

1º Simpósio  **LAVIREO**

# *Neighbor-joining*

---

**Edson Delatorre**

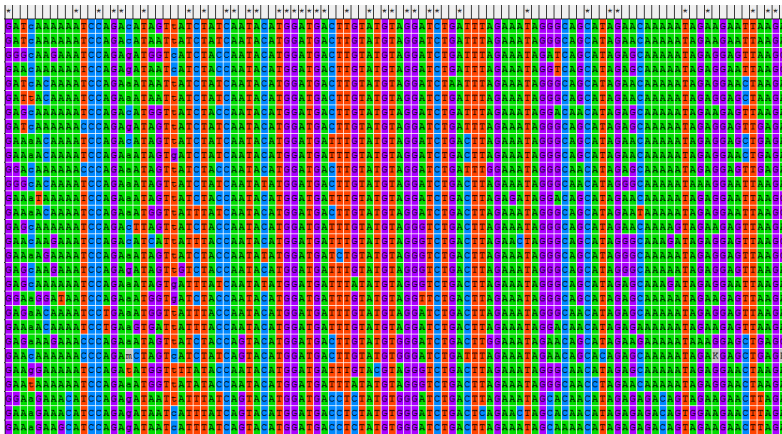
Lab. de Genética Molecular de Microrganismos  
Instituto Oswaldo Cruz/FIOCRUZ

[delatorre.ioc@gmail.com](mailto:delatorre.ioc@gmail.com)

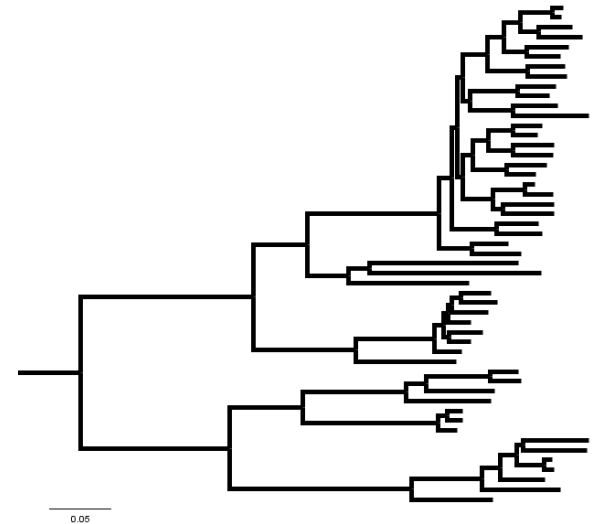
Rio de Janeiro – RJ, janeiro de 2019

# Modelando a evolução

---



Alinhamento



Árvore filogenética

# Métodos de Reconstrução de Filogenias

---

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

# Métodos baseados em distância

---

Ajustam uma árvore a uma matriz  
distâncias genéticas entre pares de  
diferentes OTUs

Algoritmos baseados em distância incluem:

- Evolução Mínima
- UPGMA
- Neighbor-joining

# Evolução mínima

---

- A árvore com o comprimento mínimo, que é a soma do comprimento de todos os ramos da árvore, é escolhida como a melhor estimativa da árvore.

$$\text{Trees} = \frac{(2n-3)!}{(2^{n-2}(n-2))!}$$

2 sequências:	1
3 sequências :	3
4 sequências :	15
5 sequências :	105
6 sequências :	954
7 sequências :	10395
8 sequências :	135135
9 sequências :	2027025
10 sequências :	34459425
51 sequências :	>10 <sup>80</sup>

# Método de NJ

---

- O **NJ** (Saitou and Nei, 1987) é um método de busca heurística para estimar a árvore de evolução mínima.
- O método de **NJ** está baseado no princípio evolução mínima e constrói em cada etapa novos nós internos juntando os vizinhos mais próximos, a partir de uma árvore tipo-estrela.

# Método de NJ

---

Matrix de distâncias genéticas  
P-distance

	A	B	C	D	E	F
A	0	5	4	7	6	8
B	5	0	7	10	9	11
C	4	7	0	7	6	8
D	7	10	7	0	5	9
E	6	9	6	5	0	8
F	8	11	8	9	8	0

# Método de NJ

Uma nova matrix é computada, corrigida com base na taxa global de divergência

	A	B	C	D	E	F
A	0	5	4	7	6	8
B	5	0	7	10	9	11
C	4	7	0	7	6	8
D	7	10	7	0	5	9
E	6	9	6	5	0	8
F	8	11	8	9	8	0



	A	B	C	D	E	F
A	0	-13	-11.5	-10	-10	-10.5
B	-13	0	-11.5	-10	-10	-10.5
C	-11.5	-11.5	0	-10.5	-10.5	-11
D	-10	-10	-10.5	0	-13	-11.5
E	-10	-10	-10.5	-13	0	-11.5
F	-10.5	-10.5	-11	-11.5	-11.5	0

$$\text{Div}(A) = \sum_i \text{dist}(A,i) = 5+4+7+6+8 = 30$$

$$\text{Div}(B) = \sum_i \text{dist}(B,i) = 5+7+10+9+11 = 42$$

$$\text{Div}(C) = \sum_i \text{dist}(C,i) = 32$$

$$\text{Div}(D) = \sum_i \text{dist}(D,i) = 38$$

$$\text{Div}(E) = \sum_i \text{dist}(E,i) = 34$$

$$\text{Div}(F) = \sum_i \text{dist}(F,i) = 44$$

$$M(i,j) = \text{dist}(i,j) - \frac{(\text{Div}(i) + \text{Div}(j))}{N - 2}$$

$$M(A,B) = 5 - (30+42)/4 = -13$$

$$M(A,C) = 4 - (30+32)/4 = -11.5$$

etc...



# Método de NJ

Os comprimentos dos ramos entre os OTUs A e B e seu ancestral são computados:

	A	B	C	D	E	F
A	0	5	4	7	6	8
B	5	0	7	10	9	11
C	4	7	0	7	6	8
D	7	10	7	0	5	9
E	6	9	6	5	0	8
F	8	11	8	9	8	0

**r**      **30**    **42**    **32**    **38**    **34**    **44**



	A	B	C	D	E	F
A	0	-13	-11.5	-10	-10	-10.5
B	-13	0	-11.5	-10	-10	-10.5
C	-11.5	-11.5	0	-10.5	-10.5	-11
D	-10	-10	-10.5	0	-13	-11.5
E	-10	-10	-10.5	-13	0	-11.5
F	-10.5	-10.5	-11	-11.5	-11.5	0

$$S(AU) = \text{dist}(A,B)/2 + (\text{Div}(B) - \text{Div}(A))/2(N-2)$$

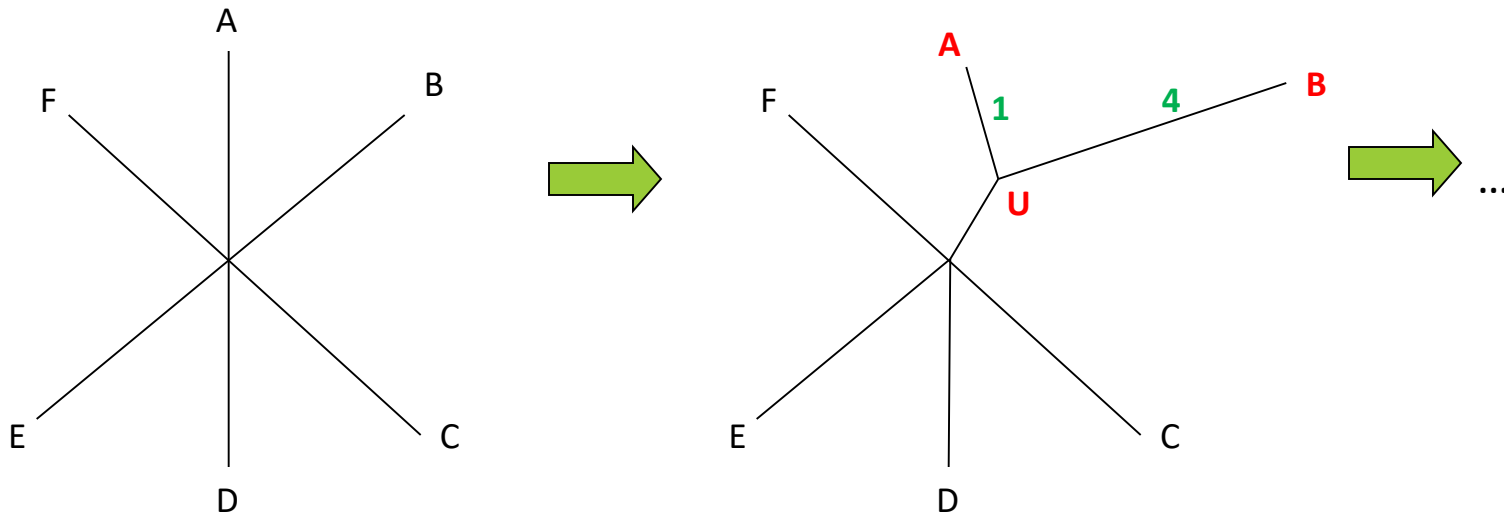
$$S(AU) = 5/2 + (42 - 30)/2(6-2) = 4$$

$$S(BU) = \text{dist}(A,B) - S(AU) = 1$$

# Método de NJ

---


Os comprimentos dos ramos entre os OTUs A e B e seu ancestral são computados:



# Método de NJ

As novas distâncias do nó U para cada um dos nós adjacentes é calculada:

	A	B	C	D	E	F
A	0	5	4	7	6	8
B	5	0	7	10	9	11
C	4	7	0	7	6	8
D	7	10	7	0	5	9
E	6	9	6	5	0	8
F	8	11	8	9	8	0
<b>r</b>	<b>30</b>	<b>42</b>	<b>32</b>	<b>38</b>	<b>34</b>	<b>44</b>



	A	B	C	D	E	F
A	0	-13	-11.5	-10	-10	-10.5
B	-13	0	-11.5	-10	-10	-10.5
C	-11.5	-11.5	0	-10.5	-10.5	-11
D	-10	-10	-10.5	0	-13	-11.5
E	-10	-10	-10.5	-13	0	-11.5
F	-10.5	-10.5	-11	-11.5	-11.5	0

$$\text{dist}(\text{C}, \text{U}) = (\text{dist}(\text{A}, \text{C}) + \text{dist}(\text{B}, \text{C}) - \text{dist}(\text{A}, \text{B}))/2 = (4 + 7 - 5)/2 = \mathbf{3}$$

$$\text{dist}(\text{D}, \text{U}) = (\text{dist}(\text{A}, \text{D}) + \text{dist}(\text{B}, \text{D}) - \text{dist}(\text{A}, \text{B}))/2 = (7 + 10 - 5)/2 = \mathbf{6}$$

$$\text{dist}(\text{E}, \text{U}) = (\text{dist}(\text{A}, \text{E}) + \text{dist}(\text{B}, \text{E}) - \text{dist}(\text{A}, \text{B}))/2 = (6 + 9 - 5)/2 = \mathbf{5}$$

$$\text{dist}(\text{F}, \text{U}) = (\text{dist}(\text{A}, \text{F}) + \text{dist}(\text{B}, \text{F}) - \text{dist}(\text{A}, \text{B}))/2 = (8 + 11 - 5)/2 = \mathbf{7}$$

# Método de NJ

Uma nova matrix de distância é computada, e posteriormente corrigida com base na taxa global de divergência

	U	C	D	E	F
U	0	3	6	5	7
C	3	0	7	6	8
D	6	7	0	5	9
E	5	6	5	0	8
F	7	8	9	8	0

**r**    **21**   **24**   **27**   **24**   **32**



	U	C	D	E	F
U	0	-12	-10	-10	-10,7
C	-12	0	-11	-10	-10,7
D	-10	-11	0	-12	-10,7
E	-10	-10	-12	0	-10,7
F	-10,7	-10,7	-10,7	-10,7	0

$$M(i,j) = \text{dist}(i,j) - \frac{(\text{Div}(i) + \text{Div}(j))}{N - 2}$$

$$M(U,C) = 3 - (21+24)/3 = -12$$

$$M(U,D) = 6 - (21+27)/3 = -10$$

etc...

# Método de NJ

Os comprimentos dos ramos entre os OTUs U e C e seu ancestral são computados:

	U	C	D	E	F
U	0	3	6	5	7
C	3	0	7	6	8
D	6	7	0	5	9
E	5	6	5	0	8
F	7	8	9	8	0

**r**    **21**   **24**   **27**   **24**   **32**



	U	C	D	E	F
U	0	-12	-10	-10	-10,7
C	-12	0	-11	-10	-10,7
D	-10	-11	0	-12	-10,7
E	-10	-10	-12	0	-10,7
F	-10,7	-10,7	-10,7	-10,7	0

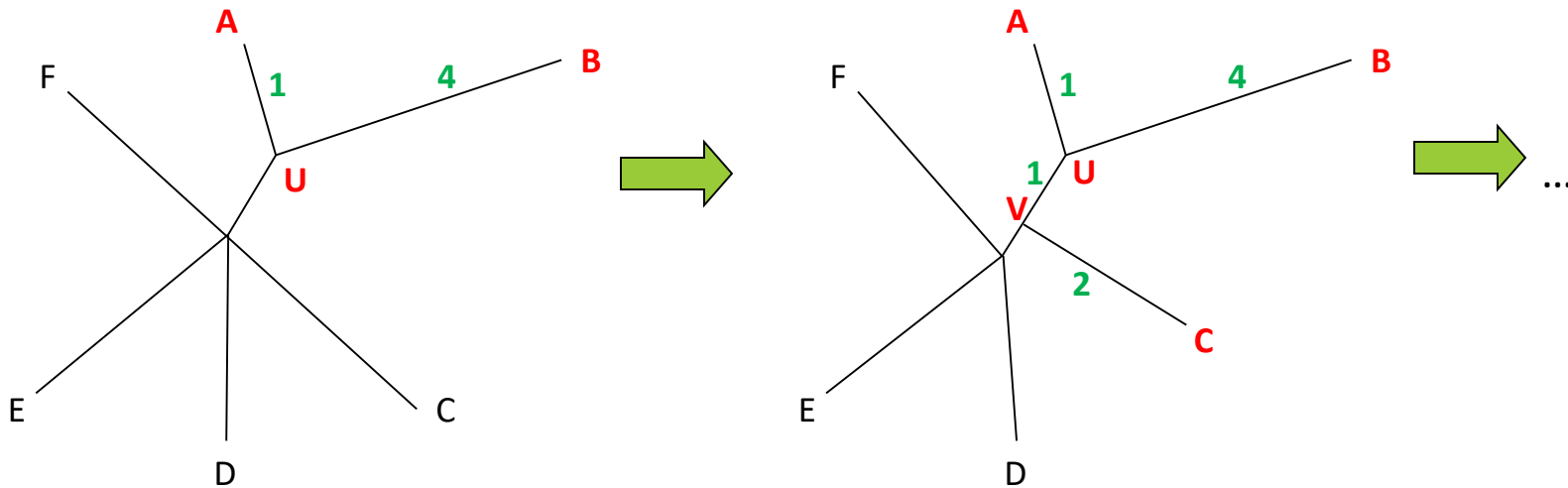
$$S(UV) = \text{dist}(C,U)/2 + (\text{Div}(U) - \text{Div}(C))/2(N-2)$$

$$S(UV) = 3/2 + (21 - 24)/2(5-2) = 1$$

$$S(CV) = \text{dist}(C,U) - S(UV) = 2$$

# Método de NJ

Os comprimentos dos ramos entre os OTUs U e C e seu ancestral são computados:



# Método de NJ

As novas distâncias do nó V para cada um dos nós adjacentes é calculada:

	U	C	D	E	F
U	0	3	6	5	7
C	3	0	7	6	8
D	6	7	0	5	9
E	5	6	5	0	8
F	7	8	9	8	0

**r**    **21**   **24**   **27**   **24**   **32**



	U	C	D	E	F
U	0	-12	-10	-10	-10,7
C	-12	0	-11	-10	-10,7
D	-10	-11	0	-12	-10,7
E	-10	-10	-12	0	-10,7
F	-10,7	-10,7	-10,7	-10,7	0

$$\text{dist}(\mathbf{V}, \mathbf{D}) = (\text{dist}(\mathbf{C}, \mathbf{D}) + \text{dist}(\mathbf{U}, \mathbf{D}) - \text{dist}(\mathbf{C}, \mathbf{U}))/2 = (7 + 6 - 3)/2 = \mathbf{5}$$

$$\text{dist}(\mathbf{V}, \mathbf{E}) = (\text{dist}(\mathbf{C}, \mathbf{E}) + \text{dist}(\mathbf{U}, \mathbf{E}) - \text{dist}(\mathbf{C}, \mathbf{U}))/2 = (6 + 5 - 3)/2 = \mathbf{4}$$

$$\text{dist}(\mathbf{V}, \mathbf{F}) = (\text{dist}(\mathbf{C}, \mathbf{F}) + \text{dist}(\mathbf{U}, \mathbf{F}) - \text{dist}(\mathbf{C}, \mathbf{U}))/2 = (8 + 7 - 3)/2 = \mathbf{6}$$

# Método de NJ

Uma nova matrix de distâncias é computada, e posteriormente corrigida com base na taxa global de divergência

	V	D	E	F
V	0	5	4	6
D	5	0	5	9
E	4	5	0	8
F	6	9	8	0
r	15	19	17	23



	V	D	E	F
V	0	-12	-12	-13
D	-12	0	-13	-12
E	-12	-13	0	-12
F	-13	-12	-12	0

$$M(i,j) = \text{dist}(i,j) - \frac{(\text{Div}(i) + \text{Div}(j))}{N - 2}$$

$$M(V,D) = 5 - (15+19)/2 = -12$$

$$M(V,E) = 4 - (15+17)/3 = -12$$

etc...



# Método de NJ

Os comprimentos dos ramos entre os OTUs D e E e seu ancestral são computados:

	V	D	E	F
V	0	5	4	6
D	5	0	5	9
E	4	5	0	8
F	6	9	8	0
<b>r</b>	<b>15</b>	<b>19</b>	<b>17</b>	<b>23</b>



	V	D	E	F
V	0	-12	-12	-13
D	-12	0	-13	-12
E	-12	-13	0	-12
F	-13	-12	-12	0

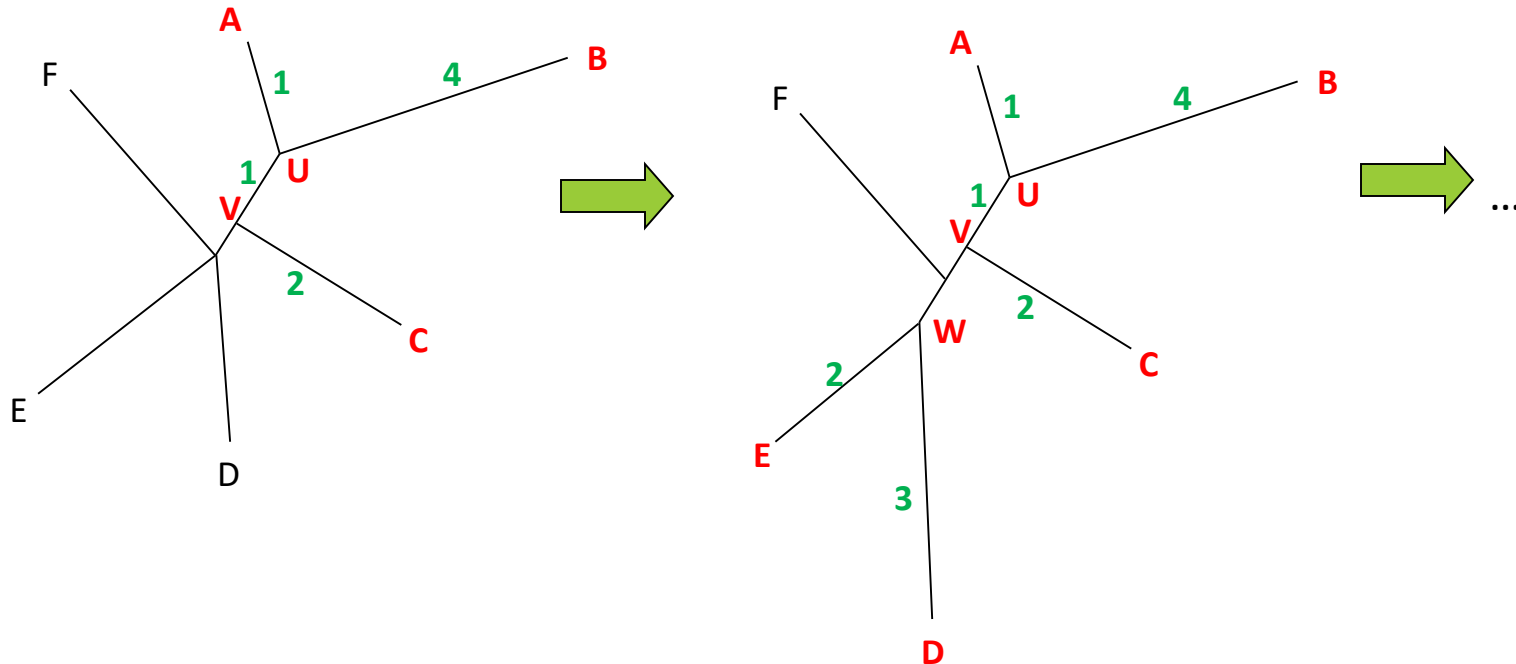
$$S(DW) = \text{dist}(D,E)/2 + (\text{div}(D) - \text{div}(E))/2(N-2)$$

$$S(DW) = 5/2 + (19 - 17)/2(4-2) = 3$$

$$S(EW) = \text{dist}(D,E) - S(DW) = 2$$

# Método de NJ

Os comprimentos dos ramos entre os OTUs D e E e seu ancestral são computados:



# Método de NJ

As novas distâncias do nó W para cada um dos nós adjacentes é calculada:

	V	D	E	F
V	0	5	4	6
D	5	0	5	9
E	4	5	0	8
F	6	9	8	0
<b>r</b>	<b>15</b>	<b>19</b>	<b>17</b>	<b>23</b>



	V	D	E	F
V	0	-12	-12	-13
D	-12	0	-13	-12
E	-12	-13	0	-12
F	-13	-12	-12	0

$$\text{dist}(V,W) = (\text{dist}(D,V) + \text{dist}(E,V) - \text{dist}(D,E))/2 = (5 + 4 - 5)/2 = 2$$

$$\text{dist}(F,W) = (\text{dist}(D,F) + \text{dist}(E,F) - \text{dist}(D,E))/2 = (9 + 8 - 5)/2 = 6$$

# Método de NJ

Uma nova matrix de distâncias é computada, e posteriormente corrigida com base na taxa global de divergência

	W	V	F
W	0	2	6
V	2	0	6
F	6	6	0
r	8	8	12



	W	V	F
W	0	-14	-14
V	-14	0	-14
F	-14	-14	0

$$M(i, j) = \text{dist}(i, j) - \frac{(\text{Div}(i) + \text{Div}(j))}{N - 2}$$

$$M(W, V) = 2 - (8+8)/1 = -14$$

$$M(W, F) = 6 - (8+12)/1 = -14$$

$$M(V, F) = 6 - (8+12)/1 = -14$$

# Método de NJ

Os comprimentos dos ramos entre os OTUs D e E e seu ancestral são computados:

	W	V	F
W	0	2	6
V	2	0	6
F	6	6	0
<b>r</b>	<b>8</b>	<b>8</b>	<b>12</b>



	W	V	F
W	0	-14	-14
V	-14	0	-14
F	-14	-14	0

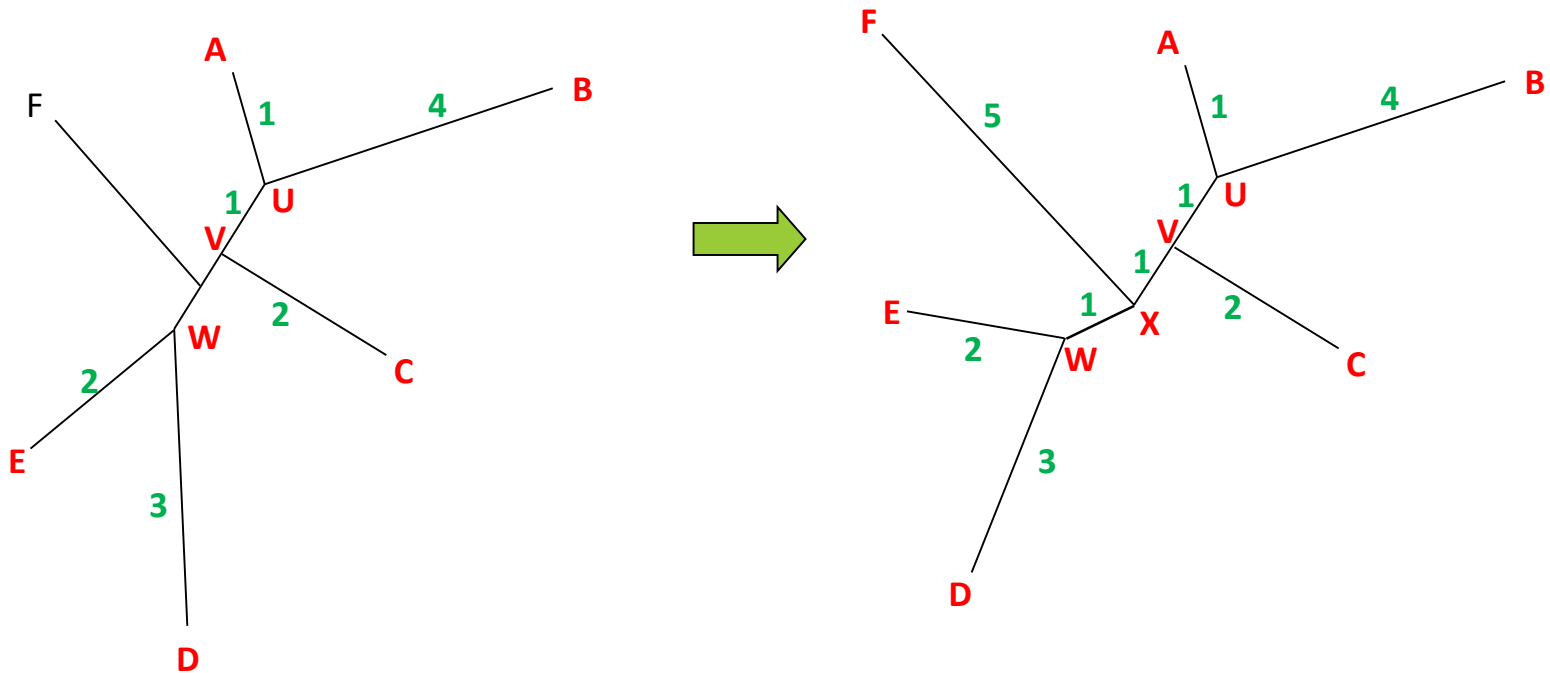
$$S(VX) = \text{dist}(F,V)/2 + (\text{div}(V) - \text{div}(F))/2(N-2)$$

$$S(VX) = 6/2 + (8 - 12)/2(3-2) = 1$$

$$S(FX) = \text{dist}(F,V) - S(VX) = 5$$

# Método de NJ

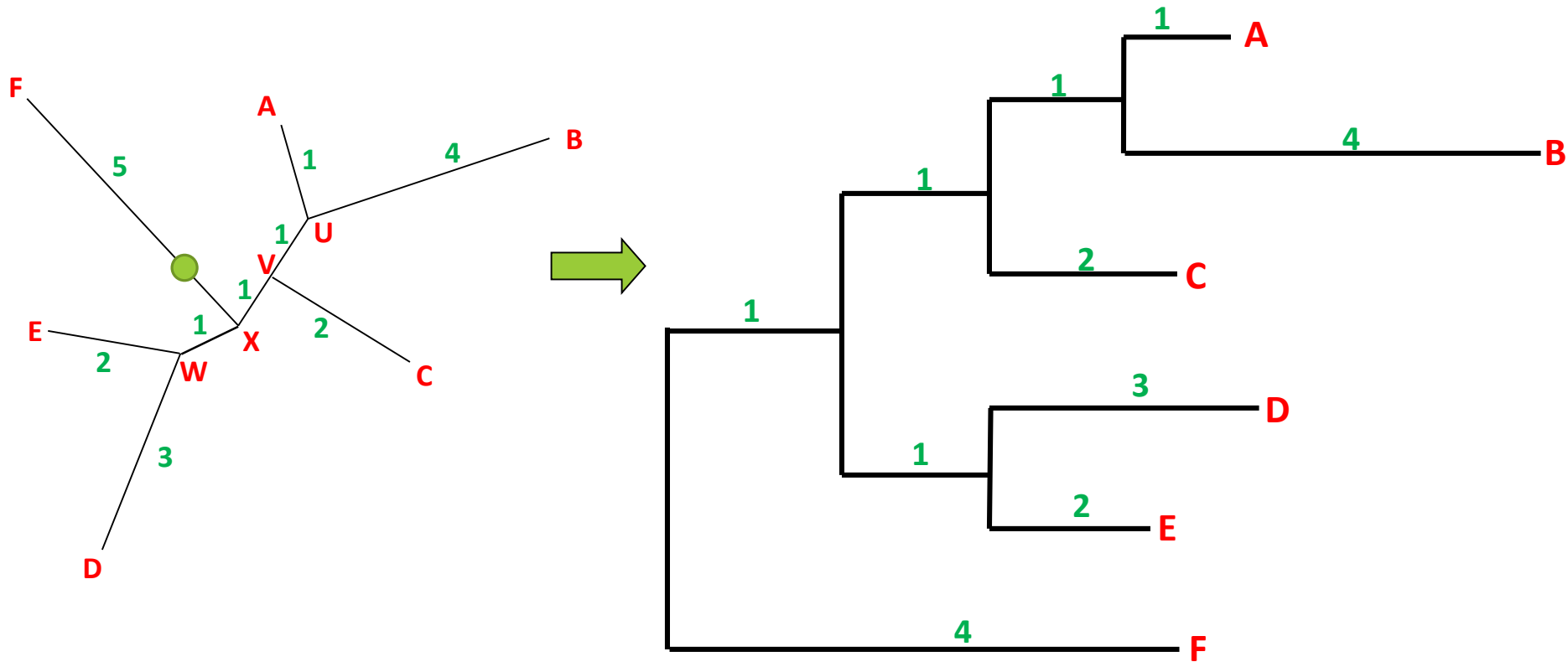
Os comprimentos dos ramos entre os OTUs D e E e seu ancestral são computados:



$$\text{dist}(W,X) = (\text{dist}(F,W) + \text{dist}(V,W) - \text{dist}(F,V))/2 = (6 + 2 - 6)/2 = 1$$

# Método de NJ

Topologia retangular enraizada



# Métodos baseados em distância

---

- Vantagens:
  - Rapidez
  - Útil para analisar grandes datasets ( $> 5000$  sequências)
- Desvantagens:
  - não temos qualquer garantia de encontrar a árvore de ME
  - só produz uma árvore e não temos ideia sobre outras árvores potenciais (igualmente prováveis)



# Confiabilidade da topologia da árvore

---

- O teste mais frequentemente utilizado para avaliar a confiabilidade da topologia da árvore filogenética é o ***bootstrap***.
- O ***bootstrap*** é uma técnica estatística que utiliza reamostragem aleatória dos dados para determinar o erro amostral ou intervalos de confiança para algum parâmetro estimado:
  - 1) É realizada uma amostragem com substituição das posições de um alinhamento de forma de criar muitas **réplicas do dataset**.
  - 2) Se constrói uma árvore filogenética de cada **dataset replicado aleatoriamente**.
  - 3) É computada a frequência com que os grupos são encontrados nas análises do conjunto de dados replicados (**valores de *bootstrap***).

# Bootstrap

## Data set

Posições:	1	2	3	4	5	6	7	8	9
Taxon A :	A	T	G	C	G	A	G	C	T
Taxon B :	A	T	A	C	T	A	G	C	T
Taxon C :	A	T	G	C	T	A	T	C	T

**Passo 1:** Pseudodatasets, repetidos de 200-1000 vezes.

Réplica 1

1 5 6 2 3 1 4 9 5 1

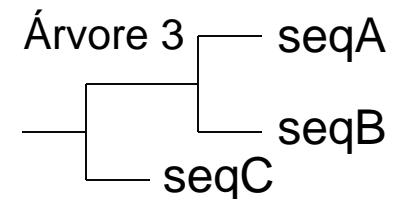
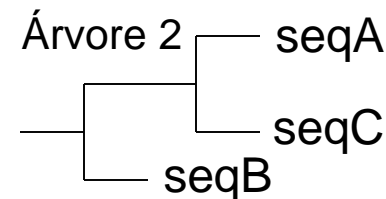
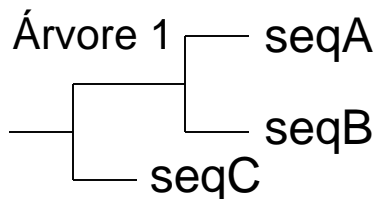
Réplica 2

5 2 3 4 9 2 4 4 1 8

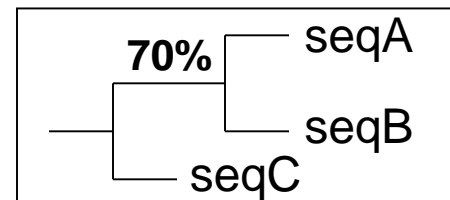
Réplica 3

5 6 0 7 7 1 8 9 0 7

**Passo 2:** Reconstruir árvores a partir de cada pseudodataset.



**Passo 3:** Sumarizar a frequência de ocorrência de cada grupo.



Árvore consenso

# Bootstrap

---

- Valores elevados de *bootstrap* ( $\geq 75\%$ ) são indicativos de forte sinal filogenético nos dados em favor de um determinado cluster filogenético.
- Em algumas circunstâncias, no entanto, um cluster pode apresentar um valor de *bootstrap* elevado, mesmo se for um erro (por exemplo, o agrupamento de sequências com viés de composição de bases similar, ou com um aumento da taxa evolutiva).
- Contrariamente, um valor de *bootstrap* baixo não significa necessariamente que a relação filogenética é falsa, só que é mal suportada (sinal filogenético fraco nos dados).