1º Simpósio LAVIREO

# *Análise Filogenética: uma introdução*

**Edson Delatorre**

Lab. de Genética Molecular de Microrganismos

Instituto Oswaldo Cruz/FIOCRUZ

delatorre.ioc@gmail.com

Rio de Janeiro – RJ, janeiro de 2019

# Filogenética

Estudo da evolução e relacionamentos de uma coleção de “coisas” (genes, proteínas, organismos) que derivam de um ancestral comum.

# Dados utilizados para análises filogenéticas

- Caracteres morfológicos
- Dados genéticos
  - RFLP
  - Frequências alélicas
  - Sequências de AA ou NT
- Uma combinação destes

# Dados utilizados para análises filogenéticas

- Caracteres morfológicos
- Dados genéticos
  - RFLP
  - Frequências alélicas
  - **Sequências de AA ou NT**
- Uma combinação destes

# Homologia x Analogia x Similaridade

- **Homologia**:
  - Termo absoluto, significa que os OTU **compartilham um ancestral comum** recente o suficiente para que a variação observada entre as sequências carregue informação suficiente para as análises. filogenéticas

- **Similaridade**:
  - Termo relativo, comparação entre duas sequências expressa comumente como uma porcentagem.

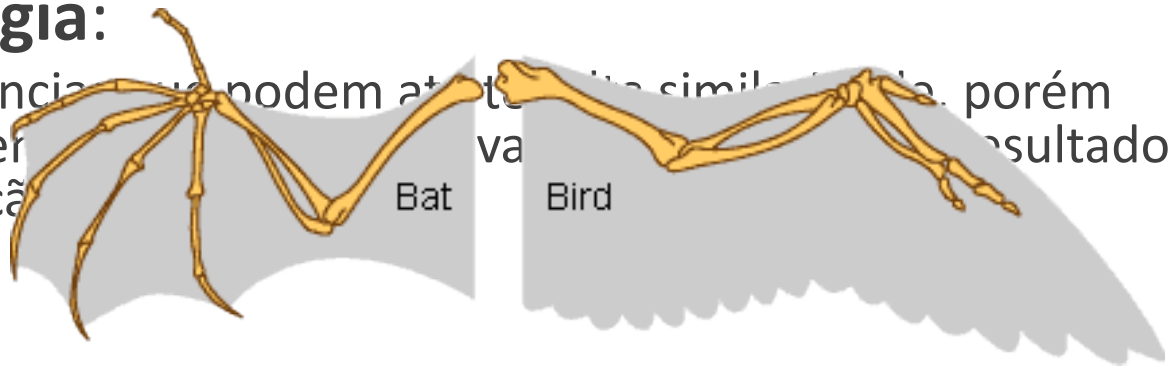
- **Analogia**:
  - Sequências que podem até ter alta similaridade, porém apresentam origens evolutivas separadas. São o resultado de evolução convergente.

# Homologia x Analogia x Similaridade

- **Homologia**:
  - Termo absoluto, significa que os OTU **compartilham um ancestral comum** recente o suficiente para que a variação observada entre as sequências carregue informação suficiente para as análises. filogenéticas

- **Similaridade**:
  - Termo relativo, comparação entre duas sequências expressa comumente como uma porcentagem.

- **Analogia**:
  - Sequência[illegible] podem at[illegible] porém apresen[illegible] va[illegible] sultado de evoluçã[illegible]



# Análises filogenéticas

- Somente sequências **homólogas** podem ser utilizadas para análises filogenéticas.

- **Recombinação**:
  - Perturba a comparação, uma vez que cada parte do genoma recombinado possui uma história evolutiva distinta.

- **Alinhamento de sequências**
  - Mutações pontuais
  - Indels: adição de “gaps” para conseguir a homologia posicional

# Alinhamento

```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                 :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                          :   :  .:      .      ..       .   :
```

# NT ou AA?

- Para genes distantemente relacionados alinhamento de NTs podem se tornar ambíguos.

- Para genes proximamente relacionados, alinhamentos de AAs podem não conter uma quantidade suficiente de substituições para construção de uma árvore confiável.

# Homologia posicional

```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                 :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                          :   :  .:      .      ..       .   :
```

# Estratégias de alinhamento

- Algoritmos de alinhamento auxiliam quando se tem uma grande quantidade de sequências e quando estas são distantemente relacionadas.

- Existem diversos algoritmos de alinhamento
  - Diferem na penalidade de gaps, pontuação e outros detalhes

- Qual algoritmo escolher?

# Estratégias de alinhamento

**Table 3.2** Typical alignment tasks and recommended procedures

| Input data | Recommendations |
|---|---|
| 2–100 sequences of typical protein length (maximum around 10 000 residues) that are approximately globally alignable | Use PROBCONS, T-COFFEE, and MAFFT or MUSCLE, compare the results using ALTAVISTA. Regions of agreement are more likely to be correct. For sequences with low percent identity, PROBCONS is generally the most accurate, but incorporating structure information (where available) via 3DCOFFEE (a variant of T-COFFEE) can be extremely helpful |
| 100–500 sequences that are approximately globally alignable | Use MUSCLE or one of the MAFFT scripts with default options. Comparison using ALTAVISTA is possible, but the results are hard to interpret with larger numbers of sequences unless they are highly similar |
| >500 sequences that are approximately globally alignable | Use MUSCLE with a faster option (we recommend maxiters-2) or one of the faster MAFFT scripts |
| Large numbers of alignments, high-throughput pipeline | Use MUSCLE with faster options (e.g. maxiters-1 or maxiters-2) or one of the faster MAFFT scripts |
| 2–100 sequences with conserved core regions surrounded by variable regions that are not alignable | Use DIALIGN |
| 2–100 sequences with one or more common domains that may be shuffled, repeated or absent | Use PRODA |
| A small number of unusually long sequences (say, >20 000 residues) | Use CLUSTALW. Other programs may run out of memory, causing an abort (e.g. a segmentation fault) |

Current Opinion in Structural Biology 2006, 16:368–373.

# Estratégias de alinhamento

**Table 3.2** Typical alignment tasks and recommended procedures

| Input data | Recommendations |
|---|---|
| 2–100 sequences of typical protein length (maximum around 10 000 residues) that are approximately globally alignable | Use PROBCONS, T-COFFEE, and MAFFT or MUSCLE, compare the results using ALTAVISTA. Regions of agreement are more likely to be correct. For sequences with low percent identity, PROBCONS is generally the most accurate, but incorporating structure information (where available) via 3DCOFFEE (a variant of T-COFFEE) can be extremely helpful |
| 100–500 sequences that are approximately globally alignable | Use MUSCLE or one of the MAFFT scripts with default options. Comparison using ALTAVISTA is possible, but the results are hard to interpret with larger numbers of sequences unless they are highly similar |
| >500 sequences that are approximately globally alignable | Use MUSCLE with a faster option (we recommend maxiters-2) or one of the faster MAFFT scripts |
| Large numbers of alignments, high-throughput pipeline | Use MUSCLE with faster options (e.g. maxiters-1 or maxiters-2) or one of the faster MAFFT scripts |
| 2–100 sequences with conserved core regions surrounded by variable regions that are not alignable | Use DIALIGN |
| 2–100 sequences with one or more common domains that may be shuffled, repeated or absent | Use PRODA |
| A small number of unusually long sequences (say, >20 000 residues) | Use CLUSTALW. Other programs may run out of memory, causing an abort (e.g. a segmentation fault) |

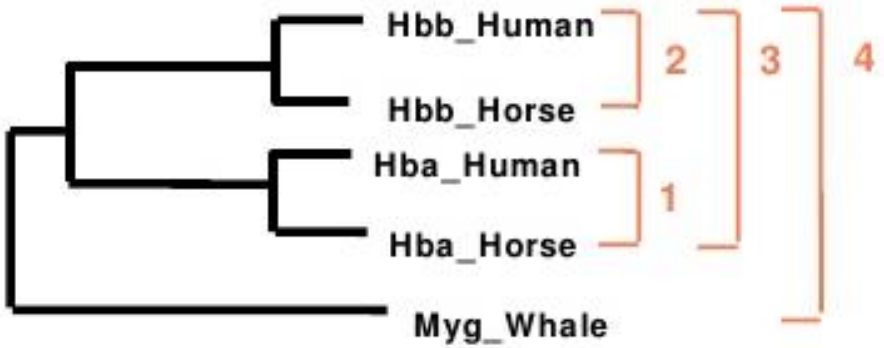
Current Opinion in Structural Biology 2006, 16:368–373.

# CLUSTAL W

Nucl. Acids Res. 1994, 22:4673-4680

| | | | | | |
|---|---|---|---|---|---|
| Hbb_Human | 1 | - | | | |
| Hbb_Horse | 2 | .17 | - | | |
| Hba_Human | 3 | .59 | .60 | - | |
| Hba_Horse | 4 | .59 | .59 | .13 | - |
| Myg_Whale | 5 | .77 | .77 | .75 | .75 | - |

Alinhamento de todos os pares.
Cálculo de uma matrix de distâncias.




A partir da matrix de distâncias
se constrói uma árvore filogenética
que servirá como guia.

alpha-helices

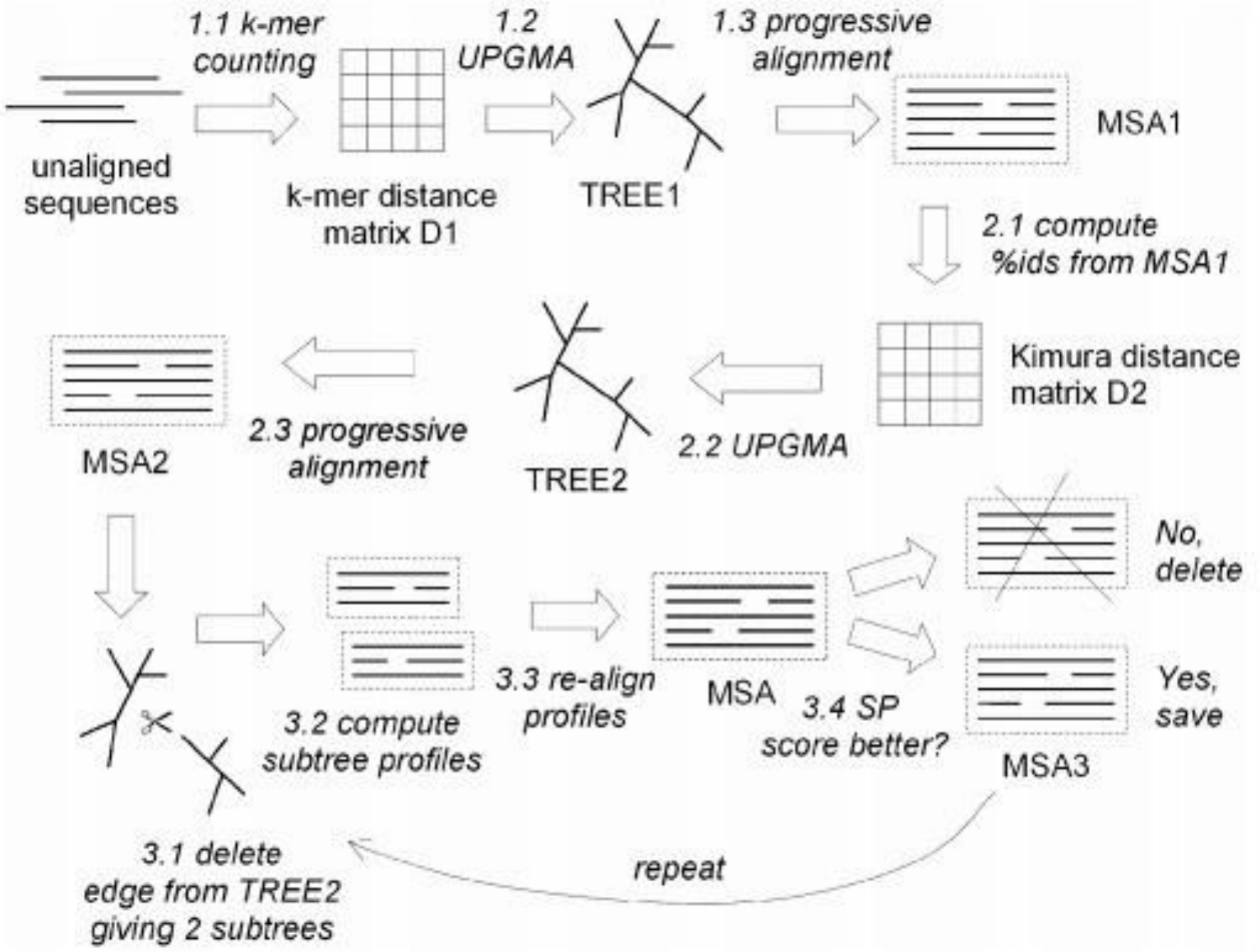
```
1   PEEKSAVTALWGKVN--VDEVGG
2   GEEKAAVLALWDKVN--EEEVGG
3    PADKTNVKAAWGKVGAHAGEYGA
4    AADKTNVKAAWSKVGGHAGEYGA
5    EHEWQLVLHVWAKVEADVAGHGQ
```

2 3 4 1

Alinhamento global progressivo
das folhas até a raiz.

# MUSCLE

Nucl. Acids Res. 2004 32:1792-1797

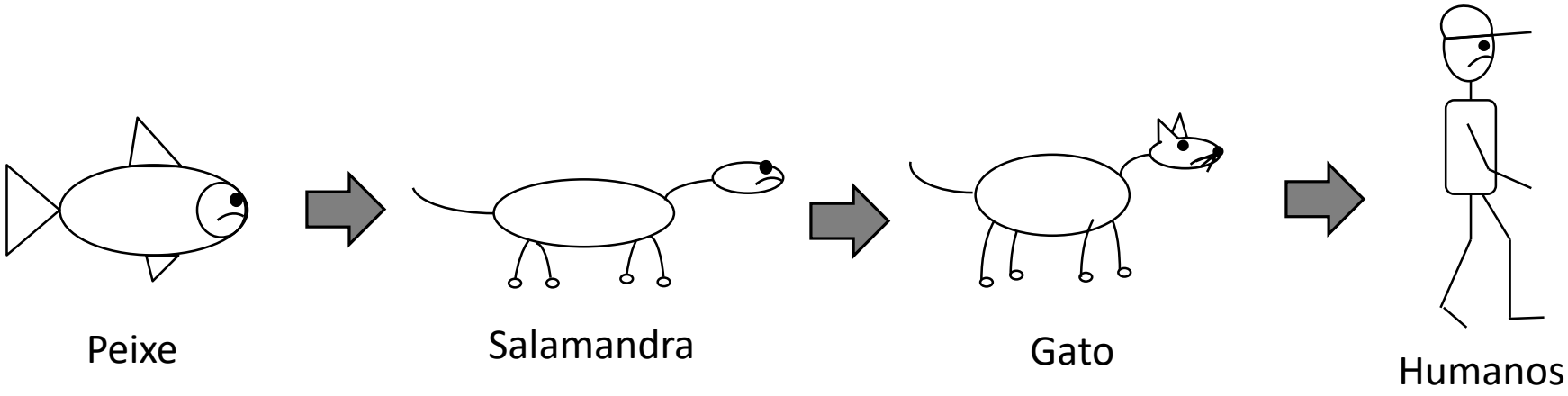



**Rascunho**

**Progressivo**

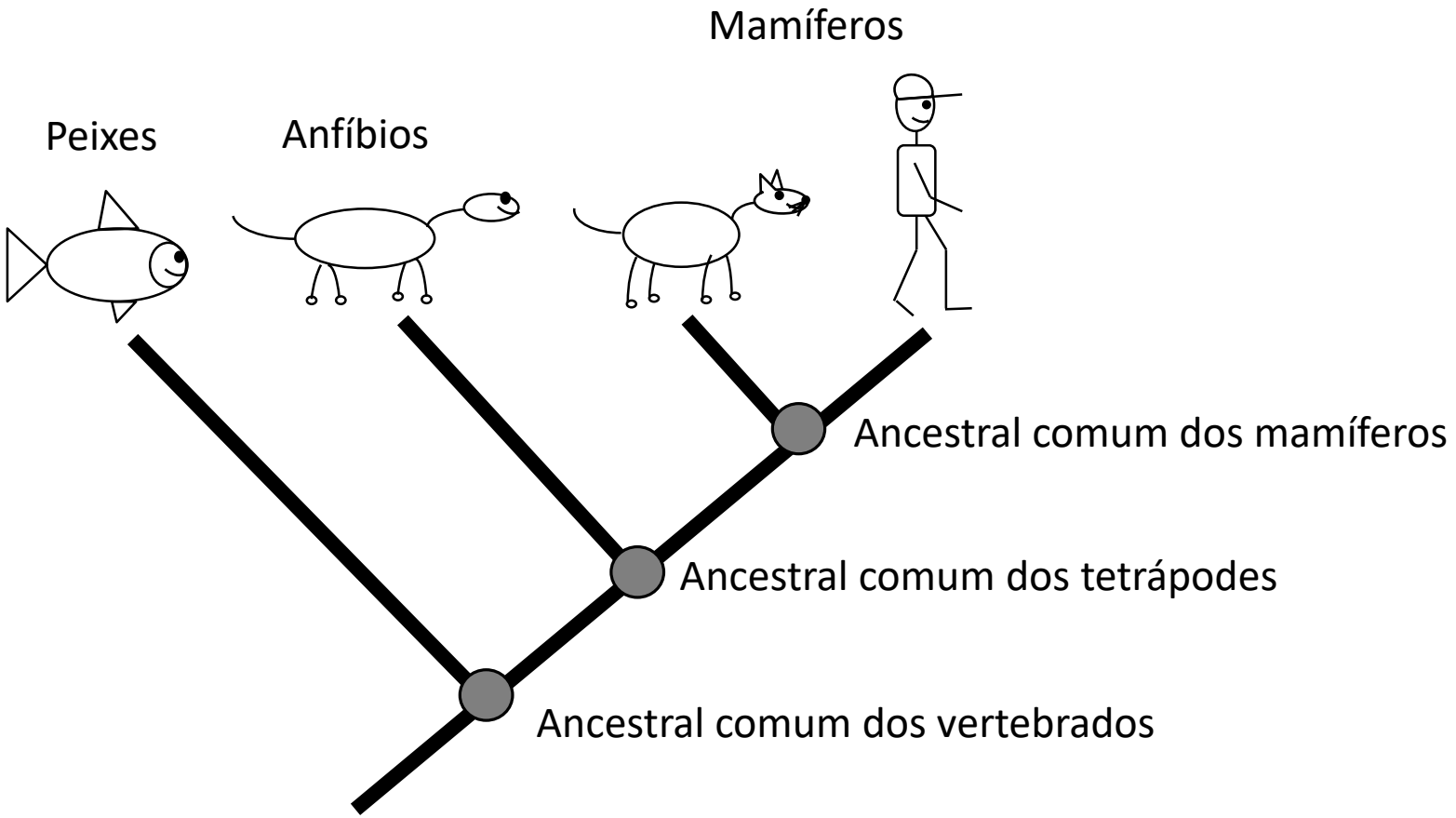
**Refinamento**

# Árvores filogenéticas e evolução



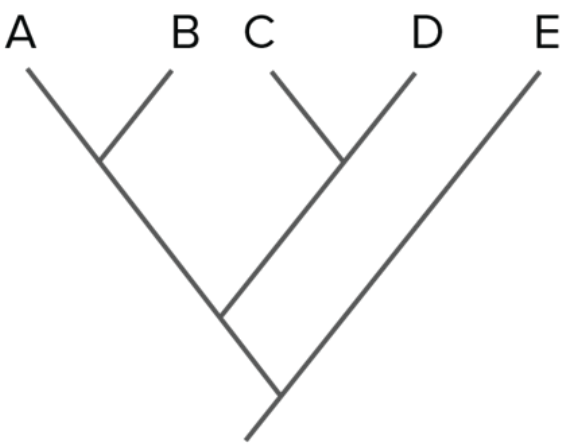

Isto **NÃO** é evolução

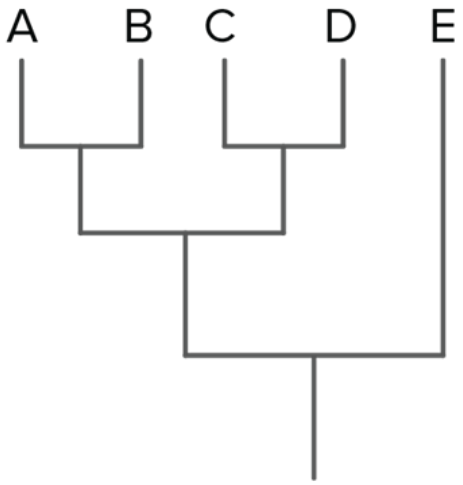
# Árvores filogenéticas e evolução

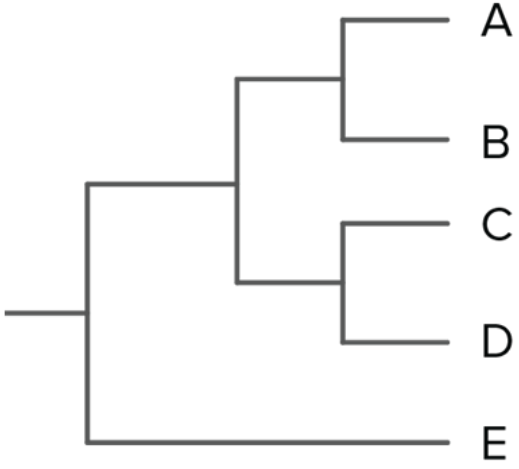



Isto **É** evolução

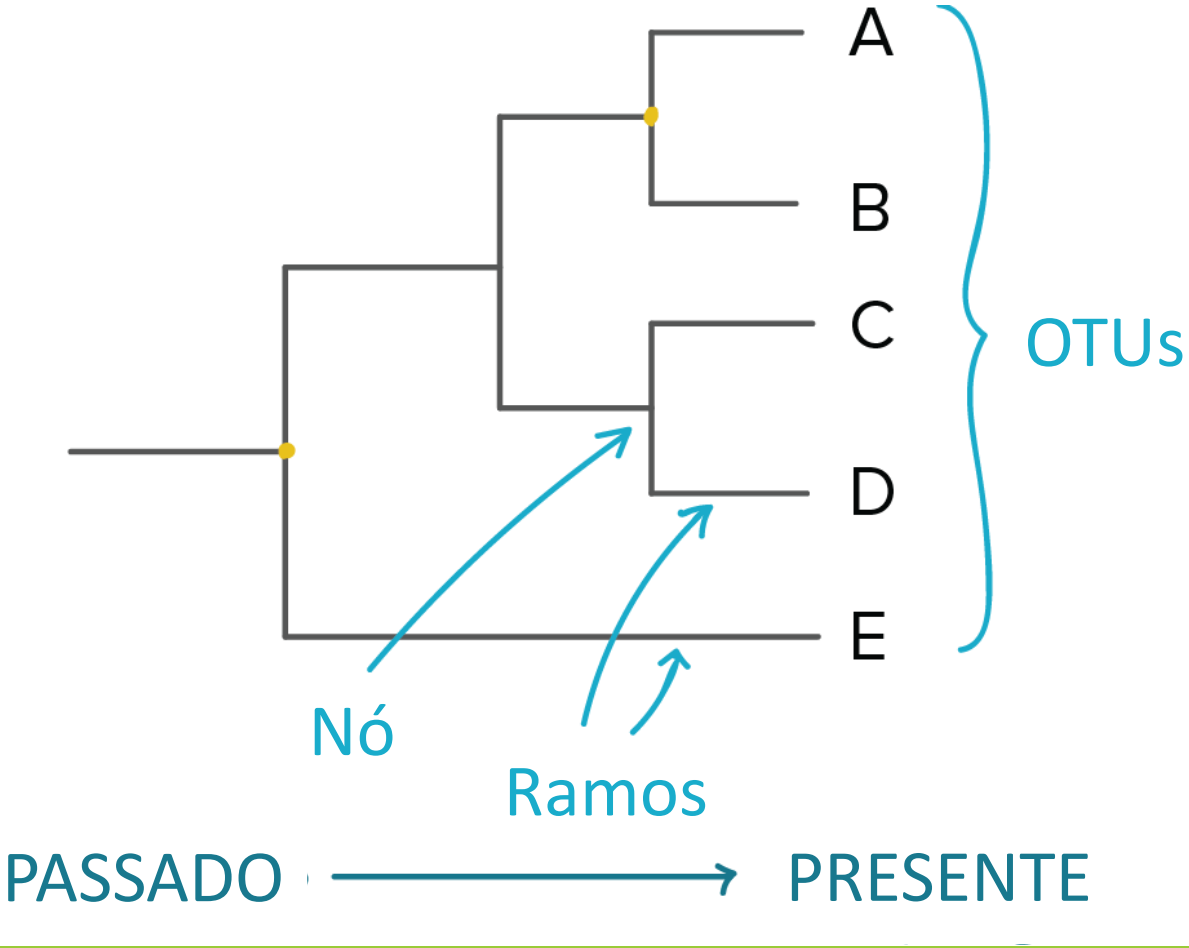
# Anatomia de uma árvore filogenética



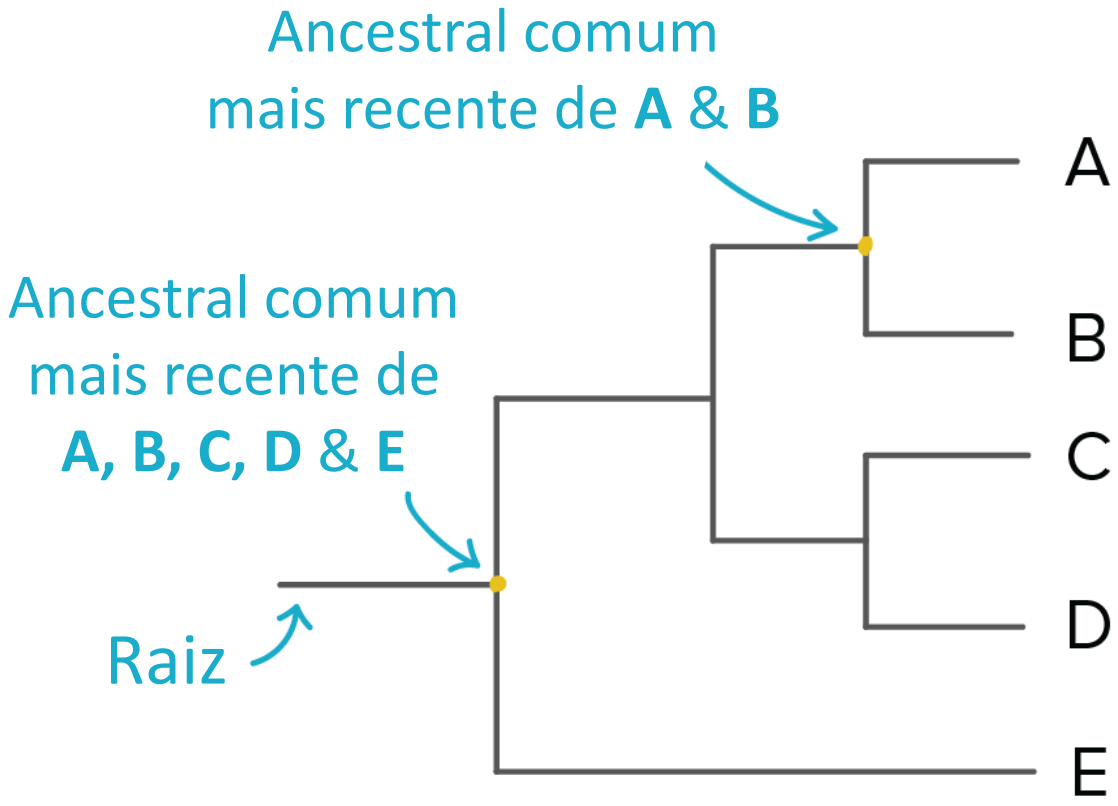



# Anatomia de uma árvore filogenética

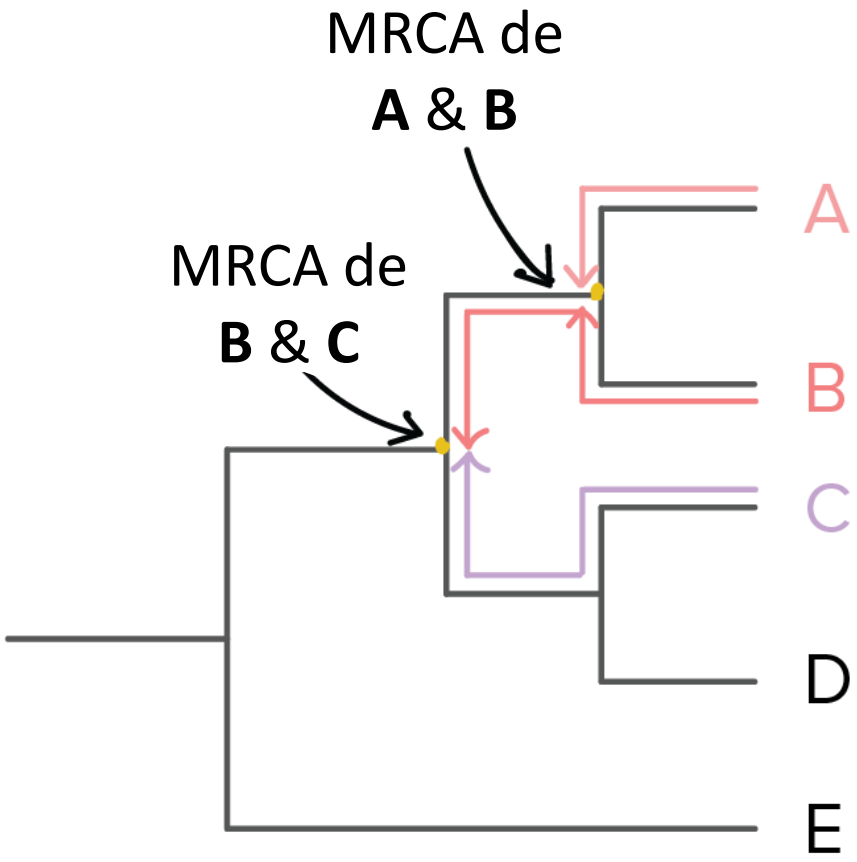


# Anatomia de uma árvore filogenética

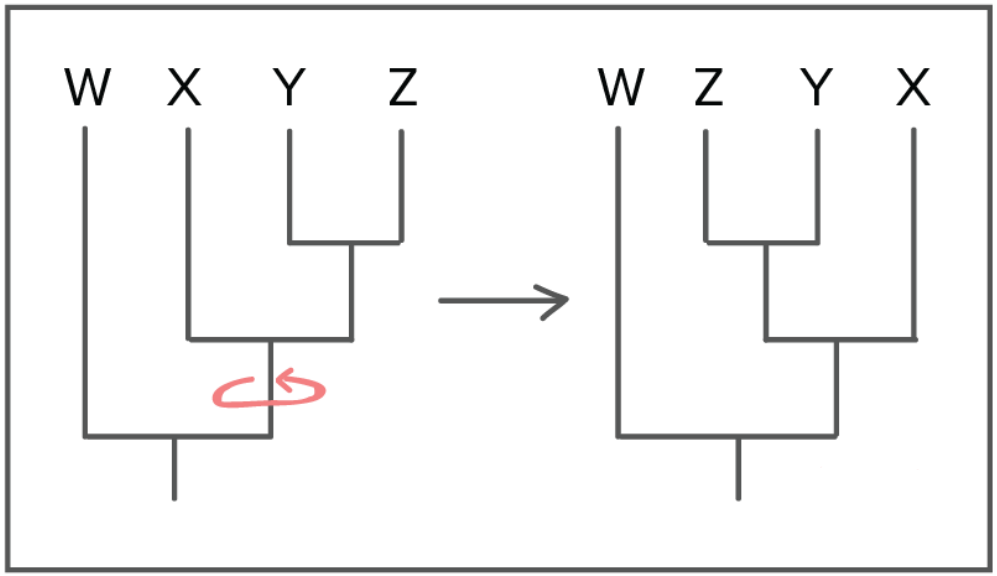


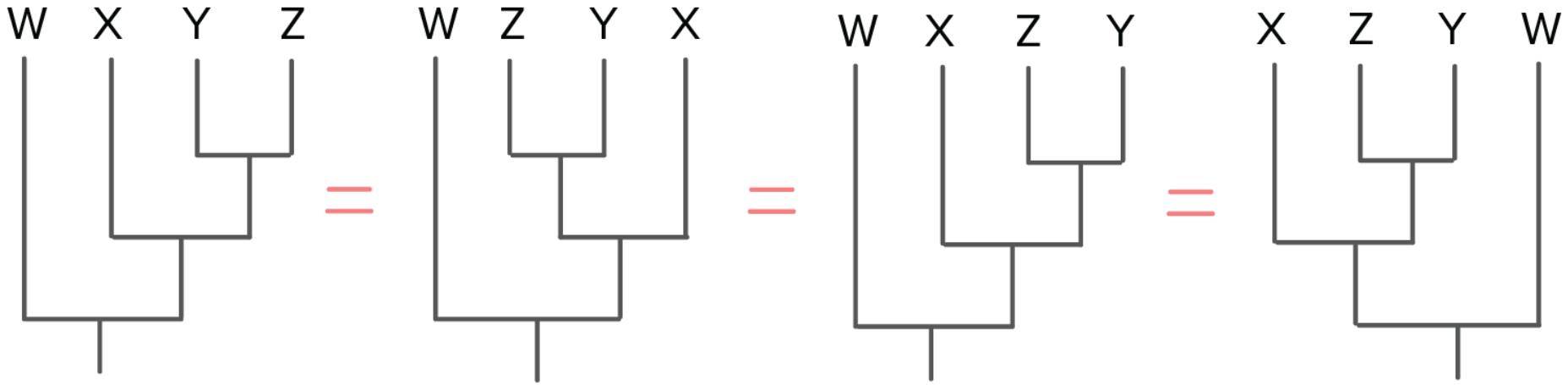

Ancestral comum mais recente = **MRCA**

# Anatomia de uma árvore filogenética

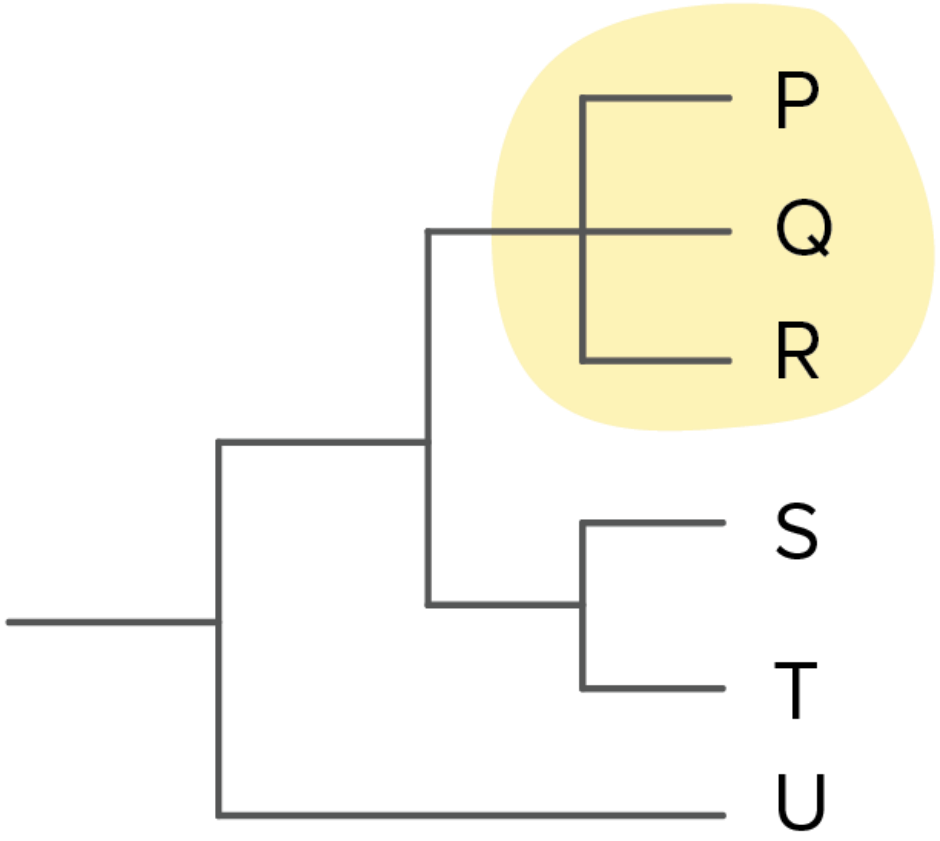


# Anatomia de uma árvore filogenética



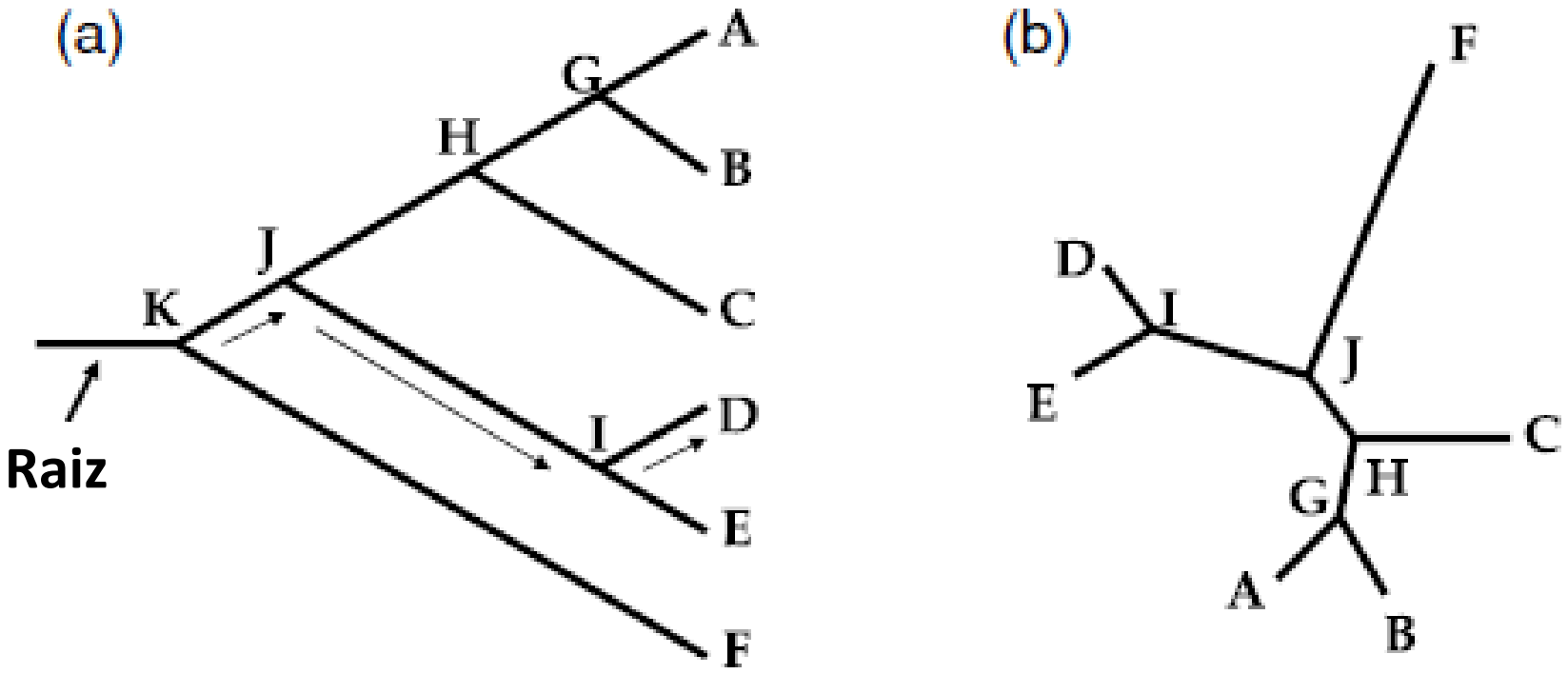



# Anatomia de uma árvore filogenética

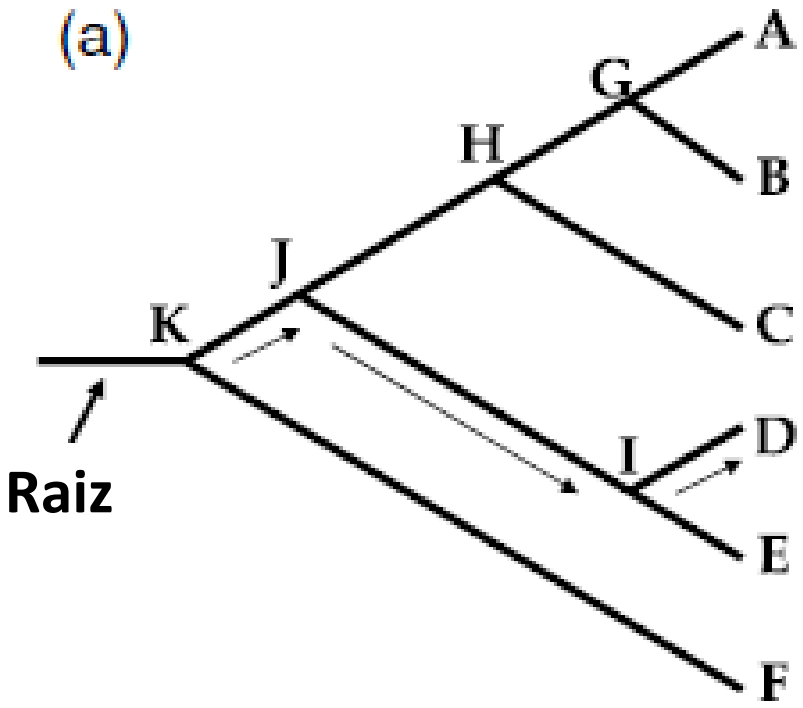


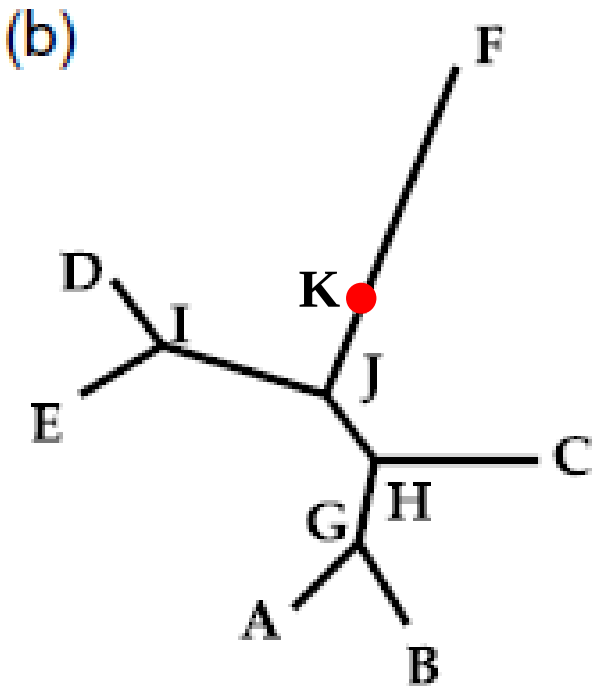

**Politomia**

# Anatomia de uma árvore filogenética

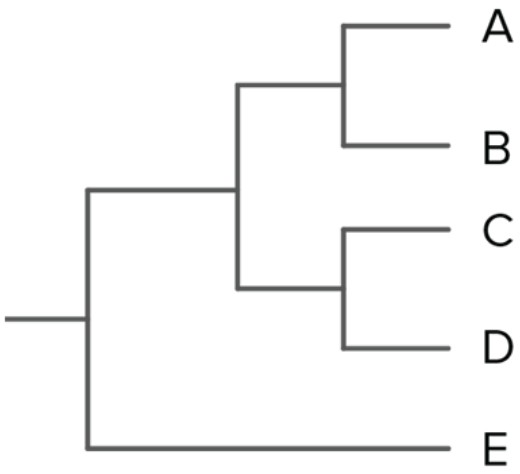


# Anatomia de uma árvore filogenética

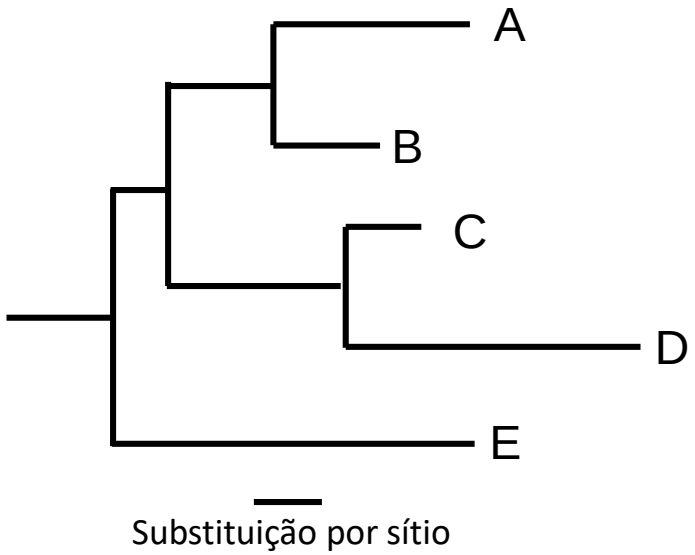


# Anatomia de uma árvore filogenética

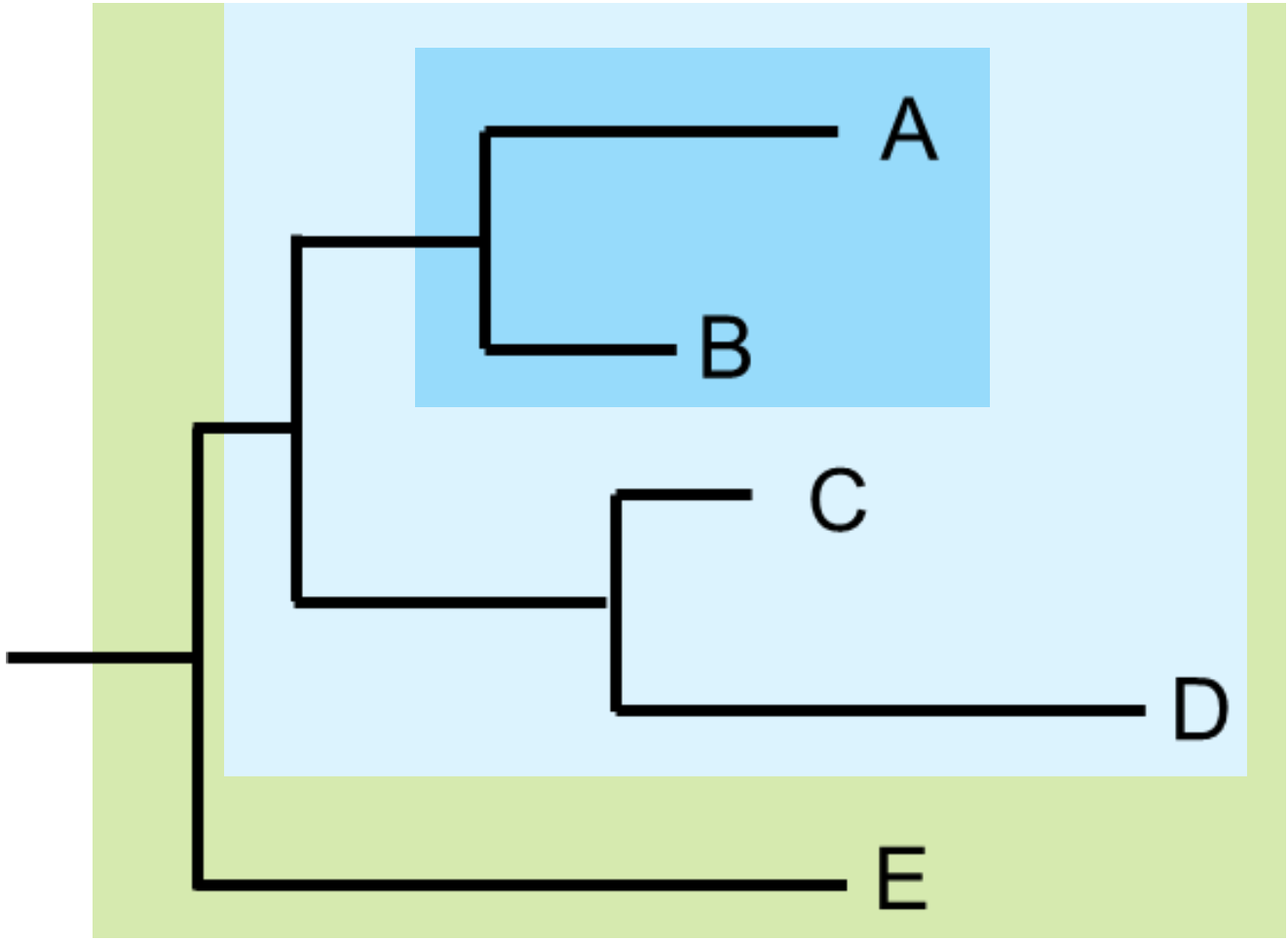



Cladograma




Filograma

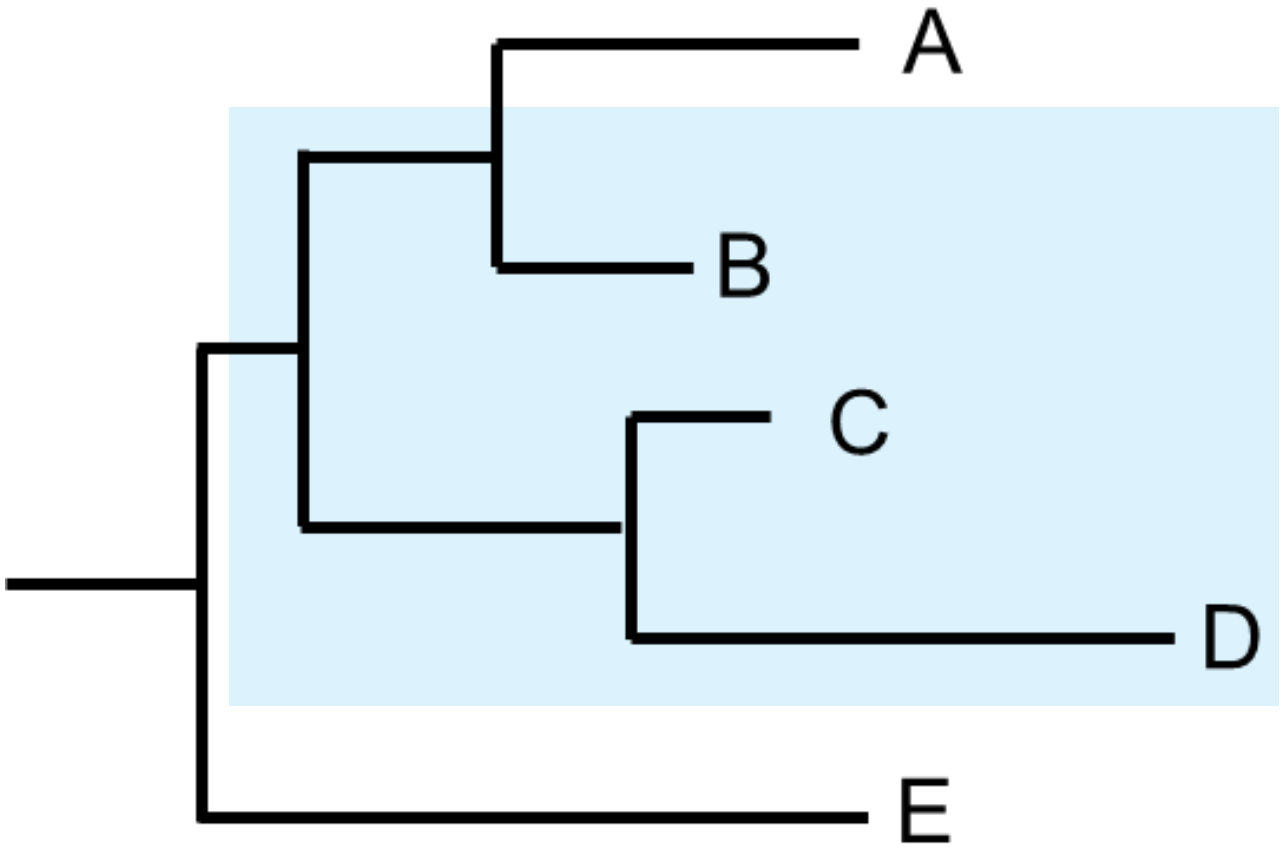
# Entendendo filogenias




**São CLADOS**

**Monofilia**

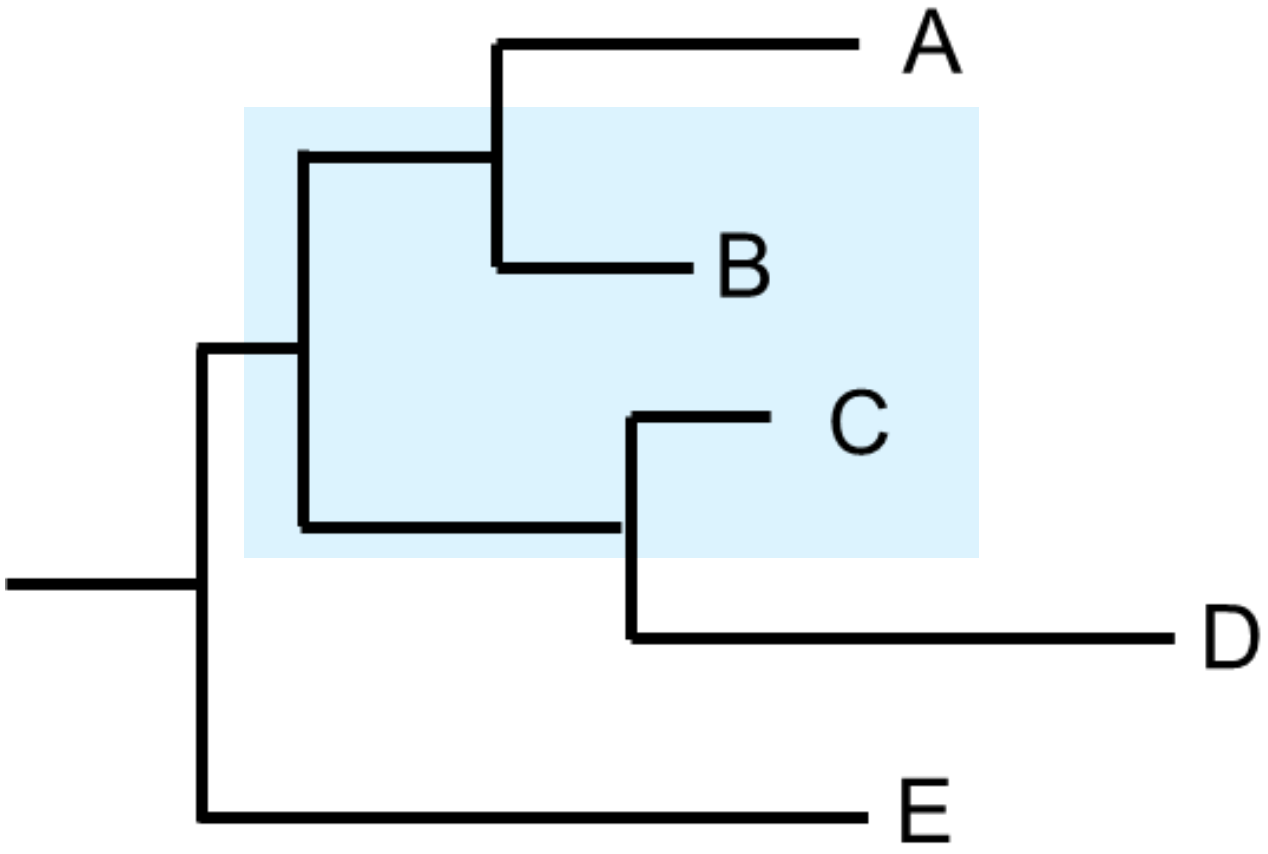
# Entendendo filogenias




**NÃO são um CLADO**

**Parafilia**

# Entendendo filogenias




**NÃO São um CLADO**

**Polifilia**