



Sequências de DNA na identificação de espécies e análise filogenética de microrganismos

Dra. Chirlei Glienke

Dra. Desirrê Petters-Vandresen

Modelos Evolutivos E Métodos Filogenéticos



ÁRVORE FILOGENÉTICA

ÁRVORE CERTA: NÃO EXISTE!!!

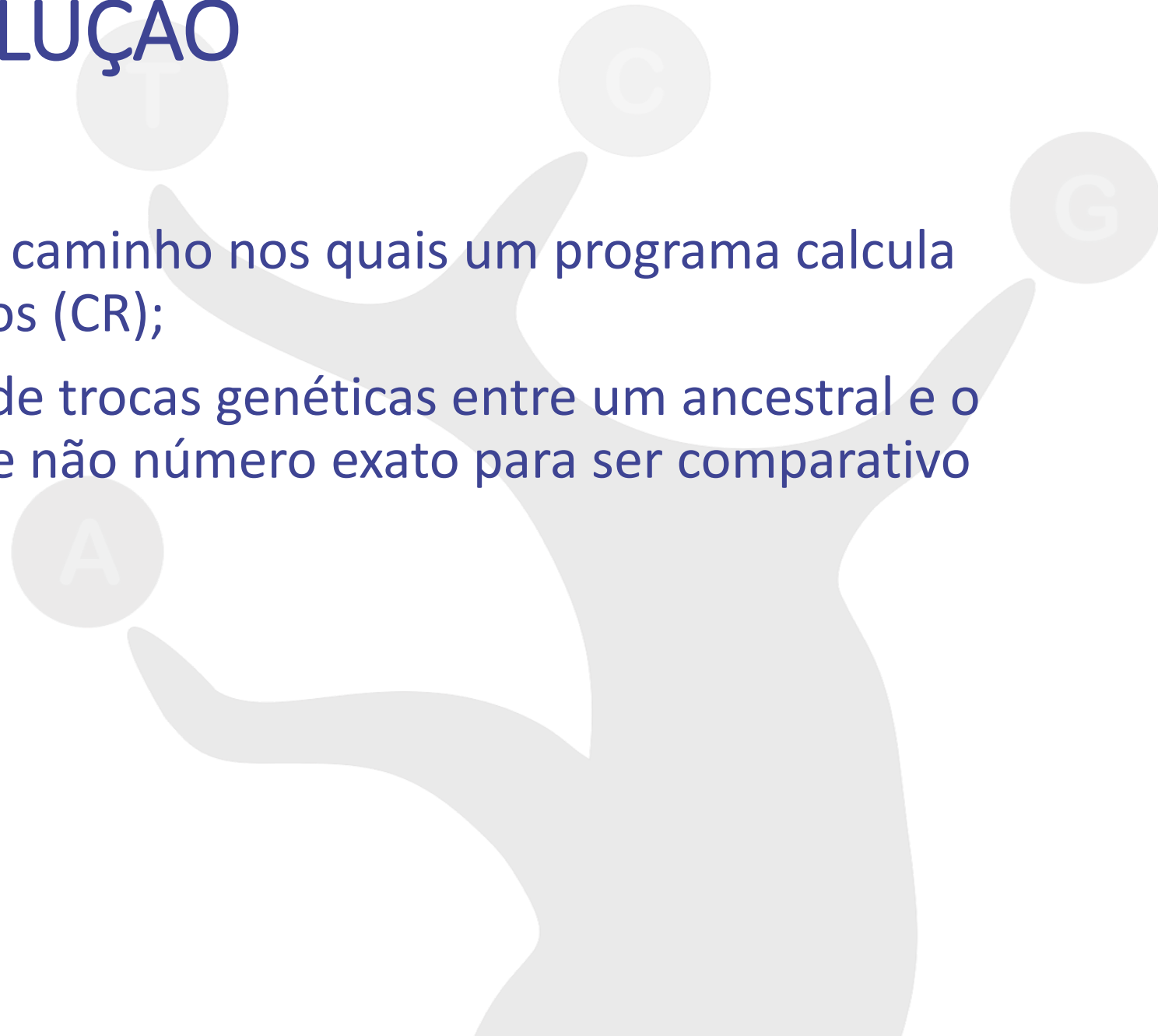
- Tenta-se deduzir a ordem na qual os taxa existentes (sequencias) divergem de um ancestral comum hipotético;
- Calcula-se a quantidade de alterações ao longo dos ramos entre os eventos divergentes;
- Obter uma árvore que melhor se aproxima com o que realmente aconteceu.

Alinhamento Múltiplo

1	A	G	G	C	A	A	G	C	C	A	T	A	G	C	T	G	T	C	C	
2	A	G	G	C	A	A	A	G	A	C	A	T	A	C	C	T	G	A	C	C
3	A	G	G	C	C	A	A	G	A	C	A	T	A	G	C	T	G	T	C	C
4	A	G	G	C	A	A	A	G	A	C	A	T	A	C	C	T	G	T	C	C

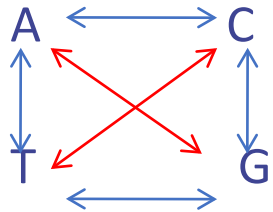
MODELOS DE EVOLUÇÃO

- Os modelos determinam o caminho nos quais um programa calcula os comprimentos dos ramos (CR);
- CR: indicam a quantidade de trocas genéticas entre um ancestral e o descendente – proporção e não número exato para ser comparativo com outros genes.



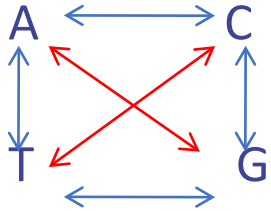
MODELOS DE EVOLUÇÃO

- Modelo 1 parâmetro (modelo de substituição de Jukes-Cantor, 1969): a probabilidade é a mesma



MODELOS DE EVOLUÇÃO

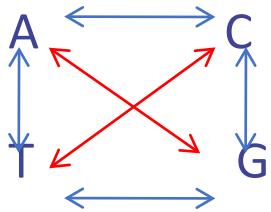
- Modelo 2 parâmetros (modelo de substituição de Kimura, 1980): as probabilidades de transições e transversões diferem



→ Transições são mais frequentes

MODELOS DE EVOLUÇÃO

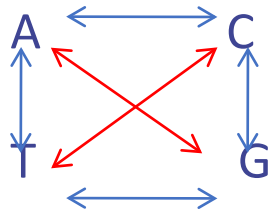
- Modelo Tamura 3 parâmetros (modelo de substituição Tamura e Nei): as probabilidades de transições e transversões diferem, e há diferenças entre as frequências das bases



—————> Transições são mais frequentes

MODELOS DE EVOLUÇÃO

- Modelo GTR (modelo de substituição Geral reversível pelo tempo – Tavaré, 1986) e no modelo HKY85: há 6 diferentes probabilidades de substituição



—————> Transições são mais frequentes

MODELOS DE EVOLUÇÃO

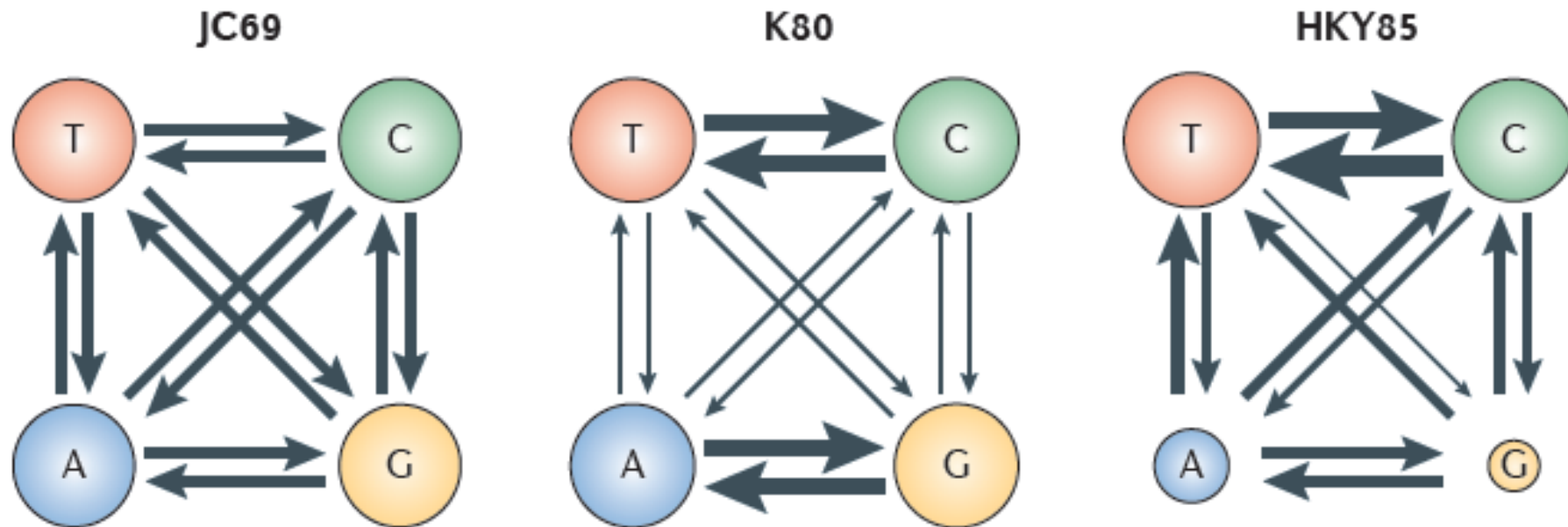
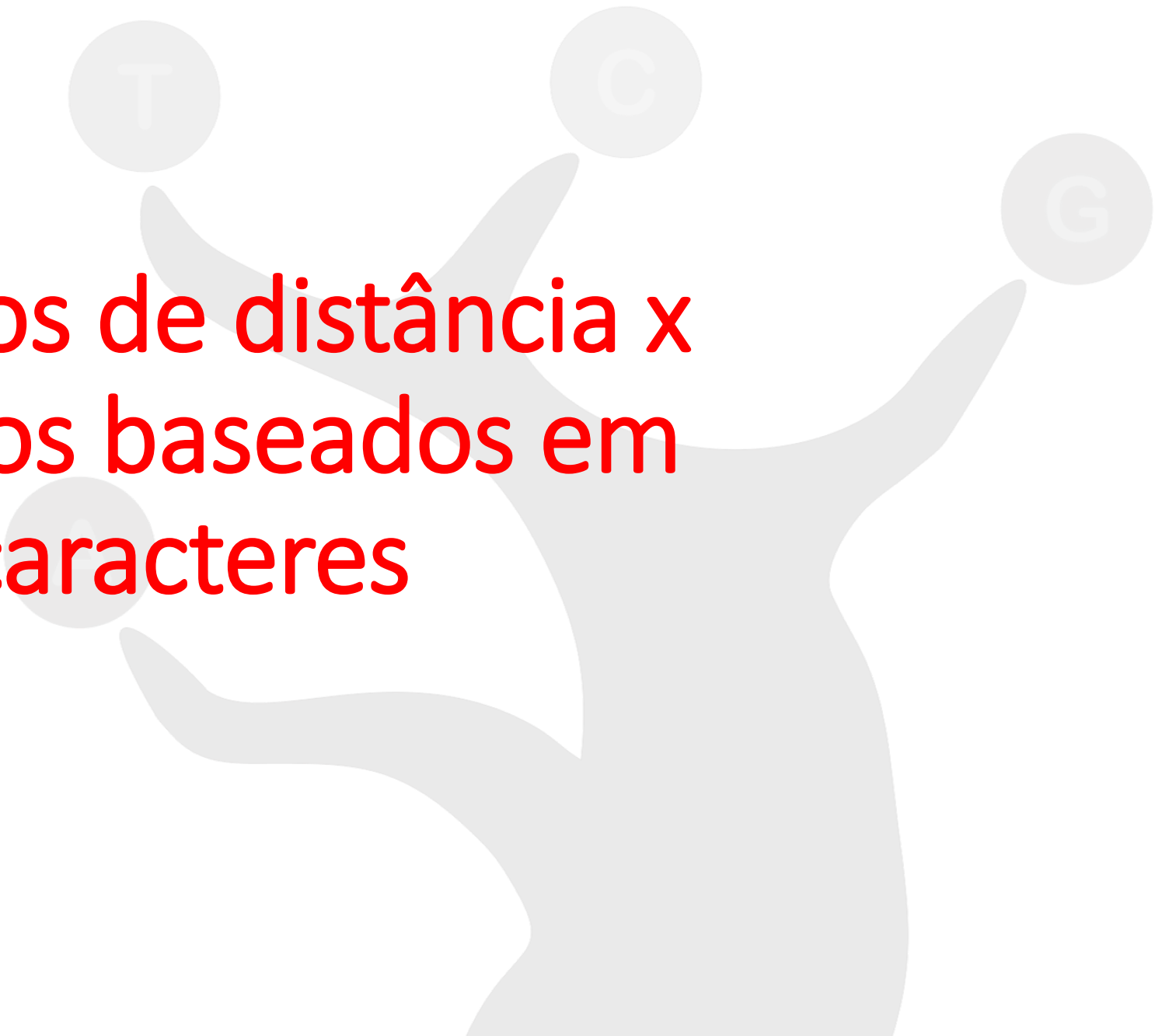


Figure 1 | **Markov models of nucleotide substitution.** The thickness of the arrows indicates the substitution rates of the four nucleotides (T, C, A and G), and the sizes of the circles represent the nucleotide frequencies when the substitution process is in equilibrium. Note that both JC69 and K80 predict equal proportions of the four nucleotides.

Métodos de Análise Filogenética

A decorative background on the right side of the slide. It features a light gray, stylized tree-like shape. At the top of the tree, there are four circular nodes containing the letters 'C', 'G', 'A', and 'T' in a light gray font, representing the four nucleotide bases in DNA. The tree branches downwards and outwards.

- Algorítmicos (distância): Utilizam algoritmos para construir 1 árvore filogenética. São mais rápidos.
 - UPGMA
 - Neighbor Joining
- Busca por árvore (baseados em caracteres): Constróem diversas árvores e usam critérios para definir qual é a melhor (ou melhor conjunto). Mais lentos.
 - Parcimônia
 - Máxima Verossimilhança
 - Inferência Bayesiana



Métodos de distância x métodos baseados em caracteres

MÉTODOS DE DISTÂNCIA: NJ E UPGMA

- Convertem sequencias alinhadas em uma matriz de distância de diferenças pareadas entre as sequencias – semelhante a % de homologia;
- Distâncias são expressas como a fração de sítios que diferem entre 2 sequencias em um alinhamento múltiplo;

MÉTODOS DE DISTÂNCIA: problemas e limitações

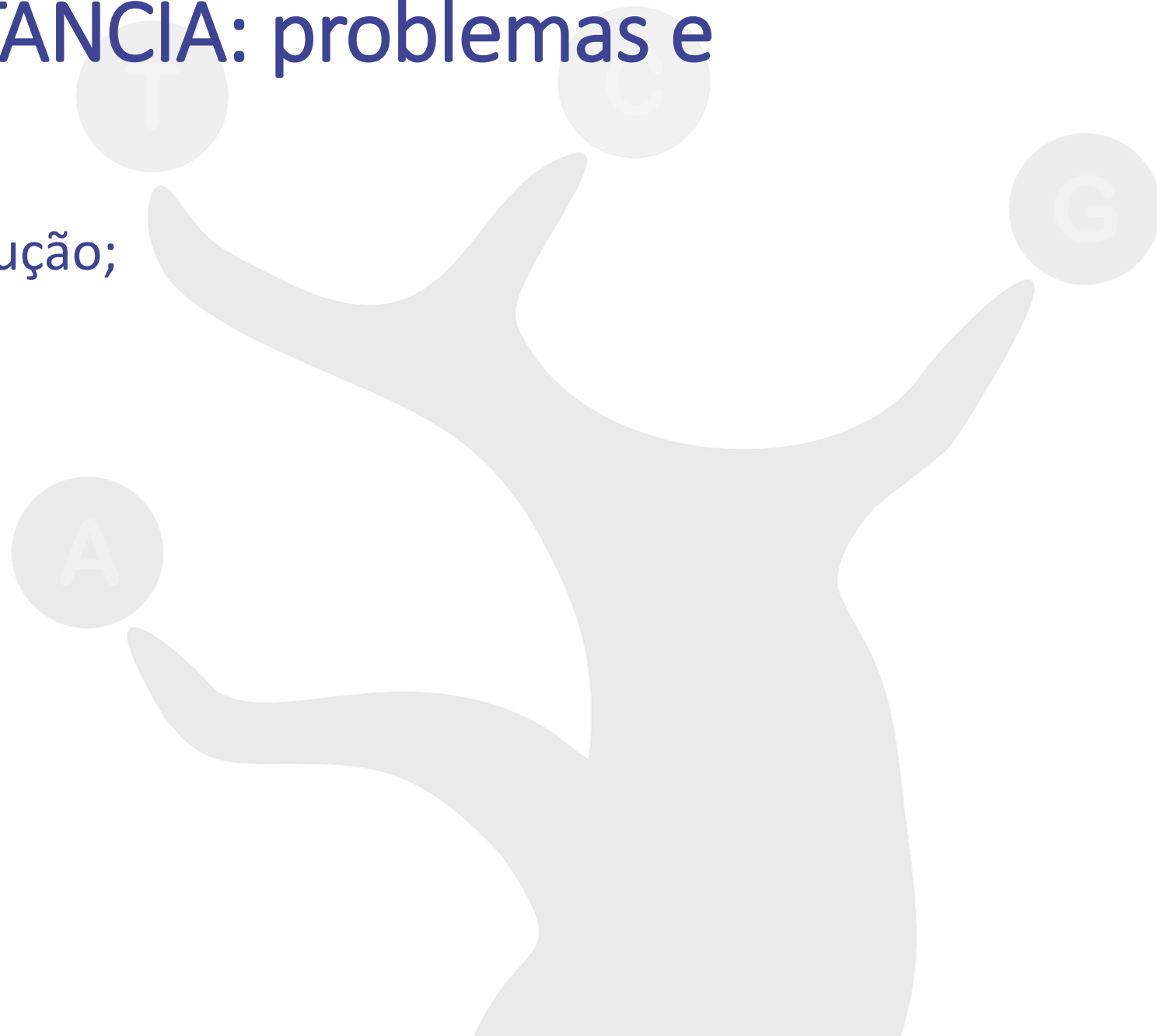
- Diferentes tempos de evolução;
- Substituições múltiplas

- 1) $A \longrightarrow A$

- 2) $A \longrightarrow C \longrightarrow$ $A = 0$

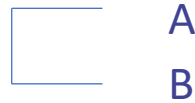
- 1) $A \longrightarrow A$

- 2) $A \longrightarrow C \longrightarrow$ $G = 1$



MÉTODOS DE DISTÂNCIA: upgma

- Método de agrupamento
- Pares de taxa com menor distância e constrói o ramo (metade da distância)



- Ruim para filogenia porque é ultramétrica: todos os taxa são igualmente distantes de 1 raiz

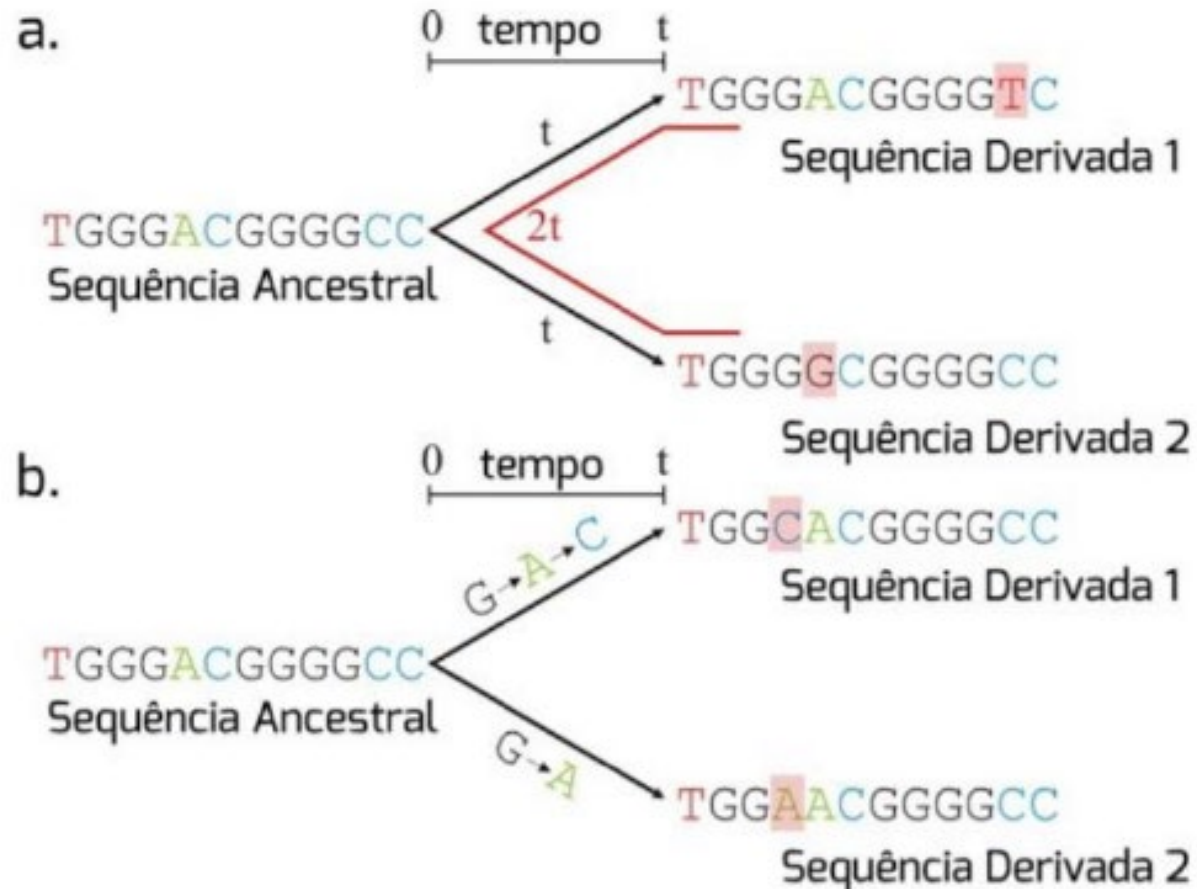
Multiple alignment

1 AGGCCAAGCCATAGCTGTCC
2 AGGCAAAGACATACCTGACC
3 AGGCCAAGACATAGCTGTCC
4 AGGCAAAGACATACCTGTCC

Distance matrix

	1	2	3	4
1	—	0.20	0.05	0.15
2		—	0.15	0.05
3			—	0.10
4				—

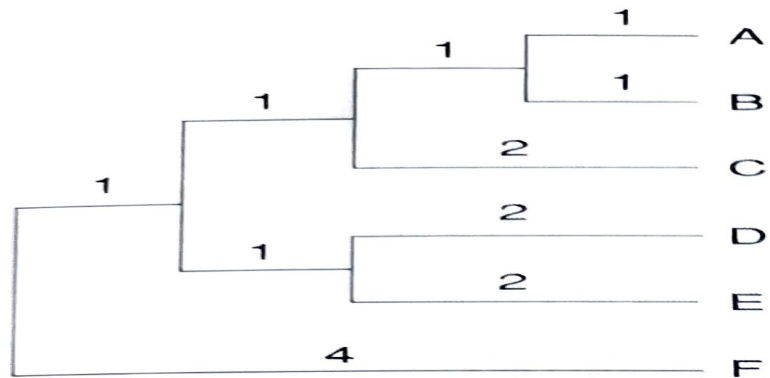
Distancia genética: o número de substituições de nucleotídeos que se acumularam nas sequências desde a divergência



- Distância p : contagem das diferenças dividida pelo número total de sítios do alinhamento
- 8 sítios diferentes
- 100pb
- Distância $p = 0,08$
- A ocorrência de múltiplas substituições ao longo do tempo na divergência de sequências homólogas pode mascarar as verdadeiras diferenças entre as sequências

EXEMPLO

Box 5.1 Cluster analysis (Sneath & Sokal, 1973)

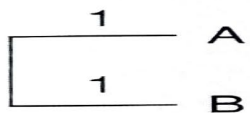


	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

N = 6

Cluster analysis proceeds as follows:

- (1) Group together (cluster) these OTUs for which the distance is minimal; in this case group together A and B. The depth of the divergence is the distance between A and B divided by 2.



(2) Compute the distance from cluster (A, B) to each other OTU

$$d_{(AB)C} = (d_{AC} + d_{BC})/2 = 4$$

$$d_{(AB)D} = (d_{AD} + d_{BD})/2 = 6$$

$$d_{(AB)E} = (d_{AE} + d_{BE})/2 = 6$$

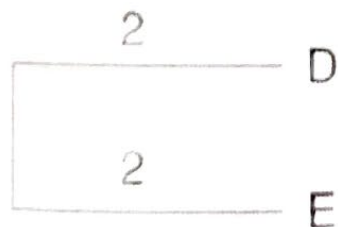
$$d_{(AB)F} = (d_{AF} + d_{BF})/2 = 8$$

	(AB)	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

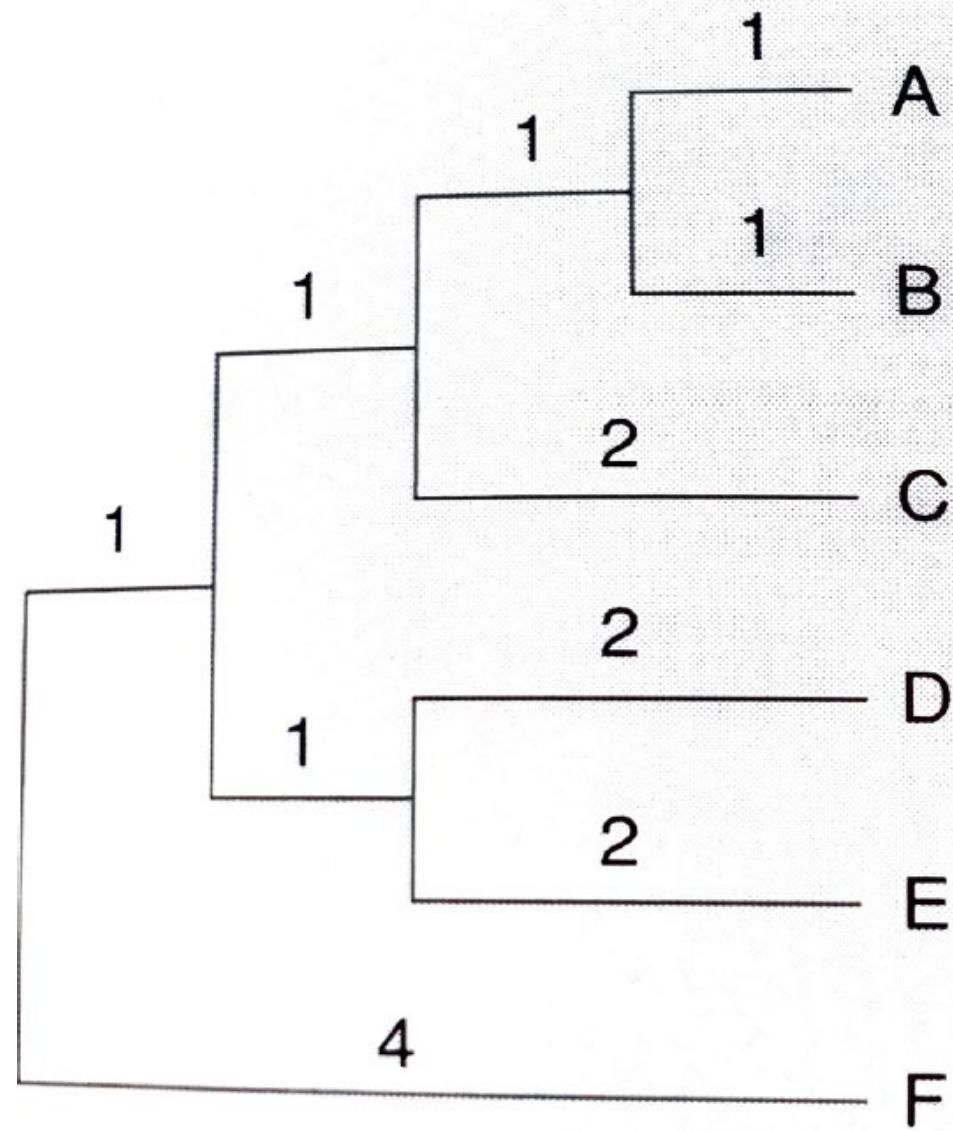
Repeat steps 1 and 2 until all OTUs are clustered (repeat until $N = 2$)

$$N = N - 1 = 5$$

(1) Group together (cluster) these OTUs for which the distance is minimal, e.g. group together D and E. Alternatively, (AB) could be grouped with C.



$$N = N - 1 = 2$$



MÉTODOS DE DISTÂNCIA: upgma

- Método de agrupamento
- Pares de taxa com menor distância e constrói o ramo (metade da distância)

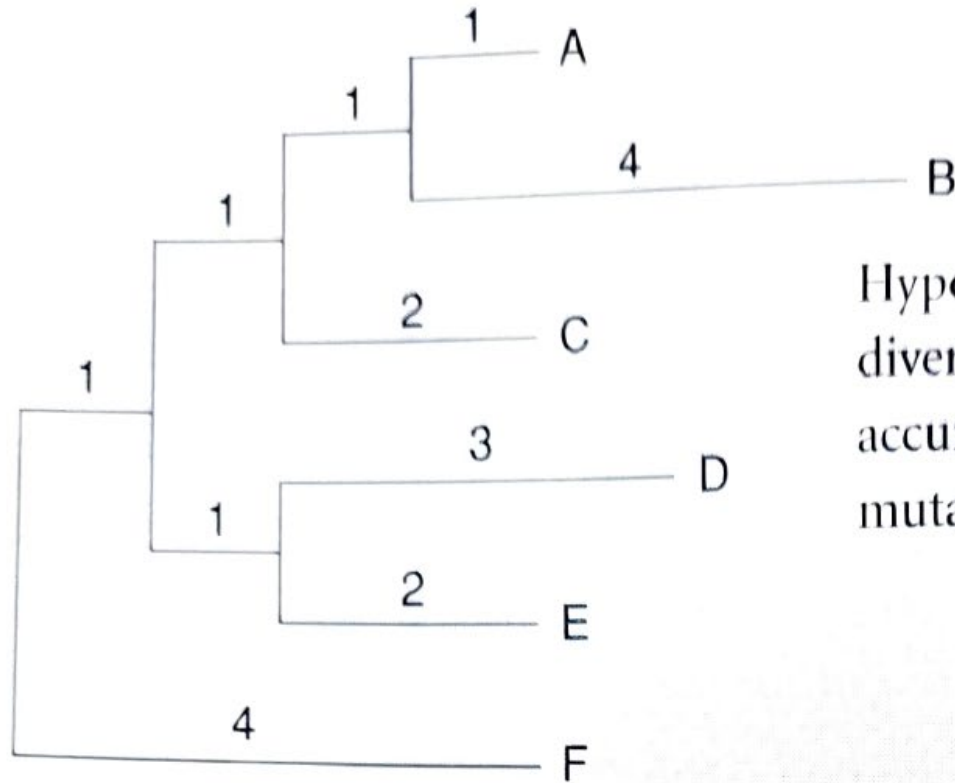


- Ruim para filogenia porque é ultramétrica: todos os taxa são igualmente distantes de 1 raiz

MÉTODOS DE DISTÂNCIA: Neighbour Joining

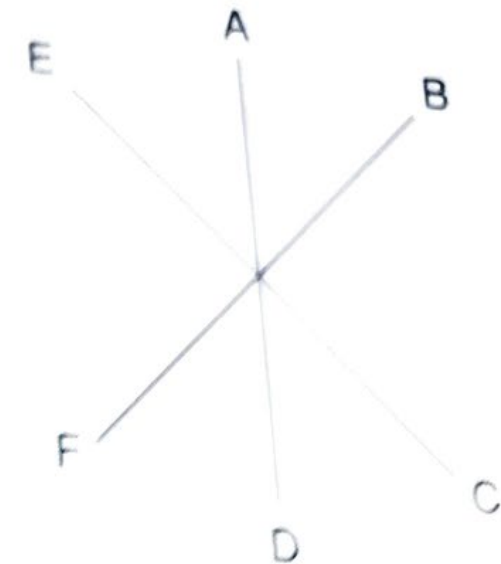
- Diferenças de UPGMA:
 - Não constrói grupos, e sim calcula as distâncias diretamente para os nós internos.
 - A partir da matriz original:
 - 1) calcula para cada taxon sua divergência de todos os outros taxa como a soma de todas distâncias individuais do taxon
 - 2) corrige a matriz original

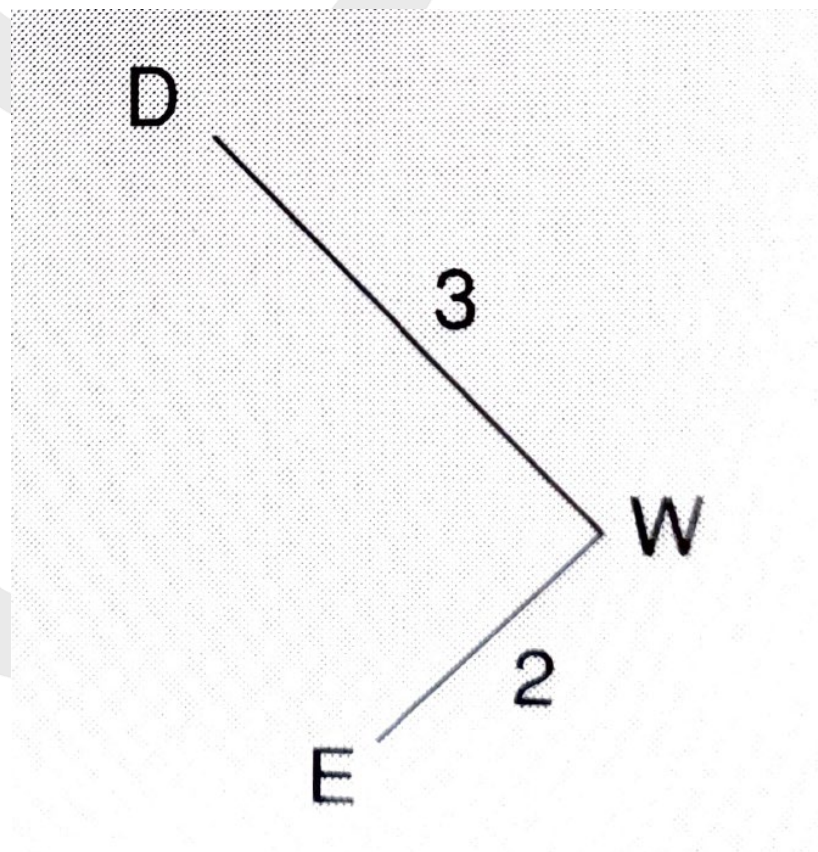
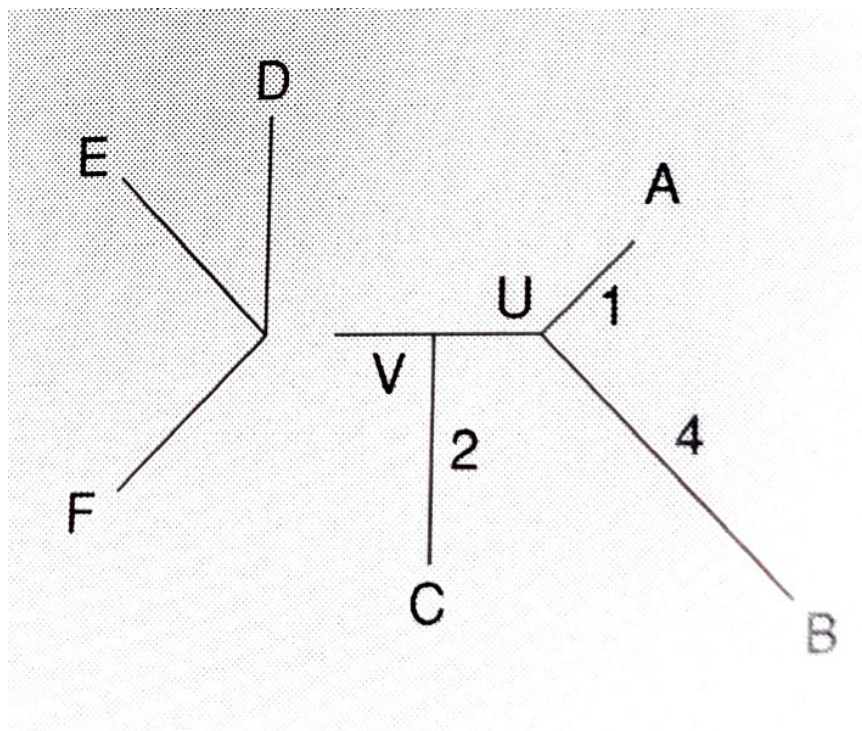
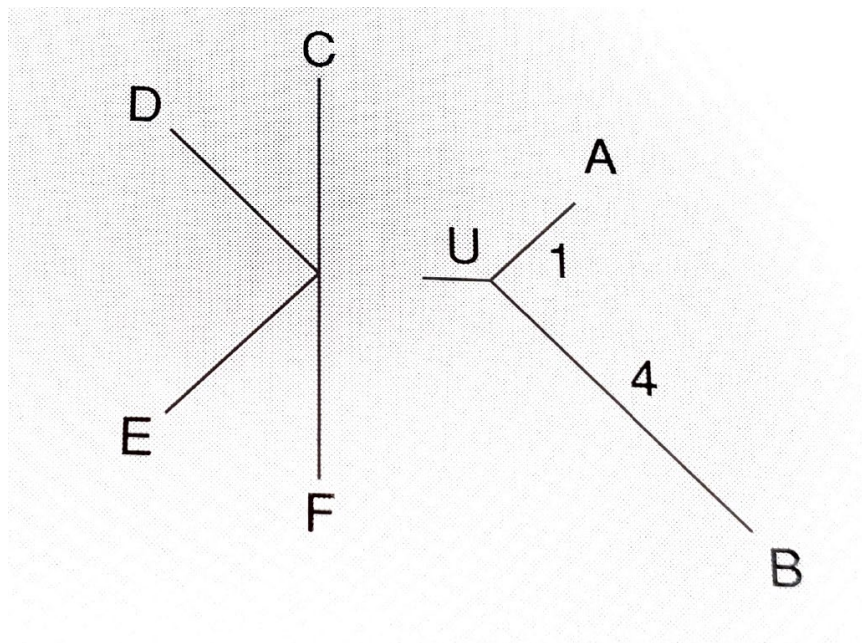
Box 5.2 The neighbor-joining method (Saitou & Nei, 1987; modified from Studier & Keppler, 1988)

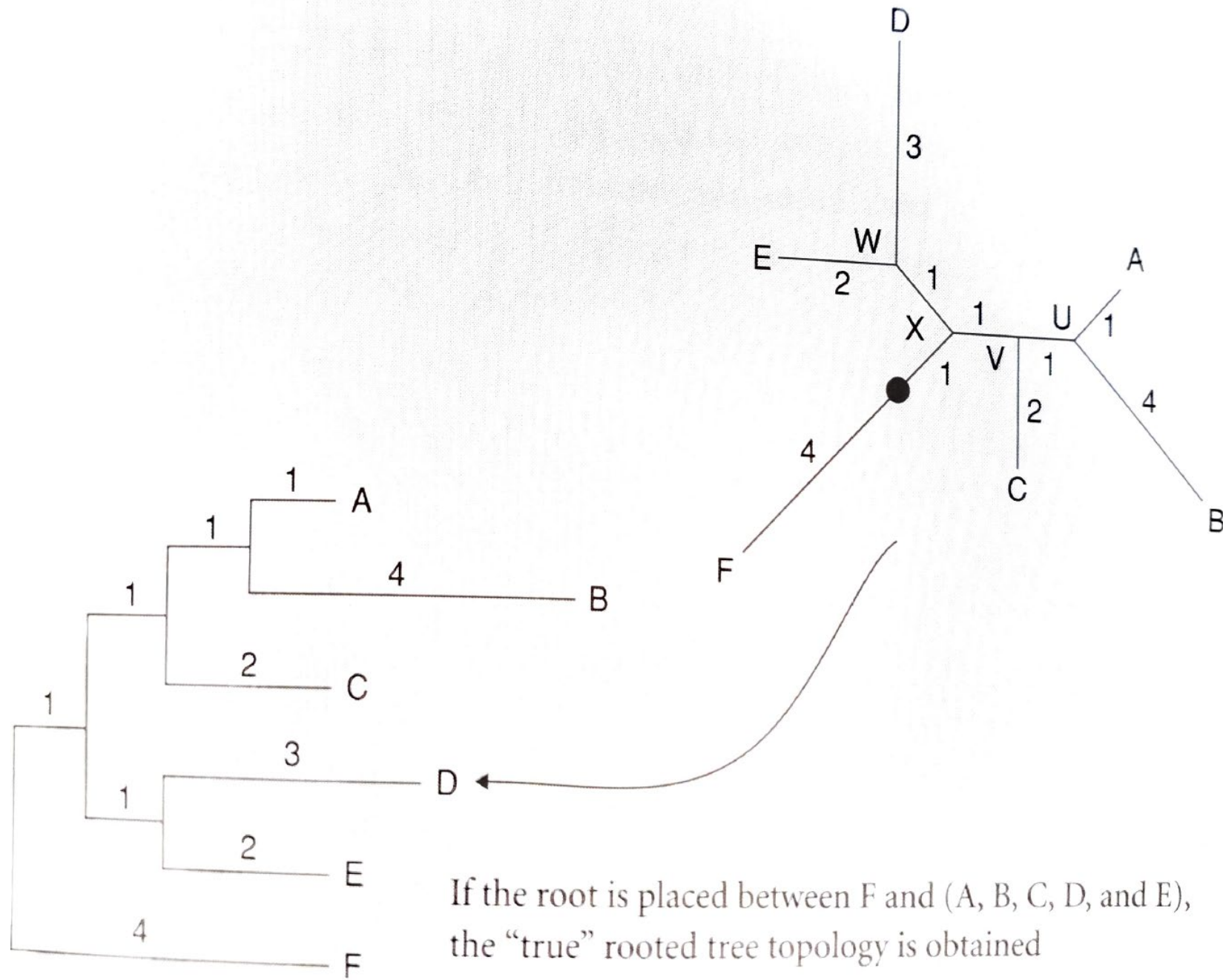


Hypothetical tree topology: since the divergence of sequences A and B, B has accumulated four times as many mutations as sequence A.

	A	B	C	D	E
B	(-13)				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	(-13)	
F	-10.5	-10.5	-11	-11.5	-11.5







If the root is placed between F and (A, B, C, D, and E), the “true” rooted tree topology is obtained

MÉTODOS DE DISTÂNCIA: Neighbour Joining

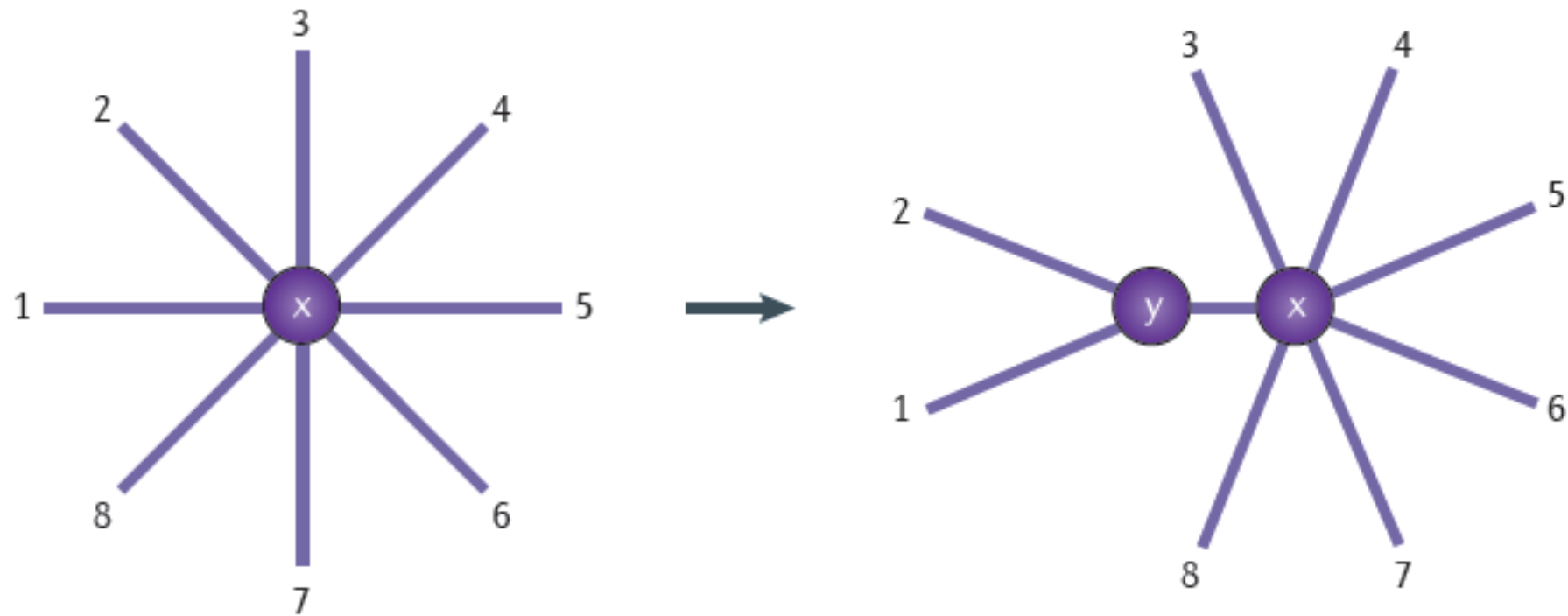
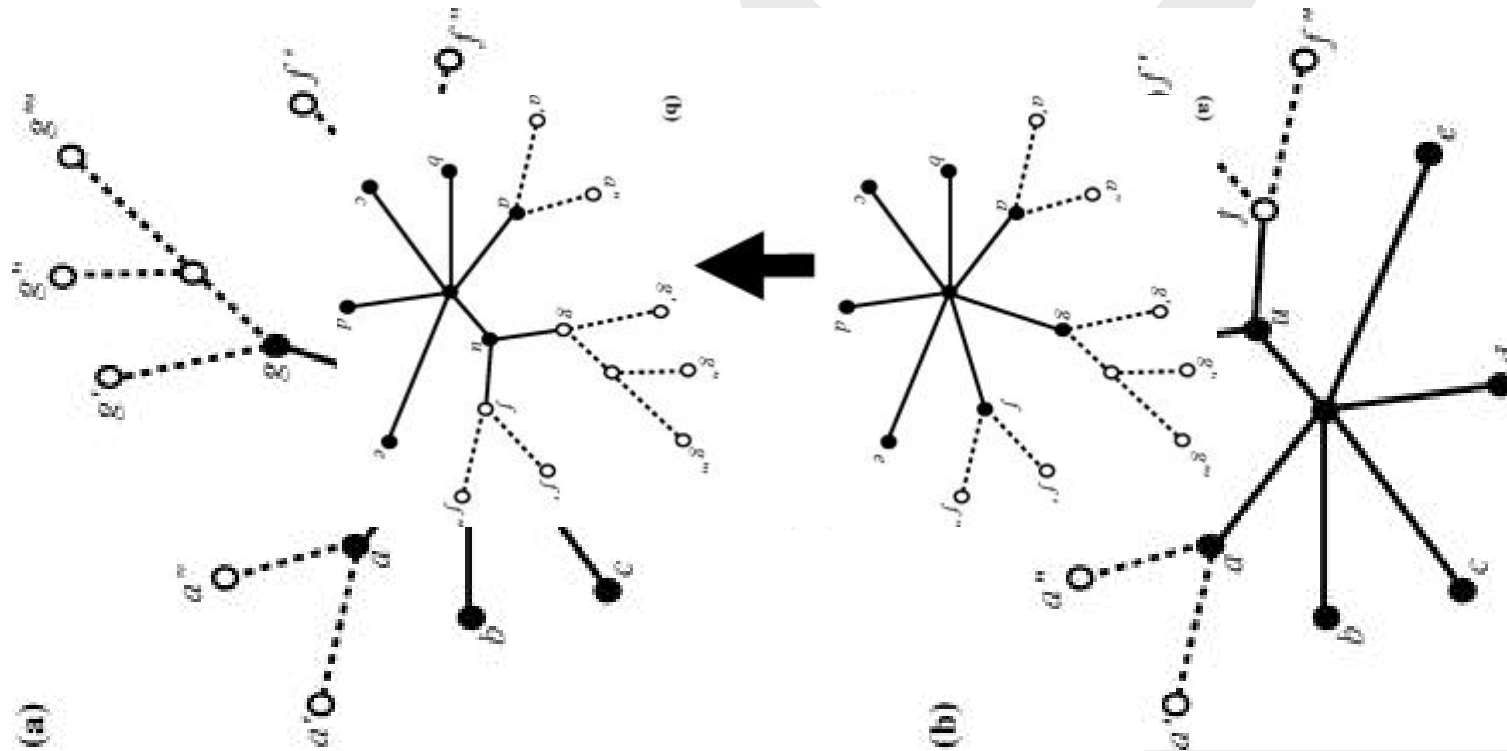


Figure 2 | **The neighbour joining algorithm.** The neighbour joining algorithm is a divisive cluster algorithm. It starts from a star tree: two nodes are then joined together on this tree (in this example, nodes 1 and 2), reducing the number of nodes at the root (node x) by one. The process is repeated until a fully resolved tree is generated.

MÉTODOS DE DISTÂNCIA: Neighbour Joining



MÉTODOS DE BUSCA POR ÁRVORES (TOPOLOGIAS) OU BASEADOS EM CARACTERES

- Parcimônia, Maxima verossimilhança e Inferência Bayesiana: usam alinhamento múltiplo diretamente pela comparação de caracteres em cada coluna (sítio) no alinhamento;
- Buscam a árvore que melhor encontra alguns critérios ótimos pela avaliação de árvores individuais
 - Busca exaustiva – qdo o número de taxa é pequeno.
 - 10 taxa: >34 milhões de árvores
 - Algoritmo de ramo vinculado
 - Heurístico

MÉTODO DE PARCIMÔNIA

- Parcimônia: árvore (s) com o número mínimo de alterações. Normalmente mais árvores.
- Baseada na suposição de que a árvore mais provável é aquela que requer o menor número de eventos evolutivos (substituições) para explicar os dados no alinhamento.
- O método calculará as probabilidades de mudanças dos nucleotídeos nos ramos da árvore.
- A parcimônia considera cada sítio do alinhamento individualmente e calcula as probabilidades de ocorrência dos quatro nucleotídeos nos táxons ancestrais
 - Premissa básica: taxa apresentam características comuns porque eles compartilham esta característica de um ancestral comum;

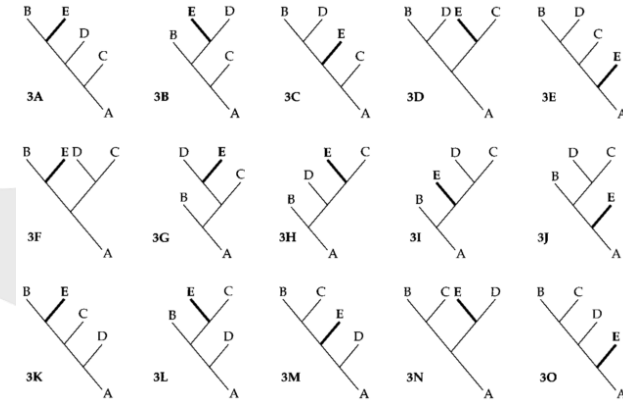
MÉTODO DE PARCIMÔNIA

Dados

AGTGAAGA
AGTGAAGA
AGTGAAGA
AGTGAAGA
AGTGGAGA
AGTGAAGA
AGTGAAGA

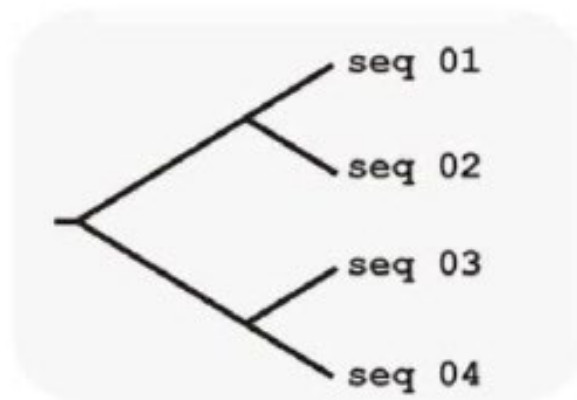


Contagem
de Passos
em
diversas
Topologias

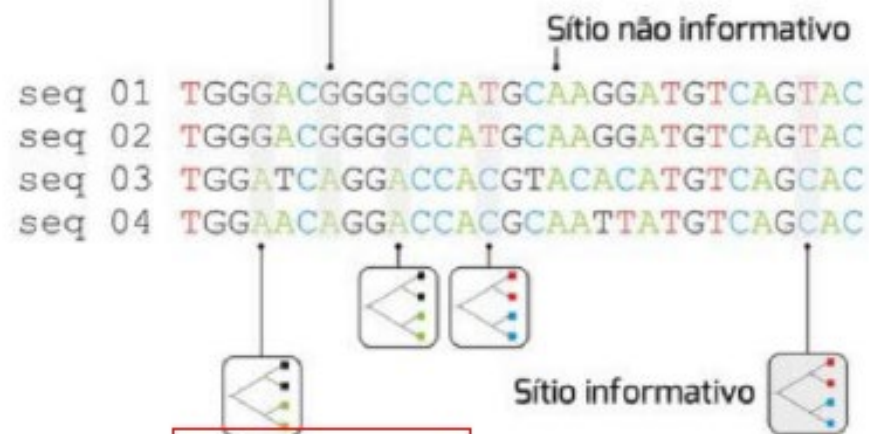


Topologia com o menor
número de passos
= escolhida

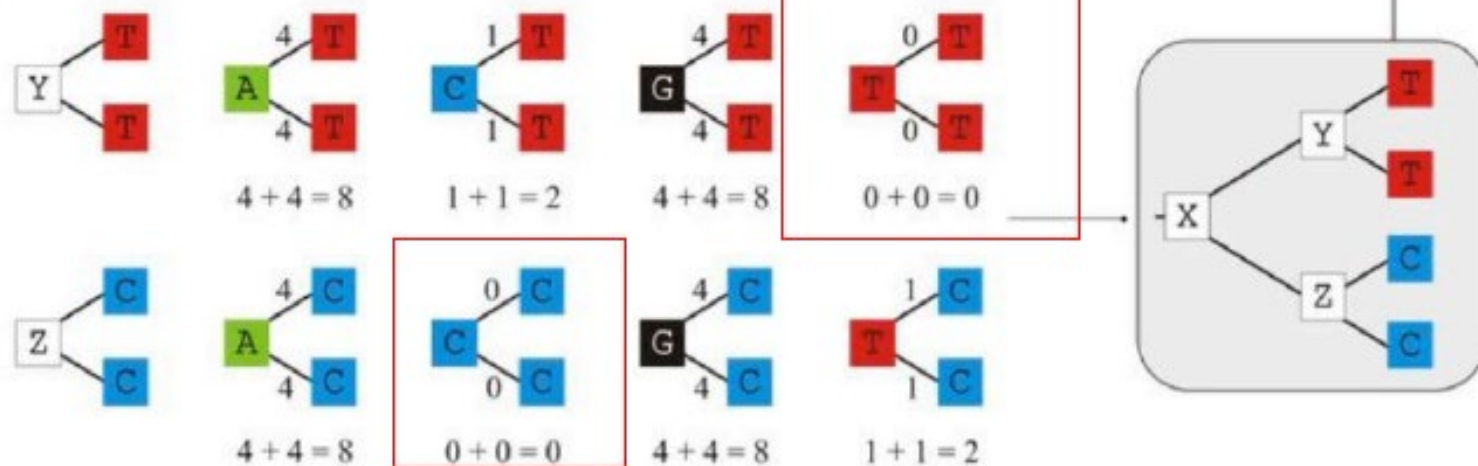
a.



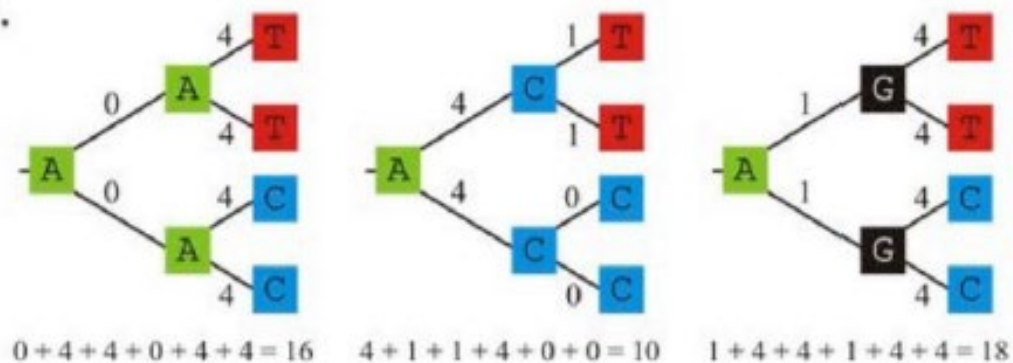
b.



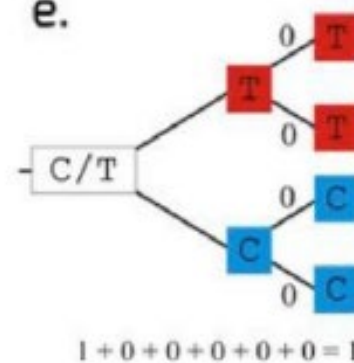
c.



d.

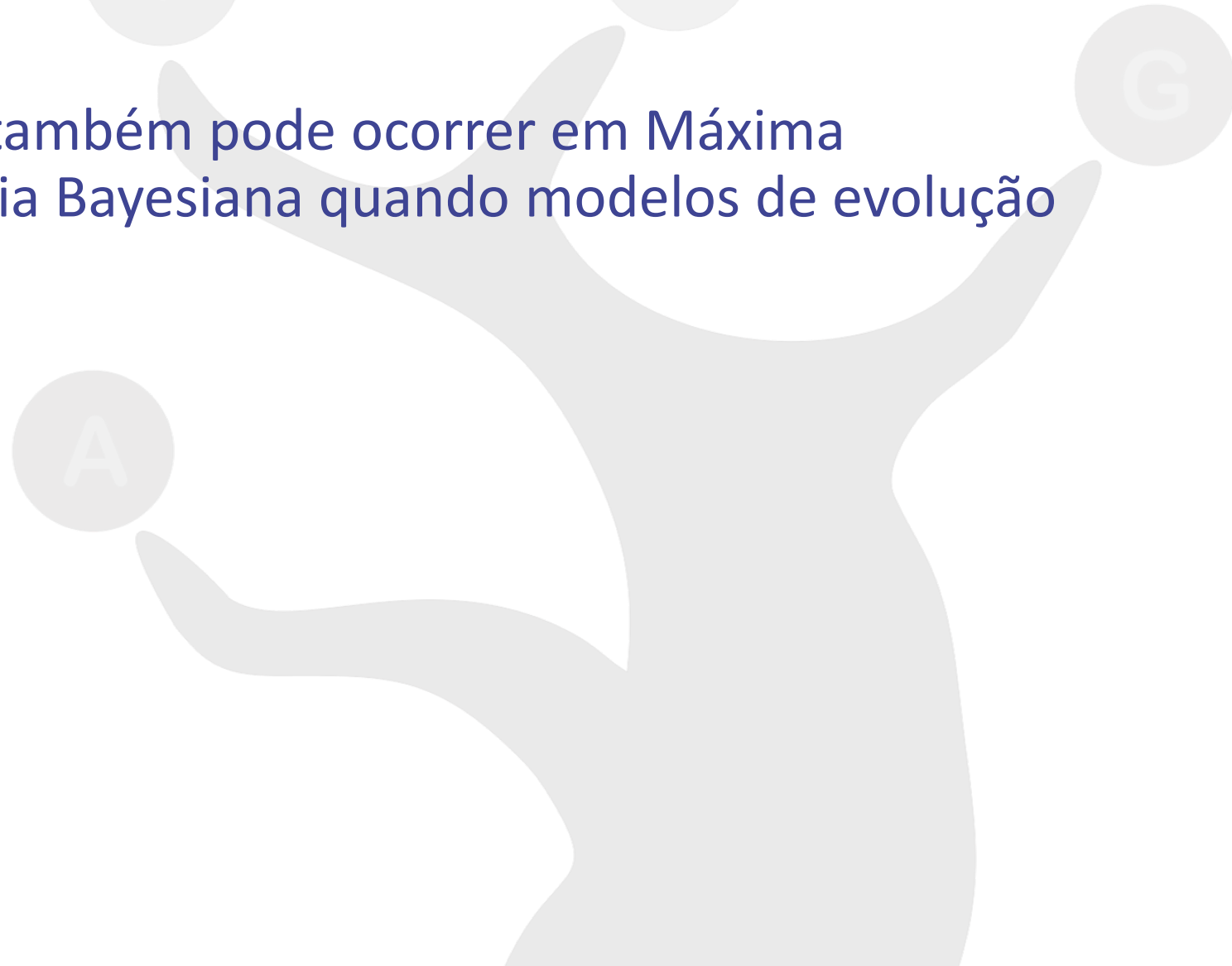


e.



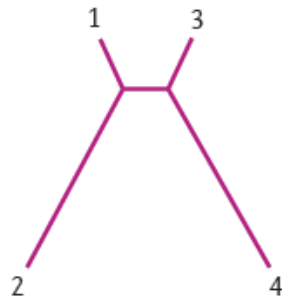
MÉTODO DE PARCIMÔNIA - Problemas

- Atração de ramos longos (também pode ocorrer em Máxima verossimilhança e Inferência Bayesiana quando modelos de evolução simples são utilizados)

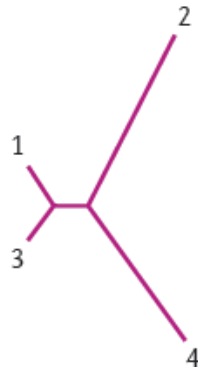


ATRAÇÃO DE RAMOS LONGOS

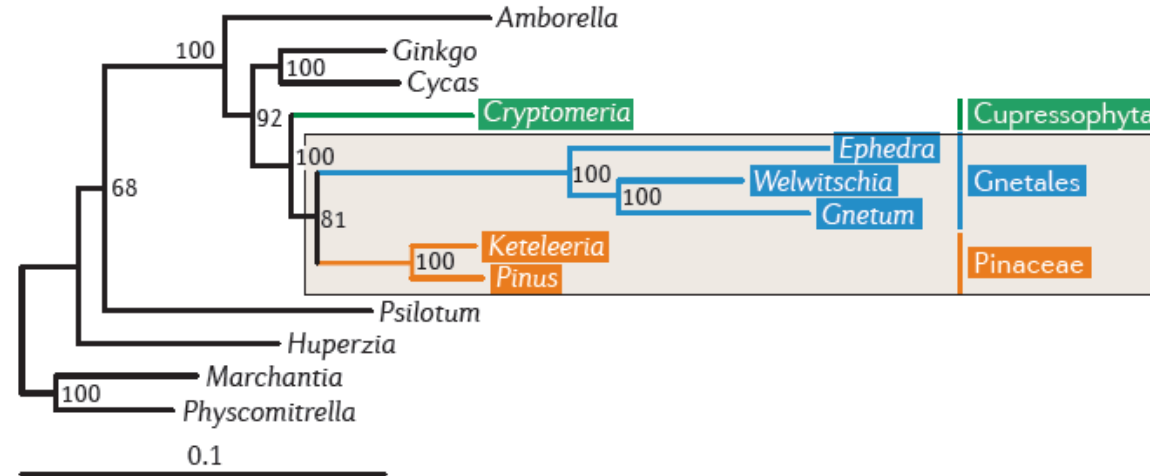
a Correct tree, T_1



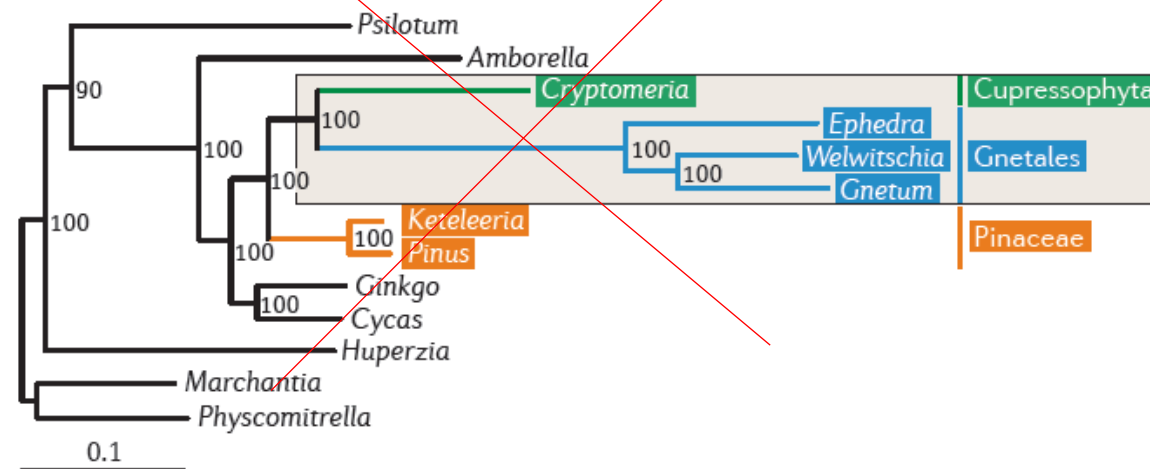
b Wrong tree, T_2



c The Gnepine tree



d The GneCup tree

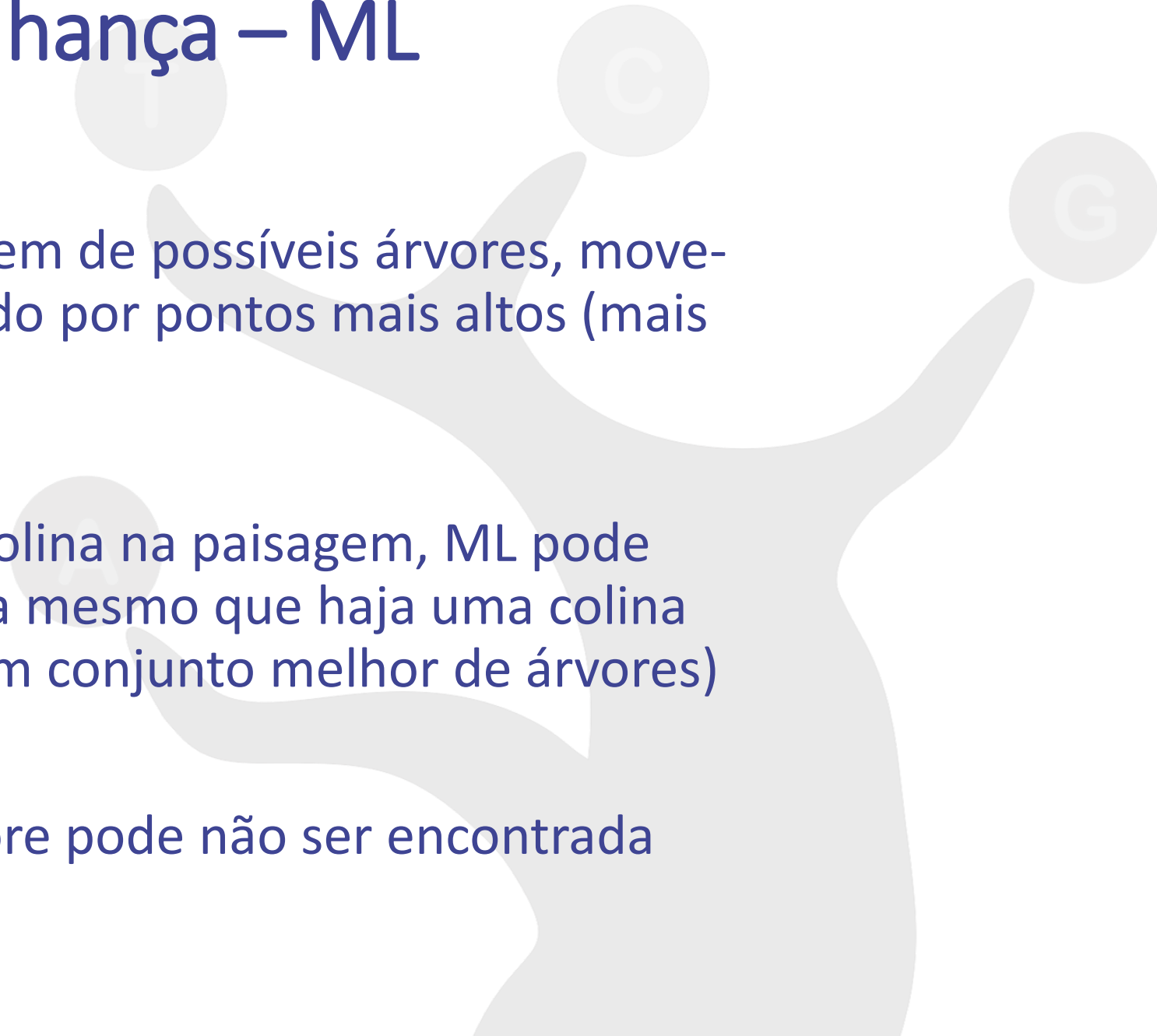


MÉTODOS DE BUSCA POR ÁRVORES (TOPOLOGIAS) OU BASEADOS EM CARACTERES

- M.L.: árvore que sob algum modelo de evolução, maximiza a probabilidade de observar os dados. Geralmente 1 árvore – a mais provável
- Inferência Bayesiana: utiliza os métodos de cadeia de Markov e Monte Carlo (MCMC) para estimar a distribuição a posteriori dos parâmetros do modelo

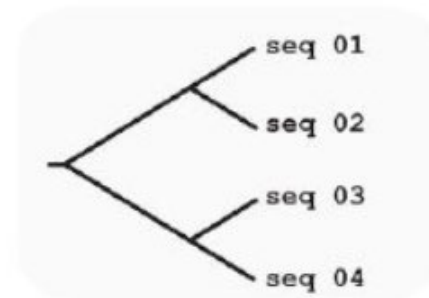
Máxima Verossimilhança – ML

- Ao pesquisar uma paisagem de possíveis árvores, move-se ponto a ponto buscando por pontos mais altos (mais prováveis árvores)
- Se houver mais de uma colina na paisagem, ML pode ficar preso em uma colina mesmo que haja uma colina maior nessa paisagem (um conjunto melhor de árvores)
- Conclusão: a melhor árvore pode não ser encontrada

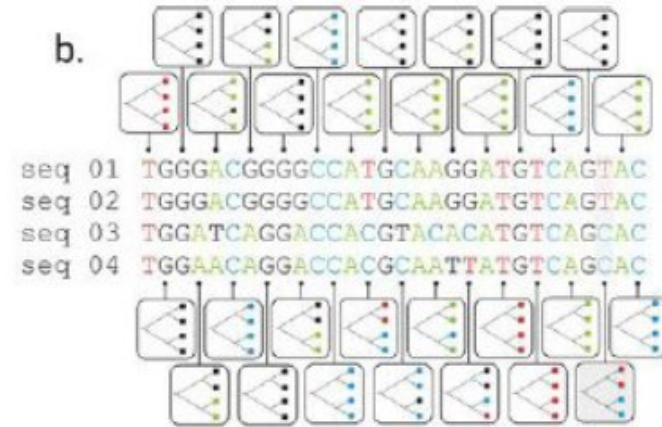


Máxima Verossimilhança – ML

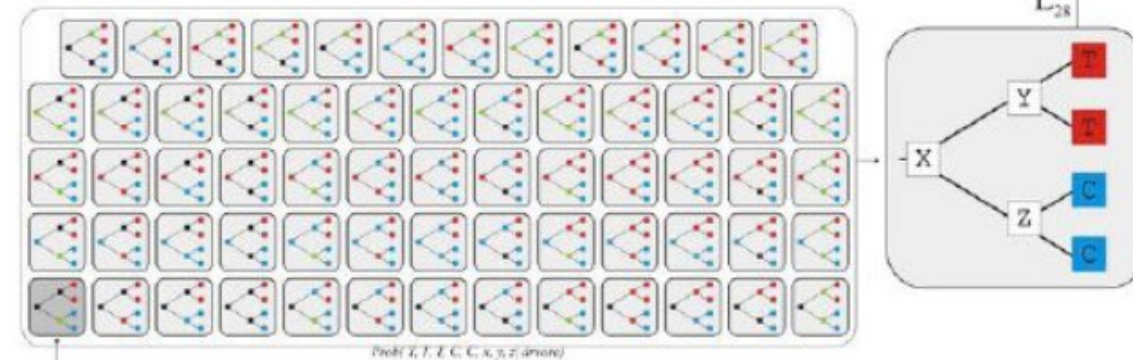
a.



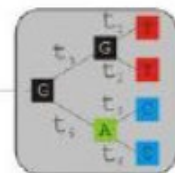
b.



c.



d.



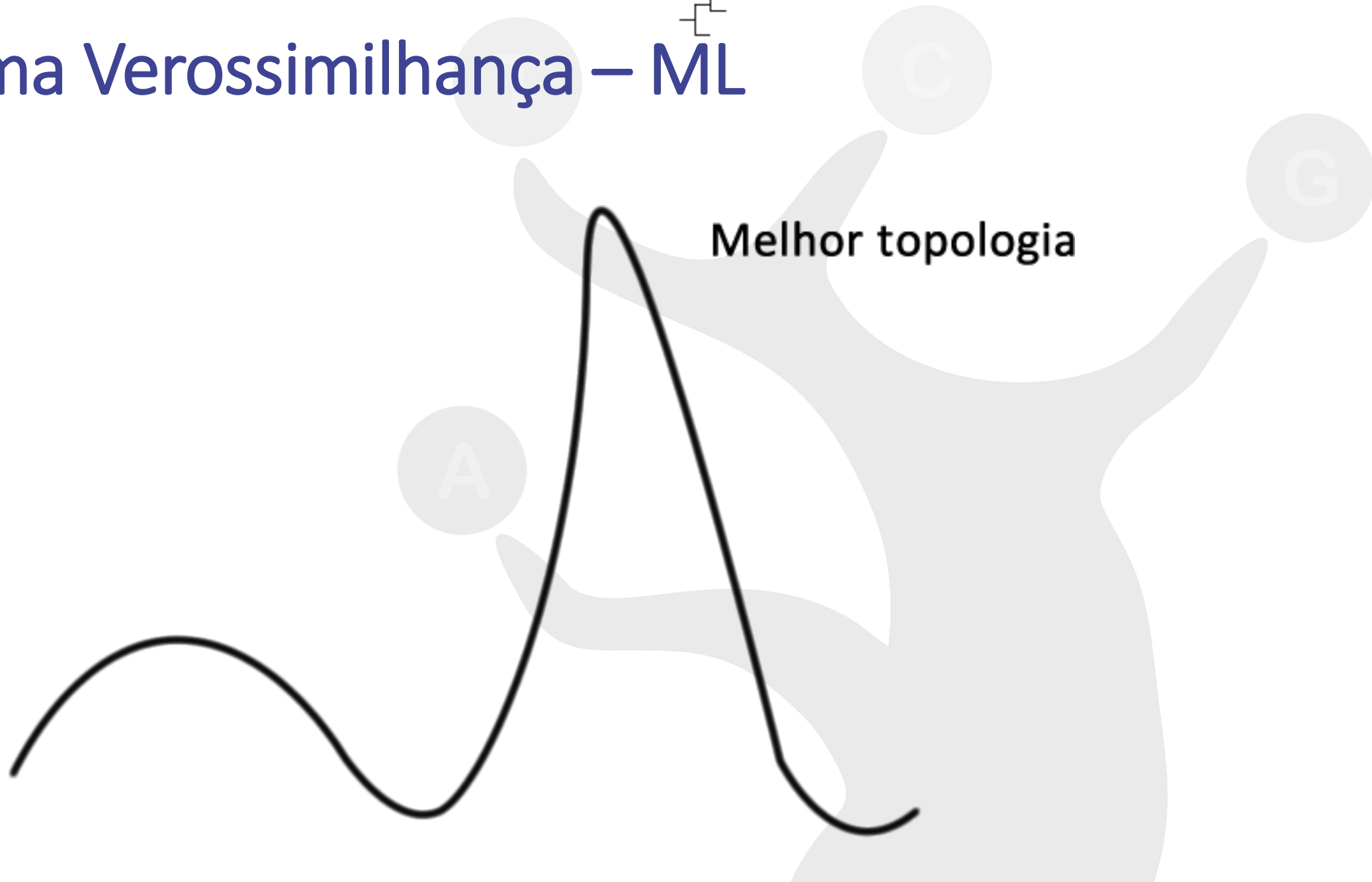
$Prob(T, T, T, C, C, x, y, z | \text{árvore}) =$

$$\pi_G \cdot P_{GG} \cdot (t_5) \cdot P_{GT} \cdot (t_1) \cdot P_{GT} \cdot (t_2) \cdot P_{GA} \cdot (t_6) \cdot P_{AC} \cdot (t_3) \cdot P_{AC} \cdot (t_4)$$

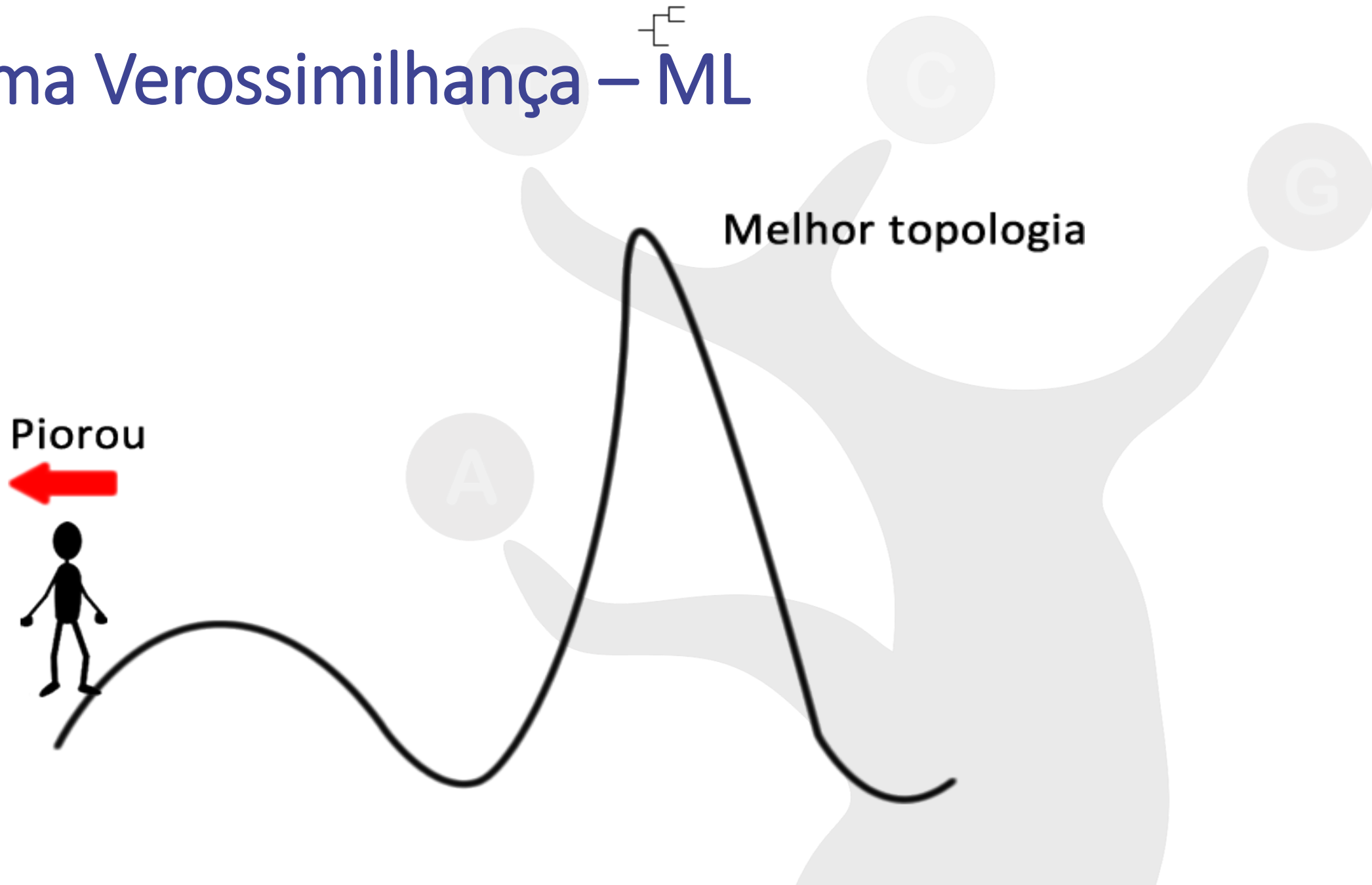
Probabilidade de ocorrência dos nucleotídeos (A, C, G, T) no nó mais ancestral é dada pela frequência estacionária.

Probabilidade de mudança do nó X para o nó Y multiplicando pelo tamanho do ramo t_i e considerando variações em parâmetros do modelo de substituição.

Máxima Verossimilhança – ML

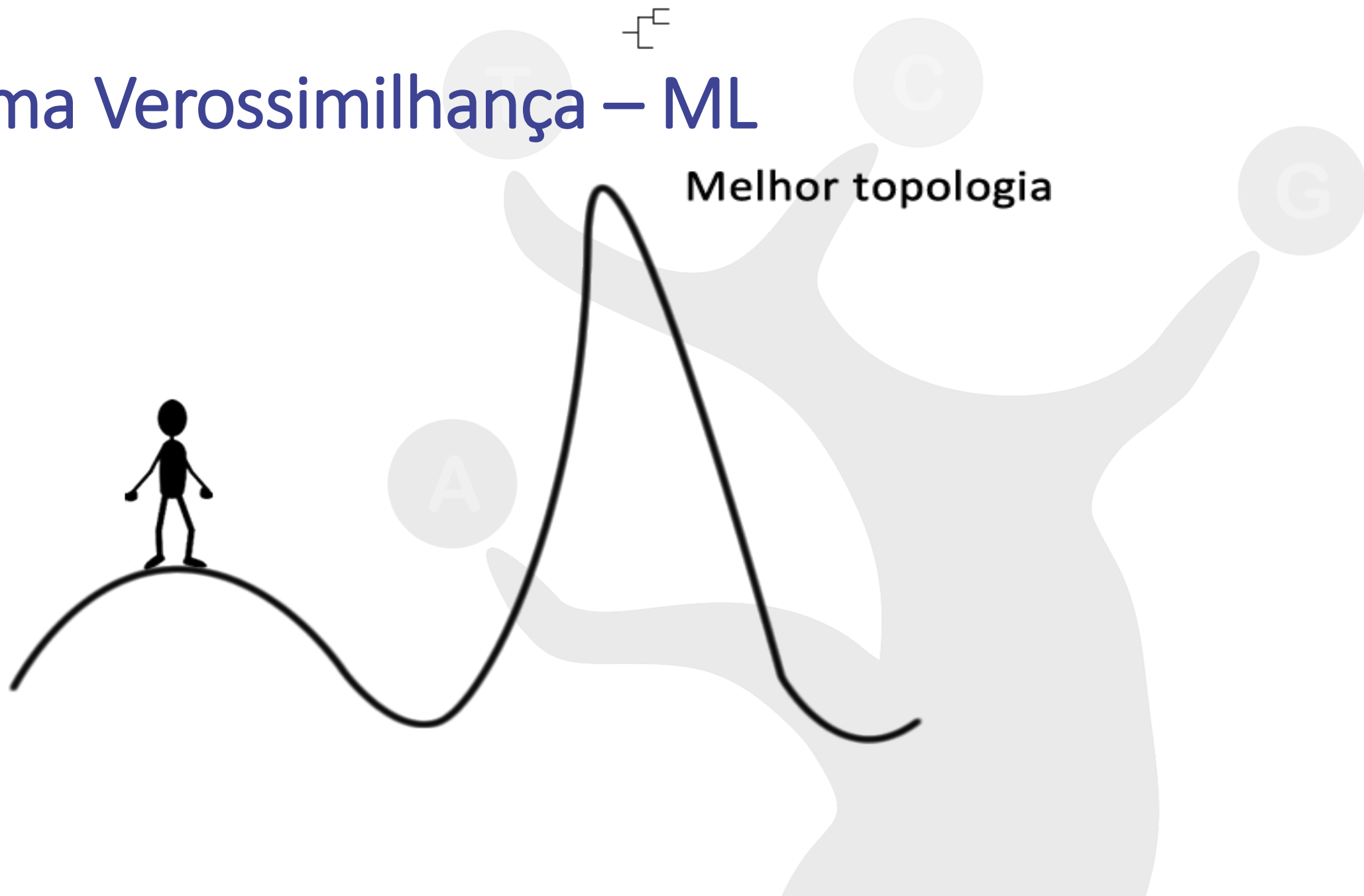


Máxima Verossimilhança – ML

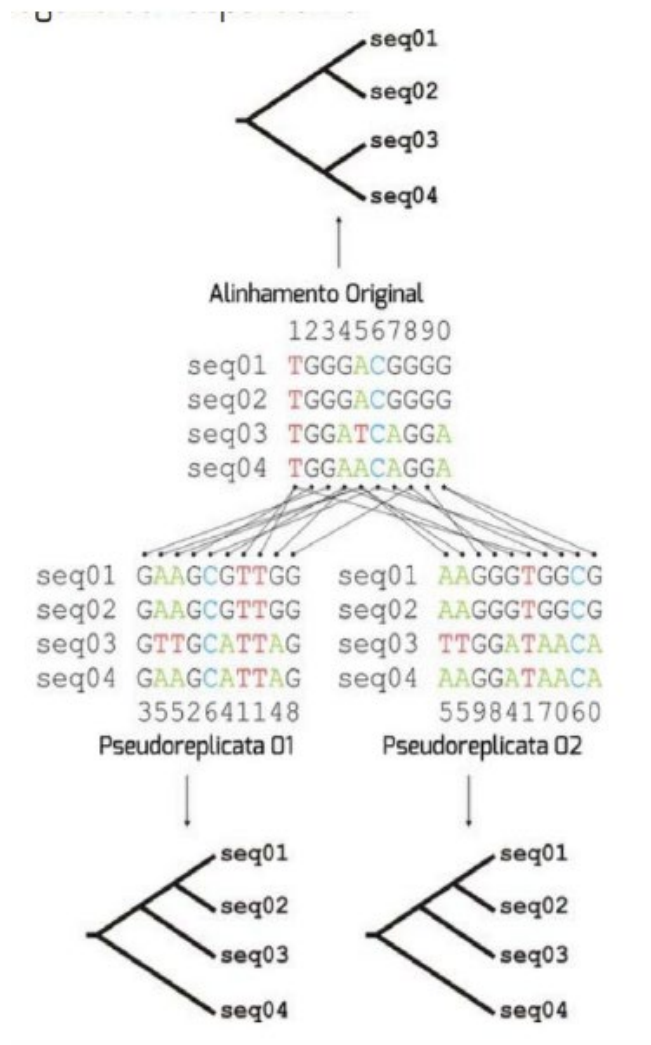


Máxima Verossimilhança – ML

Melhor topologia



Confiabilidade - *Bootstrap*



- Como é possível saber se a amostragem foi suficiente e a filogenia é confiável?
- Reamostragem – Usar novas amostras e a filogenia resultante ser a mesma
- Algoritmos: **bootstrap** – método de reamostragem
- Ao final, o algoritmo analisará os clados e automaticamente verificará a presença de determinados agrupamentos em todas as filogenias construídas.
- Exemplo: as sequências 1 e 2 formam um clado em 70% das filogenias construídas, atribuiremos a confiabilidade de 70 ao clado formado por estas duas sequências.
- Teste estatístico para medir o grau de suporte dos nós nas árvores filogenéticas pelo alinhamento das sequências

Inferência Bayesiana

A faint, light gray background graphic of a stylized tree. The tree has a central trunk and several branches. At the end of some branches are circular nodes containing the letters 'T', 'C', and 'G' in a light gray font. The 'T' is on a node near the top left, 'C' is on a node near the top right, and 'G' is on a node further to the right. The tree's structure is abstract and serves as a decorative element for the slide.

- Inferência Bayesiana: Variante mais recente de ML: árvores com a maior probabilidade com base nos dados. Probabilidades a posteriori
- Usuário define um modelo de evolução, e o software irá procurar pelas melhores árvores, que sejam coerentes com o modelo e o conjunto de dados (alinhamento)
- Produz 1 conjunto de árvores com probabilidades semelhantes – Frequencia de um clado é virtualmente a probabilidade daquele clado – sem bootstrap

Inferência Bayesiana



- É uma formalização matemática de um processo de tomada de decisão – análise de probabilidade
- Baseada na noção de probabilidades a posteriori – Probabilidades que são estimadas baseadas em algum modelo, após obter alguma informação a partir dos dados
- Exemplo: Campeão do mundo em
 - Futebol
 - Hóquei no gelo

Inferência Bayesiana

- Exemplo:

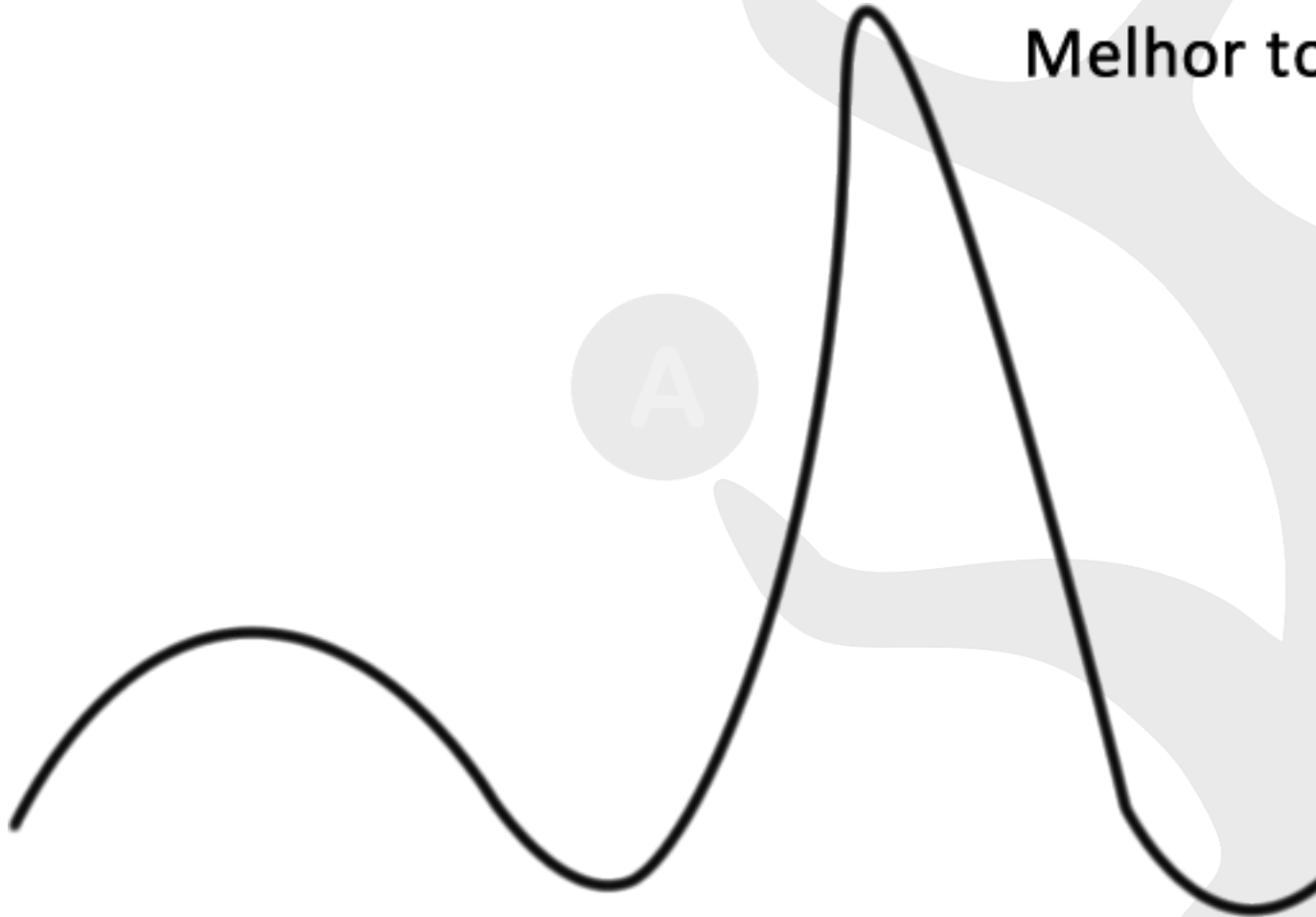
- Moedas em um saco: sabe-se que existe 90% de moedas corretas e 10% de moedas viciadas (80% das vezes sai *cara* para cima).
- Qual a probabilidade de aleatoriamente você retirar uma moeda viciada? 0.1
- Mas se você observar a seguinte sequencia: *Cara, Cara, Coroa, Cara, Cara, Coroa, Coroa, Cara, Cara, Cara*.
- Qual a probabilidade de tal sequencia, sendo essa moeda verdadeira?
 $0,5^{10}=9,76 \times 10^{-4}$
- Qual a probabilidade de tal sequencia, sendo essa moeda viciada? $0,8^7 \times 0,2^3=1,67 \times 10^{-3}$
- Qual a probabilidade a posteriori de que essa moeda seja viciada?
- De acordo com Bayes: $1,67 \times 10^{-3} \times 0,1$
- ----- = 0,13
- $(1,67 \times 10^{-3} \times 0,1) + (9,76 \times 10^{-4} \times 0,9)$

Algoritmo de *Metropolis-Hastings*



Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC)

Melhor topologia



MÉTODO DE MONTE CARLO

A decorative background featuring a stylized, light gray figure of a person with arms raised. Three circular icons are placed around the figure: a circle with the letter 'A' near the figure's head, a circle with the letter 'C' near the figure's right arm, and a circle with the letter 'G' near the figure's left arm.

- **O Método de Monte Carlo (MC)** pode ser descrito como um método **estatístico**, onde se utiliza uma sequência de **números aleatórios** para a realização de uma **simulação**
- Este método já era conhecido há séculos, mas começou a ser utilizado efetivamente, somente nas últimas décadas

MÉTODO DE MONTE CARLO

- O **método de Monte Carlo (MMC)** é um método **estatístico** utilizado em simulações estocásticas com diversas aplicações em áreas como a física, matemática e biologia e tem sido utilizado há bastante tempo como forma de obter aproximações numéricas de funções complexas.
- Envolve a geração de observações de alguma **distribuição de probabilidades** e o uso da amostra obtida para aproximar a função de interesse.
- A **simulação de Monte Carlo** é um processo de **amostragem** cujo objetivo é permitir a observação do desempenho de uma variável de interesse em razão do comportamento de variáveis que encerram elementos de incerteza.

MÉTODO DE MONTE CARLO

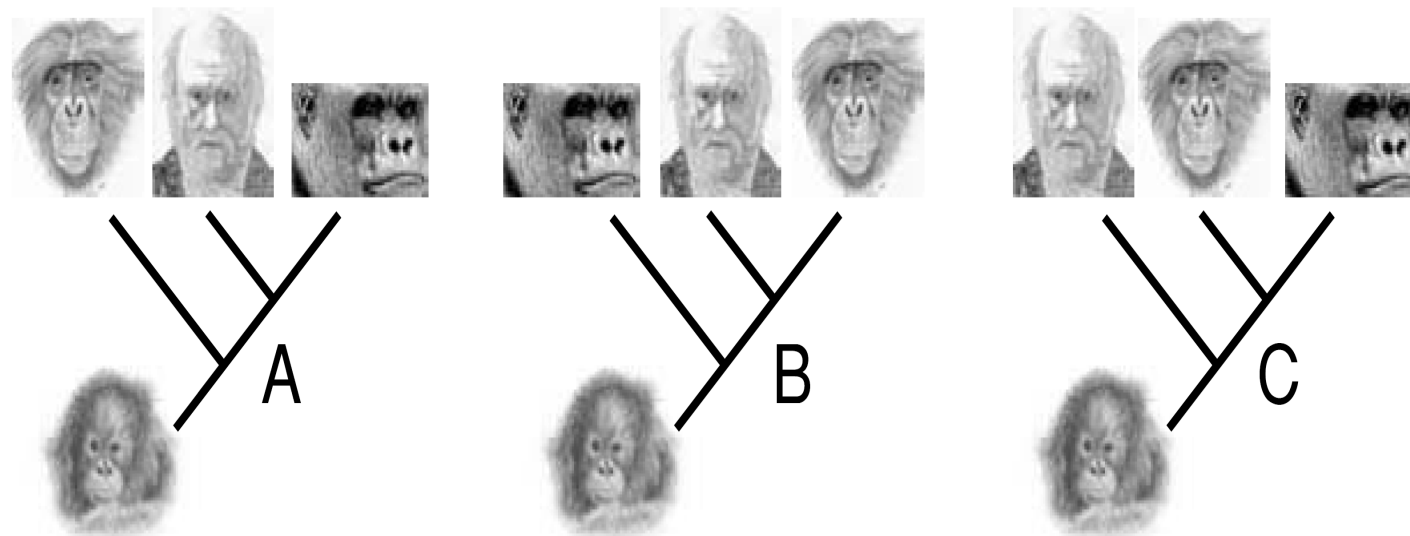
- A base para o processo de amostragem realizado nas simulações de Monte Carlo é a **geração de números aleatórios**. É a partir desse mecanismo que são produzidas as distribuições das variáveis de interesse, tomando por base as premissas e as distribuições associadas às variáveis de entrada, bem como a inter-relação entre as mesmas
- **O método de amostragem de Monte Carlo** seleciona valores aleatoriamente de forma independente de acordo com a distribuição de probabilidade definida. Em outras palavras, o número aleatório utilizado em uma rodada não influencia os próximos números aleatórios a serem utilizados.

MÉTODO DE MONTE CARLO

- **Exemplos:**
- **Estimativa da população de peixes em um lago**
- Nesse método é feita uma primeira etapa de **identificação**: são pescados diversos peixes que são marcados por meio de um anel. Nessa etapa é importante saber exatamente o número de peixes que foram identificados dessa forma.
- Na segunda etapa são pescados novamente uma certa quantidade de peixes aleatoriamente, anotando respectivamente aqueles que estavam identificados com o anel e aqueles que não estavam identificados.
- É esperado que a proporção entre peixes pescados com a identificação e o número total de peixes pescados siga a mesma proporção entre o número total de peixes com a identificação e o número total de peixes.
- **Monte Carlo e COVID-19**
- Utilizando simulação de Monte Carlo, foi possível montar uma modelagem para a dinâmica de espalhamento da COVID-19 utilizando primeiramente cenários definidos previamente e depois testando o modelo em dados da doença para a Austrália e o Reino Unido, estimando assim as datas de picos de casos para ambos os países e o número de casos, obtendo resultados consistentes com as estimativas na literatura. Tal modelo foi considerado efetivo como uma ferramenta para tomadas de decisões no combate à COVID-19 e para outras doenças que possam surgir no futuro

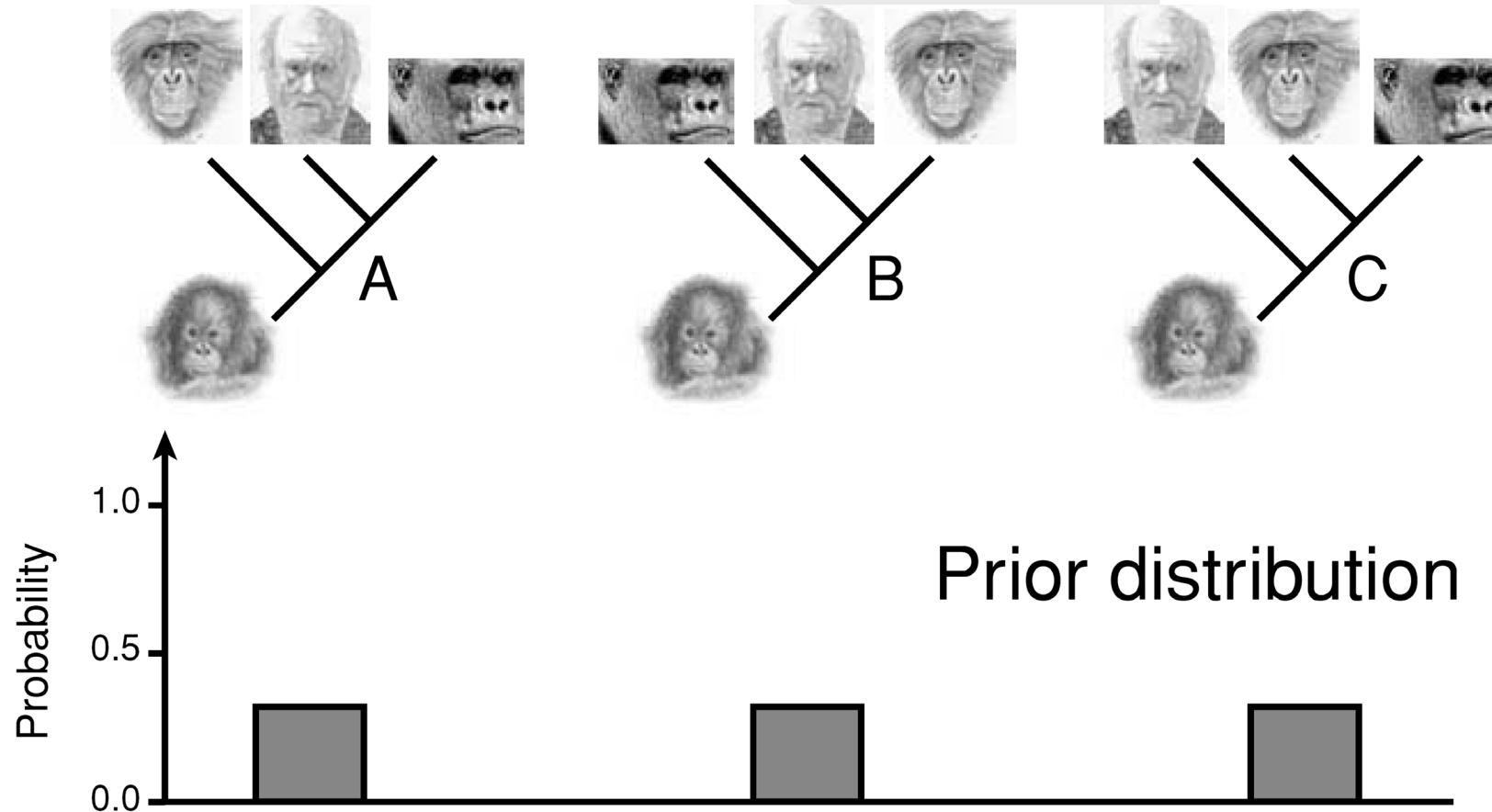
Análise filogenética por Inferência Bayesiana

- Como aplicar Inferência Bayesiana para Análise Filogenética?
- Exemplo:
 - Relação entre humanos, gorilas e chimpanzés
 - Orangotango para enraizar a árvore
 - Três árvores possíveis:
 - A: Chimpanzés e humanos; B: Gorila e humanos; C: Chimpanzés e gorilas



Análise filogenética por Inferência Bayesiana

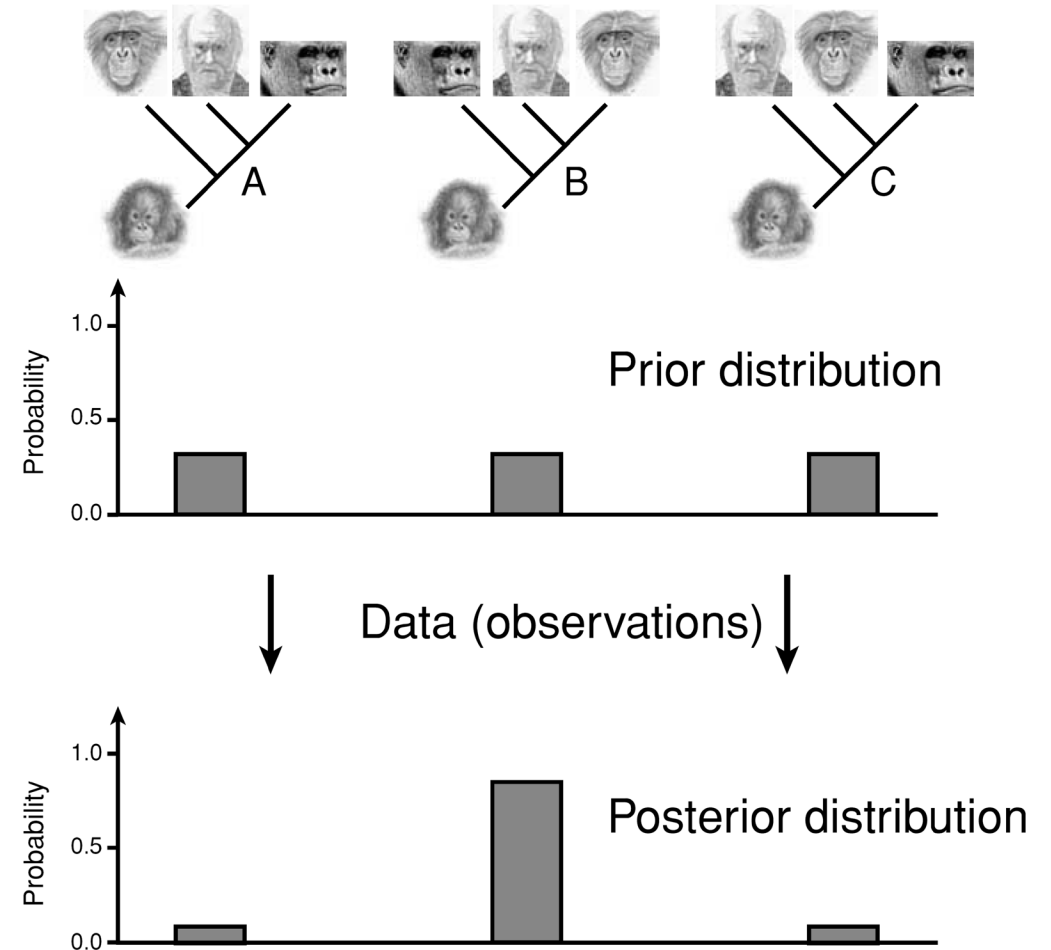
- Sem informações prévias, definimos que todas tem a mesma probabilidade
- Distribuição da probabilidade a priori é não informativa



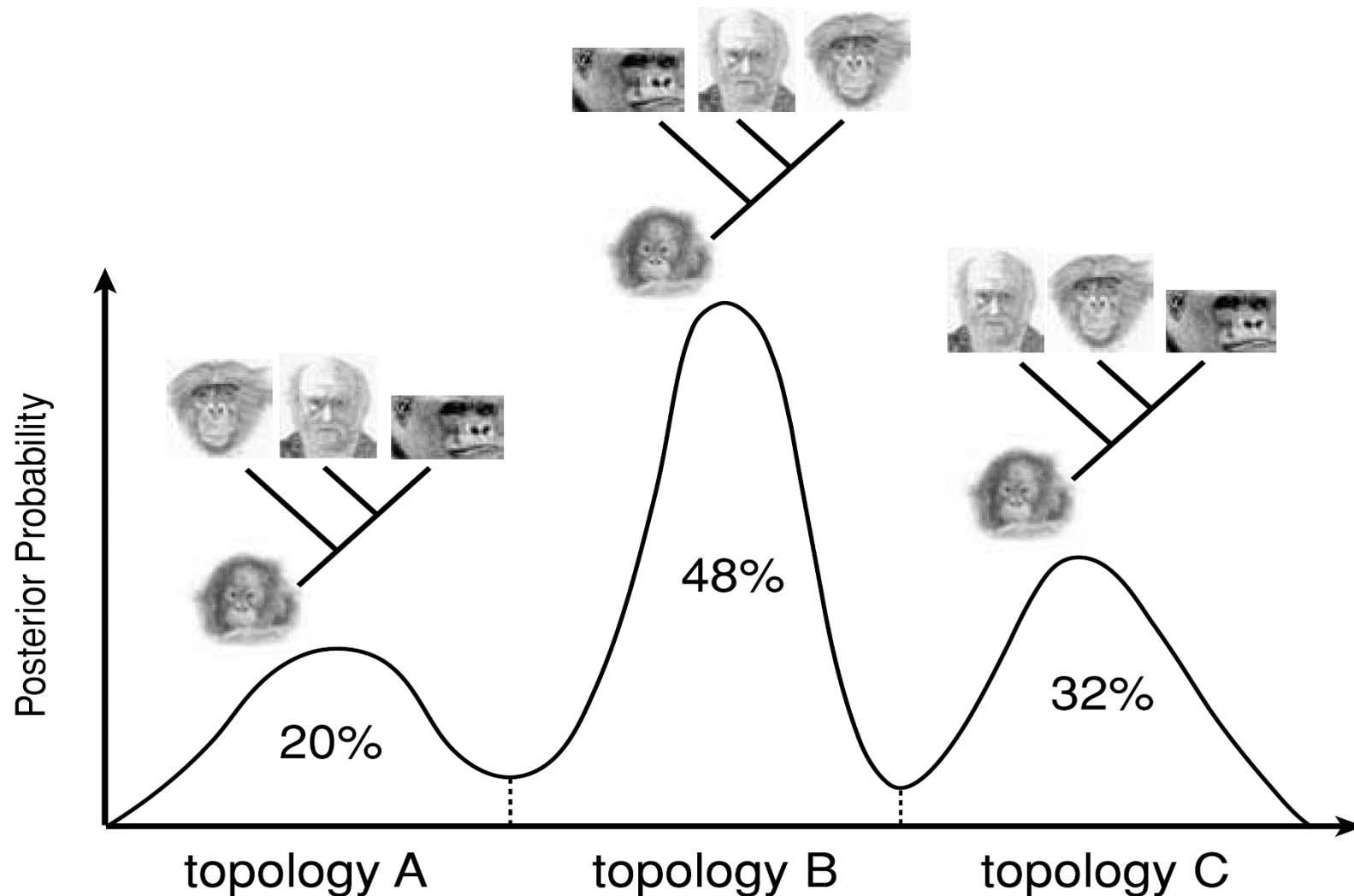
Análise filogenética por Inferência Bayesiana

Usar os dados de alinhamento e um modelo evolutivo para gerar as probabilidades a posteriori (PP)

PP: Probabilidade de cada árvore, dado o modelo, a P a priori e os dados

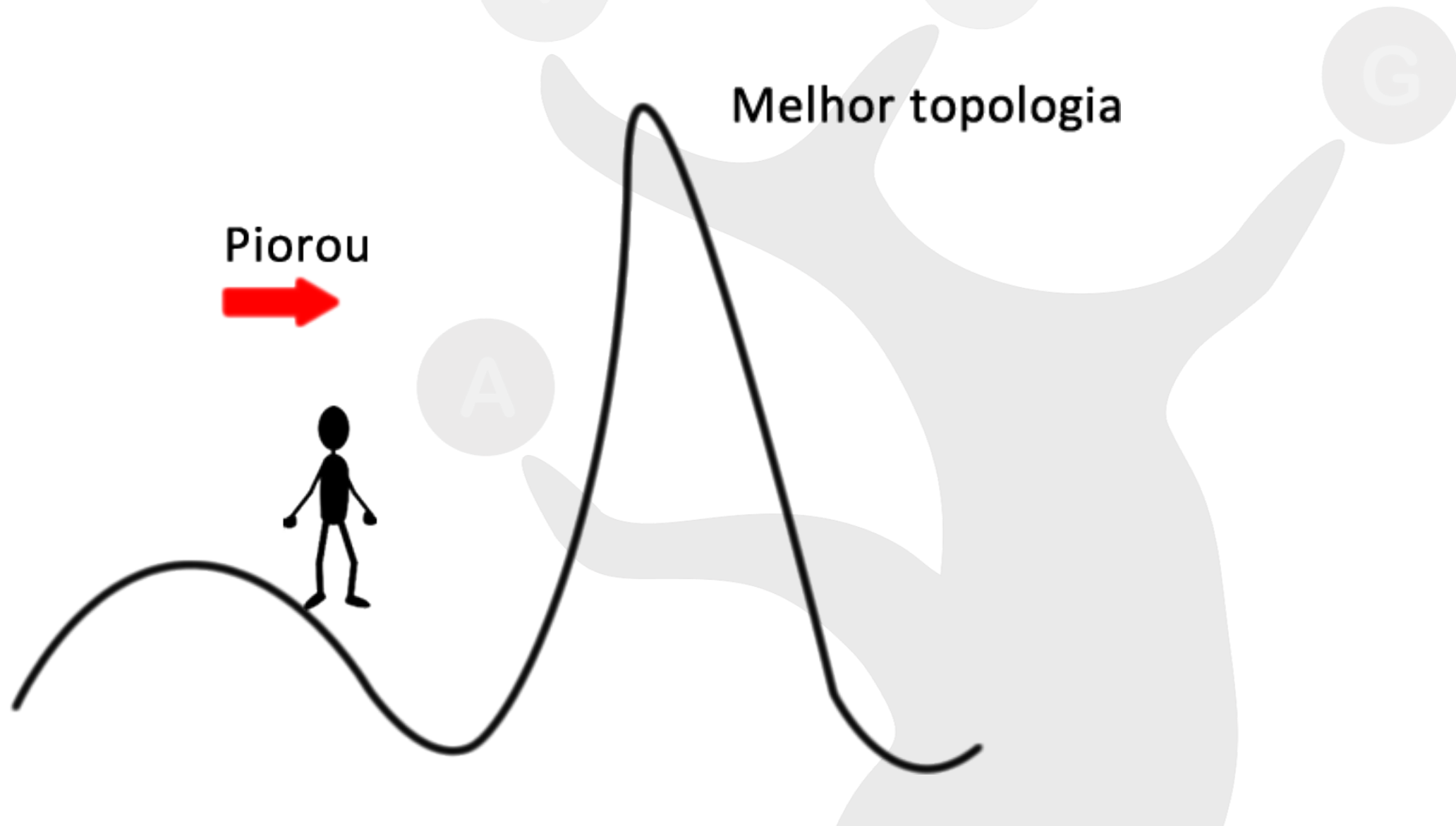


Análise filogenética por Inferência Bayesiana



Algoritmo de *Metropolis-Hastings*

Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC)

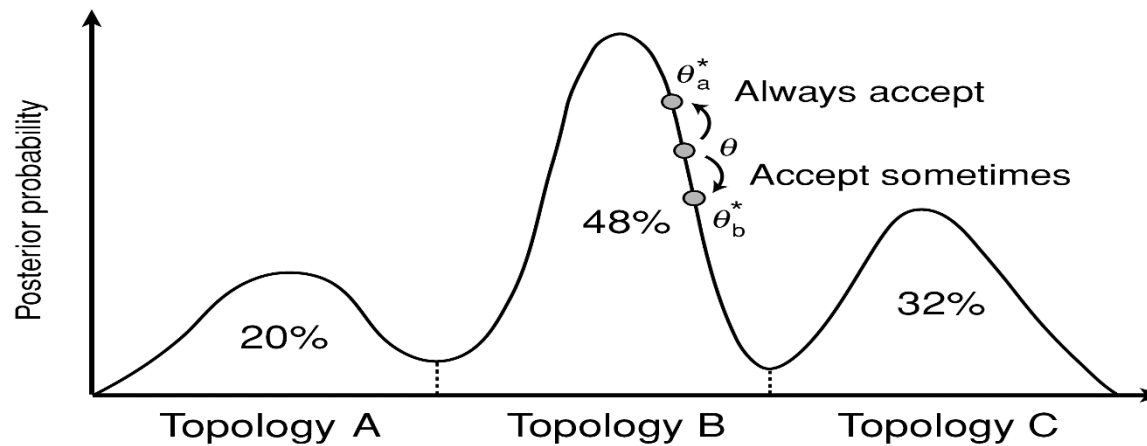


Algoritmo de *Metropolis-Hastings*

Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC)

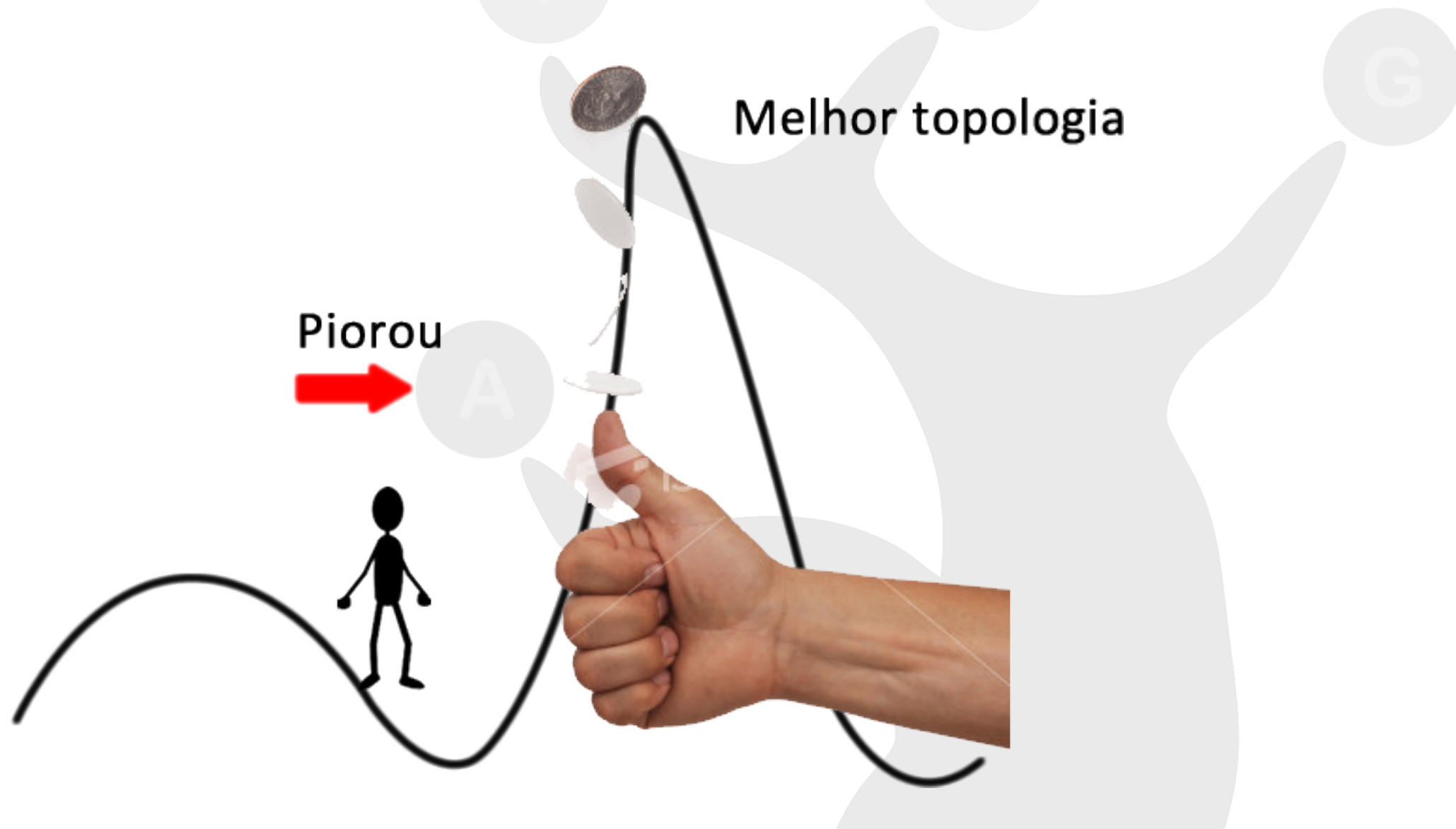
Markov chain Monte Carlo steps

1. Start at an arbitrary point (θ)
2. Make a small random move (to θ^*)
3. Calculate height ratio (r) of new state (to θ^*) to old state (θ)
 - (a) $r > 1$: new state accepted
 - (b) $r < 1$: new state accepted with probability r
if new state rejected, stay in old state
4. Go to step 2



Algoritmo de *Metropolis-Hastings*

Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC)

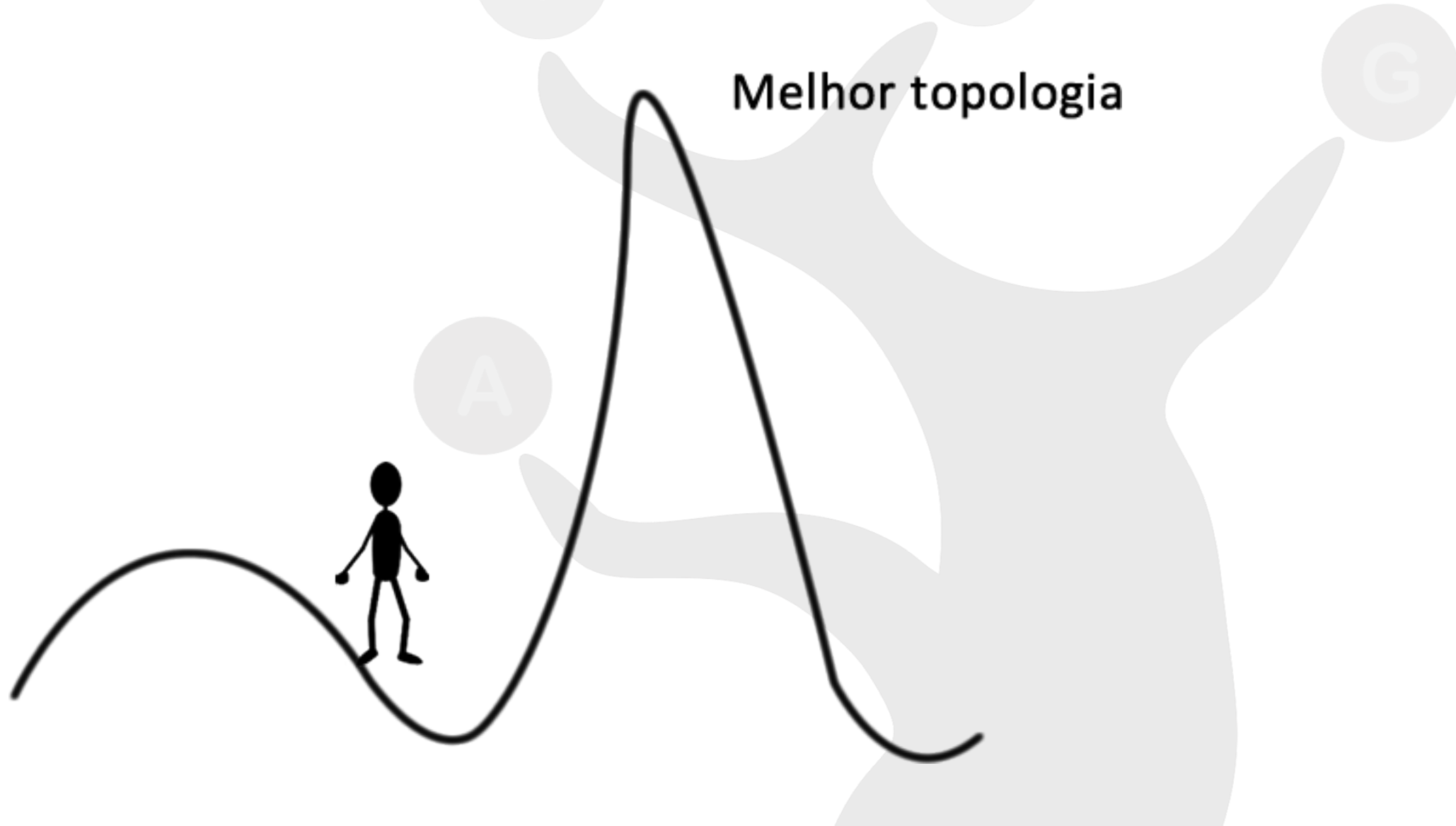


Algoritmo de *Metropolis-Hastings*



Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC)

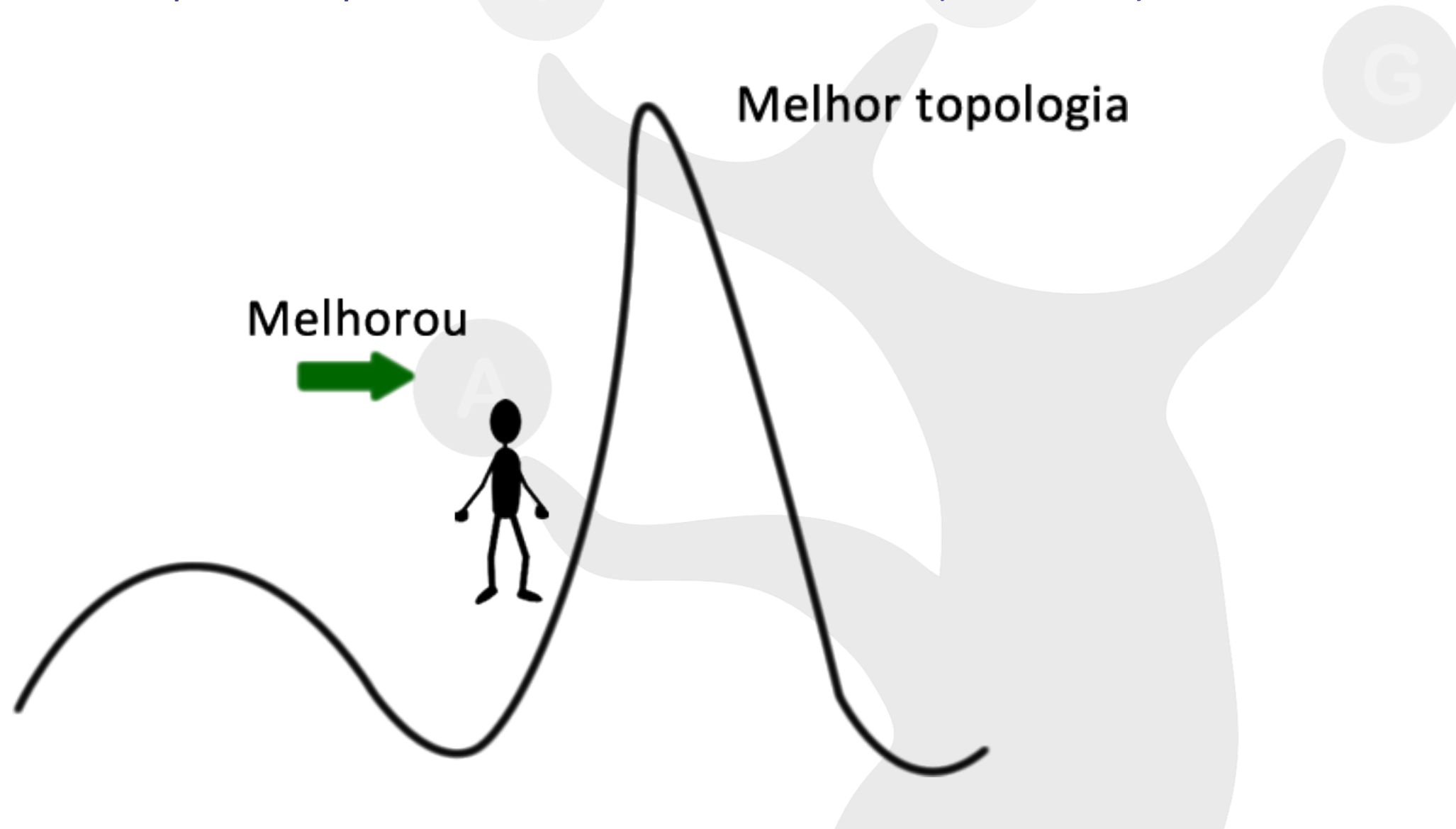
Melhor topologia



Algoritmo de *Metropolis-Hastings*



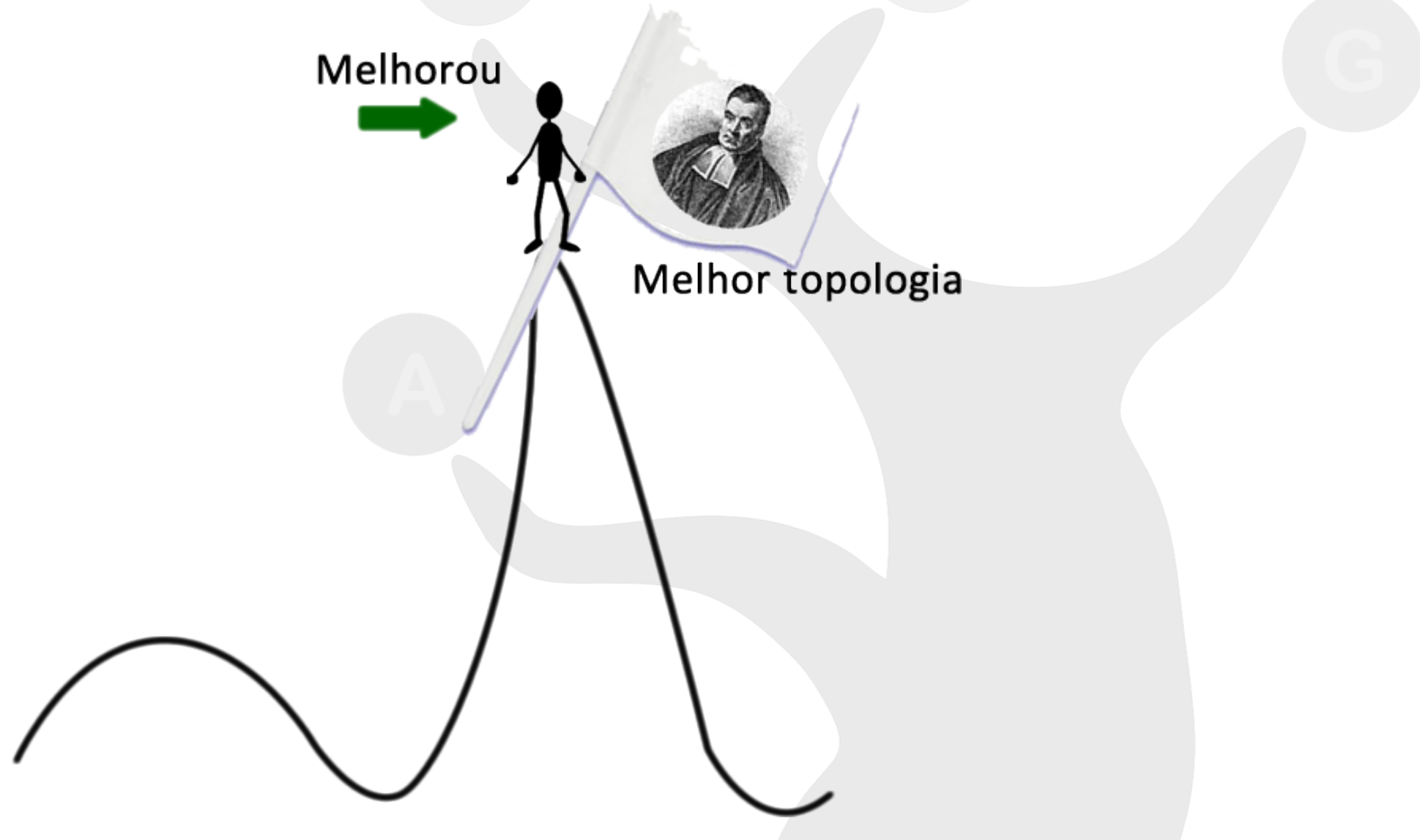
Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC)



Algoritmo de *Metropolis-Hastings*



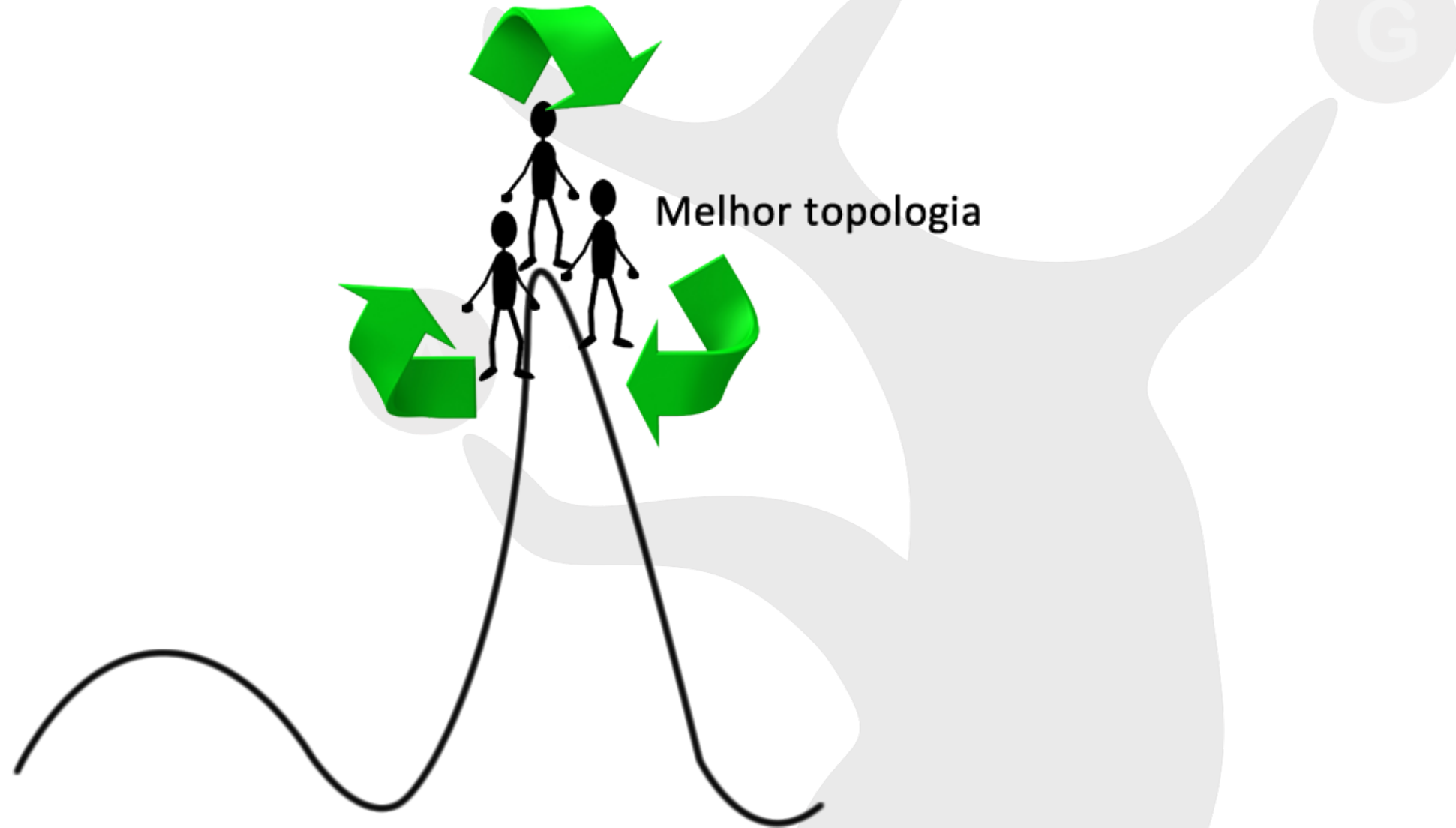
Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC)



Algoritmo de *Metropolis-Hastings*



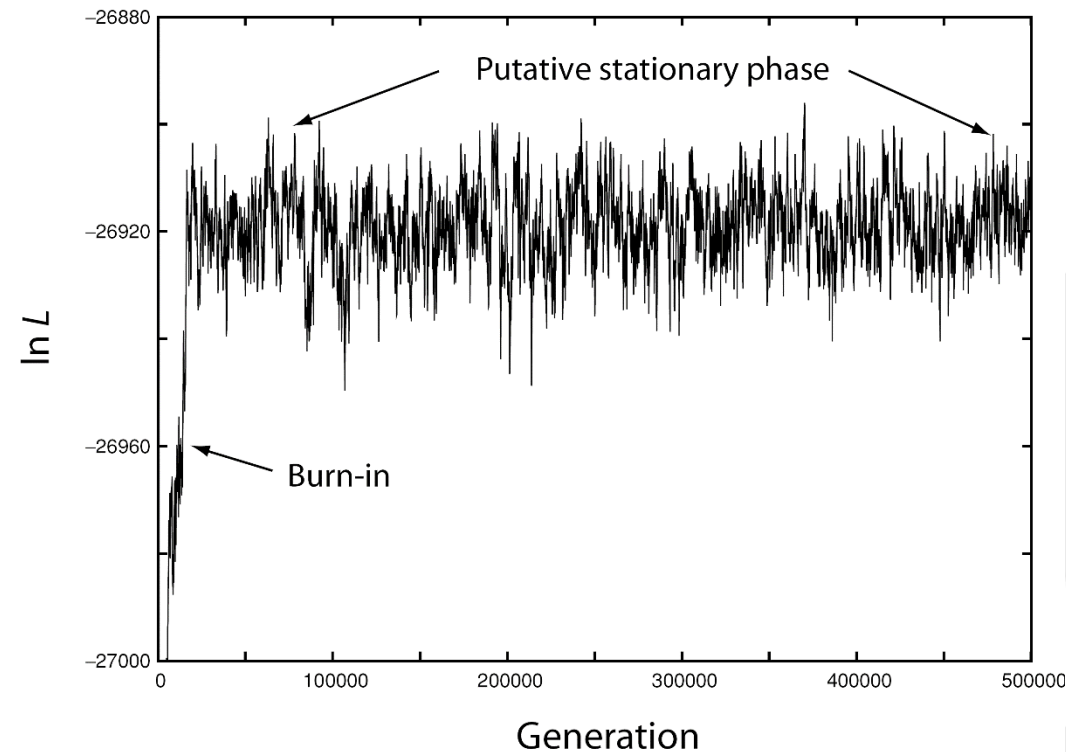
Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC)



Análise filogenética por Inferência Bayesiana - Cadeia, geração e convergência

- Método Bayesiano é aplicado a cada sítio do alinhamento
- A análise inicia com uma árvore aleatória (ou especificada pelo usuário), que é o estado inicial da cadeia:
 - Combinação de comprimento dos ramos
 - Parâmetros de substituição
 - Taxas de variação
- A cada geração, um novo estado da cadeia é proposto, e será aceito ou rejeitado, dependendo da probabilidade do novo estado, dado o estado anterior
- Novo estado da cadeia: mover um ramo, e/ou alterar o comprimento do ramo (cria uma nova árvore)
- Aumento do número de gerações: a análise gera um conjunto de árvores com probabilidades semelhantes – aceitar ou rejeitar uma alteração é quase aleatório: **CONVERGÊNCIA**

Análise filogenética por Inferência Bayesiana - Burn in



Burn in: Fase inicial da cadeia de Markov - Os valores de probabilidades aumentam rapidamente na fase inicial da análise, porque o ponto inicial geralmente é muito longe dos valores de maior probabilidade

Burn-in: Descarte de amostras iniciais – 25%

Análise filogenética por Inferência Bayesiana – Cadeias independentes

- Cadeia fria e cadeias (3) quentes: Estratégia utilizada para evitar que a análise fique presa em uma colina que não é a mais alta (como pode ocorrer em ML):
 - Roda 4 cadeias independentes, iniciando com árvores diferentes
 - As cadeias rapidamente divergem umas das outras
 - A cada geração, as cadeias podem trocar de estado (vai ocorrendo permuta entre elas – se uma cadeia ficar presa em uma colina menor, ela poderá ser retirada de lá trocando com outra que está em uma colina maior)
 - Após normalmente 100 gerações, a árvore obtida pela cadeia fria será salva