1º Simpósio LAVIREO

Modelos Evolutivos Teoria e Prática

Edson Delatorre

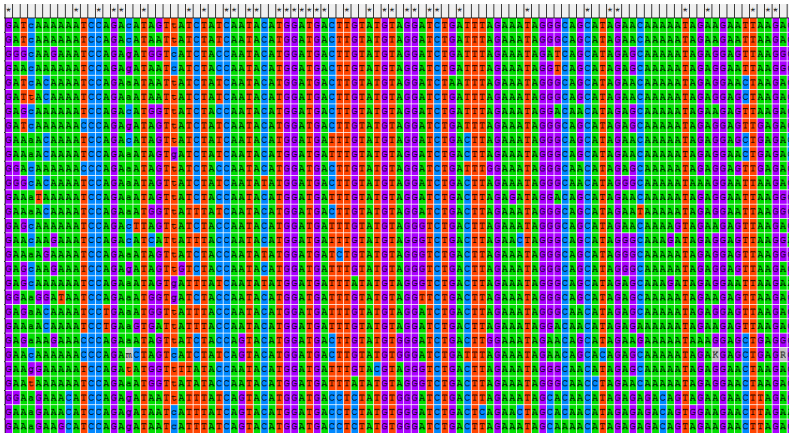
Lab. de Genética Molecular de Microrganismos

Instituto Oswaldo Cruz/FIOCRUZ

delatorre.ioc@gmail.com

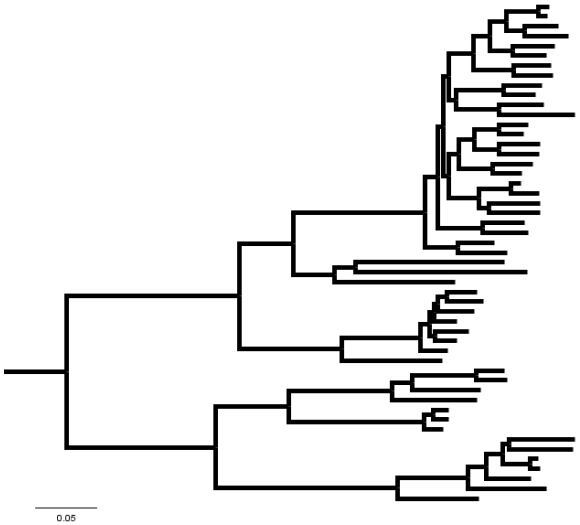
Rio de Janeiro – RJ, janeiro de 2019

Modelando a evolução



Alinhamento

Modelo



Árvore filogenética

Sequência ancestral

AACCTGTGCA

AATCTGTGTA
*** ***

ATCCTGGGTT
*** * ****

Seq1 **AATCTGTGTA**
Seq2 **ATCCTGGGTT**
**** * ***

p-distance:

número de diferenças nucleotídicas por sítio

```
Seq1    AATCTGTGTA
Seq2    ATCCTGGGTT
          **   *  *
```

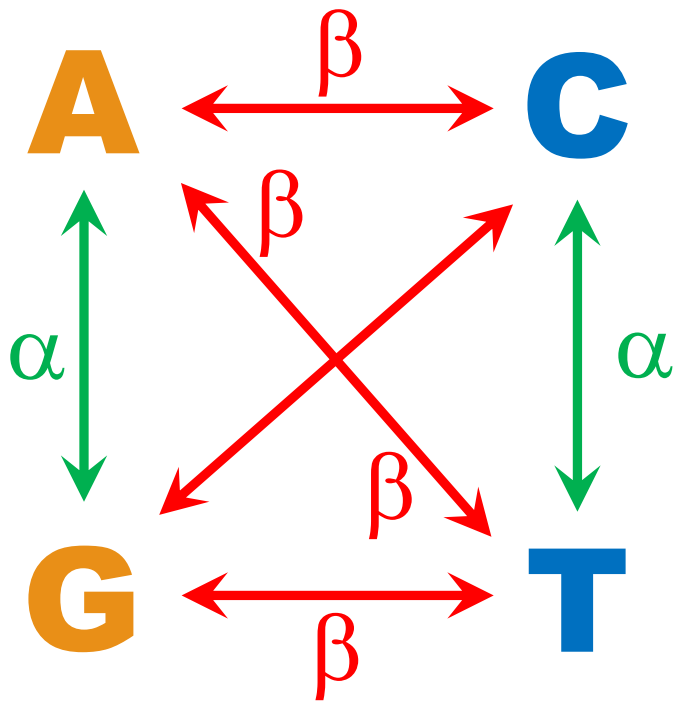
p-distance = 0.4

Normalmente subestima a "**distância verdadeira**":

d - distância genética

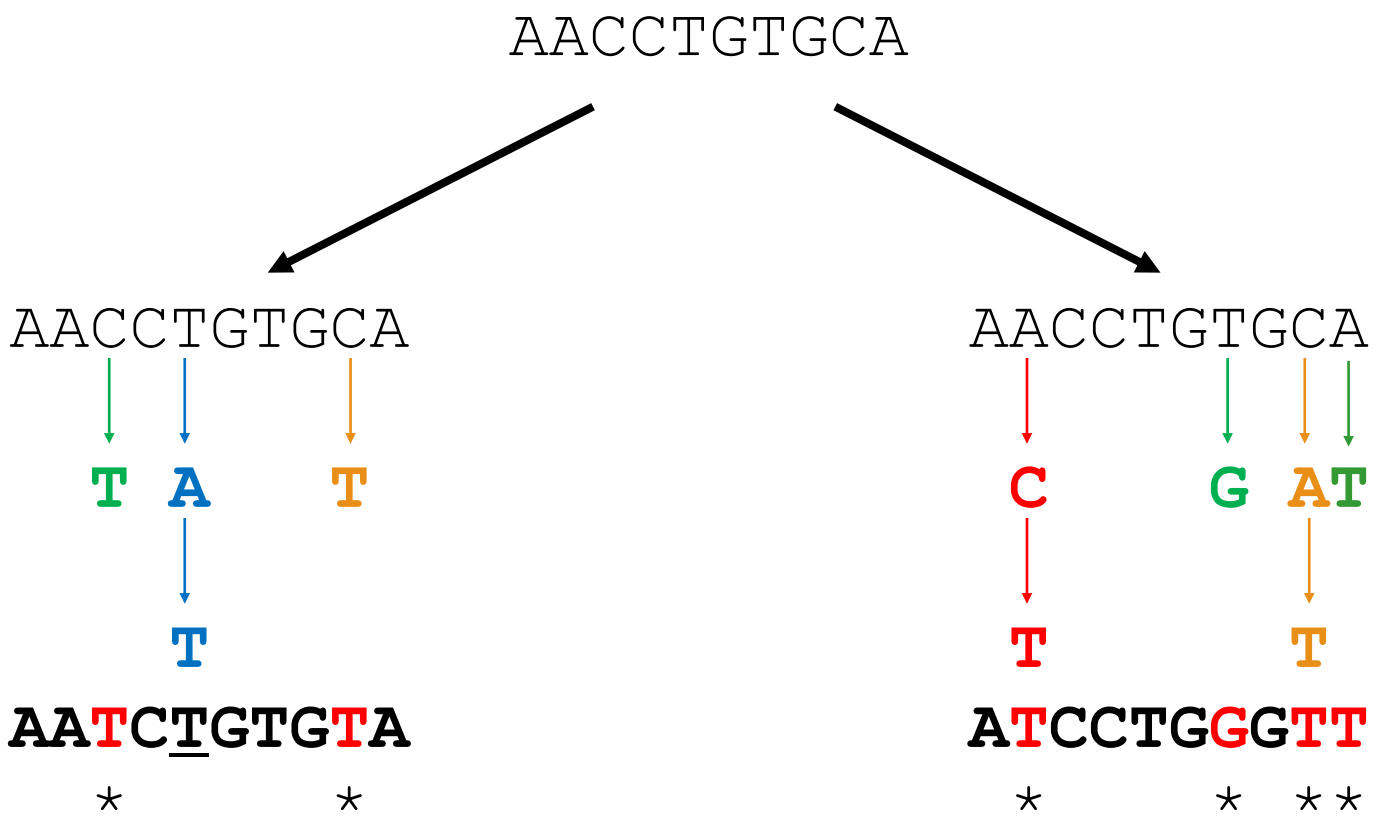
Transições (α): Pu $\leftrightarrow$ Pu (A,G) ou Py $\leftrightarrow$ Py (C,T)

Transversões (β): Pu $\leftrightarrow$ Py

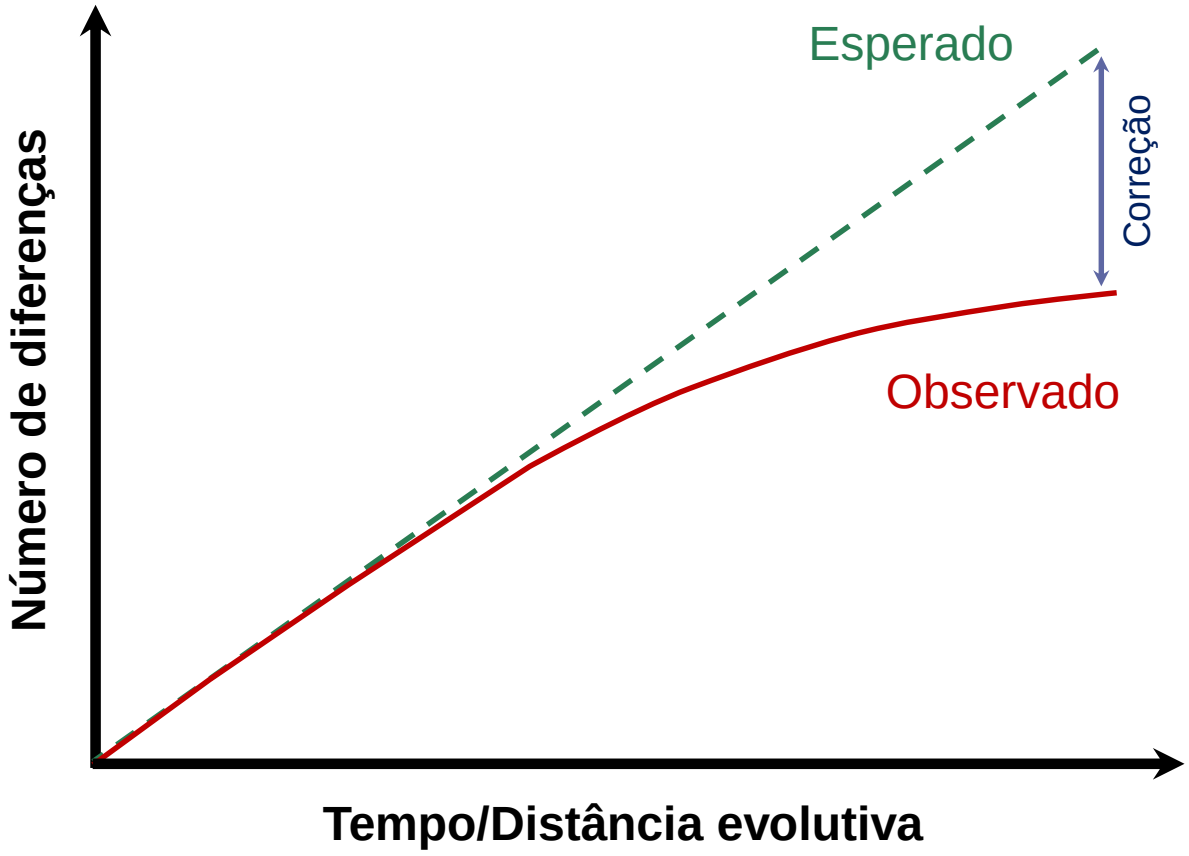


Transições são pelo menos 2 vezes mais frequentes que transversões

Sequência ancestral



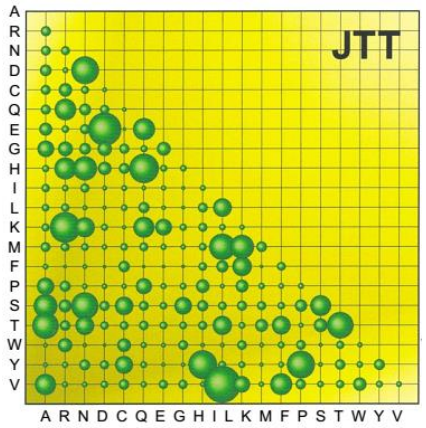
Modelos de substituição



Modelos evolutivos

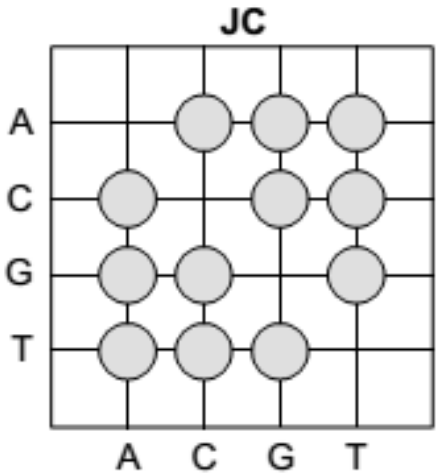
- Empíricos:

  Propriedades calculadas através de comparações de um grande número de sequências observadas.



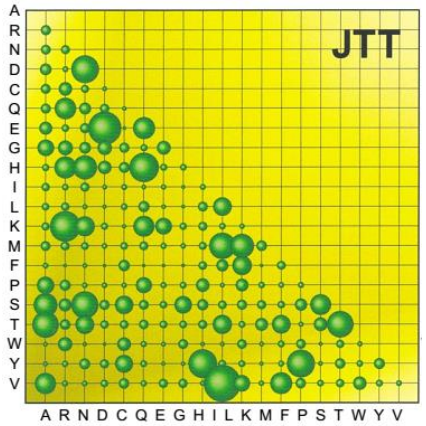
- Paramétricos:

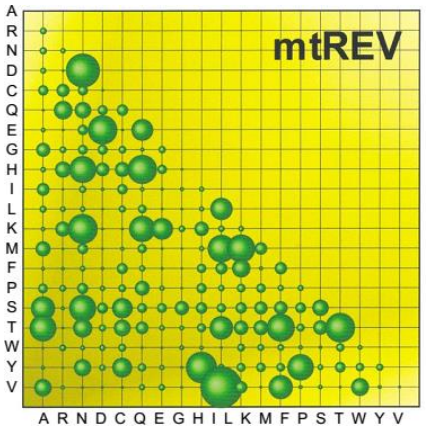
  Valores dos parâmetros são estimados a partir do dataset utilizado em cada análise particular



Modelos de substituição de aminoácidos

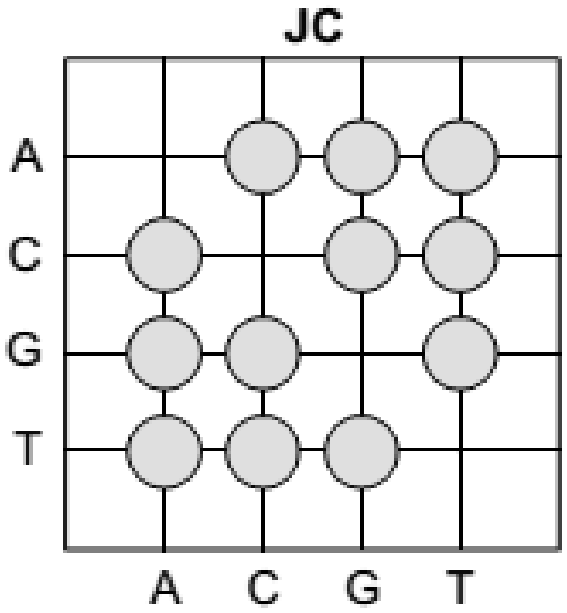
- Derivados da contagem simples de substituições de aminoácidos em grandes bancos de dados de sequências.
- Análise de sequências de proteínas intimamente relacionadas.
- Construção de matrizes específicas para grupos de proteínas.





Modelos de substituição de nucleotídeos

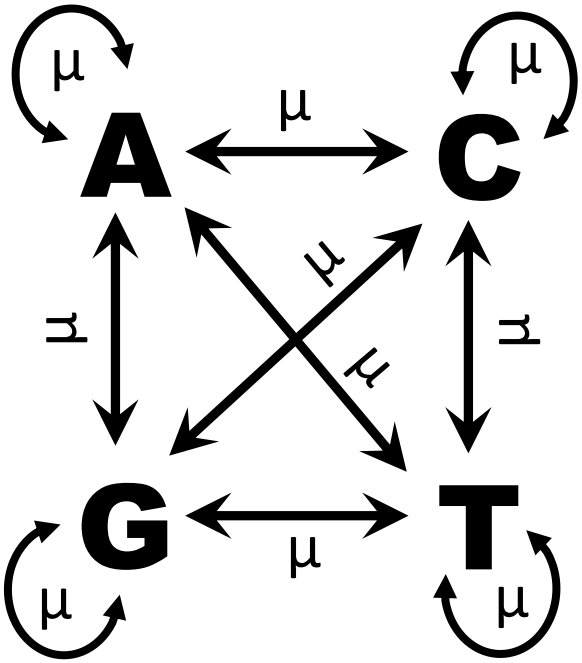
- Parâmetros
  - Frequência de bases
  - Taxa de substituição



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

Matriz

Modelo de Jukes e Cantor



Qualquer nucleotídeo *i* pode permanecer o mesmo

$i \longrightarrow i$ taxa = μ

ou mudar para qualquer dos outros nt (*j=3)*

$i \longrightarrow j$ taxa = 3·μ

Fórmula de Jukes e Cantor

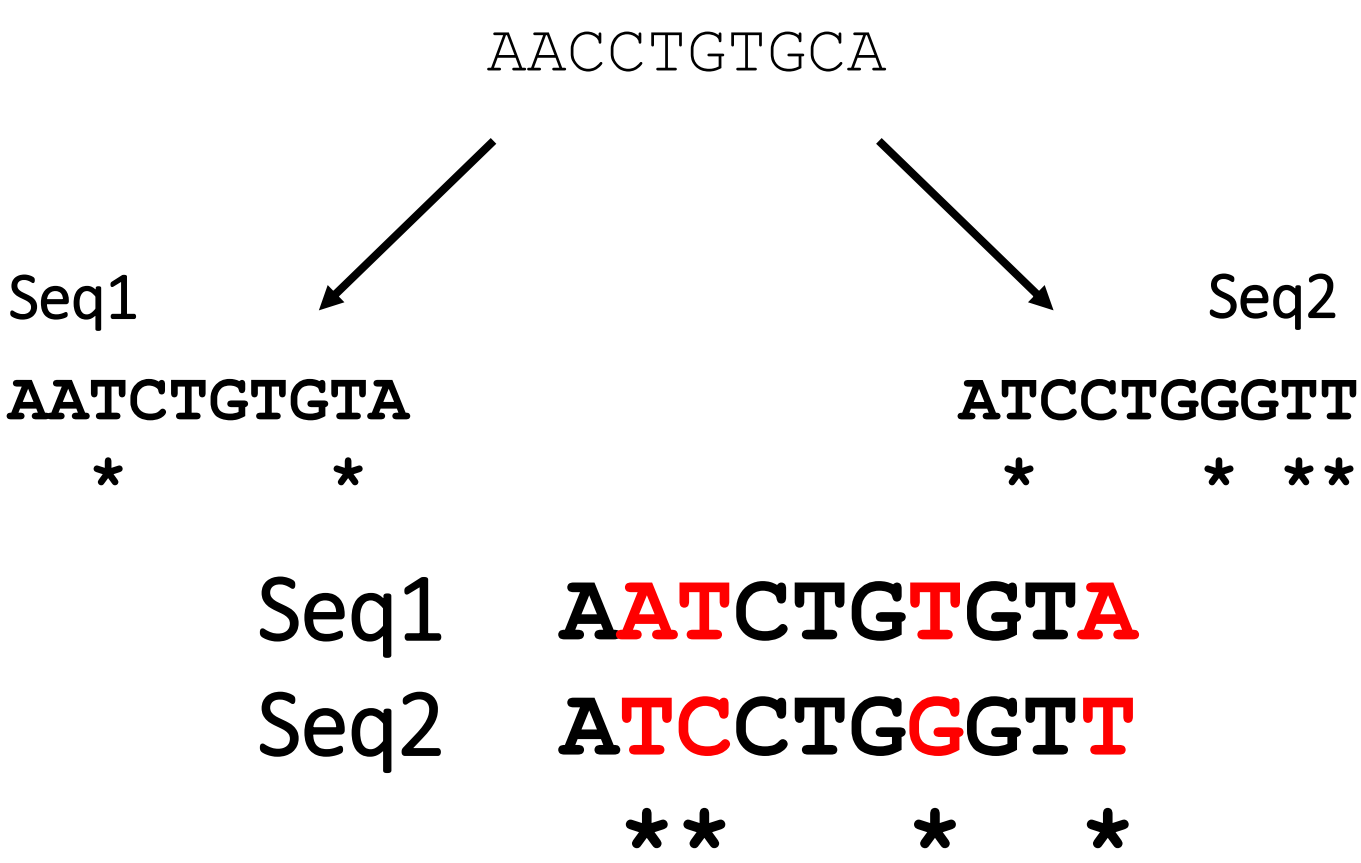
$$P = \frac{3}{4}\left(1 - e^{-4/3\mu t}\right)$$

Resolvendo por μ *t*

$$\mu t = -\frac{3}{4} \ln\left(1 - \frac{4}{3} P\right)$$

Distância genética estimada **(d)**

Distância observada *P*-distance



P-distance = **0.4**

d (JC) = - 3/4 *ln* [1- 4/3 (0.4)] = **0.5716**

Substitutições de nt como processos de Markov homogêneos

$$\boldsymbol{Q} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ a\mu\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ b\mu\pi_A & d\mu\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & f\mu\pi_T \\ c\mu\pi_A & e\mu\pi_C & f\mu\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix} \end{array}$$

Premissas

- As taxas de mudança da base *i* para a base *j* são independentes da base que ocupou o sítio *i* anteriormente.
- As taxas de substituição não mudam com o passar do tempo (homogeneidade).
- As frequências relativas de A, C, G e T (πA, πC, πG, πT) estão em equilíbrio (estacionariedade).

Substitutições de nt como processos de Markov homogêneos

$$\boldsymbol{Q} = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ a\mu\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ b\mu\pi_A & d\mu\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & f\mu\pi_T \\ c\mu\pi_A & e\mu\pi_C & f\mu\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix} \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$

As taxas de substituição do nucleotídeo $i \rightarrow j = j \rightarrow i$

Em geral: f = 1

a, b, c, d, e são estimados a partir dos dados por ML.

Matriz Q para o modelo Jukes e Cantor (JC)

Matriz de taxas = $\begin{pmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{pmatrix}$ ➔ $a=b=c=d=e=f=1$

$\pi_A = \pi_T = \pi_C = \pi_G = 0.25$

$$Q = \begin{pmatrix} -3/4\mu & 1/4\mu & 1/4\mu & 1/4\mu \\ 1/4\mu & -3/4\mu & 1/4\mu & 1/4\mu \\ 1/4\mu & 1/4\mu & -3/4\mu & 1/4\mu \\ 1/4\mu & 1/4\mu & 1/4\mu & -3/4\mu \end{pmatrix}$$

Modelos evolutivos

Modelo	Matrix de taxas	Parâmetros	Param. livres
K2P	$\begin{pmatrix} - & 1 & k & 1 \\ 1 & - & 1 & k \\ k & 1 & - & 1 \\ 1 & k & 1 & - \end{pmatrix}$	$\pi_A = \pi_C = \pi_G = \pi_T = 1/4$ Razão Ti/Tv ($b = e$ = k; $a = c = d = f$ = 1)	1
HKY85	$\begin{pmatrix} - & 1 & k & 1 \\ 1 & - & 1 & k \\ k & 1 & - & 1 \\ 1 & k & 1 & - \end{pmatrix}$	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$ RazãoTi/Tv ($b = e$ = k; $a = c = d = f$ = 1)	4
TN93	$\begin{pmatrix} - & 1 & b & 1 \\ 1 & - & 1 & e \\ b & 1 & - & 1 \\ 1 & e & 1 & - \end{pmatrix}$	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$ Razão Ti/Tv e razãoPu(Ti)/Py(Ti) ($b \neq e$; $a = c = d = f$ = 1)	5
GTR	$\begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & 1 \\ c & e & 1 & - \end{pmatrix}$	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$ ($a \neq$ b $\neq c \neq d \neq$ e; f = 1)	8

Heterogeneidade da taxa de evolução entre caracteres

- Todos os modelos assumem homogeneidade de taxas ao longo dos sítios

- Realidade:

  Nt em diferentes posições dentro dos códons possuem taxas diferentes (normalmente $3^{\circ} > 1^{\circ} > 2^{\circ}$)

  Regiões hipervariáveis de proteínas (hotspots)

Subestimação das distâncias genéticas

A distribuição Gama

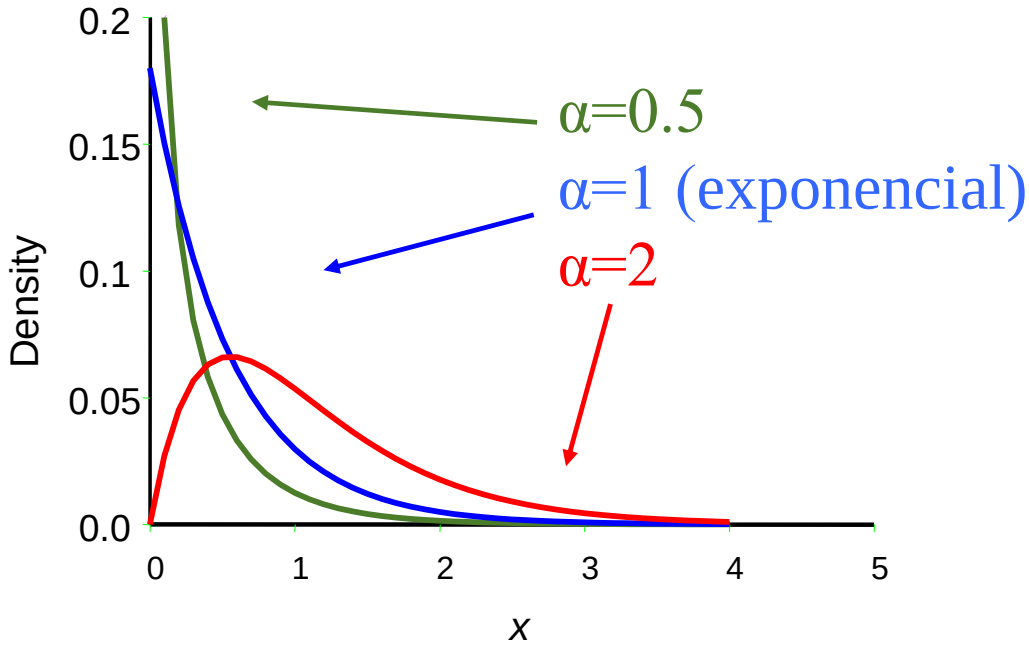
Função de densidade:

A forma da distribuição depende somente de um parâmetro *alpha*:

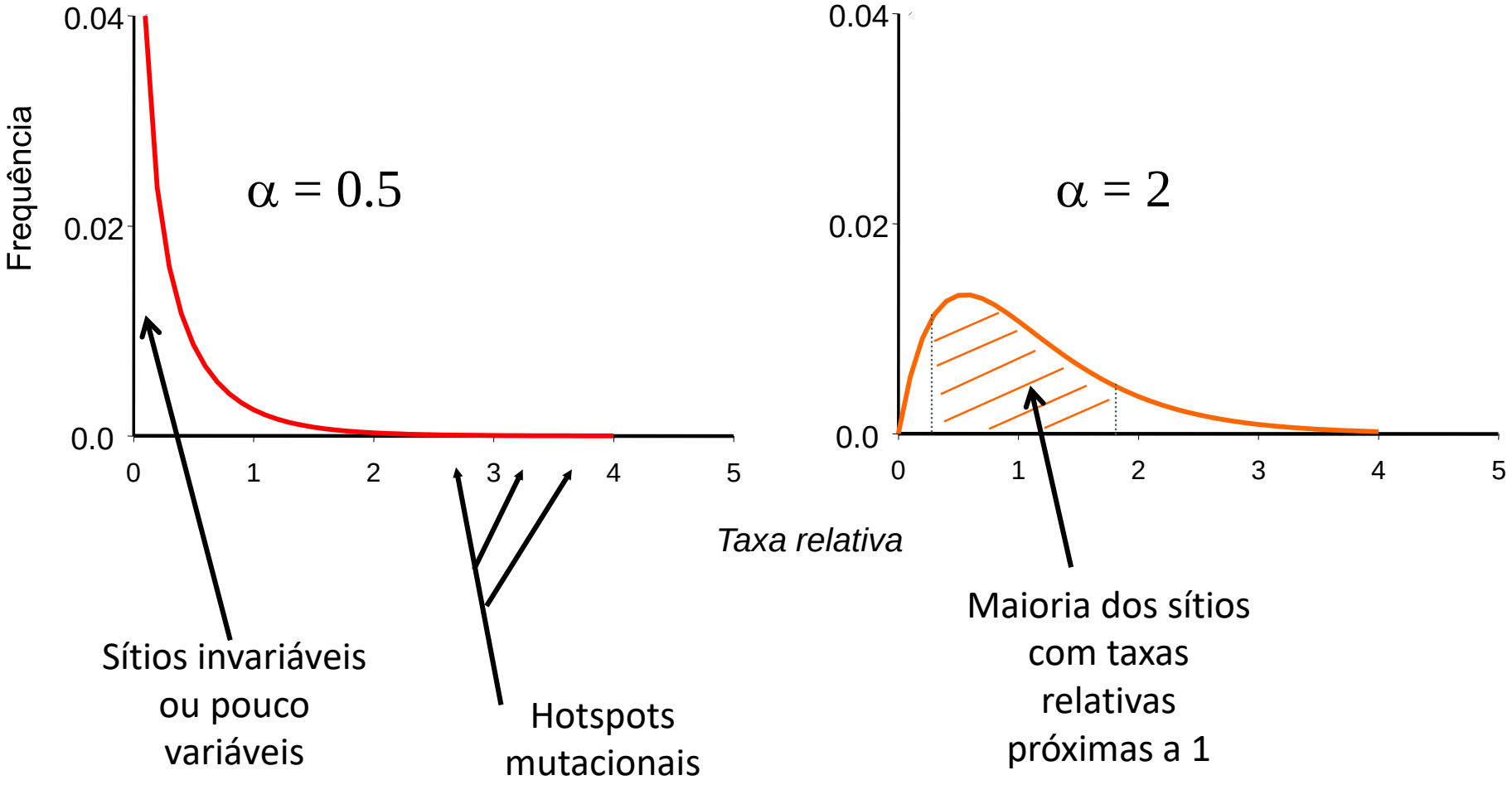
$\alpha < 1 \Rightarrow$ forma L

$\alpha > 1 \Rightarrow$ forma de sino (~Normal)

$\alpha \to \infty \Rightarrow$ um único valor no eixo x



Heterogeneidade de taxas atraves da distribuição Γ

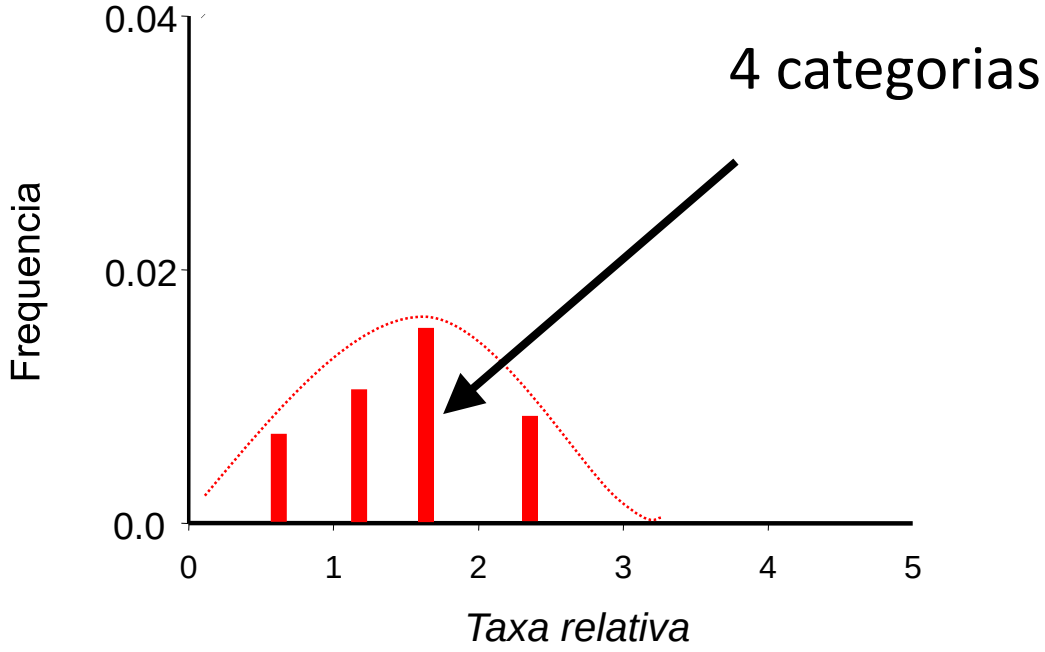


Modelo JC + Γ

$$d = \frac{3\alpha}{4}\left\{\left[1 - \frac{4}{3}p\right]^{-1/\alpha} - 1\right\}$$

A taxa de substituição varia de acordo com a distribuição Γ

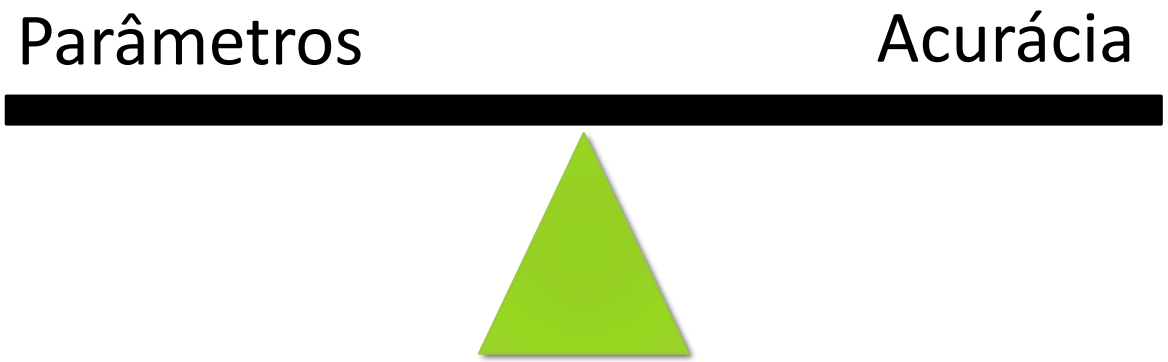
Distribuição Γ discreta



- Em geral, quanto maior as categorias, melhor a aproximação.
- 4-8 categorias são normalmente suficientes.

Como escolher o melhor modelo?

- Modelos evolutivos são sempre simplificações, muitas vezes com premissas para tornar um problema complexo em algo passível de ser computado.
- Melhor modelo é aquele que se ajusta aos dados e permite fazer predições acuradas.
- Em geral, quanto mais complexo o modelo, melhor o ajuste do mesmo aos dados.



Teste da razão de verossimilhança (LRT)

- Próprio para a comparação de hipóteses evolutivas aninhadas

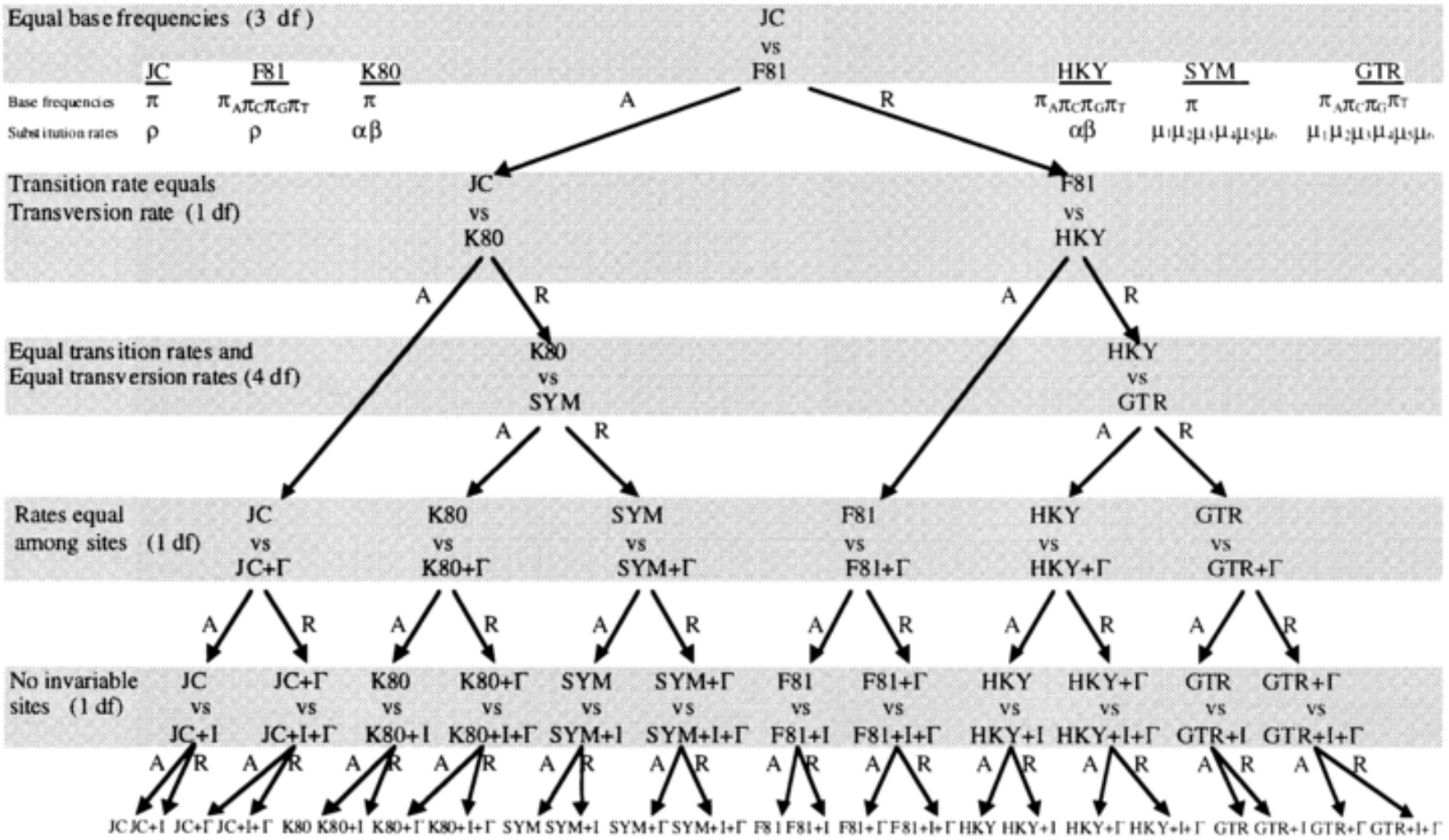
$$\Delta = 2 (\log_e L_1 - \log_e L_0)$$

L_1 – MV sob o modelo com mais parâmetros (complexo) ➔ hipótese alternativa (H_1)

L_0 – MV sob o modelo com menos parâmetros (simples) ➔hipótese nula (H_0)

Δ se distribui como X^2, com número de graus de liberdade igual a diferença entre os números de parâmetros livres entre os dois modelos.

Modelos


Equal base frequencies (3 df)
JC vs F81
JC
F81
K80
HKY
SYM
GTR
Base frequencies
Substitution rates
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

Teste da razão de verossimilhança (LRT)

$$\Delta = 2 (\log_e L_1 - \log_e L_0)$$

- Se LR é significativo ($p < 0,05$ ou $< 0,01$): a inclusão de parâmetros adicionais no modelo alternativo aumenta significativamente a verossimilhançaa dos dados.

- Quando Δ é próximo a zero ($p > 0,05$): a hipótese alternativa não se ajusta aos dados significativamente melhor do que a hipótese nula.

Seleção de modelos com jmodeltest

- Seleção de modelos baseada na LRT hierárquica.
- Akaike Information Criterion

$$AIC = -2 \ln L + 2k$$

- Corrected AIC

$$AICc = AIC + 2k(k+1)/(N-k-1)$$

- Bayesian Information Criterion

$$BIC = -2 \ln L + k \ln N$$

L – verossimilhança do modelo, k – número de parâmetros estimáveis, N – tamanho da amostra