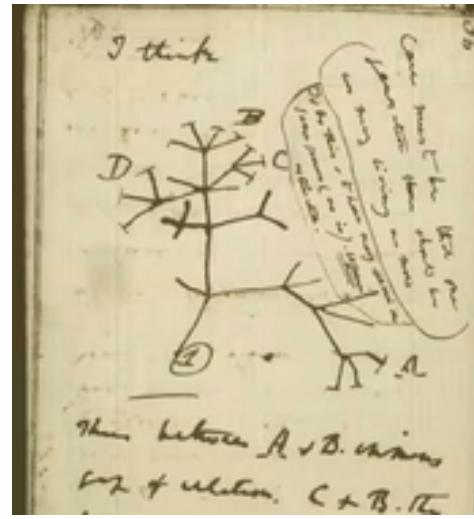


Filogenética molecular aplicada

1. Introdução



Plano de aula

- Filogenética: Principais aplicações
- Ancestralidade comum e homologia
- Terminologia
- Visão geral dos principais métodos de inferência de árvores
- Inferindo uma árvore I (métodos baseados em distância)
- Inferindo uma árvore II (métodos baseados em caracteres: máxima parcimônia)

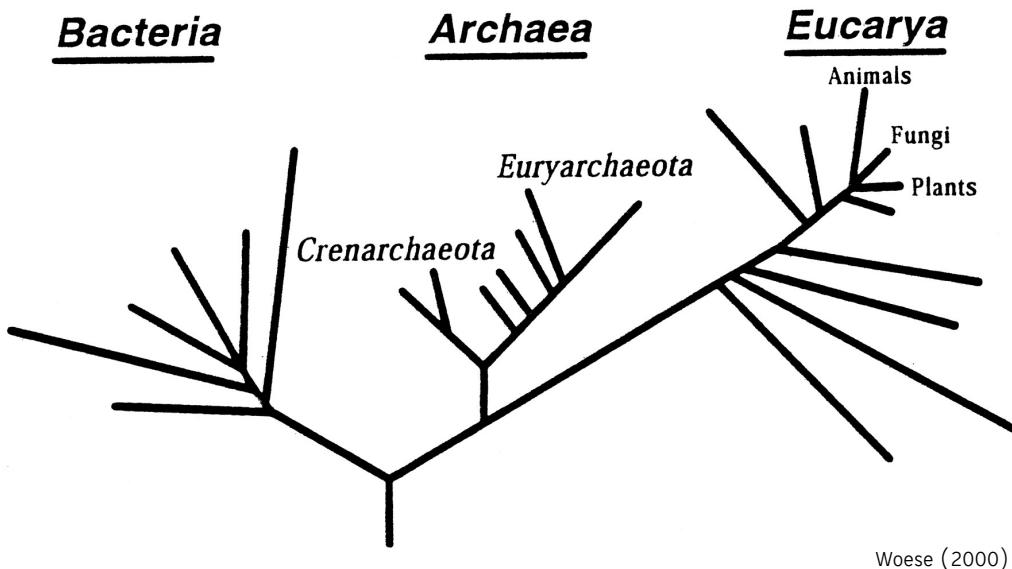
Prática: conjunto de ferramentas básico para análises filogenéticas; inferindo árvores com métodos baseados em distância e máxima parcimônia

Por que inferir árvores filogenéticas?

- Base da classificação biológica moderna
- Essencial para estudos comparativos entre espécies/grupos
- Ferramenta de predição de características biológicas

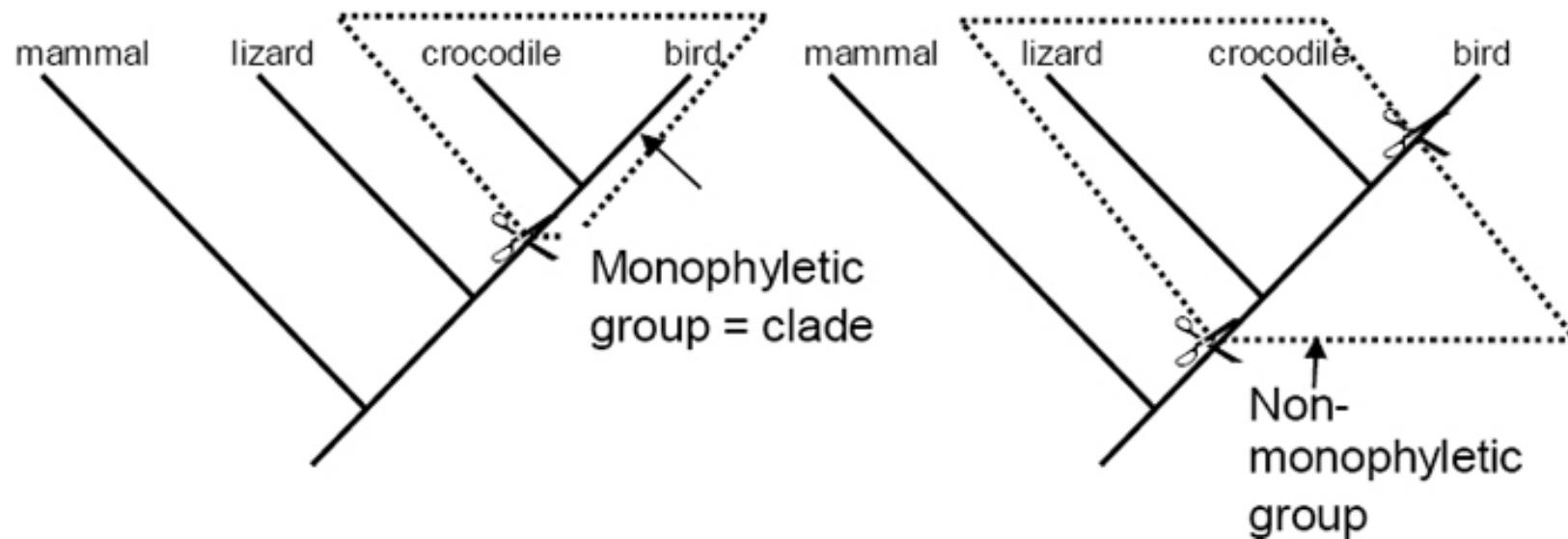
A árvore da vida

Reconstruindo a história da vida no planeta



Filogenias são a base da classificação biológica moderna

Táxons naturais = grupos monofiléticos (clados)

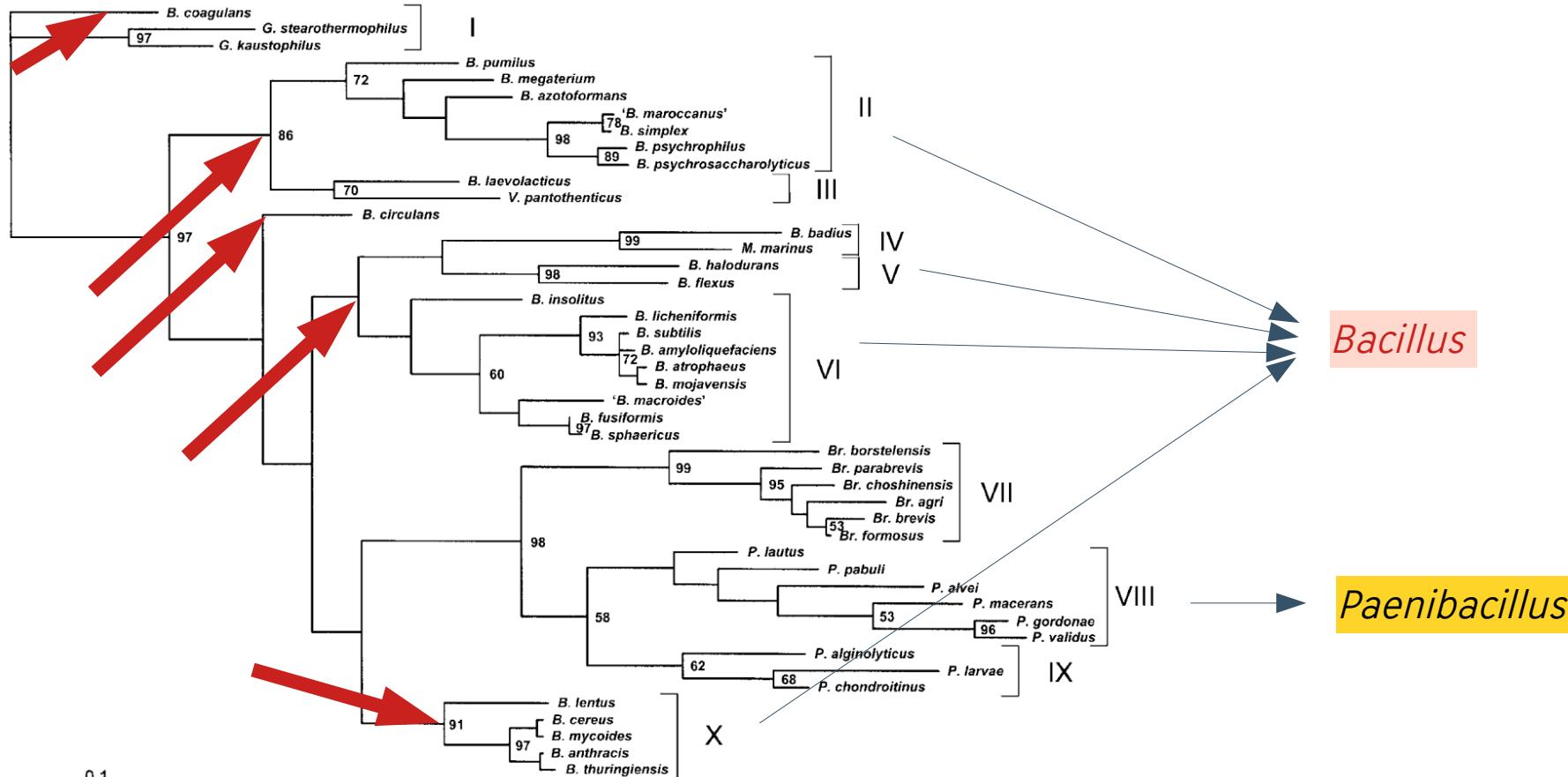


Filogenias são a base da classificação biológica moderna

Árvores filogenéticas fornecem evidência para revisões taxonômicas

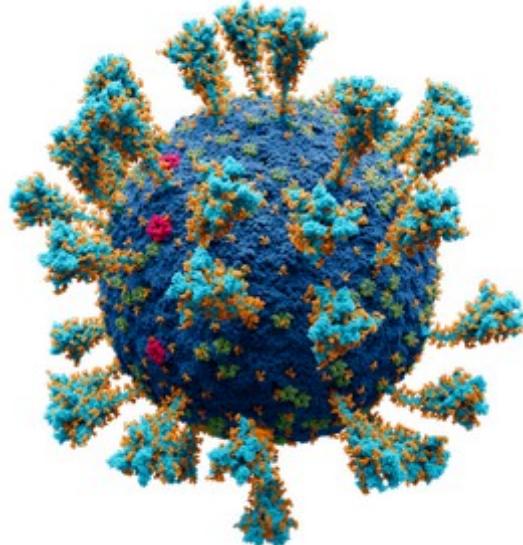
Filogenias são a base da classificação biológica moderna

Árvores filogenéticas fornecem evidência para revisões taxonômicas



Filogenias permitem testar hipóteses evolutivas

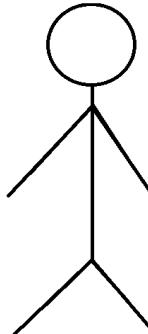
Adaptação ou coincidência? Correlações ecológico-evolutivas



Isoforma X

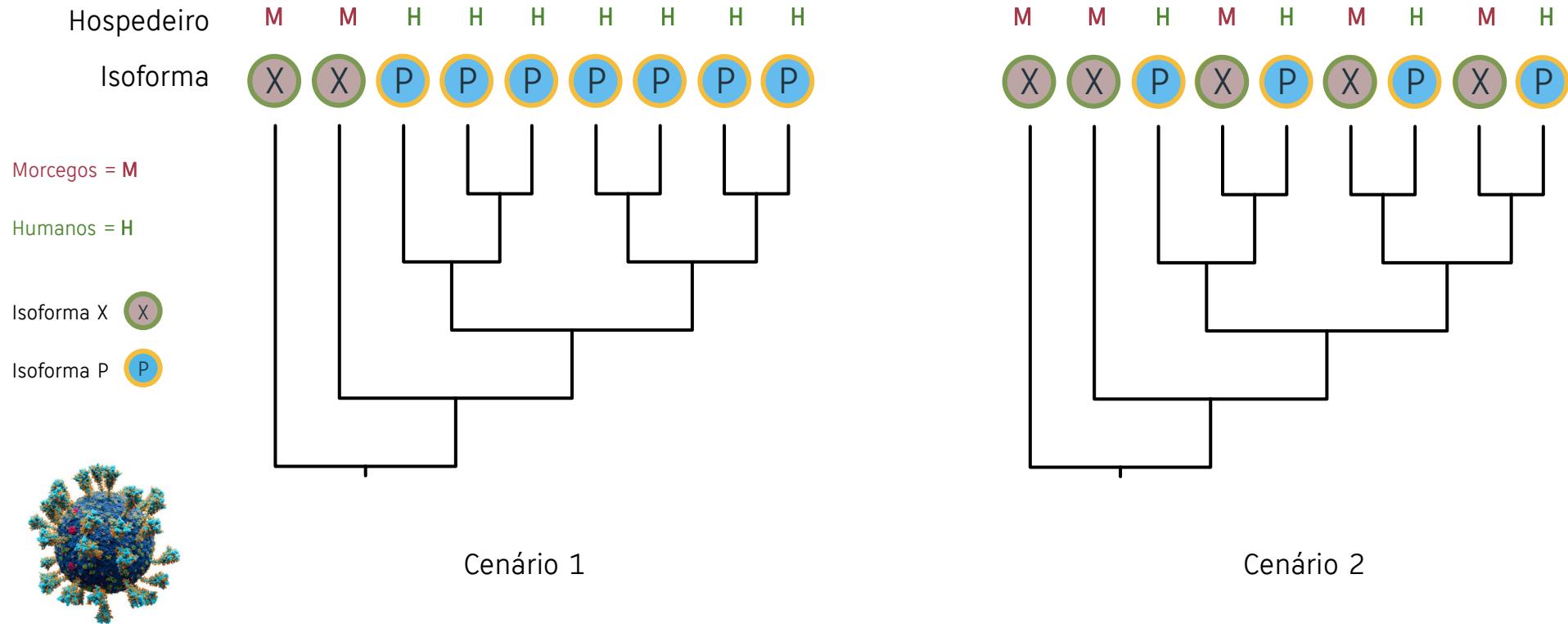


Isoforma P



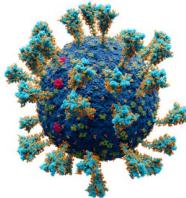
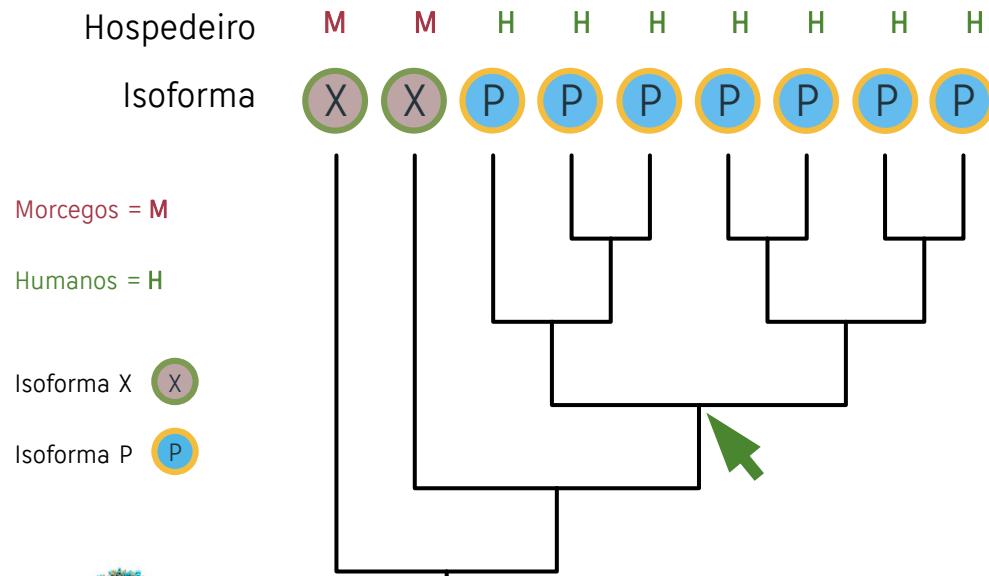
Filogenias permitem testar hipóteses evolutivas

Adaptação ou coincidência? Correlações ecológico-evolutivas

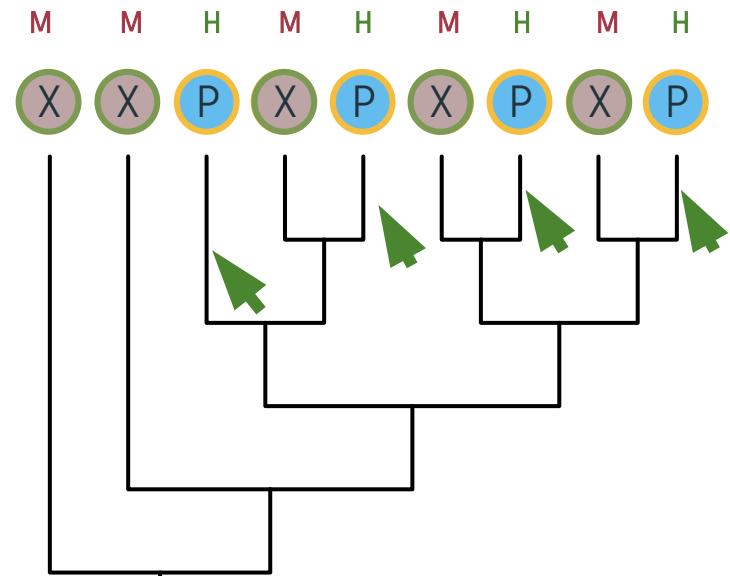


Filogenias permitem testar hipóteses evolutivas

Adaptação ou coincidência? Correlações ecológico-evolutivas



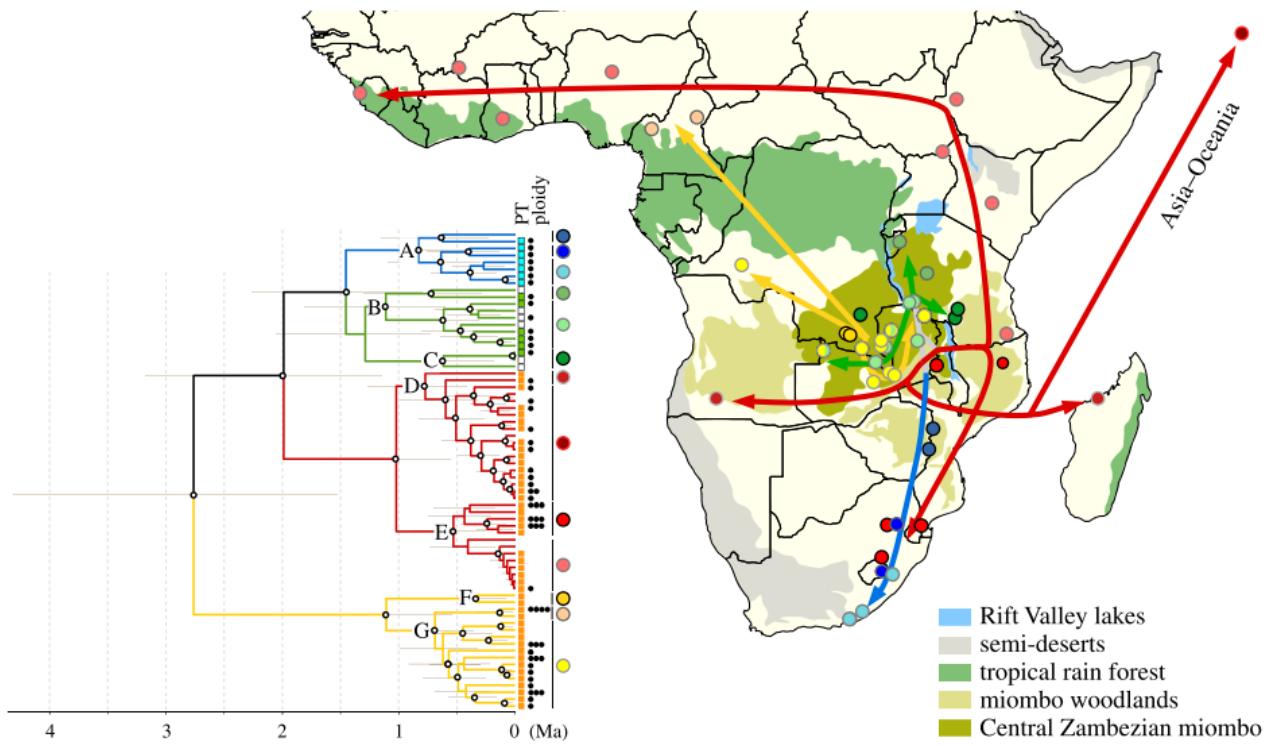
Cenário 1



Cenário 2

História da dispersão de espécies pelo planeta

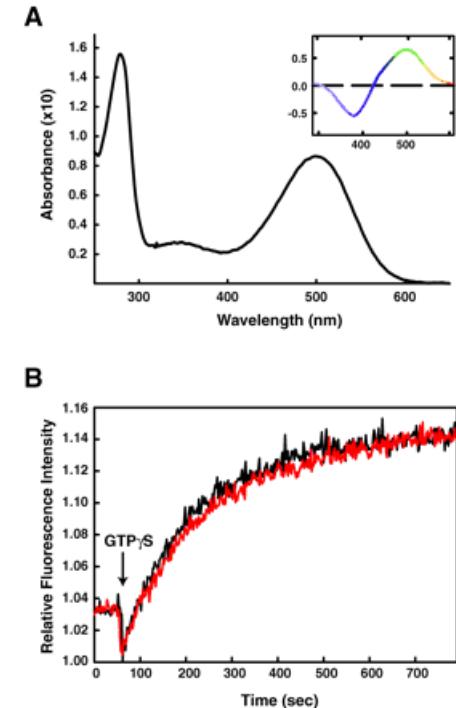
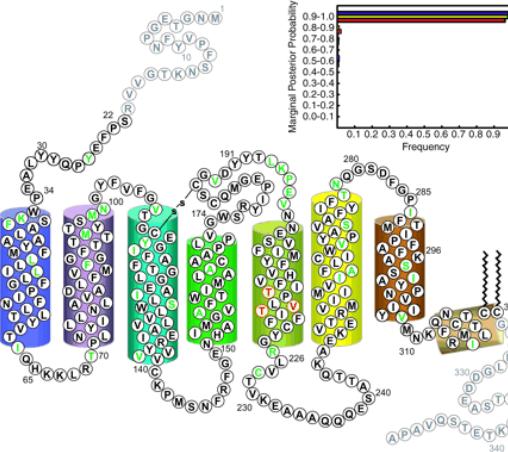
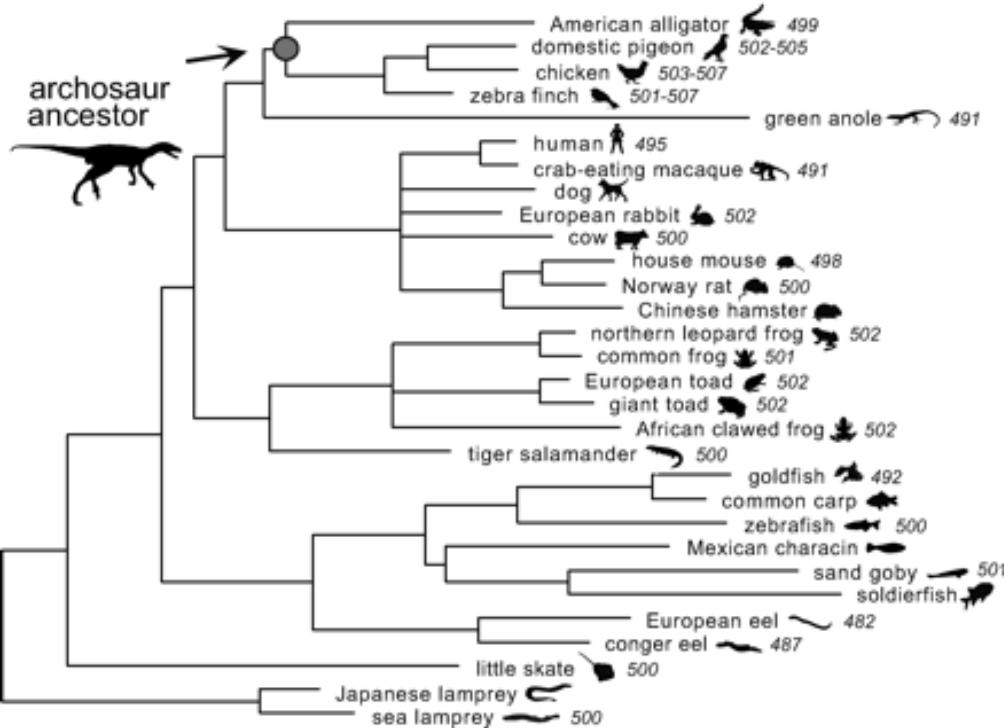
Filogeografia: reconstruindo a história de uma espécie através de filogenias



Ressuscitando proteínas de grupos extintos



Reconstrução de sequências ancestrais



Identificação biológica de amostras desconhecidas



<https://edition.cnn.com/cnn-underscored/reviews/best-plant-identification-app>

Identificação biológica de amostras desconhecidas

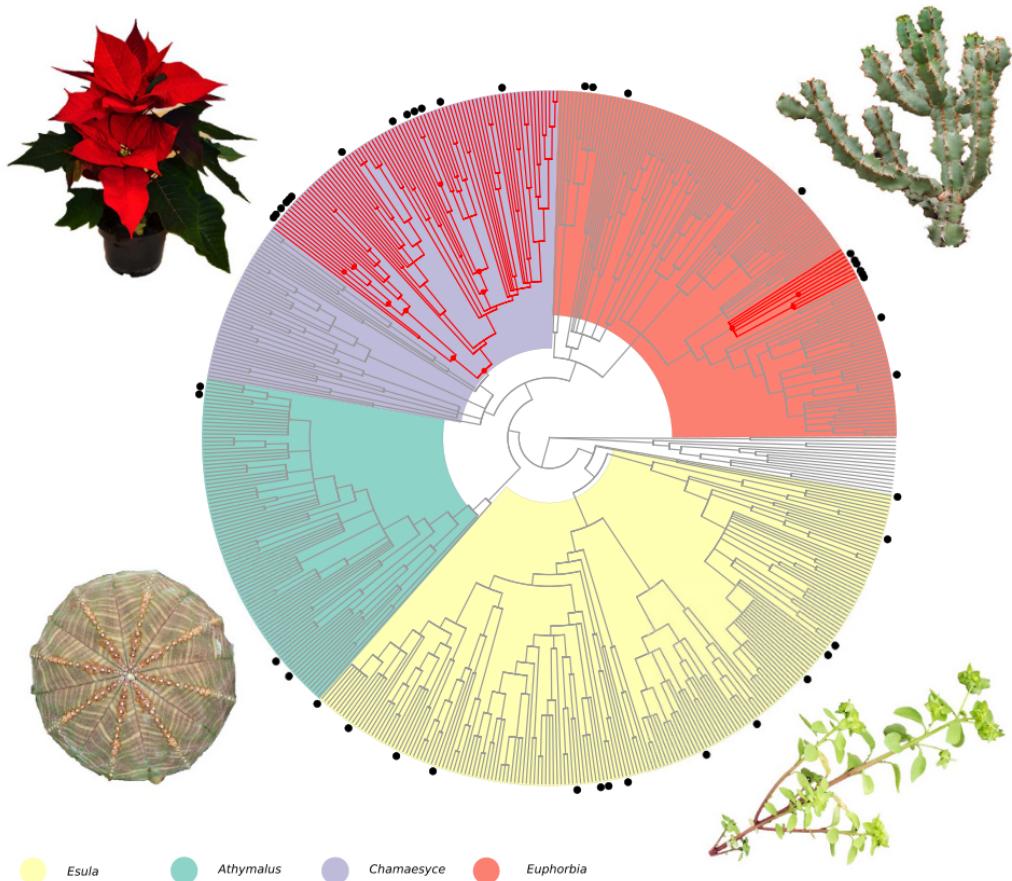


Identificação biológica de amostras desconhecidas



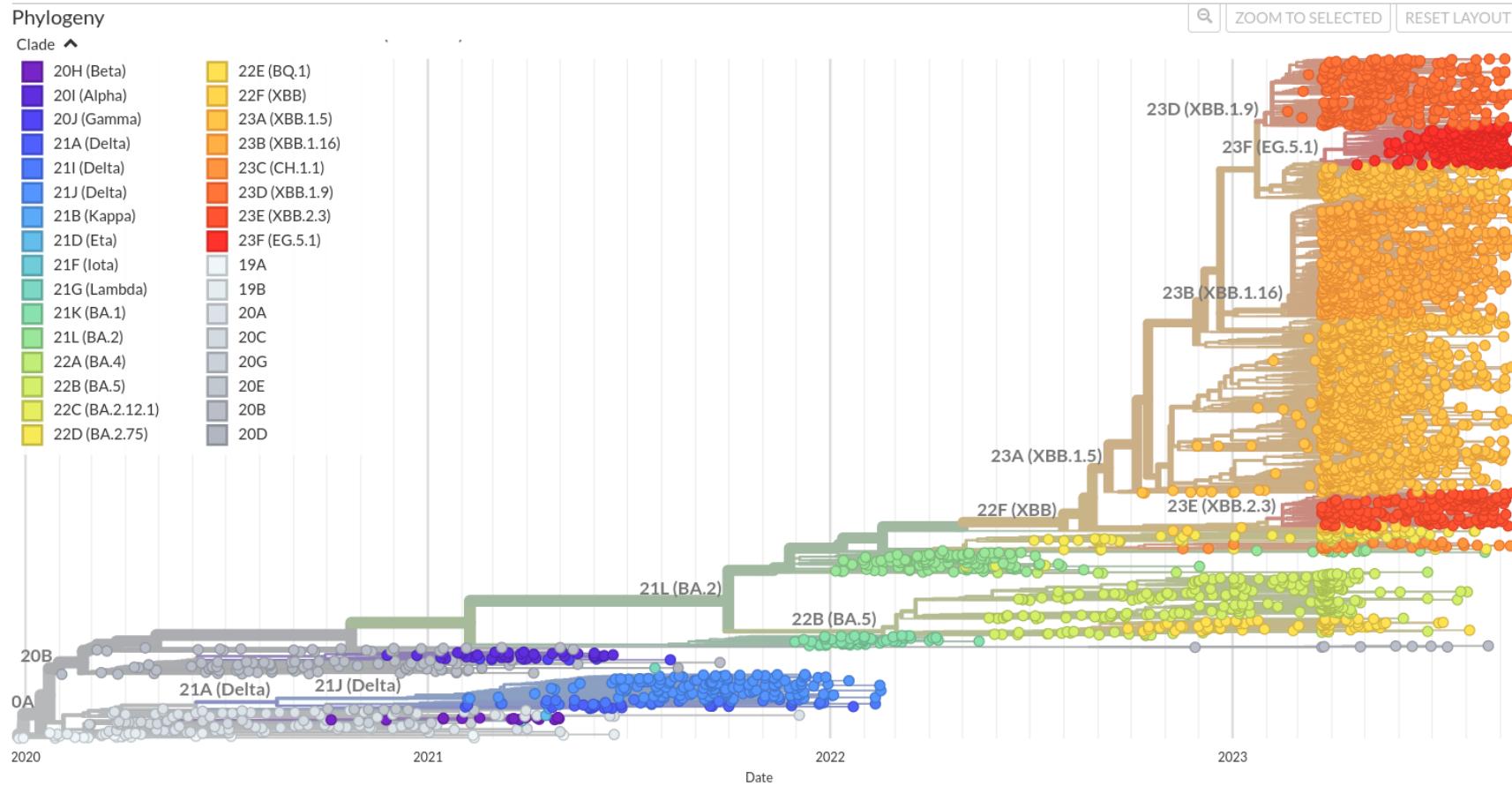
Predição de características biológicas

Prospecção de moléculas de interesse, como fármacos



- Espécies conhecidamente produtoras de substâncias antiinflamatórias

Monitoramento de epidemias



Hadfield et al. (2018)

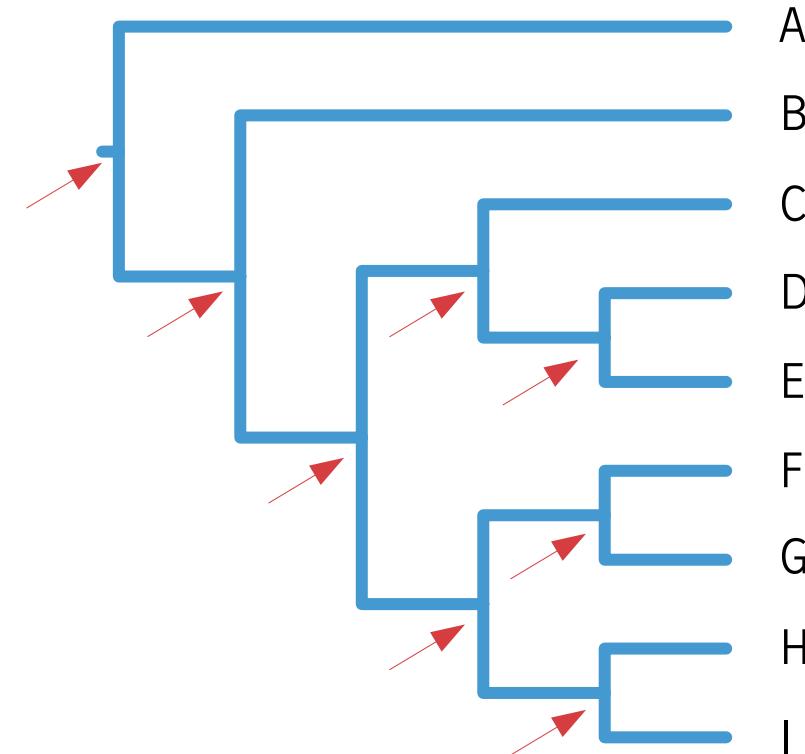
<https://nextstrain.org/>

O que são árvores filogenéticas?

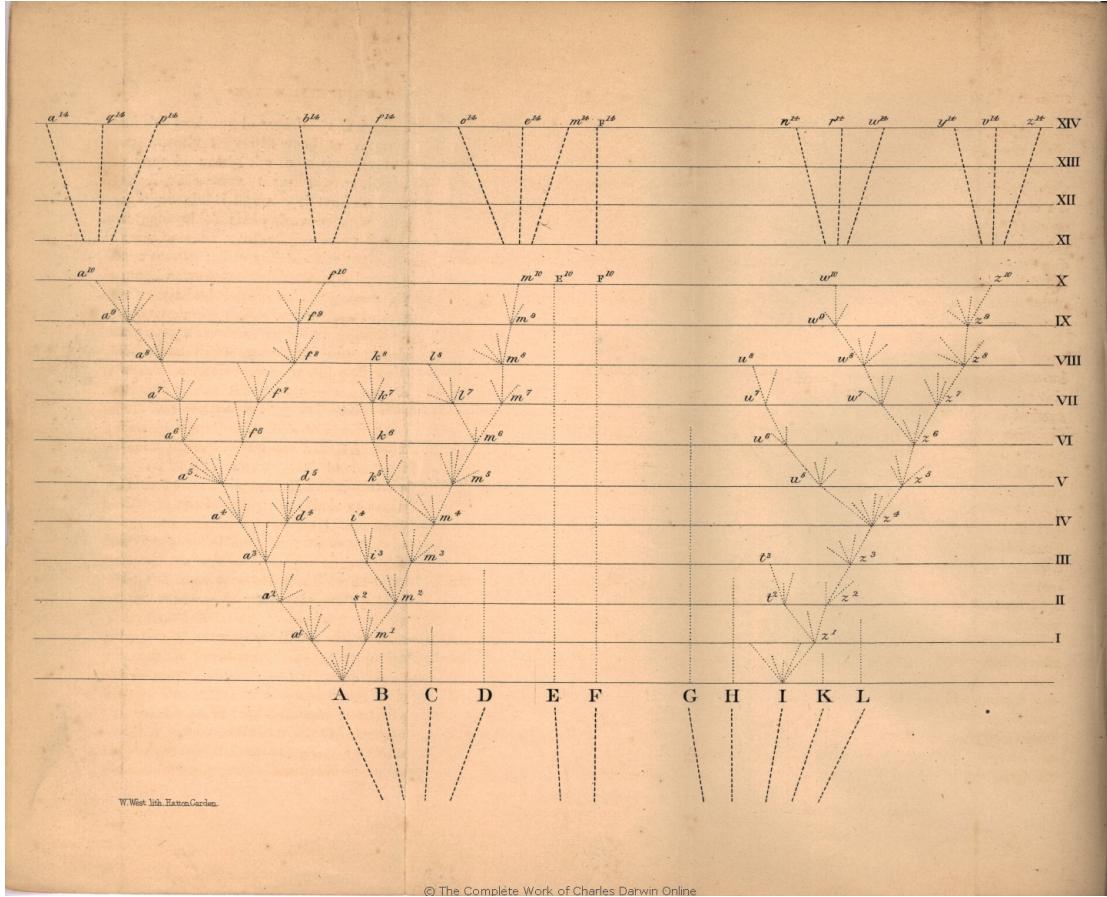
- O conceito de ancestralidade comum
- Terminologia

Representação das relações evolutivas entre táxons

- Indicam sucessivos ancestrais comuns compartilhados por linhagens distintas
- Ordenam eventos de separação entre linhagens ao longo da evolução
- Linhagens passam a evoluir independentemente após a separação

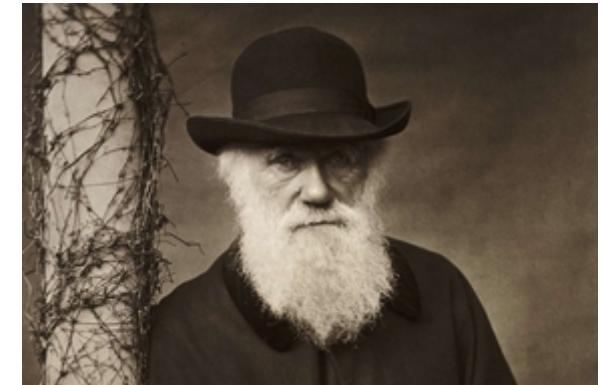


Ancestralidade comum



“Há grandeza nessa visão da vida, com suas inúmeras capacidades, ter sido originalmente respirada por algumas poucas formas ou apenas uma, (...) e que a partir de um começo simples infinitas formas de grande beleza evoluíram e continuam evoluindo.”

Charles Darwin ‘Origem das espécies’ (1859)



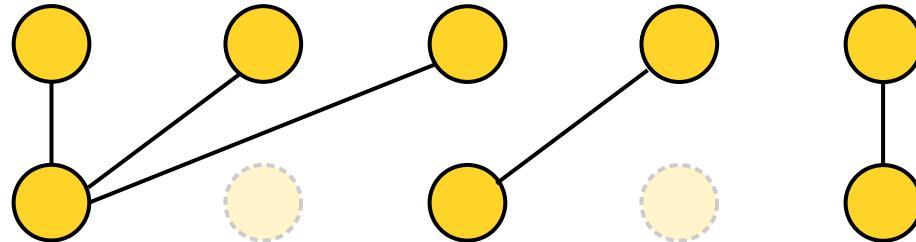
Ancestralidade comum

Filogenias são o resultado inevitável da hereditariedade



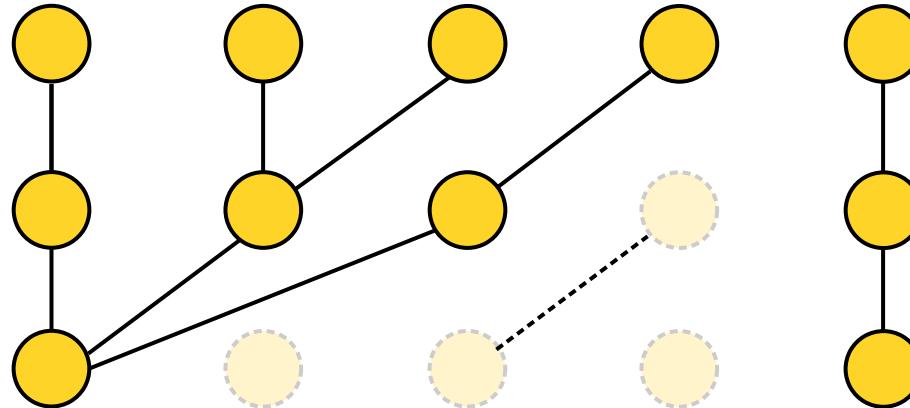
Ancestralidade comum

Filogenias são o resultado inevitável da hereditariedade



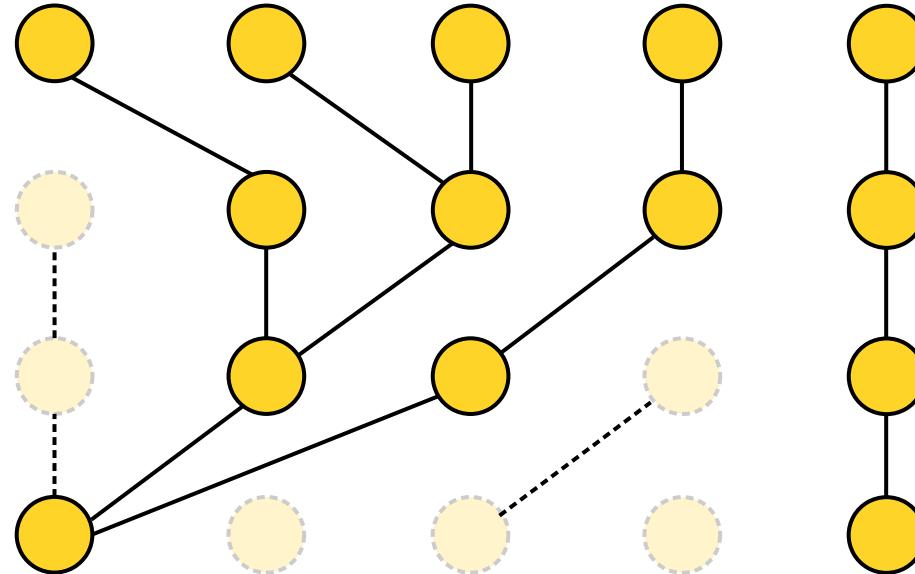
Ancestralidade comum

Filogenias são o resultado inevitável da hereditariedade



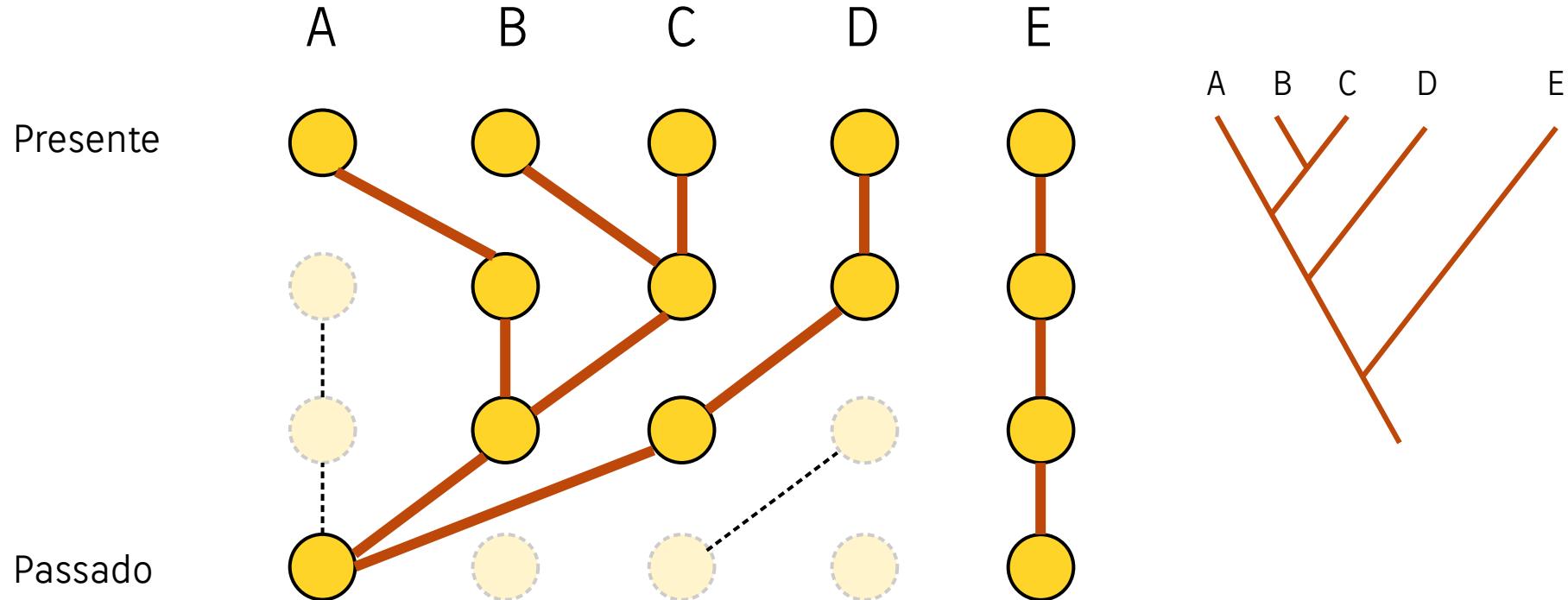
Ancestralidade comum

Filogenias são o resultado inevitável da hereditariedade



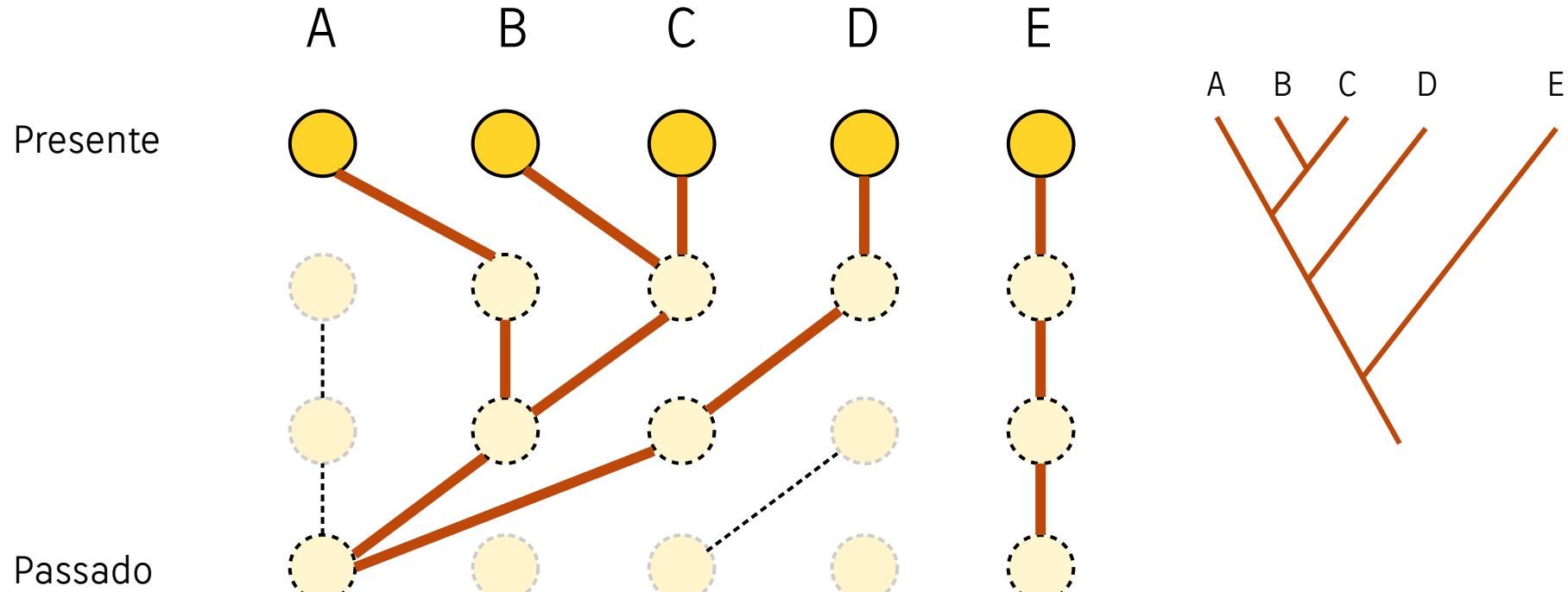
Ancestralidade comum

Filogenias são o resultado inevitável da hereditariedade



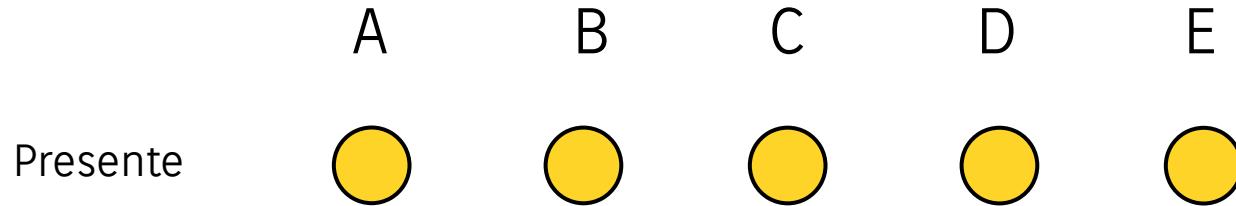
Ancestralidade comum

Não há registro sobre os ancestrais e as trajetórias de cada linhagem...



Ancestralidade comum

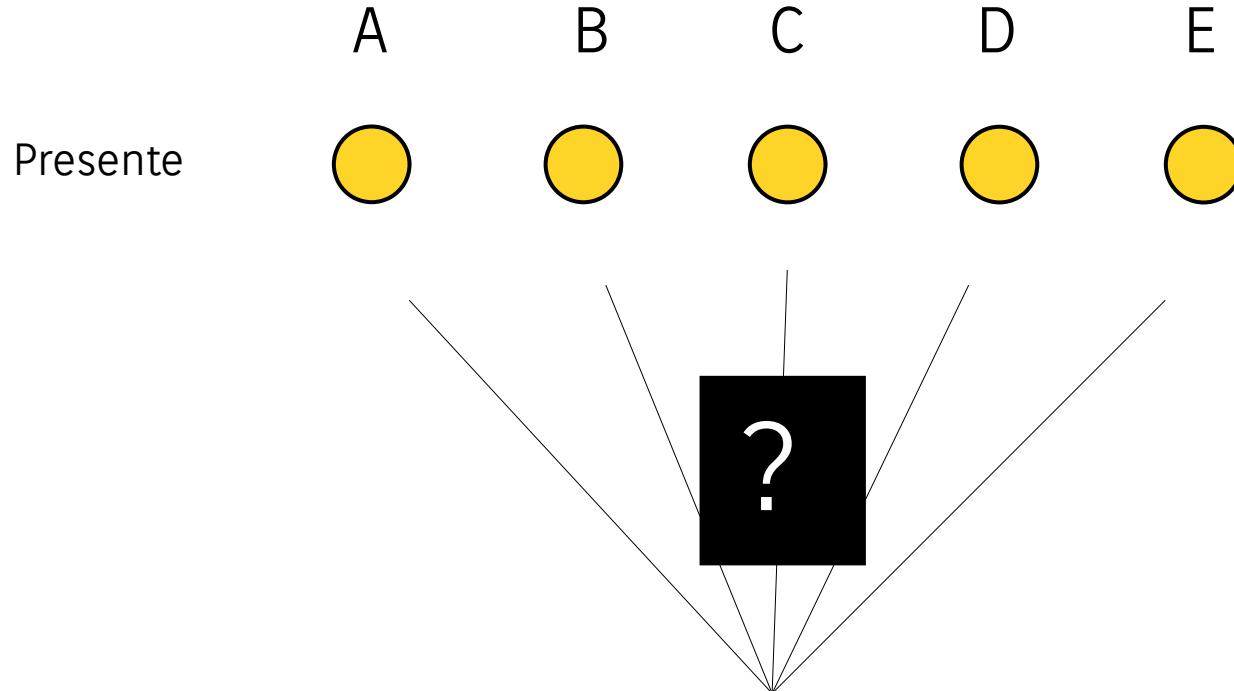
Em geral, o que temos são apenas as espécies viventes



Ancestralidade comum

Em geral, o que temos são apenas as espécies viventes

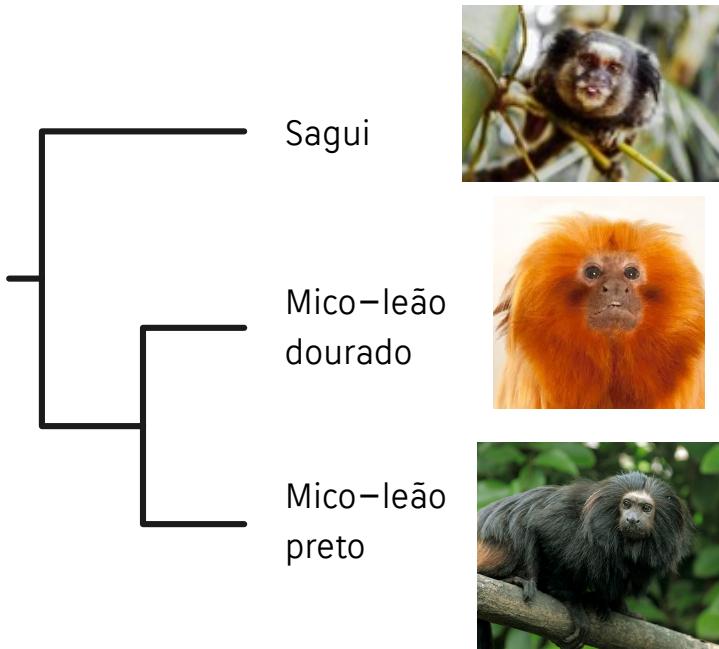
O objetivo é reconstruir a história a partir do presente



Árvores podem representar espécies ou famílias gênicas

- Árvores de espécies ('species trees') representam eventos de especiação

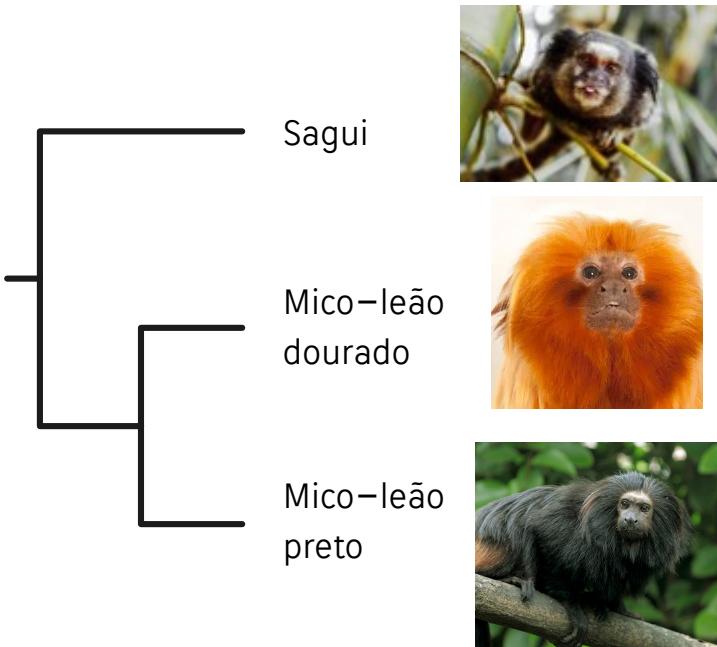
Gene G1



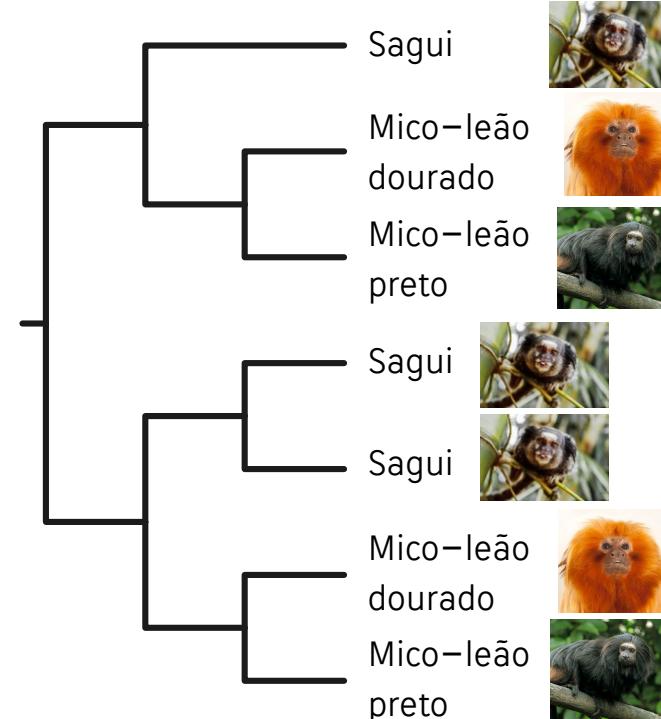
Árvores podem representar espécies ou famílias gênicas

- Árvores de espécies ('species trees') representam eventos de especiação
- Árvores gênicas ('gene trees') representam eventos de especiação ou duplicação gênica

Gene G1



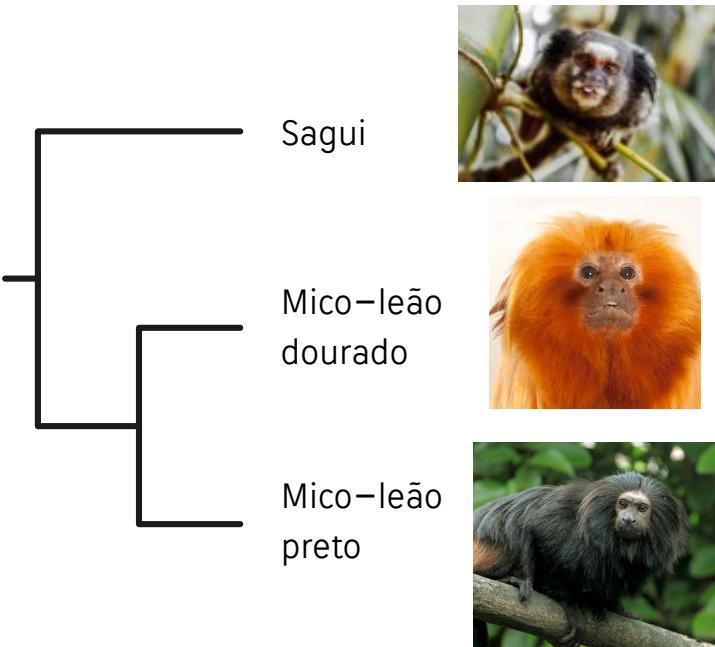
Gene E1



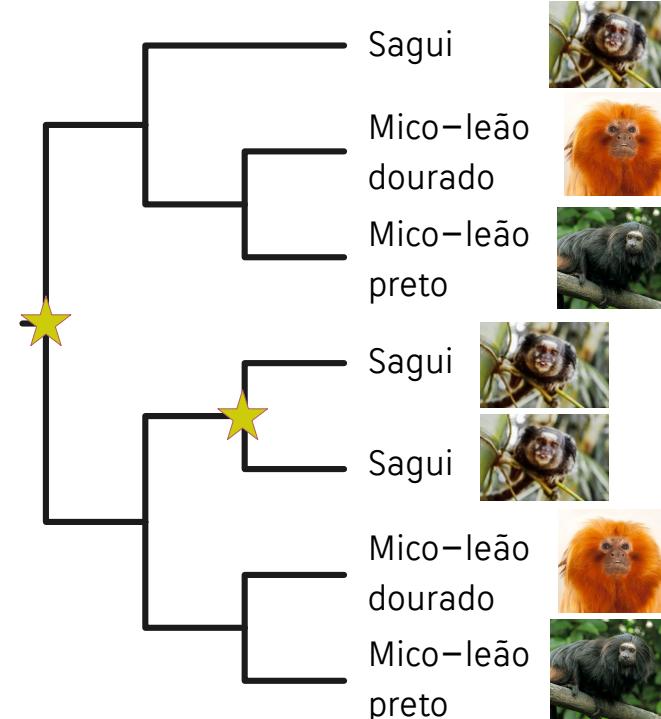
Árvores podem representar espécies ou famílias gênicas

- Árvores de espécies ('species trees') representam eventos de especiação
- Árvores gênicas ('gene trees') representam eventos de especiação ou duplicação gênica

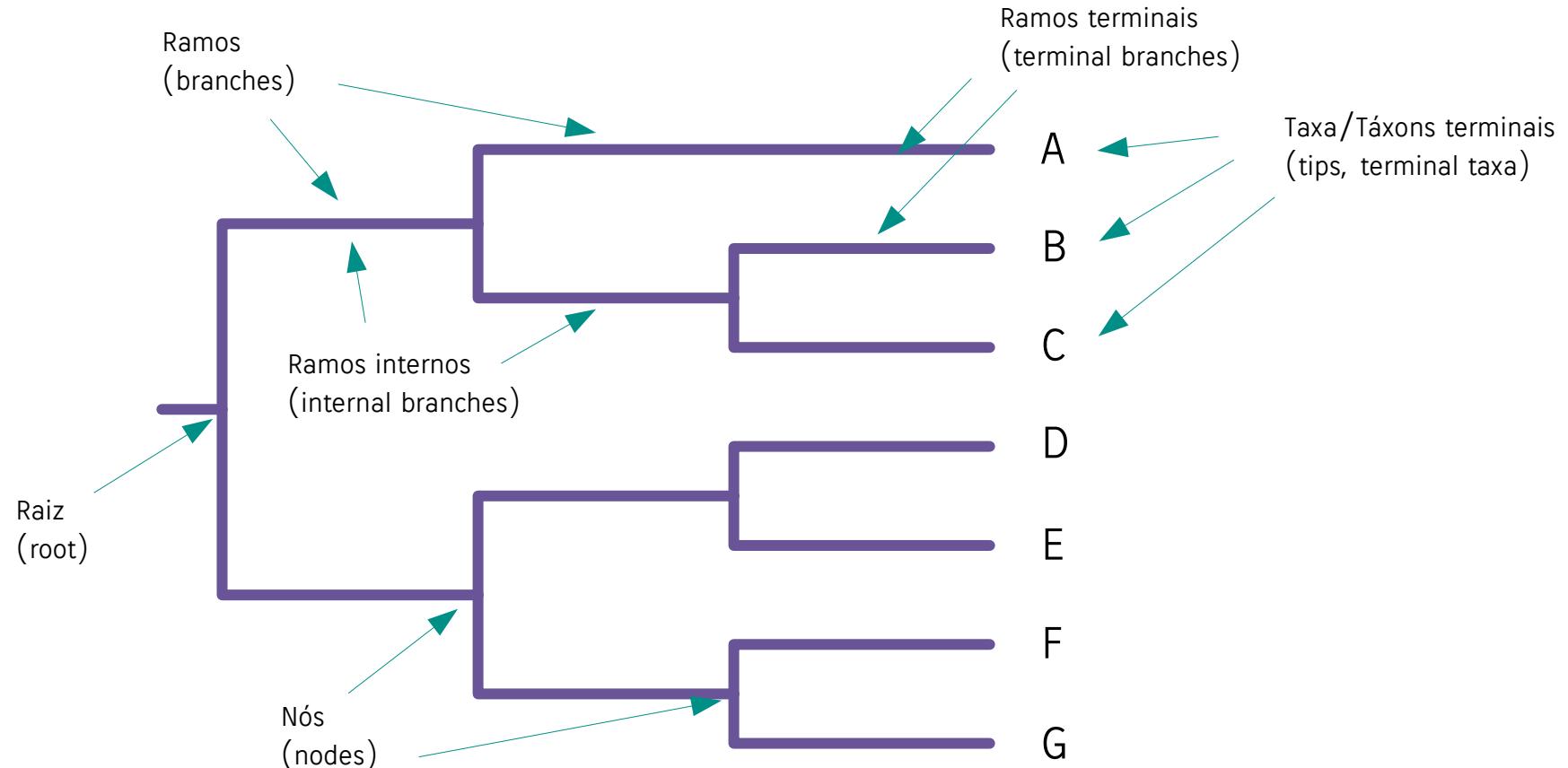
Gene G1



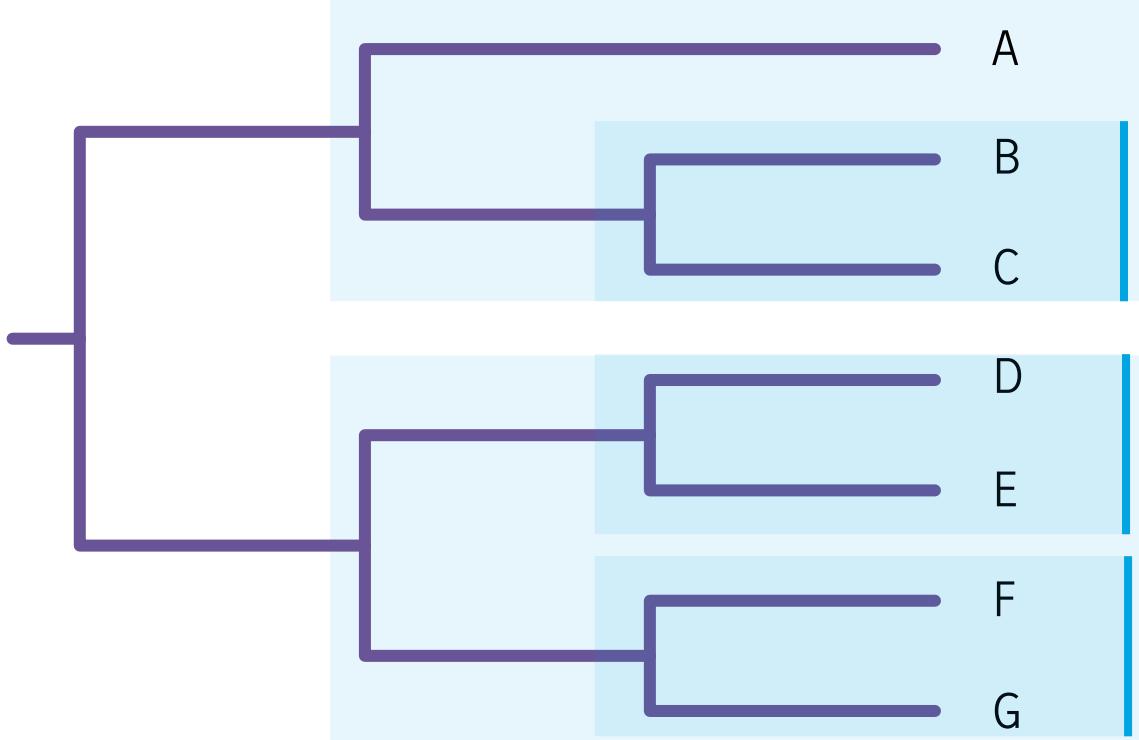
Gene E1



Terminologia

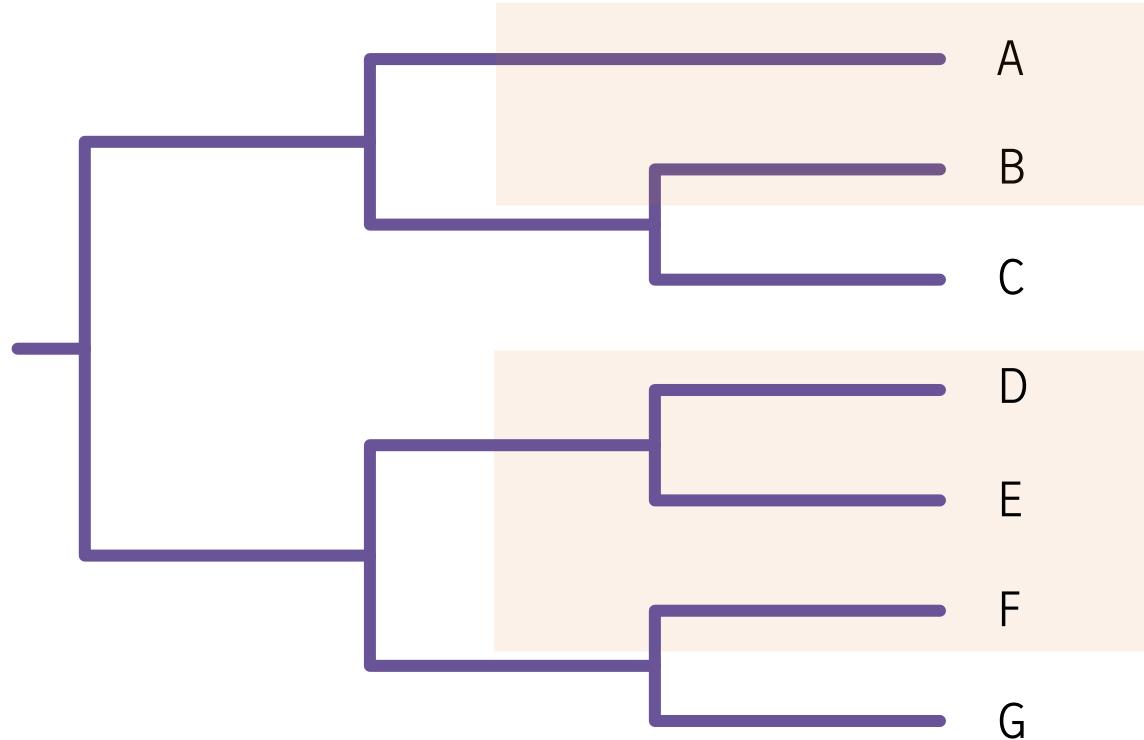


Terminología



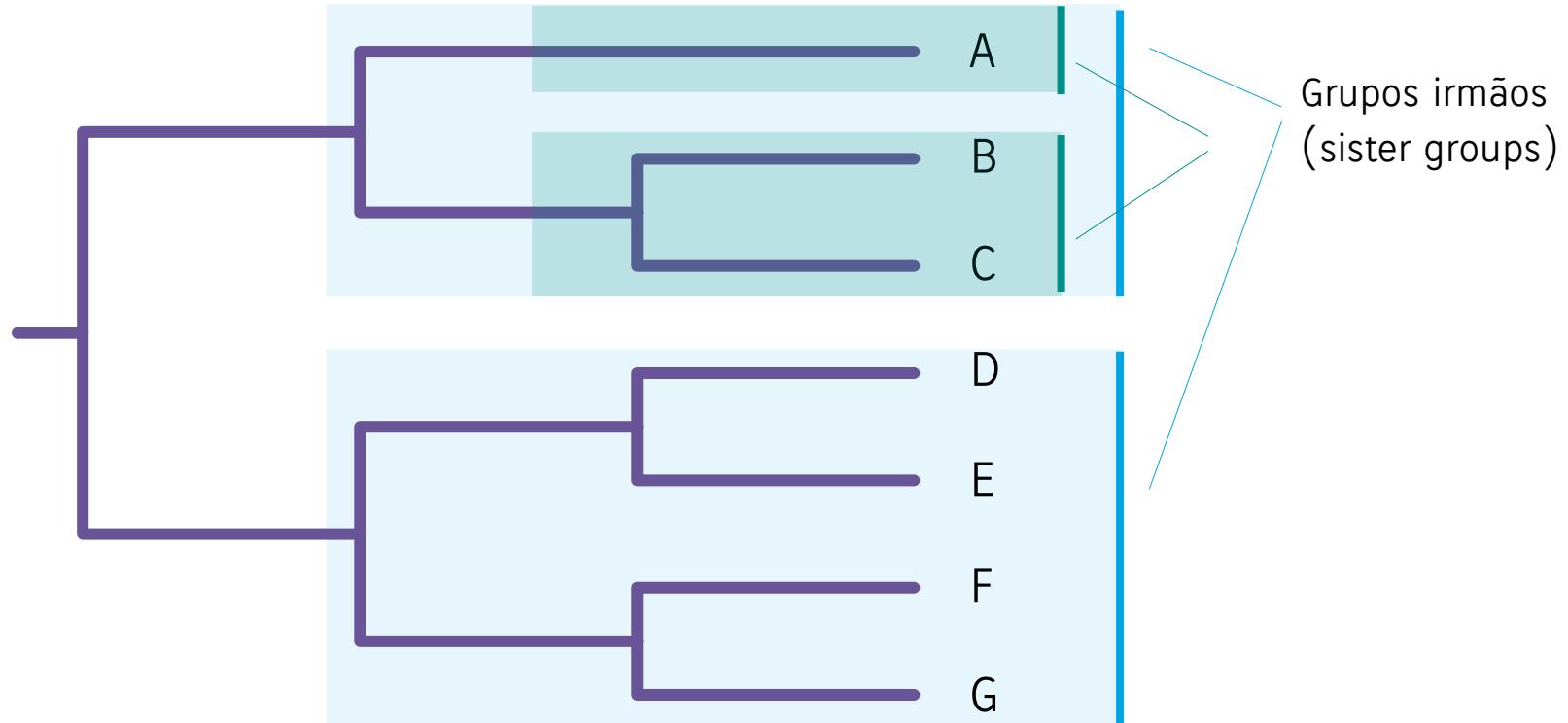
Grupos monofiléticos/clados
(monophyletic groups/clades)

Terminología



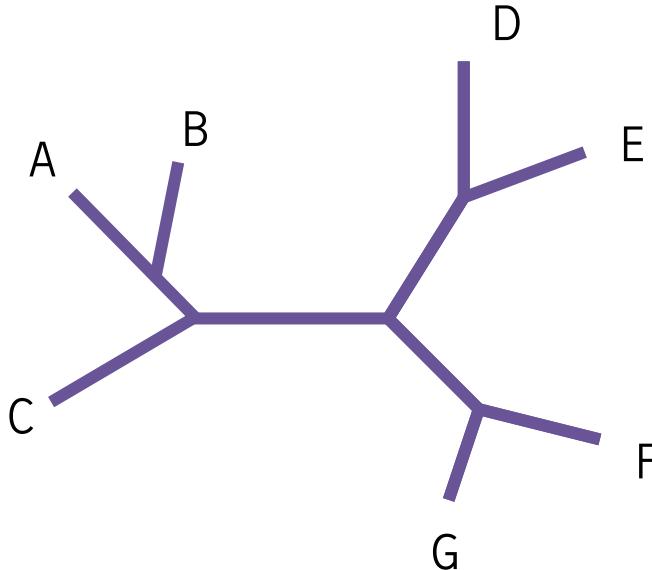
'Grupos' parafiléticos
(paraphyletic groups)

Terminologia

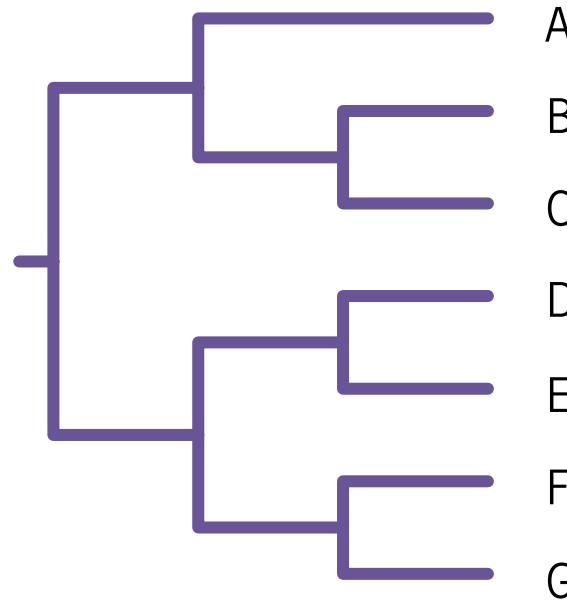


Enraizamento da árvore

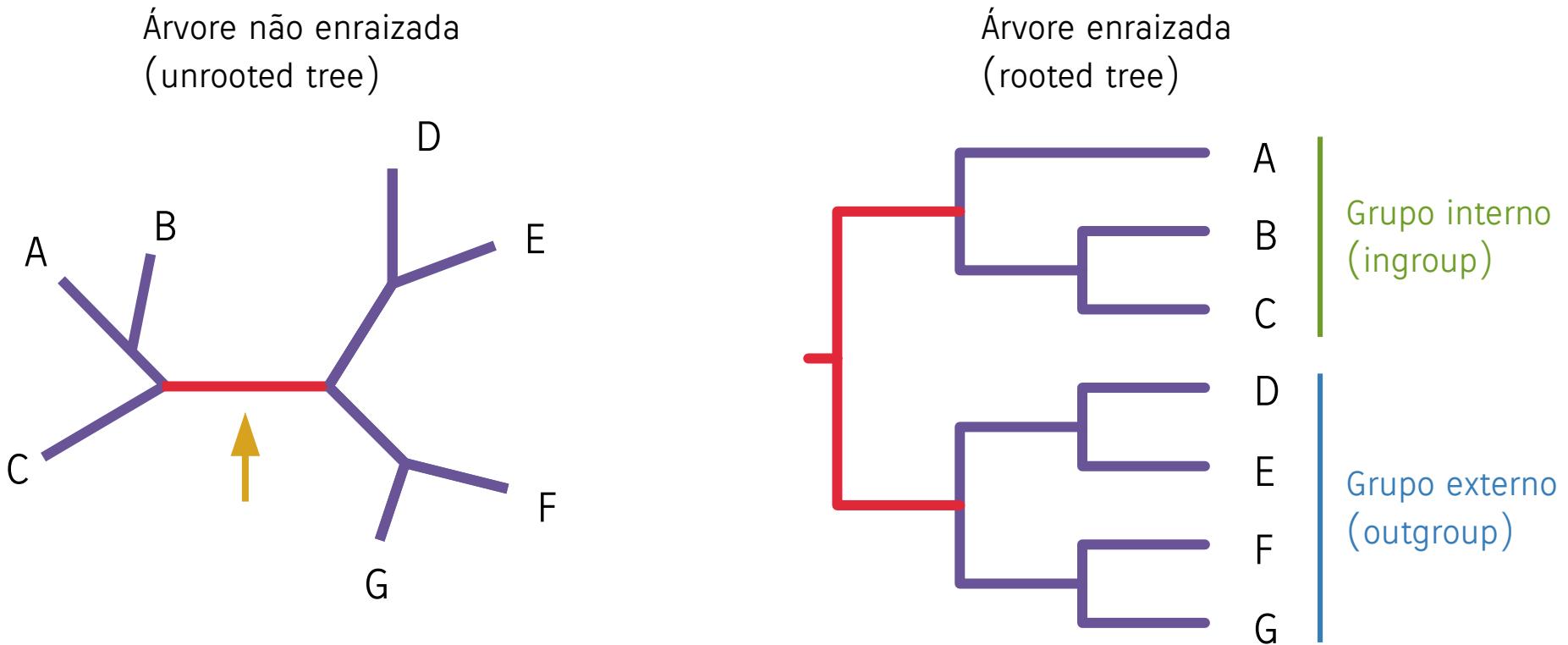
Árvore não enraizada
(unrooted tree)



Árvore enraizada
(rooted tree)

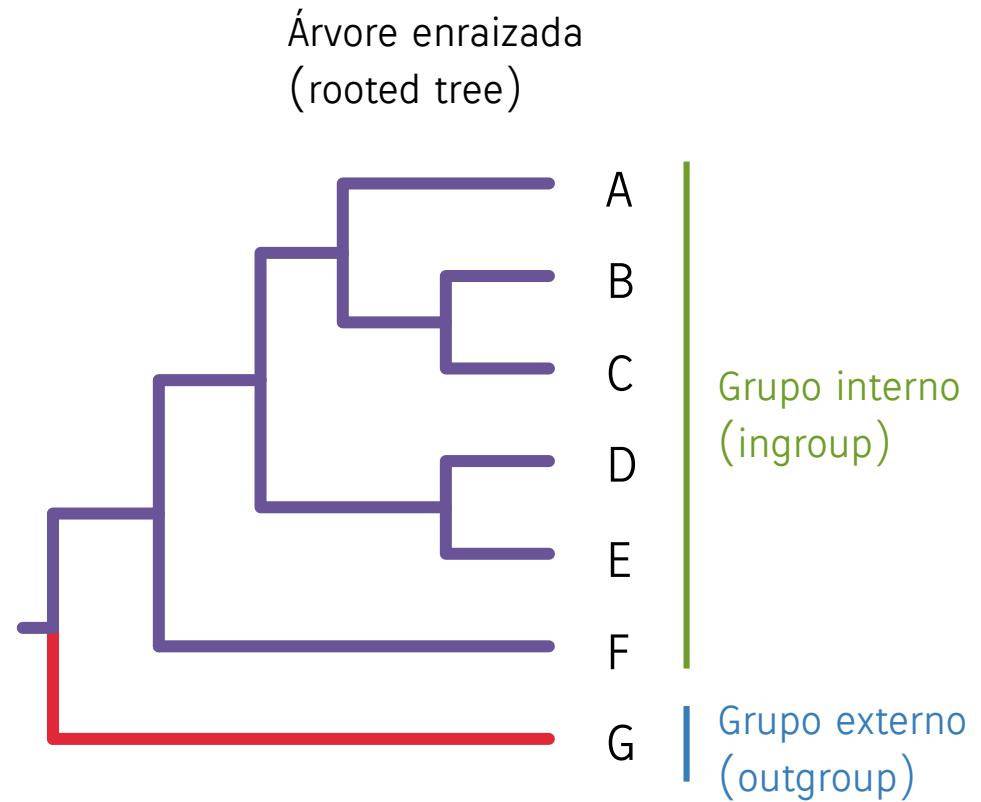
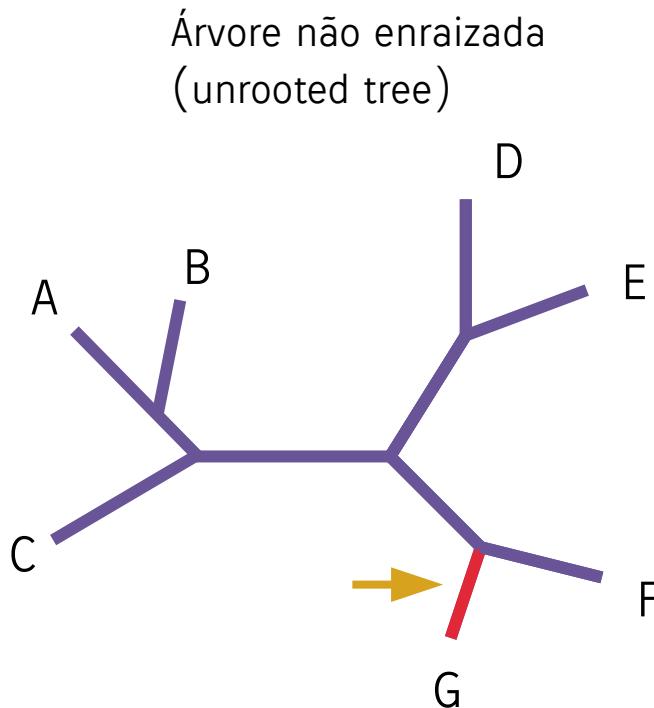


Enraizamento da árvore



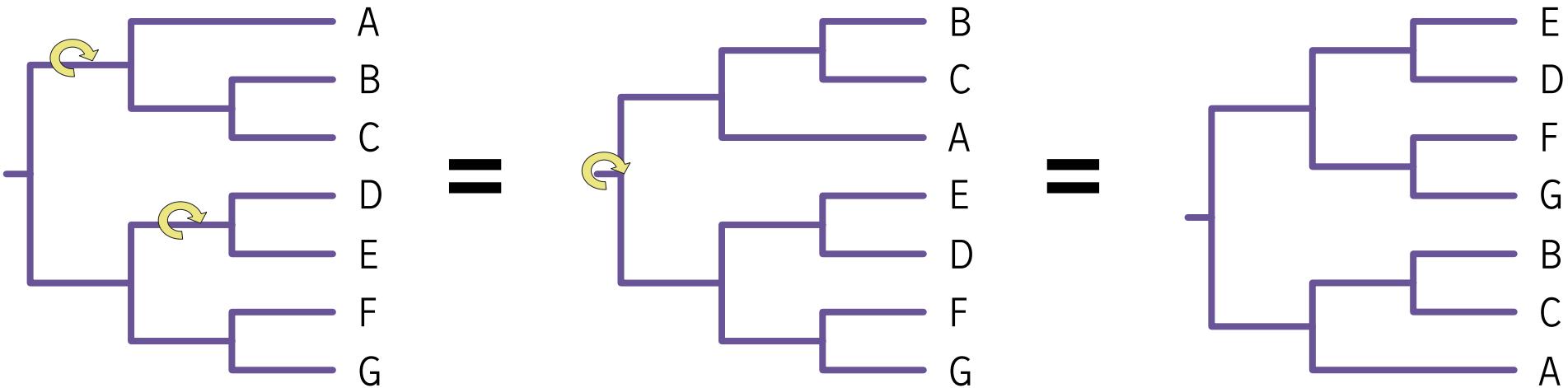
Programas de inferência filogenética geram árvores não enraizadas!
As árvores devem ser enraizadas no grupo externo pelo usuário

Enraizamento da árvore



Programas de inferência filogenética geram árvores não enraizadas!
As árvores devem ser enraizadas no grupo externo pelo usuário

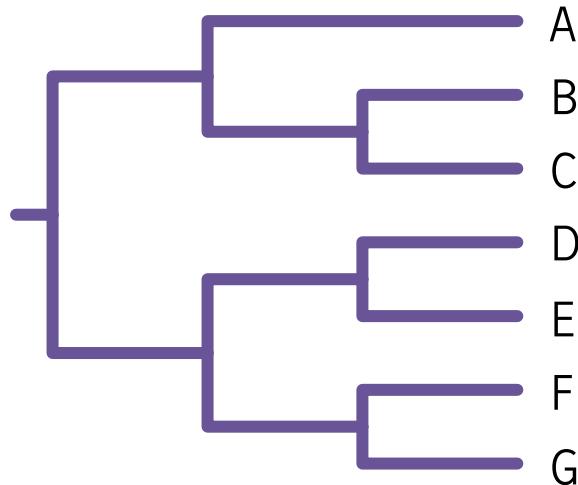
Propriedades



Tipos de árvores

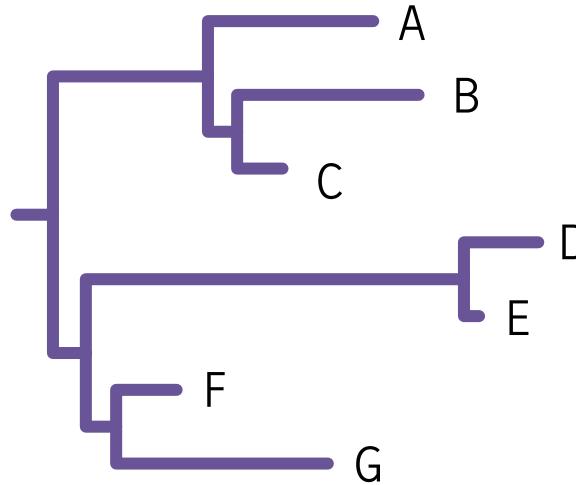
Cladograma
(cladogram)

Mostra apenas a
topologia



Filograma
(phylogram)

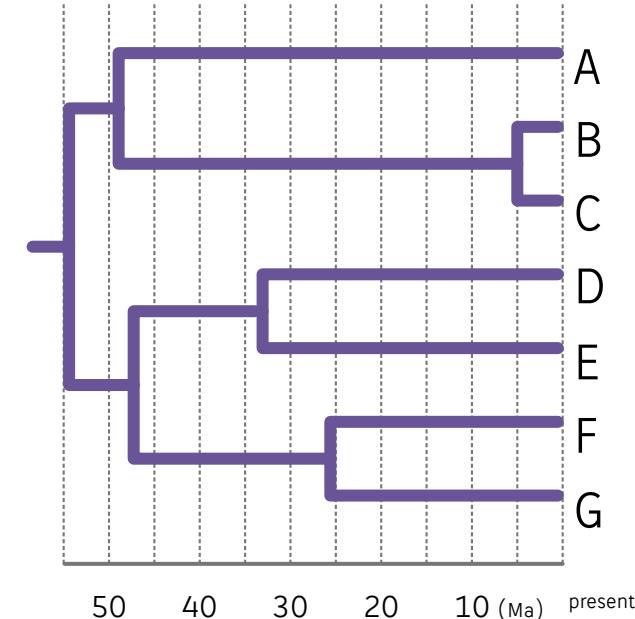
Ramos proporcionais à quantidade
de mudanças



0.05
(substituições por site)

Cronograma
(chronogram)

Ramos proporcionais ao tempo



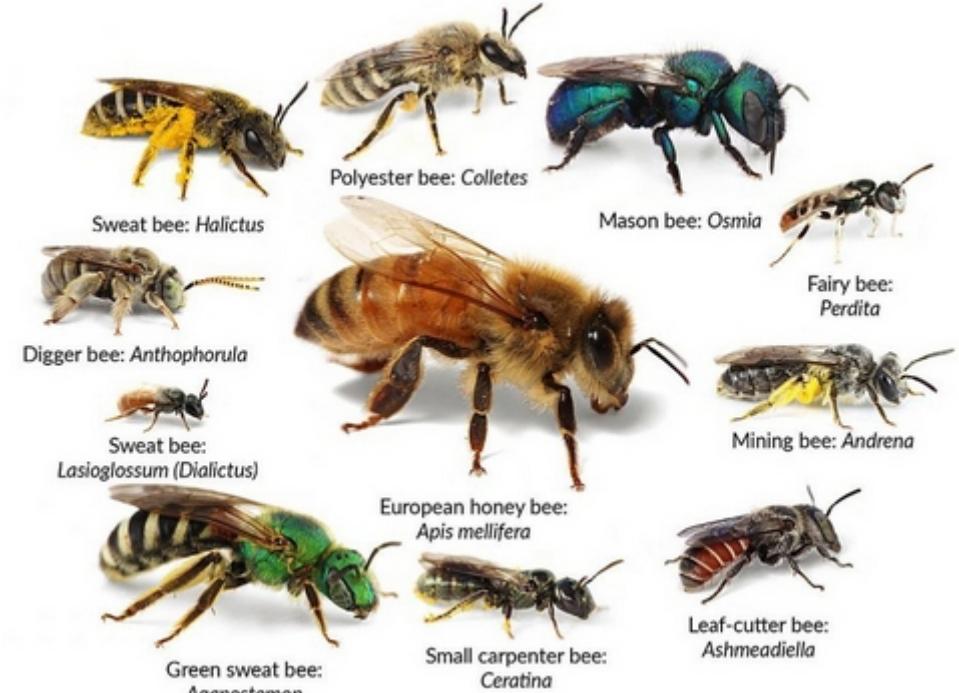
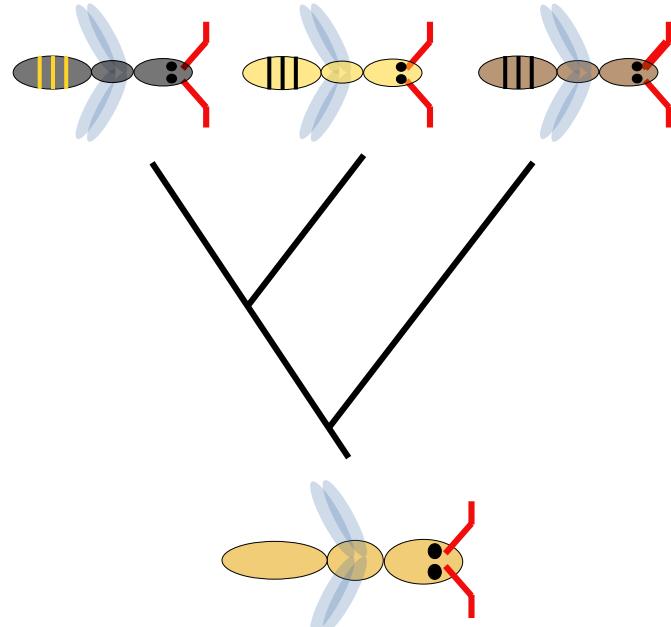
50 40 30 20 10 (Ma) presente

Como inferir árvores filogenéticas?

- Conceito de homologia
- Que dados usar?
- Dificuldades para estabelecer relações de homologia

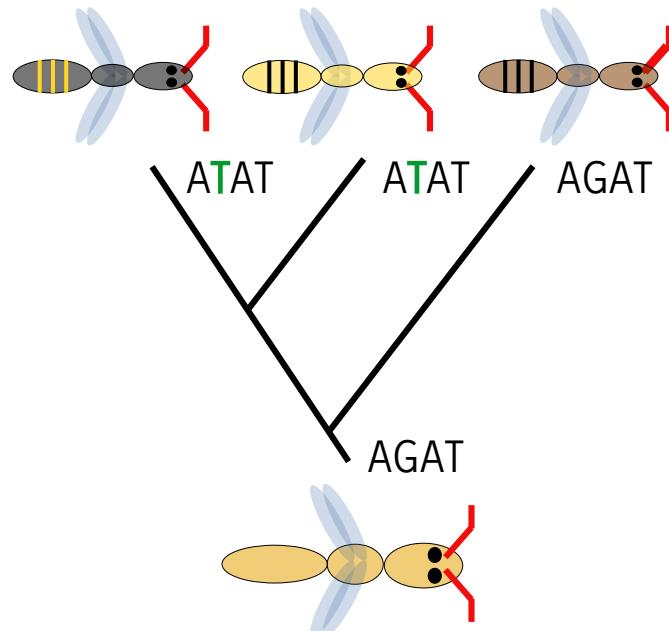
Análise comparativa de características compartilhadas

Caracteres compartilhados entre táxons derivam de um mesmo caracter que já existia no ancestral comum a eles → Homologia



Análise comparativa de características compartilhadas

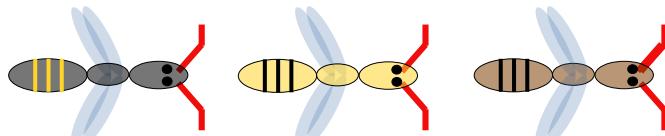
Caracteres compartilhados entre táxons derivam de um mesmo caracter que já existia no ancestral comum a eles → Homologia



Que dados utilizar?

Características compartilhadas entre táxons

- caracteres morfológicos
- sequências de DNA/proteínas

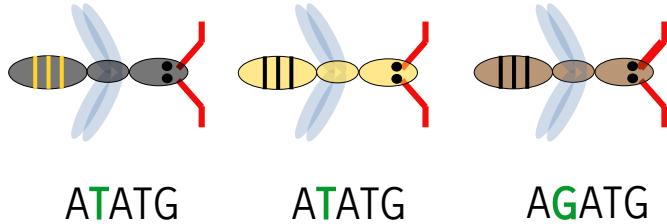


Taxa	Characters			
	0000000001 1234567890	1111111112 1234567890	2222222223 1234567890	3 1
<i>Winthemia venusta</i>	0000000000	0000000000	-000001000	0
<i>Drinomyia hokkaidensis</i>	1000100001	0100000000	-000002000	0
<i>Phorocerosoma vicarium</i>	0000100000	0010000000	-000001000	0
<i>Austrophorocera grandis</i>	0120100000	0101010001	0000003000	1
<i>A. hirsuta</i>	0020100000	0001010001	0000003000	1
<i>Bessa parallela</i>	1021101000	0001010001	1000003000	1
<i>B. remota</i>	1021101000	0001010001	1000003000	1
<i>Chaetoexorista ateripalpis</i>	0011100000	0001010000	-000003000	1

Que dados utilizar?

Características compartilhadas entre táxons

- caracteres morfológicos
- sequências de DNA/proteínas



	A	C	G	T	G	A	C	T	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	C	G	A	T	
sp1	A	C	G	T	G	A	C	T	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	C	G	A	T	
sp2	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	C	G	A	T		
sp3	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	C	G	A	T		
sp4	A	C	G	T	G	A	C	T	T	G	A	T	C	C	T	A	G	C	A	T	G	C	A	T	T	G	A	T	
sp5	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	G	A	T	G	C	A	T	C	G	A	T		
sp6	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	T	C	G	A	T	
sp7	A	C	G	A	A	A	C	T	T	G	A	T	A	G	T	A	G	C	A	T	G	C	A	T	C	G	A	T	
sp8	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	T	C	C	T		
sp9	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	A	G	A	T		
sp10	A	C	G	T	G	A	C	T	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	T	C	G	A	T
sp11	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	T	A	T	G	C	A	T	T	C	G	A	T	
sp12	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	A	C	A	T	G	C	A	T	T	C	G	A	T	
sp13	A	C	G	T	G	T	T	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	T	C	G	A	T	
sp14	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	C	A	A	A	C	A	T	T	C	G	A	T	
sp15	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	T	C	G	A	T	

Alinhamento de sequências

Alinhamento define regiões homólogas

sp1	G C G T G A T C G T A G C T G A C T G T G A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G A C T C T A C G T G T G
sp2	G C G T G A T C G T A G C T G A A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G A C G T G T C T G T A G T C
sp3	T C G T A G C T G A C T G T G A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G A C G T G T C T G T A G T C G T G A C
sp4	G C G T G A T C G T A G C T A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G A C T A C G T G T C T G T A G T C G T G A C
sp5	G C G T G A T C G T A G C T G A C T G T G C G T A G C T G A T C G A T G C T G C G T G T C G A T G C T G A C G T G T C T G T A G T C G T G A C
sp6	G C G T G A T C G T A G C T G A C T A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G A C C G T G A T C T A C G T G T G A C

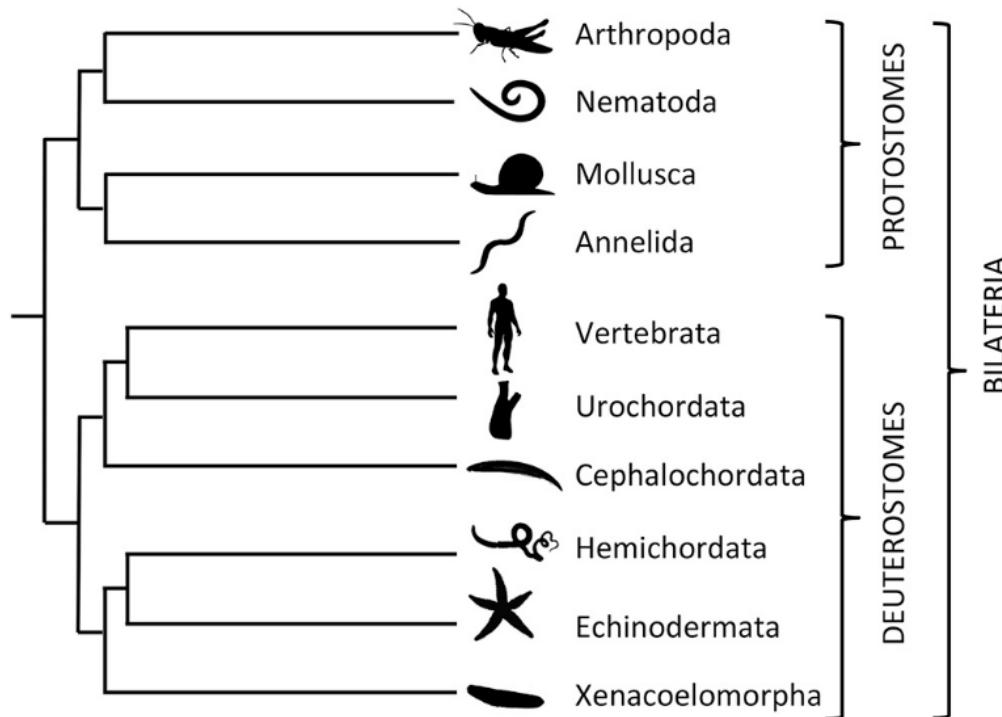


Algoritmo de alinhamento
(maximização da quantidade de
bases idênticas em cada posição)

sp1	G C G T G A T C G T A G C T G A C T G T G A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G - - - A C - - - T C T A C G T G T G - - -
sp2	G C G T G A T C G T A G C T G A - - - A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G - - - A C G T G T C T G T A G T C - - -
sp3	- - - T C G T A G C T G A C T G T G A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G - - - A C G T G T C T G T A G T C G T A G T C G T G A C
sp4	G C G T G A T C G T A G C T - - - A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G A C T A C C G T G T C T G T A G T C G T G A T C T A C G T G T G A C
sp5	G C G T G A T C G T A G C T G A C T G T - - - G T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G - - - A C G T G T C T G T A G T C G T G A T C T A C G T G T G A C
sp6	G C G T G A T C G T A G C T G A C T - - - A T C G T A G C T A G C T G A T C G A T G C T G C G T G T C G A T G C T G - - - A C - - - C G T G A T C T A C G T G T G A C

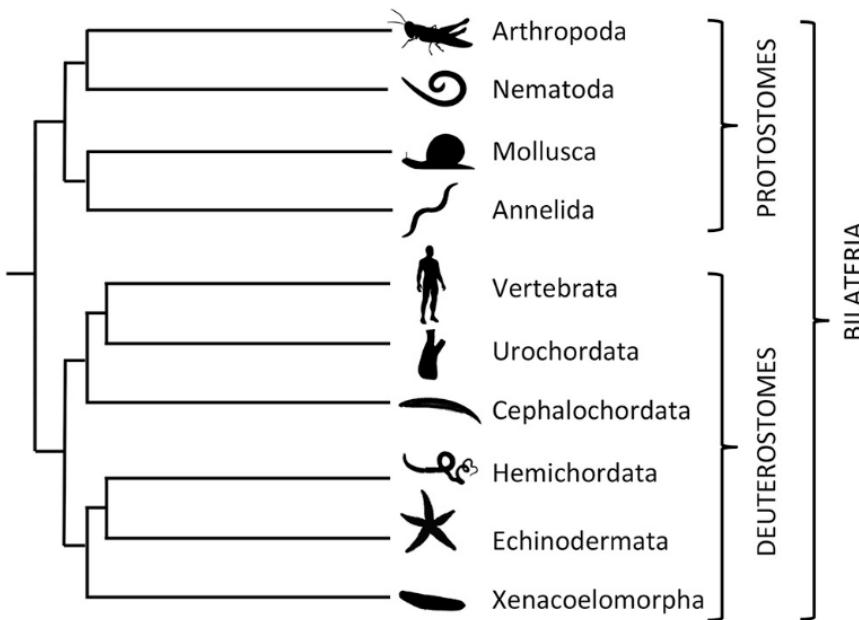
Estabelecendo homologias entre grupos distantes

- Homologias podem não ser óbvias
- Linhagens evoluem independentemente e acumulam mudanças ao longo do tempo...



Estabelecendo homologias entre grupos distantes

- Homologias podem não ser óbvias
- Linhagens evoluem independentemente e acumulam mudanças ao longo do tempo...

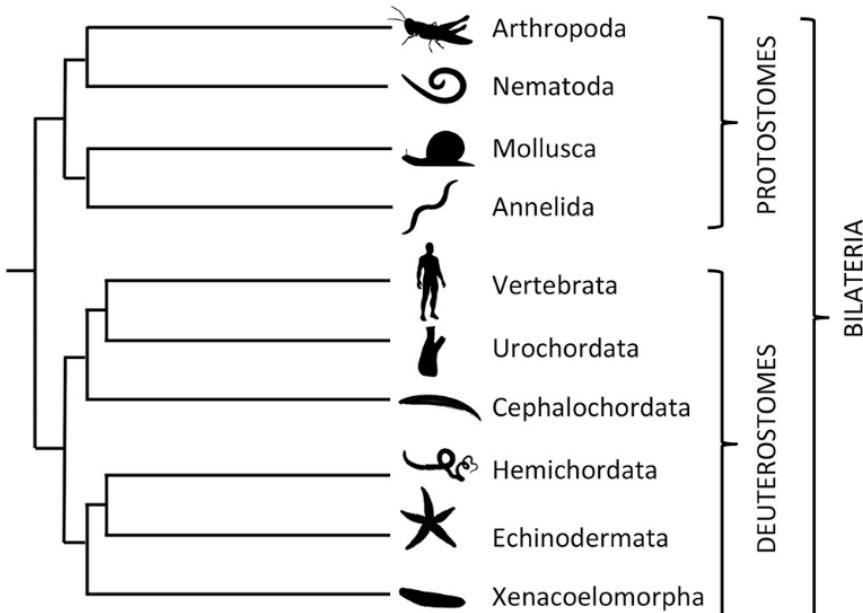


sp1	A	C	G	T	G	A	C	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A					
sp2	A	C	G	T	G	T	C	T	G	A	T	-	G	T	A	G	C	A	T	A	C	T	G	A			
sp3	A	C	G	T	G	T	C	T	G	A	T	C	G	T	A	G	A	G	T	G	C	A	T	G	A		
sp4	A	C	G	T	G	-	-	-	G	G	T	C	C	T	A	G	C	A	T	G	C	A	T	C	G	A	
sp5	A	C	G	T	G	T	C	T	G	G	A	T	C	G	T	A	G	G	A	T	G	C	A	T	C	G	A
sp6	A	C	G	T	G	T	C	T	G	G	-	-	T	A	G	C	A	T	G	C	A	T	C	G	A		
sp7	A	C	G	A	A	A	C	T	G	A	T	A	G	T	A	G	C	A	T	G	C	A	T	C	G	A	
sp8	A	C	G	T	G	T	C	T	G	C	T	C	G	T	A	G	C	A	T	G	C	A	T	C	C		
sp9	A	C	G	T	T	T	A	T	G	C	T	C	G	T	A	G	C	A	T	G	C	A	T	A	G	A	
sp10	A	C	G	T	G	A	C	T	G	T	T	A	G	T	A	G	C	A	T	G	C	A	T	C	G	A	
sp11	A	C	G	A	G	T	C	T	G	A	T	C	G	T	A	G	T	A	T	G	C	A	T	C	G	A	
sp12	A	T	A	T	G	T	C	T	G	A	A	A	A	T	A	A	C	A	T	G	C	A	T	C	G	A	
sp13	A	C	G	T	G	T	T	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	C	G	A	
sp14	A	C	G	T	G	T	A	A	A	T	T	C	G	T	A	G	C	A	A	C	T	C	G	A			
sp15	A	A	T	T	G	T	C	T	G	A	T	C	G	T	A	G	C	A	T	G	C	A	T	C	G	A	

Odekunle and Elphick (2020)

Estabelecendo homologias entre grupos distantes

- Homologias podem não ser óbvias
- Linhagens evoluem independentemente e acumulam mudanças ao longo do tempo...
- Sequências de aminoácidos são mais conservadas (o código genético é degenerado)



Odekunle and Elphick (2020)

sp1	T	R	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp2	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp3	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp4	T	R	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp5	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp6	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp7	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp8	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp9	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp10	T	R	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp11	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp12	T	S	L	I	V	A	C	I	D	R	P	I	L	D	R	S
sp13	T	S	L	I	V	A	C	I	D	R	P	M	L	E	R	S
sp14	T	S	L	I	V	A	C	I	D	R	P	M	L	D	R	S
sp15	T	S	L	I	V	A	C	I	D	H	Q	M	L	D	R	S

Métodos para inferência de árvores filogenéticas

- Principais métodos de inferência
- Métodos baseados em distância
- Métodos baseados em caracteres, parte I : Máxima parcimônia

Grupos de métodos para inferência filogenética

1. Métodos baseados em distância (distance-based methods)

Árvores são construídas a partir de uma matriz de distâncias

2. Métodos baseados em caracteres (character-based methods)

Todos os caracteres são considerados individualmente para a construção da árvore

Grupos de métodos para inferência filogenética

1. Métodos baseados em distância (distance-based methods)

Árvores são construídas a partir de uma matriz de distâncias

→ Neighbour joining (NJ)

2. Métodos baseados em caracteres (character-based methods)

Todos os caracteres são considerados individualmente para a construção da árvore

- 2.1. Máxima parcimônia (maximum parsimony, MP)
- 2.2. Máxima verossimilhança (maximum likelihood, ML)
- 2.3. Inferência Bayesiana (Bayesian inference, BI)

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

	1	2	3	4	5	6	7	8	9	10
<u>Sp1</u>	A	A	C	G	A	G	G	T	G	A
<u>Sp2</u>	A	T	C	G	A	G	T	T	G	A
<u>Sp3</u>	A	G	C	T	A	C	C	A	G	A
<u>Sp4</u>	A	G	C	T	A	C	C	G	A	A

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

	1	2	3	4	5	6	7	8	9	10		Sp1	Sp2	Sp3	Sp4
Sp1	A	A	C	G	A	G	G	T	G	A		Sp1			
Sp2	A	T	C	G	A	G	T	T	G	A		Sp2			
Sp3	A	G	C	T	A	C	C	A	G	A		Sp3			
Sp4	A	G	C	T	A	C	C	G	A	A		Sp4			

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

	1	2	3	4	5	6	7	8	9	10		Sp1	Sp2	Sp3	Sp4
Sp1	A	A	C	G	A	G	G	T	G	A	Sp1	0	2		
Sp2	A	T	C	G	A	G	T	T	G	A	Sp2				
Sp3	A	G	C	T	A	C	C	A	G	A	Sp3				
Sp4	A	G	C	T	A	C	C	G	A	A	Sp4				

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

	1	2	3	4	5	6	7	8	9	10		Sp1	Sp2	Sp3	Sp4
Sp1	A	A	C	G	A	G	G	T	G	A	Sp1	0	2	5	
Sp2	A	T	C	G	A	G	T	T	G	A	Sp2				
Sp3	A	G	C	T	A	C	C	A	G	A	Sp3				
Sp4	A	G	C	T	A	C	C	G	A	A	Sp4				

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

	1	2	3	4	5	6	7	8	9	10		Sp1	Sp2	Sp3	Sp4
Sp1	A	A	C	G	A	G	G	T	G	A	Sp1	0	2	5	6
Sp2	A	T	C	G	A	G	T	T	G	A	Sp2				
Sp3	A	G	C	T	A	C	C	A	G	A	Sp3				
Sp4	A	G	C	T	A	C	C	G	A	A	Sp4				

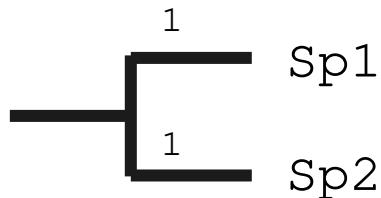
1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

	1	2	3	4	5	6	7	8	9	10		Sp1	Sp2	Sp3	Sp4
Sp1	A	A	C	G	A	G	G	T	G	A	Sp1	0	2	5	6
Sp2	A	T	C	G	A	G	T	T	G	A	Sp2	—	0	5	6
Sp3	A	G	C	T	A	C	C	A	G	A	Sp3	—	—	0	2
Sp4	A	G	C	T	A	C	C	G	A	A	Sp4	—	—	—	0

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

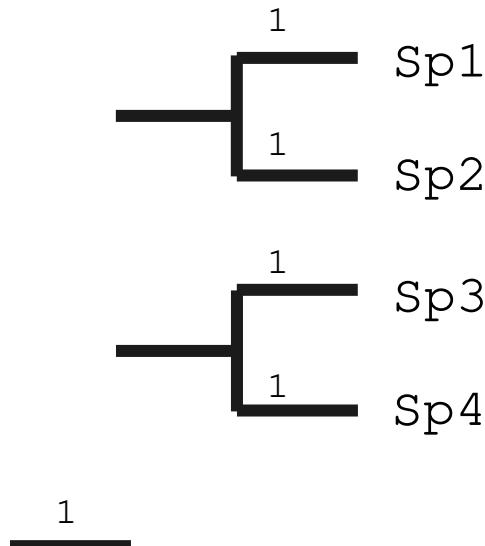


	Sp1	Sp2	Sp3	Sp4
Sp1	0	2	5	6
Sp2	-	0	5	6
Sp3	-	-	0	2
Sp4	-	-	-	0

Métodos baseados em distância: Encontre a árvore na qual as distâncias entre táxons (topologia e tamanho dos ramos) sejam o mais próximo possível das distâncias genéticas observadas (matriz)

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

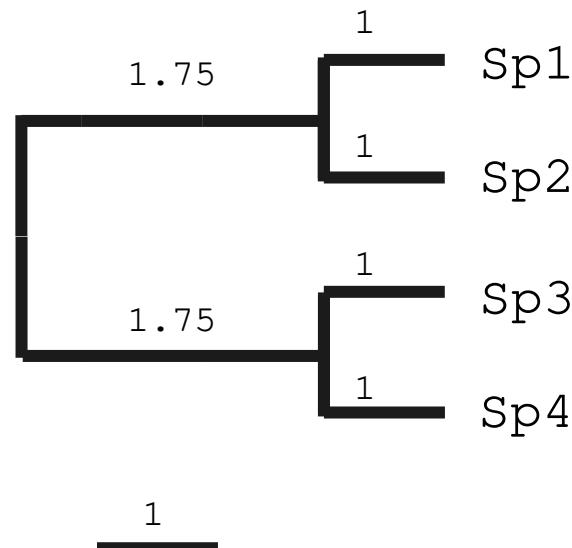


	Sp1	Sp2	Sp3	Sp4
Sp1	0	2	5	6
Sp2	—	0	5	6
Sp3	—	—	0	2
Sp4	—	—	—	0

Métodos baseados em distância: Encontre a árvore na qual as distâncias entre táxons (topologia e tamanho dos ramos) sejam o mais próximo possível das distâncias genéticas observadas (matriz)

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

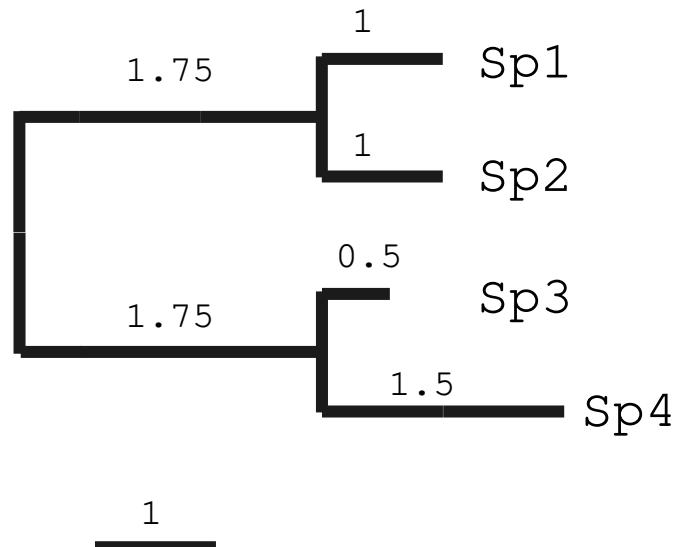


	Sp1	Sp2	Sp3	Sp4
Sp1	0	2	5	6
Sp2	—	0	5	6
Sp3	—	—	0	2
Sp4	—	—	—	0

Métodos baseados em distância: Encontre a árvore na qual as distâncias entre táxons (topologia e tamanho dos ramos) sejam o mais próximo possível das distâncias genéticas observadas (matriz)

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

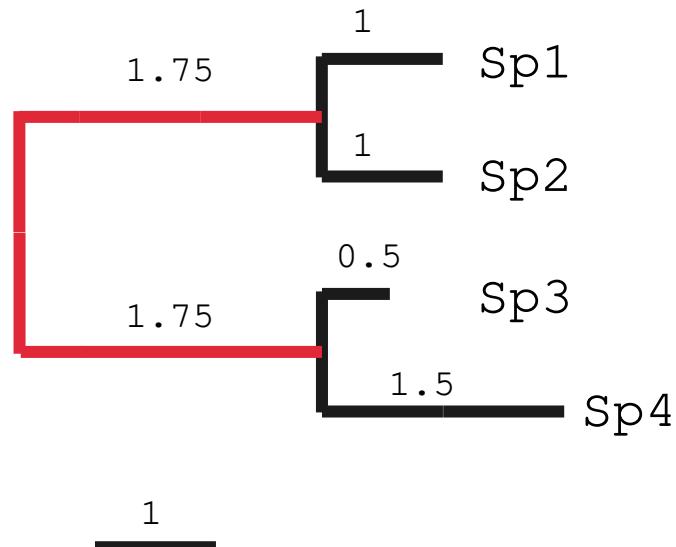


	Sp1	Sp2	Sp3	Sp4
Sp1	0	2	5	6
Sp2	—	0	5	6
Sp3	—	—	0	2
Sp4	—	—	—	0

Métodos baseados em distância: Encontre a árvore na qual as distâncias entre táxons (topologia e tamanho dos ramos) sejam o mais próximo possível das distâncias genéticas observadas (matriz)

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias

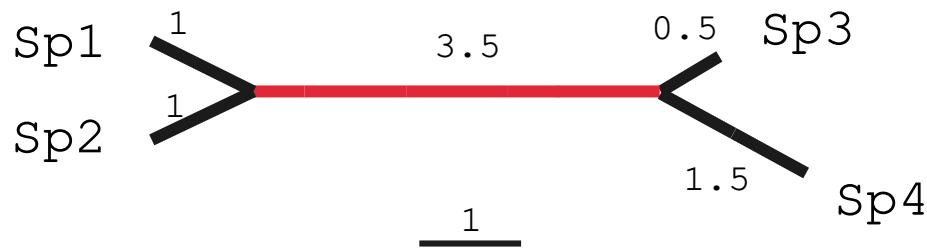


	Sp1	Sp2	Sp3	Sp4
Sp1	0	2	5	6
Sp2	—	0	5	6
Sp3	—	—	0	2
Sp4	—	—	—	0

Métodos baseados em distância: Encontre a árvore na qual as distâncias entre táxons (topologia e tamanho dos ramos) sejam o mais próximo possível das distâncias genéticas observadas (matriz)

1. Métodos baseados em distância

Uma árvore é construída diretamente a partir de uma matriz de distâncias



	Sp1	Sp2	Sp3	Sp4
Sp1	0	2	5	6
Sp2	-	0	5	6
Sp3	-	-	0	2
Sp4	-	-	-	0

Métodos baseados em distância: Encontre a árvore na qual as distâncias entre táxons (topologia e tamanho dos ramos) sejam o mais próximo possível das distâncias genéticas observadas (matriz)

1. Métodos baseados em distância

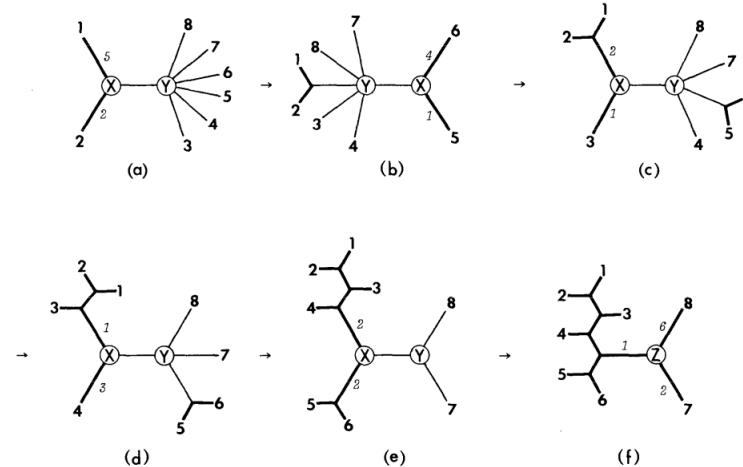
Neighbour joining (NJ) (Saitou & Nei, 1987)

- Método de clusterização

Segue uma ‘receita’ para chegar numa árvore que acomoda as distâncias observadas

- Algoritmo ultraeficiente

Não procura pela árvore ótima no universo de todas as árvores possíveis (como em métodos de otimização)



1. Métodos baseados em distância

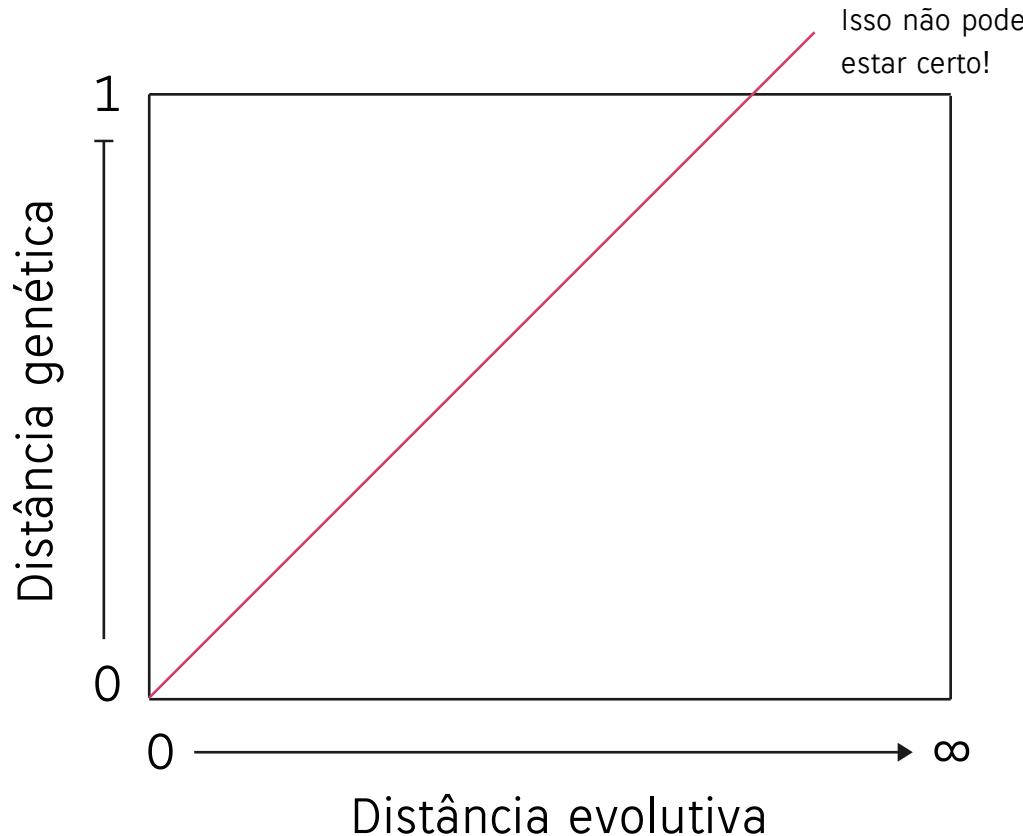
Neighbour joining (NJ) (Saitou & Nei, 1987)

Principais críticas:

- Depende de um conjunto de condições para produzir resultados confiáveis:
 - Sequências não muito distantes entre si
 - Taxa de evolução similar entre táxons
 - Taxa de evolução similar ao longo da sequência
- Pode gerar ramos com tamanho negativo
- Perde-se informação quando sequências são reduzidas a um número
- Distâncias observadas entre sequências muito divergentes são uma estimativa ruim da distância evolutiva verdadeira, e por isso devem ser corrigidas
 - ‘Problema das substituições ocultas’

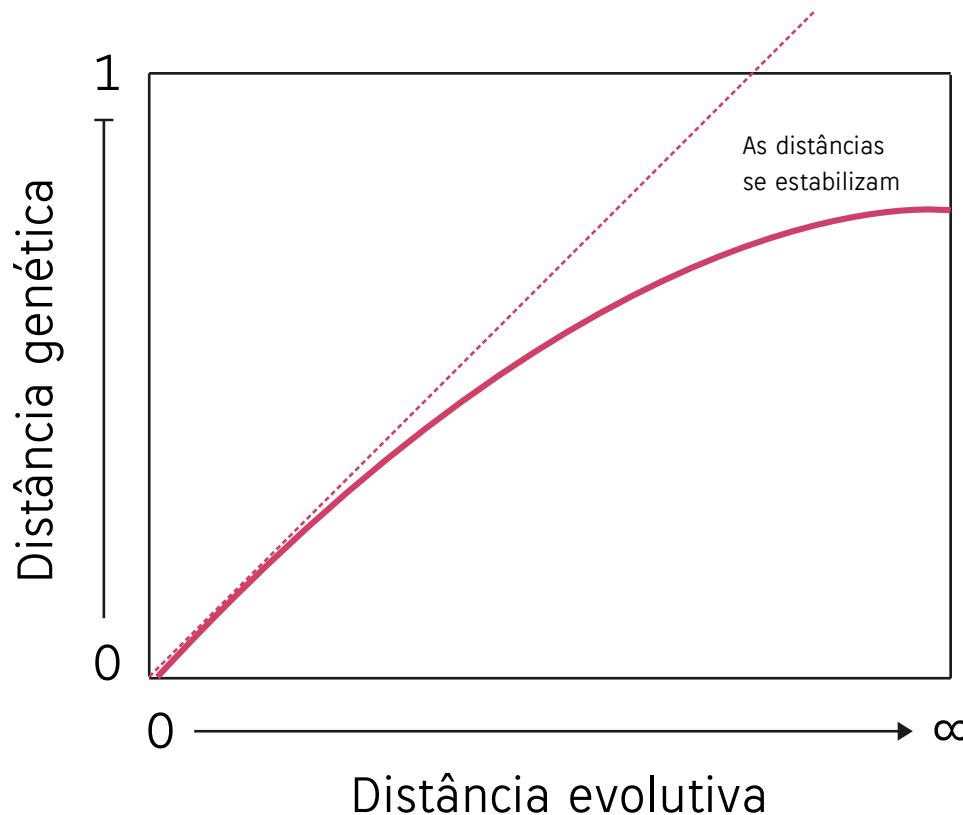
Nem todas as mutações são observáveis

Relação esperada entre distância genética e distância evolutiva



Nem todas as mutações são observáveis

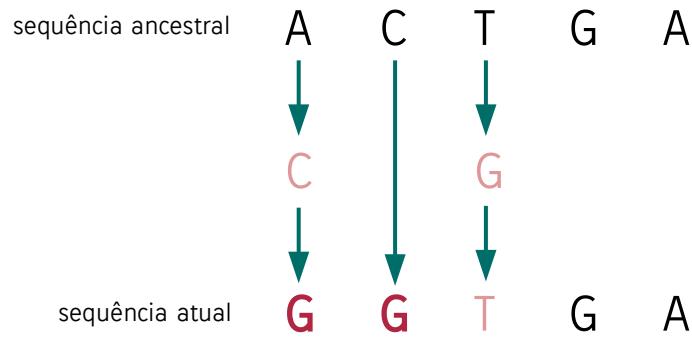
Relação esperada entre distância genética e distância evolutiva



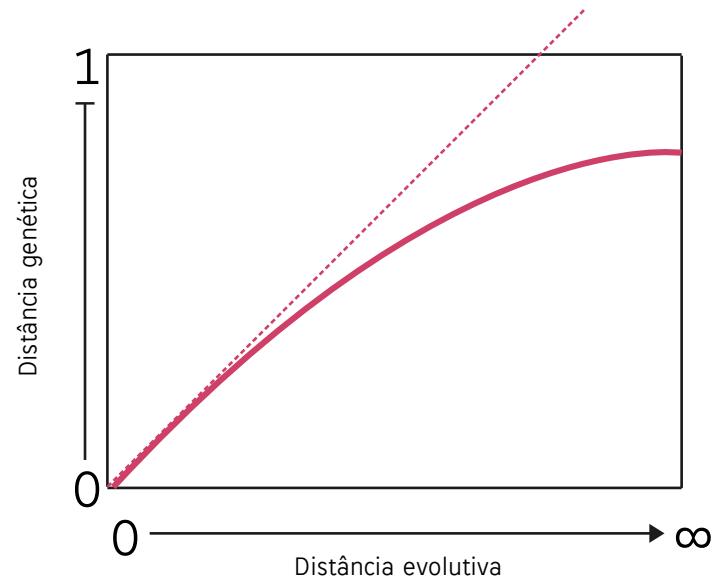
Nem todas as mutações são observáveis

O problema da saturação ('multiple hits'):

- Múltiplas substituições podem afetar a mesma posição na sequência
- Distâncias genéticas são quase sempre subestimadas
- Modelos estatísticos devem ser usados para estimar o número de substituições não observáveis e corrigir as distâncias



Mutações ocorridas = 5
Mutações observadas = 2



Nem todas as mutações são observáveis

O problema da saturação ('multiple hits'):

- Múltiplas substituições podem afetar a mesma posição na sequência
- Distâncias genéticas são quase sempre subestimadas
- Modelos estatísticos devem ser usados para estimar o número de substituições não observáveis e corrigir as distâncias

Correção de Jukes–Cantor

$$d = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p \right)$$

d = distância corrigida

p = distância observada

2. Métodos baseados em caracteres

- Todos os caracteres da sequência são considerados individualmente no processo de inferência da árvore
- Usam um critério de otimização: diferentes árvores são avaliadas, e cada uma recebe uma pontuação; aquela com melhor pontuação é escolhida como a melhor árvore* (Inferência Bayesiana produz uma distribuição de árvores igualmente prováveis)
- Principais métodos:
 - 2.1. Máxima parcimônia
 - 2.2. Máxima verossimilhança
 - 2.3. Inferência Bayesiana (usa máxima verossimilhança)

2. Métodos baseados em caracteres

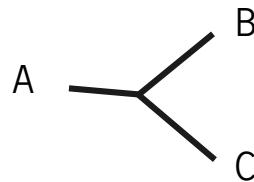
- Principal desafio dos métodos que usam critérios de otimização:
Poder computacional para avaliar as topologias possíveis

2. Métodos baseados em caracteres

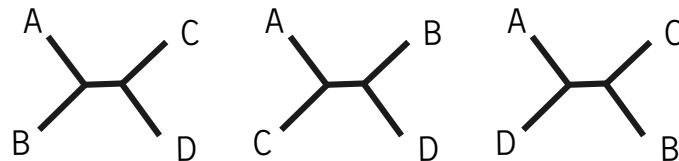
Topologias possíveis

3 táxons

Não enraizadas



4 táxons



5 táxons

15 topologias

Enraizadas



15 topologias

105 topologias

2. Métodos baseados em caracteres

Topologias possíveis

60 táxons =

$5,86 \times 10^{96}$ topologias

(maior que o número de
átomos no universo!)

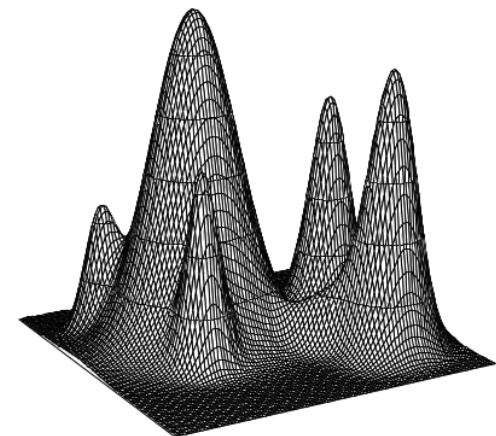
Tips	Number of unrooted (binary) trees
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625
16	213,458,046,676,875
17	6,190,283,353,629,375
18	191,898,783,962,510,625
19	6,332,659,870,762,850,625
20	22,164,309,5476,699,771,875
21	8,200,794,532,637,891,559,375
22	319,830,986,772,877,770,815,625
23	13,113,070,457,687,988,603,440,625
24	563,862,029,680,583,509,947,946,875
	> 21 moles of trees

2. Métodos baseados em caracteres

- Principal desafio dos métodos que usam critérios de otimização:
Poder computacional para avaliar as topologias possíveis
- Os programas na verdade não avaliam todas as árvores:
→ Métodos heurísticos

2. Métodos baseados em caracteres

- Principal desafio dos métodos que usam critérios de otimização:
Poder computacional para avaliar as topologias possíveis
- Os programas na verdade não avaliam todas as árvores:
→ Métodos heurísticos:
 1. Começa-se com uma árvore inicial aleatória
 2. A árvore recebe uma pontuação
 3. Avalia-se uma árvore ‘vizinha’ da árvore atual
 - a nova árvore recebe uma pontuação
 - se ela for melhor, ela substitui a primeira
 - quando não houver mais árvores vizinhas, pare (ou procure árvores em outra região do espaço)



2.1. Máxima parcimônia

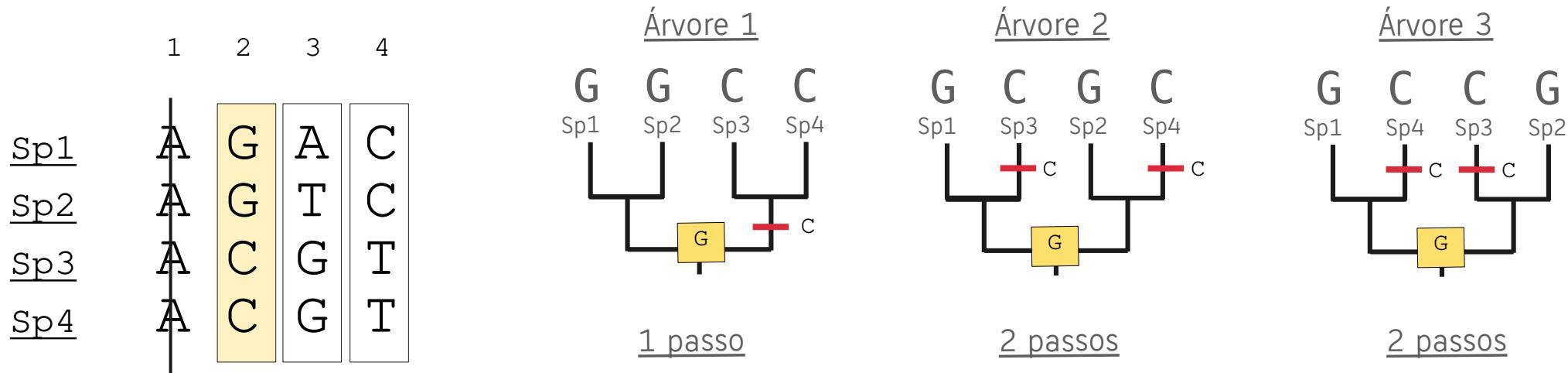
(Princípio da parcimônia: Havendo várias hipóteses igualmente plausíveis para explicar os dados, escolha a mais simples)

- Critério de ‘pontuação’ das árvores: número de passos evolutivos (i.e. mutações) necessários para explicar os dados (i.e. as bases de cada coluna em cada ramo terminal)
- Melhor árvore = menor número de passos evolutivos

2.1. Máxima parcimônia

(Princípio da parcimônia: Havendo várias hipóteses igualmente plausíveis para explicar os dados, escolha a mais simples)

- Critério de ‘pontuação’ das árvores: número de passos evolutivos (i.e. mutações) necessários para explicar os dados (i.e. as bases de cada coluna em cada ramo terminal)
- Melhor árvore = menor número de passos evolutivos



Árvore 1	-	1
Árvore 2	-	2
Árvore 3	-	2

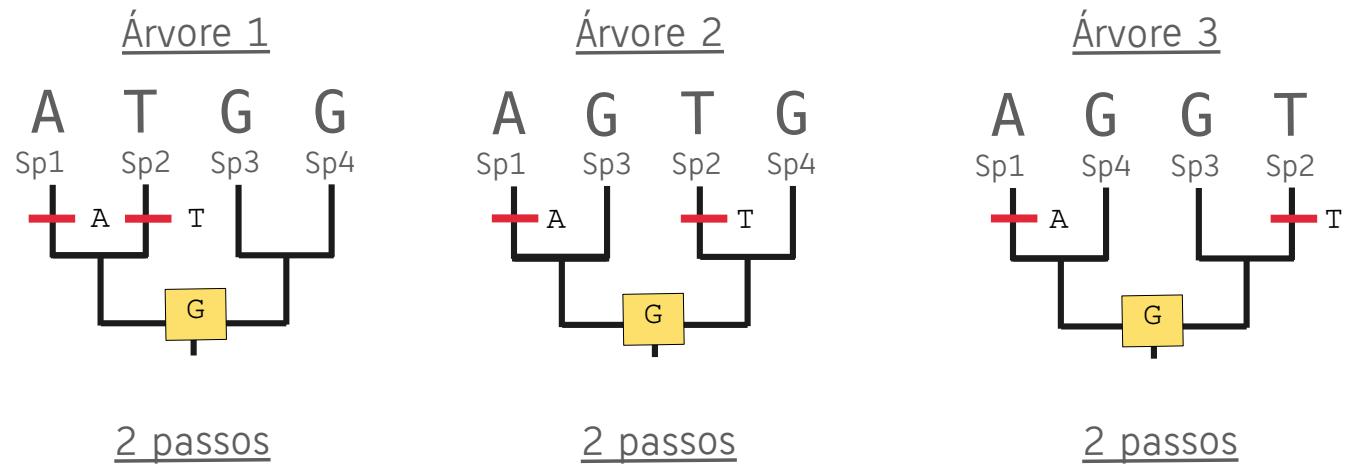
Passos evolutivos

2.1. Máxima parcimônia

(Princípio da parcimônia: Havendo várias hipóteses igualmente plausíveis para explicar os dados, escolha a mais simples)

- Critério de ‘pontuação’ das árvores: número de passos evolutivos (i.e. mutações) necessários para explicar os dados (i.e. as bases de cada coluna em cada ramo terminal)
- Melhor árvore = menor número de passos evolutivos

	1	2	3	4
<u>Sp1</u>	A	G	A	C
<u>Sp2</u>	A	G	T	C
<u>Sp3</u>	A	C	G	T
<u>Sp4</u>	A	C	G	T



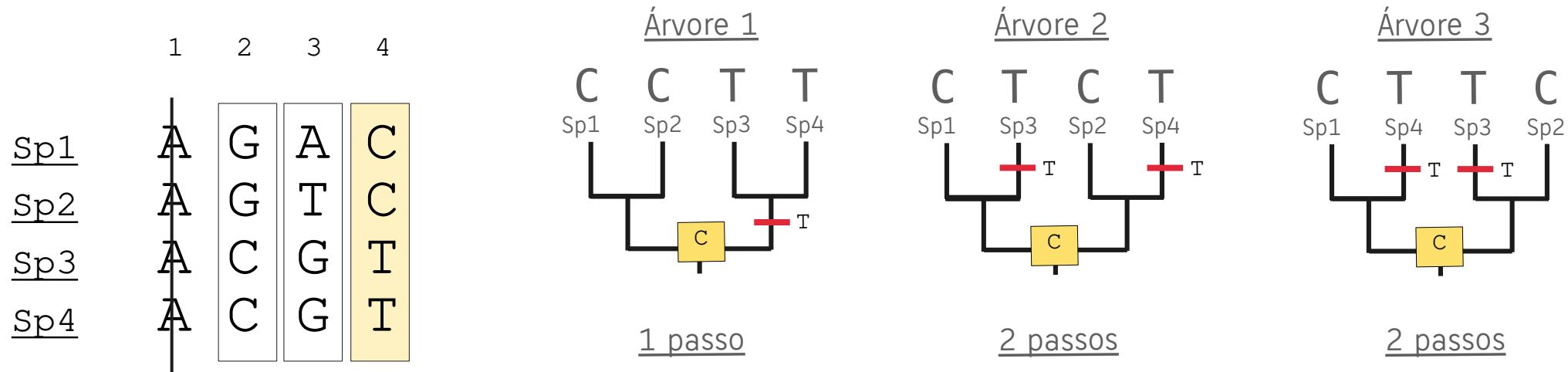
Árvore 1	-	1	2
Árvore 2	-	2	2
Árvore 3	-	2	2

Passos evolutivos

2.1. Máxima parcimônia

(Princípio da parcimônia: Havendo várias hipóteses igualmente plausíveis para explicar os dados, escolha a mais simples)

- Critério de ‘pontuação’ das árvores: número de passos evolutivos (i.e. mutações) necessários para explicar os dados (i.e. as bases de cada coluna em cada ramo terminal)
- Melhor árvore = menor número de passos evolutivos



Árvore 1	-	1	2	1
Árvore 2	-	2	2	2
Árvore 3	-	2	2	2

Passos evolutivos

2.1. Máxima parcimônia

(Princípio da parcimônia: Havendo várias hipóteses igualmente plausíveis para explicar os dados, escolha a mais simples)

- Critério de ‘pontuação’ das árvores: número de passos evolutivos (i.e. mutações) necessários para explicar os dados (i.e. as bases de cada coluna em cada ramo terminal)
- Melhor árvore = menor número de passos evolutivos

	1	2	3	4	
Sp1	A	G	A	C	
Sp2	A	G	T	C	
Sp3	A	C	G	T	
Sp4	A	C	G	T	
					Total
Árvore 1	-	1	2	1	4
Árvore 2	-	2	2	2	6
Árvore 3	-	2	2	2	6

The diagram shows three phylogenetic trees. The first tree (circled in red) has a root at the bottom, with branches leading to Sp1, Sp2, Sp3, and Sp4. The second tree has a root at the bottom, with branches leading to Sp1, Sp3, Sp2, and Sp4. The third tree has a root at the bottom, with branches leading to Sp1, Sp4, Sp3, and Sp2. The characters are represented by colored boxes: character 1 is black, character 2 is white, character 3 is yellow, and character 4 is orange.

2.1. Máxima parcimônia

- Base teórica do método é bem estabelecida ('Sistemática filogenética', proposta por Willi Hennig em 1966)
- Minimiza o número de homoplasias (surgimento independente de caracteres)
- Método recomendado para análises filogenéticas com dados morfológicos, dada a dificuldade de modelar a evolução de caracteres morfológicos (o que não é o caso em dados moleculares); no entanto, a Inferência Bayesiana também pode ser usada
- Mais rápido que máxima verossimilhança
- Funciona bem em árvores 'clock-like': taxa de evolução mais ou menos constante entre linhagens

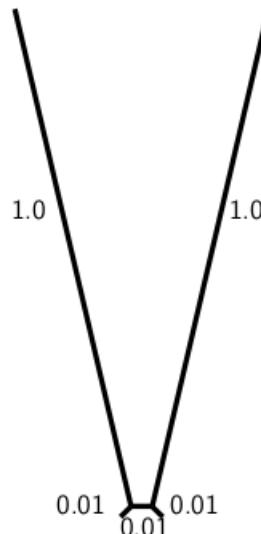
2.1. Máxima parcimônia

Principais críticas:

- Não tem consistência estatística sob algumas circunstâncias: não é garantido que a melhor árvore será encontrada a medida em que se aumenta a quantidade de dados; pelo contrário, pode resultar na árvore incorreta devido ao fenômeno da atração de ramos longos ('long branch attraction')

Situação:

- Dois grupos com taxas evolutivas muito superiores ao restante da árvore (maior número de mudanças acumuladas independentemente)
- Poucos caracteres homólogos compartilhados entre os táxons que são realmente mais próximos
- Maior probabilidade de identidade devido à homoplasia entre os táxons com maior taxa evolutiva (em DNA, apenas 4 estados possíveis, ou seja, 25% de chances de bases serem idênticas ao acaso)



Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27: 401-410.

2.1. Máxima parcimônia

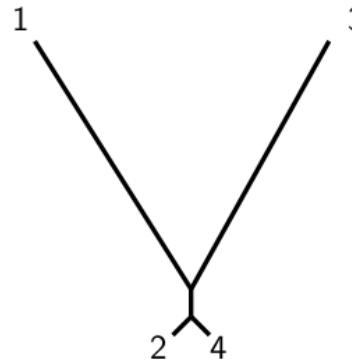
Principais críticas:

- Não tem consistência estatística sob algumas circunstâncias: não é garantido que a melhor árvore será encontrada a medida em que se aumenta a quantidade de dados; pelo contrário, pode resultar na árvore incorreta devido ao fenômeno da atração de ramos longos ('long branch attraction')

Resultado:



Árvore verdadeira



Árvore inferida

2.1. Máxima parcimônia

Principais críticas:

- Não tem consistência estatística sob algumas circunstâncias: não é garantido que a melhor árvore será encontrada a medida em que se aumenta a quantidade de dados; pelo contrário, pode resultar na árvore incorreta devido ao fenômeno da atração de ramos longos ('long branch attraction')
- A ausência de um modelo de evolução para ser testado nos métodos de parcimônia exclui a possibilidade de se incorporar informações adicionais (parâmetros) sobre a evolução, mesmo quando elas são conhecidas — 'a evolução não é parcimoniosa'

Principais pontos: neighbour joining e parcimônia

- Parte essencial do desenvolvimento dos métodos de inferência filogenética
- Mais rápidos que máxima verossimilhança e inferência bayesiana
- Funcionam bem sob condições restritas (taxa evolutiva mais ou menos constante entre ramos da árvore)
 - Podem resultar na árvore incorreta se essas condições não forem atendidas, independentemente da quantidade de dados!

Literatura sugerida

Livros:

- M Nei & S Kumar. 2000. Molecular Evolution and Phylogenetics.
- C Darwin. 1859. A origem das espécies
- RDM Page. 1998. Molecular Evolution: A Phylogenetic Approach
- J Felsenstein. 2003. Inferring Phylogenies

Revisões

- Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* 21: 428–444.
- Baum DA, Offner S. 2008. Phylogenies & Tree-Thinking. *The American Biology Teacher* 70: 222–229.