1º Simpósio LAVIREO

# Máxima verossimilhança

**Edson Delatorre**

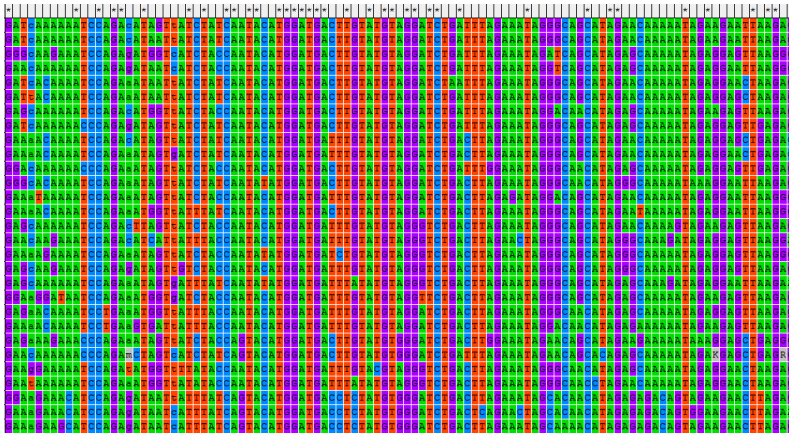
Lab. de Genética Molecular de Microrganismos

Instituto Oswaldo Cruz/FIOCRUZ

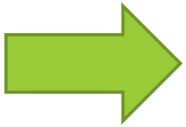
delatorre.ioc@gmail.com

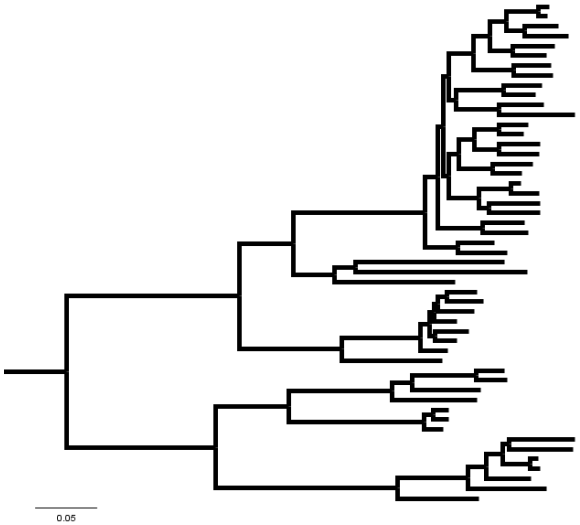
Rio de Janeiro – RJ, janeiro de 2019

# Modelando a evolução



Alinhamento






Árvore filogenética

# Métodos de Reconstrução de Filogenias

| | Métodos baseados em caracteres | Métodos baseados em distância |
|---|---|---|
| Métodos **baseados** em um modelo explícito de evolução | Máxima-verossimilhança<br>Inferência Bayesiana | Neighbor-Joining<br>Evolução mínima<br>UPGMA |
| Métodos **sem base** em um modelo explícito de evolução | Máxima parcimônia | |

# Métodos baseados em distância

O maior problema com os métodos baseados em distâncias genéticas é que eles funcionam com medidas de similaridade geral.

Problemas em situações com homoplasia

**Caracteres semelhantes que resultam de evolução independente**

# Estimativa básica da MV

- Encontrar a probabilidade dos **dados** (D) dado uma **hipótese** (H).

- Exemplo:

Estimativa do número de indivíduos parasitados em uma população:

Numa amostra de n= 10 indivíduos k= 4 são parasitadas → **D**

Vamos chamar **p** a probabilidade de ser parasitado → **H**

# Estimativa básica da MV

Probabilidade de os dados fornecidos diferentes valores de p (0.2, 0.4...)


Função de verossimilhança de *p* dado os dados k e n

$$\mathbf{pMV = 0.4}$$

# MV na reconstrução filogenética

- Encontrar a probabilidade dos **dados** (D) dado uma **hipótese** (H).
- Em filogenia molecular:

**Dados** = alinhamento de sequências

**Hipótese** = Topologia da árvore (τ)
Comprimento dos ramos (υ)
Modelo evolutivo (Φ)

$$Pr(D/H) = Pr(alinhamento/\tau, \upsilon, \Phi)$$

# MV na reconstrução filogenética

- Algumas hipóteses vão explicar os dados observados com maior probabilidade que outras.

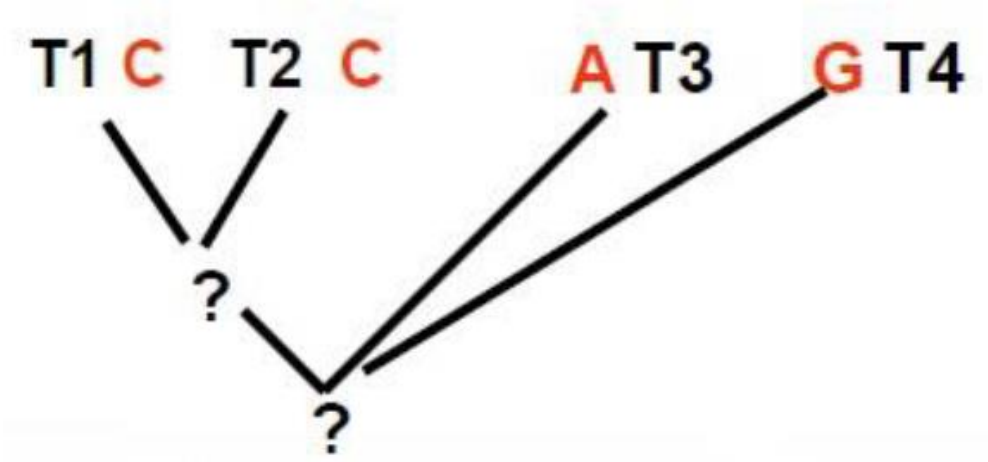
- Hipótese de Máxima Verossimilhança (MV): conjunto de valores de parâmetros que dá a maior probabilidade aos dados observados.

- A principal ideia por trás da inferência filogenética usando o critério de máxima verossimilhança é encontrar a topologia de árvore, o comprimento dos ramos e os parâmetros do modelo evolutivo que **maximizam** a **probabilidade** de observar o alinhamento de sequências que temos.

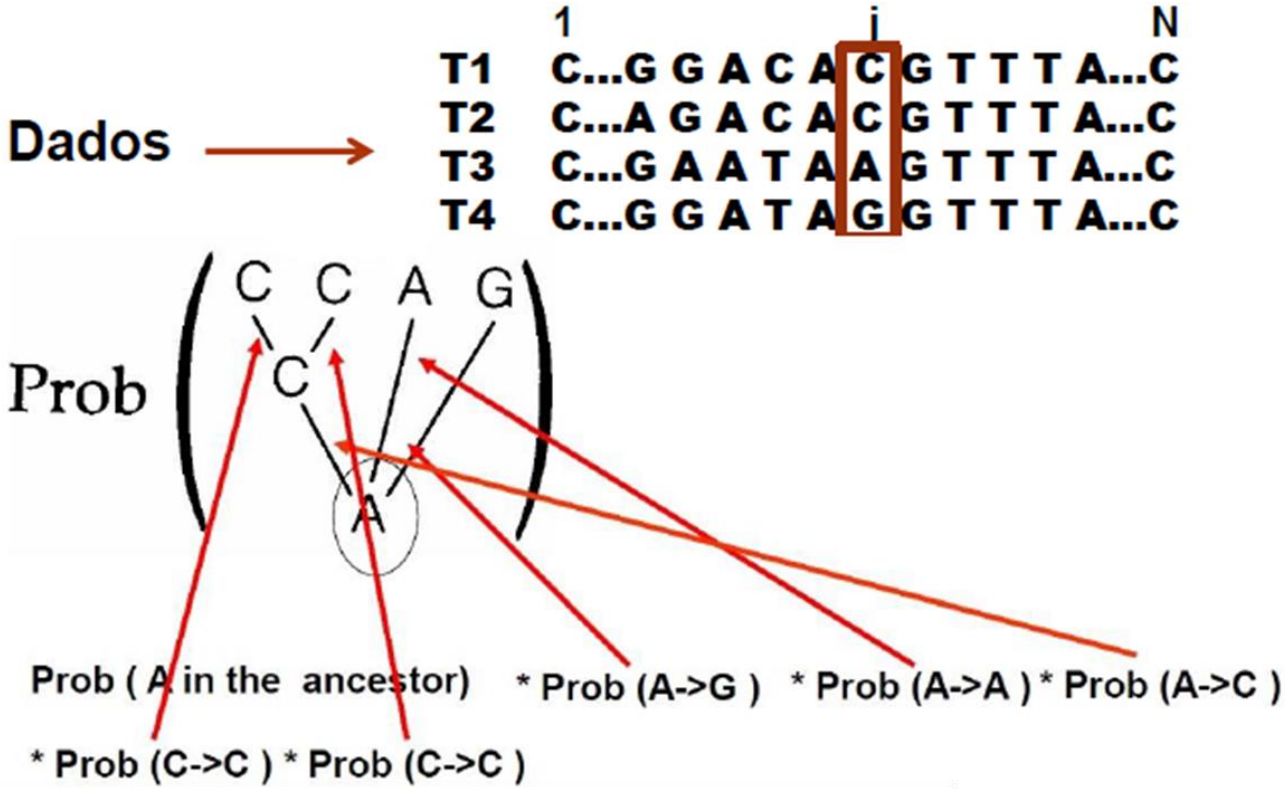
# Algoritmo para calcular a MV de uma árvore

| | 1 | | | | | j | | | | | N |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T1 | C | ... | G | G | A | C | A | C | G | T | T | T | A | ... | C |
| T2 | C | ... | A | G | A | C | A | C | G | T | T | T | A | ... | C |
| T3 | C | ... | G | A | A | T | A | A | G | T | T | T | A | ... | C |
| T4 | C | ... | G | G | A | T | A | G | G | T | T | T | A | ... | C |

Cada sítio evolui independentemente



# Algoritmo para calcular a MV de uma árvore



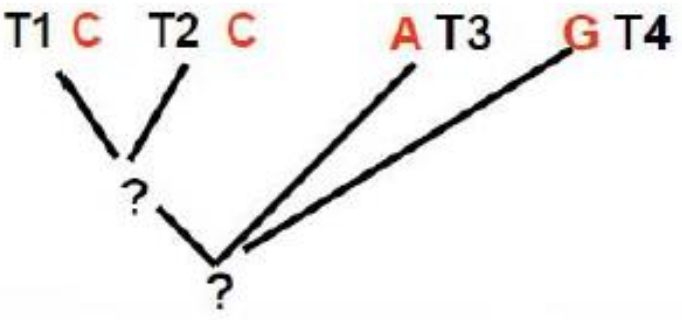

De onde vem estas probabilidades:

Ancestral: Freq nt no alinhamento

Mudanças: Modelo de substituição

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

# Algoritmo para calcular a MV de uma árvore



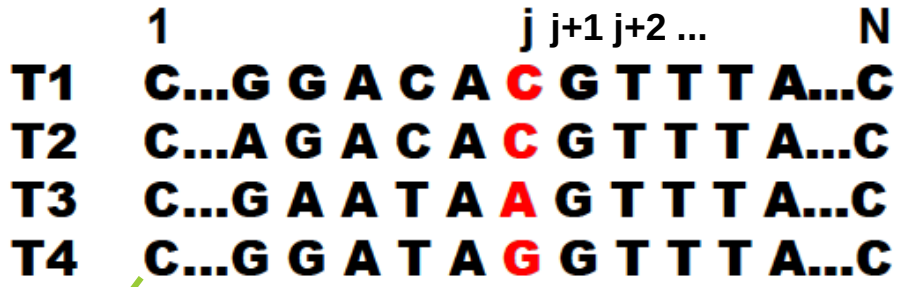

$$L_{(j)} = \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & A & & \\ & & A & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & C & & \\ & & A & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & T & & \\ & & A & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & G & & \\ & & A & \end{array}\right)$$

$$+\text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & A & & \\ & & C & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & C & & \\ & & C & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & T & & \\ & & C & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & G & & \\ & & C & \end{array}\right)$$

$$+\text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & A & & \\ & & \mathbf{T} & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & C & & \\ & & \mathbf{T} & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & T & & \\ & & \mathbf{T} & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & G & & \\ & & \mathbf{T} & \end{array}\right)$$

$$+\text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & A & & \\ & & G & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & C & & \\ & & G & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & T & & \\ & & G & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & G & & \\ & & G & \end{array}\right)$$

# Algoritmo para calcular a MV de uma árvore




$$L = L_{(1)} * L_{(2)} \ldots . L_{(j)} * L_{(j+1)} * \ldots . L_{(N)} = \prod_{i=1}^{N} L_{(i)}$$

Cada árvore tem n-1 nós internos, dos quais cada um pode ter um de quatro estados nucleotídeos

$$\mathbf{4^{n-1}}$$

# Algoritmo para calcular a MV de uma árvore

- Resumo

1. Escolha um modelo evolutivo.
2. Com base no modelo evolutivo, calcular as probabilidades por coluna.
3. Calcular a Verossimilhança da Árvore multiplicando as probabilidades para cada posição.

# Algoritmo para calcular a MV de uma árvore

- Temos um método computacionalmente viável para avaliar a probabilidade de uma determinada árvore, mas isso nos deixa com a tarefa de encontrar a árvore de MV.

- Encontrar a árvore de ML exige uma pesquisa tanto de todas as topologias possíveis, como também de todos os possíveis conjuntos de comprimentos de ramos.

**Busca heurística**

# Busca heurística

Um dos métodos mais populares de busca heurística é o denominado ***Branch swapping (Troca de ramo)***:

1) Começamos com uma árvore T (NJ) e calculamos o valor de ML.

2) Introduzimos mudanças aleatórias na topologia de T para obter uma nova árvore T'.

3) Calculamos o valor de ML da nova árvore T':

4a) Se a MV da nova árvore T' > que a MV de T, então ficamos com T' e iniciamos uma nova rodada de perturbações da topologia.

4b) Caso contrário, ficamos com a árvore inicial T e começamos uma nova rodada de perturbações da topologia.

# Busca heurística

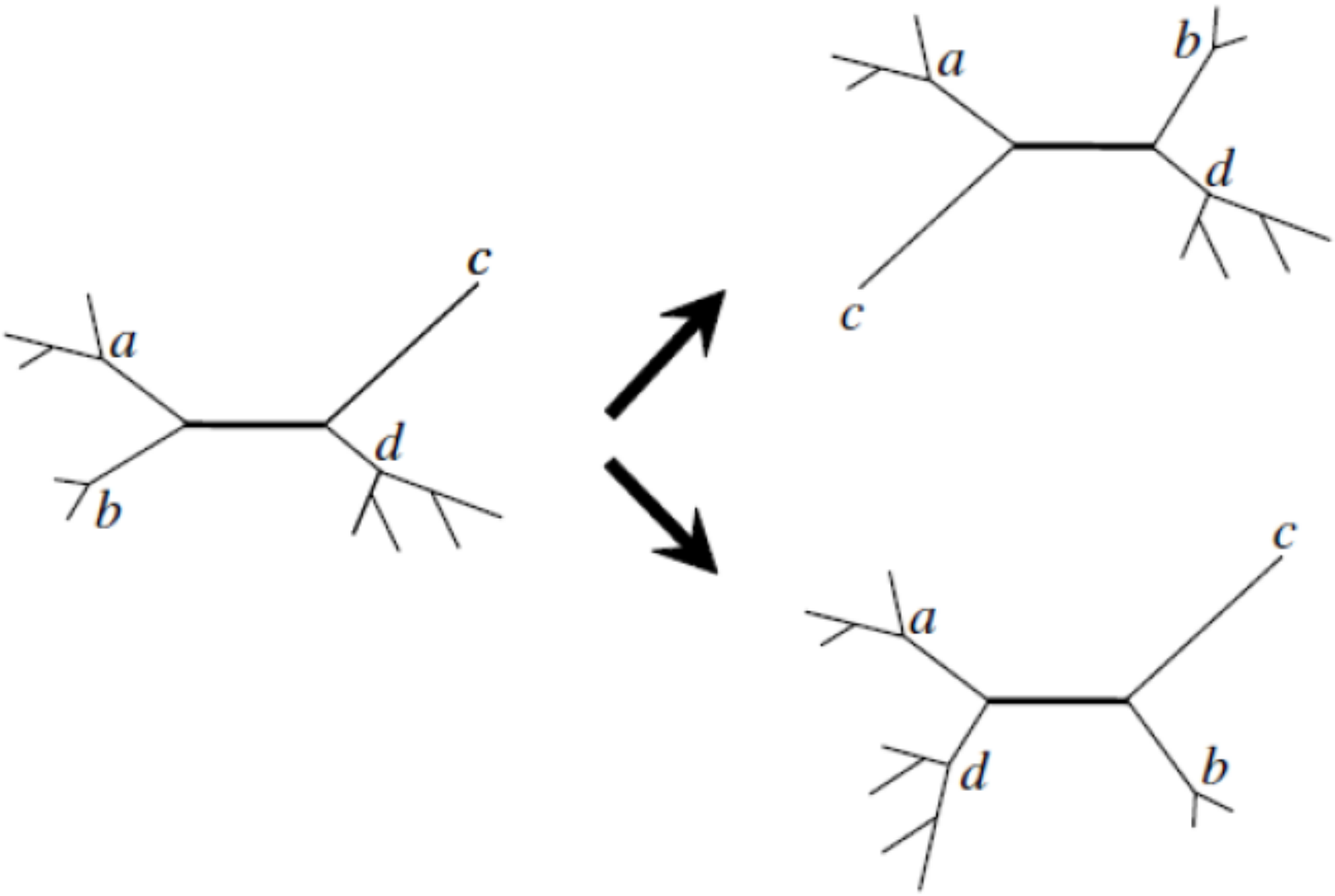
**Branch swapping (Troca de ramo):**

1) Nearest neighbor interchange (NNI)
(Troca entre vizinhos mais próximos): é o mais rápido.

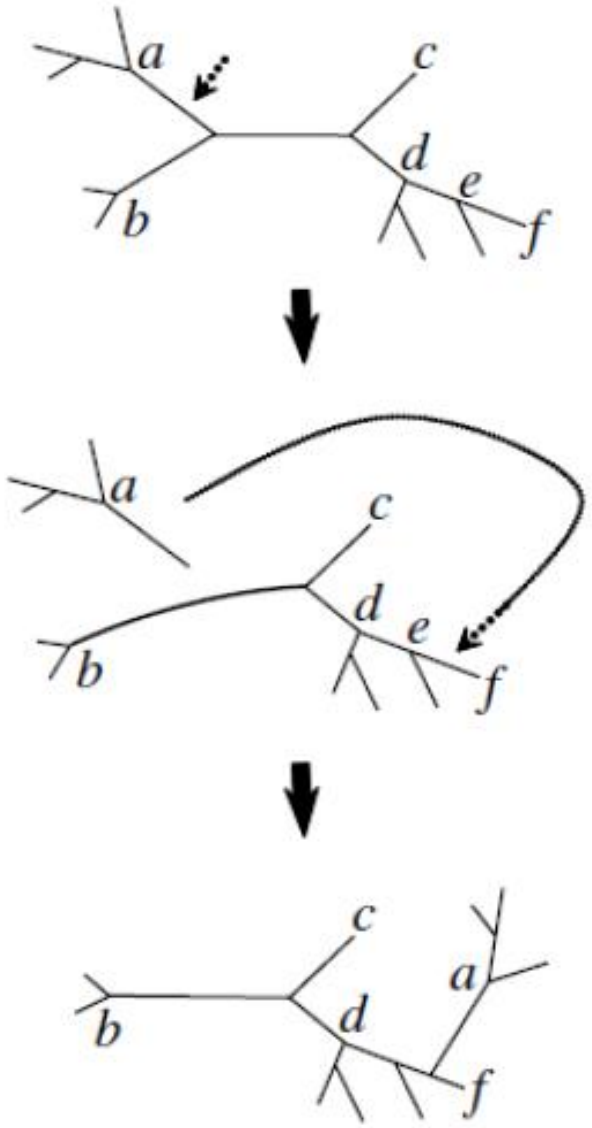
2) Subtree pruning and regrafting (SPR)
(Poda e enxerto de sub-árvores): é intermediário entre o NNI e o TBR.

3) Tree bisection and reconnection (TBR)
(Bissecção e reconexão de árvore) : é o melhor e mais completo, mas também o mais lento.

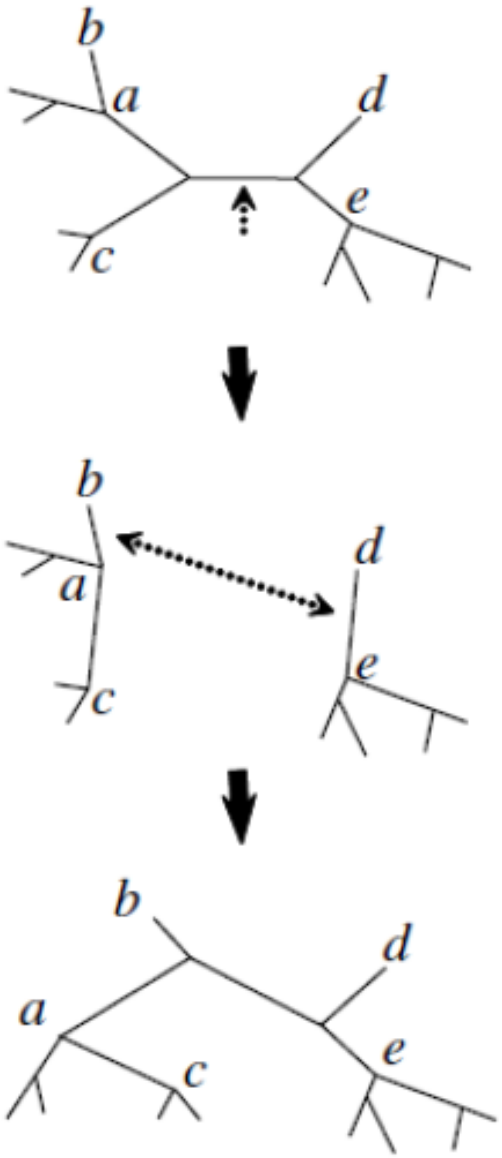
# Busca heurística: NNI (Troca entre vizinhos mais próximos)



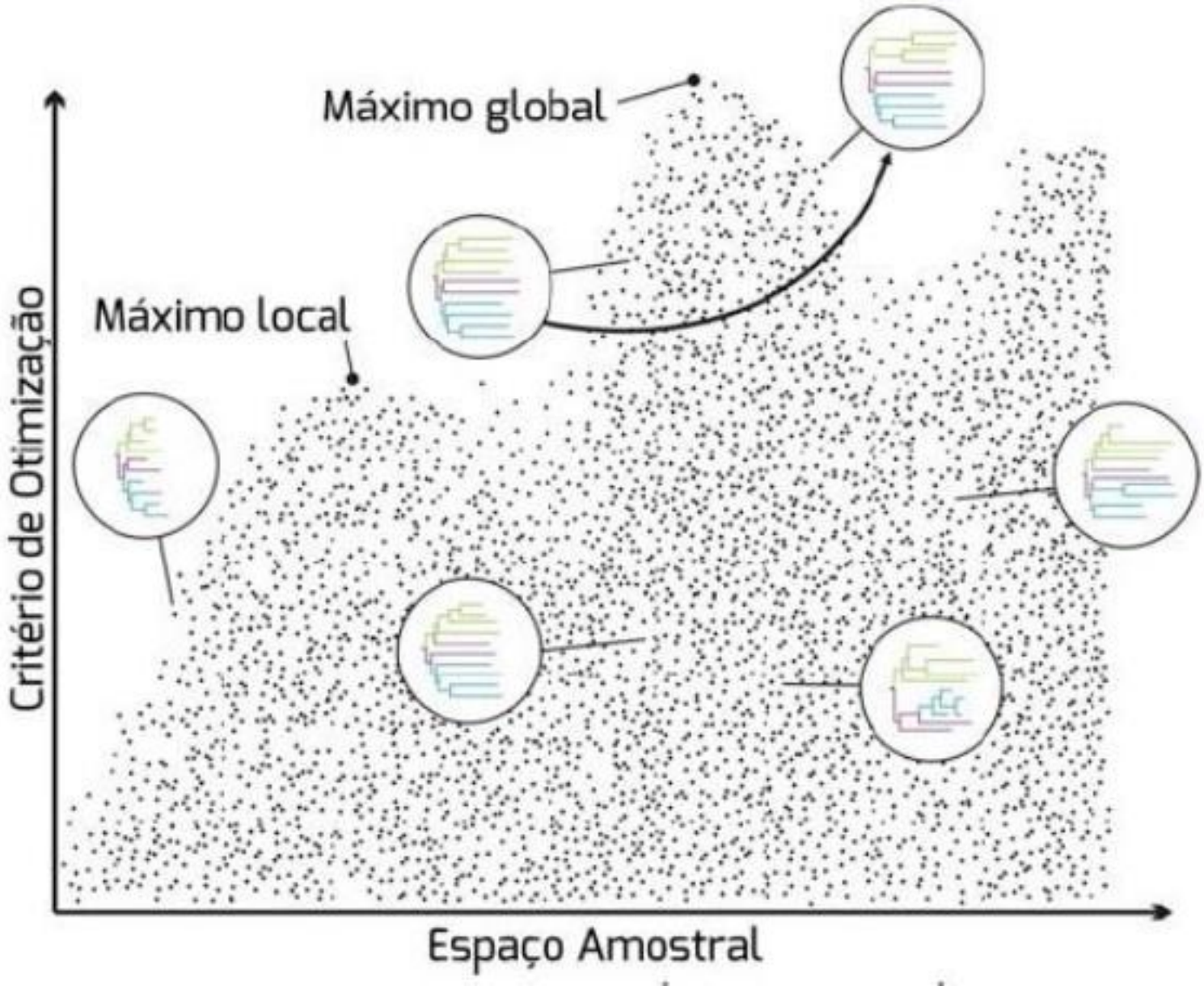
# Busca heurística: SPR (Pode e enxerto de sub-arvores)



# Busca heurística: TBR (Bissecção e reconexão de árvores)



# Espaço de árvores



# Análise de MV

Vantagens:

- permite a comparação de diferentes árvores (hipóteses) possíveis
- pode  ser utilizado para analisar datasets relativamente grandes (<5000 sequências)

Desvantagens:

- muito mais lento do que o NJ
- não temos qualquer garantia de encontrar a árvore de ML
- escolhe a partir de diferentes topologias de árvores a de maior verossimilhança e não sabemos quão diferente é essa topologia do resto das topologias pesquisadas.

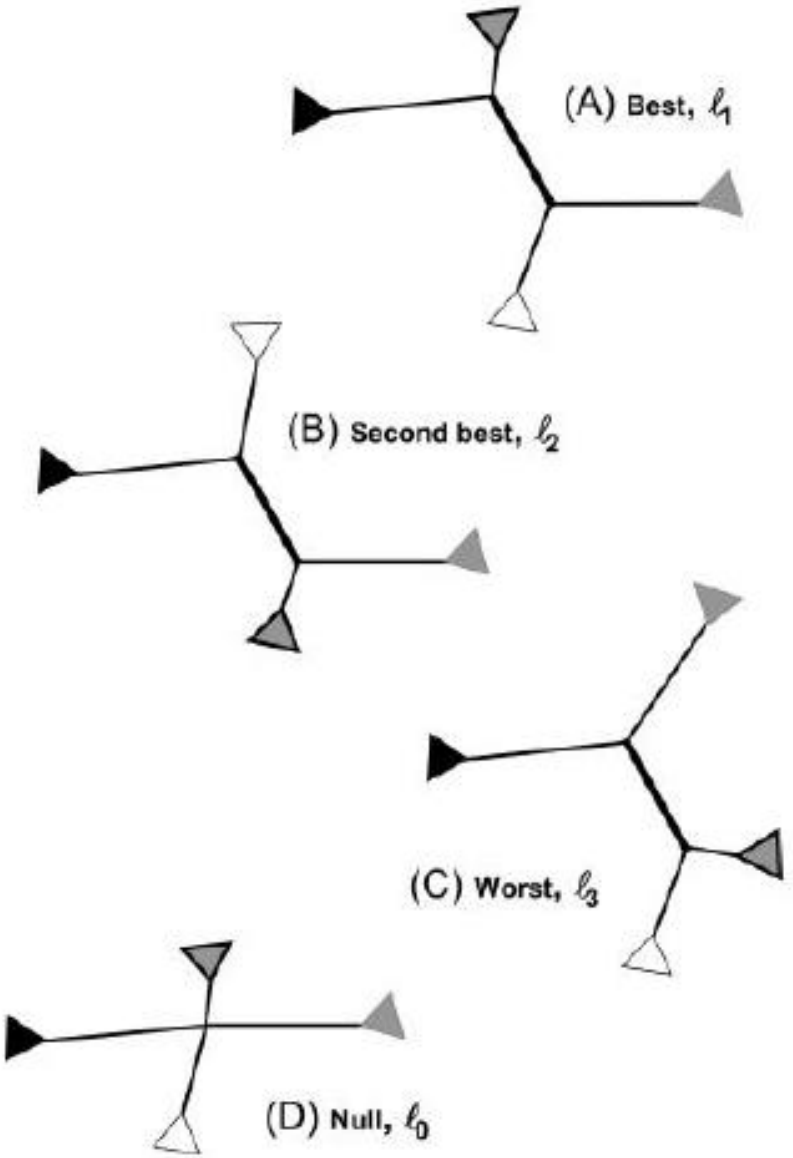
# Confiabilidade da topologia inferida por MV

## Teste de razão de verossimilhanças aproximado

(aLRT, Approximate Likelihood-Ratio Test)

- O aLRT é uma modificação do teste padrão de verossimilhança (LRT) o qual compara a hipótese alternativa de um comprimento de ramo positivo ($t \geq 0$) com a hipótese para um comprimento do ramo igual a zero ($t = 0$).
- No LRT aproximado (aLRT), a hipótese nula padrão "do ramo tem 0 comprimento" é aproximada pela hipótese mais geral "o ramo está incorrecto."
- Mais especificamente, o aLRT compara as verossimilhança do melhor e o do segundo melhor arranjo envolta do ramo de interesse.

# Teste de razão de verossimilhanças aproximado



# Teste de razão de verossimilhanças aproximado

- Valores elevados de aLRT (>0.90) é um indicativo de forte sinal filogenético nos dados em favor de um determinado cluster filogenético.

- A principal vantagem do aLRT é que ele é muito mais rápido do que o bootstrap ou a inferência Bayesiana.

- Em algumas circunstâncias, no entanto, o aLRT é um teste menos conservador (um cluster pode apresentar um valor de aLRT elevado, mesmo se estiver errado).