Modelos de Evolução de DNA –sequências de nucleotidos

1- Substituições de nucleótidos

2- modelos de substituição

3- A escolha dos modelos

Árvores filogéneticas

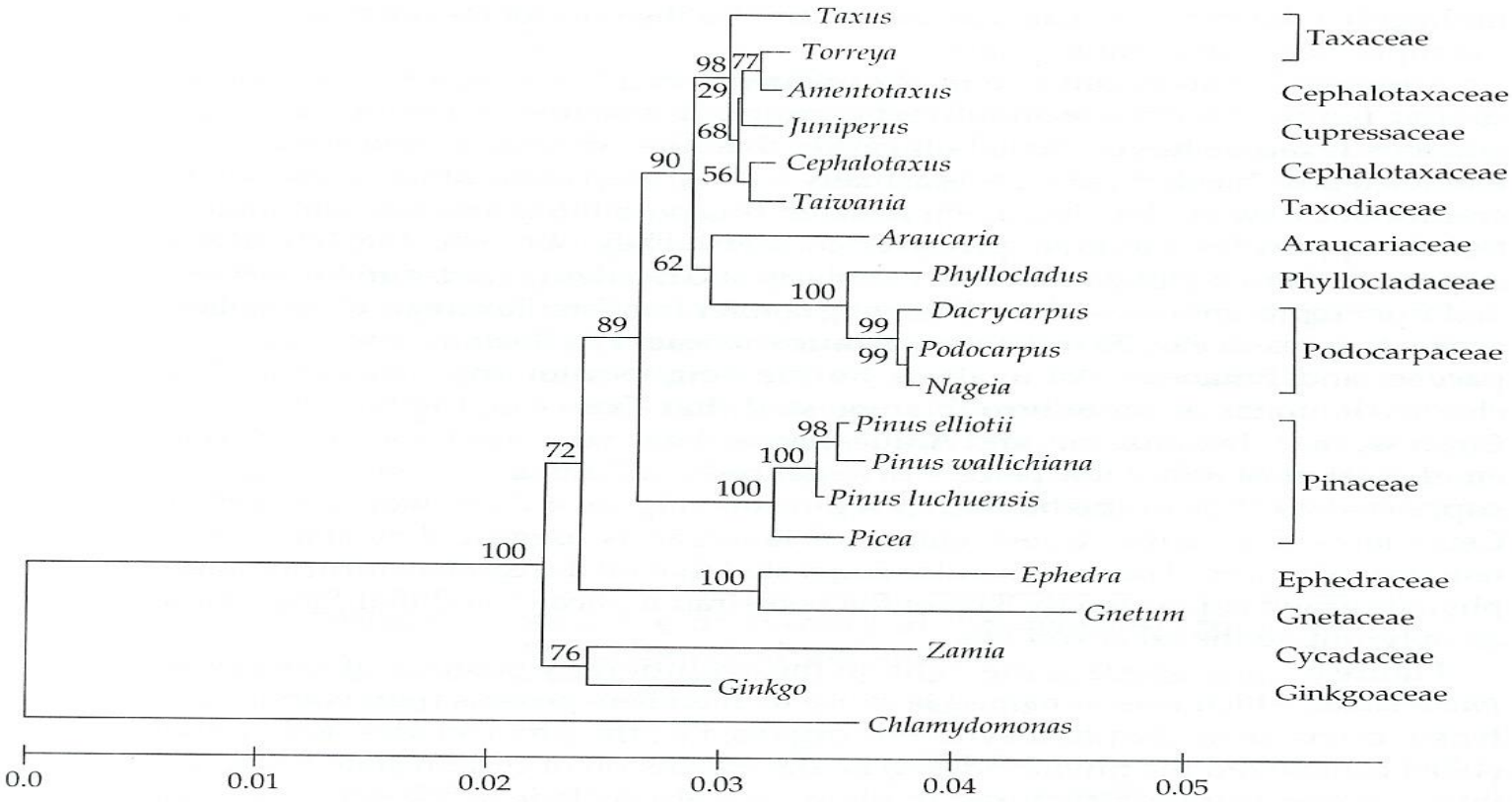


Figure 6.7 A phyloegentic tree of conifers inferred from 18S rRNA or rDNA sequences. Taxaceae, Cephalotaxaceae, Cupressacea, Taxodiaceae, Araucariaceae, Podocarpaceae, and Pinaceae are the seven families of conifers in the classification system of Pilger (1926). Whether Phyllocladaceae should be given a family status or should be a genus of Podocarpaceae is subject to dispute (see text). The other four gymnosperm species included in the tree are not conifers. The chlamydomonas was used as an outgroup. The tree was inferred by the neighbor-joining method. The number of transitional substitutions per site (d_S) and that of transversional substitutions per site (d_V) were computed using Kimura's (1980) two-parameter model and the weighted distance was computed by $\bar{d} = 0.4d_S + 0.6d_V$. The scale is given in $\bar{d}$. The numbers on the branches are bootstrap proportions. From Chaw et al. (1995).

Ou mais recentemente este tipo de legendas das árvores filogenéticas

The number of transitional substitutions per site (ds) and that of transversional substitutions (dv) were computed using Kimura (1980) two-parameter model and the weighted distance was computed by

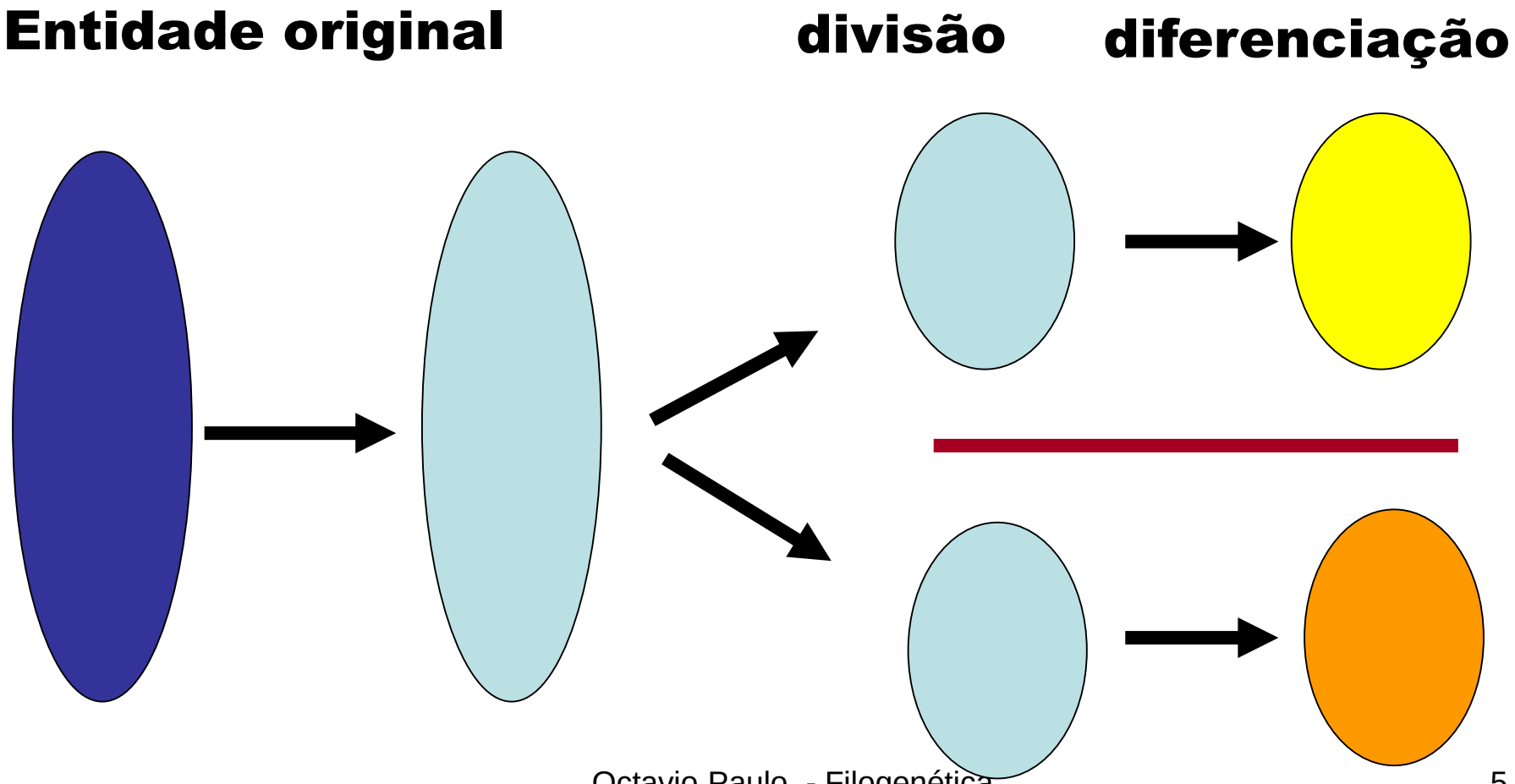
$$d = 0.4ds+0.6dv$$

Ou mais recentemente este tipo de legendas das árvores filogenéticas

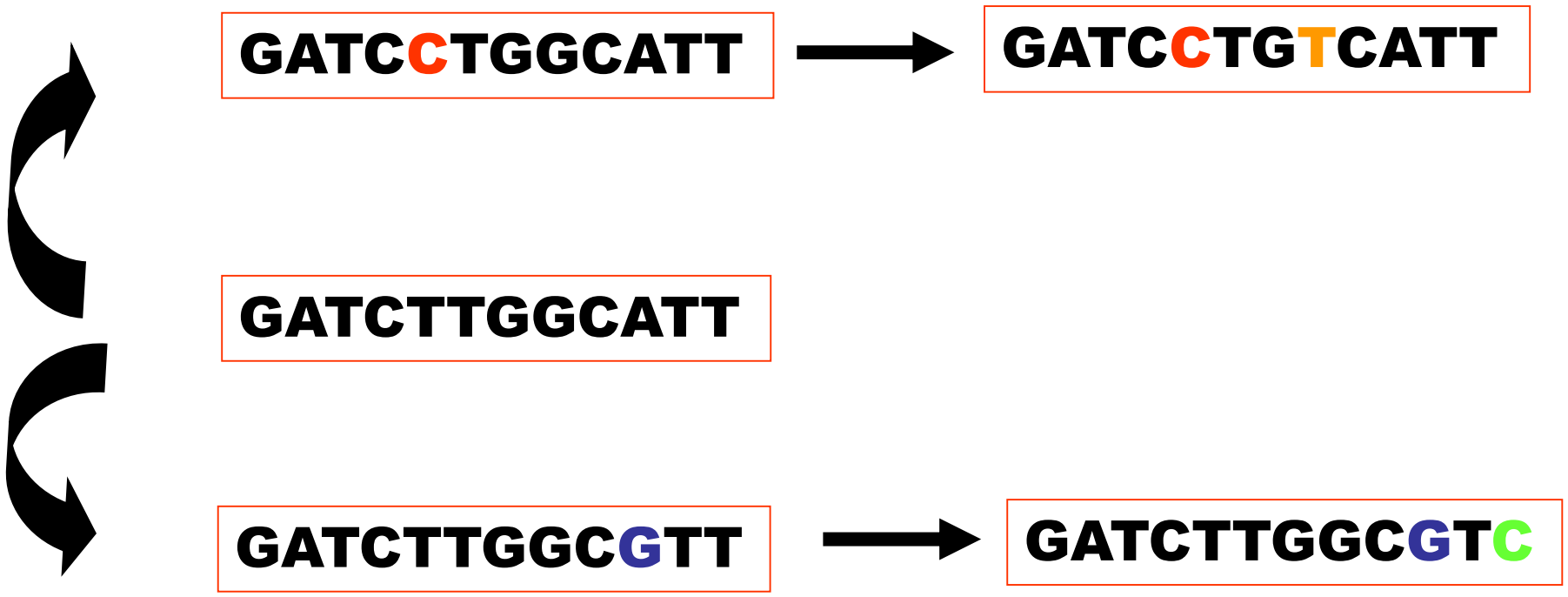
The selected evolutionary model, based on hierarchical likelihood ratio test, was the Tamura-Nei (1993) with gamma correction (TrN+G) with a G-A transition rate of 4.63957 and C-T rate of 12.0532, the proportion of invariable sites of zero and the shape parameter of the gamma distribution was 0.1394.

The Akaike Information Criterion, selected the GTR+G model (A-C transversion rate of 3.8697, G-A transition rates of 10.8879, A-T rate of 3.5091, C-G rate of 0.1615 and C-T rate of 32.0765, the proportion of invariable sites of zero and the shape parameter of the gamma distribution was 0.1469).

O acumular de substituições com o tempo...



O acumular de substituições com o tempo...

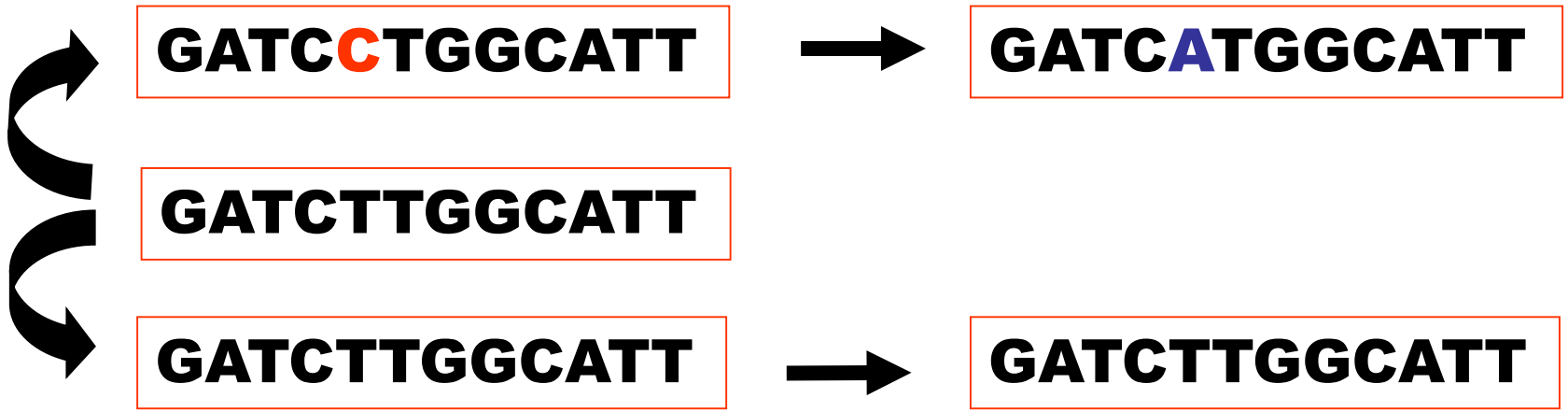


e com mais tempo...

Multiple hits I

Susbtituição simples
1 substituição 1 diferença

Susbtituição multipla
2 susbtituições 1 diferença



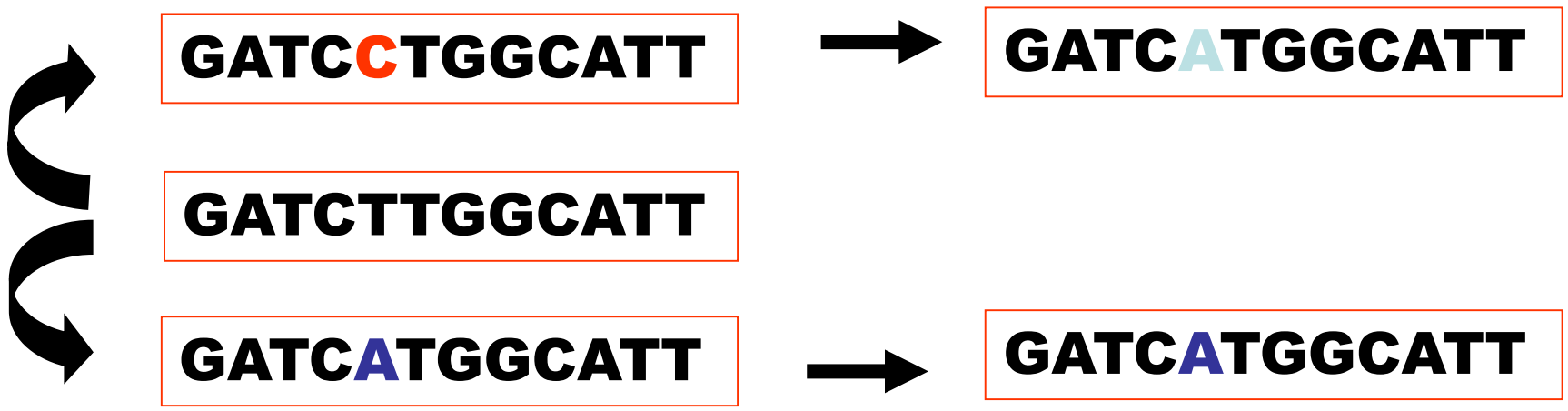
Multiple hits II

Susbtituição coincidente
2 susbtituições 1 diferença

Substituição paralela
2 susbtituições 0 diferença

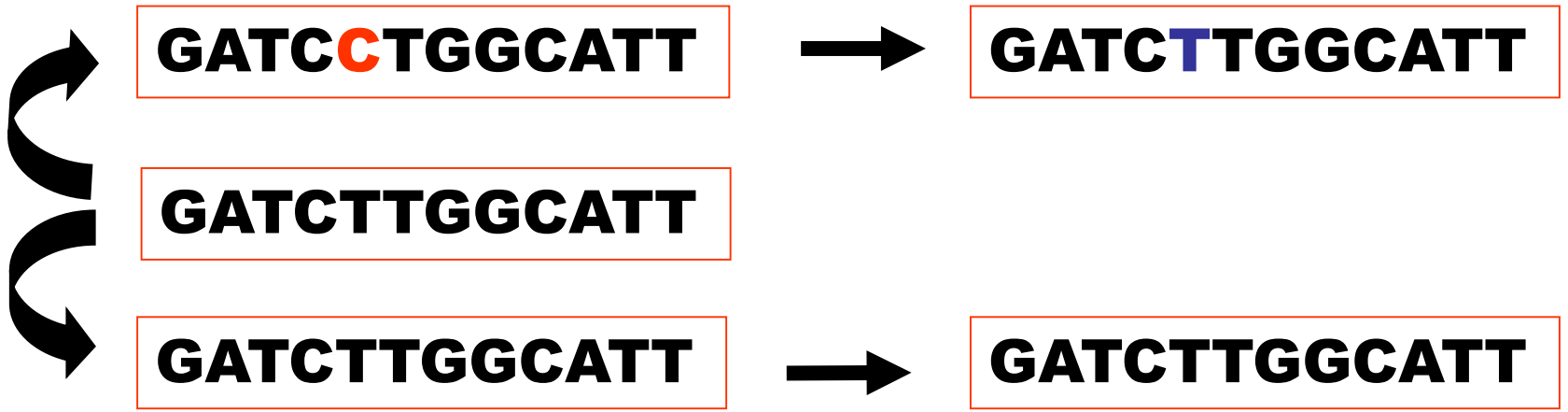
Multiple hits III

Susbtituição convergente
3 susbtituições 0 diferenças

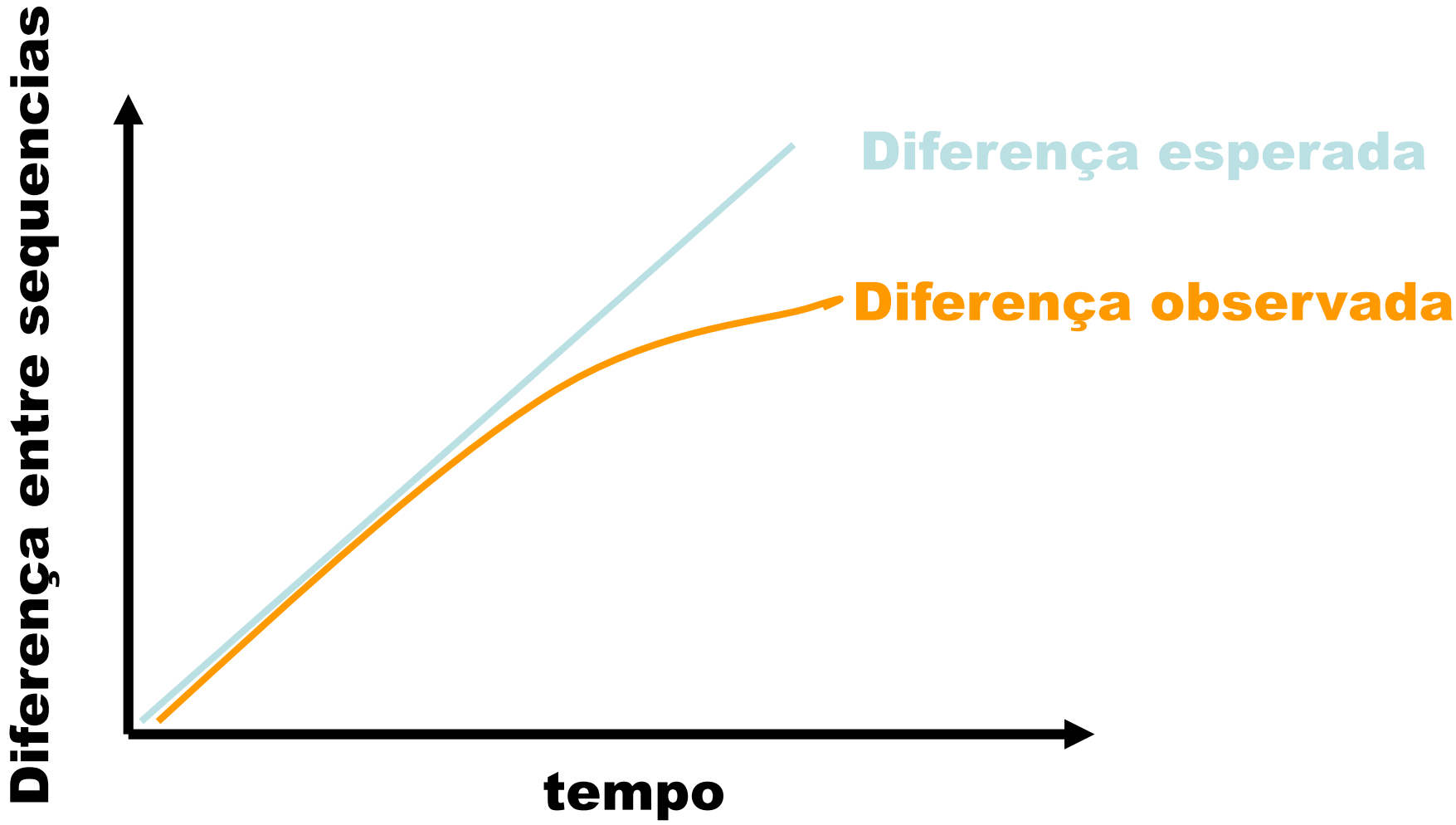


Multiple hits IV

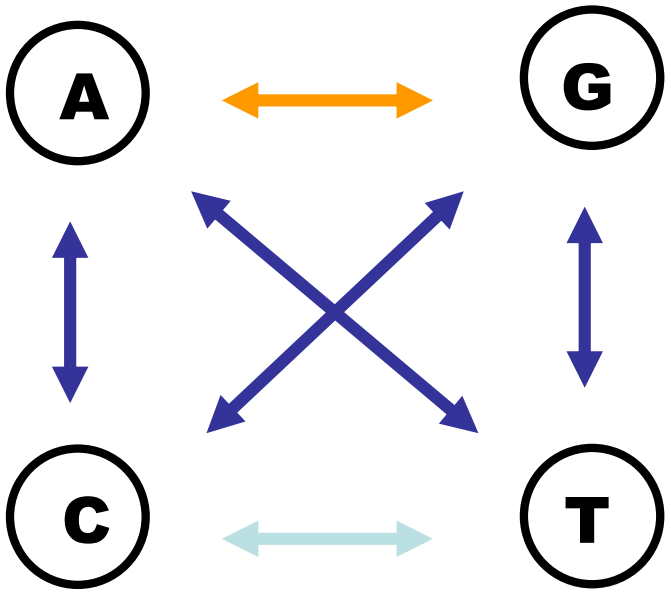
Retro (back) Susbtituição
2 substituição 0 diferença



Saturação


Diferença entre sequencias
tempo
Diferença esperada
Diferença observada

Tipos de substituições

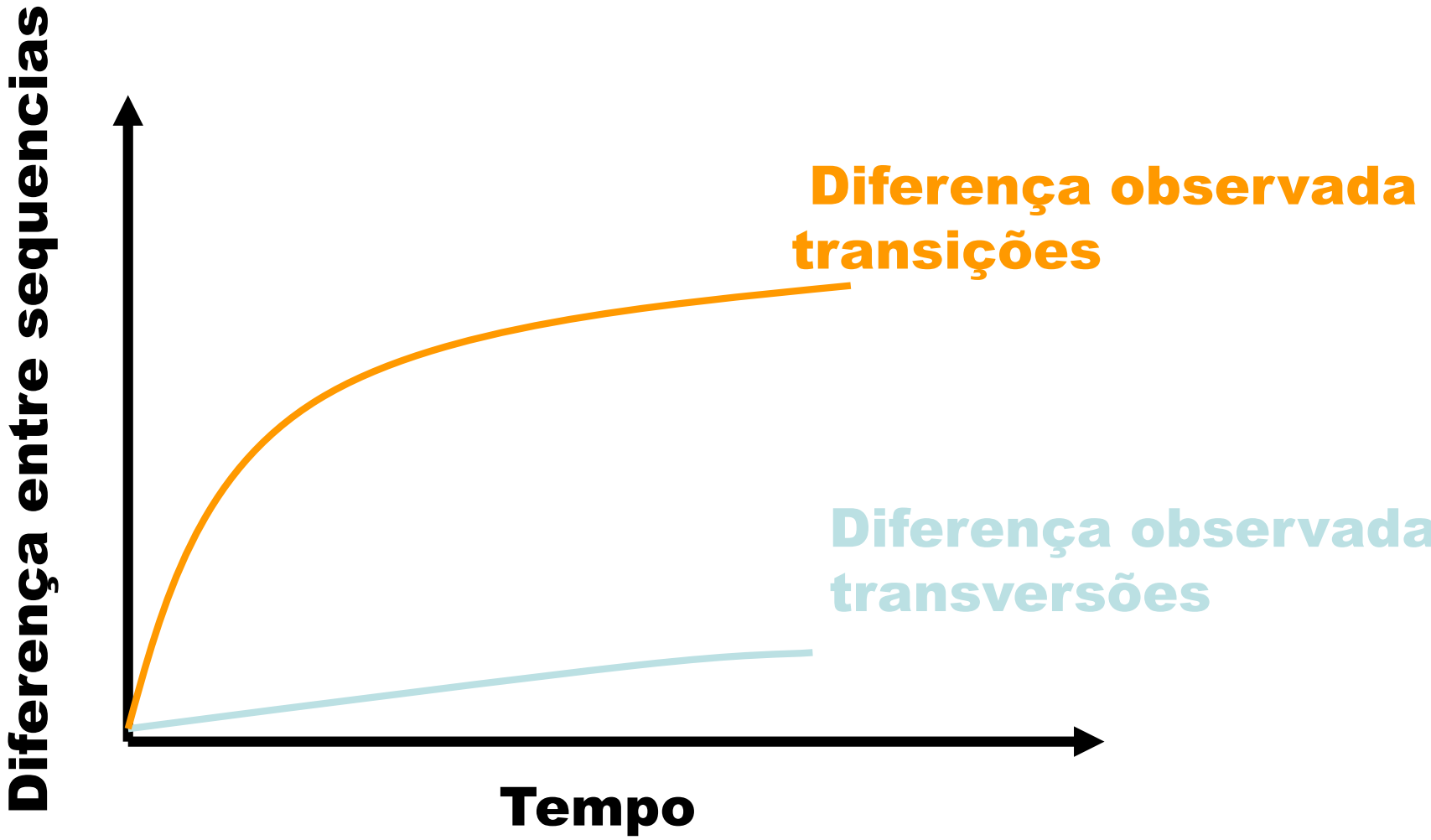


Transições (purina)

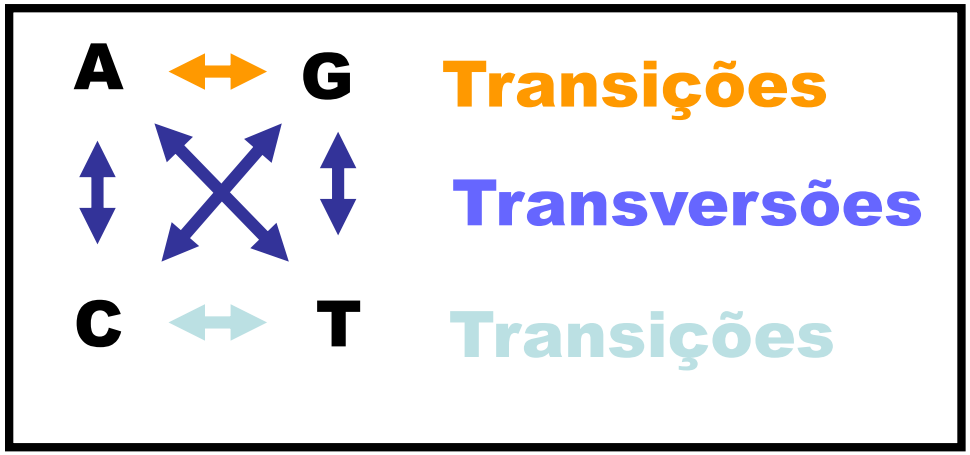
Transversões

Transições (pirimidina)

Saturação - transições e transversões



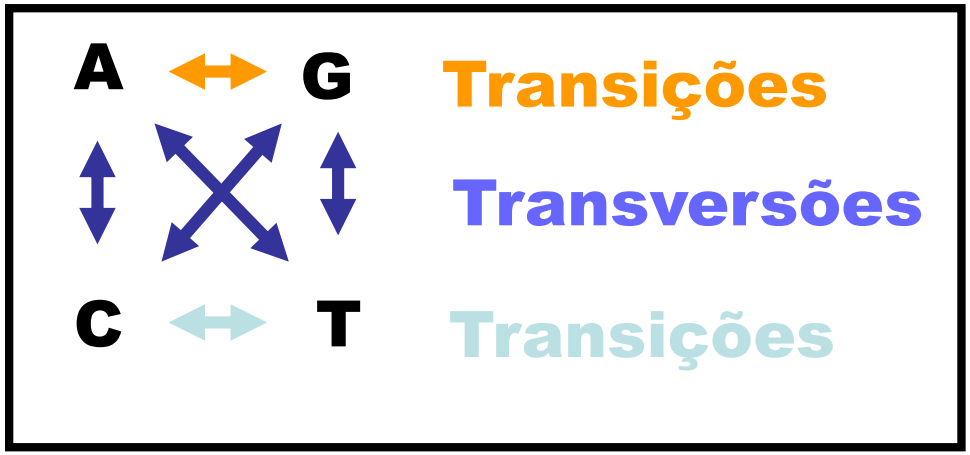
Matrix de probabilidade de substituição :



p_{AT} é a probabilidade de num dado site (posição) A passar a T durante um intervalo de tempo (t)

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

Susbtituições assumidas como um processo em cadeia de Markov



$t_0 \rightarrow t_1 \rightarrow t_2$

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

Diagonal :

p_{AA} é a probabilidade de num dado site (posição) A permancer (ou mudar para A) A durante um intervalo de tempo (t)

$$p_{AA} = 1 - (\; p_{AC} \; + \; p_{AG} \; + \; p_{AT} \;)$$

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

Vector de composição das bases em equilibrio

$$\mathbf{f} = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$$

Modelo Geral

Matrix de probabilidade de substituição

Vector de composição das bases

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

$$f = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$$

Modelo de Jukes-Cantor (1969)

$$P_t = \begin{bmatrix} p_{AA} & \alpha & \alpha & \alpha \\ \alpha & p_{CC} & \alpha & \alpha \\ \alpha & \alpha & p_{GG} & \alpha \\ \alpha & \alpha & \alpha & p_{TT} \end{bmatrix}$$

$$\mathbf{f} = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$$

$$\mathbf{f} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

$$p_{AA} = 1-3\alpha$$

Distancia de Jukes-Cantor (1969)

Distancia entre duas sequencias é dada por :

$$d = -(3/4) * \ln(1-(4/3)*p)$$

em que p é o número observado de diferenças entre as duas sequencias

Como se chega à expressão da distancia de Jukes-Cantor (1969)?

1 $p_{A(1)} = 1-3\alpha$ reverte para A

2 $p_{A(2)} = (1-3\alpha)\ p_{A(1)} + \alpha\ (1-p_{A(1)})$

3 $p_{A(t+1)} - p_{A(t)} = -3\alpha p_{A(t)} + \alpha\ (1-p_{A(t)})$

4 $\Delta p_{A(t)} = -4\alpha p_{A(t)} + \alpha$

5 $dp_{A(t)} / dt = -4\alpha p_{A(t)} + \alpha$

6 $p_{A(t)} = \frac{1}{4} + (p_{A(0)} - 1/4)e^{-4\alpha t}$

Como se chega à expressão da distancia de Jukes-Cantor (1969)? II

6 $p_{A(t)} = ¼ + (p_{A(0)} - 1/4) e^{-4\alpha t}$

7 quando $p_{A(0)} = 1$ $\quad p_{A(t)} = 1/4 + 3/4\ e^{-4\alpha t}$

8 quando $p_{A(0)} = 0$ $\quad p_{A(t)} = 1/4 - 1/4\ e^{-4\alpha t}$

Generalizando:

9 $p_{ii(t)} = 1/4 + 3/4\ e^{-4\alpha t}$

10 $p_{ij(t)} = 1/4 - 1/4\ e^{-4\alpha t}$

Como se chega à expressão da distancia de Jukes-Cantor (1969)? III

Então qual é a probabilidade de duas sequencias serem identicas ?

11 $p_{AA(t)}{}^{*}\ p_{AA(t)}$

12 $I_{(t)} = p_{AA(t)}{}^{2} + p_{AT(t)}{}^{2} + p_{AC(t)}{}^{2} + p_{AT(t)}{}^{2}$

da equação 9 e 10

13 $I_{(t)} = 1/4 + 3/4\ e^{-8\alpha t}$

14 **serem diferentes é** $p = 1 - I_{(t)}$

15 $p = 3/4\ (1 - e^{-8\alpha t})$

Como se chega à expressão da distancia de Jukes-Cantor (1969)? IV

15 $p = 3/4\ (1-e^{-8\alpha t})$

16 $8\alpha t = -\ln(1- 4p/3)$

17 sabendo que $K = 2(3\alpha t)$

Número esperado de substituições por site numa linhagem

$$K = -(3/4) * \ln(1- (4/3)*p)$$

Em que p se assume como a proporção observada dos diferentes nucleotidos entre as duas sequencias

Modelo de Kimura 2-parametros (1980)

$$P_t = \begin{bmatrix} p_{AA} & \beta & \alpha & \beta \\ \beta & p_{CC} & \beta & \alpha \\ \alpha & \beta & p_{GG} & \beta \\ \beta & \alpha & \beta & p_{TT} \end{bmatrix}$$

$$\mathbf{f} = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$$

$$\mathbf{f} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

Modelo de Felsenstein (1981)

$$P_t = \begin{bmatrix} p_{AA} & \pi_C\alpha & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & p_{CC} & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & p_{GG} & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & \pi_G\alpha & p_{TT} \end{bmatrix}$$

$$f = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$$

Modelo de HKY (1985)

$$P_t = \begin{bmatrix} p_{AA} & \pi_C \beta & \pi_G \alpha & \pi_T \beta \\ \pi_A \beta & p_{CC} & \pi_G \beta & \pi_T \alpha \\ \pi_A \alpha & \pi_C \beta & p_{GG} & \pi_T \beta \\ \pi_A \beta & \pi_C \alpha & \pi_G \beta & p_{TT} \end{bmatrix}$$

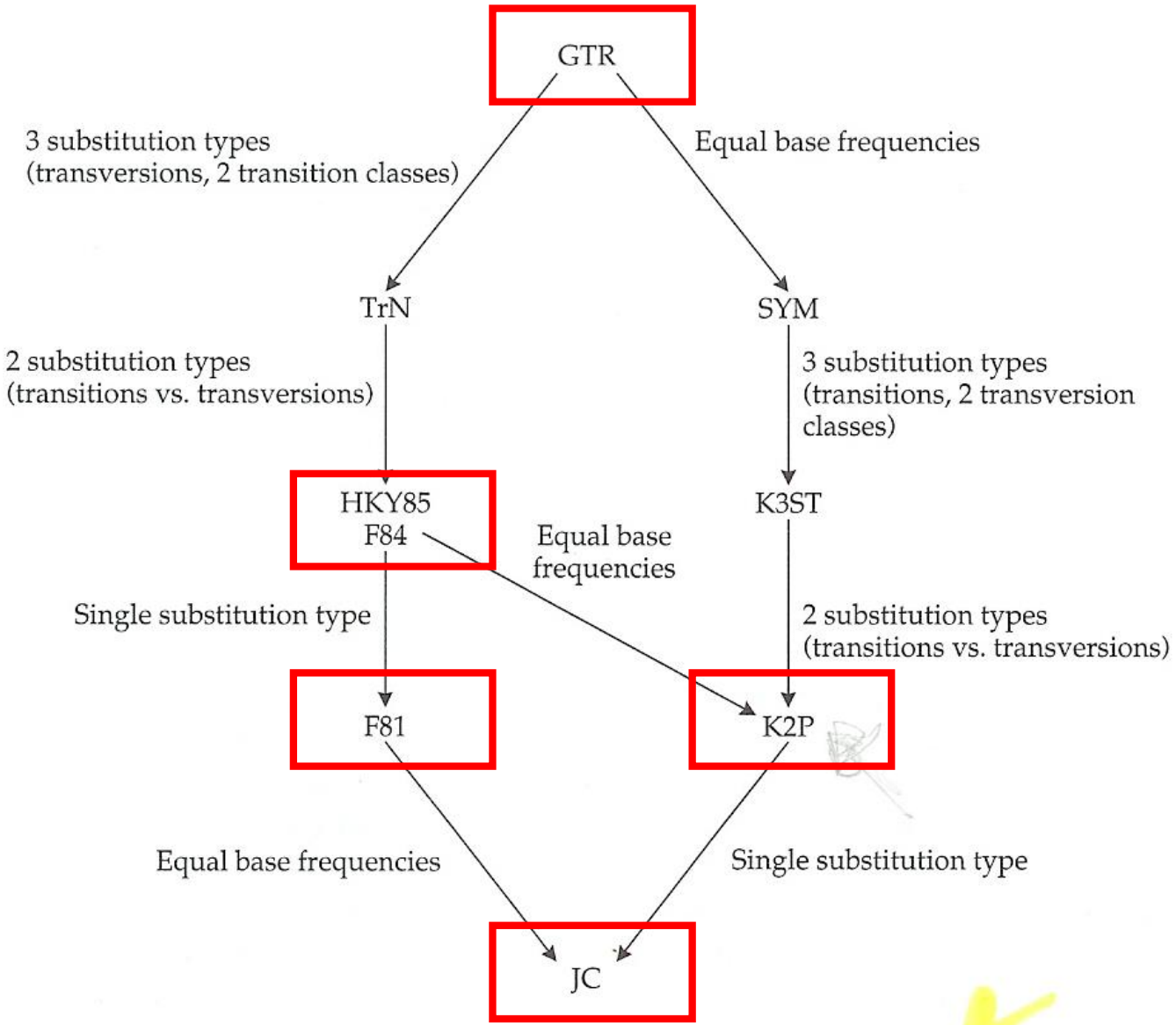
$$f = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$$

Modelo de GTR

$$P_t = \begin{bmatrix} p_{AA} & \pi_C a & \pi_G b & \pi_T c \\ \pi_A a & p_{CC} & \pi_G d & \pi_T e \\ \pi_A b & \pi_C d & p_{GG} & \pi_T f \\ \pi_A c & \pi_C e & \pi_G f & p_{TT} \end{bmatrix}$$

$$f = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$$

Modelos – diagrama compreensivo



Pressupostos dos modelos

1 - Substituição constante ao longo do tempo e entre todas as linhagens

2 - Composição das bases está equilíbrio

3 - os sites mudam independentemente

4 – probabilidade de cada nucleotido ser substituido são idênticas para todos os sites e não variam com o tempo

Gamma distribution

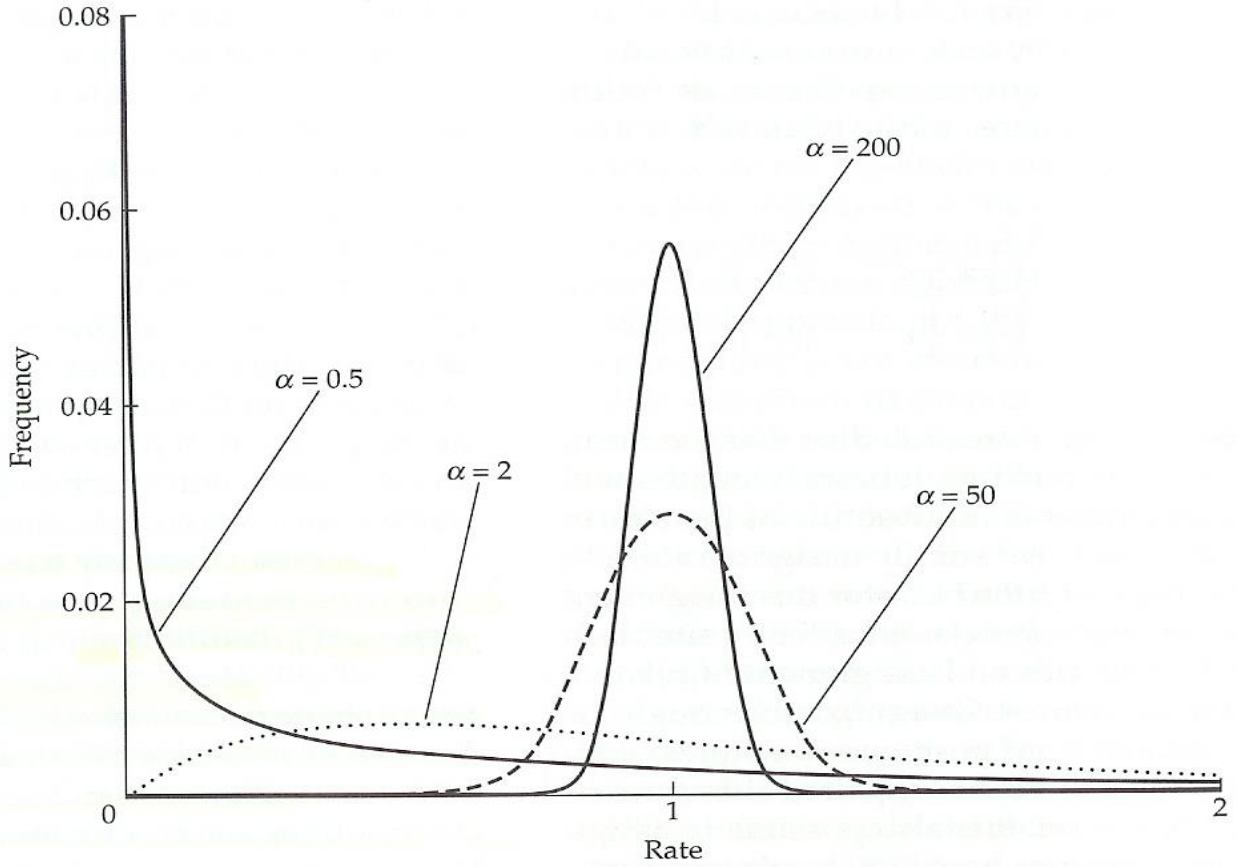
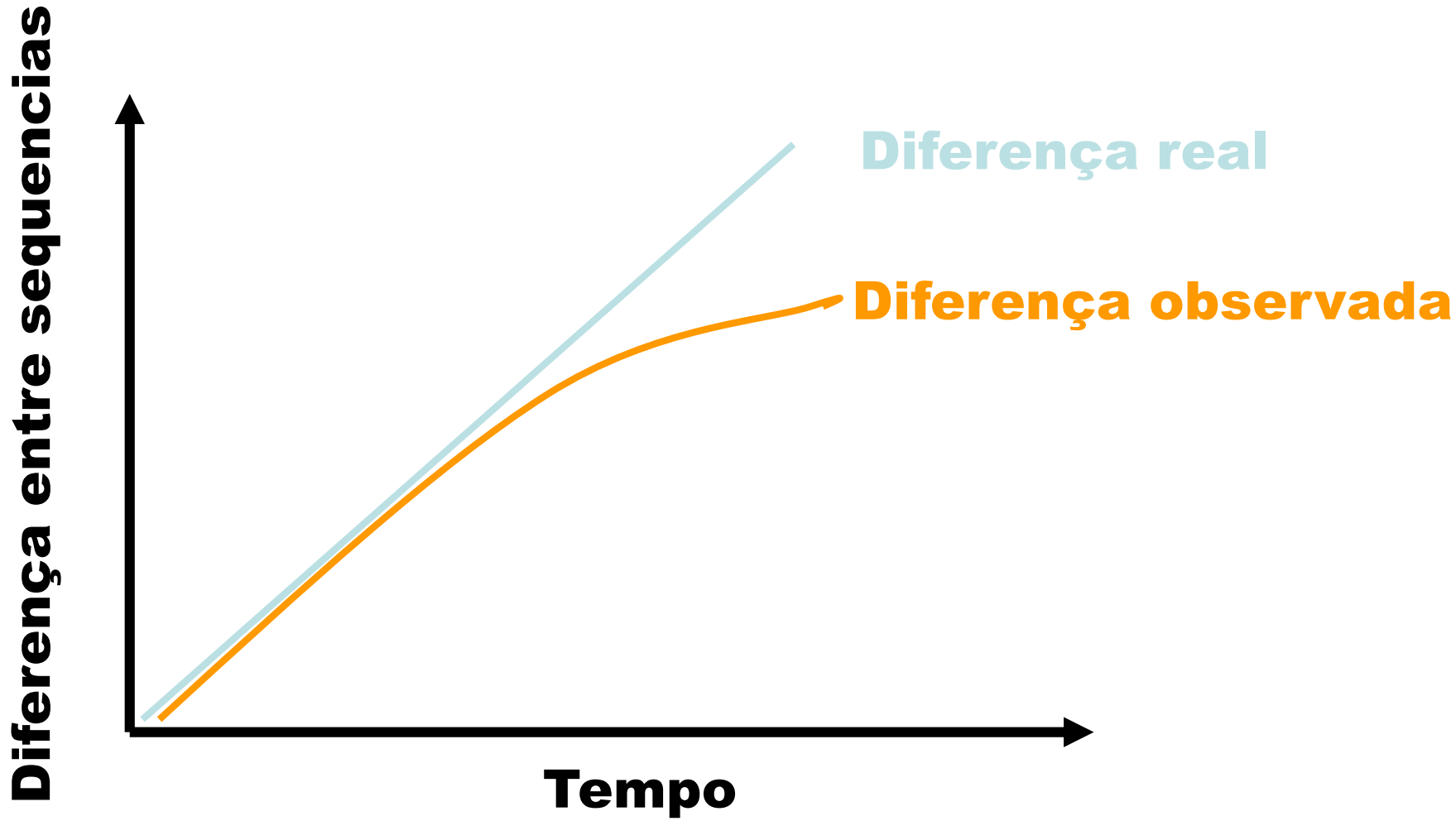


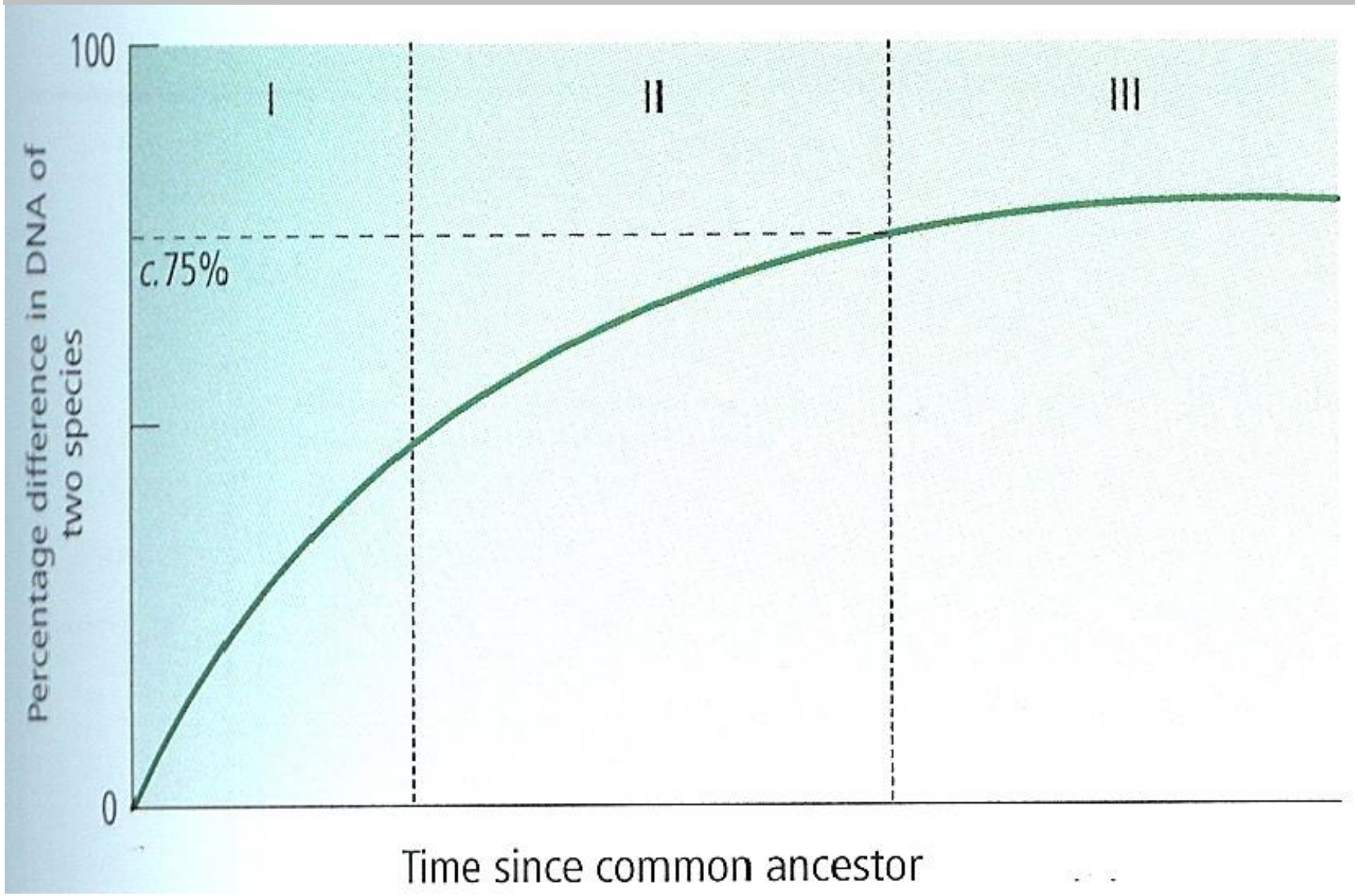
Figure 13 The gamma distribution for four different values of the shape parameter (α). When α is small, most of the sites evolve very slowly, but a few sites have moderate-to-high rates. As α increases, the distribution becomes more peaked and symmetrical about a mean rate of 1.0. When α is infinity, all sites have relative rate 1.0, so that an equal-rates model can be obtained as a special case of the gamma model.

Modelos como forma de correção das distancias



Depende da escala de tempo e da taxa de mutação do gene

Saturação


100
0
I
II
III
c.75%
Percentage difference in DNA of two species
Time since common ancestor

Que modelo escolher?

Nested models:

Frequencias de bases iguais

Transições iguais a tranversões

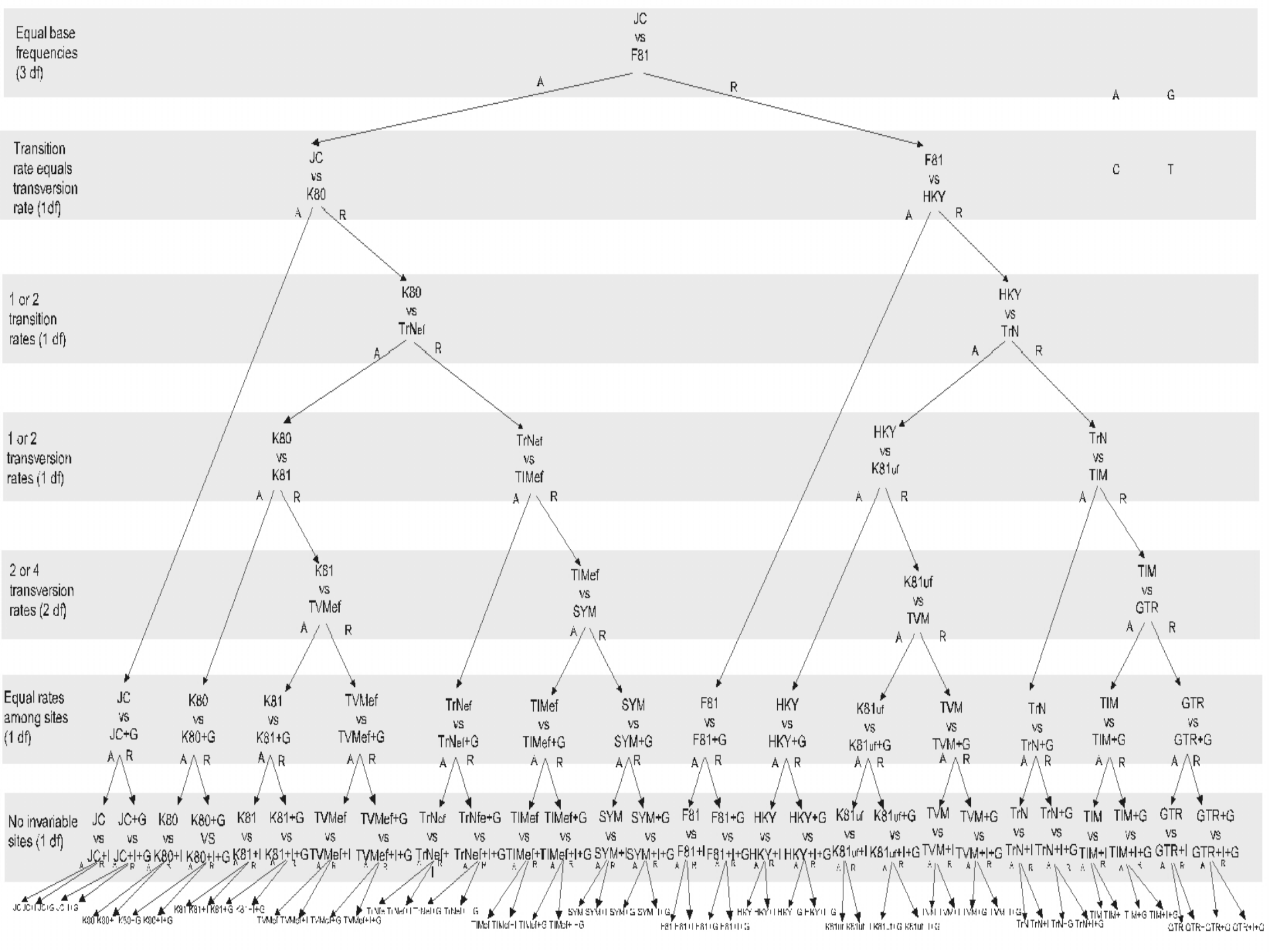
1 ou 2 taxas de transição

1 ou 2 taxas de tranversão

2 ou 4 taxas de tranversão

Taxa iguais entre sites

Numero de sites invariáveis


Equal base frequencies (3 df)
Transition rate equals transversion rate (1df)
1 or 2 transition rates (1 df)
1 or 2 transversion rates (1 df)
2 or 4 transversion rates (2 df)
Equal rates among sites (1 df)
No invariable sites (1 df)
JC vs F81
JC vs K80
F81 vs HKY
K80 vs TrNef
HKY vs TrN
K80 vs K81
TrNef vs TIMef
HKY vs K81uf
TrN vs TIM
K81 vs TVMef
TIMef vs SYM
K81uf vs TVM
TIM vs GTR
JC vs JC+G
K80 vs K80+G
K81 vs K81+G
TVMef vs TVMef+G
TrNef vs TrNef+G
TIMef vs TIMef+G
SYM vs SYM+G
F81 vs F81+G
HKY vs HKY+G
K81uf vs K81uf+G
TVM vs TVM+G
TrN vs TrN+G
TIM vs TIM+G
GTR vs GTR+G
A
R
C
T
G

Modeltest

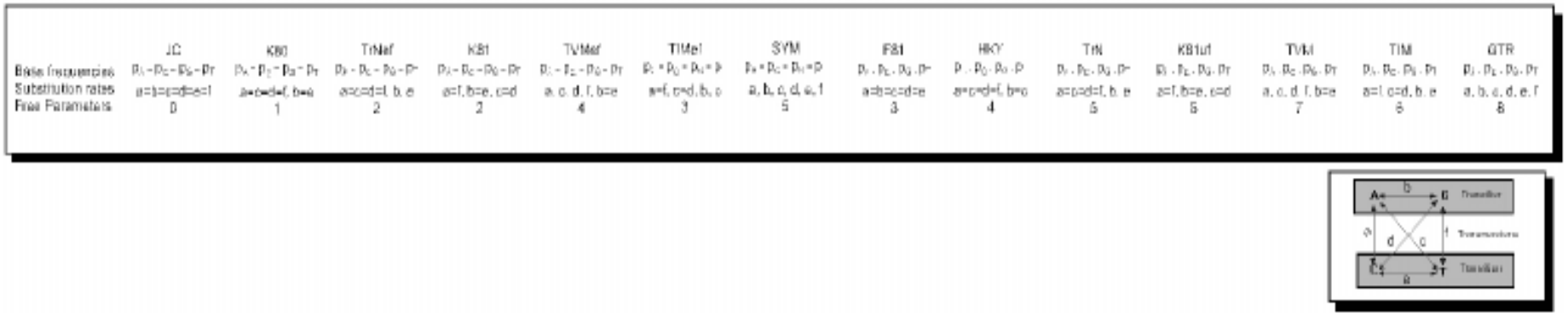


Figure 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes, Cantor, 1969), K80 (Kimura, 1980), TrNef (TrN equal base frequencies; see below), K81 (Kimura, 1981), TIMef (TIM with equal base frequencies), TIV (TIV with equal base frequencies), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa et al., 1985), TrN (Tamura, Nei, 1993), K81uf (K81 unequal base frequencies; see above), TIM, TIV, and GTR (Tavaré, 1986). G: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom.

Diversity of model selection methods

LRT

AIC

Bayesian Information Criterion

Decision theory

Likelihood ration test

Likelihood L=Pr(D/H)

The likelihood ratio test statistic is:

$$\Lambda = \frac{\max\left[L0\,(\text{Null Model} \mid \text{Data})\right]}{\max\left[L1\,(\text{Alternative Model} \mid \text{Data})\right]}$$

$$-\log \Lambda = \log L1 - \log L0$$

$$-2\log \Lambda = 2\delta = 2(\log Lgeneral - \log Lnested)$$

χ2 distribution with degrees of freedom equal to the difference in the number of free parameters in the two models

Akaike Information Criterion

Não tem de ser nested

AIC=-2log *L* +2n

n é o número de parâmetros independentes

Escolhe-se os modelos com valores mais baixos

Diversity of model selection methods

8.2 Akaike Information Criterion

The Akaike information criterion (AIC, (Akaike 1974) is an asymptotically unbiased estimator of the Kullback-Leibler information quantity (Kullback and Leibler 1951). We can think of the AIC as the amount of information lost when we use a specific model to approximate the real process of molecular evolution. Therefore, the model with the smallest AIC is preferred. The AIC is computed as:

$$AIC = -2\ell + 2K \ ,$$

where ℓ is the maximum log-likelihood value of the data under this model and K_i is the number of free parameters in the model, including branch lengths if they were estimated *de novo*. When sample size (n) is small compared to the number of parameters (say, $n/K < 40$) the use of a second-order AIC, AIC_c (Sugiura 1978; Hurvich and Tsai 1989), is recommended:

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1},$$

The AIC compares several candidate models simultaneously, it can be used to compare both nested and non-nested models, and model-selection uncertainty can be easily quantified using the AIC differences and Akaike weights (see Model uncertainty below). Burnham and Anderson (2003) provide an excellent introduction to the AIC and model selection in general.

Diversity of model selection methods

8.3 Bayesian Information Criterion

An alternative to the use of the AIC is the Bayesian Information Criterion (BIC) (Schwarz 1978):

$$BIC = -2\ell + K \log n$$

Given equal priors for all competing models, choosing the model with the smallest BIC is equivalent to selecting the model with the maximum posterior probability. Alternatively, Bayes factors for models of molecular evolution can be calculated using reversible jump MCMC (Huelsenbeck, Larget, and Alfaro 2004). We can easily use the BIC instead of the AIC to calculate BIC differences or BIC weights.

modelos

(TrN+G+I)
G-A transition rate of 4.63957
C-T rate of 12.0532,
the proportion of invariable sites of zero
and the shape parameter of the gamma distribution was 0.1394.

The Akaike Information Criterion:
selected the GTR+G model
A-C transversion rate of 3.8697
G-A transition rates of 10.8879,
A-T rate of 3.5091
C-G rate of 0.1615
C-T rate of 32.0765,
the proportion of invariable sites of zero
shape parameter of the gamma distribution was 0.1469).