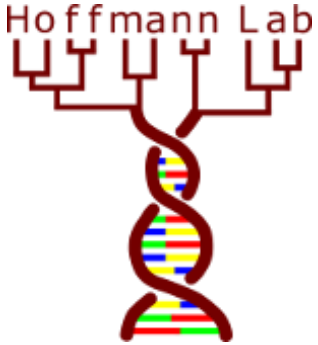
Evolución de familias multigénicas 2020

Introducción al análisis filogenético



Filogenias e inferencia sobre filogenias

Observaciones:

- Tomamos el alineamiento como paso previo a la filogenia.
- Pero el alineamiento busca establecer homología de sitios entre secuencias, y la homología es un concepto filogenético (origen en un ancestro común).

Dos aspectos principales del análisis filogenético:

- Inferencia de las relaciones filogenéticas. Descripción de un árbol filogenético:
  - Topología (sin raíz, con raíz).
  - Tiempos de divergencia (absolutos, relativos).
  - Longitudes de las ramas (cantidad de cambio).
- Inferencias a partir del árbol filogenético:
  - Tiempos de divergencia (absolutos, relativos).
  - Longitudes de las ramas (cantidad de cambio).
  - Estimación patrones de cambios de caracteres sobre el árbol:
    - Caracteres usados para la inferencia filogenética.
    - Otros caracteres.

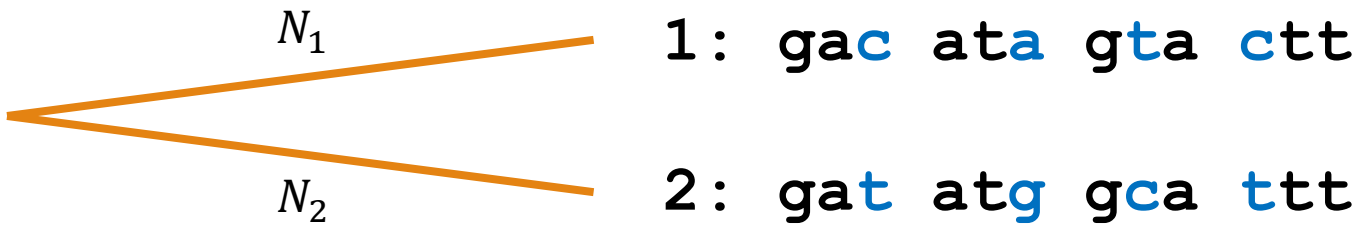
Algunos métodos de inferencia filogenética

Método	Variantes	Criterio de optimización	Uso de variación no observada
Parsimonia	máxima parsimonia	minimizar el número de pasos requeridos para obtener los datos	No
Distancias	- evolución mínima	minimizar los cambios requeridos para obtener las distancias estimadas entre taxa	Sí (incorporadas en las distancias)
	- unión de vecinos (neighbor joining)	una aproximación al árbol de evolución mínima	ídem
Inferencia estadística	Máxima verosimilitud	maximizar la verosimilitud de observar los datos, dado un árbol y un modelo de evolución molecular.	Sí (considerando todos los estados posibles en los nodos).
	Inferencia bayesiana		ídem

Algunos métodos de inferencia filogenética

Método	Variantes	Criterio de optimización	¿Optimización requiere muestreo de árboles?
Parsimonia	máxima parsimonia	minimizar el número de pasos requeridos para obtener los datos	Sí
Distancias	- evolución mínima	minimizar los cambios requeridos para obtener las distancias estimadas entre taxa	
	- unión de vecinos (neighbor joining)	una aproximación al árbol de evolución mínima	No
Inferencia estadística	Máxima verosimilitud	maximizar la verosimilitud de observar los datos, dado un árbol y un modelo de evolución molecular.	Sí
	Inferencia bayesiana		Sí

Cambios nucleotídicos observados y estimados

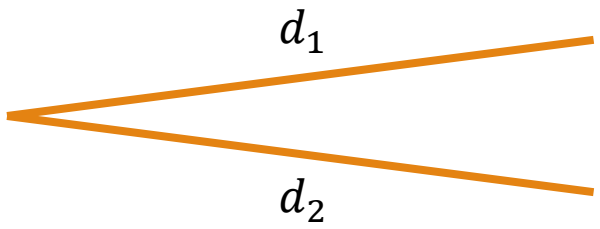


Diferencias observadas: $N_o = 4$

Cambios ocurridos en la evolución: N= $N_1 + N_2 \geq N_o$

Necesitamos de un modelo de evolución para estimar *N* a partir de N_o

Distancias observadas y estimadas



1: gac ata gta ctt

2: gat atg gca ttt

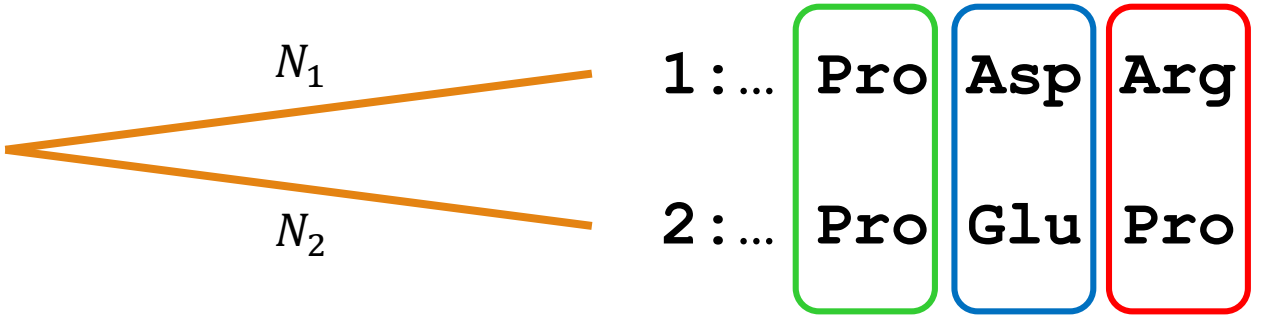
Diferencias observadas: $d_o = {}^{4}/_{12}$

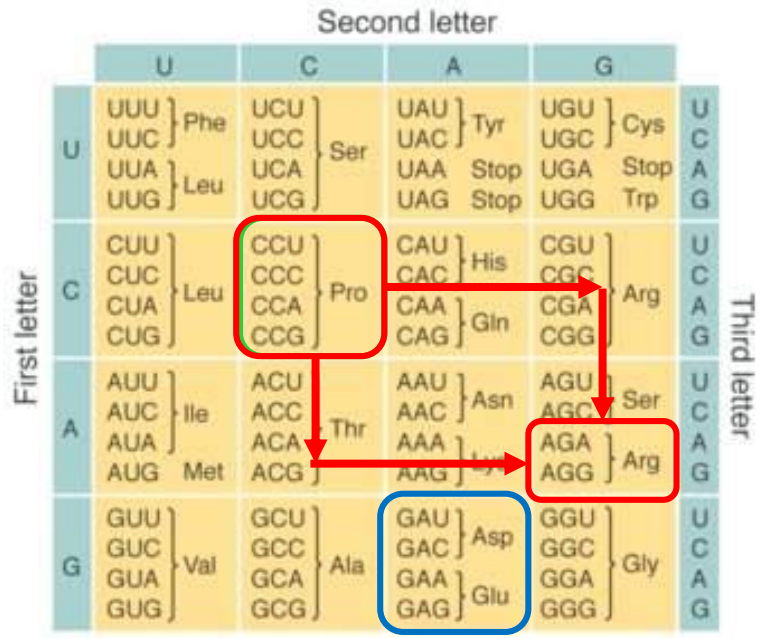
Cambios ocurridos en la evolución: d= $d_1 + d_2 \geq d_o$

Necesitamos de un modelo de evolución para estimar *d* a partir de d_o

Distancia estimada $d_e \geq d_o$

Cambios aminoacídicos observados y estimados





Cambios ocurridos en la evolución: N= $N_1 + N_2 \geq N_o$

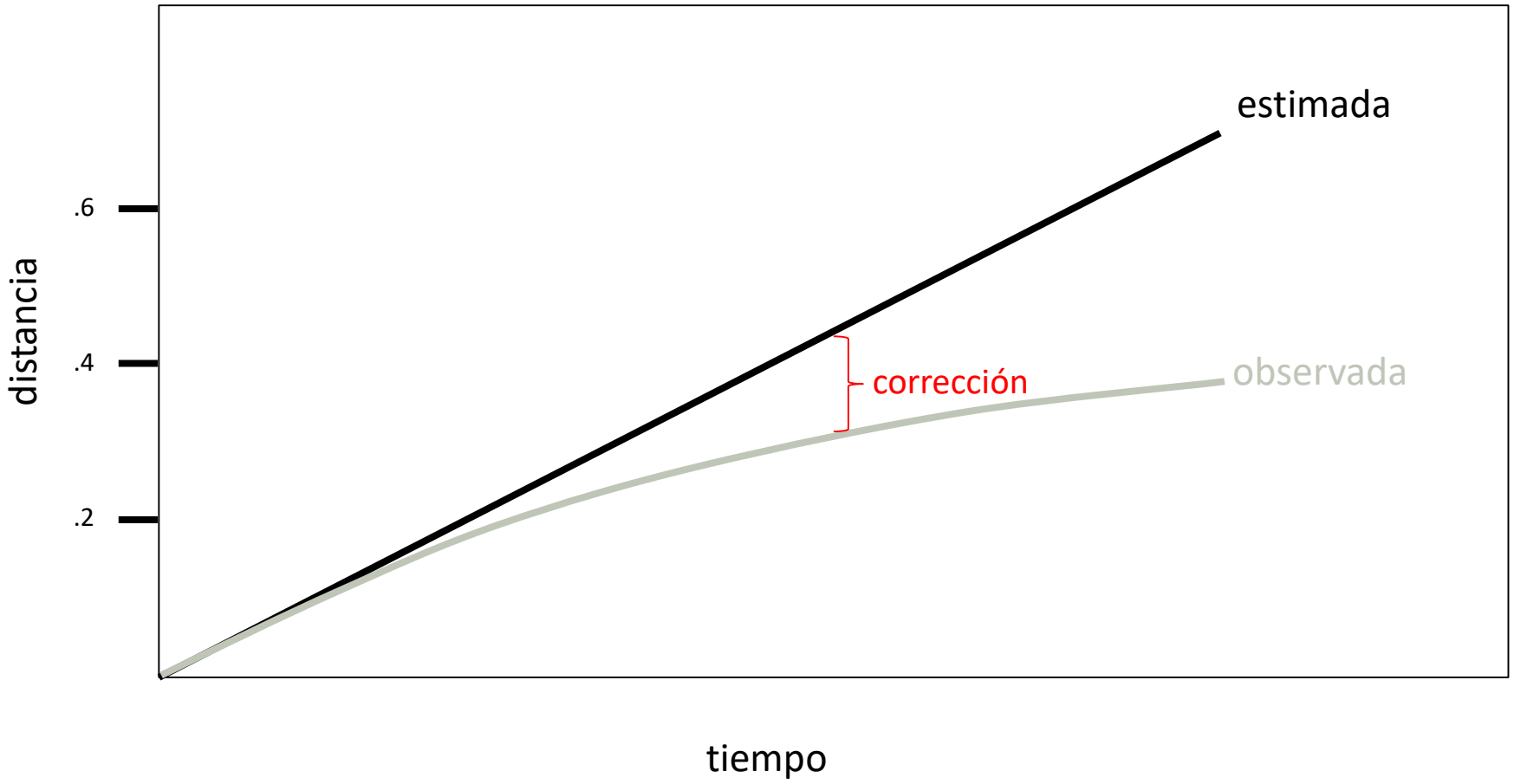
Necesitamos de un modelo de evolución para estimar *N* a partir de N_o

Aún sin un modelo, sabemos que algunos sitios implicaron más de un cambio.

Optimización: parsimonia vs. distancias

	Parsmionia	Distancias
Criterio de optimización	Minimizar la longitud (número de pasos) del árbol	Minimizar la longitud (suma de todas las ramas, medidas como distancias) del árbol
Efecto de la homoplasia	Los mejores árboles requieren más pasos que el mínimo ideal	1) Las distancias estimadas son mayores a las observadas. 2) El árbol óptimo requiere un largo total mayor al mínimo requerido por las distancias.

¿Por qué estimar distancias?



Principales clases de modelos de evolución molecular

Modelos de sitios (nucleótido, aminoácido):

- Cada sitio (o clase de sitio si se agrupan) evoluciona de acuerdo al modelo seleccionado.
- Para nucleótidos, esos modelos pueden involucrar:
  - Una matriz de transición de 4 x 4
  - Frecuencias nuceotídicas.
  - Otros parámetros (por ejemplo, modulando la matriz de transición para clases de sitios)
- Para aminoácidos, de manera análoga:
  - Una matriz de 20 x 20
  - Frecuencias aminoacídicas
  - Otros parámetros….

Modelos de codones:

- De manera análoga… (matriz de 64 x 64, de la que suele excluirsen los codones Stop), etc.

Hay modelos más complejos, que incluyen información estructural, por ejemplo.

Relación entre distancias y modelos de evolución molecular

Para calcular la distancia estimada (corregida con una estimación de los cambios no observados), necesitamos la distancia observada y un modelo de evolución molecular.

Ejemplos: modelo de Jukes y Cantor:

	G	A	T	C
G	1-3μ	0	μ	μ
A	μ	1-3 μ	μ	μ
T	μ	μ	1-3 μ	μ
C	μ	μ	μ	1-3 μ

La distancia correspondiente es $d = \frac{3}{4} \ln(1 - 4/3\ p)$, siendo p la distancia observada (proporción de sitios diferentes).

Relación entre distancias y modelos de evolución molecular

El modelo de Jukes y Cantor tiene un solo parámetro (μ).

	G	A	T	C
G	1-3μ	μ	μ	μ
A	μ	1-3 μ	μ	μ
T	μ	μ	1-3 μ	μ
C	μ	μ	μ	1-3 μ

Un modelo popular de dos parámetros es el de Kimura, que distingue transiciones y transversiones:

	G	A	T	C
G	1-α -2β	α	β	β
A	α	1-α-2β	β	β
T	β	β	1-α 2β	α
C	β	β	α	1-α-2β

Filogenias usando distancias

	Caracteres									
	1	2	3	4	5	6	7	8	9	10
Especie A	c	a	a	g	t	c	c	g	t	a
Especie B	.	.	t	.	.	t	.	a	.	.
Especie C	.	.	t	.	.	.	t	a	.	.
Especie D	t	g	.	.	c	.	.	.	.	g
Especie E	t	g	.	a	c	.	.	t	.	.

Distancia asociada a un modelo

Distancias	A	B	C	D	E
A		0,3	0,3	0,4	0,5
B			0,2	0,6	0,7
C				0,7	0,7
D					0,3
E					

algoritmo

Árbol

Método de unión de vecinos (“neighbor joining”)

NJ es una aproximación simple al criterio de evolución mínima:

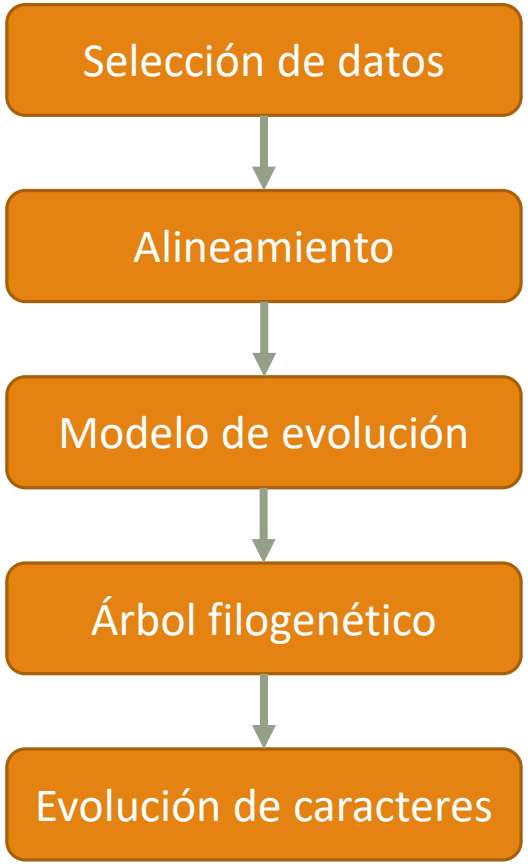
El árbol óptimo es aquel que requiere la menor longitud total (suma de todas las ramas, medidas como distancias moleculares).

El algoritmo de NJ es sumamente eficiente, aunque es criticado como un algoritmo arbitrario.

Al optar por optimizar sobre distancias,

- Resumimos la información sobre las OTUs (caracteres y sus estados) en distancias entre pares de OTUs. (A diferencia de la parsimonia, que trabaja con los caracteres originales)
- Las distancias pueden ser las observadas (número o fracción de sitios diferentes) o, más comúnmente, distancias corregidas en base a un modelo de evolución para incorporar cambios no observados. (A diferencia de la parsimonia, que no usa caracteres no observados, y solamente incorpora cambios adicionales cuando así lo requieren los árboles más parsimoniosos).

Un esquema más general



Como primera aproximación, podemos considerar que estos son pasos ordenados linealmente.

Pero, como vimos en el taller de selección de datos y alineamiento, es un proceso de iteración entre fases más complejo (“de ida y vuelta”).

Table 1 | **Features of different orthology prediction and sequence alignment programs**

Software	Method	Key features	Link	Ref.
Orthology prediction				
OMA standalone	Graph based	Infers OMA groups (that is, sets of genes for which all pairs are orthologous) and hierarchical orthologous groups; predicts gene function annotations, phylogenetic profiling (patterns of gene presence or absence across species); has been used to infer orthologous groups across the tree of life (https://omabrowser.org)	https://omabrowser.org/standalone/	27
OrthoDB pipeline	Graph based	Infers hierarchical orthologous groups; has been used to estimate orthologous groups across the tree of life (https://www.orthodb.org/) as well as near-universal single-copy orthologues (that is, BUSCO genes) for different clades (https://busco.ezlab.org/); the BUSCO genes completeness is used as a quality metric for genome and transcriptome samples	https://www.orthodb.org/?page=software	29
OrthoFinder	Graph based	Corrects a previously undetected bias related to BLAST hits and gene lengths; offers options for different BLAST algorithms, alignments, building gene and species trees and comparative genomic statistics; easy to add or subtract species	https://github.com/davidemms/OrthoFinder	26
OrthoMCL	Graph based	Uses BLAST and Markov clustering to group proteins of multiple species into groups of putative orthologues	https://orthomcl.org/orthomcl/	25
InParanoid/HieranoiDB	Graph based	Infers hierarchical orthologous groups for multiple genomes by combining pairwise orthology analysis by InParanoid and a guide tree	http://hieranoidb.sbc.su.se/	28
PhylomeDB	Tree based	Infers phylogenies of homologous genes using a two-step optimization (neighbour joining and maximum likelihood); each node of the phylogeny is identified as a duplication or speciation event based on the overlap of species	http://phylomedb.org/	20

Kalpi et al. 2020

Alignment inference				
Clustal	Progressive	Infers alignment of DNA or protein sequences and can take into account protein structural information; provides a graphical interface	http://www.clustal.org/	53
Muscle	Progressive	Infers alignment of DNA or protein sequences; similar to (but faster than) Clustal	http://www.drive5.com/muscle	52
ProbCons	Consistency	Infers alignment of amino acid sequences using a combination of probabilistic modelling and consistency-based techniques; provides alignment quality scores per site	http://probcons.stanford.edu/	56
MAFFT	Progressive and consistency	Infers alignment of DNA or protein sequences; implements several different algorithms for accommodating low sequence similarity as well as long internal and terminal branches	https://mafft.cbrc.jp/alignment/software/	54
PRANK	Evolution based	Infers alignment of DNA, codon or amino acids in a maximum likelihood framework; it is based on a guide tree (either provided by the user or inferred using neighbour joining) and it explicitly models insertions and deletions; it breaks ties randomly and therefore the result may differ across runs	https://code.google.com/archive/p/prank-msa/	62
Bali-Phy	Evolution based	Jointly estimates the alignment (DNA, codon or amino acid sequences) and the phylogeny in a Bayesian framework; assumes an explicit model for insertions and deletions; performs ancestral sequence reconstruction	http://www.bali-phy.org/	58
StatAlign	Evolution based	Jointly estimates the alignment (DNA or amino acid sequences) and the phylogeny in a Bayesian framework; assumes an explicit model for insertions and deletions; offers a graphical interface and can take into account protein structure information	https://statalign.github.io/	59

BLAST, Basic Local Alignment Search Tool; BUSCO, Benchmarking Universal Single-Copy Orthologues; OMA, Orthologous MAtrix.

Kalpi et al. 2020

Algunas referencias

Kalpi et al. 2020. Phylogenomic tree building in the genomic age. Nature Reviews Genetics: : https://doi-org.libproxy.unm.edu/10.1038/s41576-020-0233-0.

Nascimento, F. B., dos Reis, M., & Yang, Z. 2017. A biologist's guide to Bayesian phylogenetic analysis. Nature Ecology and Evolution: DOI: 10.1038/s41559-017-0280-x