

I think

Models of evolution

Màster Genètica i Genòmica
Marta Riutort

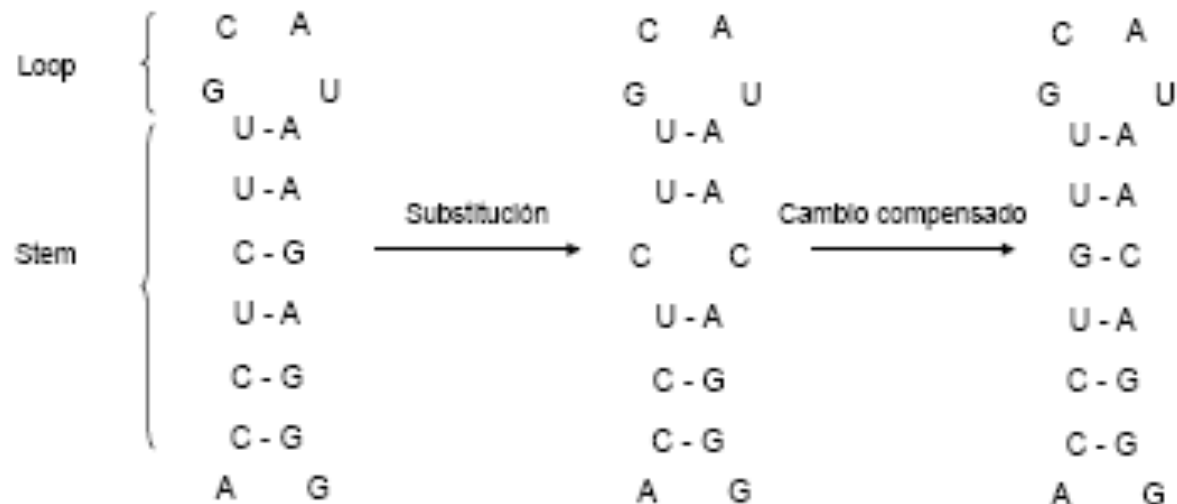


When and why we need a model of evolution

- All inference methods make a series of assumptions:
 - All sites change independently
 - Evolutionary rates are constant through time and lineages
 - Base composition is homogeneous
 - Base substitution likelihood is the same for all sites and does not change through time.

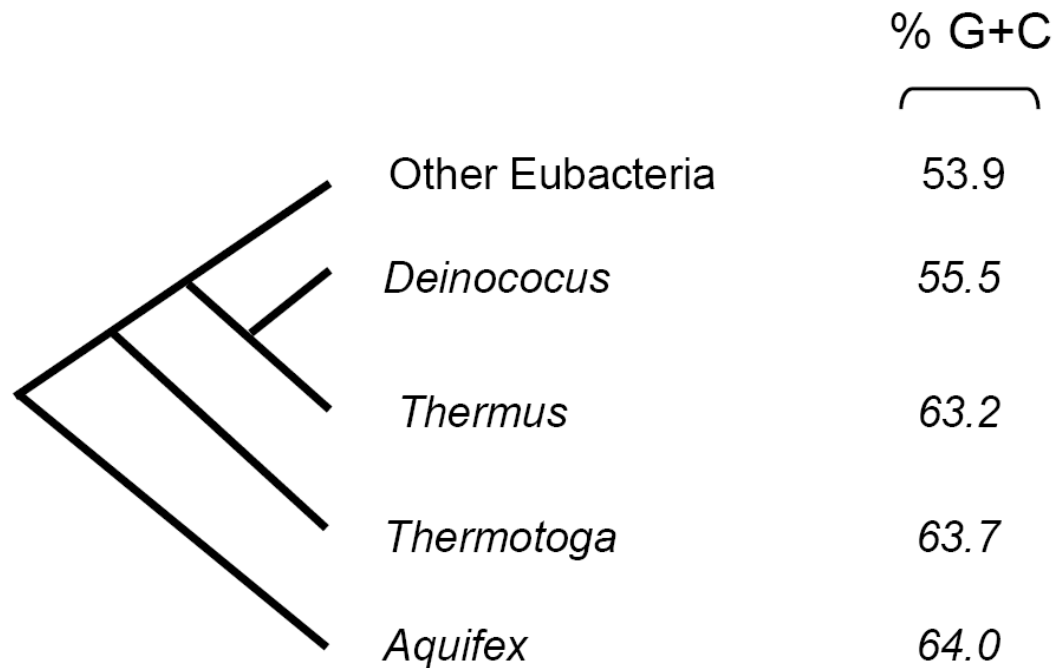
When and why we need a model of evolution

~~All sites change independently~~



When and why we need a model of evolution

~~Base composition is homogeneous~~



When and why we need a model of evolution

~~Evolutionary rates are constant through time and lineages~~

mutations per nucleotide
per 10^9 years

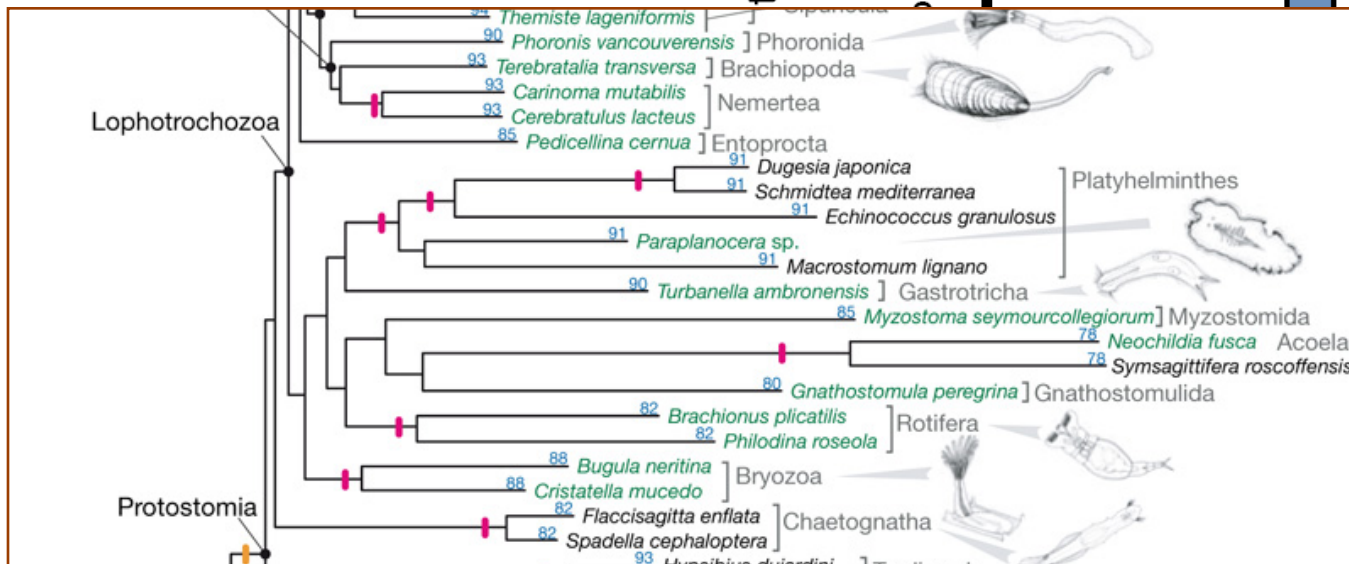
8
7
6
5
4
3
2

Introns

3' untranslated region

3' flanking region

Pseudogenes

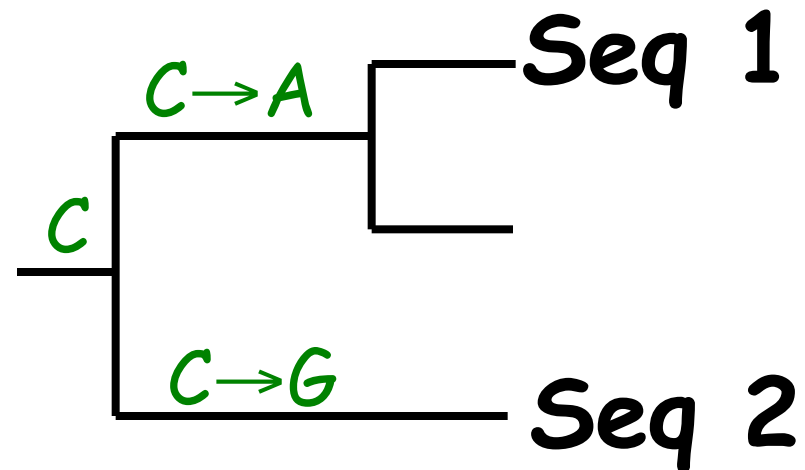
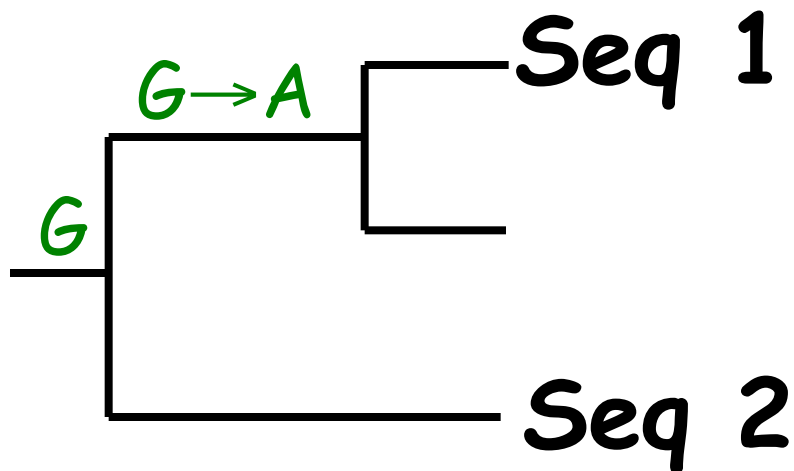


When and why we need a model of evolution

- Multiple changes at a single site - hidden changes

Seq 1 **AGCGAG**

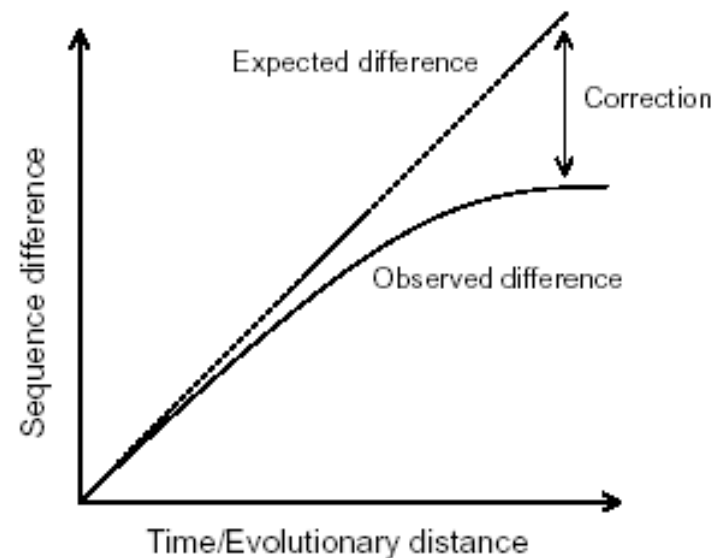
Seq 2 **AGCGGG**



When and why we need a model of evolution

- As a consequence of hidden changes sequences become progressively saturated: most of the sites changing have already changed before.
- As a result, more recent substitutions make little or no impact on the total number of observed differences.

Correcting for unobserved mutations

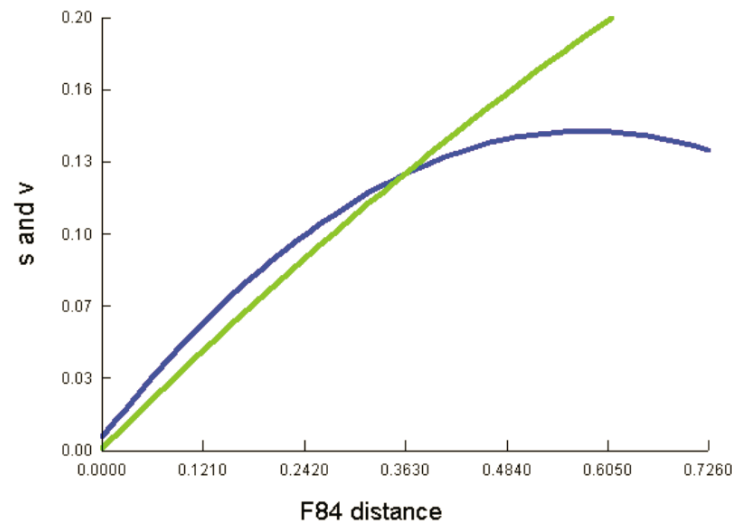


How to correct the difference?

Apply a substitution model that tries to estimate the correct number of substitutions.

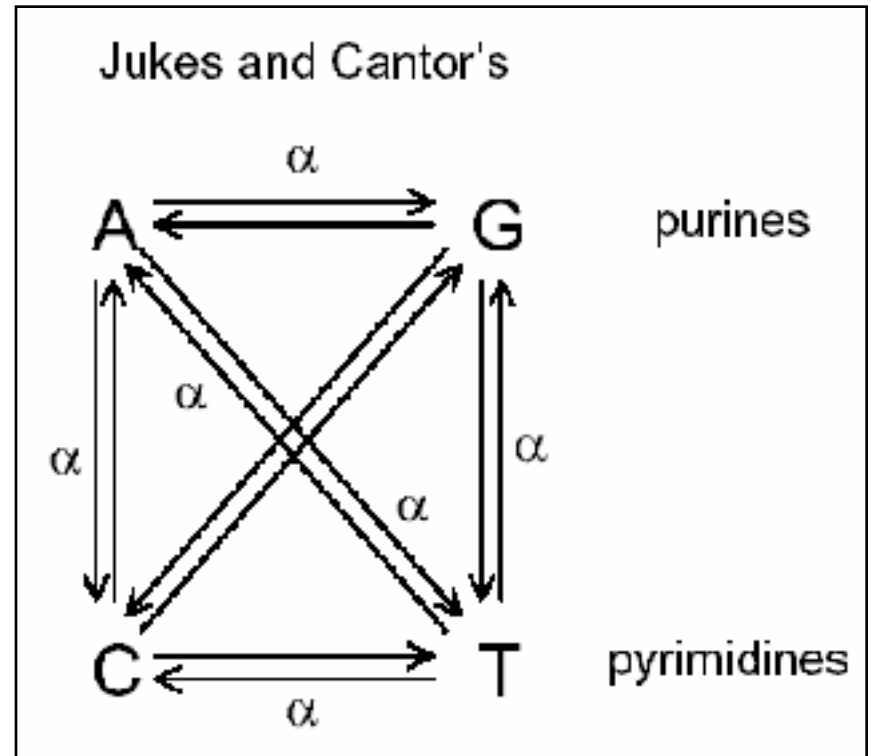
When and why we need a model of evolution

- In a phylogenetic context, models are used to predict the substitution process of the sequences through the tree branches. More explicitly, substitution (or evolutionary) models describe probabilistically the process by which homologous (aligned) character sequences (nucleotides or aminoacids) change through time



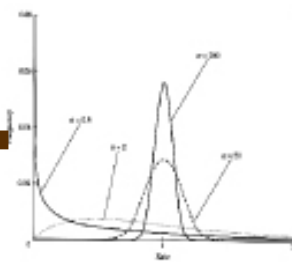
When and why we need a model of evolution

- A correction for multiple changes at a single site
 - hidden changes:
 - Jukes Cantor model: assumes all changes equally likely, equal base frequencies
 - $d = -3/4 \ln(1 - 4/3 p)$



The models

- The models generally are made of:
 - The composition: frequency of nucleotides or of aminoacids.
 - The substitution process: rate of change from one character state to another character state.
 - Other parameters

$$\text{Model} = \begin{Bmatrix} a & b & c & d \\ b & a & e & f \\ c & e & a & g \\ d & c & f & a \end{Bmatrix} + \pi = [a, c, g, t] + \text{graph}$$


Composition

- The frequency of nucleotides is represented by:

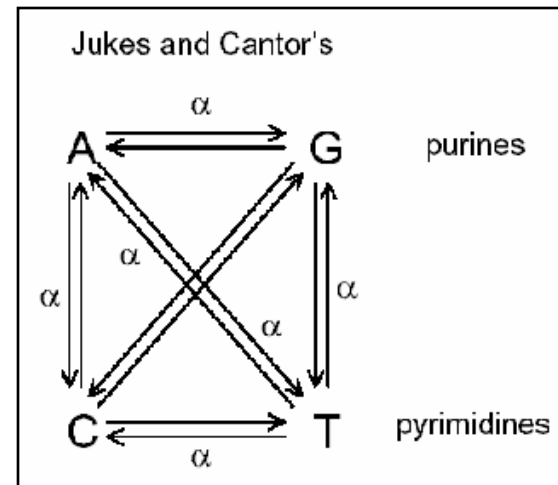
$$\pi = [0.25 \quad 0.25 \quad . \quad .]$$

- It is usually estimated from the data
- In the case of proteins it would be the frequency of aminoacids

Substitution process

- The substitution process is represented by a matrix
- For nucleotide sequences, there are 16 possible ways to describe substitutions - a 4x4 matrix.

$$\begin{array}{c}
 A \\
 C \\
 G \\
 T
 \end{array}
 P = \begin{array}{c}
 \begin{array}{cccc}
 A & C & G & T \\
 \left\{ \begin{array}{cccc}
 a & b & c & d \\
 e & f & g & h \\
 i & j & k & l \\
 m & n & o & p
 \end{array} \right.
 \end{array}
 \end{array}$$



- For nucleotides the substitution rate can be estimated from the data

Substitution matrix - an example

$$P = \begin{matrix} & \begin{matrix} \textcolor{blue}{a} & \textcolor{blue}{c} & \textcolor{blue}{g} & \textcolor{blue}{t} \end{matrix} \\ \begin{matrix} \textcolor{blue}{a} \\ \textcolor{blue}{c} \\ \textcolor{blue}{g} \\ \textcolor{blue}{t} \end{matrix} & \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix} \end{matrix}$$

- In this matrix, the probability of an a changing to a c is 0.01 and the probability of a c remaining the same is 0.983, etc.
- The rows of this matrix sum to 1 - meaning that for every nucleotide, we have covered all the possibilities of what might happen to it.
- The columns do not sum to anything in particular.

Substitution matrix

- For amino acids, the matrix Q is fixed and does not contain any free parameters.
- There are different models:
 - **Dayhoff** (Dayhoff et al., 1978) and **JTT** (Jones et al. , 1992) for use with proteins encoded on nuclear DNA,
 - **mtREV24** (Adachi and Hasegawa, 1996) for use with proteins encoded on mtDNA,
 - **BLOSUM 62** (Henikoff and Henikoff, 1992) and the **WAG** model (Whelan and Goldman, 2001) for use with distantly related amino acid sequences.
- For the different models, amino acid substitution rates are estimated based on empirical data.

Substitution matrix - an example

- For amino acids, the matrix Q is fixed and does not contain any free

BLOSUM 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Models – some examples

Jukes and Cantor (JC69):

All base compositions equal (0.25 each), rate of change from one base to another is the same

Kimura 2-Parameter (K2P):

All base compositions equal (0.25 each), different substitution rate for transitions and transversions.

Hasegawa-Kishino-Yano (HKY):

Like the K2P, but with base composition free to vary.

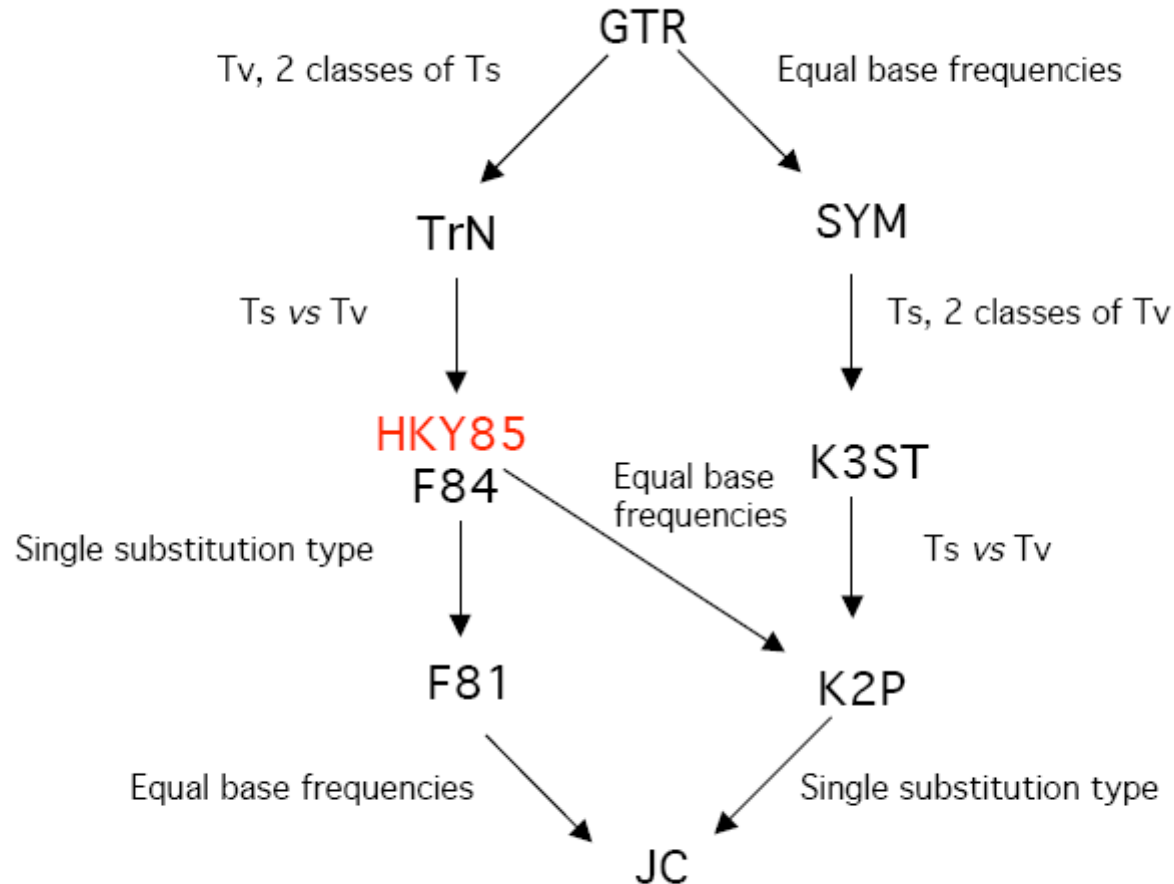
General Time Reversible (GTR):

Base composition free to vary, all possible substitutions can differ (6 types of substitution; symmetrical, so time-reversible).

$$R = \begin{bmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{bmatrix}$$

How to select a model

All these models are nested

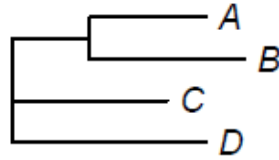


Comparison of likelihoods

- The likelihood of observing a given dataset is calculated for each model
- A test (likelihood ratio test, Akaike information criteria, Bayesian information criteria) is performed
- The simplest model that gives the best likelihood is chosen

An example

- We have some DNA data, and this tree



- Evaluate with JC (Jukes-Cantor) model: log likelihood is **-1008.587**
- Evaluate with K2P (Kimura 2-parameter) model: log likelihood is **-1008.268**
- The K2P model has one more parameter than the JC model, the tRatio.
- We got a better log likelihood with the extra parameter, by **0.319**
- Is the extra parameter worth adding?

An example

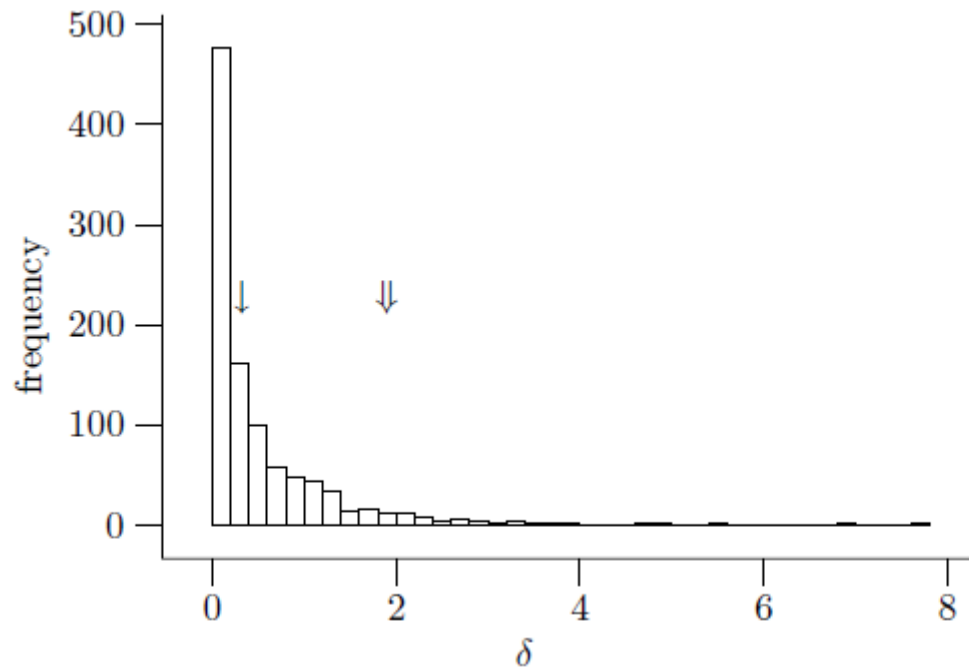
- Null hypothesis (generally): the extra parameter does not make any difference
- **Null hypothesis (specifically):** the tree and the JC model
 - We need to know how much of an improvement in likelihood we can expect *due to noise alone* when we add the parameter

An example

- Null hypothesis: the tree and the JC model
 - We need a null distribution, which we can get by simulating fake data many times under the null hypothesis
 - Evaluate the likelihood of each simulated data set with both the JC and the K2P models
 - Keep the log likelihood differences—they are the null distribution

An example

- We have generated many true null hypothesis data sets and evaluated them under the JC model and the K2P model. 95% of the differences are under 2.
- The statistic for our original data set was 0.319, and so it is *not* significant.
- In this case it is *not* worthwhile to add the extra parameter (tRatio).



Increasing the sophistication of models

- So far, the models we have dealt with assume that change is equally likely at all positions and that the rate of change is constant for the entire duration of the phylogeny.
- Selection intensity is rarely uniform across sites, so it is desirable to model site-by-site rate variation.

Increasing the sophistication of models

- **Invariable sites:**

- For a given dataset we can assume that a certain proportion of sites are not free to vary - purifying selection (related to function) prevents these sites from changing.
- We can therefore observe invariable positions either because they are under this selective constraint or because they have not had a chance to vary or because there is homoplasy in the dataset and a reversal (say) has caused the site to appear constant.

Increasing the sophistication of models

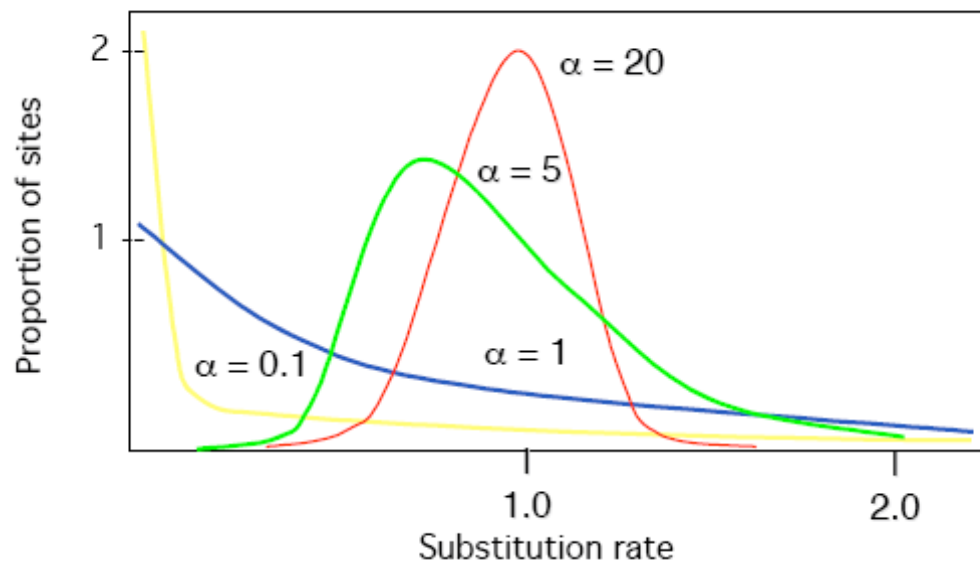
- The likelihood that a site is invariable can be calculated by incorporating this possibility into our model and calculating for every site the likelihood that it is an invariable site.
- It might improve the likelihood of the dataset if we remove a certain proportion of invariable sites.

Increasing the sophistication of models

- Variable sites:
 - Obviously other sites in the dataset are free to vary.
 - Selection intensity on these sites is rarely uniform, so it is desirable to model site-by-site rate variation.
 - This is done in two ways:
 - Site specific (codon position, or alpha helix etc.)
 - Using a discrete approximation to a continuous distribution (gamma distribution).

Increasing the sophistication of models

- Site rate heterogeneity:
 - A gamma distribution can be used to model site rate heterogeneity.



The shape of the gamma distribution for different values of alpha.

Increasing the sophistication of models

- Heterotachy:
 - The rates of substitution for each position can vary along the time
 - Covarion

Increasing the sophistication of models

- To an analysis I can apply different combinations of a model plus the complementary parameters:
 - i.e.
 - J&C model + gamma + I
 - GTR model + gamma