

# La Estadística detrás del Análisis Filogenético

Hugo Flores Arguedas

Departamento Ciencias y Matemáticas  
Arkansas State University Campus Queretaro  
hfloresarguedas@astate.edu

Escuela de Otono, Biología Matemática, Oct 7-11, 2024



# Árboles Filogenéticos

¿Qué es un árbol filogenético?

## ¿Qué es un árbol filogenético?

Un Árbol Filogenético es una representación gráfica de la historia evolutiva de secuencias biológicas, que nos permite visualizar las relaciones evolutivas entre ellas.






## ¿Qué es un árbol filogenético?

Un Árbol Filogenético es una representación gráfica de la historia evolutiva de secuencias biológicas, que nos permite visualizar las relaciones evolutivas entre ellas.

[nature](#) > [nature ecology & evolution](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 12 July 2024

### The nature of the last universal common ancestor and its impact on the early Earth system

[Edmund R. R. Moody](#) , [Sandra Álvarez-Carretero](#), [Tara A. Mahendrarajah](#), [James W. Clark](#), [Holly C. Betts](#), [Nina Dombrowski](#), [Lénárd L. Szánthó](#), [Richard A. Boyle](#), [Stuart Daines](#), [Xi Chen](#), [Nick Lane](#), [Ziheng Yang](#), [Graham A. Shields](#), [Gergely J. Szöllősi](#), [Anja Spang](#), [Davide Pisani](#) , [Tom A. Williams](#) , [Timothy M. Lenton](#)  & [Philip C. J. Donoghue](#) 

[Nature Ecology & Evolution](#) **8**, 1654–1666 (2024) | [Cite this article](#)

**102k** Accesses | **7** Citations | **1381** Altmetric | [Metrics](#)

Luca Nature

¿Qué es un árbol filogenético?

## ¿Qué es un árbol filogenético?

Un Árbol Filogenético es una representación gráfica de la historia evolutiva de secuencias biológicas, que nos permite visualizar las relaciones evolutivas entre ellas. Tenemos diferentes formas para hacer árboles:

## ¿Qué es un árbol filogenético?

Un Árbol Filogenético es una representación gráfica de la historia evolutiva de secuencias biológicas, que nos permite visualizar las relaciones evolutivas entre ellas. Tenemos diferentes formas para hacer árboles:

- **Distance-Based**

- Unweighted Pair Group Method using Arithmetic average (UPGMA)
- Neighbor Joining (NJ)



## ¿Qué es un árbol filogenético?

Un Árbol Filogenético es una representación gráfica de la historia evolutiva de secuencias biológicas, que nos permite visualizar las relaciones evolutivas entre ellas. Tenemos diferentes formas para hacer árboles:

- **Distance-Based**

- Unweighted Pair Group Method using Arithmetic average (UPGMA)
- Neighbor Joining (NJ)

- **Character-Based**

- Parsimony
- Maximum Likelihood
- Bayesian Inference

**Distance-Based:**

**Distance-Based:**

Los métodos de construcción basados en distancia involucran calcular distancias evolutivas entre secuencias usando modelos de substitución, los cuáles son usados a su vez para construir una matriz de distancia.

**Distance-Based:**

Los métodos de construcción basados en distancia involucran calcular distancias evolutivas entre secuencias usando modelos de substitución, los cuáles son usados a su vez para construir una matriz de distancia.

Los dos métodos basados en distancia más populares son UPGMA y NJ. Estos métodos están inspirados en **técnicas de clustering**.

- Unweighted Pair Group Method using Arithmetic average (UPGMA)
- Neighbor Joining (NJ)

## Métodos basados en distancia

**UPGMA** es el más simple de los métodos basados en distancia que construye un árbol filogenético con raíz usando un clustering secuencial.

## Métodos basados en distancia

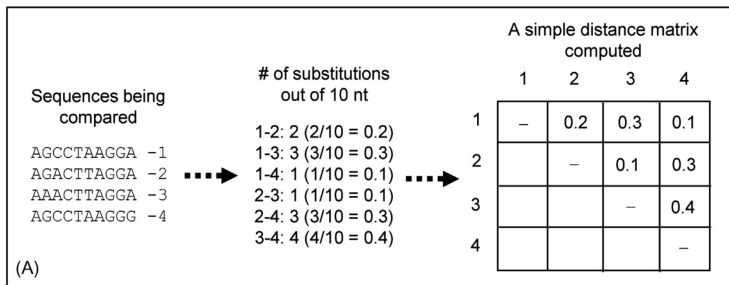
**UPGMA** es el más simple de los métodos basados en distancia que construye un árbol filogenético con raíz usando un clustering secuencial.

Primero, todas las secuencias son comparadas usando el resultado del alineamiento para calcular la matriz de distancia.

# Métodos basados en distancia

**UPGMA** es el más simple de los métodos basados en distancia que construye un árbol filogenético con raíz usando un clustering secuencial.

Primero, todas las secuencias son comparadas usando el resultado del alineamiento para calcular la matriz de distancia.



## Métodos basados en distancia

**UPGMA** es el más simple de los métodos basados en distancia que construye un árbol filogenético con raíz usando un clustering secuencial.



## Métodos basados en distancia

**UPGMA** es el más simple de los métodos basados en distancia que construye un árbol filogenético con raíz usando un clustering secuencial.

Primero, todas las secuencias son comparadas usando el resultado del alineamiento para calcular la matriz de distancia.

## Métodos basados en distancia

**UPGMA** es el más simple de los métodos basados en distancia que construye un árbol filogenético con raíz usando un clustering secuencial.

Primero, todas las secuencias son comparadas usando el resultado del alineamiento para calcular la matriz de distancia.

Usando esta matriz, **las dos secuencias con la menor distancia por pares se agrupan como un solo par**. Un nodo se coloca en el punto medio entre ellas.

## Métodos basados en distancia

**UPGMA** es el más simple de los métodos basados en distancia que construye un árbol filogenético con raíz usando un clustering secuencial.

Primero, todas las secuencias son comparadas usando el resultado del alineamiento para calcular la matriz de distancia.

Usando esta matriz, **las dos secuencias con la menor distancia por pares se agrupan como un solo par**. Un nodo se coloca en el punto medio entre ellas.

Checa el siguiente video, empezando en el minuto 1:55,

[https://www.youtube.com/watch?v=\\_b82GJhx8VM](https://www.youtube.com/watch?v=_b82GJhx8VM)

## Métodos basados en distancia

O bien, considera el siguiente ejemplo:

## Métodos basados en distancia

O bien, considera el siguiente ejemplo:

	A	B	C	D
A	0			
B	3	0		
C	5	4	0	
D	7	1	2	0

Matrix 1

## Métodos basados en distancia

O bien, considera el siguiente ejemplo:

	A	B	C	D
A	0			
B	3	0		
C	5	4	0	
D	7	1	2	0

Matrix 1

Usando esta matriz, **las dos secuencias con menor distancia son B y D**. Así, estas dos secuencias se agrupan en un par.

# Métodos basados en distancia

O bien, considera el siguiente ejemplo:

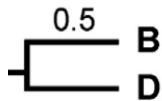
	A	B	C	D
A	0			
B	3	0		
C	5	4	0	
D	7	1	2	0

Matrix 1

Usando esta matriz, **las dos secuencias con menor distancia son B y D**. Así, estas dos secuencias se agrupan en un par.

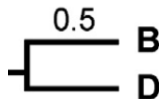
Posteriormente, la distancia entre este par y todas las otras secuencias se recalculan para formar una nueva matriz (ver siguiente diapositiva).

## Métodos basados en distancia





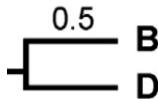
## Métodos basados en distancia



$$d(A, BD) = \{d(A, B) + d(A, D)\} / 2 = (3 + 7) / 2 = 5$$

$$d(BD, C) = \{d(B, C) + d(C, D)\} / 2 = (4 + 2) / 2 = 3$$

## Métodos basados en distancia



$$d(A, BD) = \{d(A, B) + d(A, D)\} / 2 = (3 + 7) / 2 = 5$$

$$d(BD, C) = \{d(B, C) + d(C, D)\} / 2 = (4 + 2) / 2 = 3$$

	A	BD	C
A	0		
BD	5	0	
C	5	3	0

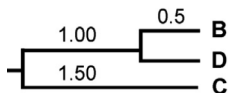
Matrix 2

## Métodos basados en distancia

Con esta nueva matriz, se identifica y se agrupa la secuencia más cercana al primer par.

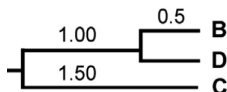
## Métodos basados en distancia

Con esta nueva matriz, se identifica y se agrupa la secuencia más cercana al primer par.



# Métodos basados en distancia

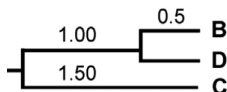
Con esta nueva matriz, se identifica y se agrupa la secuencia más cercana al primer par.



El proceso se repite hasta que todas las secuencias hayan sido posicionadas en el árbol.

## Métodos basados en distancia

Con esta nueva matriz, se identifica y se agrupa la secuencia más cercana al primer par.



El proceso se repite hasta que todas las secuencias hayan sido posicionadas en el árbol.

$$d(A, BDC) = \{d(A, B) + d(A, D) + d(A, C)\} / 3 = (3 + 7 + 5) / 3 = 5$$

## Métodos basados en distancia

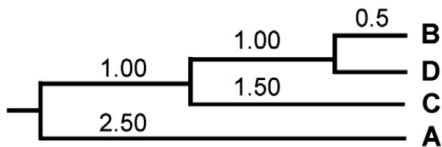
	A	BDC
A	0	
BDC	5	0

Matrix 3

## Métodos basados en distancia

	A	BDC
A	0	
BDC	<b>5</b>	0

Matrix 3



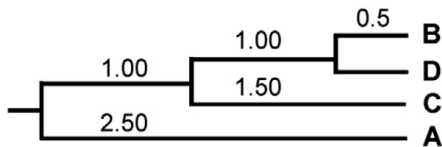
The Tree



## Métodos basados en distancia

	A	BDC
A	0	
BDC	5	0

Matrix 3



The Tree

El método UPGMA **asume** que la tasa evolutiva de todos los taxons es **constante**, por lo que son equidistantes a la raíz.

## Métodos basados en distancia

El método neighbor-joining (NJ) es **el más comúnmente utilizado entre los métodos basados en distancia.**

## Métodos basados en distancia

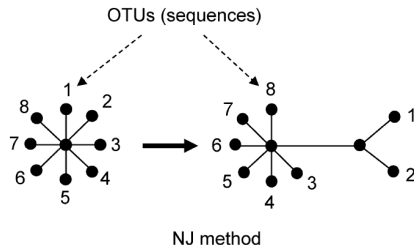
El método neighbor-joining (NJ) es **el más comúnmente utilizado entre los métodos basados en distancia**.

Es similar al método UPGMA , sin embargo, no asume la tasa constante, por lo que produce un árbol sin raíz.

## Métodos basados en distancia

El método neighbor-joining (NJ) es **el más comúnmente utilizado entre los métodos basados en distancia**.

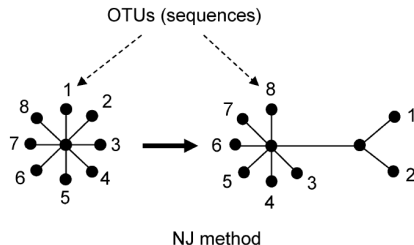
Es similar al método UPGMA , sin embargo, no asume la tasa constante, por lo que produce un árbol sin raíz.



# Métodos basados en distancia

El método neighbor-joining (NJ) es **el más comúnmente utilizado entre los métodos basados en distancia**.

Es similar al método UPGMA , sin embargo, no asume la tasa constante, por lo que produce un árbol sin raíz.



Checa el siguiente video, empezando en el minuto 7:51 para más detalles,

[https://www.youtube.com/watch?v=\\_b82GJhx8VM](https://www.youtube.com/watch?v=_b82GJhx8VM)

## Métodos basados en distancia

Para un ejemplo concreto utilizando Python, se puede consultar el repositorio:

<https://github.com/hugofloresar/EOBM2024>

Carga el notebook PhylogeneticTrees.ipynb en Google Colab:

<https://colab.research.google.com>

## Métodos basados en posición (Character-Based)

Los métodos character-based involucran analizar las secuencias propiamente. Estos métodos evalúan todas las secuencias a la vez analizando un sitio a la vez.

## Métodos basados en posición (Character-Based)

Los métodos character-based involucran analizar las secuencias propiamente. Estos métodos evalúan todas las secuencias a la vez analizando un sitio a la vez.

Estos métodos son generalmente **considerados más exactos** que los métodos basados en distancia. Sin embargo, son **computacionalmente más intensivos** y requieren modelos estadísticos sofisticados.



## Métodos basados en posición (Character-Based)

Los métodos character-based involucran analizar las secuencias propiamente. Estos métodos evalúan todas las secuencias a la vez analizando un sitio a la vez.

Estos métodos son generalmente **considerados más exactos** que los métodos basados en distancia. Sin embargo, son **computacionalmente más intensivos** y requieren modelos estadísticos sofisticados.

Máxima parsimonia (MP) y máxima verosimilitud (ML) son los métodos más conocidos.

# Máxima Verosimilitud

The Maximum Likelihood (ML) method also uses each position in an alignment and evaluates all possible trees.

It calculates the likelihood for each tree and seeks the one with the maximum likelihood.

The likelihood is determined by evaluating the probability that a certain evolutionary model has generated the observed data.

The likelihood's for each site are then multiplied to provide likelihood for each tree.

ML method is the slowest and most computationally intensive method.