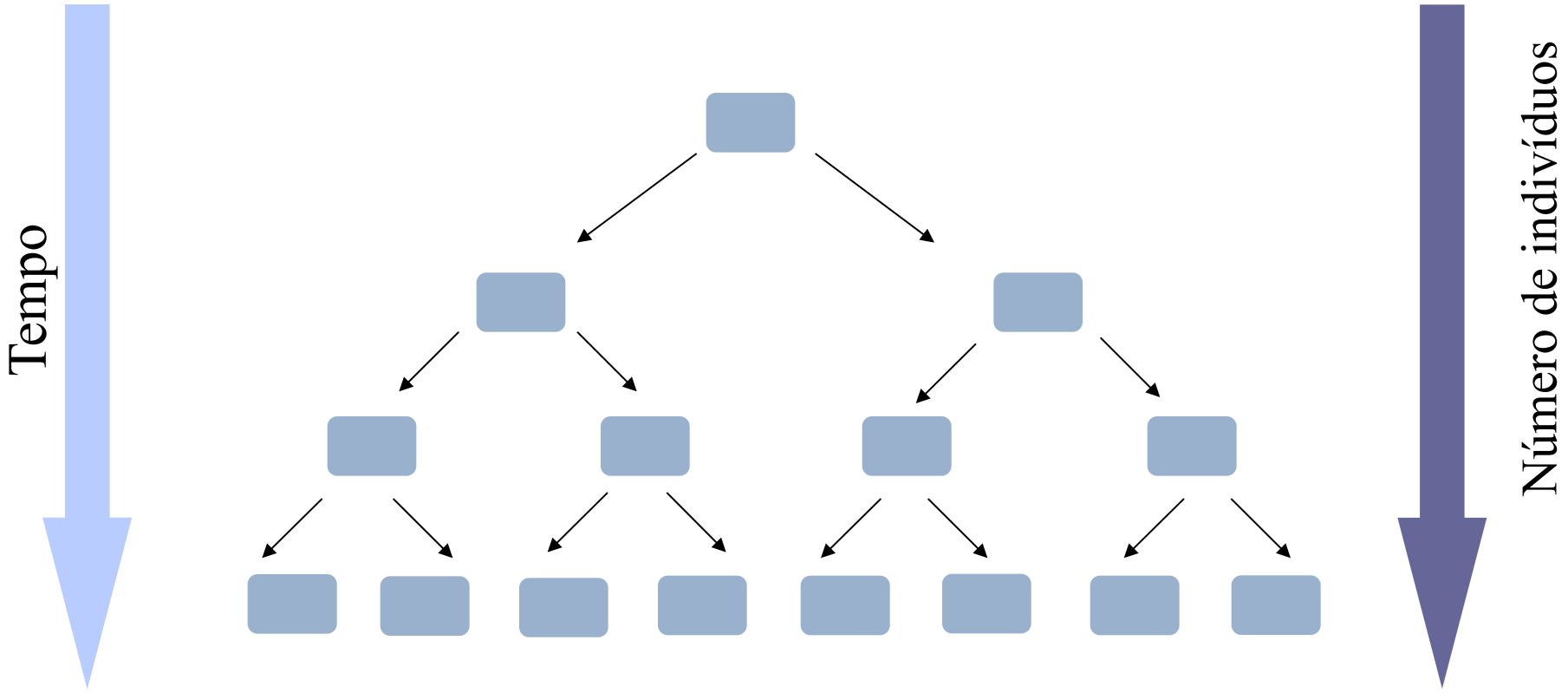
Introduzindo filogenias

Um olhar íntimo nos sistemas vivos

- Os sistemas vivos apresentam as seguintes características básicas
  - Se reproduzem (com o sem sexo);
  - Mantém a informação que carregam ao longo das gerações (hereditariedade);

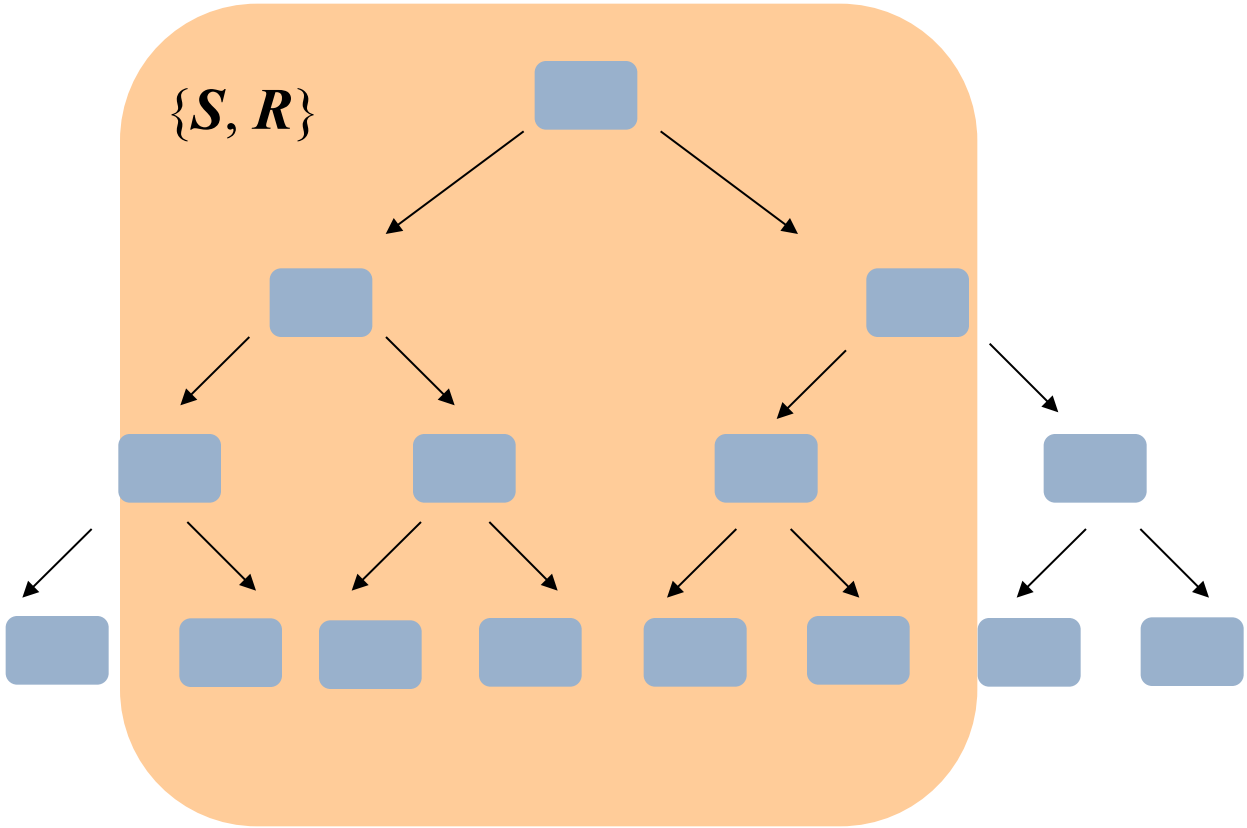
Conseqüentemente..



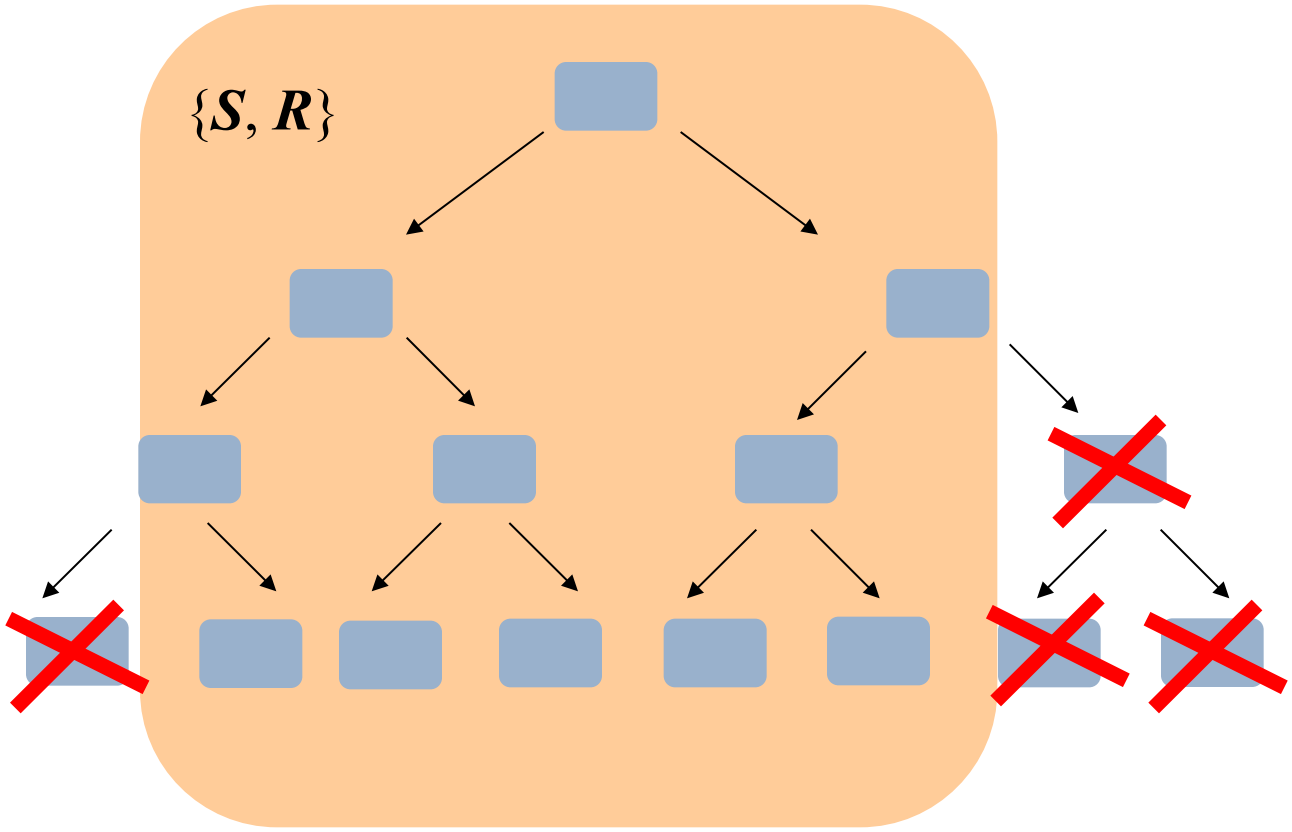
Teoricamente, o número de indivíduos tenderia ao infinito

Isso não aconteceu. Porque?

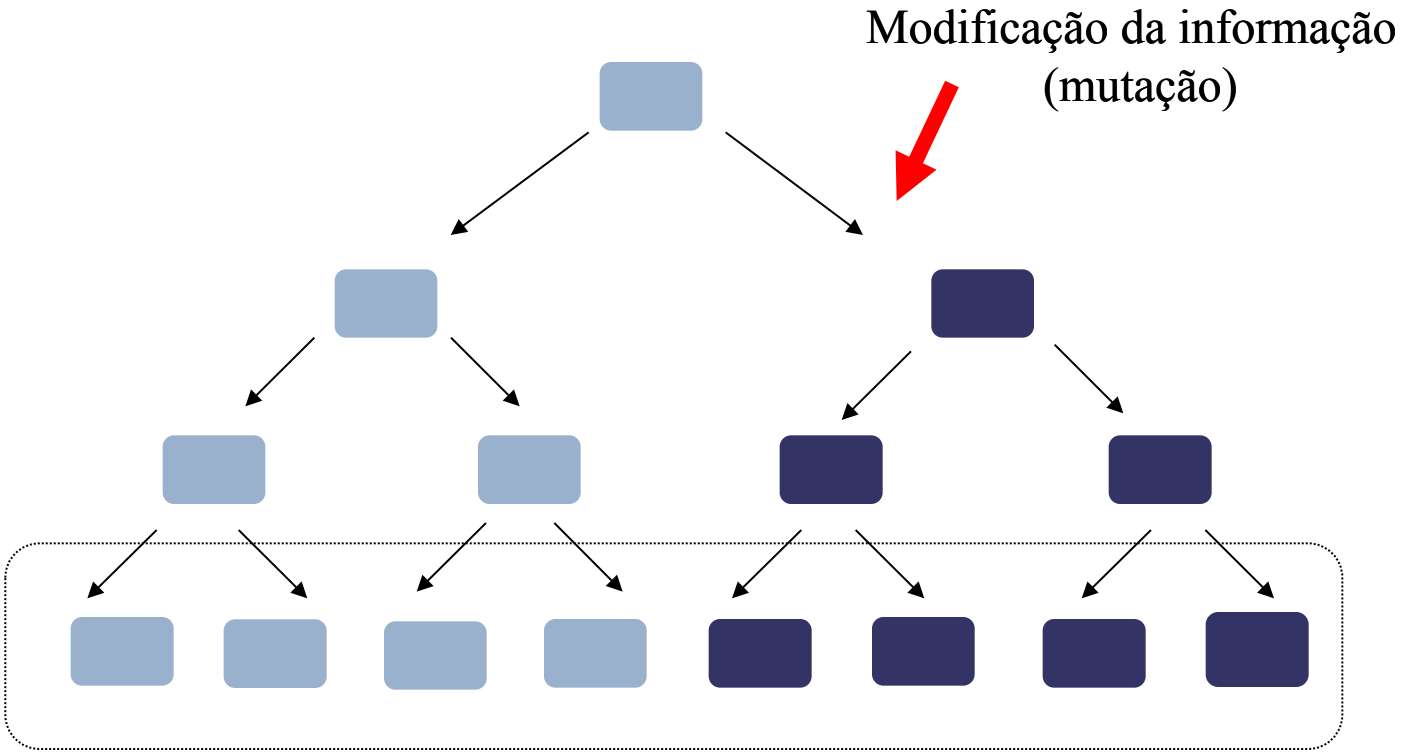
- Existe um número máximo de indivíduos ocupantes da mesma região *S* do espaço, dividindo os recursos *R*.



Portanto...

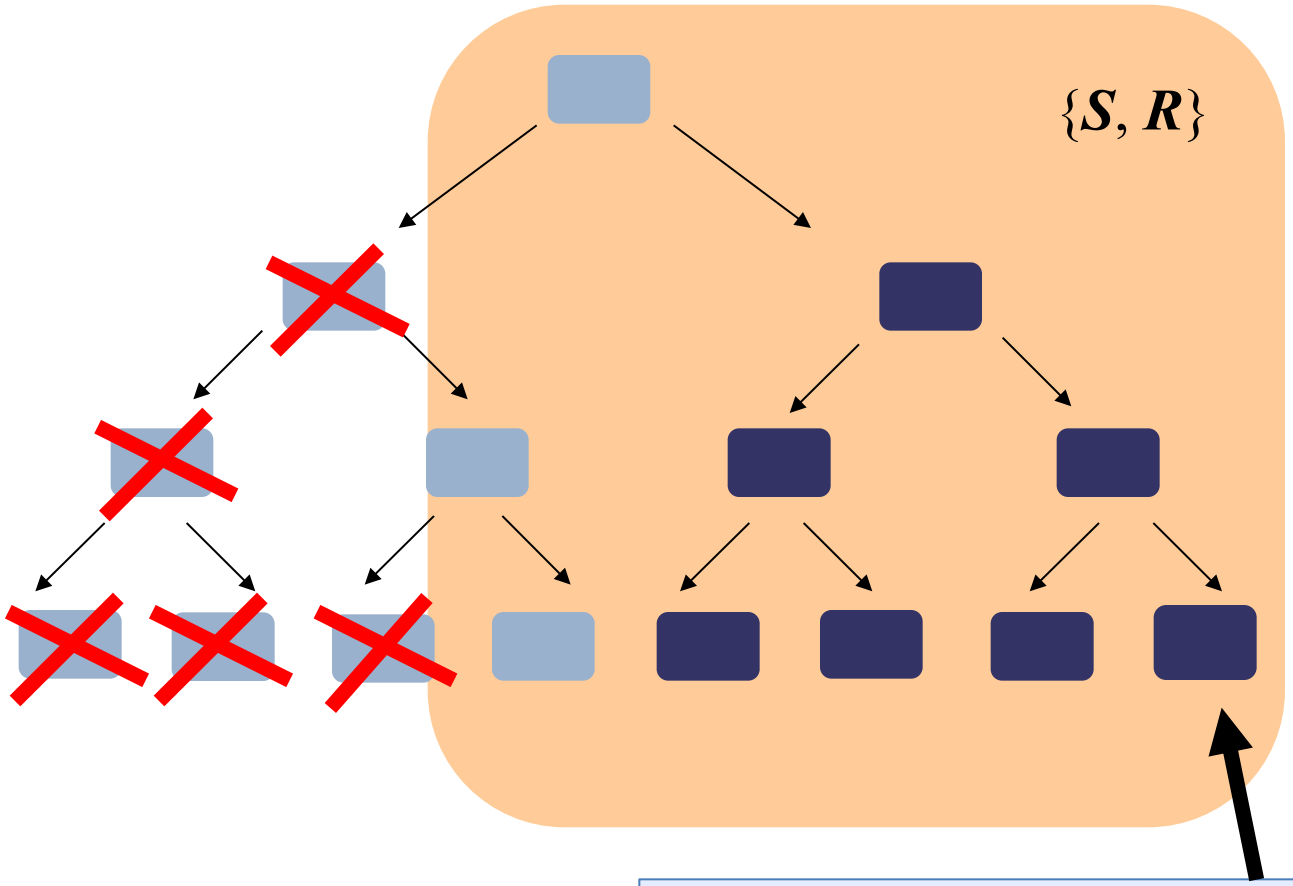


Além disso...



Variação
Propriedade de um grupo de indivíduos

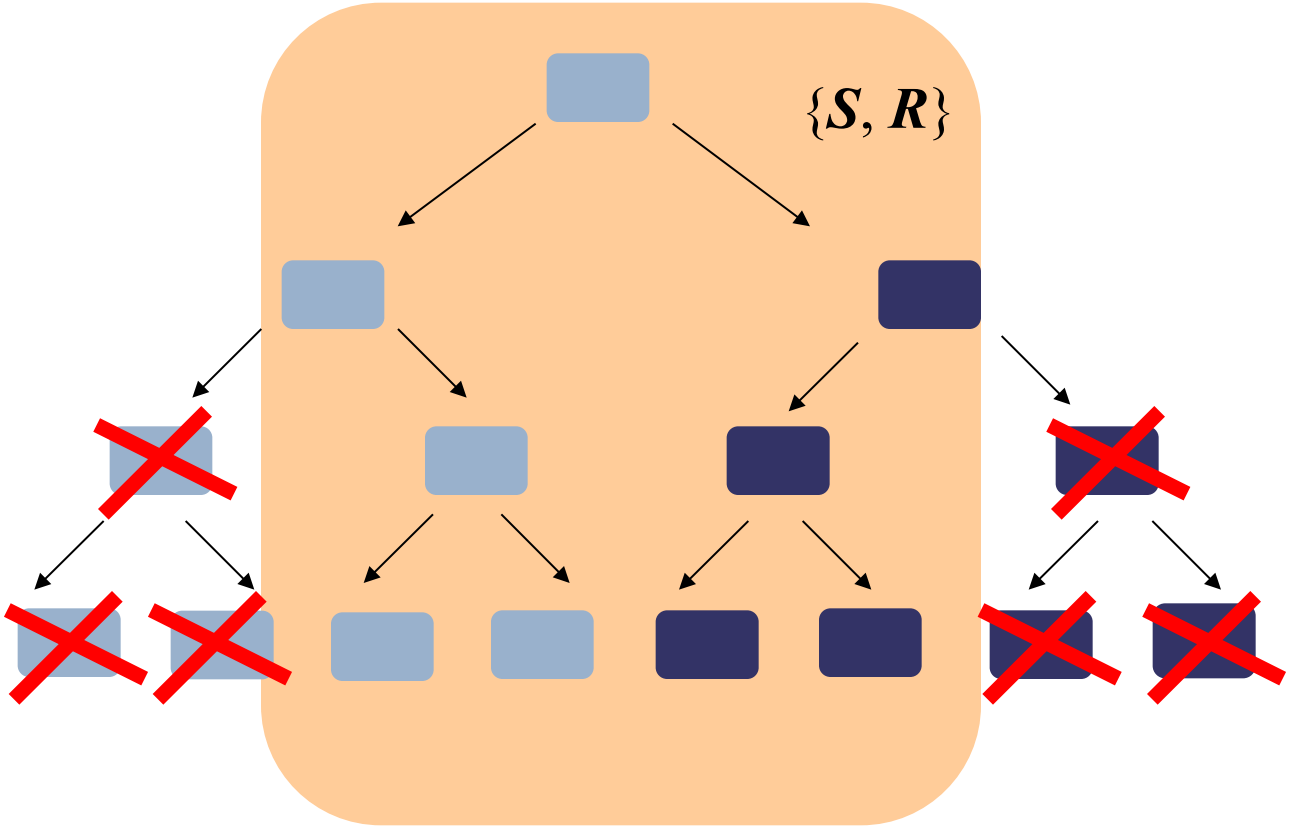
Então...



Concluindo

- Grupos de **sistemas informacionais** que se reproduzem, apresentam variação e que dividem o hiperespaço { *S*, *R* } estão sujeitos ao sucesso reprodutivo diferencial, ou seja, **seleção natural**.

Podemos também admitir o seguinte



O sucesso reprodutivo é equivalente em ambos os tipos

Nesse caso...

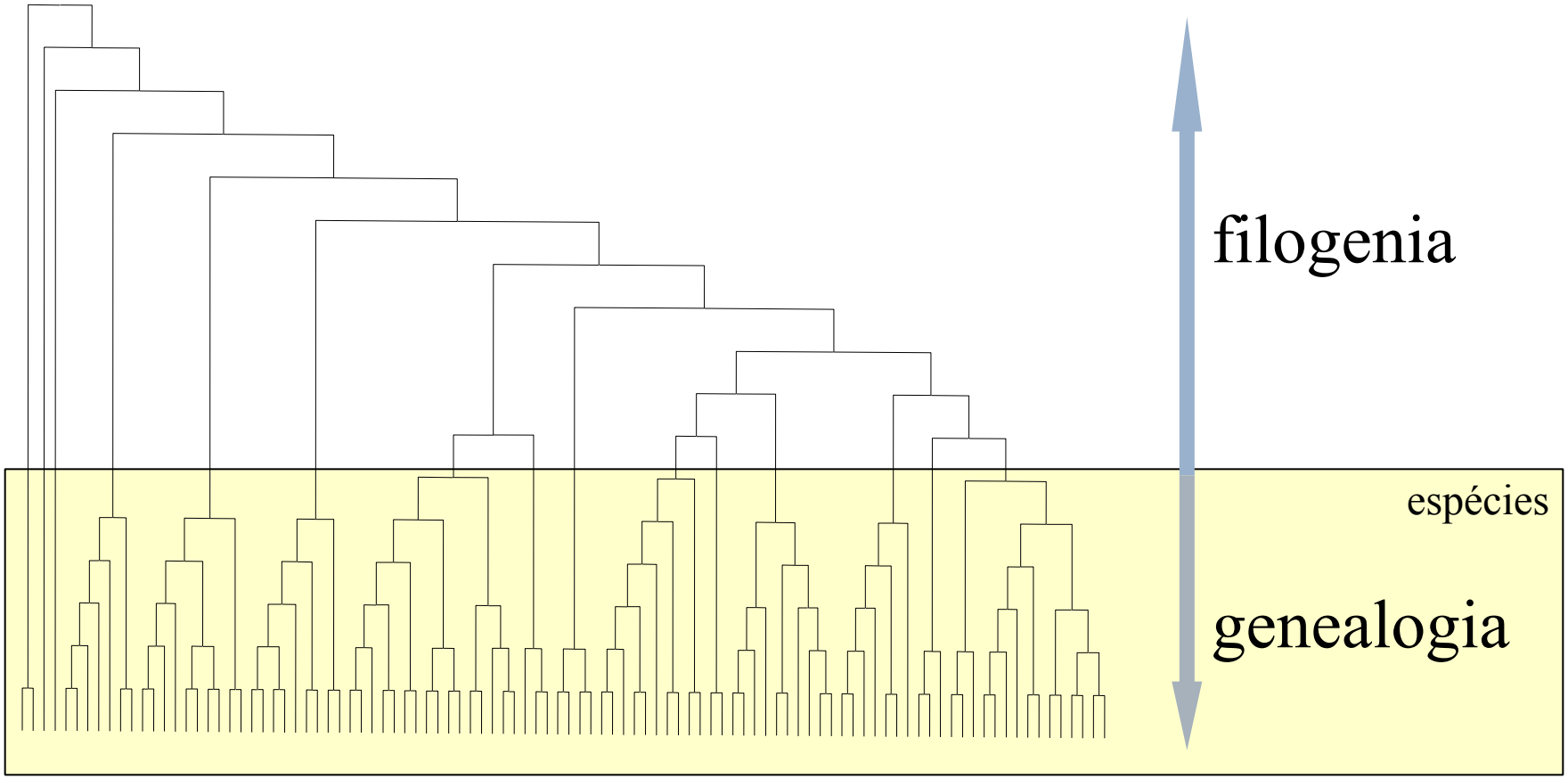
- A modificação das frequências dos tipos será randômica e dependente do número de indivíduos existentes em {*S*, *R*}
- Esse fenômeno recebe o nome de **deriva genética**.

Tornando a conversa mais concreta

- O esquema anterior pode ser usado para representar, por exemplo, as relações ancestral-descendente de uma população clonal de bactéria.
- Tecnicamente, mostramos a **genealogia** dessas bactérias
- Essa é uma representação das relações de parentesco. Poderíamos chamar isso de filogenia?

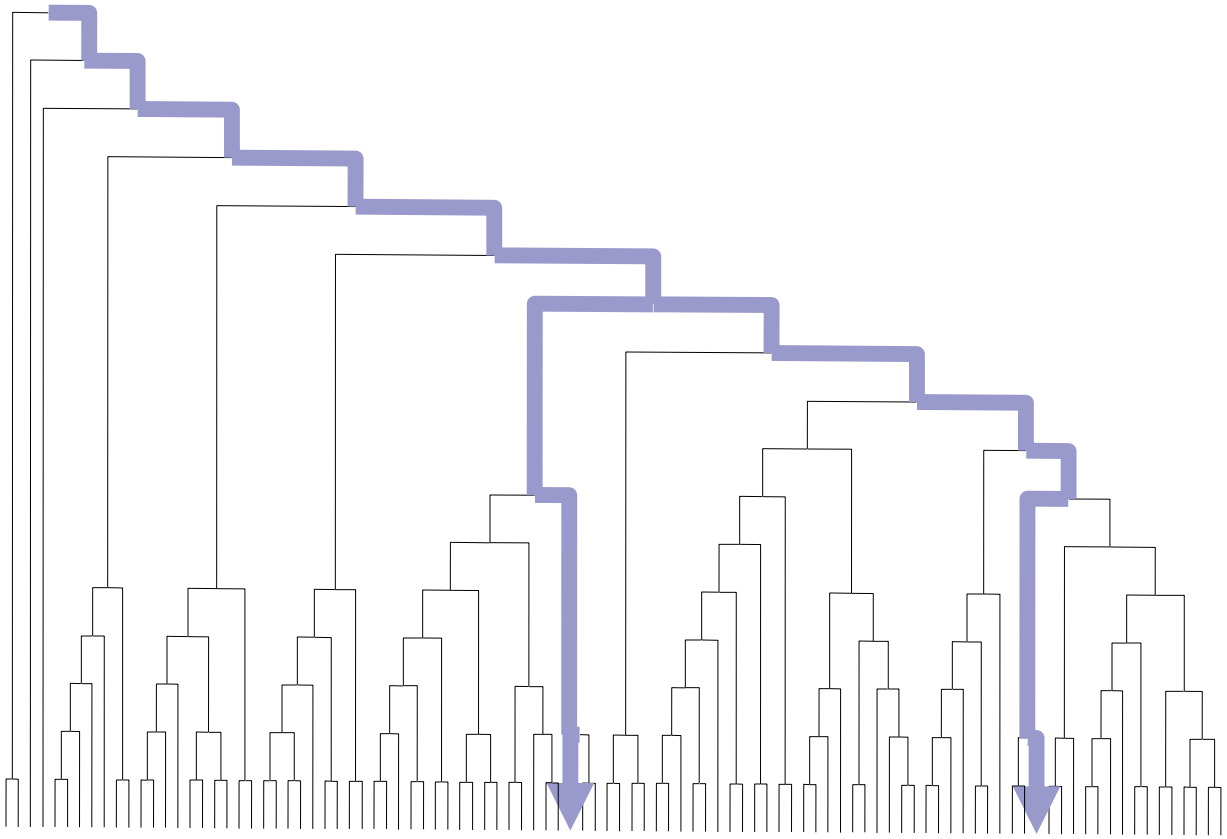
Genealogia e Filogenia

- A diferença entre genealogia e filogenia é tênue.



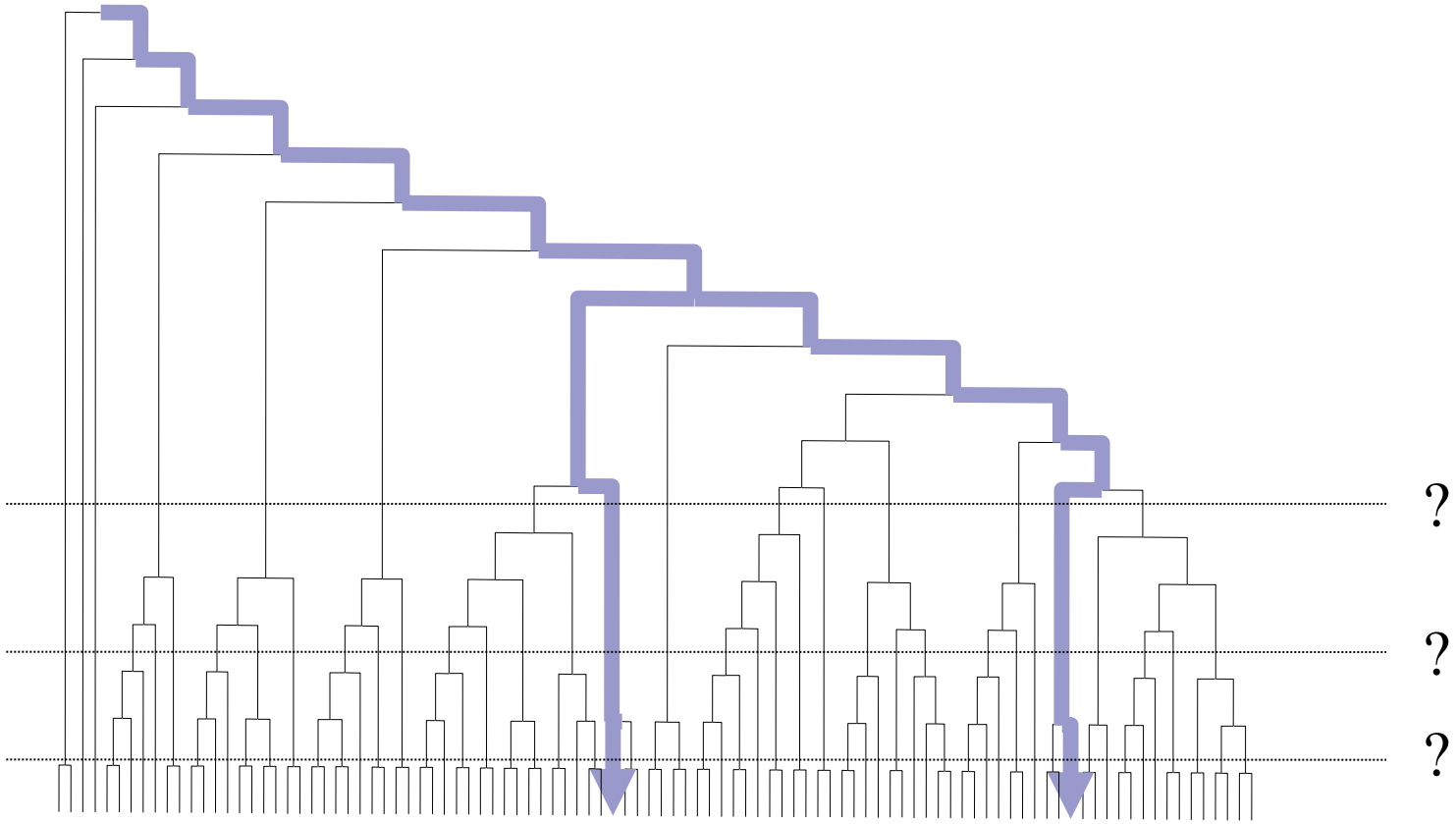
Por que?

- O fluxo de informação é **contínuo** no tempo.



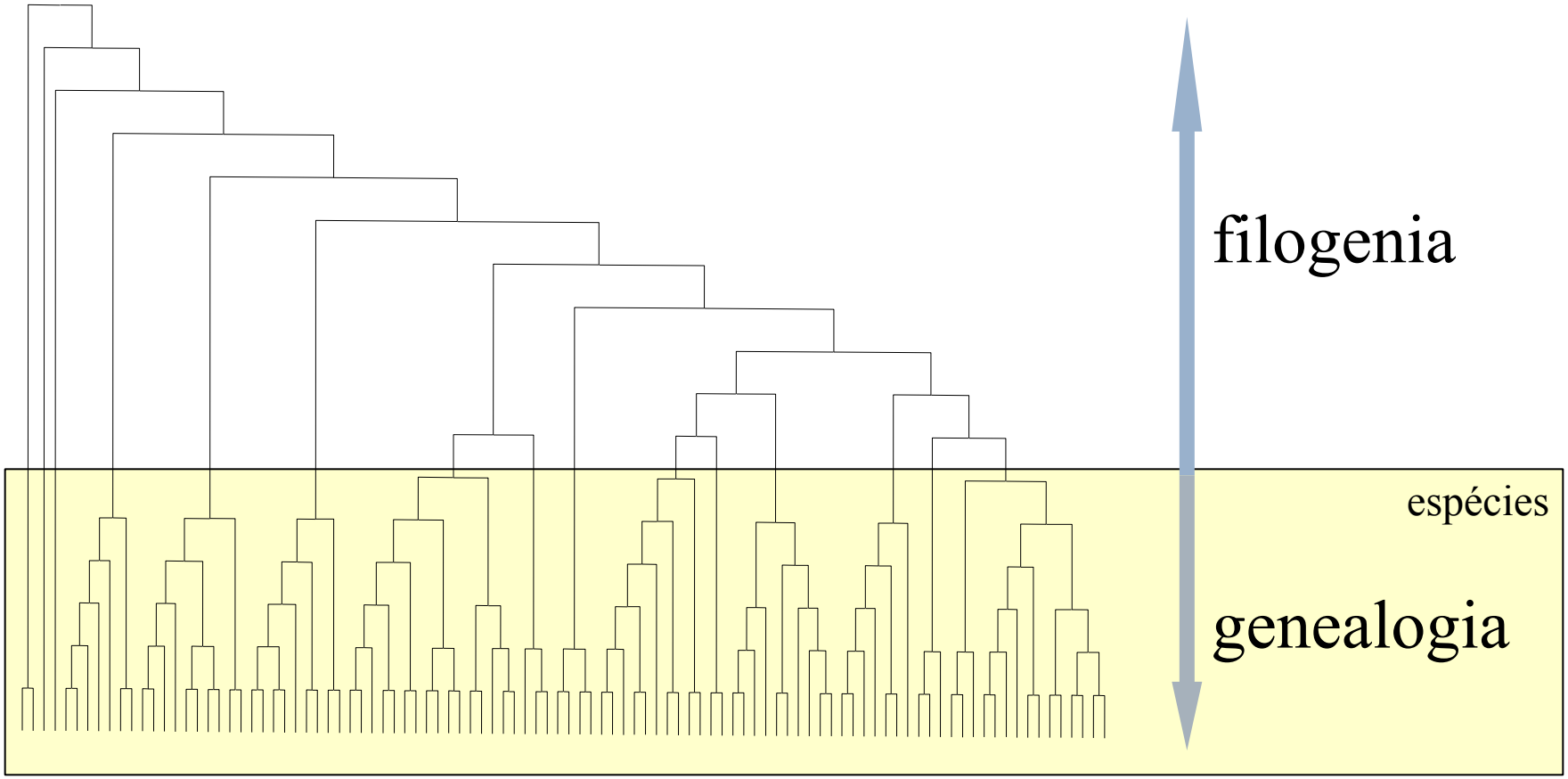
Portanto...

- Qualquer “corte” é necessariamente arbitrário



Entreteanto...

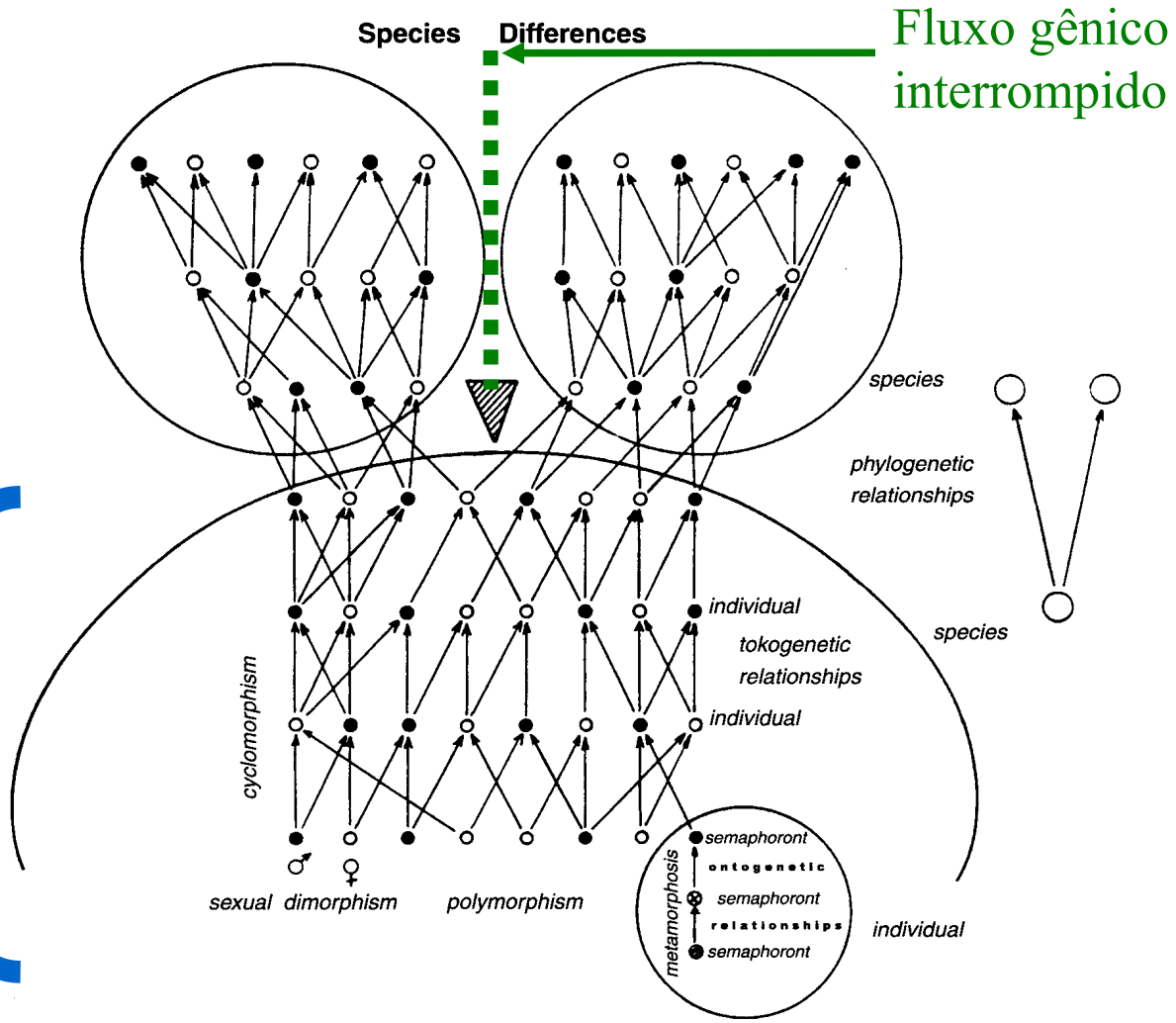
- Se pensarmos em sistemas que trocam informação (ex, indivíduos de uma espécie com reprodução sexuada), o “corte” fica melhor definido.

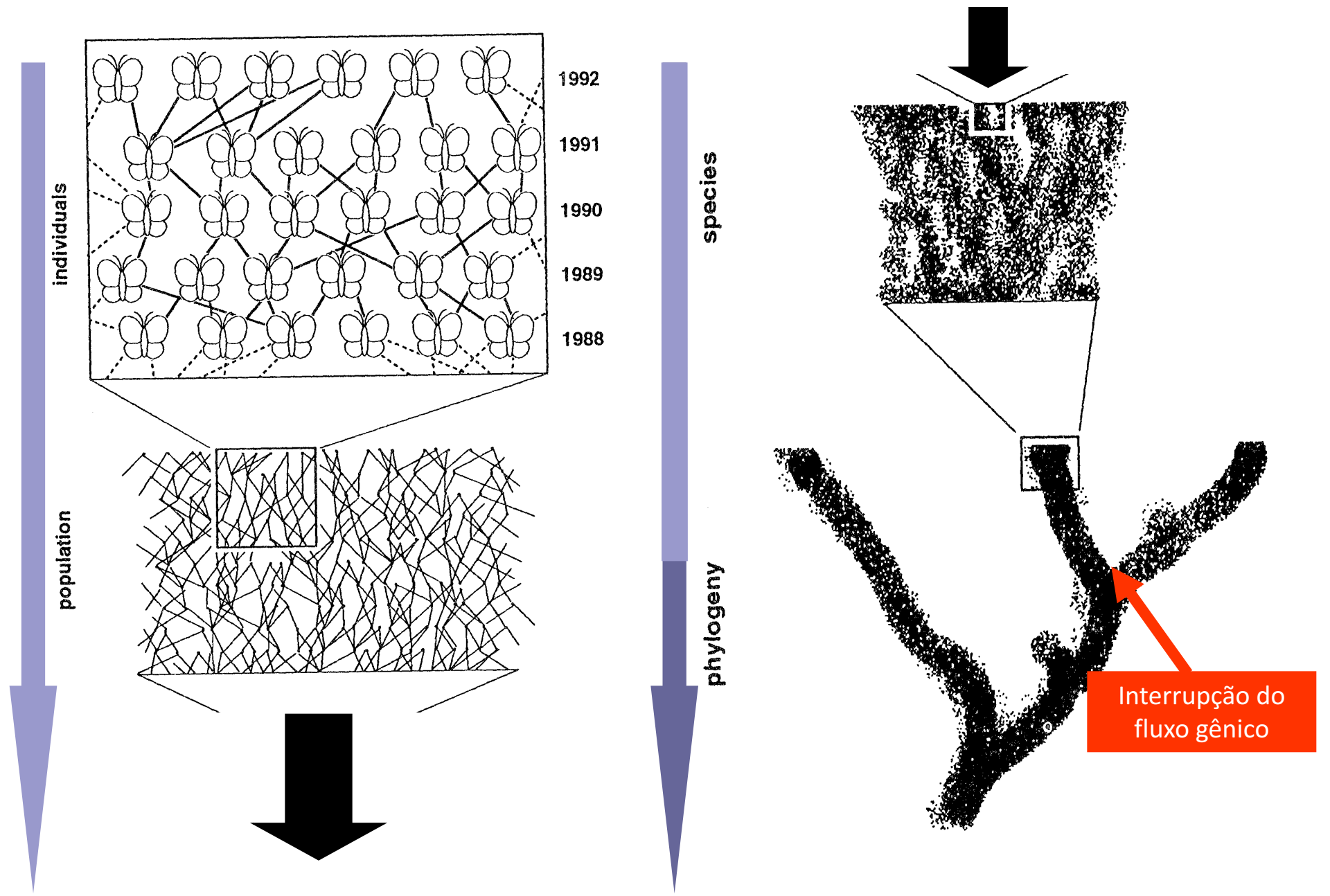


Um *zoom* nas relações de parentesco

filogenia

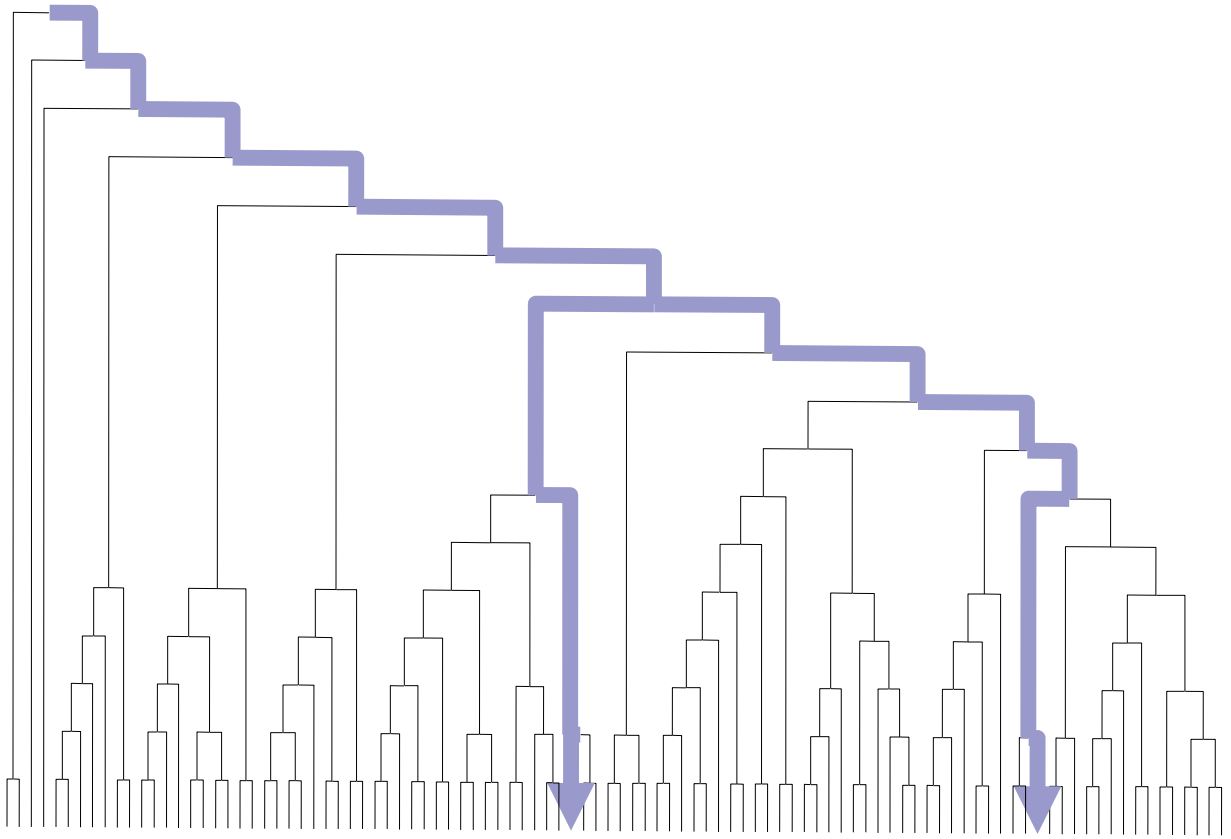
genealogia




individuals
1992
1991
1990
1989
1988
population
species
phylogeny
Interrupção do fluxo gênico

Afastando mais o *zoom*

Entretanto, mesmo em espécies sexuadas, o fluxo vertical não é quebrado



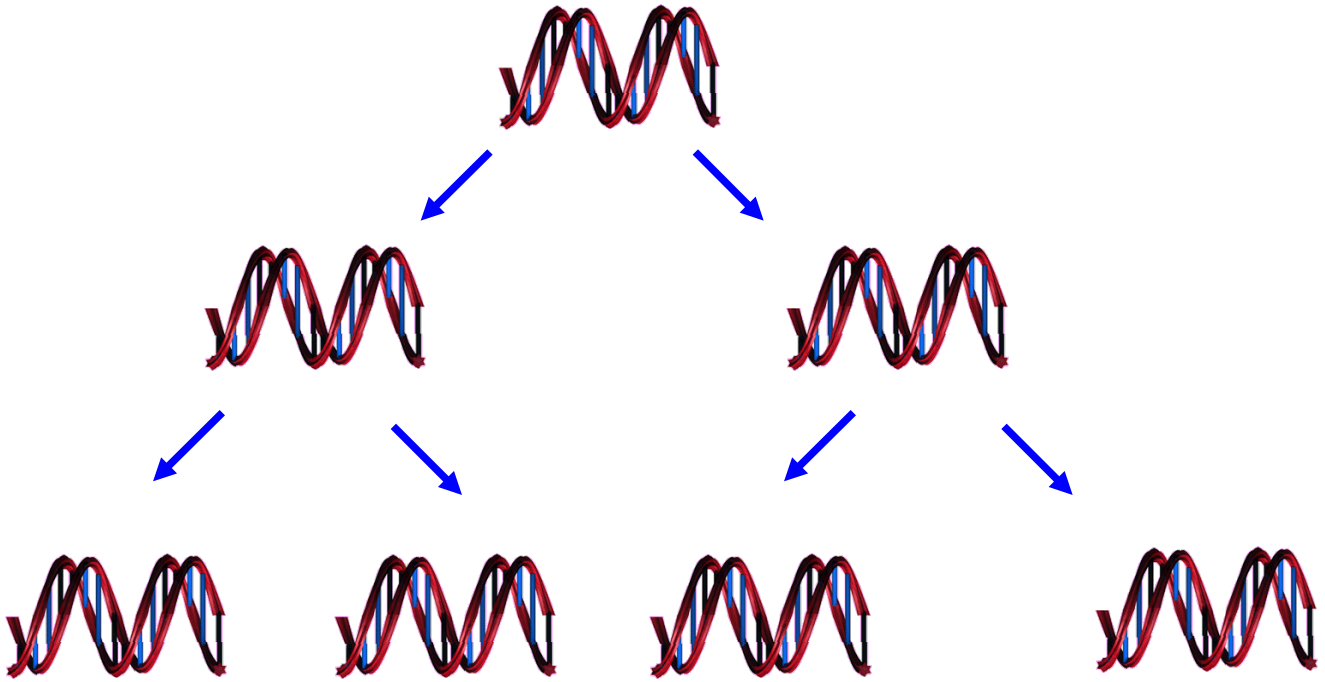
Uma forma unificada de acompanhar o fluxo de informação nos sistemas vivos

- Já que a molécula que carrega a informação nos sistemas vivos é o DNA, porque não acompanhar as relações ancestral-descendente através dela?

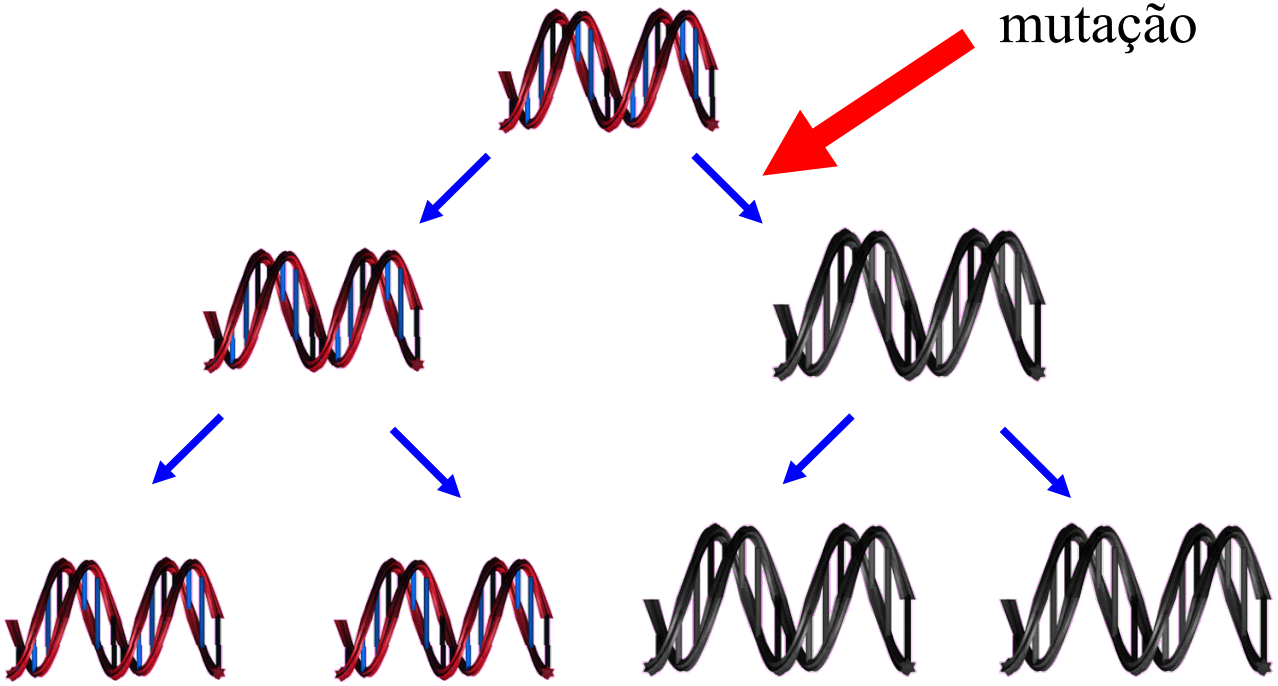
- Entretanto, “uma molécula de DNA” é algo abstrato.. a unidade fundamental da informação é o **gene**.

Ao invés de “indivíduos” vamos considerar genes

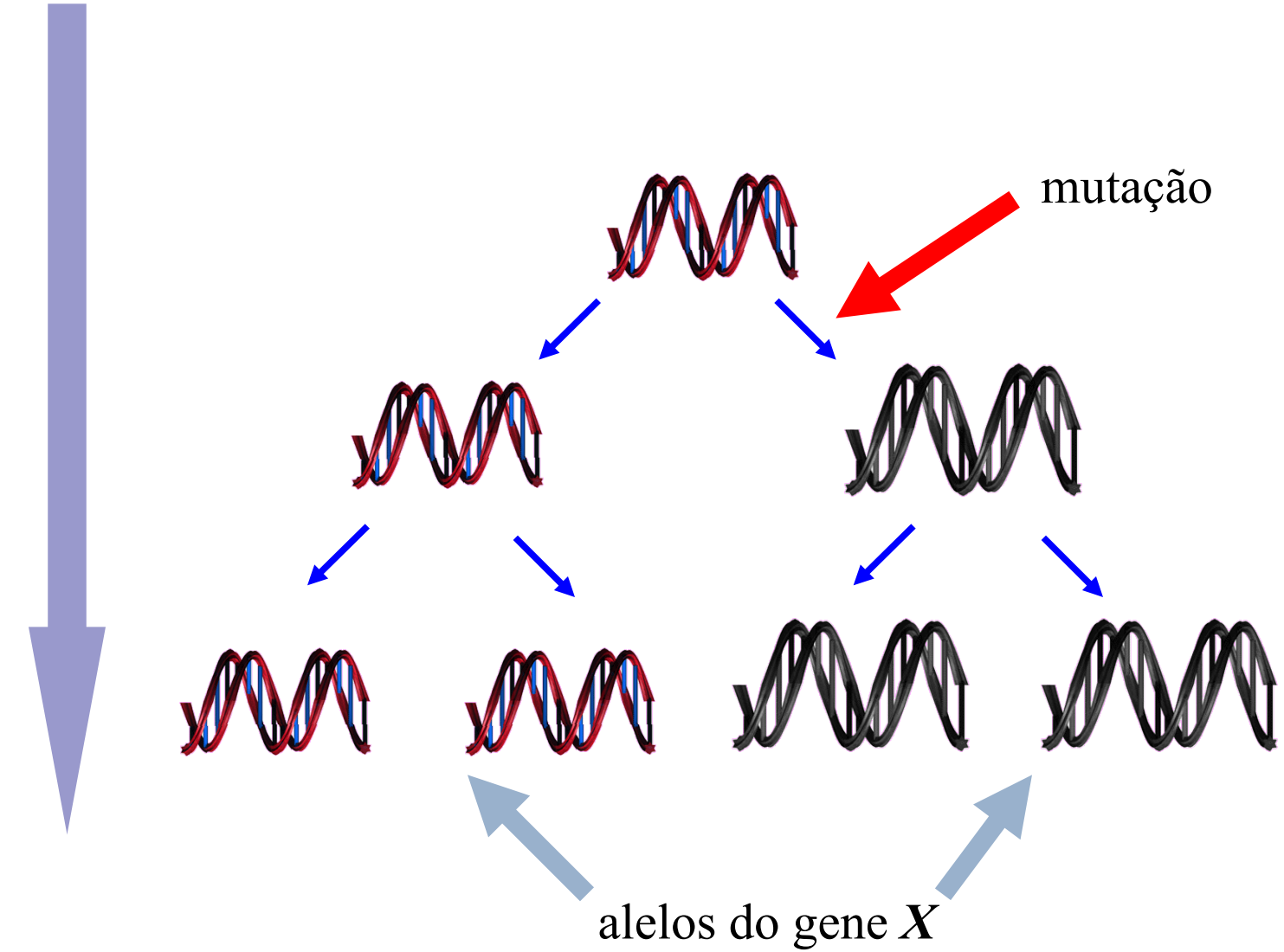
- Por exemplo, vamos acompanhar um gene ***X*** ao longo do tempo.



Eventualmente...



Alelos



Definindo o problema









?

Qual é a filogenia desses organismos ?

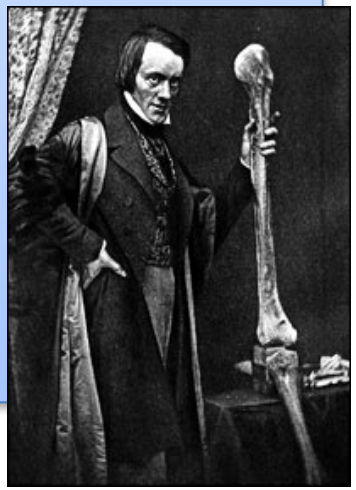
Passo 0: Homologia

- Se você quer estudar as relações evolutivas do vertebrados, você deverá comparar regiões homólogas do corpo

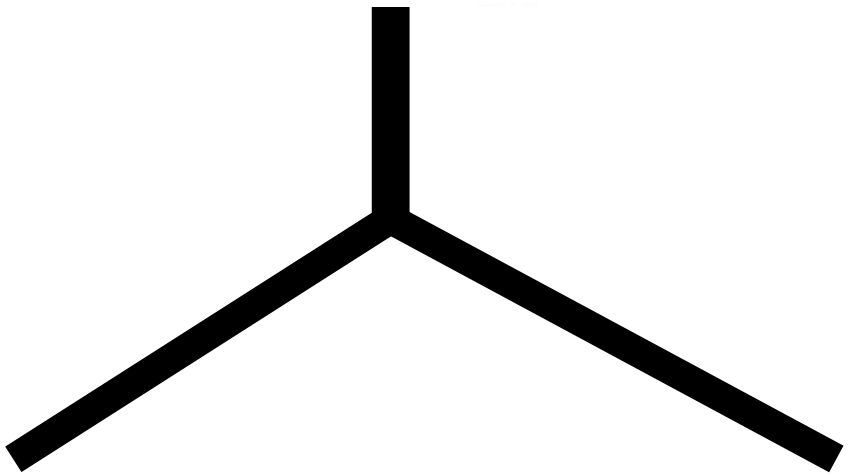
Definindo homologia

- Inicialmente, o conceito de homologia foi estabelecido por Richard Owen

- Duas estruturas são ditas homólogas se elas são originárias de uma mesma estrutura ancestral

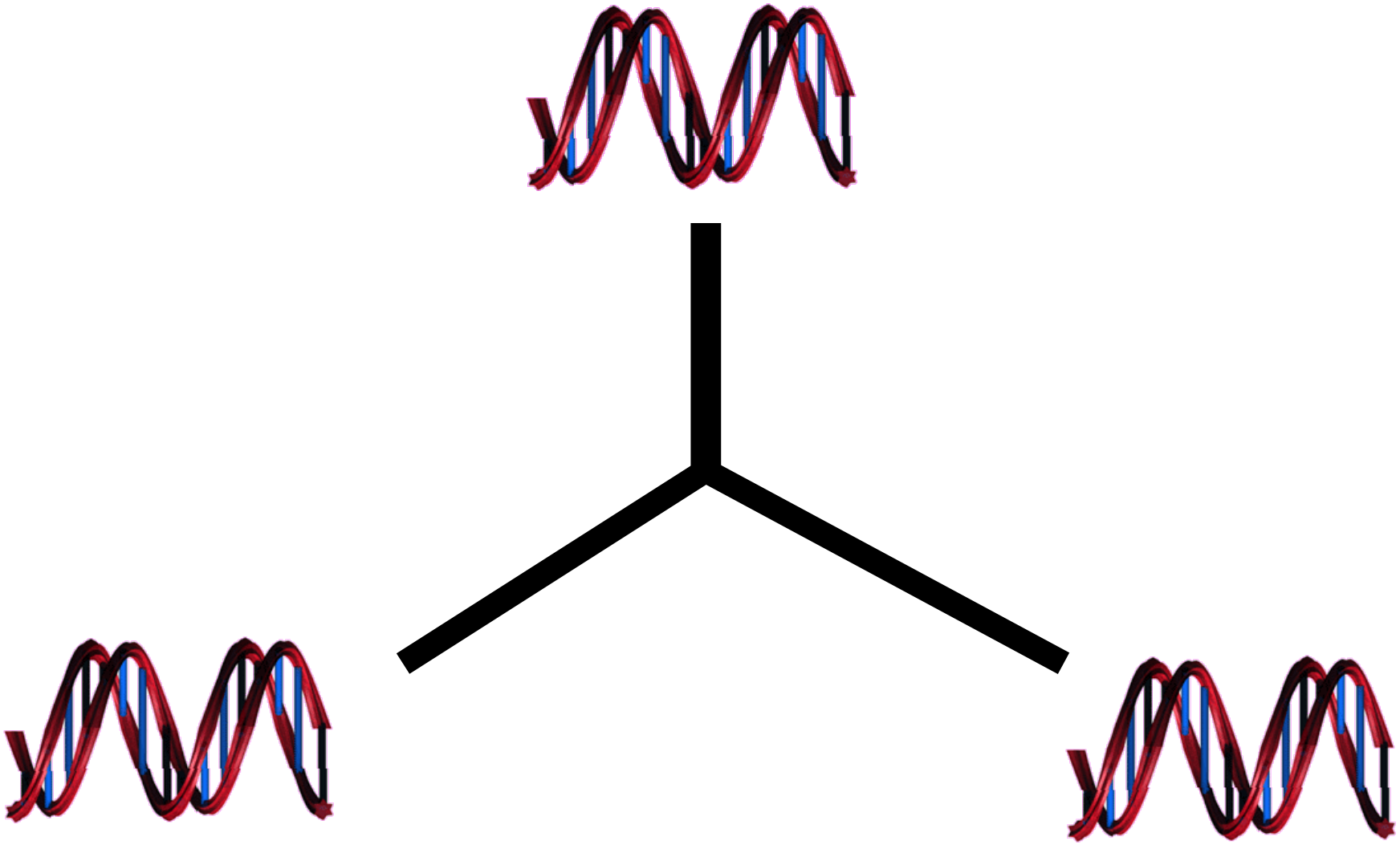


Portanto...

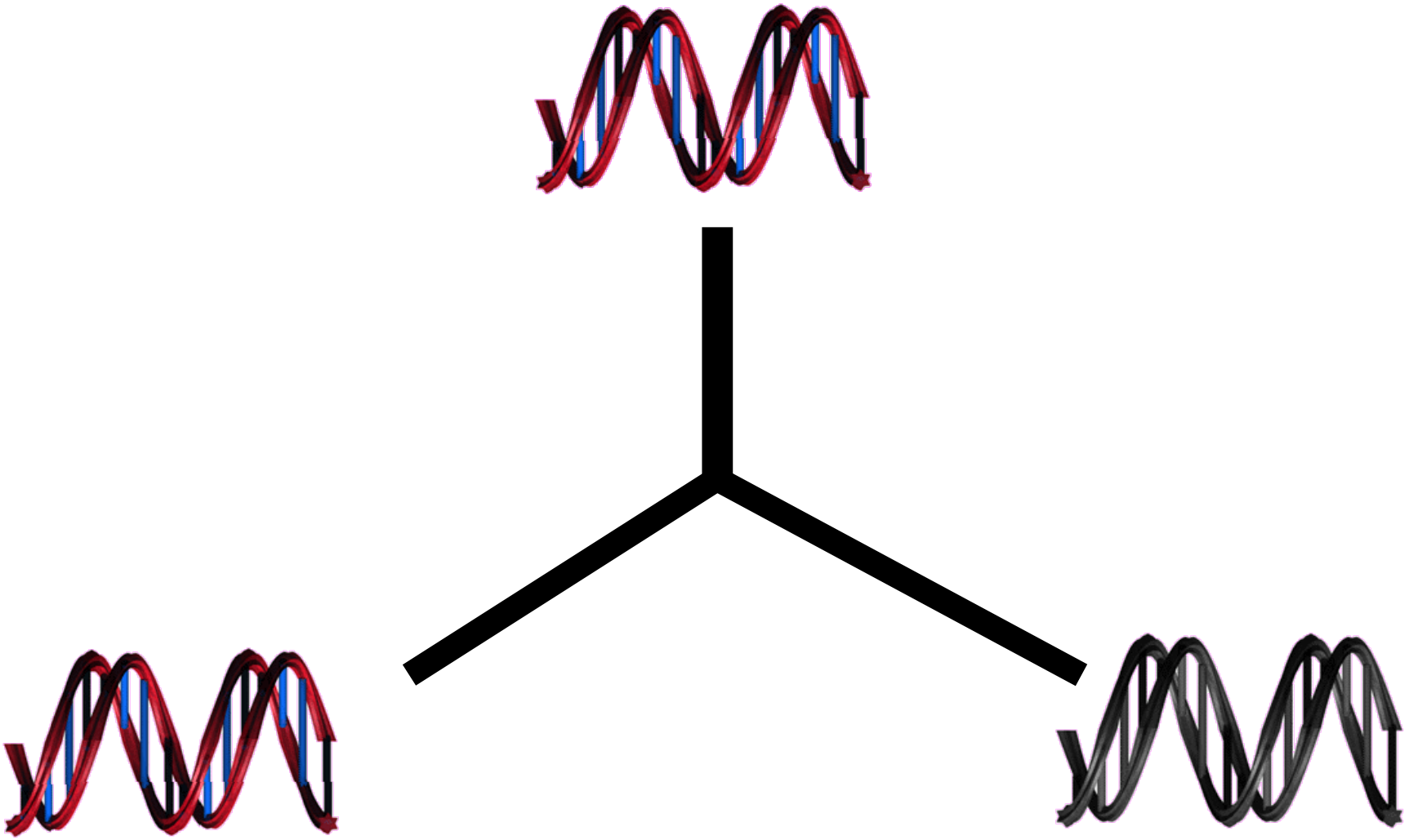


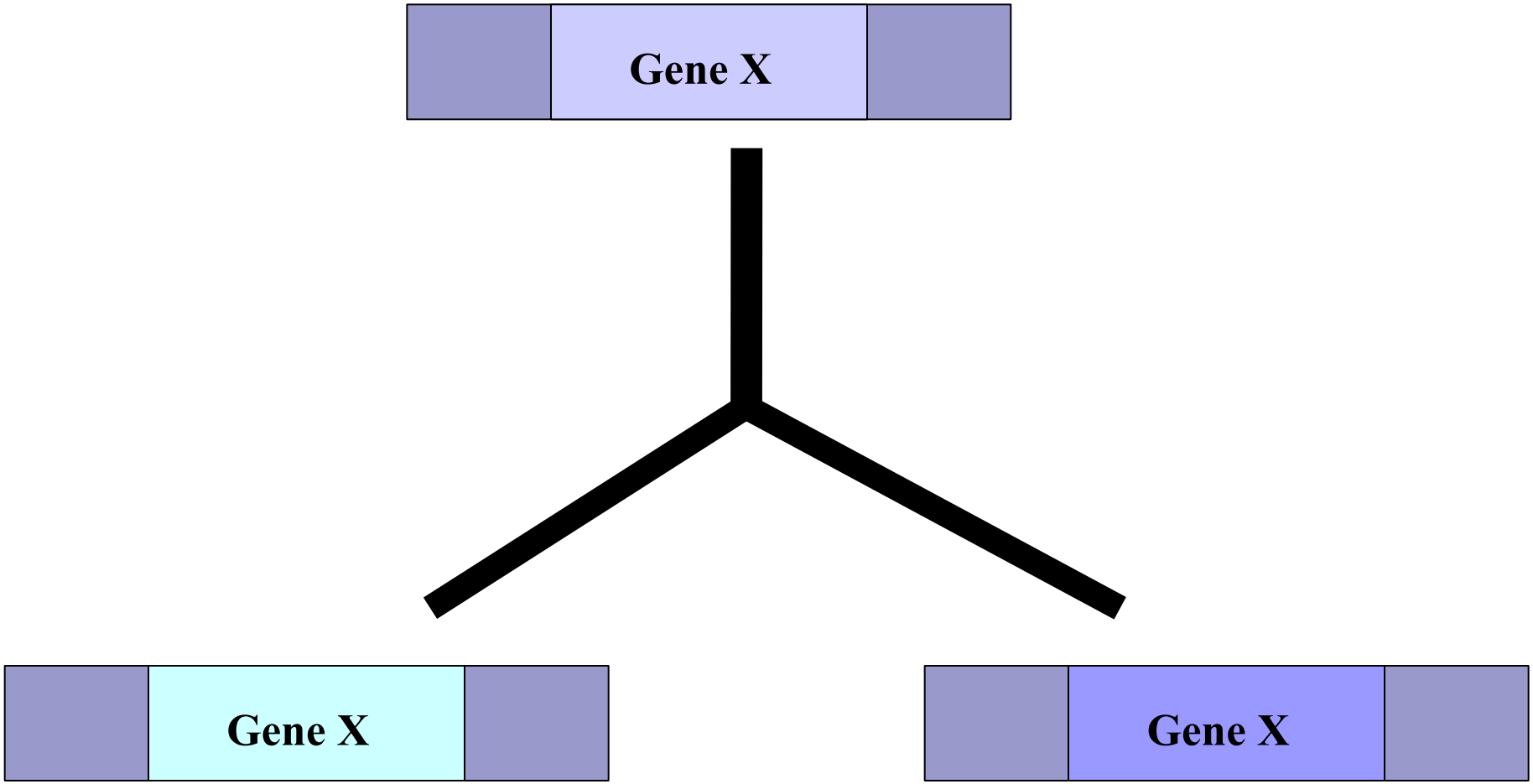
crânio dos primatas

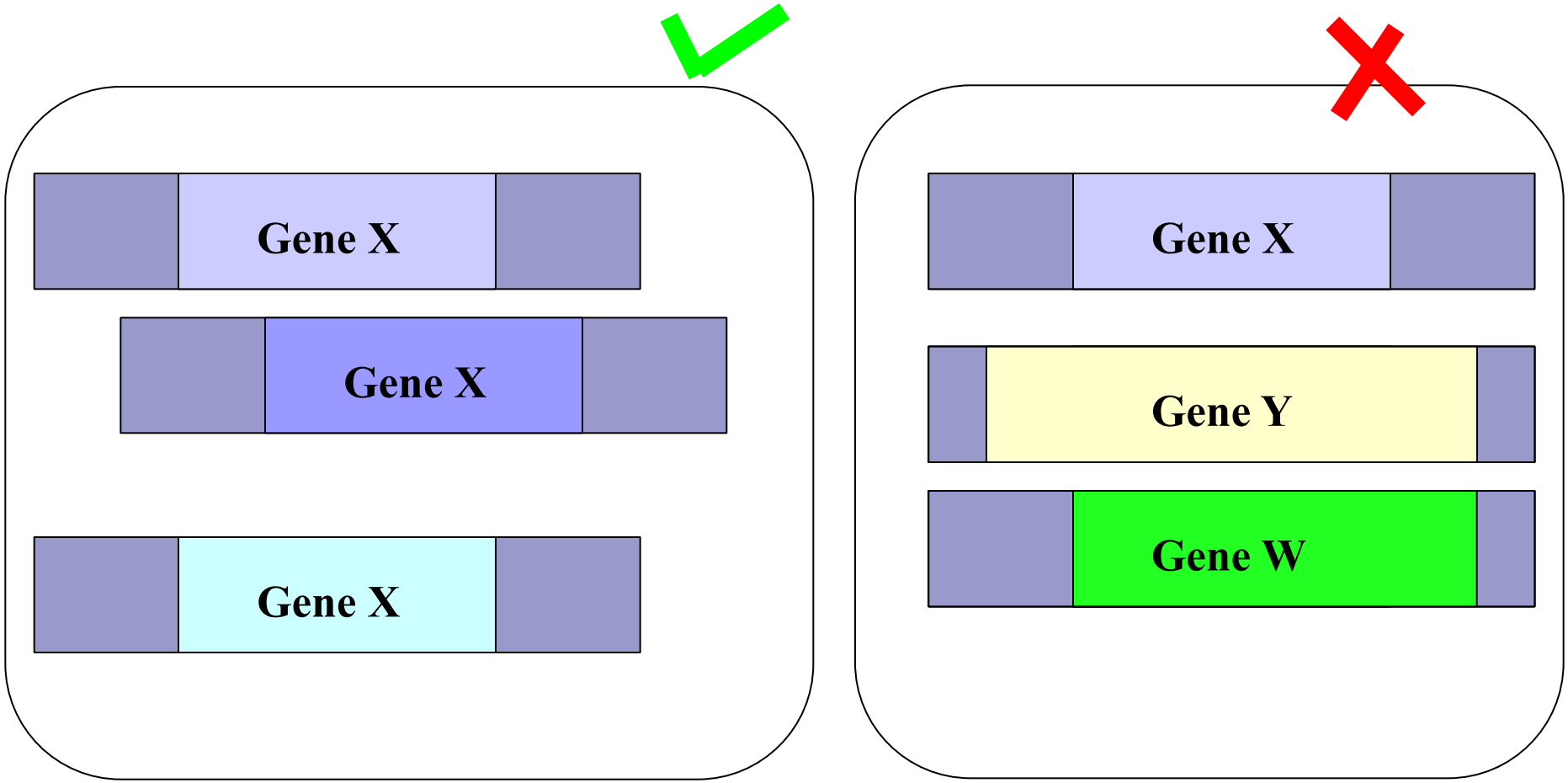
Molecularmente



Alelos são homólogos




Gene X
Gene X
Gene X


Gene X
Gene X
Gene X
Gene X
Gene Y
Gene W

Voltando ao problema...









?

Qual é a filogenia desses organismos ?

Uma solução..





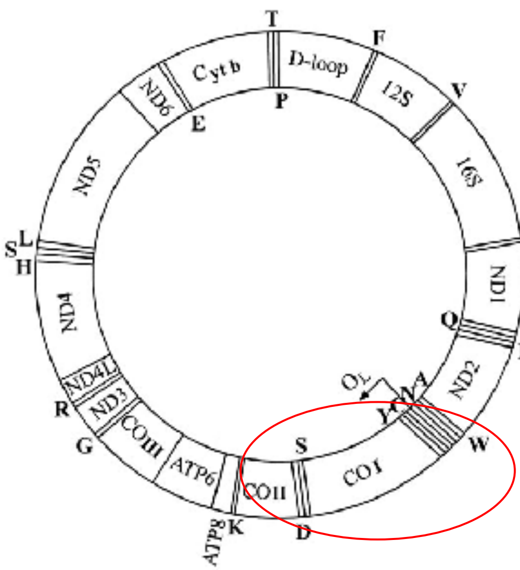


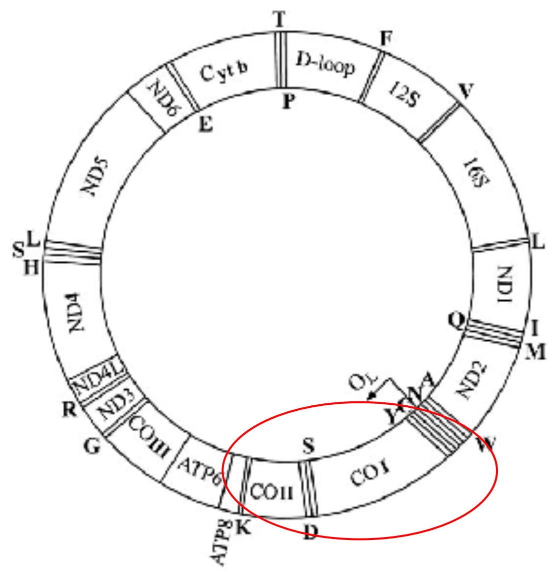


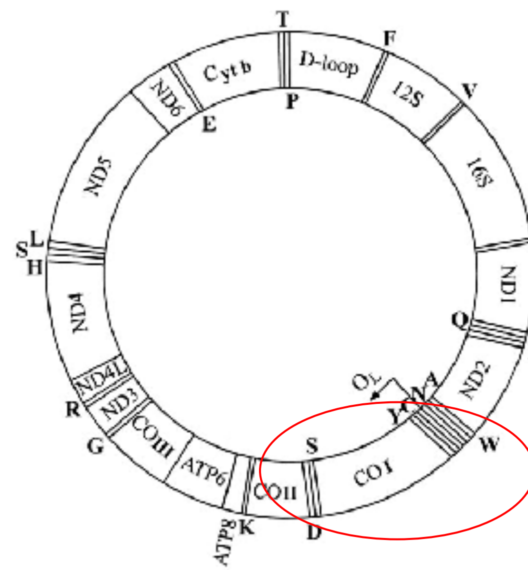
Sequence a homologous gene

Por exemplo, o *COI*



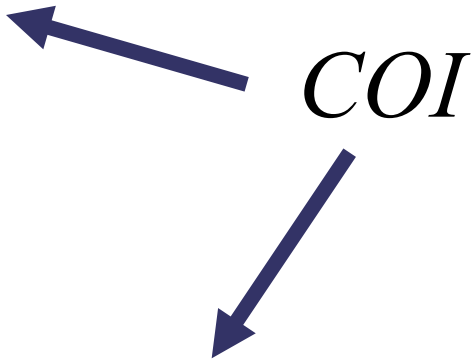






Após sequenciamento

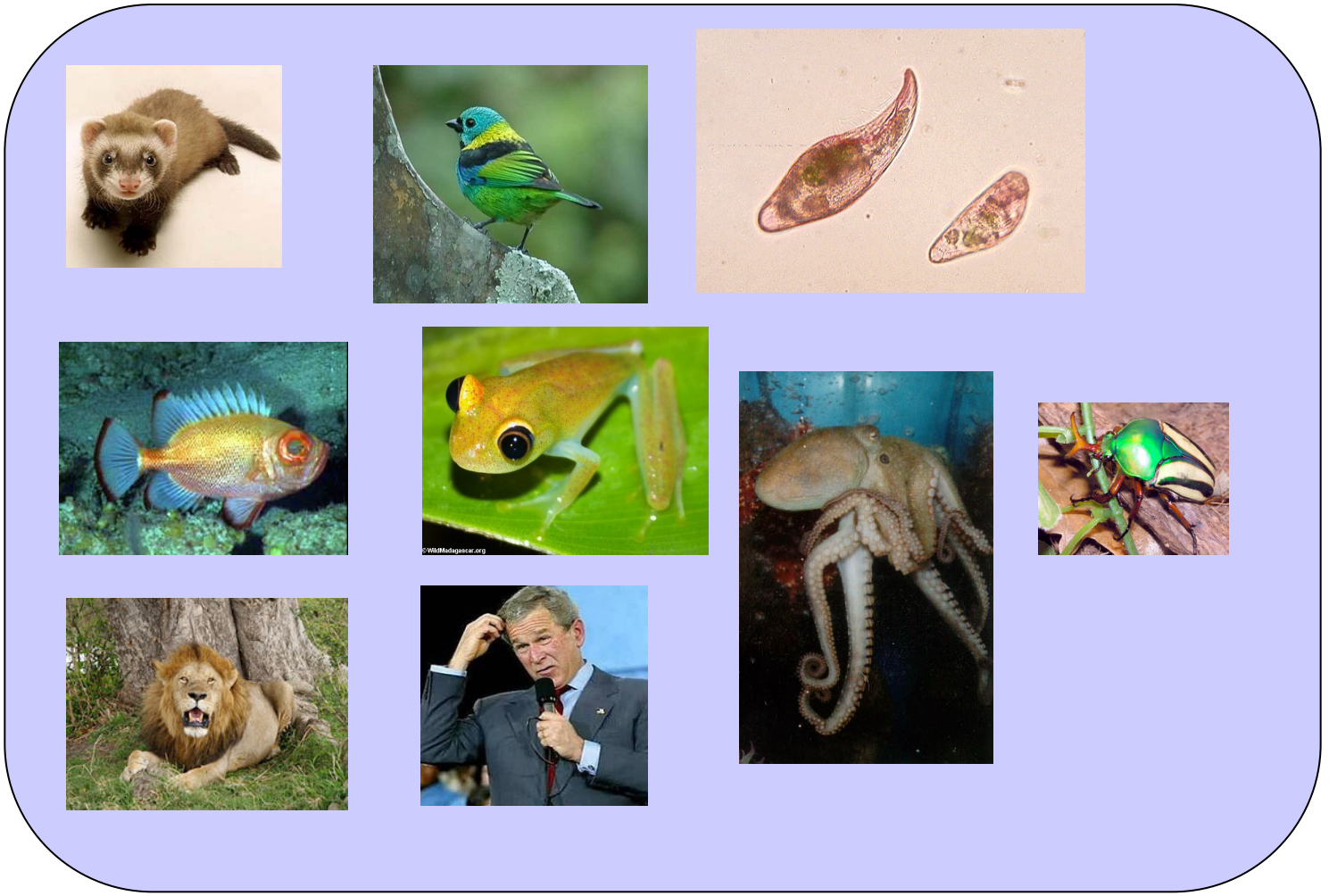
>bush
ATGTTCGCCGACCGTTGACTATTCTCTACAAACCACAAAGACATTGGAACACTATACCTATTATTCGGCG
CATGAGCTGGAGTCCTAGGCACAGCTCTAAGCCTCCTTATTCGAGCCGAGCTGGGCCAGCCAGGCAACCT
TCTAGGTAACGACCACATCTACAACGTTATCGTCACAGCCCATGCATTTGTAATAATCTTCTTCATAGTA
ATACCCATCATAATCGGAGGCTTTGGCAACTGACTAGTTCCCCTAATAATCGGTGCCCCCGATATGGCGT
TTCCCCGCATAAACAACATAAGCTTCTGACTCTTACCTCCCTCTCTCCTACTCCTGCTCGCATCTGCTAT
AGTGGAGGCCGGAGCAGGAACAGGTTGAACAGTCTACCCTCCCTTAGCAGGGAACTACTCCCACCCTGGA
GCCTCCGTAGACCTAACCATCTTCTCCTTACACCTAGCAGGTGTCTCCTCTATCTTAGGGGCCATCAATT
TCATCACAACAATTATCAATATAAAACCCCCTGCCATAACCCAATACCAAACGCCCCTCTTCGTCTGATC
CGTCCTAATCACAGCAGTCCTACTTCTCCTATCTCTCCCAGTCCTAGCTGCTGGCATCACTATACTACTA
ACAGACCGCAACCTCAACACCACCTTCTTCGACCCCGCCGGAGGAGGAGACCCCATTCTATACCAACACC
TATTCTGATTTTTCGGTCACCCTGAAGTTTATATTCTTATCCTACCAGGCTTCGGAATAATCTCCCATAT
TGTAACTTACTACTCCGGAAAAAAAGAACCATTTGGATACATAGGTATGGTCTGAGCTATGATATCAATT
GGCTTCCTAGGGTTTATCGTGTGAGCACACCATATATTTACAGTAGGAATAGACGTAGACACACGAGCAT
ATTTCACCTCCGCTACCATAATCATCGCTATCCCCACCGGCGTCAAAGTATTTAGCTGACTCGCCACACT
CCACGGAAGCAATATGAAATGATCTGCTGCAGTGCTCTGAGCCCTAGGATTCATCTTTCTTTTCACCGTA
GGTGGCCTGACTGGCATTGTATTAGCAAACTCATCACTAGACATCGTACTACACGACACGTACTACGTTG
TAGCTCACTTCCACTATGTCCTATCAATAGGAGCTGTATTTGCCATCATAGGAGGCTTCATTCACTGATT
TCCCCTATTCTCAGGCTACACCCTAGACCAAACCTACGCCAAAATCCATTTCACTATCATATTCATCGGC
GTAAATCTAACTTTCTTCCCACAACACTTTCTCGGCCTATCCGGAATGCCCCGACGTTACTCGGACTACC
CCGATGCATACACCACATGAAACATCCTATCATCTGTAGGCTCATTCATTTCTCTAACAGCAGTAATATT
AATAATTTTCATGATTTGAGAAGCCTTCGCTTCGAAGCGAAAAGTCCTAATAGTAGAAGAACCCTCCATA
AACCTGGAGTGACTATATGGATGCCCCCCACCCTACCACACATTCGAAGAACCCGTATACATAAAATCTA



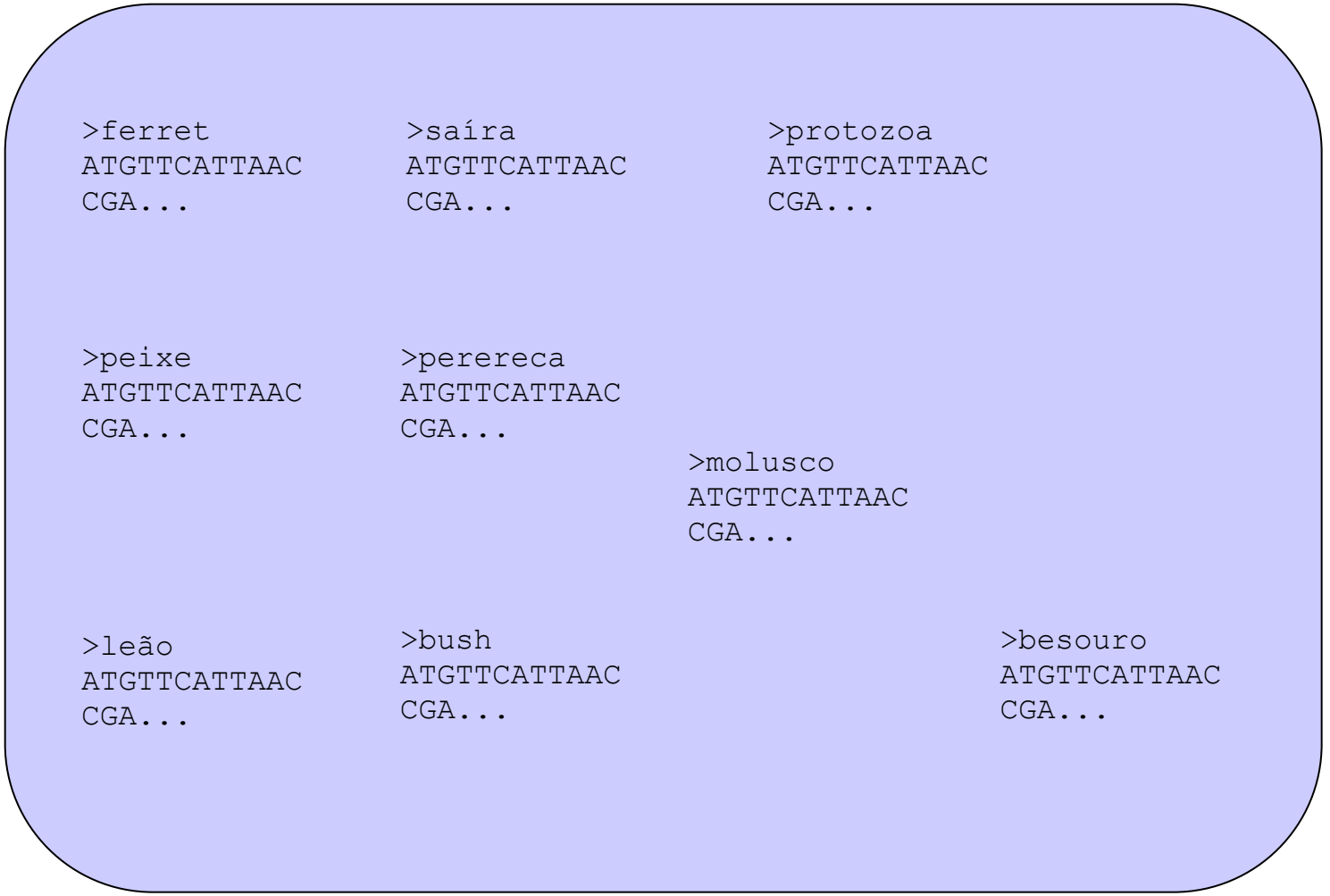
>ferret
ATGTTCATTAACCGATGACTGTTCTCCACTAATCACAAGGATATTGGTACTTTATACTTACTATTTGGAG
CATGAGCCGGTATAGTAGGCACTGCTTTGAGCCTCCTCATCCGAGCCGAACTAGGTCAGCCCGGTACTTT
ACTAGGTGACGATCAAATTTATAATGTCATCGTAACCGCCCATGCTTTCGTAATAATCTTCTTCATAGTC
ATGCCCATCATAATTGGGGGCTTTGGAAACTGACTAGTGCCGTTAATAATTGGTGCTCCGGACATGGCAT
TCCCCCGAATAAATAACATGAGCTTCTGACTCCTTCCTCCATCCTTTCTTCTACTATTAGCATCTTCTAT
GGTAGAAGCAGGTGCAGGAACGGGATGAACCGTATACCCCCCACTGGCTGGCAATCTGGCCCATGCAGGA
GCATCCGTTGACCTTACAATTTTCTCCTTACACTTAGCCGGAGTCTCTTCTATTTTAGGGGCAATTAATT
TCATCACTACTATTATCAACATAAAACCCCCTGCAATATCCCAGTATCAAACTCCCCTGTTTGTATGATC
AGTACTAATTACAGCAGTTCTACTCTTACTATCCCTGCCTGTACTGGCTGCTGGAATTACAATACTTTTA
ACAGACCGGAATCTTAATACAACATTTTTTGATCCCGCTGGAGGAGGAGACCCTATCCTATATCAACACC
TATTCTGATTCTTCGGACATCCTGAAGTTTACATTCTTATCCTGCCCGGATTCGGAATAATTTCTCACAT
TGTCACTTACTACTCAGGGAAAAAAGAGCCTTTCGGTTATATAGGAATAGTATGAGCAATAATATCTATT
GGGTTTTTAGGCTTTATCGTATGAGCTCACCATATGTTTACCGTAGGAATAGATGTAGACACACGAGCGT
ACTTTACGTCCGCCACTATAATTATCGCTATTCCAACGGGAGTAAAAGTATTTAGTTGACTGGCAACACT
TCATGGAGGCAATATTAAATGATCTCCAGCTATGCTATGAGCTTTAGGGTTTATTTTCTTATTTACAGTA
GGCGGGTTAACAGGTATTGTCCTAGCTAATTCGTCCTTAGACATCGTTCTTCATGATACATATTATGTTG
TGGCTCATTTTCACTATGTGCTTTCAATAGGAGCAGTTTTTGCCATTATGGGAGGATTTGCCCACTGATT
CCCTTTATTCTCAGGTTATACTCTTAACGATACTTGAGCAAAGATTCACTTTACAATTATGTTTGTGGGA
GTAAATATAACTTTCTTCCCTCAACATTTCCTAGGTTTATCTGGAATACCTCGTCGATACTCTGACTACC
CAGATGCATATACTACCTGAAATACCGTCTCCTCTATAGGATCGTTTATCTCGCTTACAGCGGTGATGCT
TATAATTTTTATGATCTGGGAAGCCTTTGCATCCAAACGAGAAGTTGCTATAGTAGAACTTACTACAACT
AACATTGAGTGACTACATGGATGTCCCCCTCCATACCACACGTTCGAAGAACCTACATATGTGATCCAAA
AATAA

O problema foi reformulado

Isso:



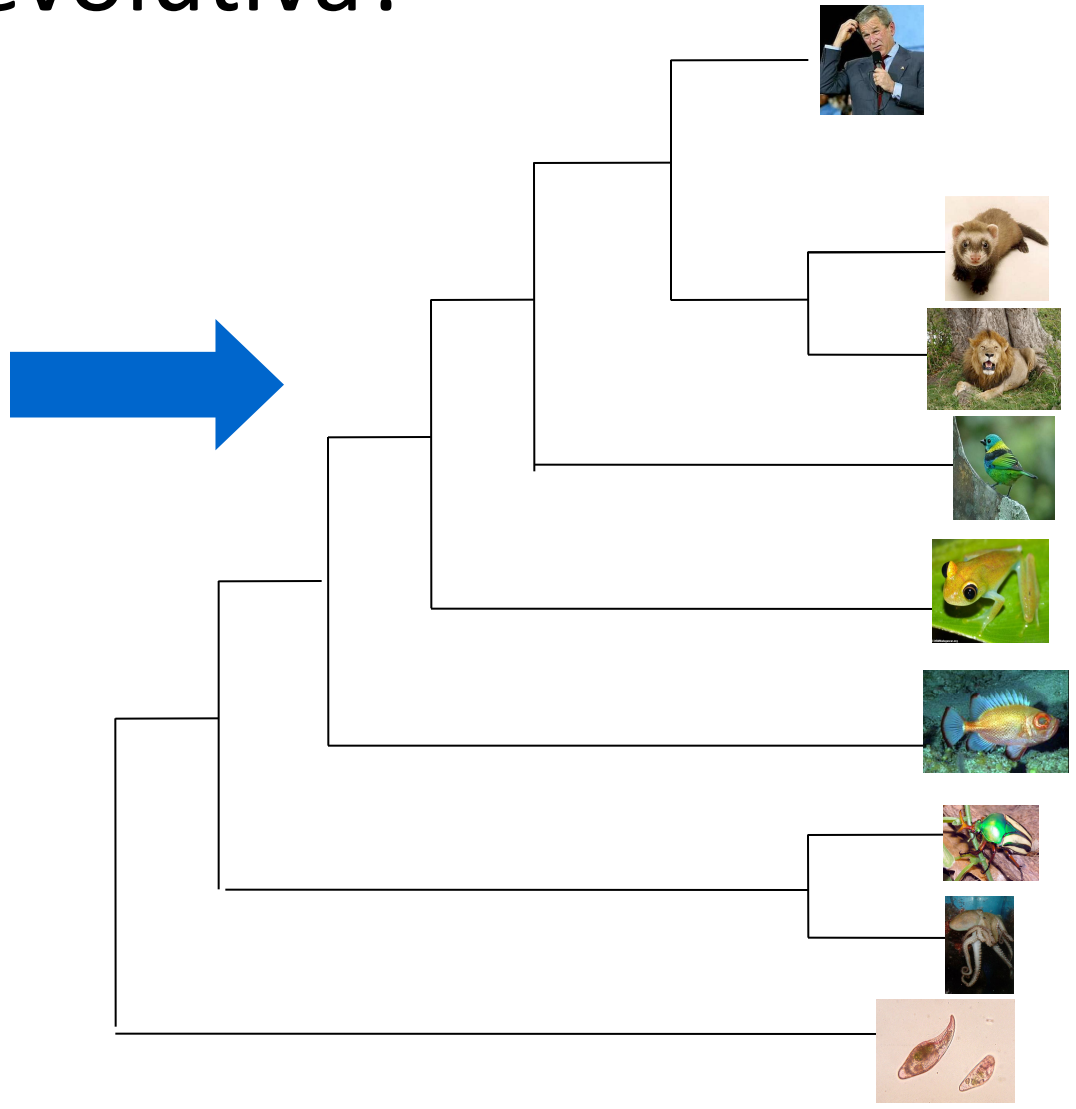
Virou isso:



COI

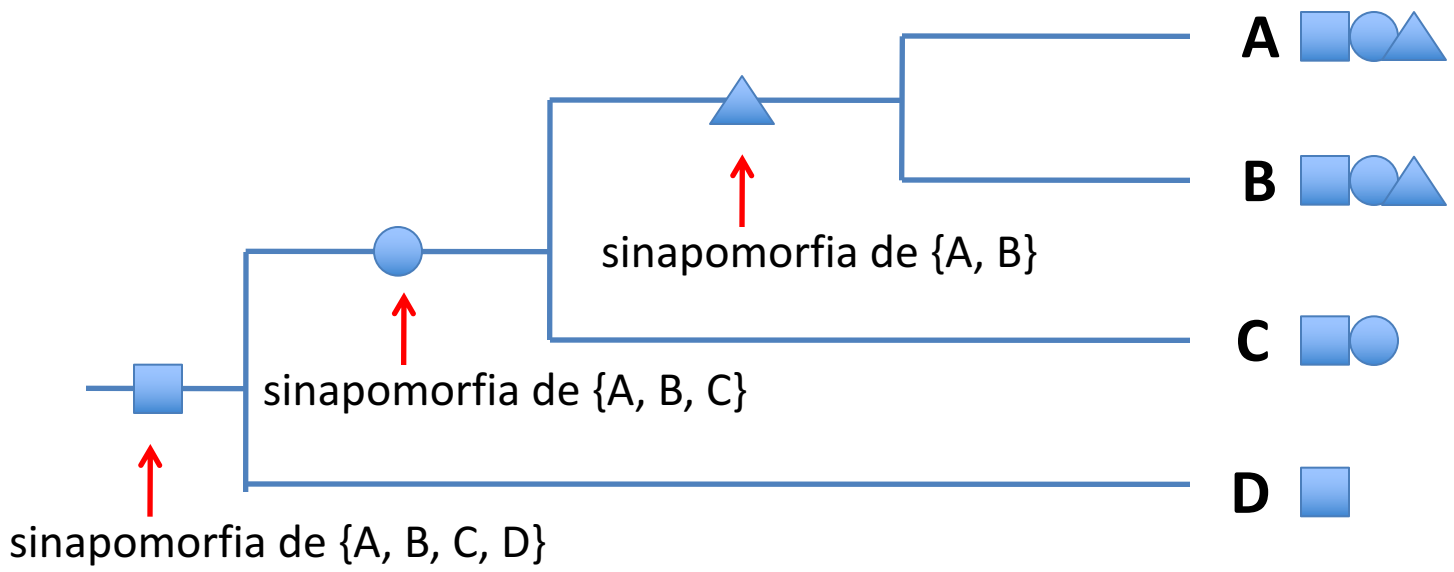
Como usar a informação das sequencias para fazer uma árvore evolutiva?

```
>ferret
ATGTTCATTAAC
CGA...
>saíra
ATGTTCATTAAC
CGA...
>peixe
ATGTTCATTAAC
CGA...
>perereca
ATGTTCATTAAC
CGA...
>protozoa
ATGTTCATTAAC
CGA...
>molusco
ATGTTCATTAAC
CGA...
>besouro
ATGTTCATTAAC
CGA...
>leão
ATGTTCATTAAC
CGA...
>bush
ATGTTCATTAAC
CGA...
```



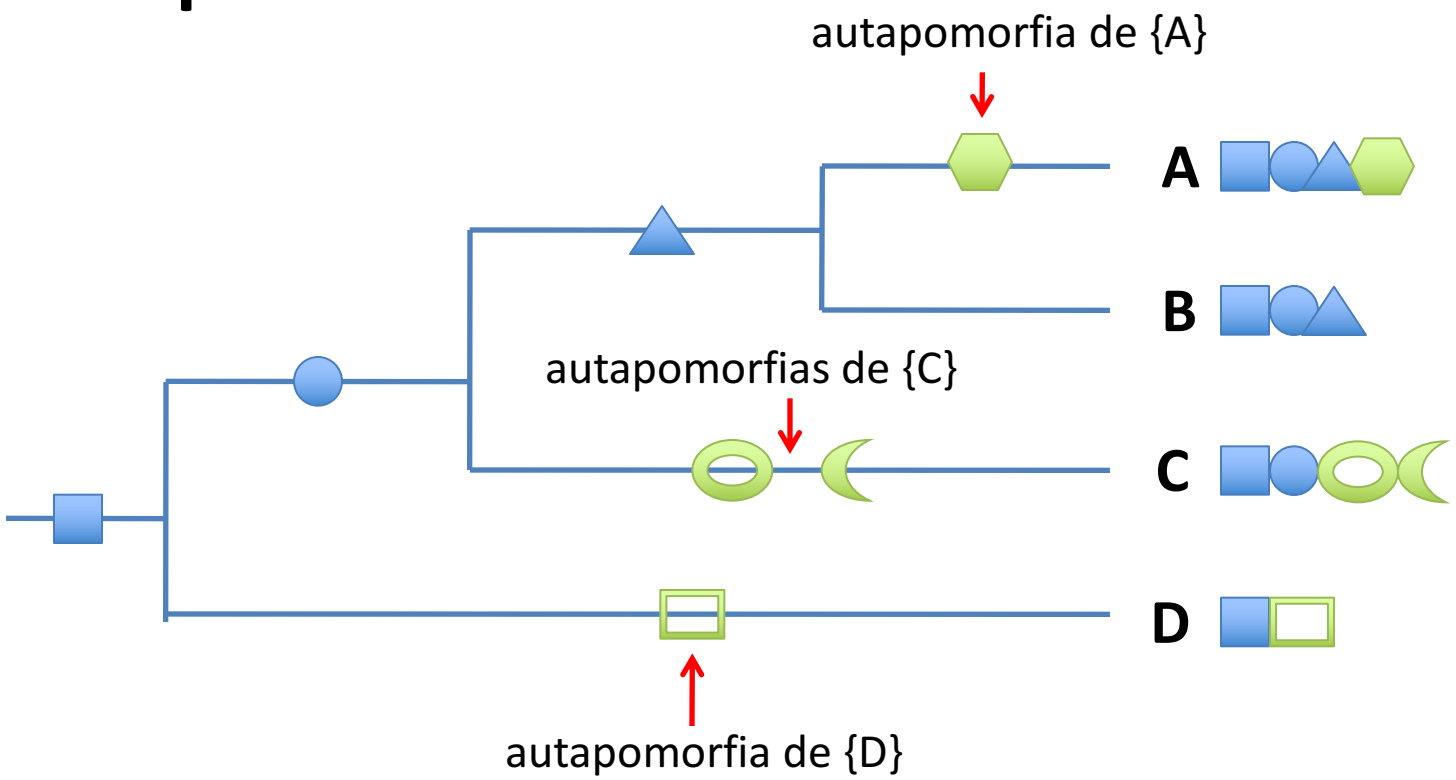
O problema filogenético

- A reconstrução filogenética objetiva recuperar as relações de ancestralidade entre as linhagens
- Isso é possível, pois linhagens irmãs compartilham características, chamadas de **sinapomorfias**



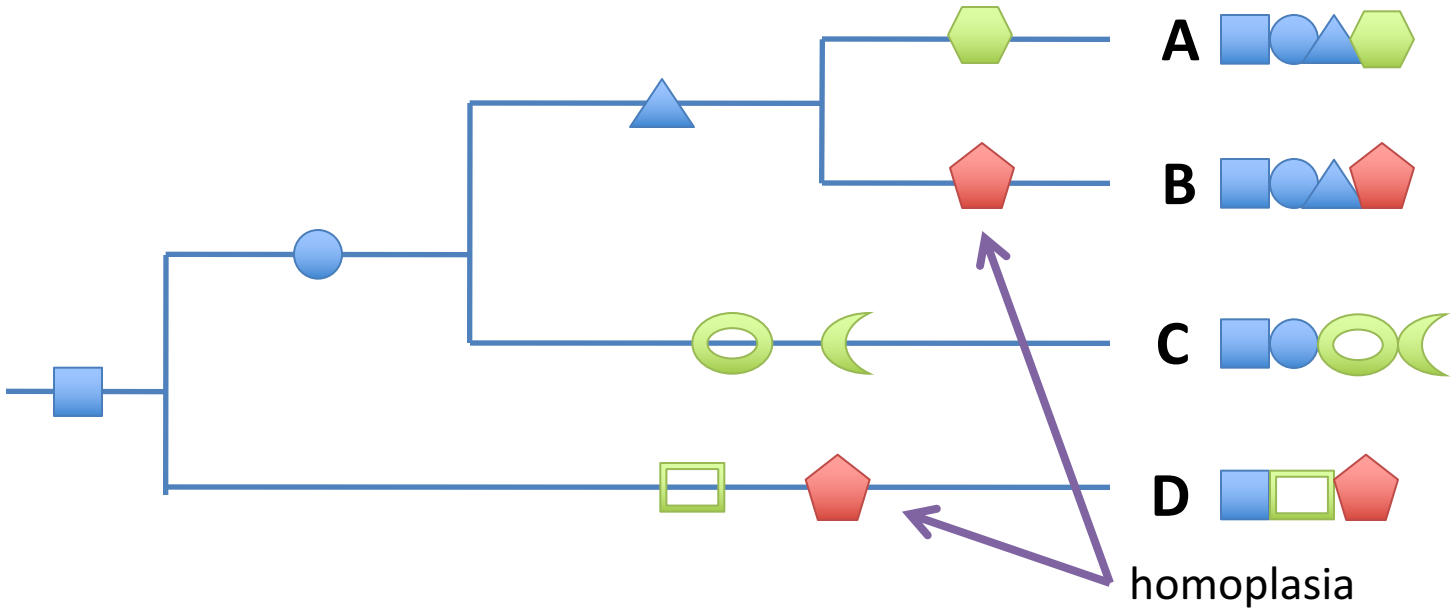
O Problema filogenético

- Algumas linhagens acumulam mudanças exclusivas, elas são chamadas de **autapomorfias**



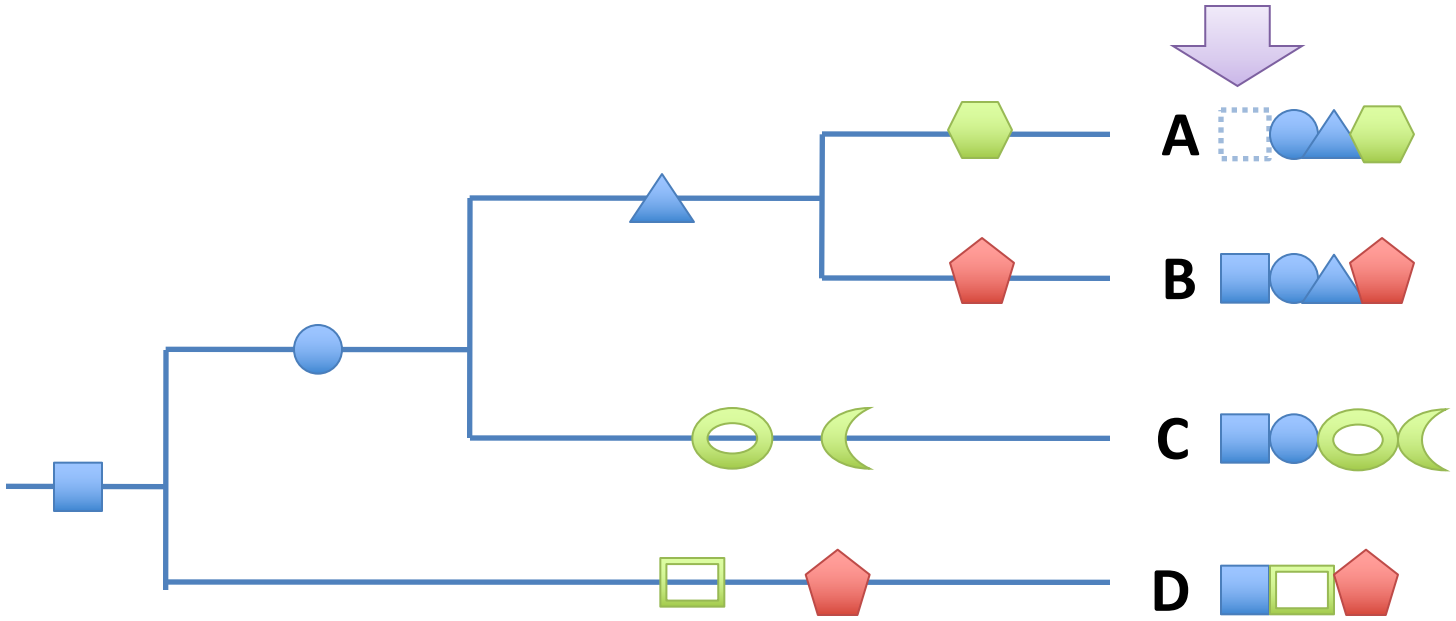
O Problema filogenético

- Às vezes, algumas características aparecem em linhagens independentes como resultado de evolução convergente. Elas são chamadas de **homoplasias**



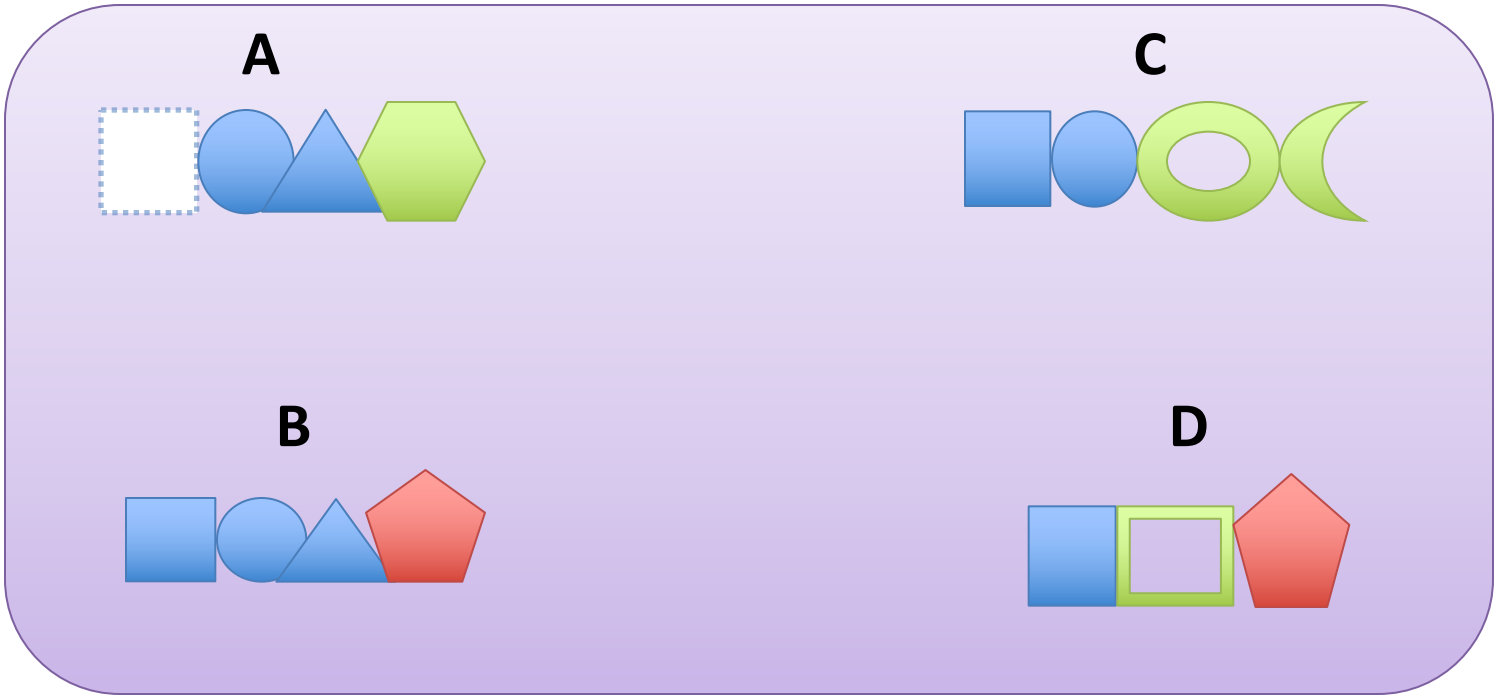
O Problema filogenético

- Eventualmente algumas características são perdidas ao longo do processo evolutivo



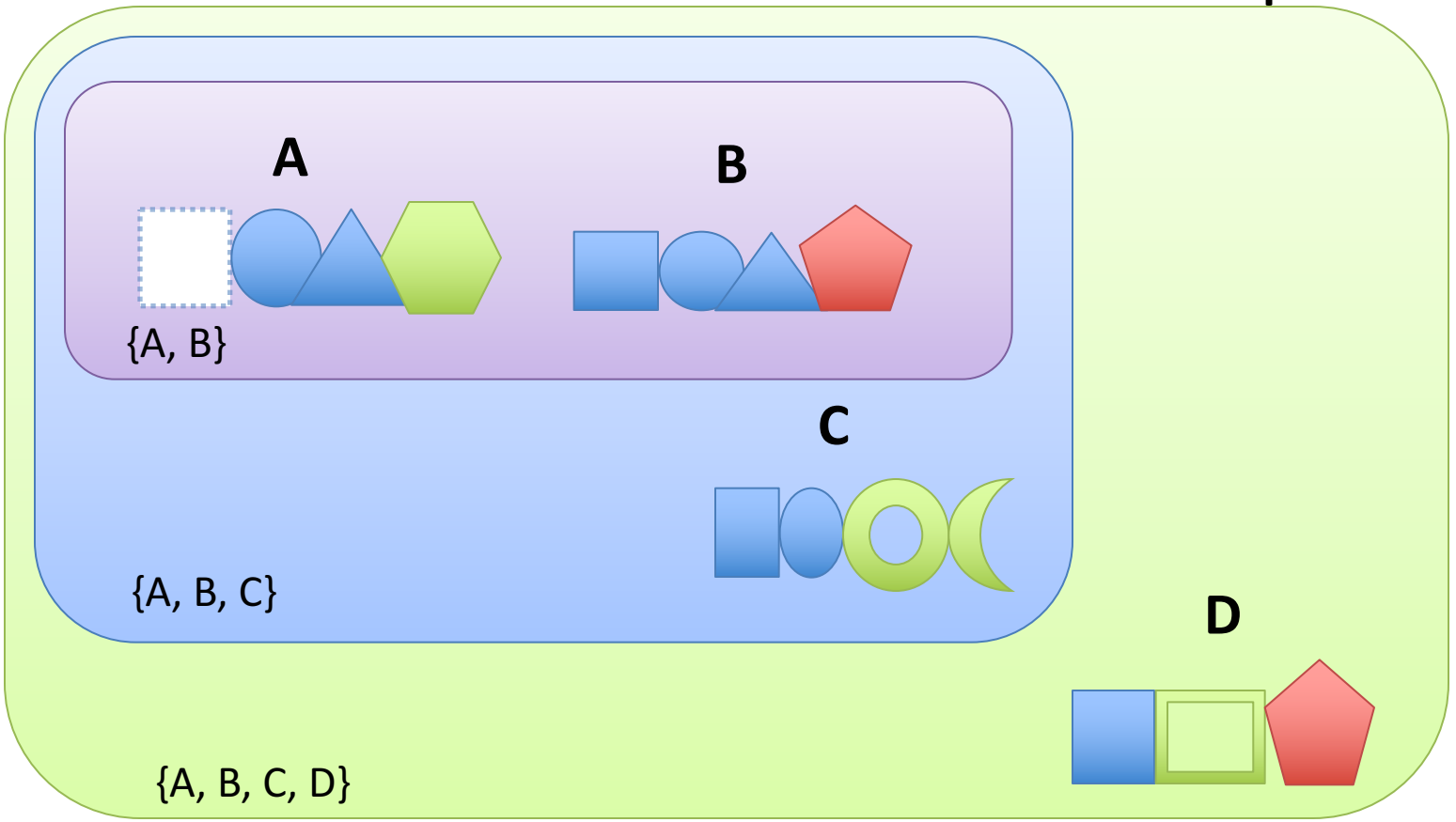
Organismos são mosaicos de características

- O processo evolutivo gera um mosaico de sinapomorfias, autapomorfias e homplasias.



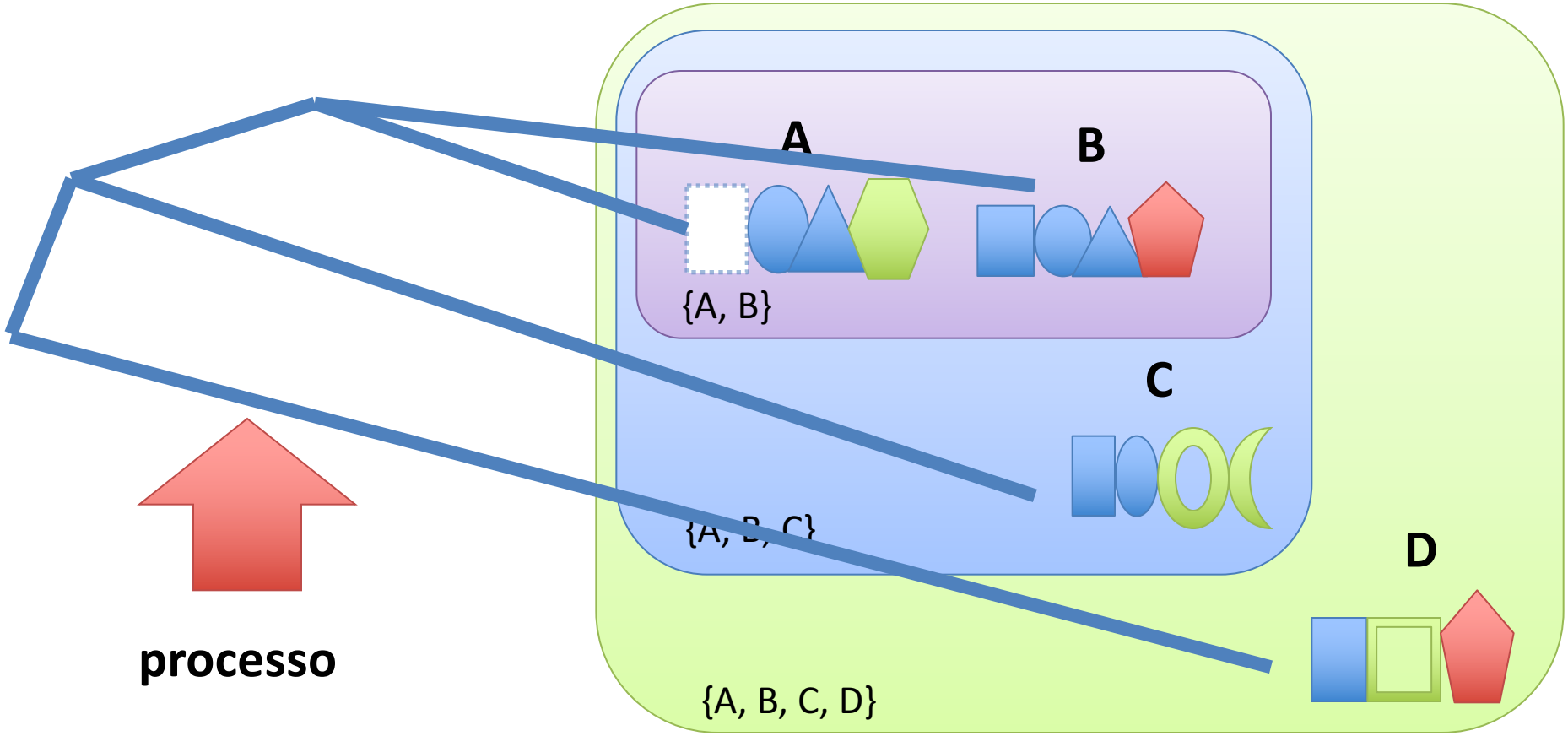
A diversidade é hierárquica

- A natureza do processo de geração dos caracteres torna a diversidade hierárquica

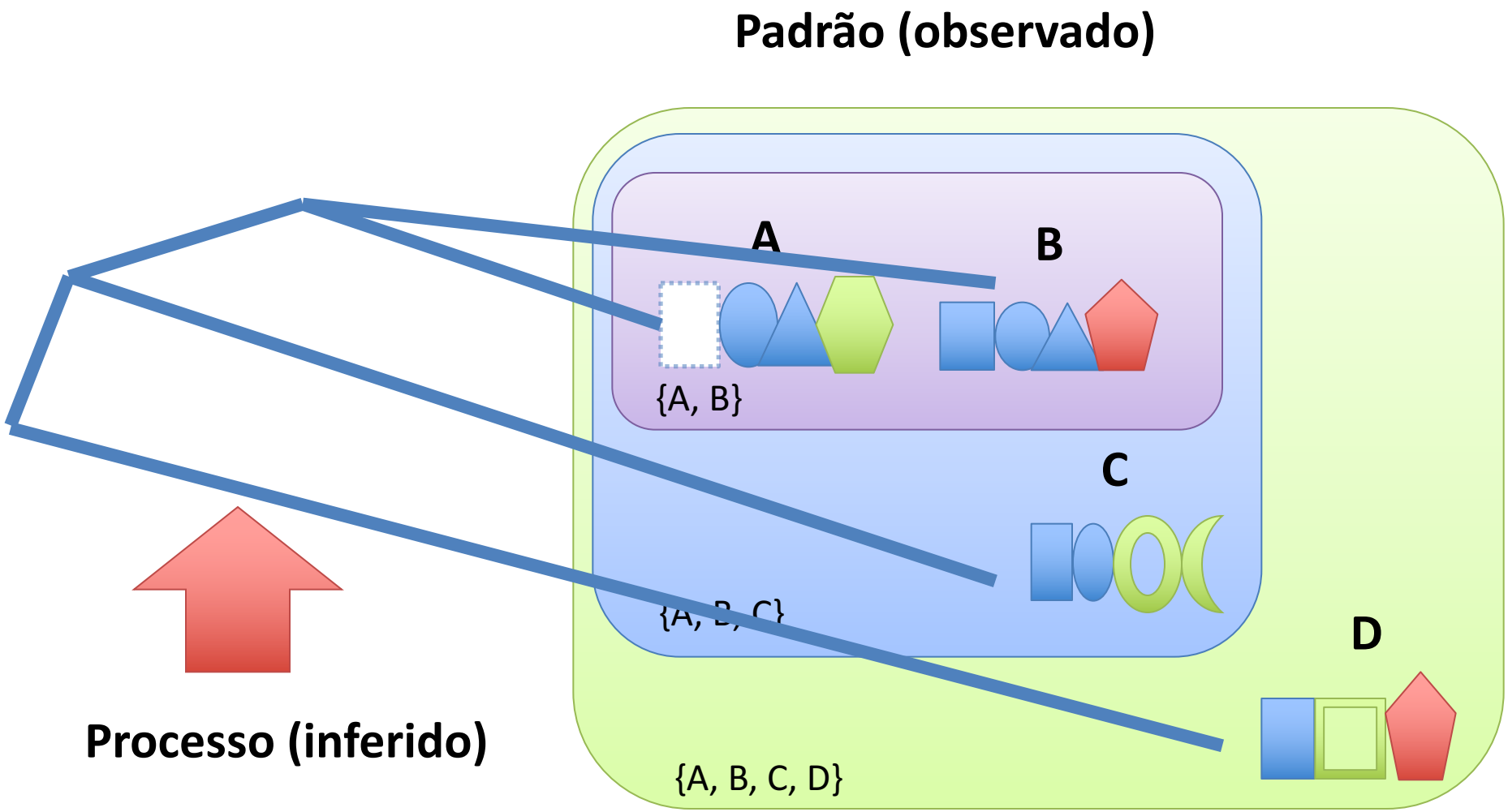


A informação da diversidade

- A hierarquia que observamos (**padrão**) nos oferece informação sobre o **processo** que a gerou

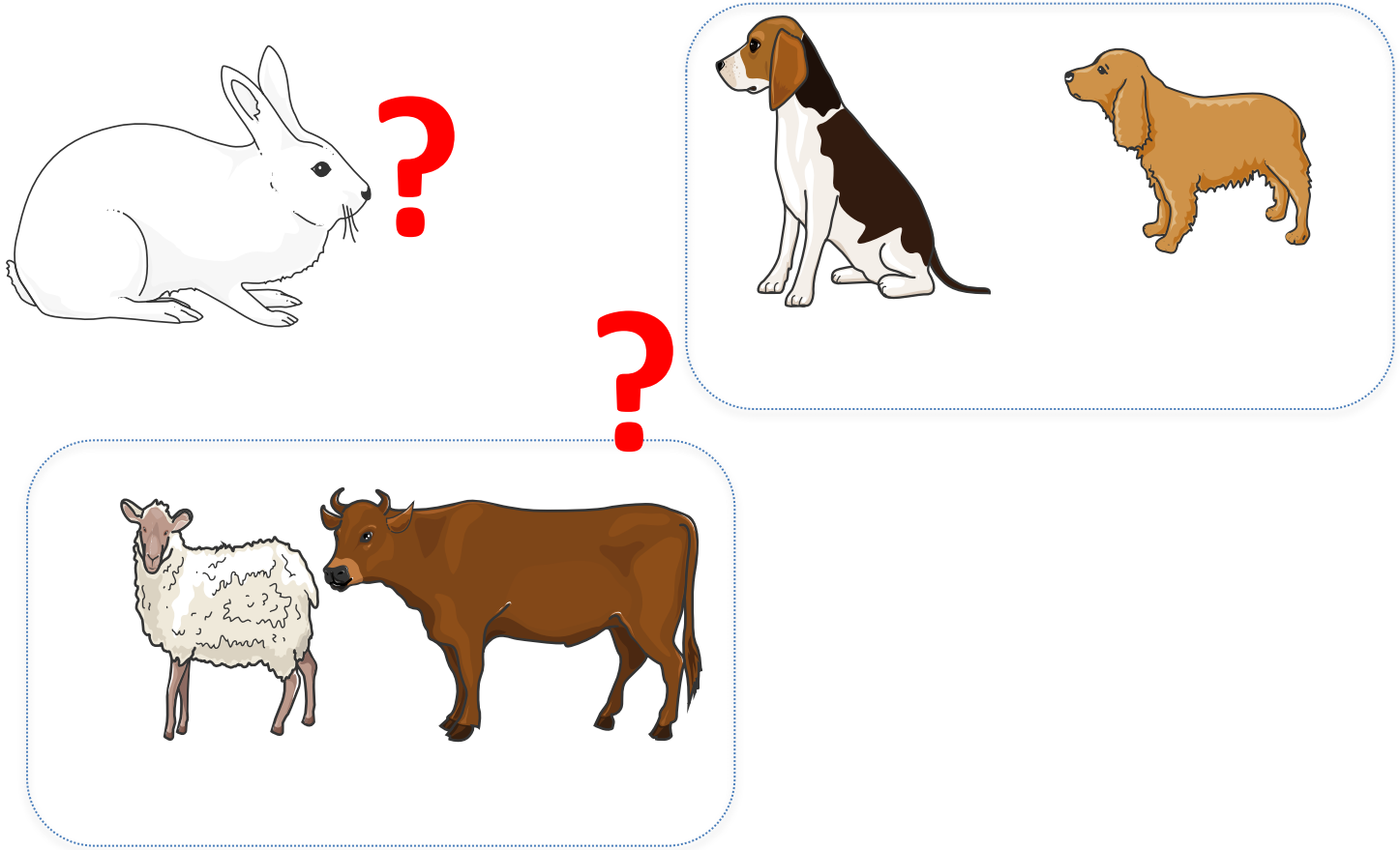


O Processo é inferido, o padrão observado



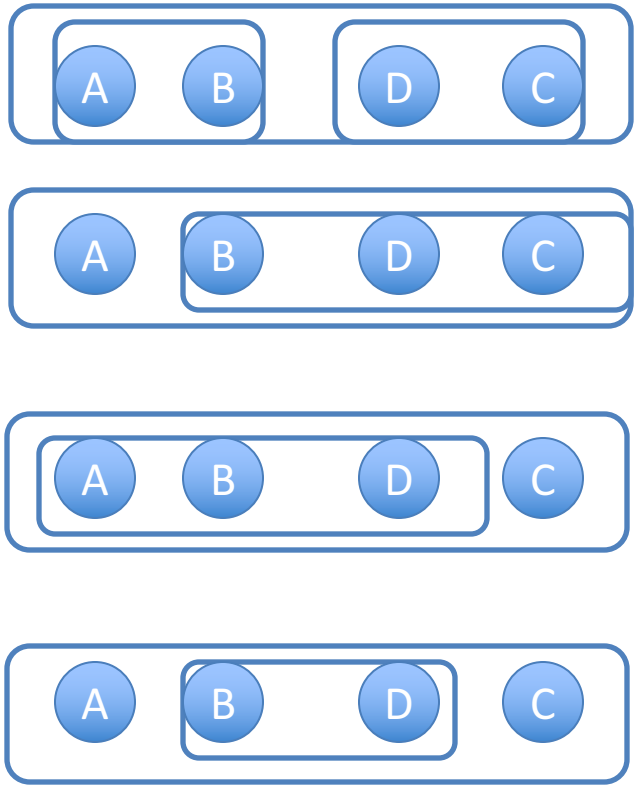
Mas as coisas não são tão simples...

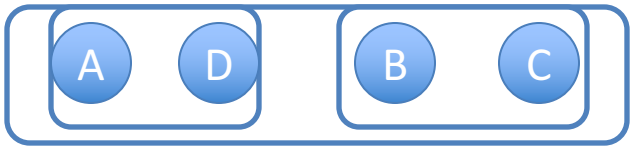
- Embora teoricamente simples, não é sempre fácil observar os padrões da diversidade



O Número de grupos possíveis é enorme

- Para quatro organismos, existem 15 grupos possíveis:

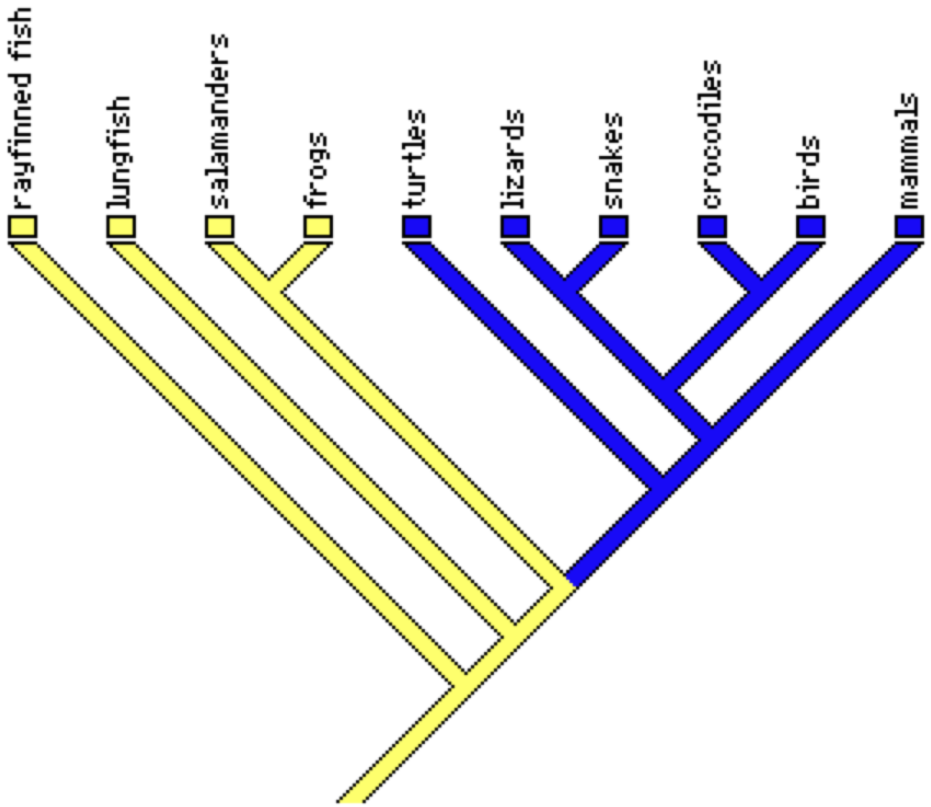




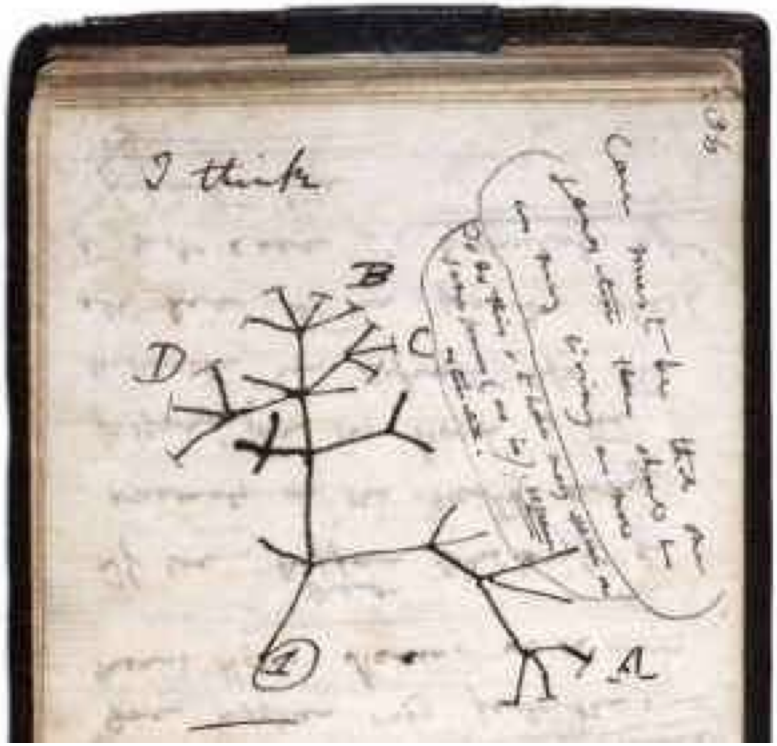
etc

A representação

- A forma tradicional de representar as relações evolutivas é através de árvores



A representação em árvores é antiga



Anotações de Charles Darwin

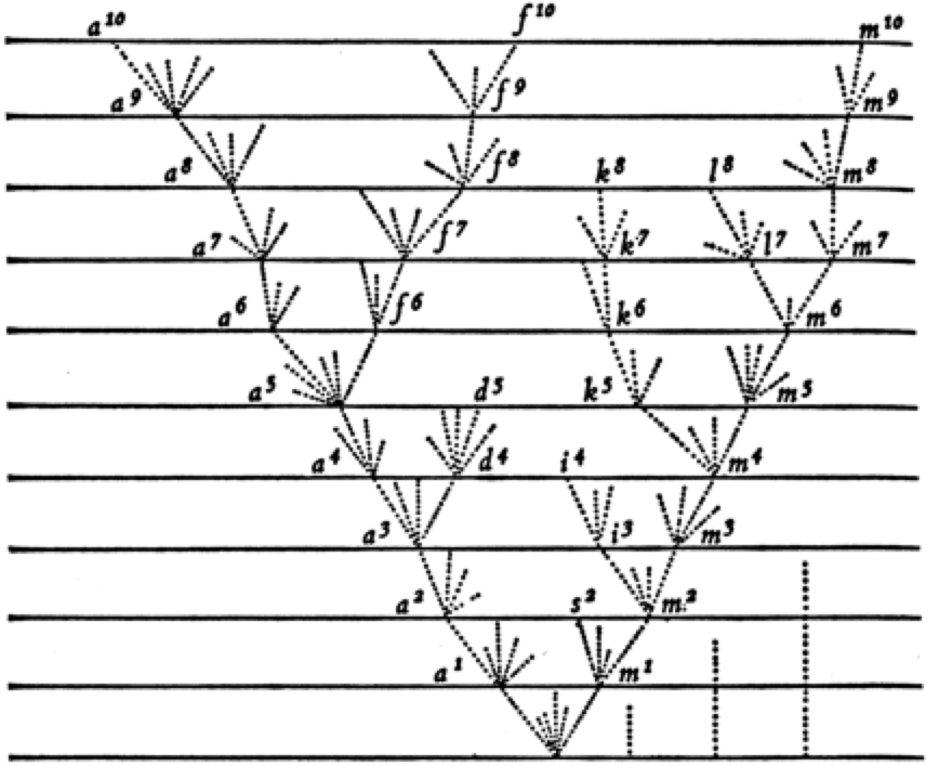
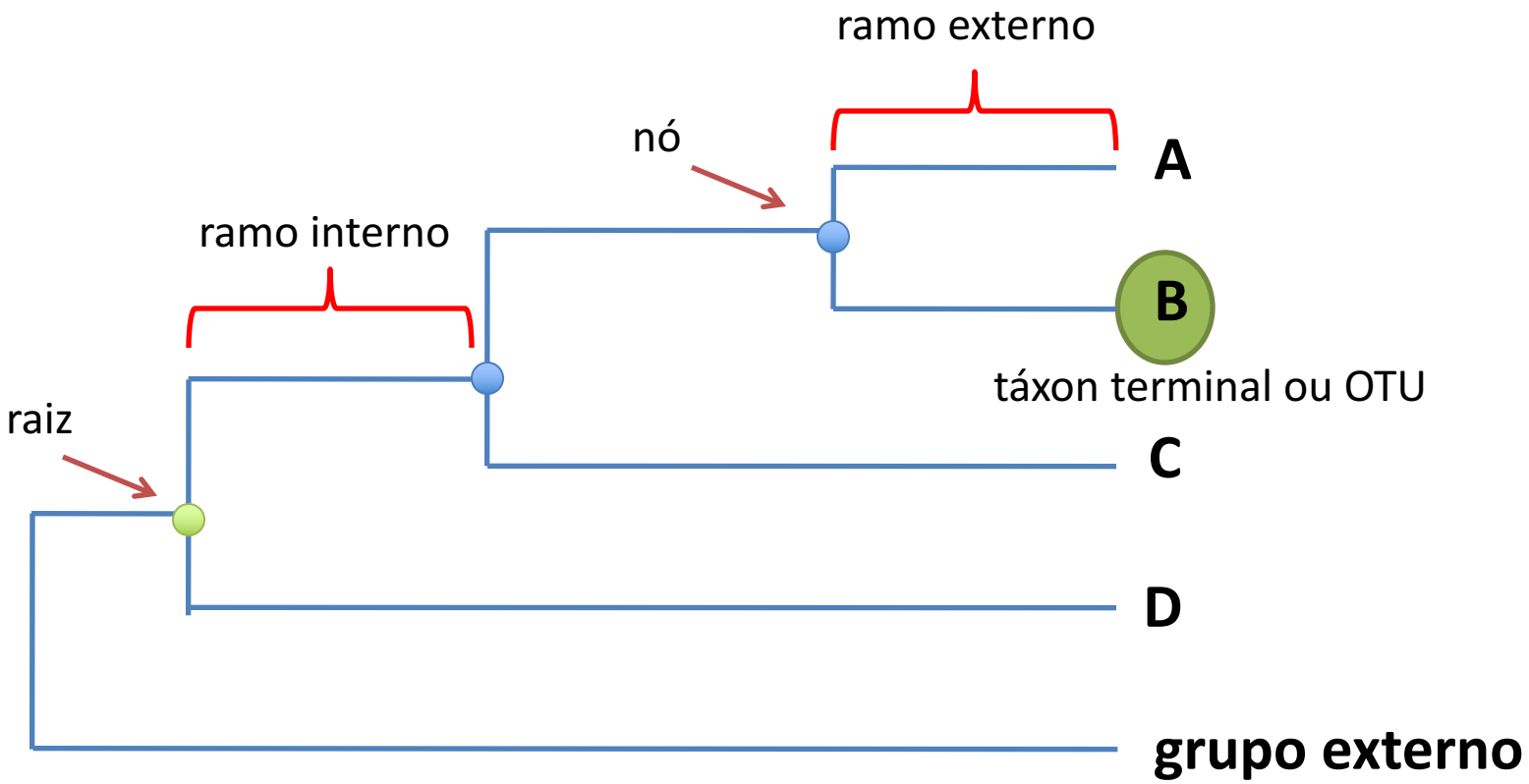
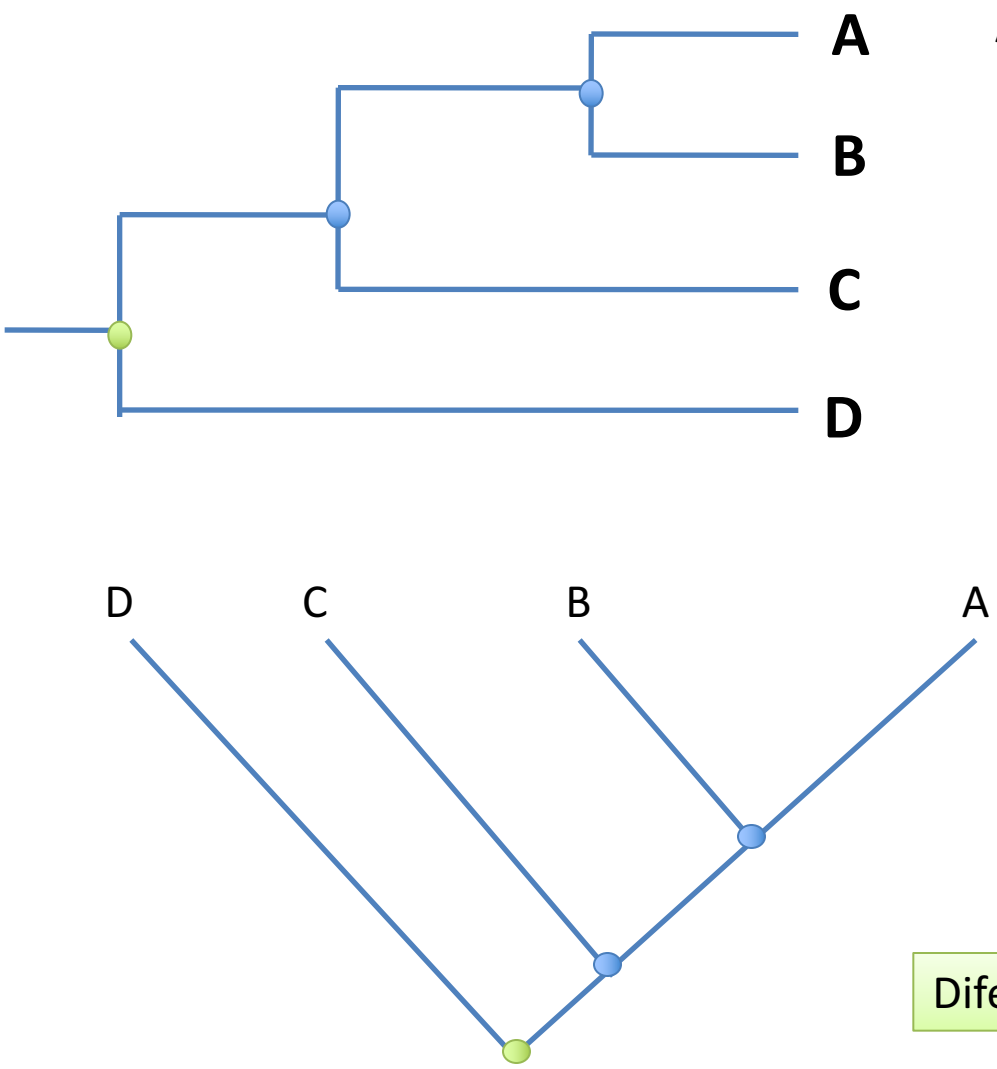


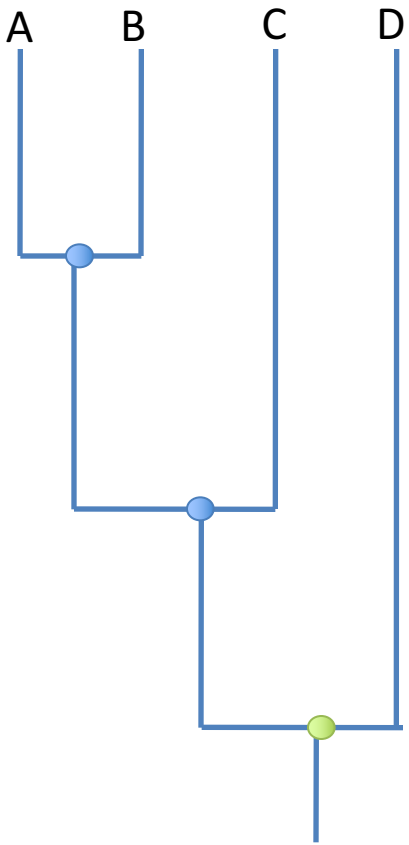
Figura de *A Origem das Espécies*

As árvores filogenéticas



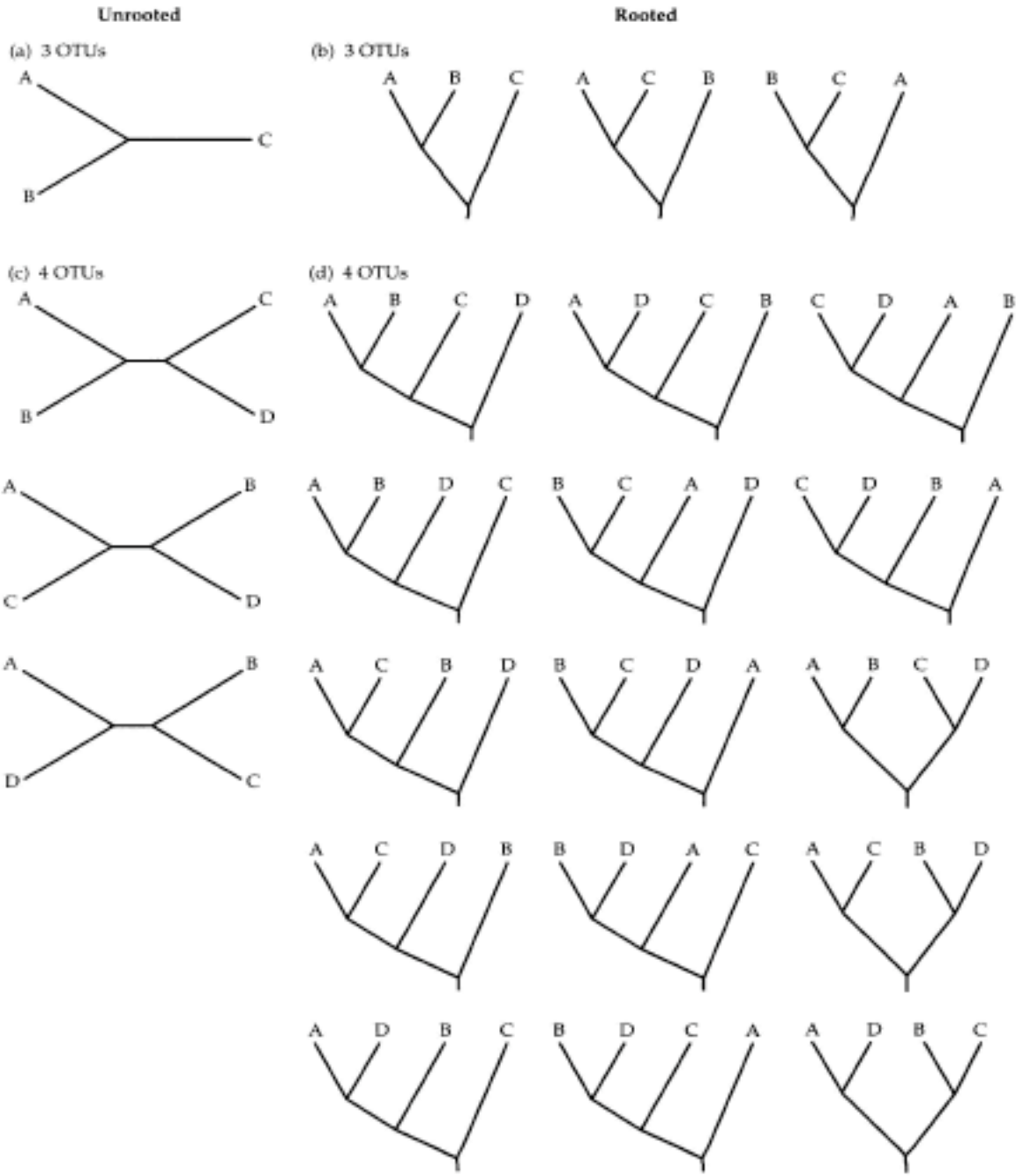
As formas não importam





Diferentes árvores com a mesma topologia

O Número de topologias possíveis é muito grande



O problema

- Dado regiões homólogas do genoma, como construir a árvore filogenética?

	Sítios									
	1	2	3	4	5	6	7	8	9	10
OTU 1	A	C	G	C	G	G	C	A	A	T
OTU 2	A	C	G	C	G	G	T	A	A	T
OTU 3	A	T	G	C	G	G	A	A	A	T
OTU 4	A	G	T	C	G	G	G	A	A	T
OTU 5	A	G	T	C	G	G	G	A	A	C

Esse problema pode ser resolvido de diversas formas

- Existem vários **algoritmos** de reconstrução filogenética:
- **Métodos de distância**
  . UPGMA (Unweighted Pair Group Method with Arithmetic Mean)
  . Neighbor joining
- **Métodos de parcimônia**
- **Métodos de máxima verossimilhança**

Os métodos de distância

- Para construir árvores filogenéticas por métodos de distância devemos primeiro obter a matriz de distância das seqüências analisadas

Uma matriz de distância

	OTU 1	OTU 2	OTU 3	OTU 4	OTU 5
OTU 1					
OTU 2	d_{12}				
OTU 3	d_{13}	d_{23}			
OTU 4	d_{14}	d_{24}	d_{34}		
OTU 5	d_{15}	d_{25}	d_{35}	d_{45}	

Portanto...

- Voltando aos dados.

	Sítios									
	1	2	3	4	5	6	7	8	9	10
OTU 1	A	C	G	C	G	G	C	A	A	T
OTU 2	A	C	G	C	G	G	T	A	A	T
OTU 3	A	T	G	C	G	G	A	A	A	T
OTU 4	A	G	T	C	G	G	G	A	A	T
OTU 5	A	G	T	C	G	G	G	A	A	C

$$d_p = \frac{n}{N} \rightarrow d_{12} = \frac{1}{10} = 0{,}1$$

A matriz de distância *p* dos dados

	OTU 1	OTU 2	OTU 3	OTU 4	OTU 5
OTU 1					
OTU 2	0,1				
OTU 3	0,2	0,2			
OTU 4	0,3	0,3	0,3		
OTU 5	0,4	0,4	0,4	0,1	

E agora?

- Intuitivamente, seqüências com menor distância são geneticamente mais próximas...

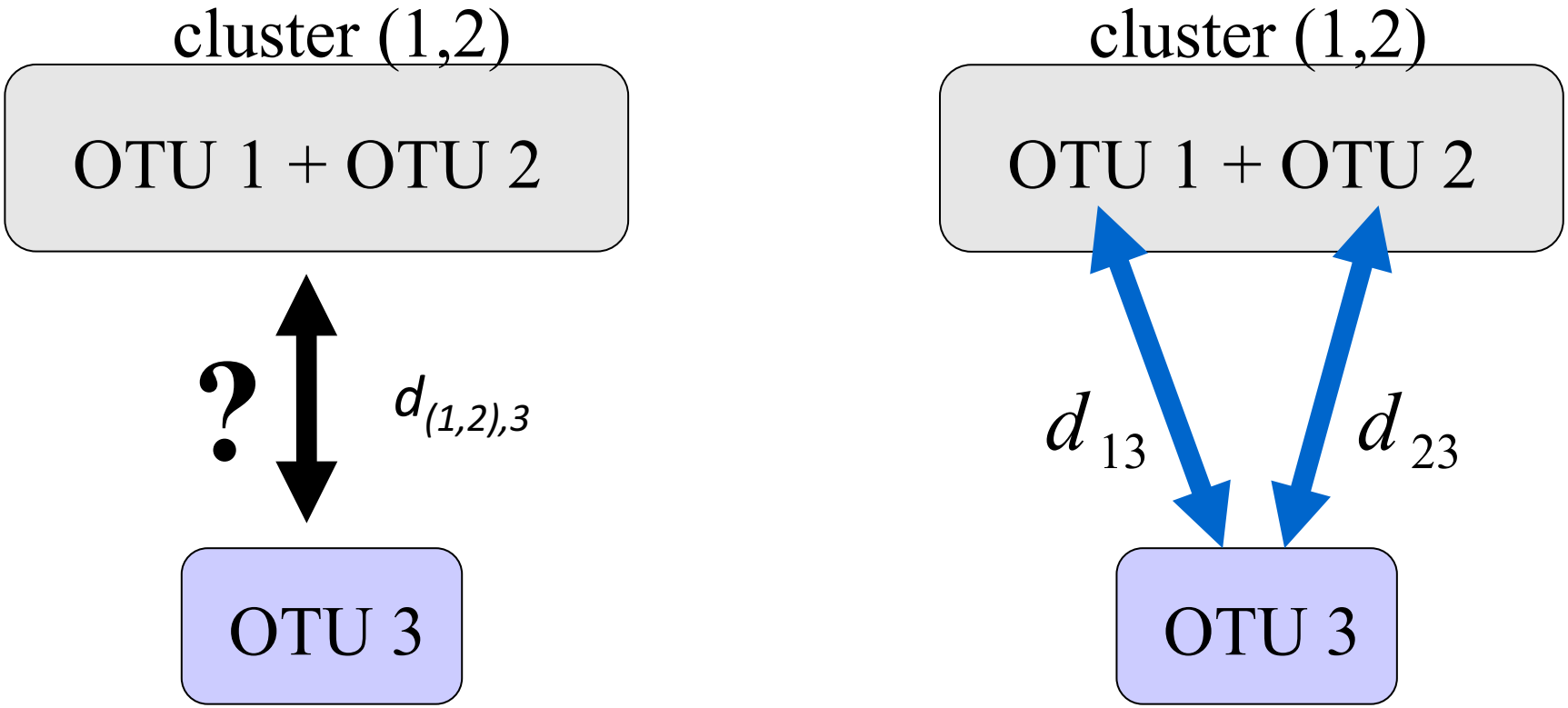
	OTU 1	OTU 2	OTU 3	OTU 4	OTU 5
OTU 1					
OTU 2	0,1				
OTU 3	0,2	0,2			
OTU 4	0,3	0,3	0,3		
OTU 5	0,4	0,4	0,4	0,1	

O algoritmo UPGMA

- A idéia do UPGMA é agrupar OTUs progressivamente de acordo com a distância genética

- OTUs agrupadas passam a compor um *cluster*

- A distância entre uma OTU e o *cluster* é a média entre as distâncias individuais



$$d_{(1,2)3} = \frac{d_{13} + d_{23}}{2}$$

UPGMA em ação

	OTU 1	OTU 2	OTU 3	OTU 4	OTU 5
OTU 1					
OTU 2	0,1				
OTU 3	0,2	0,2			
OTU 4	0,3	0,3	0,3		
OTU 5	0,4	0,4	0,4	0,1	

$$d_{(1,2)3} = \frac{d_{13} + d_{23}}{2}$$

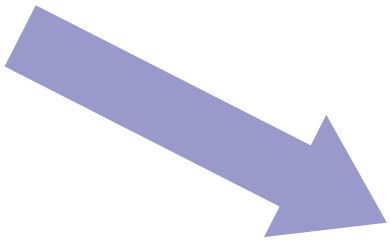
	cluster(1,2)	OTU 3	cluster(4,5)
cluster(1,2)			
OTU 3	0,2		
cluster(4,5)	0,35	0,35	

$$d_{(1,2)(4,5)} = \frac{d_{14} + d_{15} + d_{24} + d_{25}}{4}$$

$$d_{(4,5)3} = \frac{d_{43} + d_{53}}{2}$$

UPGMA ainda em ação...

	cluster(1,2)	OTU 3	cluster(4,5)
cluster(1,2)			
OTU 3	0,2		
cluster(4,5)	0,35	0,35	



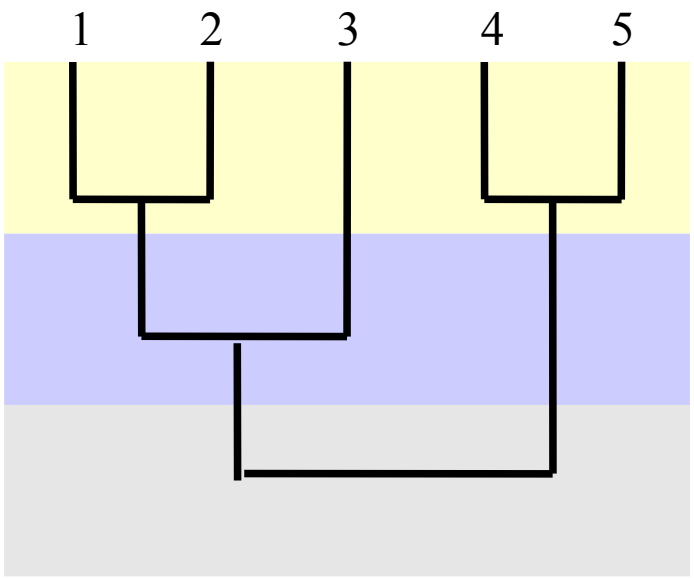
	cluster(1,2,3)	cluster(4,5)
cluster(1,2,3)		
cluster(4,5)	0,35	

$$d_{(1,2,3)(4,5)} = \frac{d_{14} + d_{15} + d_{24} + d_{25} + d_{34} + d_{35}}{6}$$

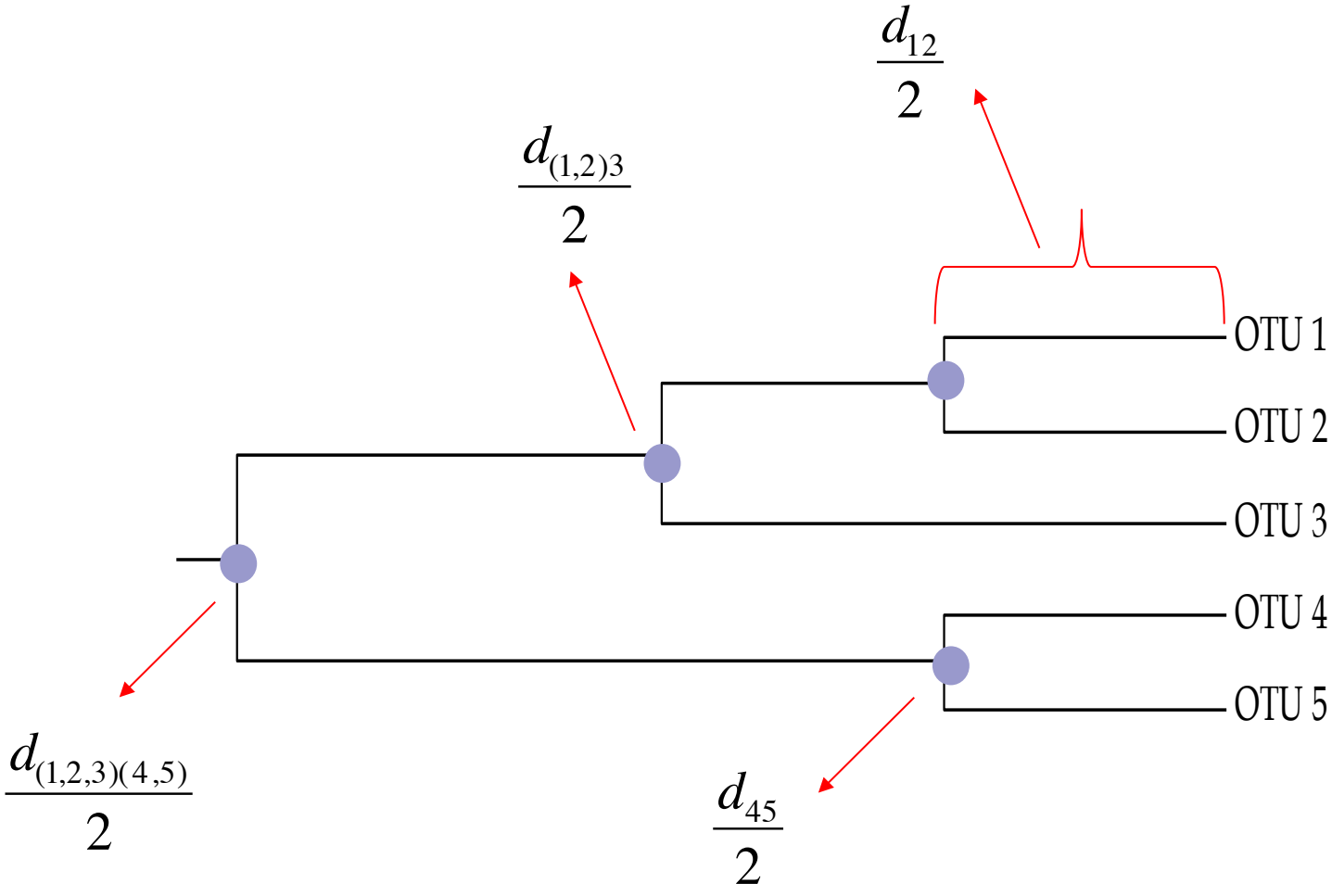
A árvore de UPGMA

- A ordem de agrupamento foi

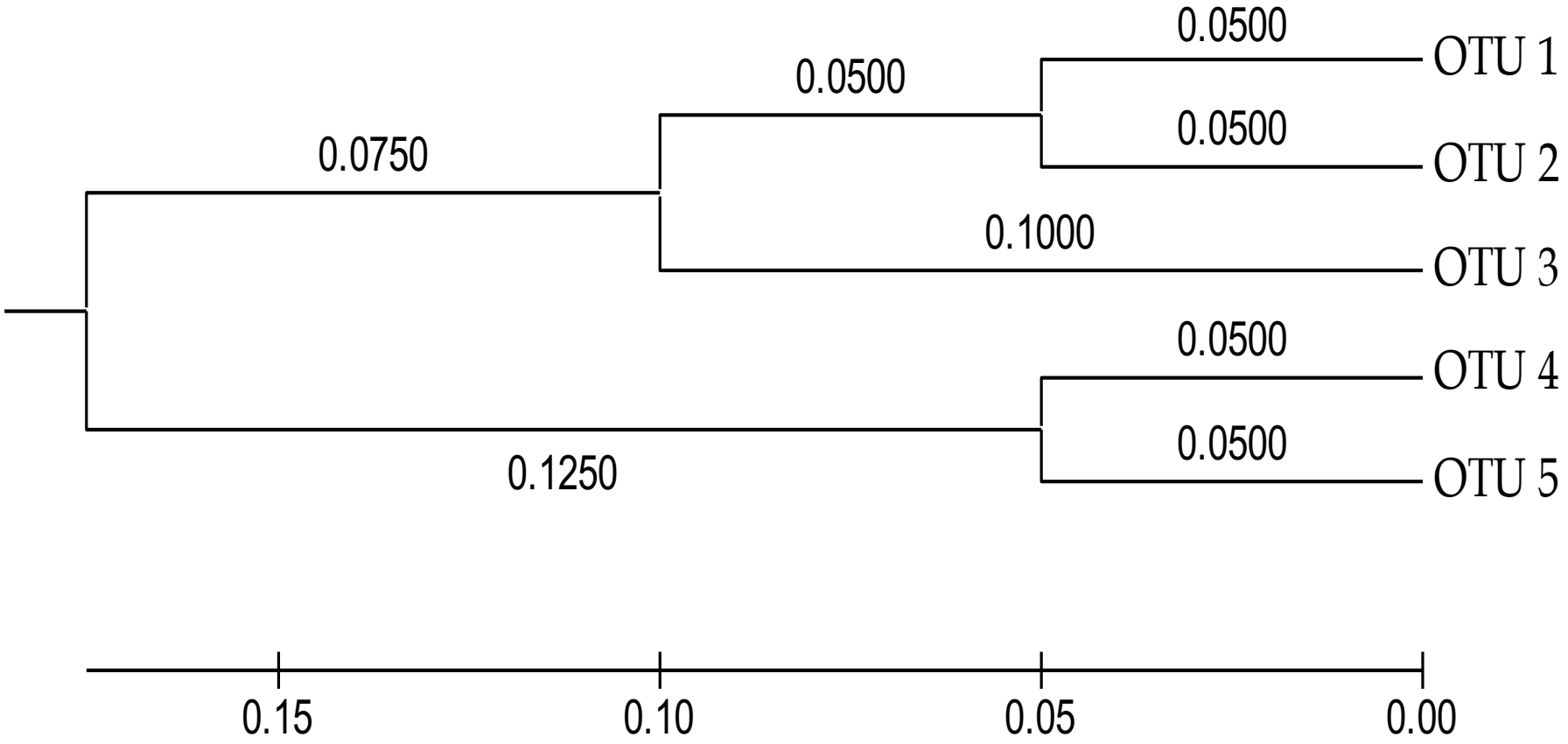
Passo	Grupos
1	(1,2) & (4,5)
2	(1,2,3) & (4,5)
3	(1,2,3,4,5)



Pontos de quebra



Finalmente


0.0500
OTU 1
0.0500
0.0500
OTU 2
0.0750
0.1000
OTU 3
0.0500
OTU 4
0.0500
OTU 5
0.1250
0.15
0.10
0.05
0.00