Wellcome trust advanced course
“Molecular Evolution”

Markovian Models for Molecular Phylogenetics

Manolo Gouy
CNRS - Université de Lyon

Hinxton 29 March - 9 April 2009

Markovian models of DNA sequence evolution

The evolution of a sequence site is modelled as follows:
there are substitution rates *i->j*, per time unit, that apply at any time during the evolutionary process.
Matrix M of <u>instantaneous substitution rates</u>:

$$M = \begin{array}{|c|c|c|c|c|}
\hline
\swarrow & \mathrm{A} & \mathrm{T} & \mathrm{C} & \mathrm{G} \\
\hline
\mathrm{A} & -\lambda_A & m_{TA} & m_{CA} & m_{GA} \\
\hline
\mathrm{T} & m_{AT} & -\lambda_T & m_{CT} & m_{GT} \\
\hline
\mathrm{C} & m_{AC} & m_{TC} & -\lambda_C & m_{GC} \\
\hline
\mathrm{G} & m_{AG} & m_{TG} & m_{CG} & -\lambda_G \\
\hline
\end{array}$$

m_{ij} = rate of i→j substitution per time unit.

λ_i are such that column sums = *0*
(λ_i = total mutation rate of *i*)

Here, M follows the convention $m_{\text{colum, row}}$. The other convention $m_{\text{row,column}}$ is often used in the literature.

This most general model contains <u>12 free parameters</u>.

Markovian models of DNA sequence evolution *(continued)*

Any matrix M of instantaneous substitution rates possesses two major properties:

1) If *F(t)* is the vector of base frequencies at time *t*

$$\frac{dF(t)}{dt} = MF(t)$$

2) If *P(t)* is the matrix of conditional substitution probabilities after *t* time units of evolution, $P(t) = e^{Mt}$.

ancestor: $i \xrightarrow{t \text{ time units}} j$: descendant

$P_{ij}(t)$ = proba *j* at *t* when *i* at *0*

Equilibrium frequencies of a Markovian model

Any realistic Markovian model possesses its own set of equilibrium frequencies F_{eq}:

$$\text{such that} \quad \frac{dF_{eq}(t)}{dt} = 0 \quad \text{or} \quad MF_{eq} = 0$$

[F_{eq} is the eigenvector associated to the eigenvalue 0 of M]

If any sequence evolves with constant substitution rates, it will reach a fixed composition, its equilibrium composition

$$\mathrm{F}_{\mathrm{eq}} = (\pi_A, \pi_T, \pi_C, \pi_G)$$

that will then remain unchanged.

Reversibility of Markovian evolutionary models

Mathematical definition :

$$\forall i,j \;\; \pi_j m_{ji} = \pi_i m_{ij} \quad <=> \quad \forall i,j,t \;\; \pi_j P_{ji}(t) = \pi_i P_{ij}(t)$$

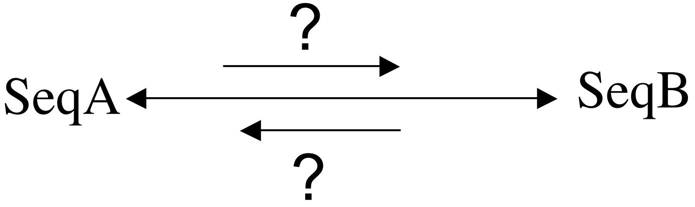
(at base equilibrium frequencies)

Conceptual definition :

for any pair of nucleotides (i,j), $i \rightarrow j$ flux = $j \rightarrow i$ flux

Consequences :

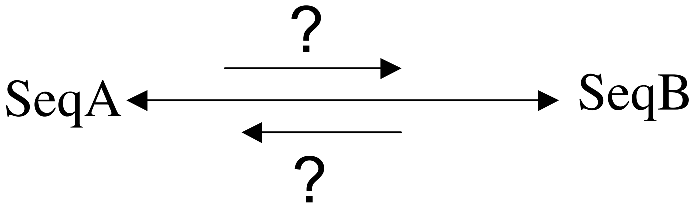
- The observation of two sequences at the extremities of a branch (during which evolution followed a constant Markovian model) contains no information about the direction of evolution.



- Computations can be done as though the tree is unrooted.

But there is <u>no biological justification</u> for believing that the molecular evolutionary process is reversible.

Reversibility of Markovian evolutionary models *(continued)*



Only constraint: reversibility

General Time Reversible (9 parameters)

$$M = \begin{array}{|c|c|c|c|c|}
\hline
\swarrow & A & T & C & G \\
\hline
A & -\lambda_A & a\pi_A & b\pi_A & c\pi_A \\
\hline
T & a\pi_T & -\lambda_T & d\pi_T & e\pi_T \\
\hline
C & b\pi_C & d\pi_C & -\lambda_C & f\pi_C \\
\hline
G & c\pi_G & e\pi_G & f\pi_G & -\lambda_G \\
\hline
\end{array}$$

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$

10 symbols but 9 parameters because $\pi_A + \pi_T + \pi_C + \pi_G = 1$

Tamura & Nei 93 (6 parameters)

$$M = \begin{array}{|c|c|c|c|c|} \hline \swarrow & A & T & C & G \\ \hline A & -\lambda_A & \beta\pi_A & \beta\pi_A & \alpha_R \frac{\pi_A}{\pi_R} + \beta\pi_A \\ \hline T & \beta\pi_T & -\lambda_T & \alpha_Y \frac{\pi_T}{\pi_Y} + \beta\pi_T & \beta\pi_T \\ \hline C & \beta\pi_C & \alpha_Y \frac{\pi_C}{\pi_Y} + \beta\pi_C & -\lambda_C & \beta\pi_C \\ \hline G & \alpha_R \frac{\pi_G}{\pi_R} + \beta\pi_G & \beta\pi_G & \beta\pi_G & -\lambda_G \\ \hline \end{array}$$

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$

This is the most parameter-rich reversible model for which one can compute analytically the matrix $P(t) = e^{Mt}$ of conditional substitutions.

[Tamura & Nei (1993) *MolBiolEvol* 10:512]

Jukes & Cantor (1 parameter)

$$M = \begin{array}{|c|c|c|c|c|}
\hline
\swarrow & \mathrm{A} & \mathrm{T} & \mathrm{C} & \mathrm{G} \\
\hline
\mathrm{A} & -\lambda_A & r & r & r \\
\hline
\mathrm{T} & r & -\lambda_T & r & r \\
\hline
\mathrm{C} & r & r & -\lambda_C & r \\
\hline
\mathrm{G} & r & r & r & -\lambda_G \\
\hline
\end{array}$$

Eq. (1/4, 1/4, 1/4, 1/4)

The Jukes & Cantor model has been historically the first one to be introduced.
Justification: simplicity.

Kimura (2 parameters)

$$M = \begin{array}{|c|c|c|c|c|}
\hline
\swarrow & \mathrm{A} & \mathrm{T} & \mathrm{C} & \mathrm{G} \\
\hline
\mathrm{A} & -\lambda_A & r & r & \alpha\, r \\
\hline
\mathrm{T} & r & -\lambda_T & \alpha\, r & r \\
\hline
\mathrm{C} & r & \alpha\, r & -\lambda_C & r \\
\hline
\mathrm{G} & \alpha\, r & r & r & -\lambda_G \\
\hline
\end{array}$$

Eq. (1/4, 1/4, 1/4, 1/4)

Kimura’s 2-parameter model aims at reflecting the fact that transitions are more frequent than transitions.

Felsenstein 81 (4 parameters)

$$M = \begin{array}{|c|c|c|c|c|}
\hline
\swarrow & A & T & C & G \\
\hline
A & -\lambda_A & r\pi_A & r\pi_A & r\pi_A \\
\hline
T & r\pi_T & -\lambda_T & r\pi_T & r\pi_T \\
\hline
C & r\pi_C & r\pi_C & -\lambda_C & r\pi_C \\
\hline
G & r\pi_G & r\pi_G & r\pi_G & -\lambda_G \\
\hline
\end{array}$$

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$

Felsenstein's 1981 model allows for any arbitrary set of equilibrium frequencies.

Felsenstein 84 (5 parameters)

$$M = \begin{array}{|c|c|c|c|c|}
\hline
\swarrow & A & T & C & G \\
\hline
A & -\lambda_A & \beta\pi_A & \beta\pi_A & \alpha\frac{\pi_A}{\pi_R} + \beta\pi_A \\
\hline
T & \beta\pi_T & -\lambda_T & \alpha\frac{\pi_T}{\pi_Y} + \beta\pi_T & \beta\pi_T \\
\hline
C & \beta\pi_C & \alpha\frac{\pi_C}{\pi_Y} + \beta\pi_C & -\lambda_C & \beta\pi_C \\
\hline
G & \alpha\frac{\pi_G}{\pi_R} + \beta\pi_G & \beta\pi_G & \beta\pi_G & -\lambda_G \\
\hline
\end{array}$$

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$

$\pi_R = \pi_{A+}\pi_G \quad \pi_Y = \pi_{C+}\pi_T$

Felsenstein's 1984 model was a pioneering attempt to incorporate both transition/transversion bias and an arbitrary set of equilibrium frequencies.

HKY-Hasegawa,Kishino,Yano: 5 params

$$M = \begin{array}{|c|c|c|c|c|}\hline \swarrow & A & T & C & G \\ \hline A & -\lambda_A & \pi_A b & \pi_A b & \pi_A a \\ \hline T & \pi_T b & -\lambda_T & \pi_T a & \pi_T b \\ \hline C & \pi_C b & \pi_C a & -\lambda_C & \pi_C b \\ \hline G & \pi_G a & \pi_G b & \pi_G b & -\lambda_G \\ \hline \end{array}$$

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$

The HKY model is another way to incorporate both transition/transversion bias and an arbitrary set of equilibrium frequencies.
HKY and F84 are very similar models.

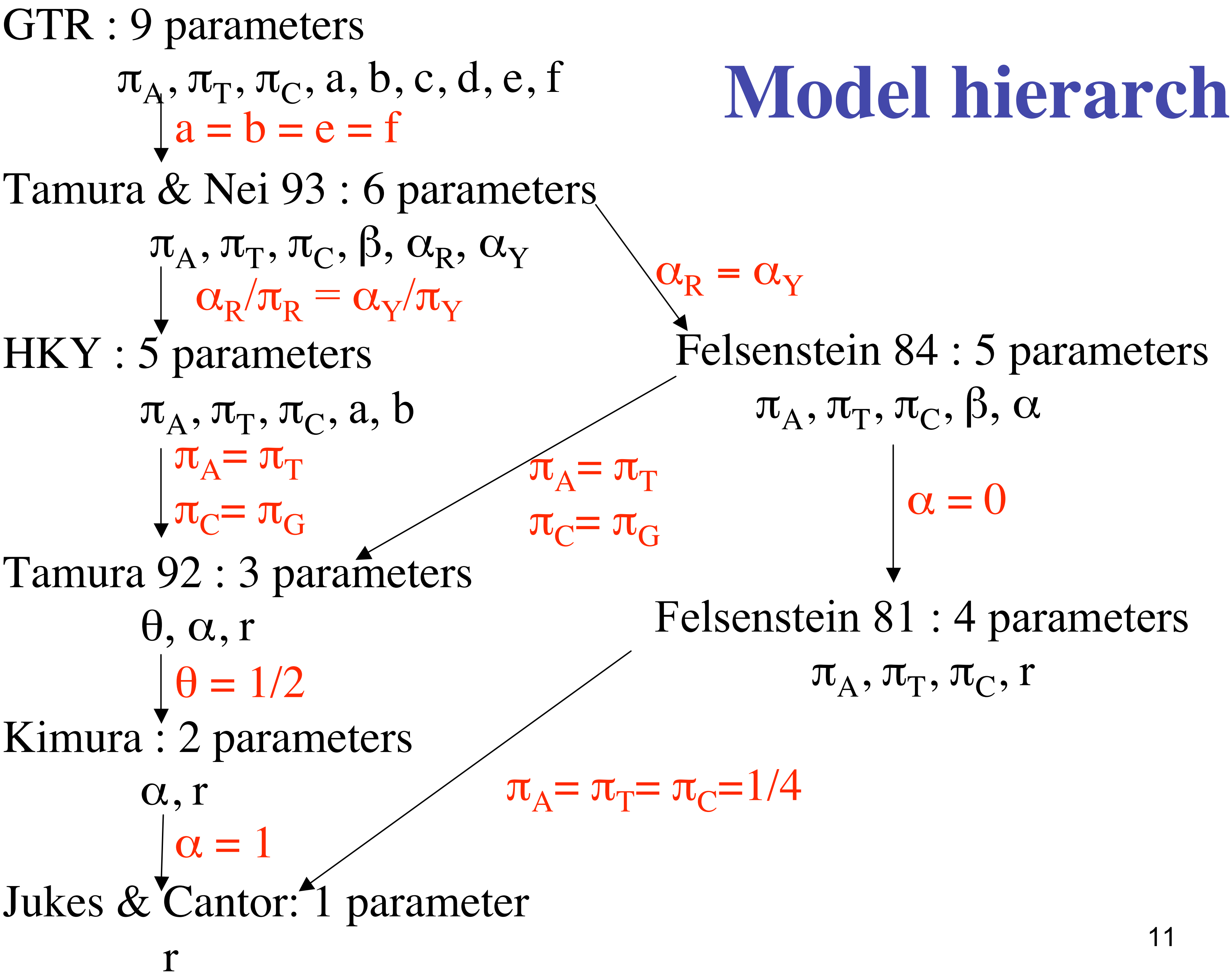
Tamura 92 (3 parameters)

$$M = \begin{array}{|c|c|c|c|c|}\hline \swarrow & A & T & C & G \\ \hline A & -\lambda_A & \frac{1-\theta}{2}r & \frac{1-\theta}{2}r & \alpha\frac{1-\theta}{2}r \\ \hline T & \frac{1-\theta}{2}r & -\lambda_T & \alpha\frac{1-\theta}{2}r & \frac{1-\theta}{2}r \\ \hline C & \frac{\theta}{2}r & \alpha\frac{\theta}{2}r & -\lambda_C & \frac{\theta}{2}r \\ \hline G & \alpha\frac{\theta}{2}r & \frac{\theta}{2}r & \frac{\theta}{2}r & -\lambda_G \\ \hline \end{array}$$

Eq.$((1-\theta)/2, (1-\theta)/2, \theta/2, \theta/2)$

This model aims at representing two phenomena :
- sequence G+C content
- transition/transversion bias

Model hierarchy



GTR : 9 parameters

$\pi_A, \pi_T, \pi_C, a, b, c, d, e, f$

$a = b = e = f$

Tamura & Nei 93 : 6 parameters

$\pi_A, \pi_T, \pi_C, \beta, \alpha_R, \alpha_Y$

$\alpha_R/\pi_R = \alpha_Y/\pi_Y$

$\alpha_R = \alpha_Y$

HKY : 5 parameters

$\pi_A, \pi_T, \pi_C, a, b$

Felsenstein 84 : 5 parameters

$\pi_A, \pi_T, \pi_C, \beta, \alpha$

$\pi_A = \pi_T$

$\pi_C = \pi_G$

$\pi_A = \pi_T$

$\pi_C = \pi_G$

$\alpha = 0$

Tamura 92 : 3 parameters

θ, α, r

Felsenstein 81 : 4 parameters

π_A, π_T, π_C, r

$\theta = 1/2$

Kimura : 2 parameters

α, r

$\pi_A = \pi_T = \pi_C = 1/4$

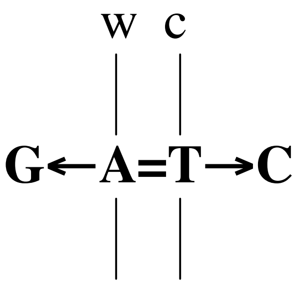
$\alpha = 1$

Jukes & Cantor: 1 parameter

r

A biologically-motivated non-reversible Markovian model

Assumption: both DNA strands are replicated under the same conditions.



$m_{AG} = w_{AG} + c_{TC}$, more generally, $m_{ij} = w_{ij} + c_{\bar{i}\bar{j}}$

Symmetry of both stands *vis a vis* replication : $w_{ij} = c_{ij}$

$\forall i, j \;\; m_{ij} = w_{ij} + c_{\overline{ij}} = c_{ij} + w_{\overline{ij}} = m_{\overline{ij}}$

Lobry & Sueoka 95 (6 param.)

$$M = \begin{array}{|c|c|c|c|c|}\hline \swarrow & \mathrm{A} & \mathrm{T} & \mathrm{C} & \mathrm{G} \\ \hline \mathrm{A} & -\lambda_A & a & d & b \\ \hline \mathrm{T} & a & -\lambda_T & b & d \\ \hline \mathrm{C} & e & c & -\lambda_C & f \\ \hline \mathrm{G} & c & e & f & -\lambda_G \\ \hline \end{array}$$

Eq. $(u/2v, u/2v, (v-u)/2v, (v-u)/2v)$

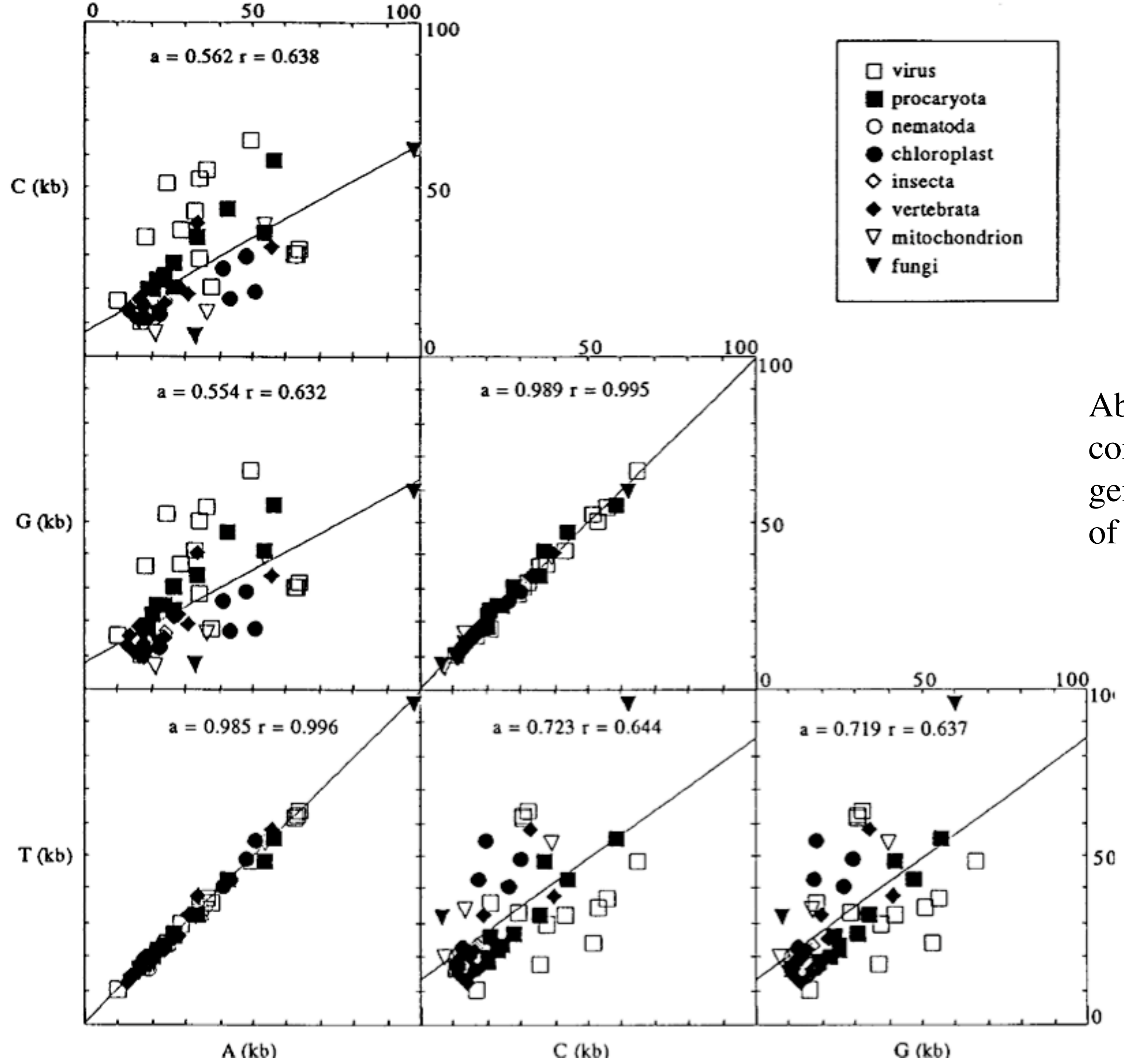
$u = b+d; \; v = b+c+d+e$

Non-reversible : $\pi_A m_{AC} \neq \pi_C m_{CA}$

at equilibrium : [A]=[T] et [C]=[G]

Sueoka (1995) JMolEvol 40:318

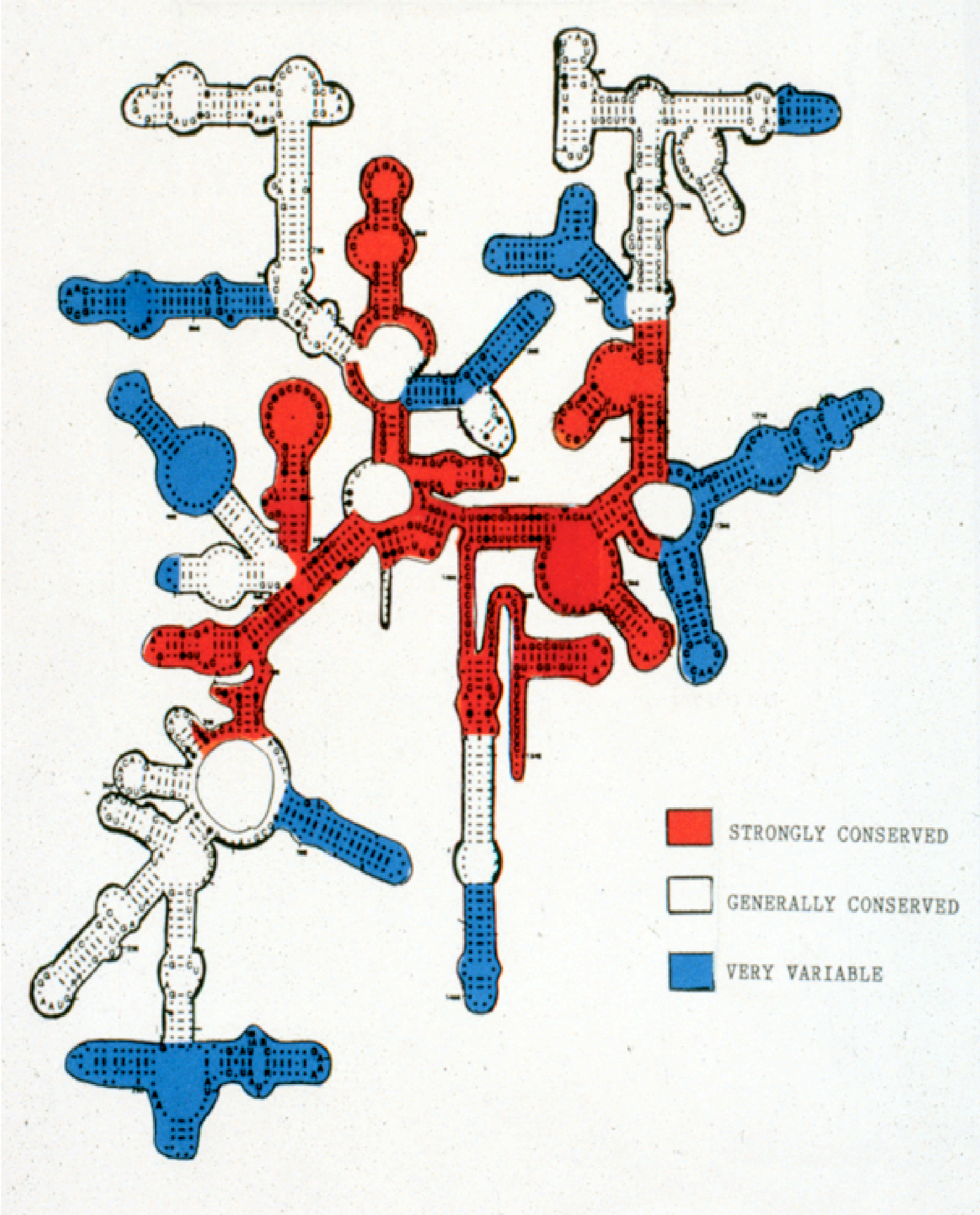
Lobry (1995) JMolEvol 40:326



Absolute base compositions of genomic fragments of length ≥ 50 kb.

Lobry 1995
JMolEvol 40:326

Across sites evolutionary rate variation



Small subunit ribosomal RNA (18S or 16S)

Modelling across sites evolutionary rate variation

Density *f(r)* of the gamma distribution :

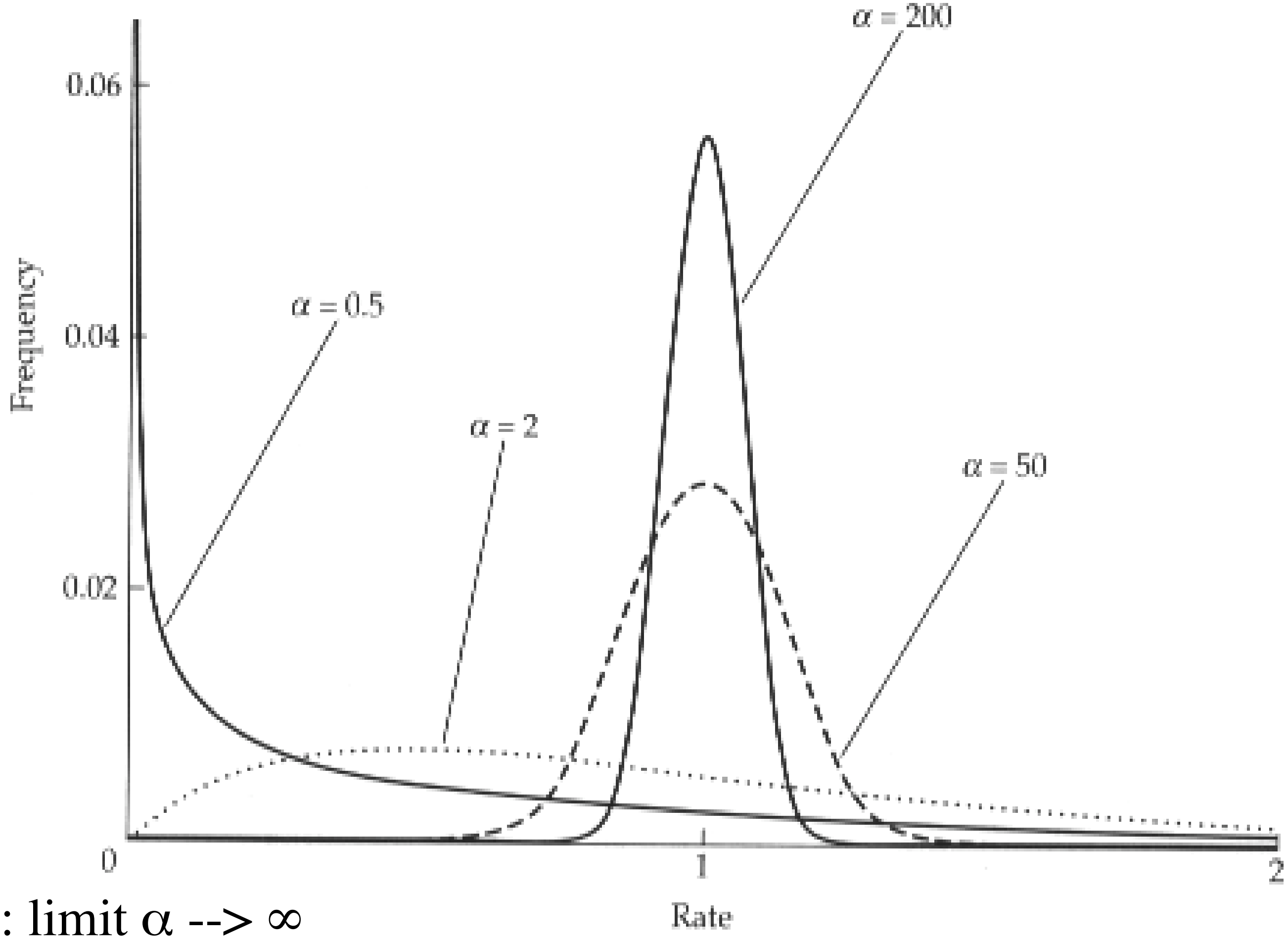
$$f(r) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} r^{\alpha-1} e^{-r/\beta}$$

α: shape parameter
β: scale parameter
mean: $\alpha\beta$
variance: $\alpha\beta^2$

Taking $\beta=1/\alpha$,
mean = 1
variance = $1/\alpha$

This allows to model the distribution of evolutionary rates around the mean rate.

The gamma distribution has no biological justification, it was chosen for its convenience.

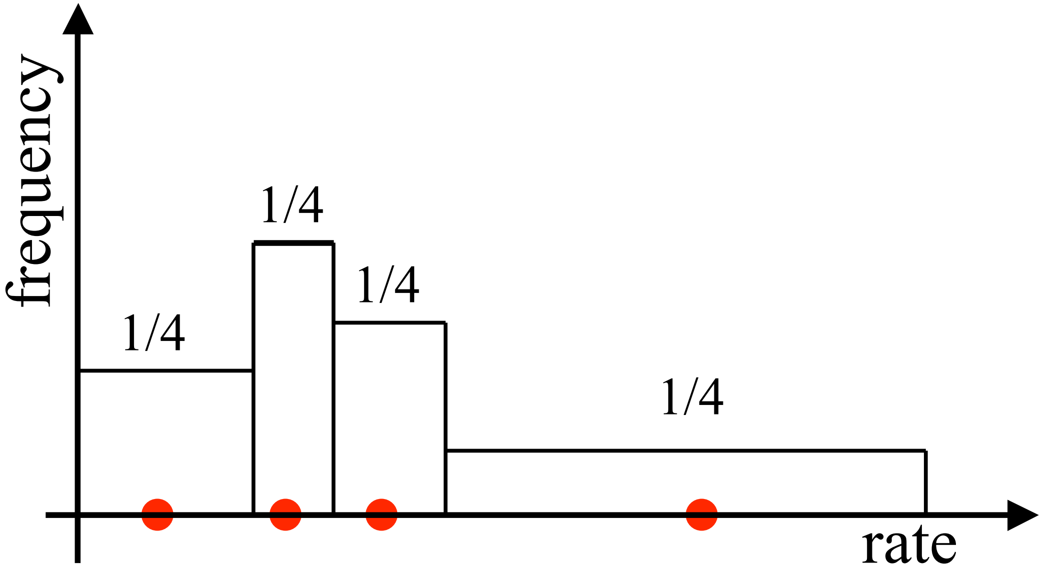


No variation across sites : limit $\alpha \to \infty$

Modelling across sites evolutionary rate variation *(continued)*

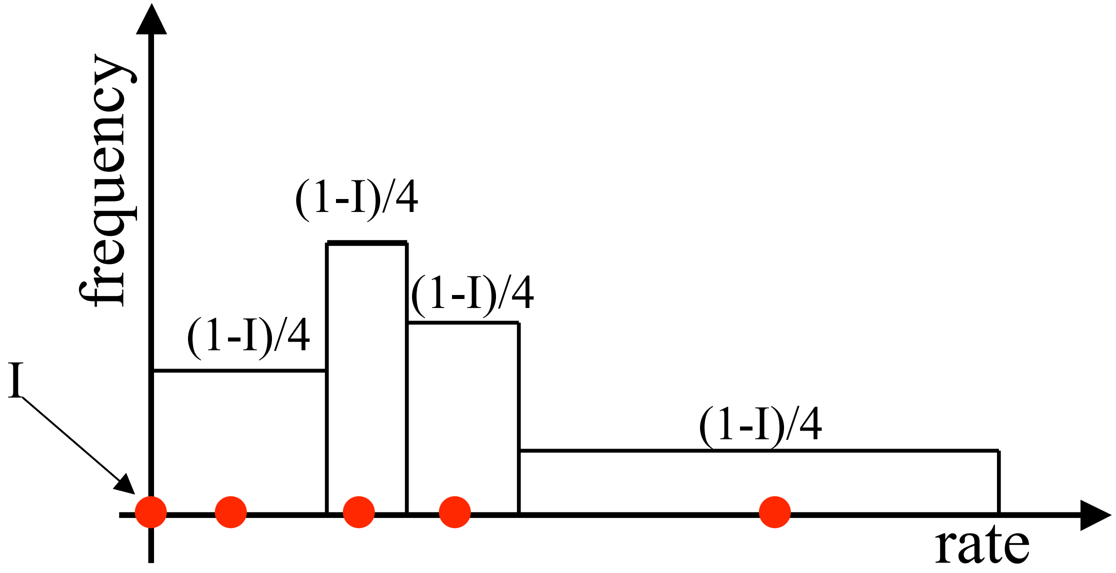
In many contexts, the gamma distribution is simplified by discretization to allow easy computations.

Example of discretization in 4 classes of equal weight:



Modelling across sites evolutionary rate variation *(continued)*

Frequently, an additional class of sites is allowed : invariable sites.
This is the G + I model.
The fraction I of invariable sites needs to be estimated from the data.



Markovian models of protein sequence evolution

The evolutionary process is modelled by a matrix Q of the rates $q_{i,j}$ of amino acid replacements per unit time :

$$Q = (q_{i,j})_{\; i=1,..,20,\;\; j=1,..,20}$$

As with nucleotide evolutionary models, there are equilibrium amino acid frequencies:

$$(\pi_i)_{\;\; i = 1,\ldots,20}$$

Reversible Markovian models of protein sequence evolution

General Time Reversible for DNA

↙	A	T	C	G
A	—	$a\pi_A$	$b\pi_A$	$c\pi_A$
T	$a\pi_T$	—	$d\pi_T$	$e\pi_T$
C	$b\pi_C$	$d\pi_C$	—	$f\pi_C$
G	$c\pi_G$	$e\pi_G$	$f\pi_G$	—

Eq. $(\pi_A, \pi_T, \pi_C, \pi_G)$

More generally, for a reversible Markovian substitution process :

$$q_{ij} = s_{ij} \,.\, \pi_j, \quad s_{ij} = s_{ji}\,, \qquad \text{for } i \neq j$$

Thus q_{ij} can be decomposed in two components :
s_{ij} represents the exchangeability of amino acids i and j
and π_i, the equilibrium frequency of amino acid i.

There are 190 free parameters in such a model.

Empirical models of protein sequence evolution

190 free parameters are too many for them to be estimated from a single protein sequence alignment.

Thus, empirically-derived values of exchangeabilities (s_{ij}) are used.

These have been computed from very large sets of homologous proteins :

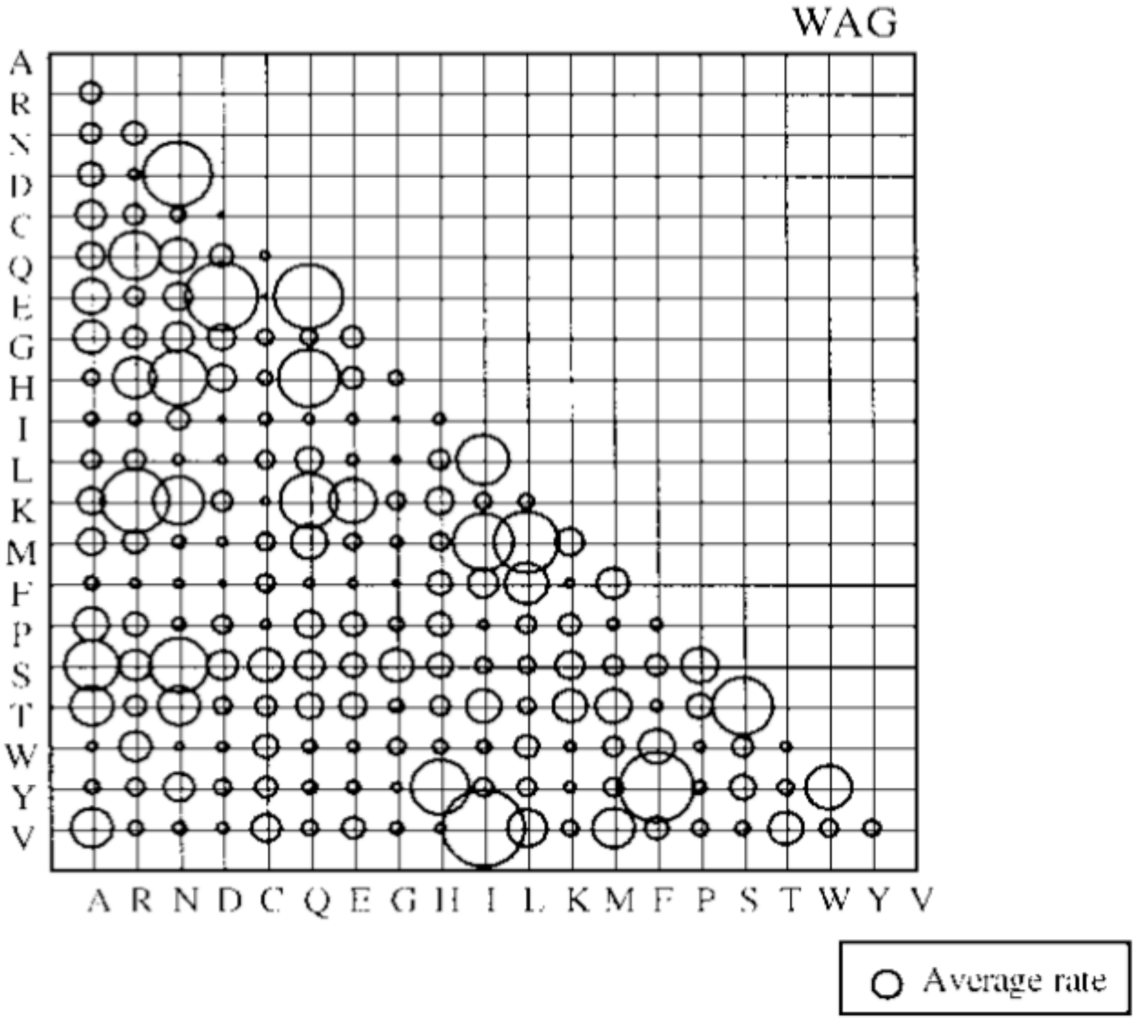
- The PAM model (Dayhoff, 1978) was built from 1,300 highly similar sequences (≥ 85 % identity) belonging to 71 families.

- The JTT model (Jones et al., 1992) was built from 16,300 sequences (≥ 85 % identity).

- The WAG model (Whelan & Goldman, 2001) was built from 3,905 proteins belonging to 182 families using a procedure that allowed for multiple replacements on a single branch at a single site.

- The LG model (Le & Gascuel, 2008) was built from 49,637 proteins of 3,912 families and improved by accounting for across-sites evolutionary rate variation.

Schematic representation of the WAG amino acid replacement matrix



The area of each bubble represents the amino acid exchangeability parameter (s_{ij}) for the replacement of amino acid *i* by amino acid *j* or vice versa.

Jargon :

Model JTT means s_{ij} are those from Jones *et al.* and π_i were as in proteins compiled by Jones *et al*.

Applying WAG + F to a protein data set means that Whelan and Goldman’s empirical exchangeability values (s_{ij}) were used and that equilibrium frequencies π_i were set to average amino acid frequencies of the data set.