly and others changing slowly.

This difference could be because of

- functional constraints on the protein or on the RNA.
- different positions within the codon
- even differences in chromosomal location.

Variation in rate among sites

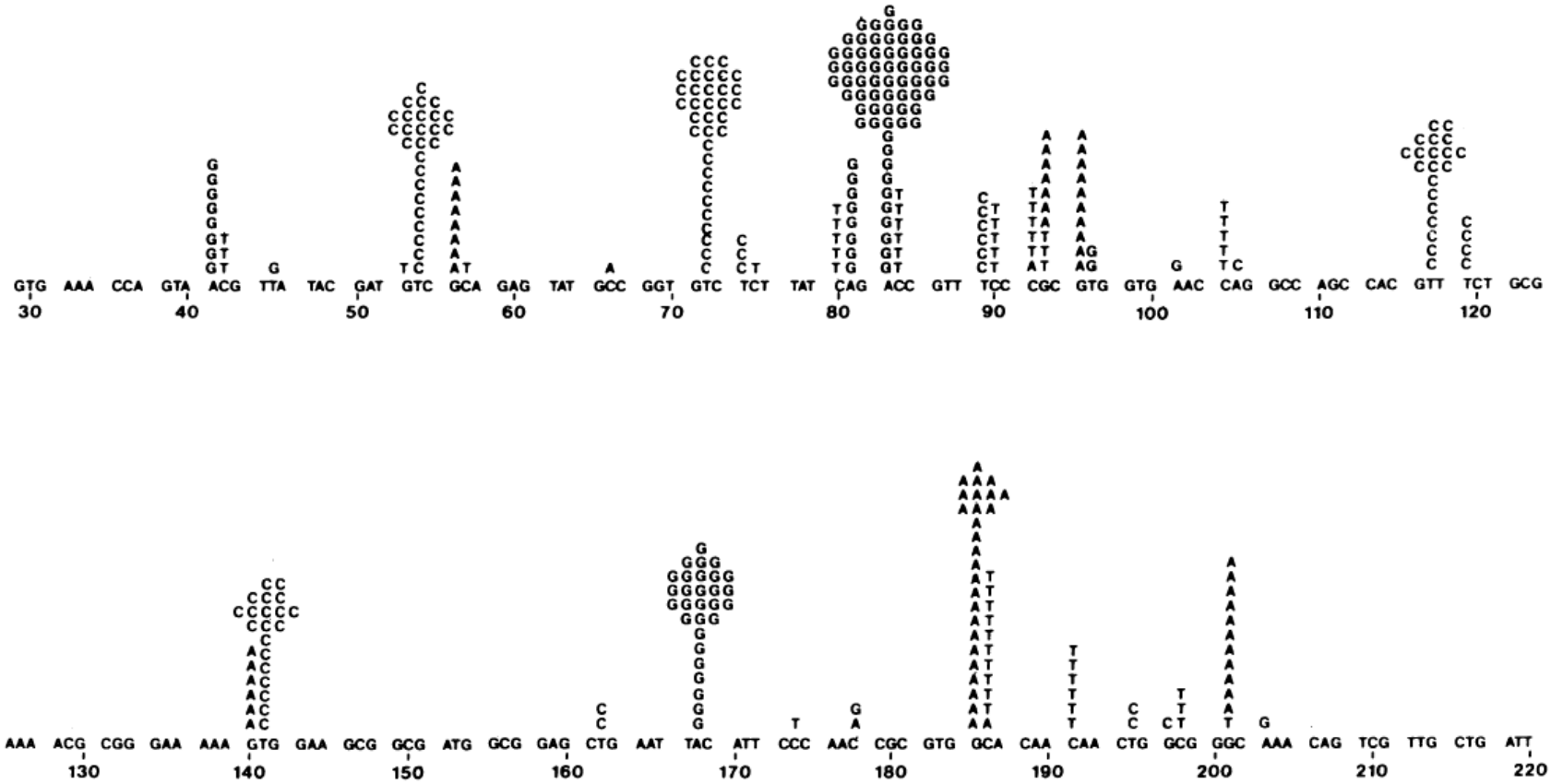


FIG. 1. Spectrum of 365 dominant spontaneous base substitutions produced by *E. coli mutH*, *mutL*, and *mutS* strains in the N-terminal part of the *lacI* gene. The figure contains the combined data from Table 3.

From: Schaaper & Dunn 1987 PNAS 84:6220

Variation in rate among sites

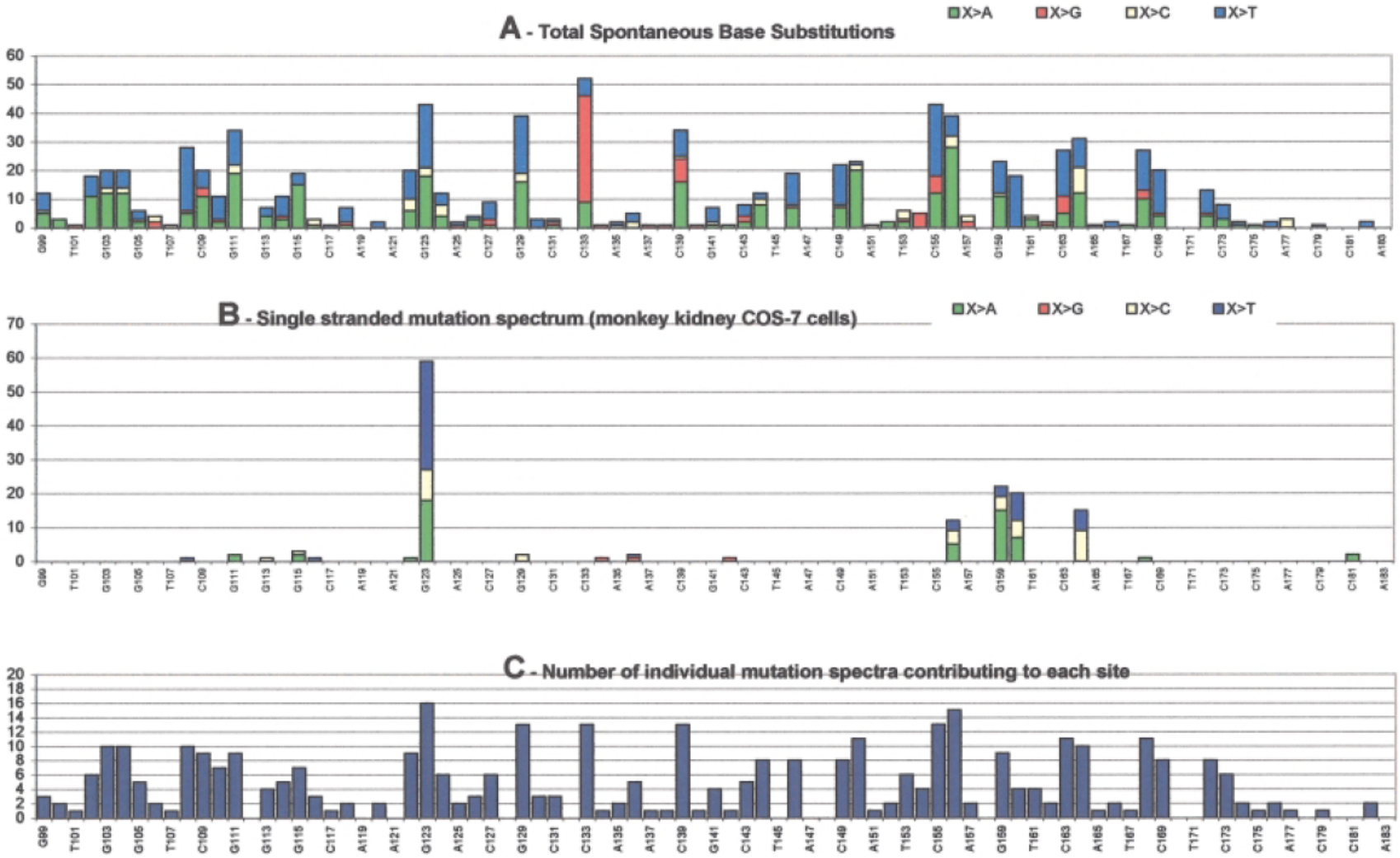


Fig. 1. *supF* mutation spectra where the tRNA nucleotide sequence for the NT strand is represented on the *x*-axis (nt 99–183) and frequency of data (actual numbers) is shown on the *y*-axis. Each base substitution is colour coded: a nucleotide change to A is green, to G is red and to Y is blue. X can be A, C, G or T depending on the nucleotide position. (**A**) Pooled *supF* spontaneous mutation spectra. Each nucleotide position shows the number of base substitutions observed over all mutant data. The types of substitution are colour coded, for example, at C133 52 substitutions were observed: nine were C→A (green), 38 were C→G (red) and seven were C→T (blue). (**B**) Mutation spectrum for spontaneous substitutions after single-stranded *supF* was transfected into monkey kidney COS-7 cells. (**C**) Spectrum showing the number of individual mutation spectra contributing mutant data to each nucleotide site, for example, mutations at G123 were derived from 16 references, whereas no substitutions were recorded from any reference at A147.

From: Lewis et al. 2001 Mutagenesis 16:503

Variation in rate among sites

The previous distances all assume that the rate of nucleotide substitution is the same for all sites.

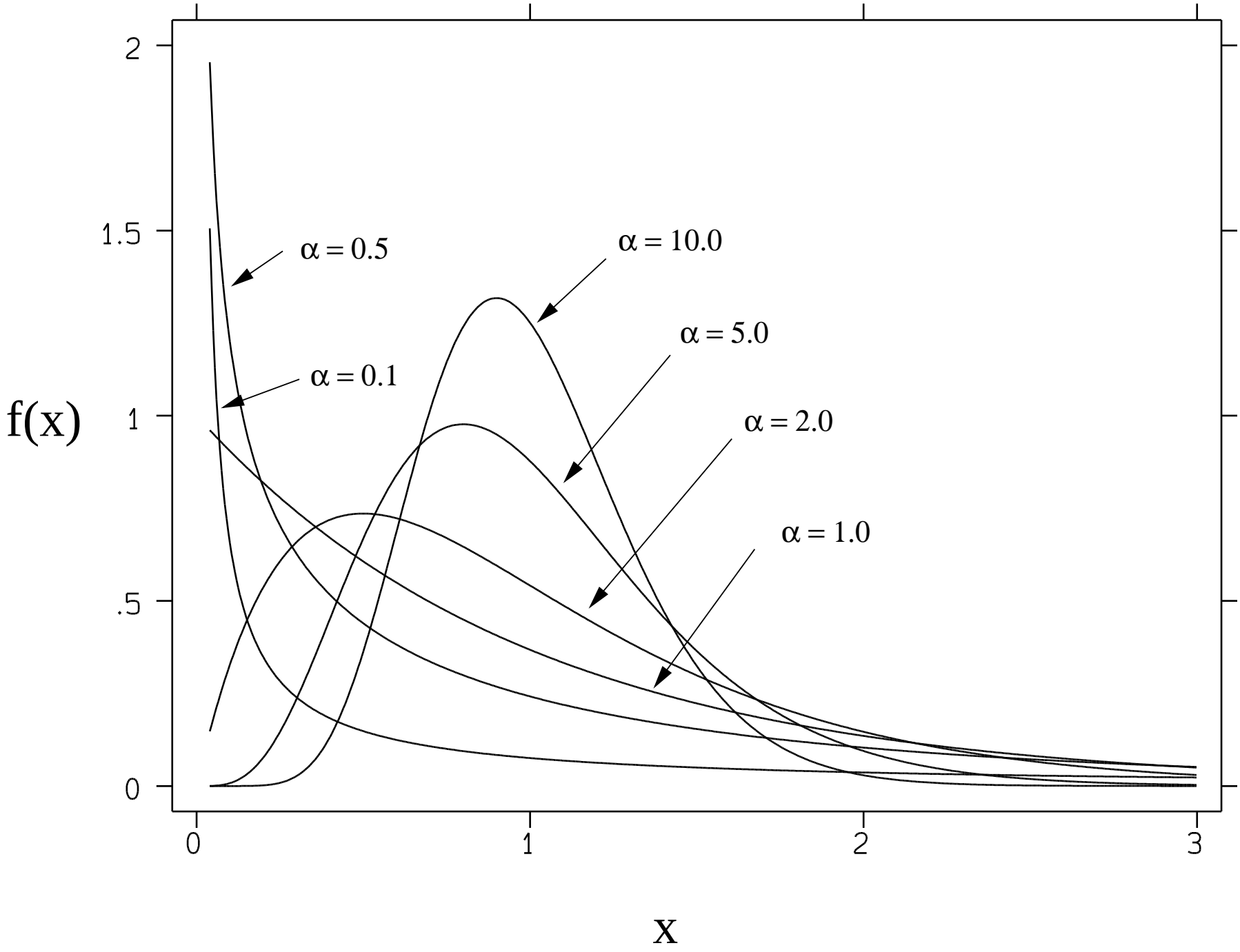
What if it is different? Some sites changing rapidly and others changing slowly.

This difference could be because of

- functional constraints on the protein or on the RNA.
- different positions within the codon
- even differences in chromosomal location.

This is dealt with by fitting a variably shaped distribution to the observed changes and then recalculating the distances.

Gamma (Γ) distribution



Gamma (Γ) distribution

$$f(x) = \left\{ \frac{\beta^{\alpha}}{\Gamma(\alpha)} \right\} x^{(\alpha-1)} e^{-\beta x}$$

with
$x > 0$

$$\Gamma(\alpha) = \int_0^{\infty} x^{(\alpha-1)} e^{-x} dx$$

For biological purposes, in general,

$\alpha = \mu^2 / Var(\mu)$
$\beta = \alpha$

where μ is the observed mean.

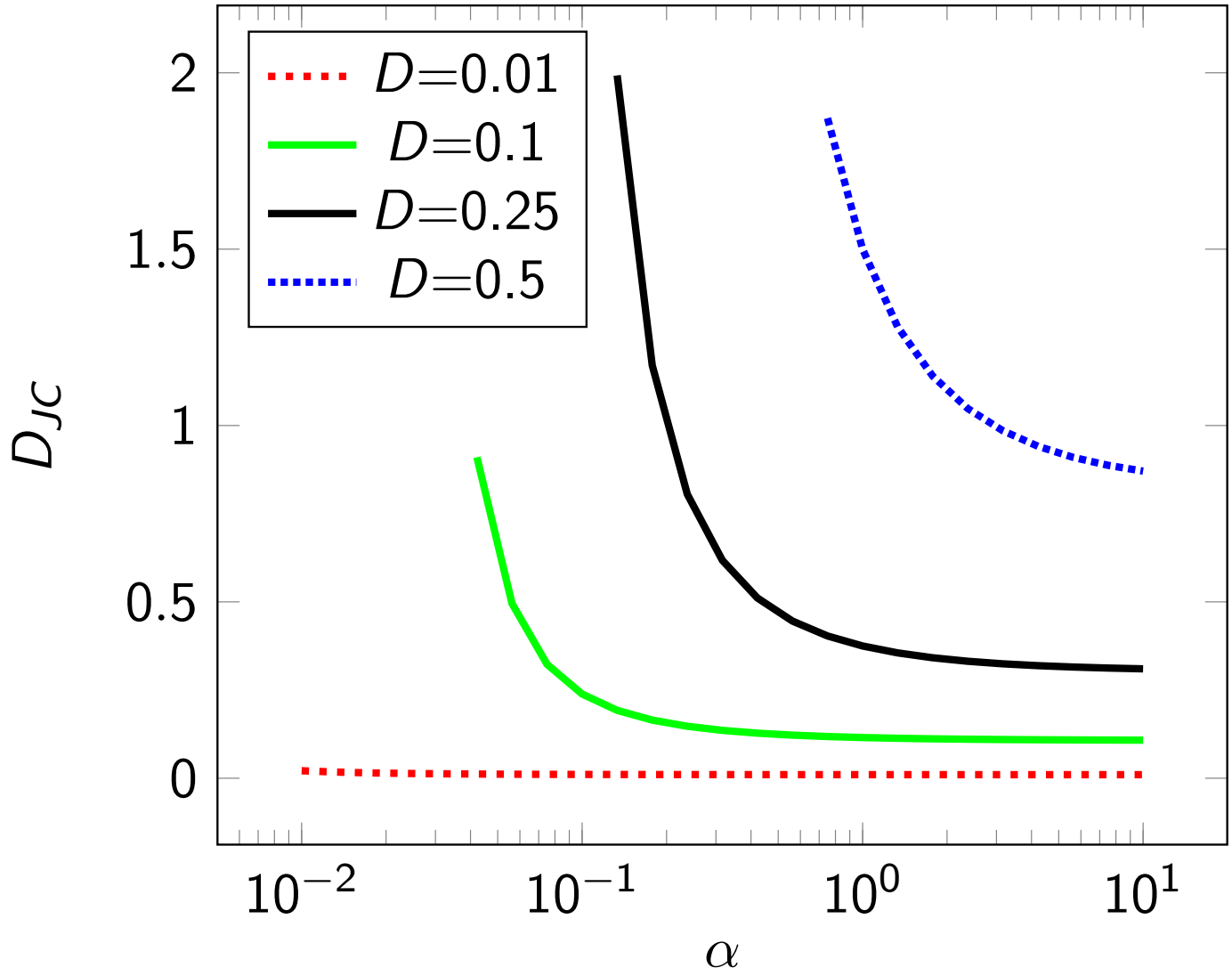
Gamma (Γ) correction

Jukes Cantor distance

$$D_{JC} = \frac{3}{4}\alpha\left\{(1-\frac{4}{3}D)^{-1/\alpha}-1\right\}$$

$$Var(D_{JC}) = D(1-D)\left\{(1-\frac{4}{3}D)^{-2(1/\alpha+1)}\right\}/n$$

Corrected Jukes-Cantor with Gamma distributed rates



Plotted for four values of Hamming distance, D.

Gamma (Γ) correction

Estimates of the Gamma Shape Parameter (α) from 13 Proteins Encoded by Mammalian Mitochondrial Genomes

GENES	Sullivan 95	YK 96	Likelihood
Atp6	2.02	1.06	0.55
Atp8	51.02	43.60	0.92
Co1	0.65	0.39	0.27
Co2	1.76	0.99	0.49
Co3	0.51	0.32	0.19
Cytb	0.94	0.55	0.36
Ndl	2.77	1.27	0.57
Nd2	20.75	3.35	0.90
Nd3	1.86	1.03	0.34
Nd4	2.06	1.11	0.57
Nd41	5.31	1.77	0.79
Nd5	2.37	1.18	0.61
Nd6	12.76	2.86	0.90

From: Gu & Zhang 1997 MBE 14:1106

The Codes used!

As you read papers in the scientific literature about genetic distances you will read code that states they used the

$$\text{“HKY} + \Gamma + \text{I”}$$

model. As explained in this section there are several models that can be used to estimate distances. These range from simple to very complex. Since in general, it is not good to over-parameterize a model, a simpler model is preferred if it adequately fits the data. To test this fit a series of hierarchical tests can be applied and have been implemented by Posada & Crandall 1998.

The Codes used!

The models that they test include

JC	Jukes and Cantor (1969)
K80	Kimura (1980) (=K2P)
HKY	Hasegawa, Kishino, Yano (1985)
TN	Tamura and Nei (1993)
TNef	Tamura-Nei equal frequencies
K81	Two transversion-parameters model 1 (=K81=K3P) (Kimura, 1981)
K81uf	K81 with unequal frequencies
TIMef	Transitional model equal frequencies
TIM	Transitional model
TVMef	Transversional model equal frequencies
TVM	Transversional model
SYM	Symmetrical model (Zharkikh, 1994)
GTR	General time reversible (=REV) (Tavare, 1986)

In addition to these, the rates can be gamma distributed, Γ, and the model might include some sites that are considered invariant.

Hence you arrive at codes such as "HKY + Γ + I; an HKY model with gamma distributed rates and invariant sites".

Gaps

In general (with exceptions) gaps are ignored.

```
1  T  A  -  C  -  A  G  T  T  A  C  C  T  C  -  A  T  T  C  C
2  T  T  T  C  C  A  -  T  T  A  -  C  T  C  A  A  T  G  -  C
3  T  A  T  C  C  -  G  T  T  T  C  -  T  C  T  A  T  T  C  C
```

```
1  T  A     C        T  T  A     T  C     A  T  T     C
2  T  T     C        T  T  A     T  C     A  T  G     C
3  T  A     C        T  T  T     T  C     A  T  T     C
```

Gaps

Distances are often done pairwise, in which case

```
1  T  A  -  C  -  A  G  T  T  A  C  C  T  C  -  A  T  T  C  C
2  T  T  T  C  C  A  -  T  T  A  -  C  T  C  A  A  T  G  -  C
3  T  A  T  C  C  -  G  T  T  T  C  -  T  C  T  A  T  T  C  C
```

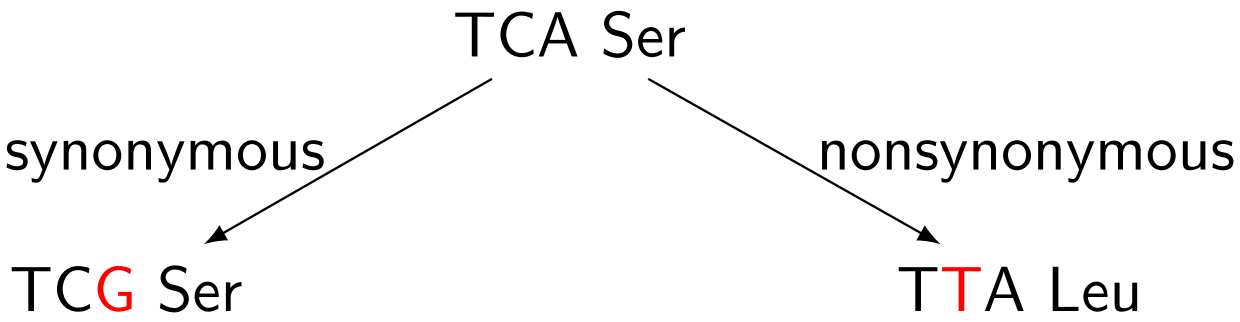
```
1  T  A     C     A     T  T  A     C  T  C     A  T  T     C
2  T  T     C     A     T  T  A     C  T  C     A  T  G     C

1  T  A     C        G  T  T  A  C     T  C     A  T  T  C  C
3  T  A     C        G  T  T  T  C     T  C     A  T  T  C  C

2  T  T  T  C  C        T  T  A        T  C  A  A  T  G     C
3  T  A  T  C  C        T  T  T        T  C  T  A  T  T     C
```

Synonymous and non-synonymous substitutions

Any substitution in a coding sequence leading to (1) the conservation of the same amino-acid is a synonymous substitution, (2) a amino acid change is a non-synonymous substitution.



Li, Wu and Luo's method relies on the classification of every site from two sequences.

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

Four fold degenerate sites:

These are third codon positions where all changes are synonymous regardless of the nucleotide present.
e.g. GUN codes for Val regardless of N.
32 of the third positions from 61 sense codons is a four fold degenerate site.

Four fold degenerate sites:

These are third codon positions where all changes are synonymous regardless of the nucleotide present.
e.g. GUN codes for Val regardless of N.
32 of the third positions from 61 sense codons is a four fold degenerate site.

Two fold degenerate sites:

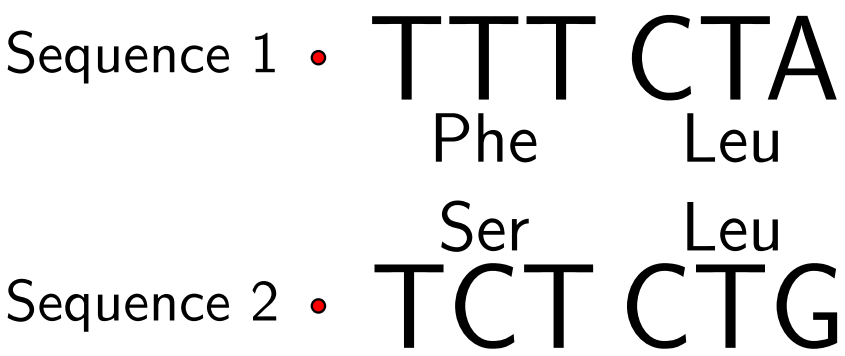
These are third codon positions where one possible change is synonymous.
e.g. CAY codes for His regardless of Y.
24 of the third positions from 61 sense codons is a two fold degenerate site.

Non-degenerate sites:

All possible changes at the third codon position are non-synonymous.
e.g. AUG for Met and UGG for Trp

Note: All second codon positions are non-synonymous
Most first codon positions are non-synonymous.

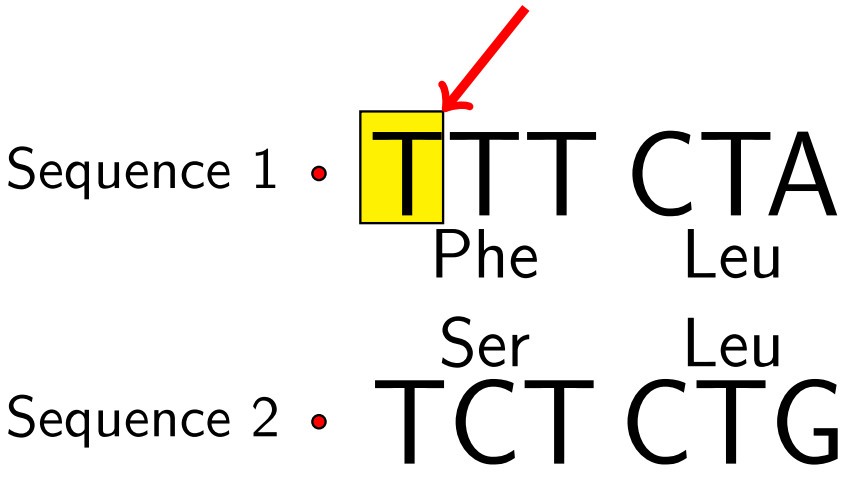
Li, Wu, Lou



Li, Wu, Lou

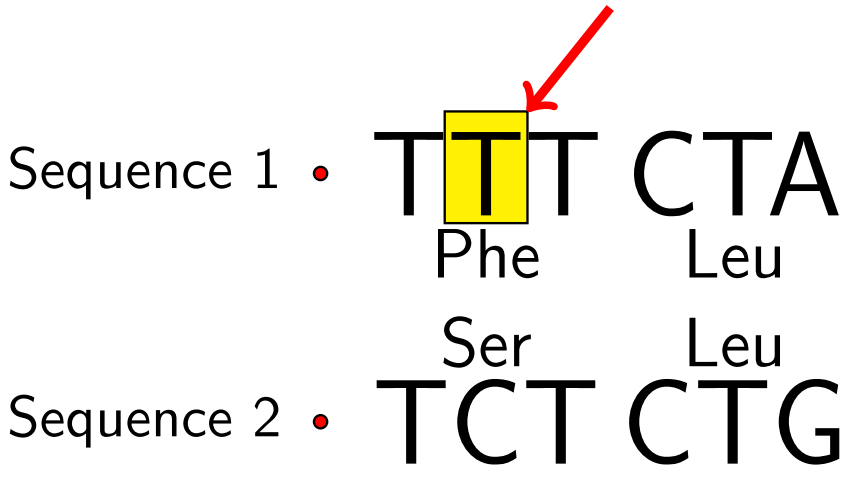
$n_n = 1$

$n_s = 0$



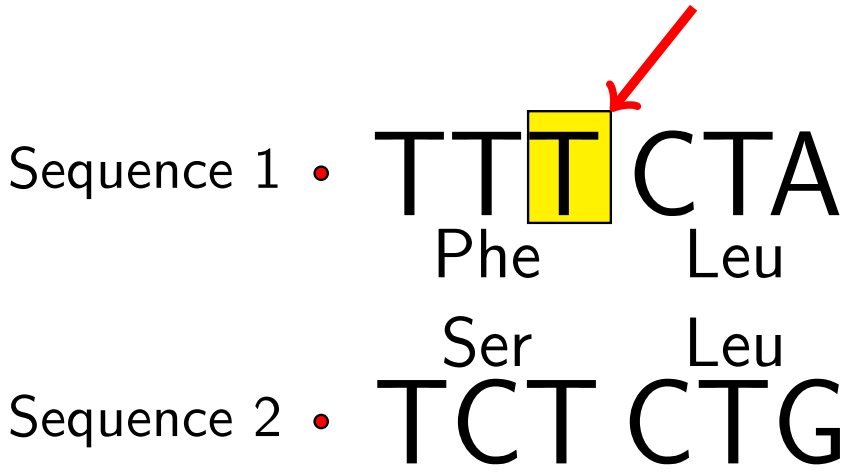
Li, Wu, Lou

$$n_n = 1 + 1$$
$$n_s = 0 + 0$$



Li, Wu, Lou

$$n_n = 1 + 1 + \frac{2}{3}$$
$$n_s = 0 + 0 + \frac{1}{3}$$



Li, Wu, Lou

Sequence 1 • TTT CTA
Phe Leu

$$\begin{cases} n_n = 4\frac{1}{3} \\ n_s = 1\frac{2}{3} \end{cases}$$

Ser Leu
Sequence 2 • TCT CTG

$$\begin{cases} n_n = 3\frac{2}{3} \\ n_s = 2\frac{1}{3} \end{cases}$$

Li, Wu, Lou

Sequence 1 • TTT CTA (Phe Leu)

$$\begin{cases} n_n = 4\frac{1}{3} \\ n_s = 1\frac{2}{3} \end{cases}$$

Sequence 2 • TCT CTG (Ser Leu)

$$\begin{cases} n_n = 3\frac{2}{3} \\ n_s = 2\frac{1}{3} \end{cases}$$

$$obs_n = 1 \quad n_n = (4\frac{1}{3} + 3\frac{2}{3})/2$$
$$obs_s = 1 \quad n_s = (1\frac{2}{3} + 2\frac{1}{3})/2$$

Li, Wu, Lou

Sequence 1 • TTT CTA (Phe Leu)

$$\begin{cases} n_n = 4\frac{1}{3} \\ n_s = 1\frac{2}{3} \end{cases}$$

Sequence 2 • TCT CTG (Ser Leu)

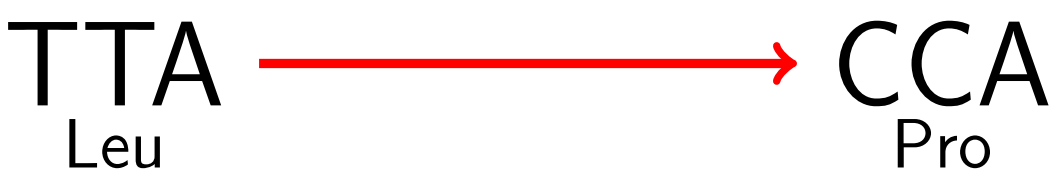
$$\begin{cases} n_n = 3\frac{2}{3} \\ n_s = 2\frac{1}{3} \end{cases}$$

$$K_a = obs_n/n_n = 1/4$$

$$K_s = obs_s/n_s = 1/2$$

Li, Wu, Lou – another complication

What about?



Li, Wu, Lou – another complication

What about?