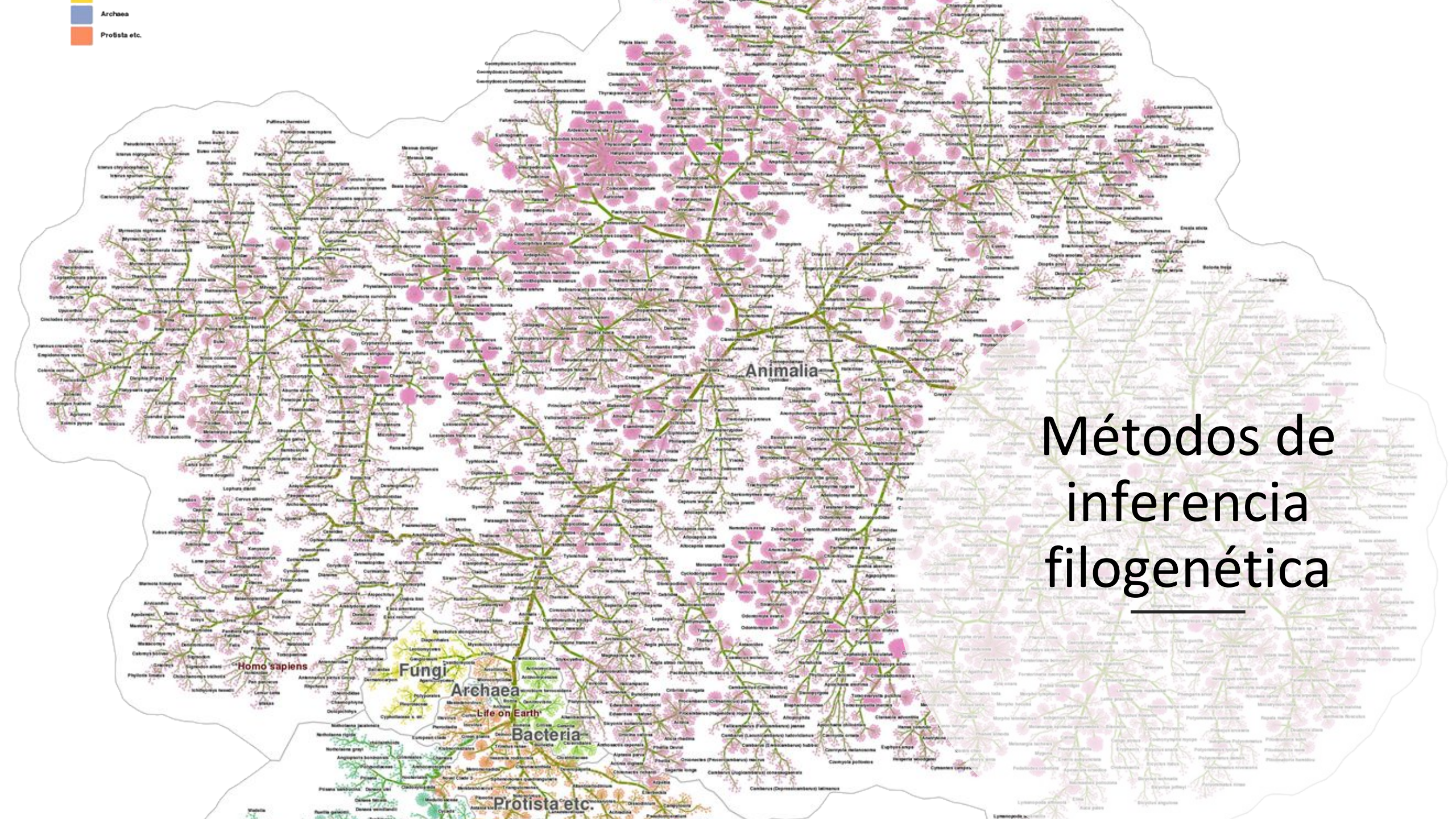
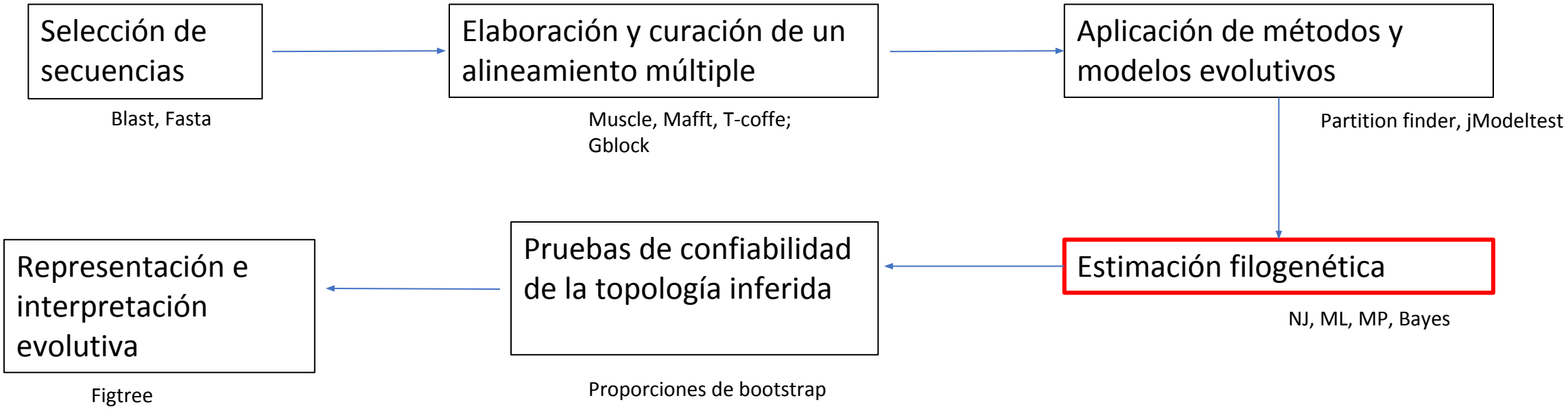
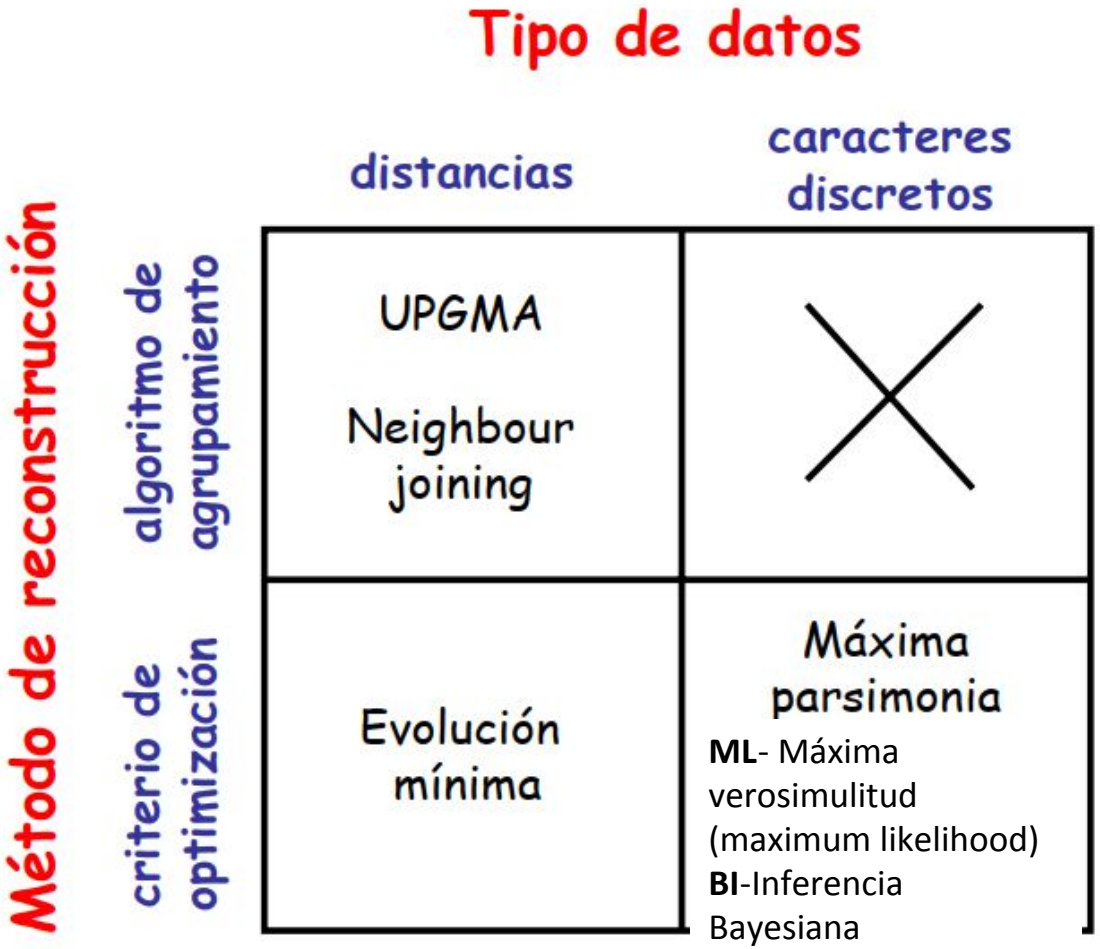


Métodos de inferencia filogenética

Cómo se construye un árbol filogenético



Clasificación de métodos



Podemos clasificar a los métodos de reconstrucción filogenética en base al tipo de datos que emplean (caracteres discretos vs. distancias) y si usan un método algorítmico o un método de búsqueda basado en un criterio de optimización para encontrar la topología óptima bajo el criterio seleccionado

Métodos de distancia

- Son métodos rápidos con algoritmos de computación simple
- Estos métodos operan en dos pasos. En un primer lugar, se calculan las distancias genéicas entre todos los pares de secuencias, y se resume esta información en una matriz de distancias. Posteriormente, se utilizan los valores de esta matriz para reconstruir un árbol filogenético. Idealmente, las distancias genéticas han de reflejar las divergencias evolutivas reales.

- Los métodos de distancia se basan en la idea de que si conociésemos las distancias evolutivas entre los taxas, podríamos reconstruir adecuadamente su historia evolutiva. Esto deriva de la relación existente entre distancias y árboles:
  - la distancia evolutiva representa una escala métrica topológica y por lo tanto define un árbol

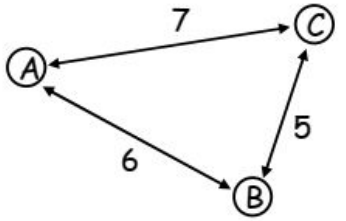
- En la práctica, empero, las distancias rara vez son métricos topológicos exactos.

Métodos basados en distancia

Para que una distancia pueda reflejar adecuadamente la filogenia debe cumplir dos requisitos: ha de ser métrica y aditiva

· **Distancias métricas** (condiciones):

1. $d(a,b) \geq 0$	(no-negatividad)
2. $d(a,b) = d(b,a)$	(simetría)
3. $d(a,c) \leq d(a,b) + d(b,c)$	(inecualidad triangular)
4. $d(a,b) = 0$ sólo si $a = b$	(distinción)

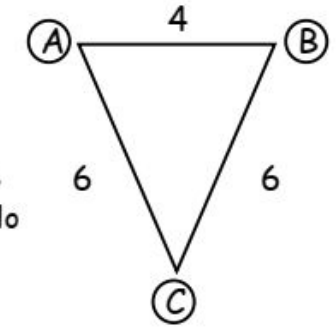


La dist. entre cualquier par de secs. no puede ser mayor que la existente entre ellas y una tercera

· **Distancias ultramétricas** (condiciones):

$d(a,b) \leq$ máximo $[d(a,c), d(b,c)]$ (distancias más largas definen un triángulo isósceles)

La inecualidad ultramétrica. Las dos distancias pareadas más largas [$d(a,c)$ y $d(b, c)$] son iguales, y por lo tanto la ultrametricidad define un triángulo isósceles



Las distancias ultramétricas tienen la virtud de **implicar igual tasa de evolución entre taxas a lo largo de toda la filogenia**

UPGMA

Unweighted pair group method with arithmetic means

- Este es uno de los pocos métodos que construye árboles ultramétricos (todas las hojas equidistantes de la raíz), es decir asume un reloj molecular perfecto a lo largo de toda la topología (evolución a tasa constante)
- Los árboles construidos por este método son llamados fenogramas, debido a que originalmente eran usados para representar el grado de semejanza de un grupo de especies en taxonomía numérica.
- A partir de una matriz de distancia van agrupando los taxones que presenten menor distancia entre si, ya que asumiendo una tasa de cambio constante serán los que han divergido más recientemente.
- Utiliza promedios aritméticos, siempre entre pares.

UPGMA

Unweighted pair group method with arithmetic means

Secuencias: A, B, C y D

matriz de distancias

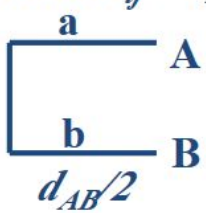
	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

d_{ij}: distancia entre secuencias i y j

```
A – GCTTGTCCGTTACGAT
B – ACTTGTCTGTTACGAT
C – ACTTGTCCGAAACGAT
D – ACTTGACCGTTTCCTT
E – AGATGACCGTTTCGAT
F – ACTACACCCTTATGAG
```

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

1ª unión *Si min* $d_{ij} = d_{AB}$



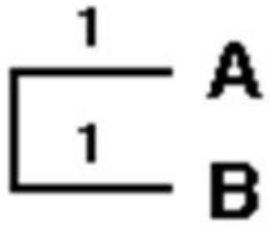
Aditividad $d_{AB} = a + b$

Ultrametricidad $a = b$

→ $d_{AB} = 2a = 2b$

$a = b = d_{AB}/2$

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8



UPGMA

Unweighted pair group method with arithmetic means

2ª unión $Si\ min\ d_{ij} = d_{(AB)C}$

	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

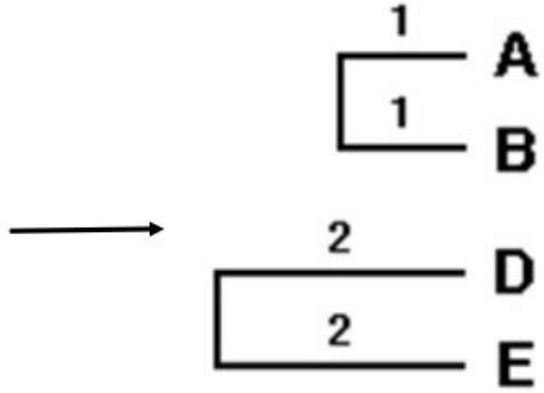
$d_{(ij)k} = (d_{ik} + d_{jk})/2$ para UPGMA y WPGMA
$d_{(ij)k} = \min(d_{ik}, d_{jk})$ para single linkage
$d_{(ij)k} = \max(d_{ik}, d_{jk})$ para complete linkage

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8

	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8



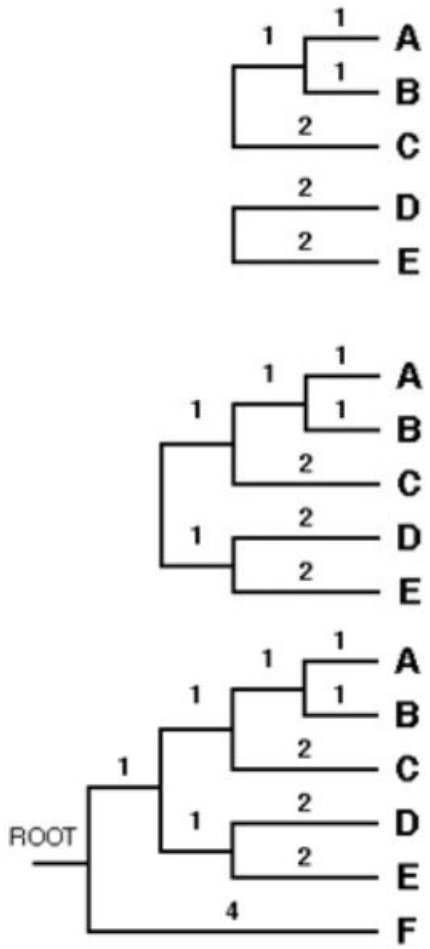
UPGMA

Unweighted pair group method with arithmetic means

	A,B	C	D,E
C	**4**		
D,E	6	6	
F	8	8	8

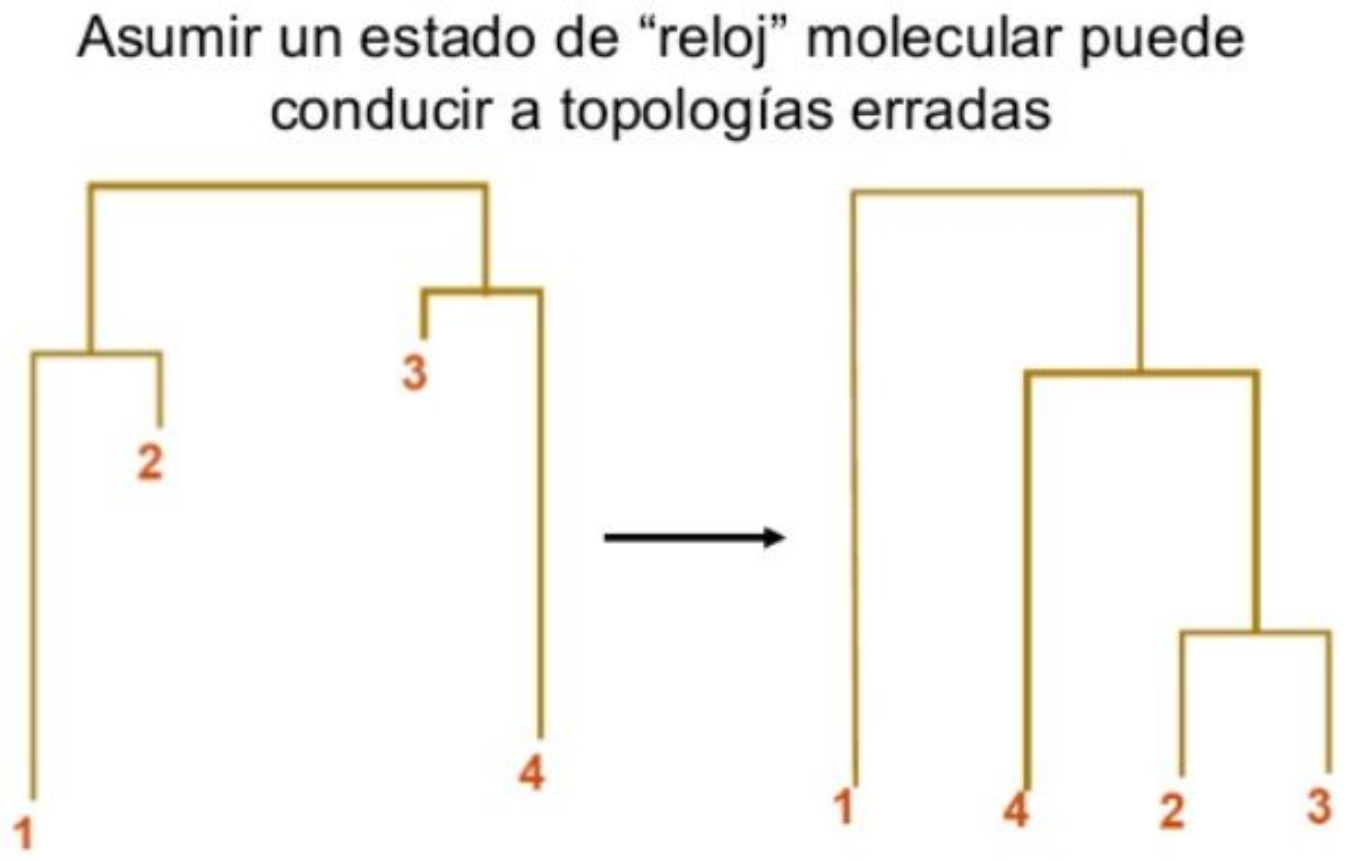
	AB,C	D,E
D,E	**6**	
F	8	8

	ABC,DE
F	**8**



UPGMA

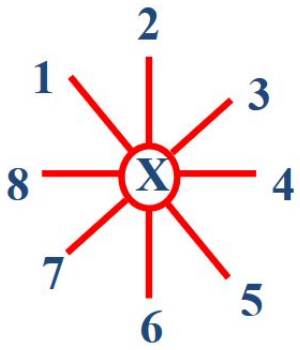
Unweighted pair group method with arithmetic means



Método neighbor-joining (**NJ**)

- Se trata de un método puramente algorítmico, representando una buena aproximación heurística para encontrar el árbol de evolución mínima más corto.
- Secuencialmente encuentra vecinos que minimizan la longitud total del árbol
- No asume ultrametricidad
- Es muy rápido y proporciona un solo árbol

Método neighbor-joining (**NJ**)



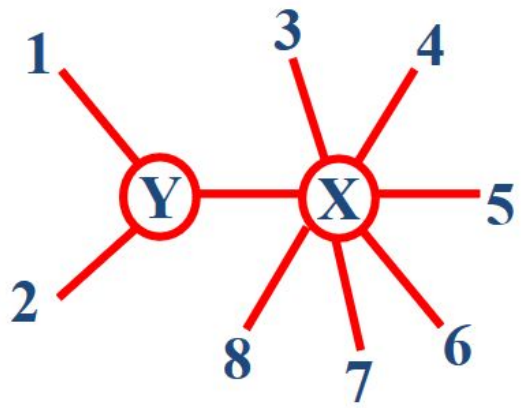
Árbol de estrellas para N OTUS

Saitou y Nei proponen que la longitud de las ramas del nuevo árbol que agrupa el par 1 y 2 (S_{12}) viene dada por:

$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^{N} (d_{1k} + d_{2k}) + \frac{1}{2} d_{12} + \frac{1}{N-2} \sum_{3 \le i \le j \le N} d_{ij}$$

S: suma de las ramas de un árbol
d_{ij}: distancia entre las secuencias i y j

Calculamos las distancias para todos los posibles árboles que agrupan todas las combinaciones posibles de pares de taxones (S_{12}, S_{13}, …, S_{78}) y Saitou y Nei demuestran que el par de taxones que minimiza la longitud del árbol es el par de vecinos más próximos.



Se recalculan las longitudes de las ramas 1 a Y (L_{1Y}) y de 2 a Y (L_{2Y}):

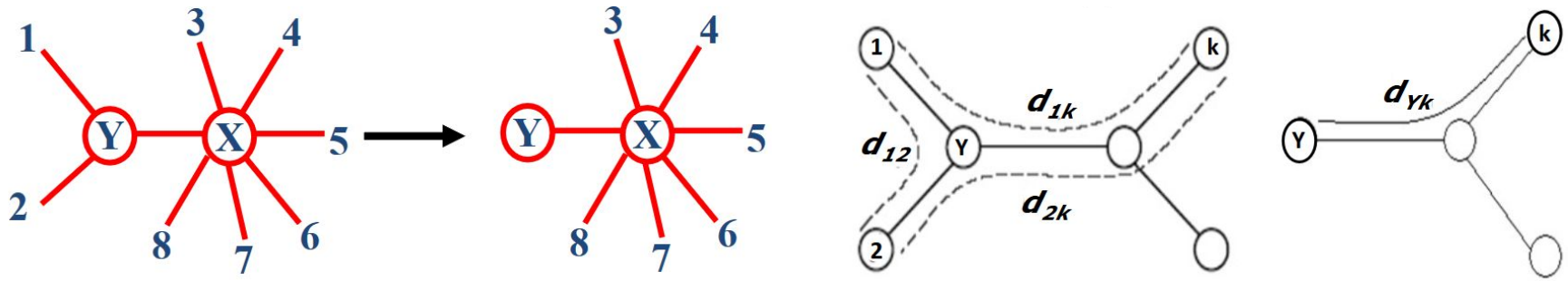
$$L_{1Y} = \frac{1}{N-2} \sum_{k=3}^{N} d_{1k} + \frac{1}{2} d_{12} - \frac{1}{N-2} \sum_{k=3}^{N} d_{2k}$$

$$L_{2Y} = \frac{1}{N-2} \sum_{k=3}^{N} d_{2k} + \frac{1}{2} d_{12} - \frac{1}{N-2} \sum_{k=3}^{N} d_{1k}$$

Método neighbor-joining (**NJ**)

Una vez identificados los vecinos más próximos, se agrupan en un nodo y se recalculan las distancias de Y al resto de los taxones. Este nuevo nodo Y se tratara como un taxon más, según la ecuación:

$$d_{Yk} = \frac{1}{2}(d_{1k} + d_{2k} - d_{12})$$



Método neighbor-joining (**NJ**)

Ejemplo tomado del artículo original de Saitou y Nei (1987)

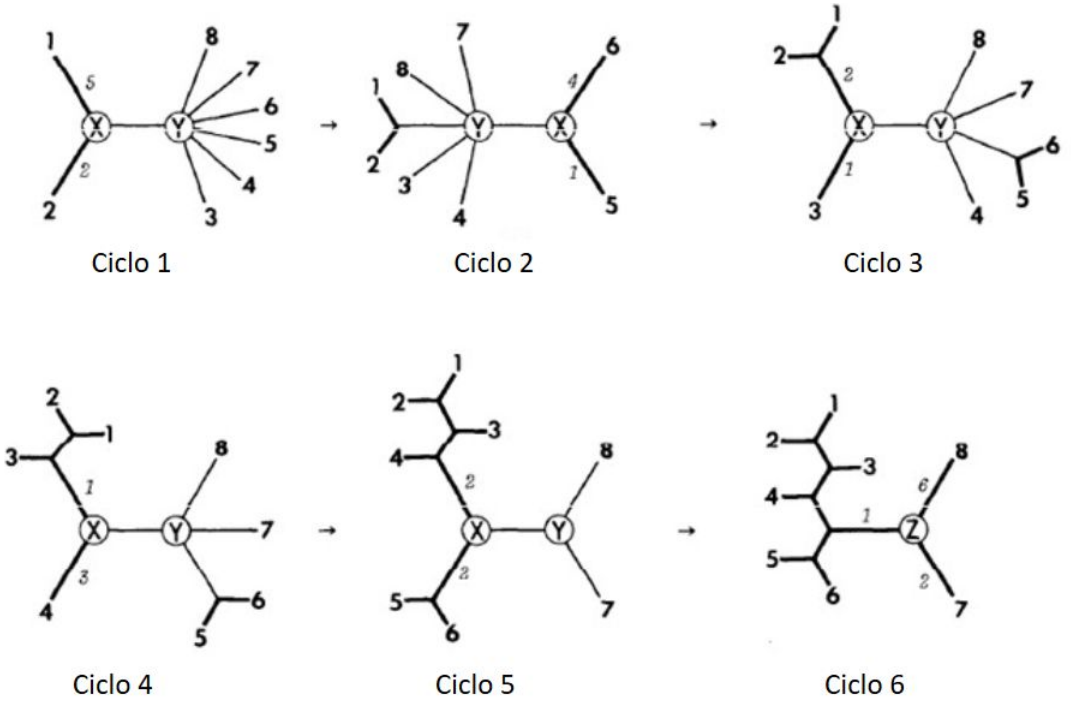
Table 2
S_{ij} Matrices for Two Cycles of the NJ Method for the Data in Table 1

A. Cycle 1: Neighbors = [1, 2]

	OTU						
OTU	1	2	3	4	5	6	7
2 ..	36.67						
3 ..	38.33	38.33					
4 ..	39.00	39.00	38.67				
5 ..	40.33	40.33	40.00	39.67			
6 ..	40.33	40.33	40.00	39.67	37.00		
7 ..	40.17	40.17	39.83	39.50	38.83	38.83	
8 ..	40.17	40.17	39.83	39.50	38.83	38.83	37.67

B. Cycle 2: Neighbors = [5, 6]

	OTU					
OTU	1-2	3	4	5	6	7
3 ..	31.50					
4 ..	32.30	32.30				
5 ..	33.90	33.90	33.70			
6 ..	33.90	33.90	33.70	31.30		
7 ..	33.70	33.70	33.50	33.10	33.10	
8 ..	33.70	33.70	33.50	33.10	33.10	31.90

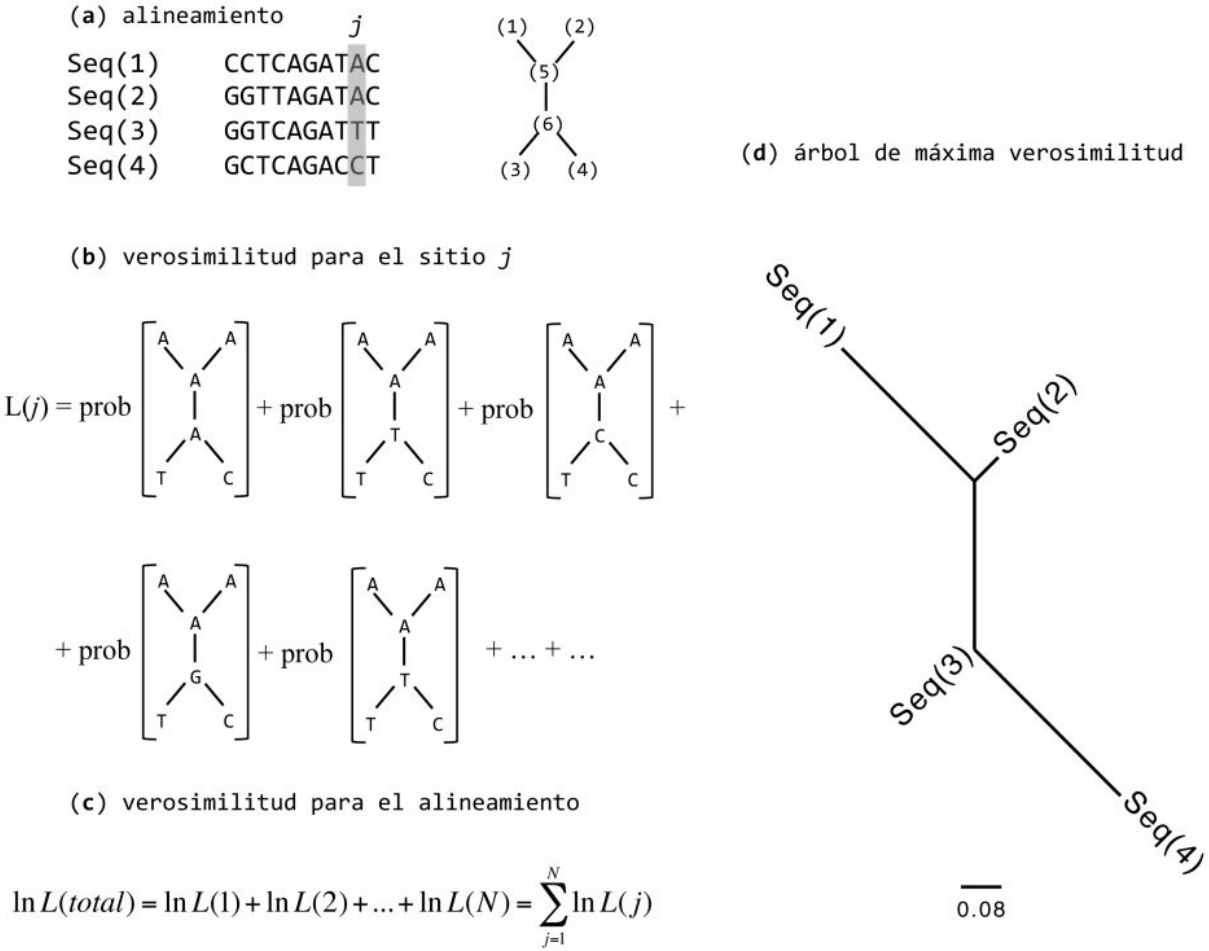


Métodos basados en carácteres

- Métodos probabilísticos que tratan la inferencia filogenética como un problema estadístico, y utilizan modelos explícitos de evolución molecular para el cálculo de probabilidades.
- A diferencia de los métodos basados en distancias, los basados en caracteres tienen en cuenta los cambios que ocurren en cada posición del alineamiento, y hacen por tanto un uso más eficiente de la información
- De forma general, aunque no garantizan encontrar el mejor árbol globalmente, permiten evaluar un número significativo de árboles más probables.
- Se utilizan algoritmos que partiendo de un árbol (por ejemplo, aleatorio o reconstruido con NJ) realizan pequeñas reorganizaciones locales para encontrar árboles más probables. Durante las búsquedas heurísticas, las distintas reorganizaciones se aceptan o rechazan con ayuda del criterio de optimización, y la búsqueda continúa hasta que no se encuentren más mejoras significativas
- Los métodos probabilísticos son los que mejor aprovechan la información filogenética contenida en las secuencias, los más avanzados y generalmente los más fiables.

ML (Máxima verosimilitud)

- ML utiliza un **criterio estadístico y computacionalmente intensivo** para evaluar una hipótesis evolutiva:
  - Toma un alineamiento múltiple (observación)
  - Formula todos los árboles posibles para cada columna (*partición*) del alineamiento
  - Calcula la probabilidad de todas las topologías posibles, basándose en un modelo evolutivo seleccionado por el usuario
  - Combina la información para cada partición
  - Identifica el árbol con la mayor probabilidad general, cómo la filogenia más probable



$$\ln L(total) = \ln L(1) + \ln L(2) + \ldots + \ln L(N) = \sum_{j=1}^{N} \ln L(j)$$

ML (Máxima verosimilitud)

- Este Método nos permiten visitar "un cierto número de árboles entre todos los posibles, y valorar si un árbol es mejor o peor en función de un criterio de optimización, en este caso su verosimilitud.
- El árbol de ML será aquél que tiene el mayor valor de verosimilitud (menor ln *L*) entre todos los árboles evaluados durante la búsqueda heurística.
- Esperamos que la búsqueda heurística nos permita encontrar el mejor árbol entre todos los posibles (alcanzar el máximo global).
- Sin embargo, nos asegurara que hemos encontrado el mejor árbol, pues durante el proceso de optimización puede quedar atrapada en un máximo local, lo cual nos generara un árbol subóptimo.

Inferencia Bayesiana

- Al igual que ML, es **puramente estadístico y computacionalmente intensivo**
- Se basa en la regla de Bayes: Puede introducir supuestos sobre la probabilidad inicial ("conocimiento previo")
- El método bayesiano utiliza parámetros aleatorios en el modelo aplicado al árbol, mientras que ML utiliza constantes fijadas y de valor desconocido
- Utilizando MCMC (*Markov Chain Monte Carlo*), se genera una *muestra* de árboles que representan la distribución de las probabilidades posteriores. En cuanto más grande la muestra, más confiable el resultado.

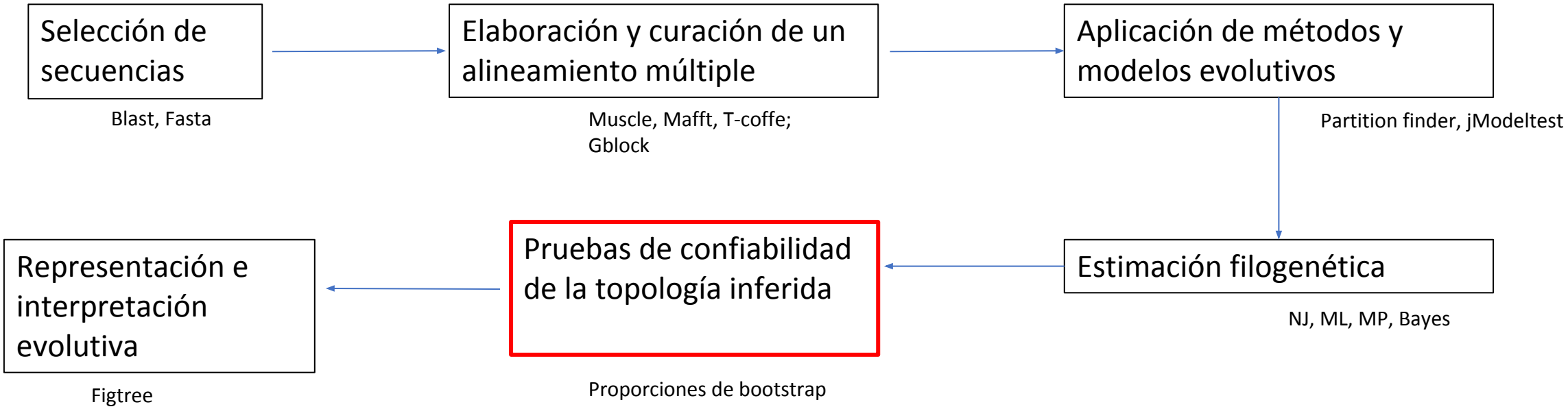
Table 1 | **Functionalities of a few commonly used phylogenetic programs**

Name	Brief description	Link	Refs
Bayesian evolutionary analysis sampling trees (BEAST)	A Bayesian MCMC program for inferring rooted trees under the clock or relaxed-clock models. It can be used to analyse nucleotide and amino acid sequences, as well as morphological data. A suite of programs, such as Tracer and FigTree, are also provided to diagnose, summarize and visualize results	http://beast.bio.ed.ac.uk	135
Genetic algorithm for rapid likelihood inference (GARLI)	A program that uses genetic algorithms to search for maximum likelihood trees. It includes the GTR + Γ model and special cases and can analyse nucleotide, amino acid and codon sequences. A parallel version is also available	http://code.google.com/p/garli	55
Hypothesis testing using phylogenies (HYPHY)	A maximum likelihood program for fitting models of molecular evolution. It implements a high-level language that the user can use to specify models and to set up likelihood ratio tests	http://www.hyphy.org	136
Molecular evolutionary genetic analysis (MEGA)	A Windows-based program with a full graphical user interface that can be run under Mac OSX or Linux using Windows emulators. It includes distance, parsimony and likelihood methods of phylogeny reconstruction, although its strength lies in the distance methods. It incorporates the alignment program ClustalW and can retrieve data from GenBank	http://www.megasoftware.net	37
MrBayes	A Bayesian MCMC program for phylogenetic inference. It includes all of the models of nucleotide, amino acid and codon substitution developed for likelihood analysis	http://mrbayes.net	71
Phylogenetic analysis by maximum likelihood (PAML)	A collection of programs for estimating parameters and testing hypotheses using likelihood. It is mostly used for tests of positive selection, ancestral reconstruction and molecular clock dating. It is not appropriate for tree searches	http://abacus.gene.ucl.ac.uk/software	137
Phylogenetic analysis using parsimony* and other methods (PAUP* 4.0)	PAUP* 4.0 is still a beta version (at the time of writing). It implements parsimony, distance and likelihood methods of phylogeny reconstruction	http://www.sinauer.com/detail.php?id=8060	
PHYLIP	A package of programs for phylogenetic inference by distance, parsimony and likelihood methods	http://evolution.gs.washington.edu/phylip.html	
PhyML	A fast program for searching for the maximum likelihood trees using nucleotide or protein sequence data	http://www.atgc-montpellier.fr/phyml/binaries.php	53
RAxML	A fast program for searching for the maximum likelihood trees under the GTR model using nucleotide or amino acid sequences. The parallel versions are particularly powerful	http://scoh-its.org/exelixis/software.html	54
Tree analysis using new technology (TNT)	A fast parsimony program intended for very large data sets	http://www.zmuc.dk/public/phylogeny/TNT	42

Note: all programs can run on Windows, Mac OSX and Unix or Linux platforms. Except for PAUP*, which charges a nominal fee, all packages are free for download. See Felsenstein's comprehensive list of programs at http://evolution.genetics.washington.edu/phylip/software.html. GTR, general time reversible; MCMC, Markov chain Monte Carlo.

Yang & Rannala, 2012

Cómo se construye un árbol filogenético



Robustez de un árbol filogenético y contraste de árboles alternativos

- Uno de los aspectos más importantes del análisis filogenético es el de poder determinar cuán fiable es el árbol que hemos obtenido.
- La inferencia filogenética mediante métodos de ML produce un único árbol, el árbol que con mayor verosimilitud ha dado lugar a los datos observados. Por esta razón, no conocemos de manera directa si existe otro árbol con una verosimilitud similar no significativamente diferente desde el punto de vista estadístico.
  La aproximación más corriente es utilizar el bootstrapping no paramétrico, que nos informa acerca de la estabilidad de las relaciones filogenéticas del árbol de ML obtenido (aunque no puede informar de la presencia de otros árboles subóptimos).

Bootstrapping