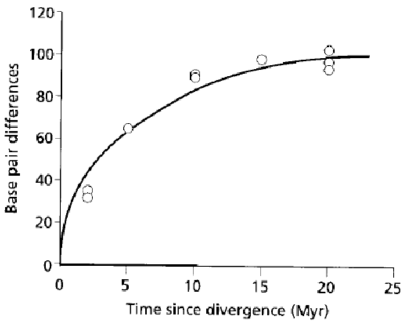
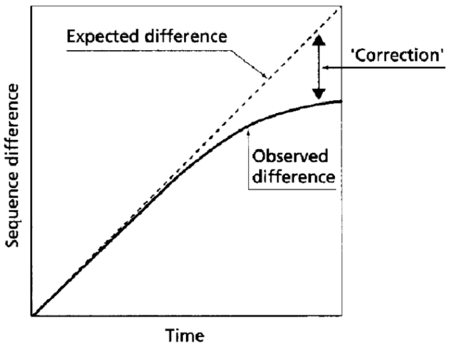
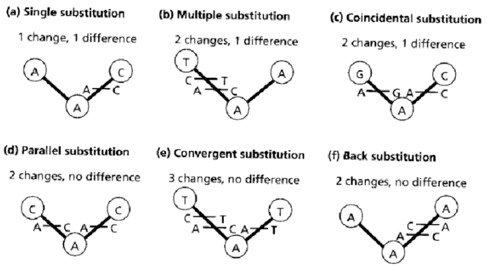
Questions?
Lab reports

Models of DNA substitutions Rate matrices and choosing the appropriate model

Sequence divergence “saturates” with time

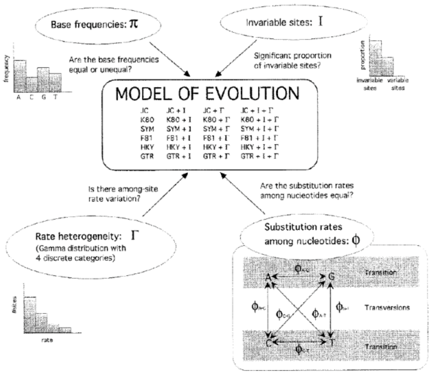






Factors to consider

- Base frequencies
- Substitution rates between nucleotides
- Invariable sites
- Rate heterogeneity


Base frequencies: π
Are the base frequencies equal or unequal?
frequency
A C G T
Invariable sites: I
Significant proportion of invariable sites?
proportion
invariable sites
variable sites
MODEL OF EVOLUTION
Is there among-site rate variation?
Are the substitution rates among nucleotides equal?
Rate heterogeneity: Γ
(Gamma distribution with 4 discrete categories)
#sites
rate
Substitution rates among nucleotides: ϕ
ϕ
A
G
C
T
Transition
Transversions
Transition

JC	JC + I	JC + Γ	JC + I + Γ
K80	K80 + I	K80 + Γ	K80 + I + Γ
SYM	SYM + I	SYM + Γ	SYM + I + Γ
F81	F81 + I	F81 + Γ	F81 + I + Γ
HKY	HKY + I	HKY + Γ	HKY + I + Γ
GTR	GTR + I	GTR + Γ	GTR + I + Γ

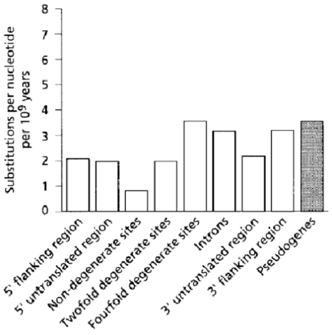
Are base frequencies equal? GC content

$$\mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}].$$

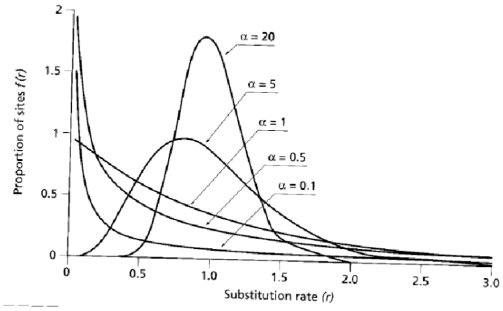
Site variation

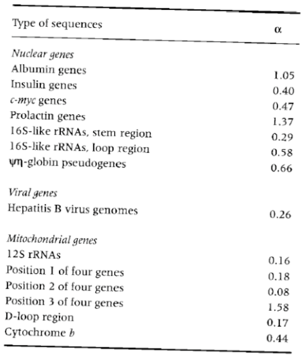
- Do all sites evolve at same rate?
  - Synonymous sites (codons)
  - Ribosomal RNA
  - Coding and noncoding

Variation in rates between sites is common



Model variation in sites with gamma parameter



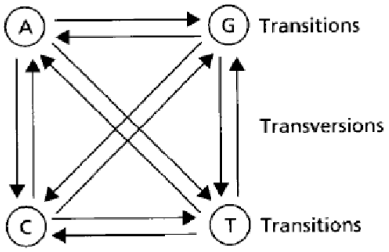


Type of sequences	α
Nuclear genes	
Albumin genes	1.05
Insulin genes	0.40
c-myc genes	0.47
Prolactin genes	1.37
16S-like rRNAs, stem region	0.29
16S-like rRNAs, loop region	0.58
ψη-globin pseudogenes	0.66
Viral genes	
Hepatitis B virus genomes	0.26
Mitochondrial genes	
12S rRNAs	0.16
Position 1 of four genes	0.18
Position 2 of four genes	0.08
Position 3 of four genes	1.58
D-loop region	0.17
Cytochrome *b*	0.44

Substitution rates among nucleotides

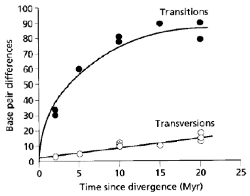
- Equal rates
- Transitions versus transversions

Transitions observed more often,
Even though more opportunities for transversions



But transitions and transversions have different rates
Ratio is often called k.

Type of sequences	Transition/transversion ratio (κ)
mtDNA	9.0
12S rRNA	1.75
α- and β-globins	0.66
Pseudo η-globin	2.70



$$\mathbf{P}_t = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

The substitution probability matrix can be computed or determined using simulations.

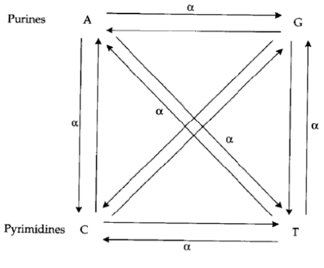
Matrix diagonal equal probability to maintain stationary frequencies of base pairs

Common models of DNA evolution

- Jukes-Cantor (JC)
  - Equal substitution rates and base freq.
- Kimura 2 parameter (K2P)
  - Transitions/transversion rates, equal base
- Felsensten 1981 (F81)
  - Unequal base freqs, equal subs. rate
- Hasegawa, Kisino and Yano (HKY)
  - Unequal base, transition/transversions rate
- General reversible (REV)
  - Unequal base freqs, all substitutions different rate

$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\,\tfrac{1}{4}\,\tfrac{1}{4}\,\tfrac{1}{4}]$$

Jukes Cantor model:
D = -3/4ln(1-4/3p)



K2P model

$$\mathbf{P}_i = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}].$$

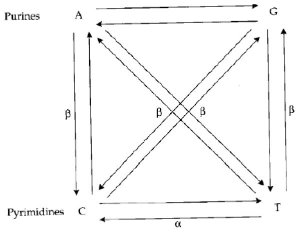
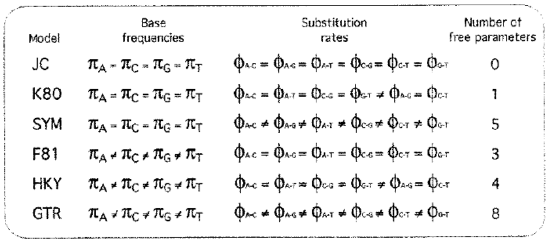
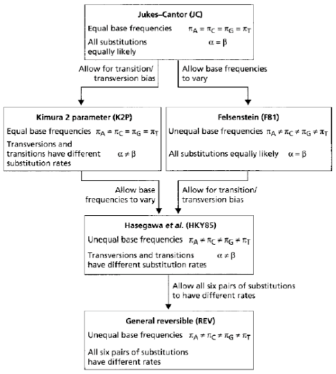


FIGURE 3.4 Two-parameter model of nucleotide substitution. The rate of transition (α) may not be equal to the rate of transversion (β).

How to choose appropriate model?



Model	Base frequencies	Substitution rates	Number of free parameters
JC	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}G} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{C\text{-}T} = \phi_{G\text{-}T}$	0
K80	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{G\text{-}T} \neq \phi_{A\text{-}G} = \phi_{C\text{-}T}$	1
SYM	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} \neq \phi_{A\text{-}G} \neq \phi_{A\text{-}T} \neq \phi_{C\text{-}G} \neq \phi_{C\text{-}T} \neq \phi_{G\text{-}T}$	5
F81	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}G} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{C\text{-}T} = \phi_{G\text{-}T}$	3
HKY	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{G\text{-}T} \neq \phi_{A\text{-}G} = \phi_{C\text{-}T}$	4
GTR	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} \neq \phi_{A\text{-}G} \neq \phi_{A\text{-}T} \neq \phi_{C\text{-}G} \neq \phi_{C\text{-}T} \neq \phi_{G\text{-}T}$	8



Jukes–Cantor (JC)
Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$
All substitutions equally likely $\alpha = \beta$

Allow for transition/ transversion bias

Allow base frequencies to vary

Kimura 2 parameter (K2P)
Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$
Transversions and transitions have different substitution rates $\alpha \neq \beta$

Felsenstein (F81)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
All substitutions equally likely $\alpha = \beta$

Allow base frequencies to vary

Allow for transition/ transversion bias

Hasegawa *et al.* (HKY85)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
Transversions and transitions have different substitution rates $\alpha \neq \beta$

Allow all six pairs of substitutions to have different rates

General reversible (REV)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
All six pairs of substitutions have different rates

What assumptions did we make?

- All nucleotides change independently
- The subsituion rate is constant over time and in different lineages
- The base composition is at equilibrium
- The conditional probabilities of nucleotide substitutions are the same for all sites and do not change over time..

How do we choose an appropriate model?

- Calculate likelihood of model given data
- Compare likelihoods of nested models

What is maximum likelihood?

Comparison to probability theory:

Probability of # heads in 5 coin tosses

Heads	Prob.
0	.03
1	.16
2	.31
3	.31
4	.16
5	.03

$P(x) = (n!/(n-x)!)p^x q^{n-x}$

Same calculation for coins with different bias.

	Bias of coin towards Heads				
Heads	.1	.3	.5	.7	.9
0	.59	.17	.03	.00	.00
1	.33	.36	.16	.03	.00
2	.07	.31	.31	.13	.01
3	.01	.13	.31	.31	.07
4	.00	.03	.16	.36	.33
5	.00	.00	.03	.17	.59

Maximum Likelihood (Box 5.1)

- L = Pr(D|H)
- Likehood = probability data (d) given the hypothesis (hypothesis).
- Often expressed as natural logarithms, or log-likelihood.

Calculating

- Difficult to solve all possibilities, but some some problems this is possible.
- Large number of calculations
  - Tree numbers
- Need to estimate

One way to get the likelihood is to estimate them using Makov Chain Monte Carlo methods.

-analogy to walking up hill.

Markov Chain Monte Carlo

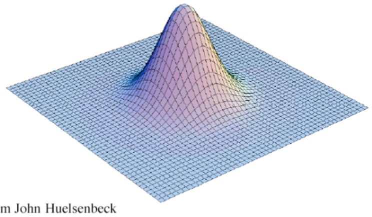
- Start with proposed state
- Perturb old state and calculate probability of new state
- Test if new state is better than old state, accept if ratio of new to old is greater than a randomly drawn number between 0 and 1.
- Move to new state if accepted, if not stay at old state
- Start over

Caveats: The proposal mechanism is at the discretion of the programmer, but must satisfy a few basic requirements: all states must be reachable, the chain must be aperiodic, and the mechanism must be stochastic.

Markov chain Monte Carlo

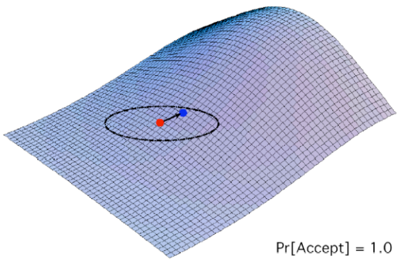
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing Machines. *J. Chem. Phys*. 21:1087–1091.

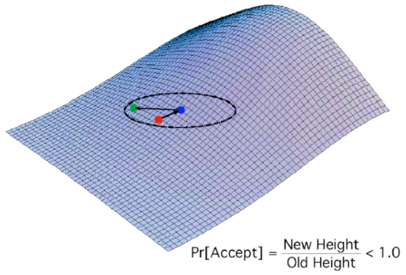
Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.



-Graphs from John Huelsenbeck

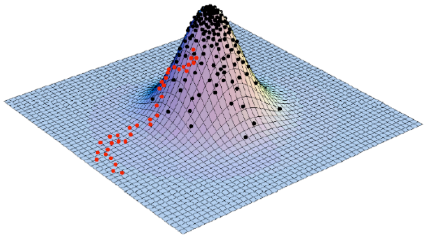
Circle represents amount of potential proposed change.





$$\text{Pr[Accept]} = \frac{\text{New Height}}{\text{Old Height}} < 1.0$$

Red dots = "burn in" period.



Likelihood ratio test statistic:

$$\Delta l = \log\left(\frac{\max\{L(\text{neutral model})\}}{\max\{L(\text{selection model})\}}\right)$$

$$= \log(\max\{L(\text{neutral model})\}) - \log(\max\{L(\text{selection model})\})$$

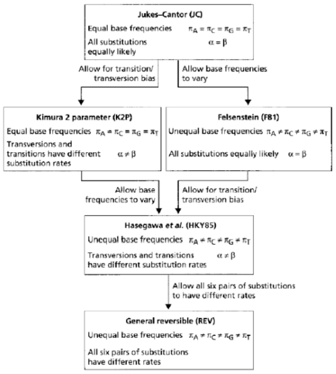
$-2\Delta l$ approximates χ^2 with *n* degrees freedom, where *n* is the difference in number of parameters between the nested models.

Test models using ModelTest:

Posada and Crandall 2001, *Syst. Biol.* **50**: 580-601

Test models

- Pair more general model with more complex
- Compare likelihoods of data given the model
- Test using LRT
- Continue testing more complex models



Jukes–Cantor (JC)

Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$

All substitutions equally likely $\alpha = \beta$

Allow for transition/transversion bias

Allow base frequencies to vary

Kimura 2 parameter (K2P)

Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$

Transversions and transitions have different substitution rates $\alpha \neq \beta$

Felsenstein (F81)

Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

All substitutions equally likely $\alpha = \beta$

Allow base frequencies to vary

Allow for transition/transversion bias

Hasegawa *et al.* (HKY85)

Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

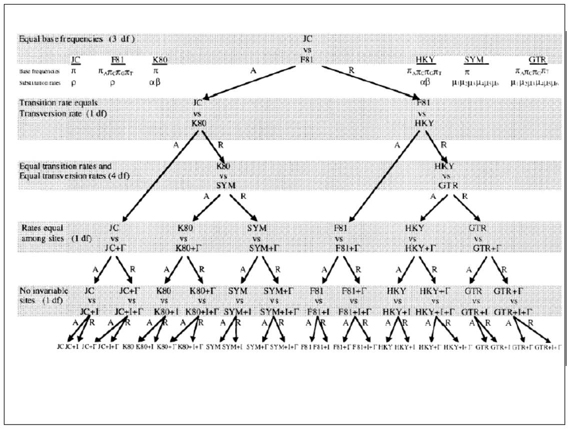
Transversions and transitions have different substitution rates $\alpha \neq \beta$

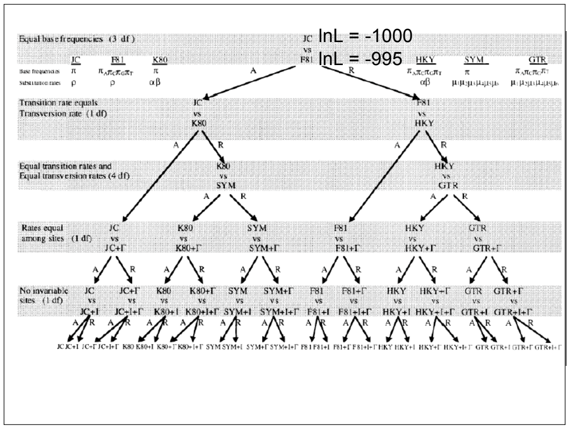
Allow all six pairs of substitutions to have different rates

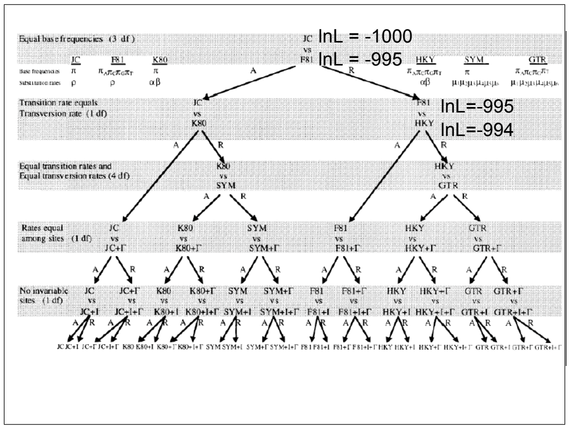
General reversible (REV)

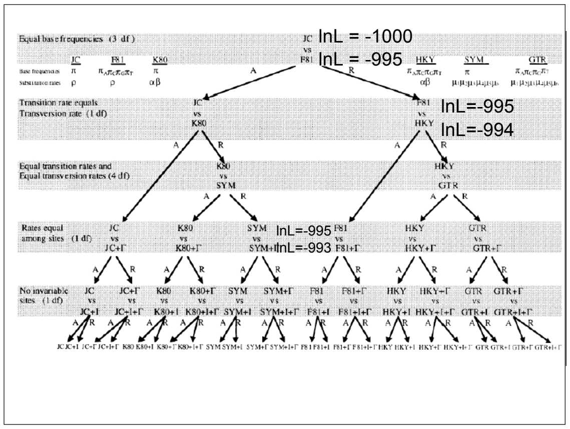
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

All six pairs of substitutions have different rates


Equal base frequencies (3 df)
JC vs F81
JC F81 K80 HKY SYM GTR
Base frequencies π π_Aπ_Cπ_Gπ_T π π_Aπ_Cπ_Gπ_T π π_Aπ_Cπ_Gπ_T
Substitution rates ρ ρ αβ αβ μ1μ2μ3μ4μ5μ6 μ1μ2μ3μ4μ5μ6
A R
Transition rate equals Transversion rate (1 df)
JC vs K80
F81 vs HKY
Equal transition rates and Equal transversion rates (4 df)
K80 vs SYM
HKY vs GTR
Rates equal among sites (1 df)
JC vs JC+Γ
K80 vs K80+Γ
SYM vs SYM+Γ
F81 vs F81+Γ
HKY vs HKY+Γ
GTR vs GTR+Γ
No invariable sites (1 df)
JC vs JC+I
JC+Γ vs JC+I+Γ
K80 vs K80+I
K80+Γ vs K80+I+Γ
SYM vs SYM+I
SYM+Γ vs SYM+I+Γ
F81 vs F81+I
F81+Γ vs F81+I+Γ
HKY vs HKY+I
HKY+Γ vs HKY+I+Γ
GTR vs GTR+I
GTR+Γ vs GTR+I+Γ
JC JC+I JC+Γ JC+I+Γ K80 K80+I K80+Γ K80+I+Γ SYM SYM+I SYM+Γ SYM+I+Γ F81 F81+I F81+Γ F81+I+Γ HKY HKY+I HKY+Γ HKY+I+Γ GTR GTR+I GTR+Γ GTR+I+Γ


Equal base frequencies (3 df)
JC vs F81
lnL = -1000
lnL = -995
JC F81 K80 HKY SYM GTR
Base frequencies π π_Aπ_Cπ_Gπ_T π π_Aπ_Cπ_Gπ_T π π_Aπ_Cπ_Gπ_T
Substitution rates ρ ρ αβ αβ μ1μ2μ3μ4μ5μ6 μ1μ2μ3μ4μ5μ6
A R
Transition rate equals Transversion rate (1 df)
JC vs K80
F81 vs HKY
Equal transition rates and Equal transversion rates (4 df)
K80 vs SYM
HKY vs GTR
Rates equal among sites (1 df)
JC vs JC+Γ
K80 vs K80+Γ
SYM vs SYM+Γ
F81 vs F81+Γ
HKY vs HKY+Γ
GTR vs GTR+Γ
No invariable sites (1 df)
JC vs JC+I
JC+Γ vs JC+I+Γ
K80 vs K80+I
K80+Γ vs K80+I+Γ
SYM vs SYM+I
SYM+Γ vs SYM+I+Γ
F81 vs F81+I
F81+Γ vs F81+I+Γ
HKY vs HKY+I
HKY+Γ vs HKY+I+Γ
GTR vs GTR+I
GTR+Γ vs GTR+I+Γ
JC JC+I JC+Γ JC+I+Γ K80 K80+I K80+Γ K80+I+Γ SYM SYM+I SYM+Γ SYM+I+Γ F81 F81+I F81+Γ F81+I+Γ HKY HKY+I HKY+Γ HKY+I+Γ GTR GTR+I GTR+Γ GTR+I+Γ


Equal base frequencies (3 df)
JC vs F81
lnL = -1000
lnL = -995
JC F81 K80 HKY SYM GTR
Base frequencies π π_Aπ_Cπ_Gπ_T π π_Aπ_Cπ_Gπ_T π π_Aπ_Cπ_Gπ_T
Substitution rates ρ ρ αβ αβ μ1μ2μ3μ4μ5μ6 μ1μ2μ3μ4μ5μ6
A R
Transition rate equals Transversion rate (1 df)
JC vs K80
F81 vs HKY
lnL=-995
lnL=-994
Equal transition rates and Equal transversion rates (4 df)
K80 vs SYM
HKY vs GTR
Rates equal among sites (1 df)
JC vs JC+Γ
K80 vs K80+Γ
SYM vs SYM+Γ
F81 vs F81+Γ
HKY vs HKY+Γ
GTR vs GTR+Γ
No invariable sites (1 df)
JC vs JC+I
JC+Γ vs JC+I+Γ
K80 vs K80+I
K80+Γ vs K80+I+Γ
SYM vs SYM+I
SYM+Γ vs SYM+I+Γ
F81 vs F81+I
F81+Γ vs F81+I+Γ
HKY vs HKY+I
HKY+Γ vs HKY+I+Γ
GTR vs GTR+I
GTR+Γ vs GTR+I+Γ
JC JC+I JC+Γ JC+I+Γ K80 K80+I K80+Γ K80+I+Γ SYM SYM+I SYM+Γ SYM+I+Γ F81 F81+I F81+Γ F81+I+Γ HKY HKY+I HKY+Γ HKY+I+Γ GTR GTR+I GTR+Γ GTR+I+Γ


Equal base frequencies (3 df)
JC vs F81
lnL = -1000
lnL = -995
JC F81 K80 HKY SYM GTR
Base frequencies π π_Aπ_Cπ_Gπ_T π π_Aπ_Cπ_Gπ_T π π_Aπ_Cπ_Gπ_T
Substitution rates ρ ρ αβ αβ μ1μ2μ3μ4μ5μ6 μ1μ2μ3μ4μ5μ6
A R
Transition rate equals Transversion rate (1 df)
JC vs K80
F81 vs HKY
lnL=-995
lnL=-994
Equal transition rates and Equal transversion rates (4 df)
K80 vs SYM
HKY vs GTR
Rates equal among sites (1 df)
JC vs JC+Γ
K80 vs K80+Γ
SYM vs SYM+Γ
F81 vs F81+Γ
lnL=-995
lnL=-993
HKY vs HKY+Γ
GTR vs GTR+Γ
No invariable sites (1 df)
JC vs JC+I
JC+Γ vs JC+I+Γ
K80 vs K80+I
K80+Γ vs K80+I+Γ
SYM vs SYM+I
SYM+Γ vs SYM+I+Γ
F81 vs F81+I
F81+Γ vs F81+I+Γ
HKY vs HKY+I
HKY+Γ vs HKY+I+Γ
GTR vs GTR+I
GTR+Γ vs GTR+I+Γ
JC JC+I JC+Γ JC+I+Γ K80 K80+I K80+Γ K80+I+Γ SYM SYM+I SYM+Γ SYM+I+Γ F81 F81+I F81+Γ F81+I+Γ HKY HKY+I HKY+Γ HKY+I+Γ GTR GTR+I GTR+Γ GTR+I+Γ

Next week reading will be on website by Friday noon

- Want printout, stop by office