Filogenias Moleculares
Introducción a la reconstrucción filogenética

Mariana G. López
Instituto de Biomedicina de Valencia (IBV-CSIC)

20 Junio 2019

Unidad de Bioinformática y Bioestadística

PRINCIPE FELIPE CENTRO DE INVESTIGACION

CSIC Consejo Superior de Investigaciones Científicas

WODA
WEB-BASED OMICS DATA ANALYSIS

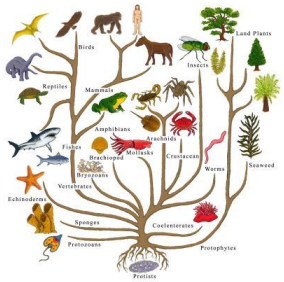
Introducción

✗ El término filogenia fue propuesto en 1866 por el embriólogo alemán Ernst Haeckel con la intención de recoger los conceptos evolucionistas de Darwin y Wallace

✗ Filogenia = árbol evolutivo
es la reconstrucción de la historia evolutiva de distintos grupos de organismos, es una hipótesis evolutiva

✗ Las relaciones evolutivas entre grupos no pueden observarse directamente, de modo que se infieren a partir de los datos de los que disponemos (morfológicos, moleculares, etc); es así que la filogenética se basa en aplicar métodos estadísticos basados en modelos de evolución para reconstruir de forma óptima una filogenia

✗ Una filogenia es un modelo de historia genealógica que no solo se aplica al estudio de grupos taxonómicos



Introducción

En sus orígenes (era pre-secuenciación), los árboles filogenéticos se utilizaban exclusivamente en taxonomía y sistemática

Actualmente las filogenias moleculares se utilizan para:

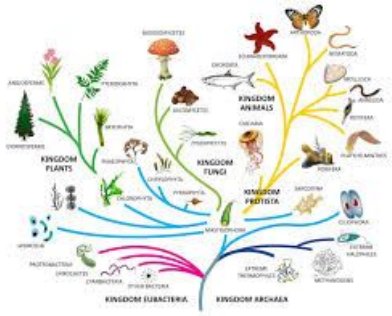
* relaciones entre grupos (árbol de la vida)
* relación entre parálogos de una familia génica (árboles de genes)
* historia de poblaciones (genética de poblaciones/coalescencia/filogeografía)
* dinámica de patógenos (evolución y epidemiología)
* diferenciación y desarrollo de células somáticas (cáncer)
* relaciones entre proteínas (inferir cambios en los patrones de selección)
* comparación de genomas
    - metagenómica
    - identificación de genes, elementos regulatorios, ncRNA, etc
    - reconstrucción de genomas antiguos

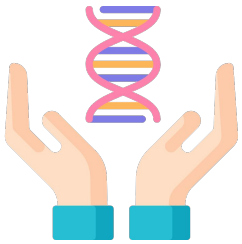
Introducción

- Tipos de filogenias según los caracteres/datos utilizados

  - **morfológicas** (antiguas)

  - **moleculares***
    - basadas en marcadores (RAPDs, RFLP, AFLPs, microsat)
    - secuencias
      - genomas completos (nucleares, organelas)
      - genes específicos
      - ARN
      - proteínas





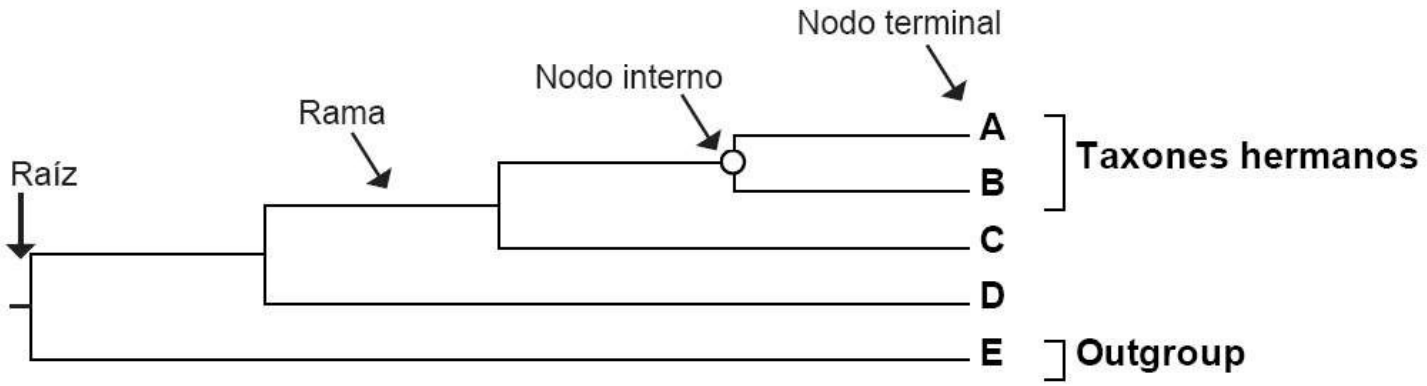
Introducción

Las filogenias se representan con un árbol (cladograma)

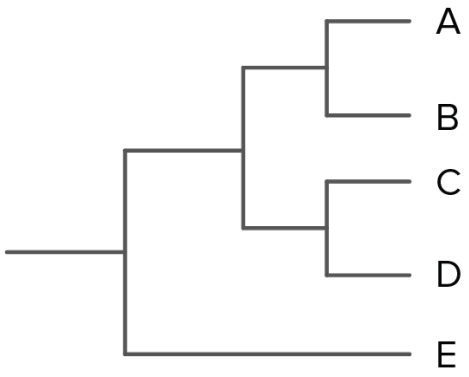
1. los nodos se conectan por ramas

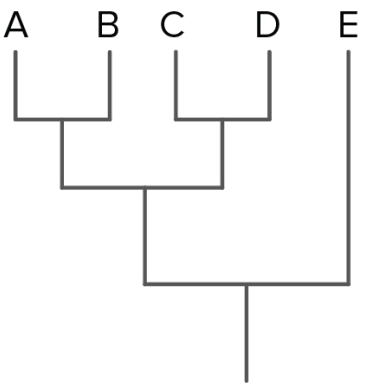
2. las longitud de ramas representan la evolución de un linaje, se expresan como tasa de sustitución x sitio (longitud desconocida)

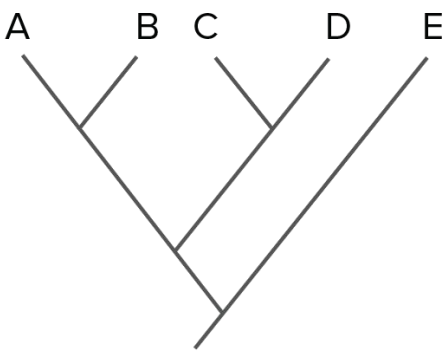
3. los nodos representan el ancestro común de dos ramas o el nacimiento de un nuevo linaje

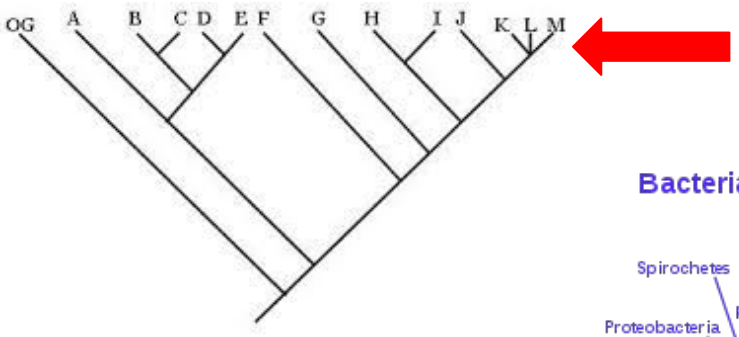


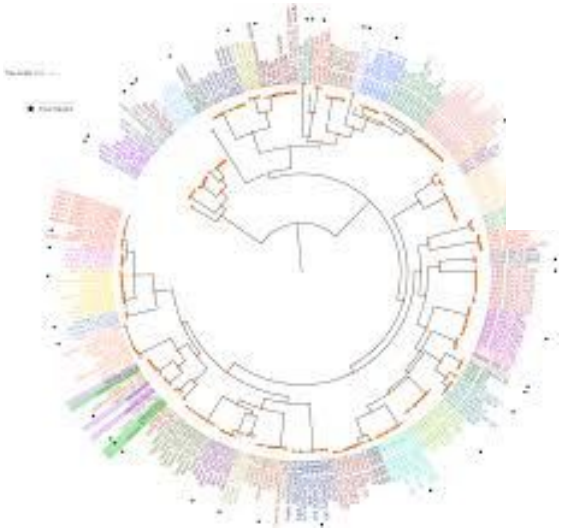
Introducción

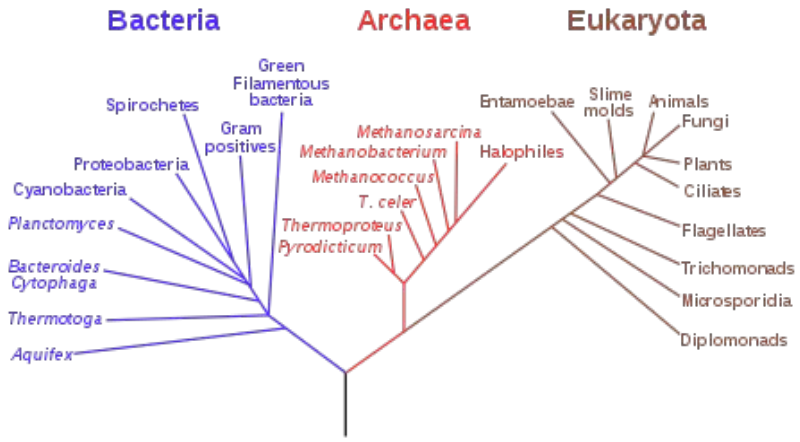








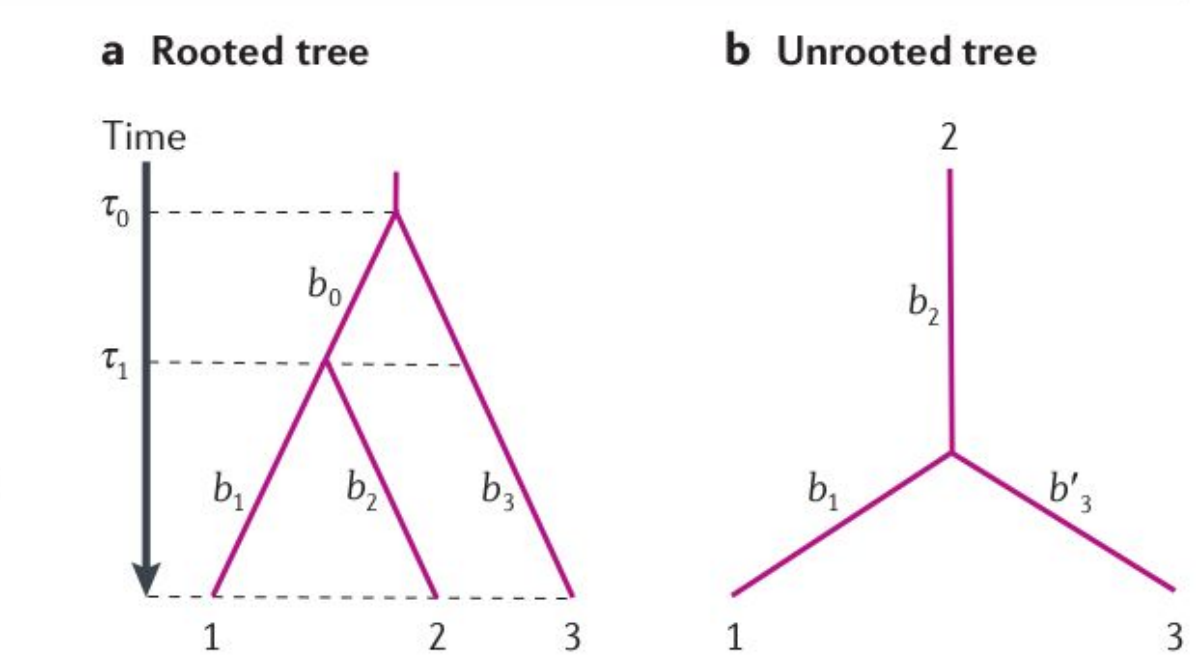




Introducción
Conceptos básicos

Enraizar un árbol

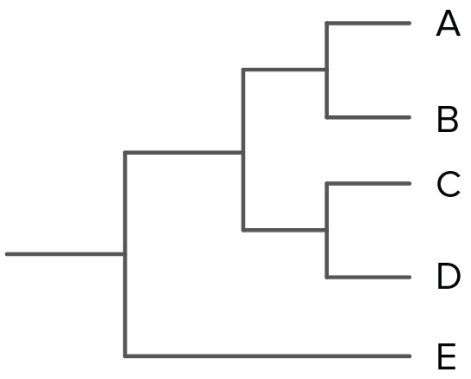
- definir el ancestro de la filogenia, dar una polaridad a la filogenia, puede hacerse incluyendo en el análisis un grupo relacionado pero distante al grupo estudiado
- árboles enraizados vs árboles no enraizados

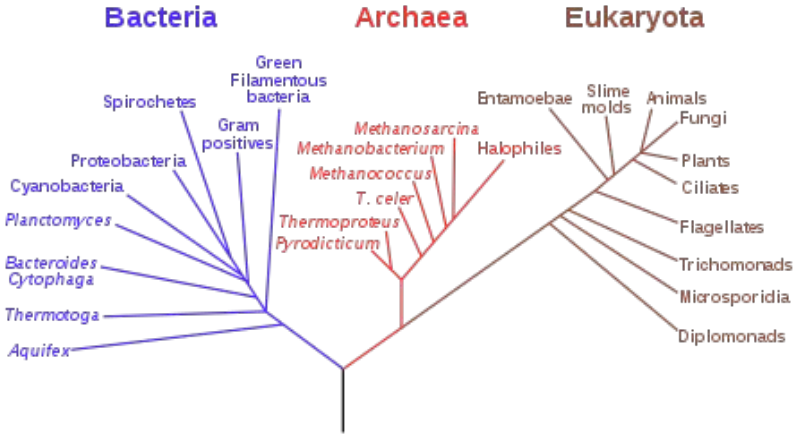


Introducción
Conceptos básicos

Reloj molecular

- se asume cuando la tasa de sustitución es constante en el tiempo o entre linajes (no varía dentro de las distintas rama del árbol); caso contrario cuando la tasa de sustitución es variable entre ramas, no puede asumirse reloj molecular
- árboles ultramétricos vs árboles de tasa evolutiva variable



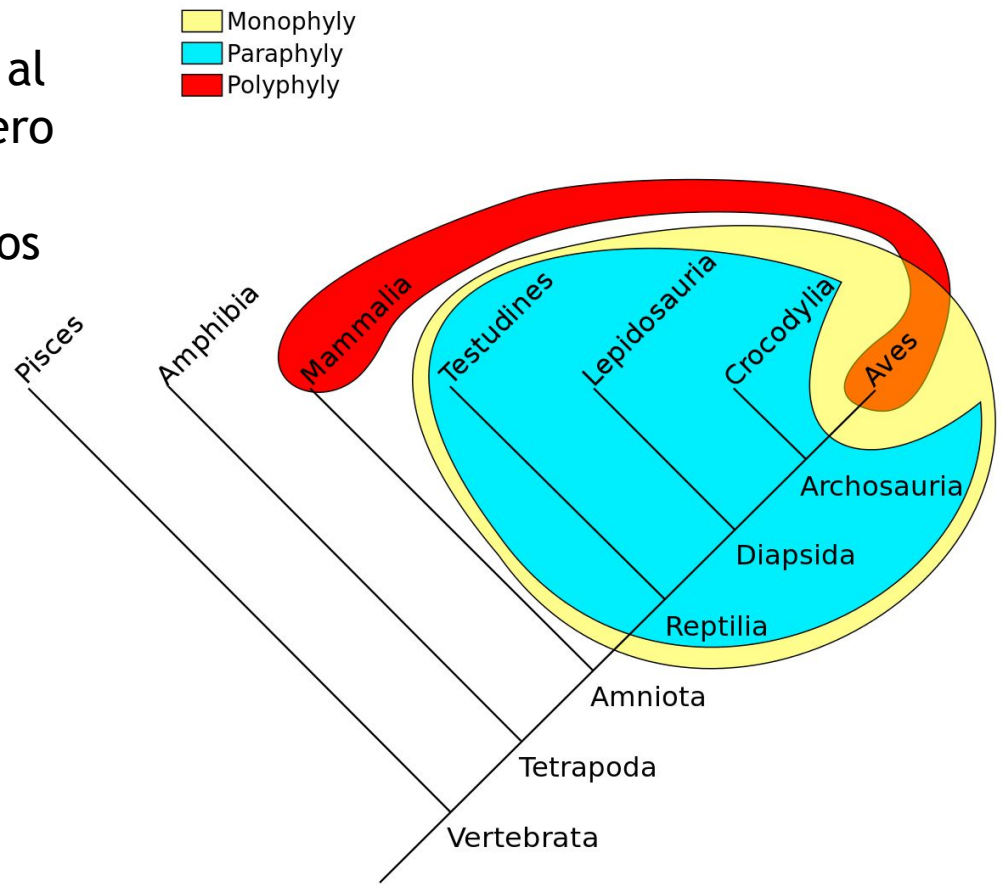


Introducción
Conceptos básicos

Tipos de grupos (árboles de especies)

- *Monofilético*: grupo que incluye al ancestro común y a todos sus descendientes (grupo natural)
- *Parafilético*: grupo que incluye al ancestro común de un grupo pero NO a todos sus descendientes
- *Polifilético*: grupos con distintos ancestros comunes

Tipos de grupos (árboles de especies)

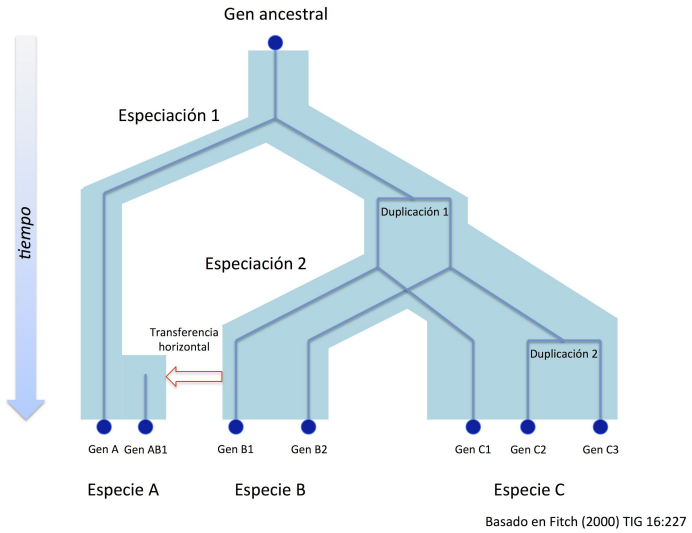


Introducción
Conceptos básicos

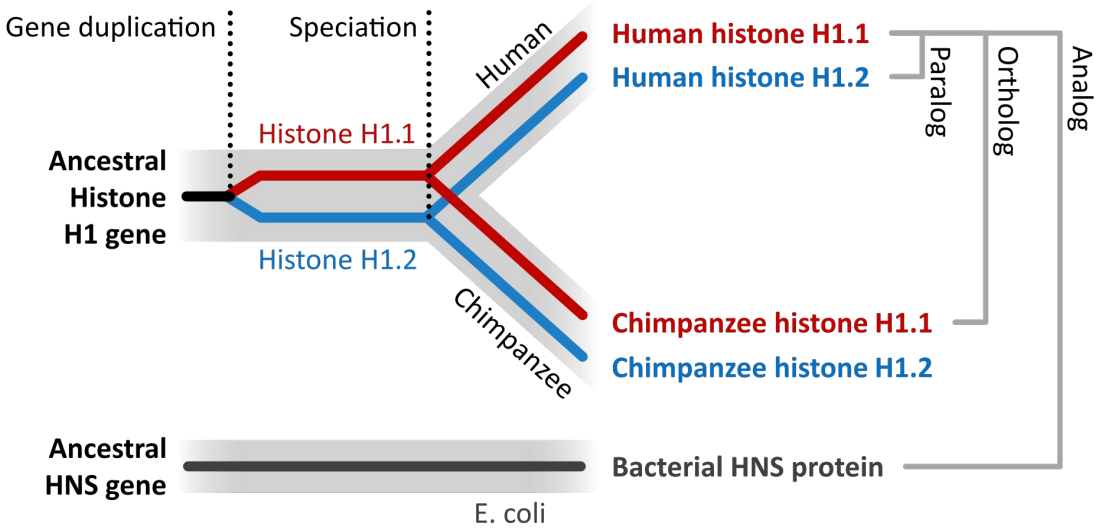
Tipos de "caracteres" (árboles de genes)

- *Homólogos*: caracteres similares (o genes) cuya similitud se debe a ancestralidad común (historia evolutiva)
  - ortólogos: se originan por un evento de especiación
  - parálogos: se originan por un evento de duplicación génica
  - xenólogos: se originan por transferencia horizontal

- *Análogos*: caracteres (o genes) cuya similitud es producto de convergencia evolutiva (homoplasias)



Basado en Fitch (2000) TIG 16:227



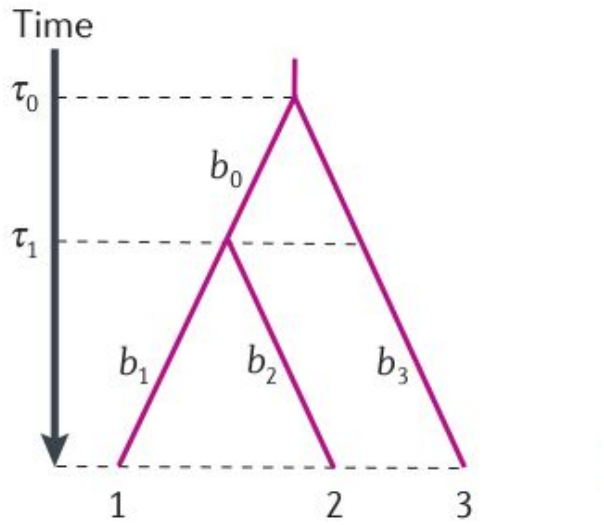
Introducción

Interpretación de los árboles

- si se trata de un árbol de especies, entonces los nodos representan eventos de especiación

- si se trata de árboles de genes obtenidos de una población, los nodos representan nacimientos de los ancestros de las terminales

- si se trata de una filogenia de genes parálogos, los nodos representan eventos de duplicación génica



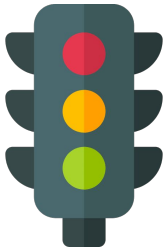
Limitaciones

✗ Transferencia horizontal de genes, genes con historia evolutiva diferente dentro de un mismo organismo

✗ Problemas de muestreo; cuando se construyen filogenias con escaso número de taxa o pocos caracteres, las relaciones pueden no resolverse completamente

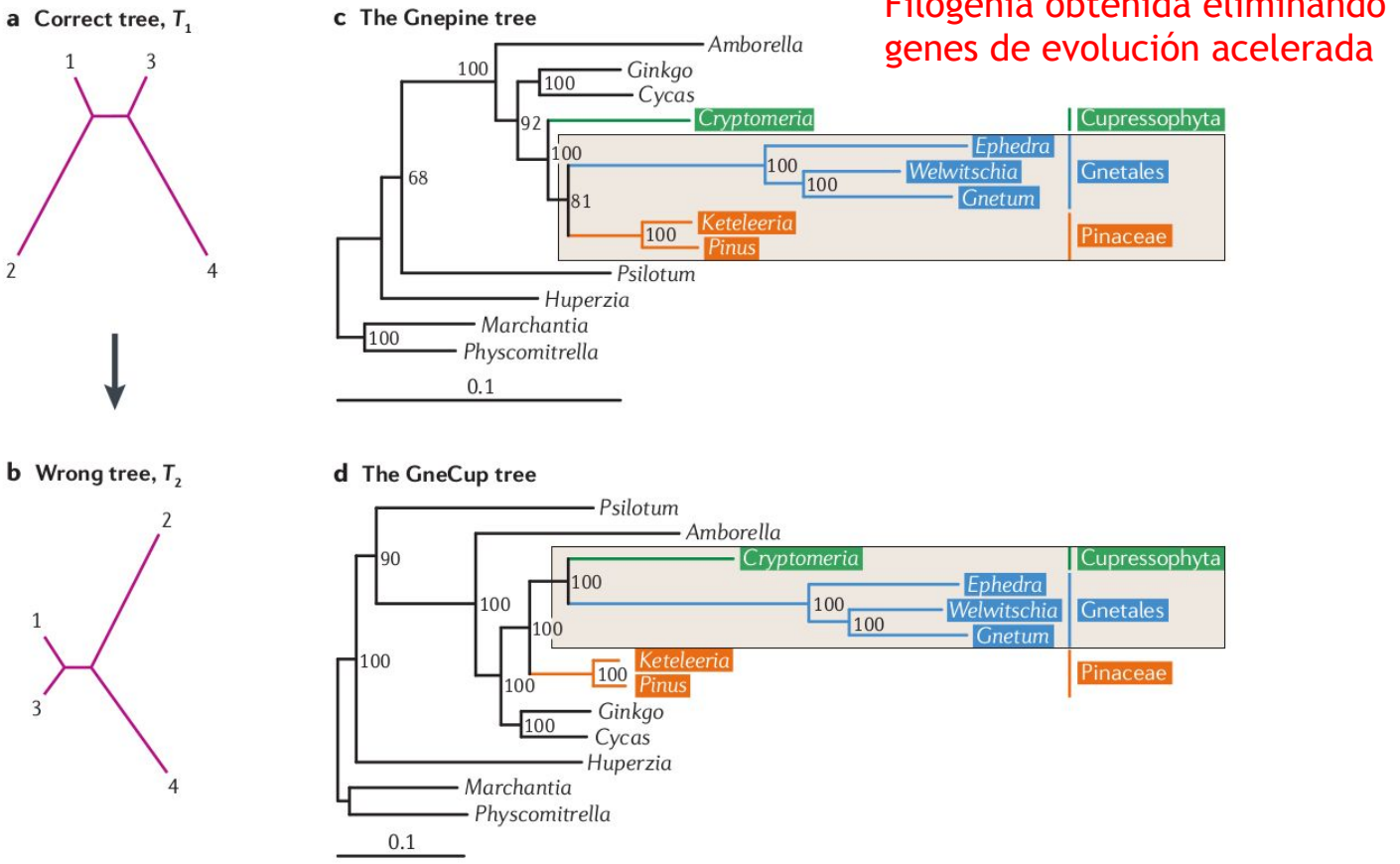
✗ Saturación; es el resultado de sustituciones múltiples en la misma posición, provoca subestimación de la divergencia entre taxa

✗ Atracción de ramas largas; como consecuencia de la saturación linajes muy distantes o divergentes aparecen en la filogenia como muy cercanos. Se subestima la tasa de sustitución



Limitaciones

Atracción de ramas largas



Filogenia obtenida eliminando genes de evolución acelerada

Reconstrucción filogenética

Datos

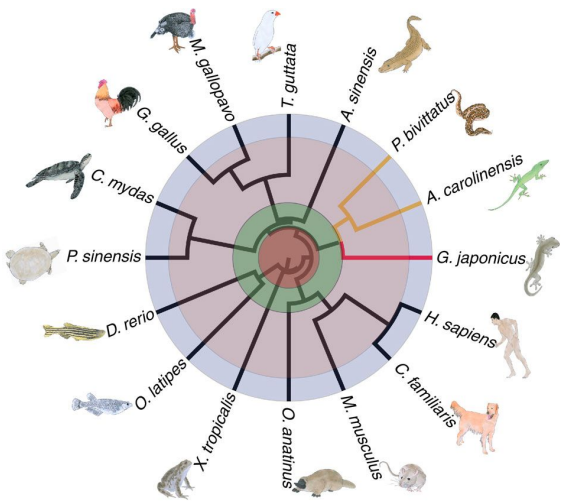
Alineamiento (ordenamiento de los datos)

Selección de modelo de sustitución (ajuste de datos a un modelo evolutivo)

Selección de método de reconstrucción filogenética

Evaluación de los árboles

Visualización



Reconstrucción filogenética
1. DATOS

✗ **Tipos de datos**
- Genomas completos (núcleo/organelas)
- Genes específicos
- ARN
- Proteínas

✗ **Origen**
- Generados por el usuario
- Obtenidos a partir de bases de datos (NCBI, Ensembl) **BLAST**
    - Búsqueda general
    - Búsqueda de secuencias relacionadas a datos propios

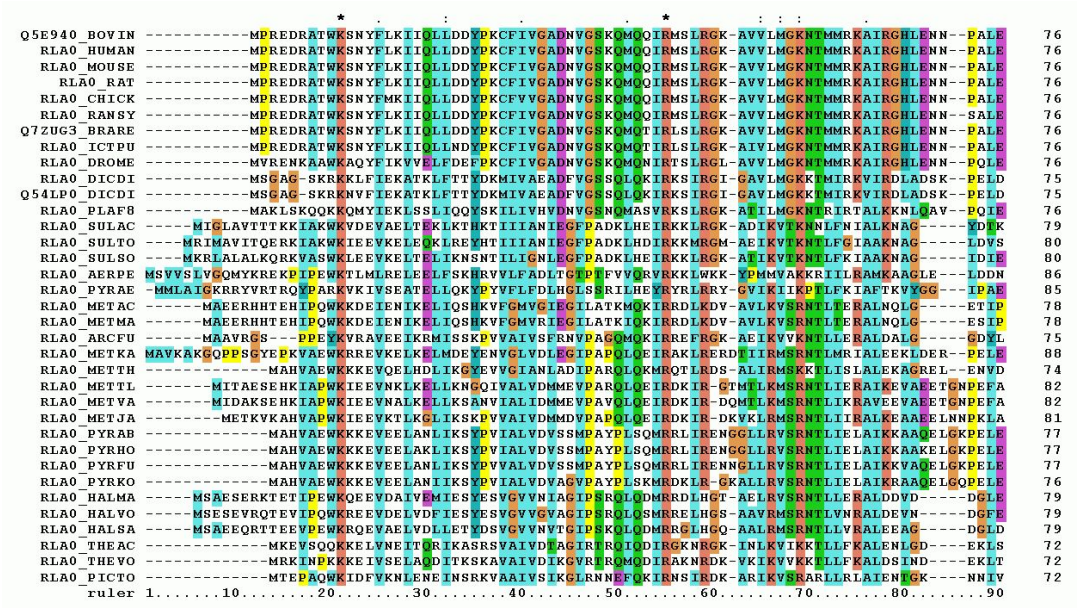
✗ **Consideraciones**
- Muestreo amplio (taxa, secuencias), evitar limitaciones

Reconstrucción filogenética
2. ALINEAMIENTOS

✗ Un alineamiento (múltiple) es el ordenamiento de las posiciones de todas las secuencias (ADN, ARN o Proteínas)

✗ Lleva implícito que las posiciones son “homólogas”, es decir que comparten un ancestro común, no se trata sólo de “similitud” entre las secuencias

✗ Un cambio en una posición supone la existencia de una mutación puntual (sustitución de nt por otro); o la desaparición de una posición (InDel)



Reconstrucción filogenética
2. ALINEAMIENTOS

✗ Obtener un buen alineamiento es imprescindible para reconstruir una filogenia de forma correcta

✗ Existen numerosos programas para realizar alineamientos múltiples

- Clustal http://www.clustal.org/ (https://www.ebi.ac.uk/Tools/msa/clustalo/)
- Kalign http://msa.sbc.su.se/cgi-bin/msa.cgi
- MAFFT http://mafft.cbrc.jp/alignment/software/
- MUSCLE http://www.drive5.com/muscle/
- T-Coffee http://www.tcoffee.org/Projects/tcoffee/
- PROBCONS http://toolkit.tuebingen.mpg.de/probcons
- ProtTest (alineamiento de proteínas)

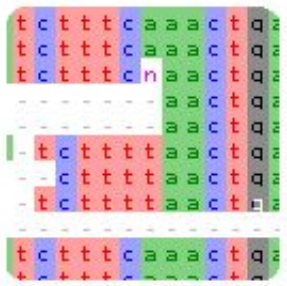
✗ Es evidente que a mayor divergencia entre las secuencias a alinear más complejo es el alineamiento y menos exacto el resultado

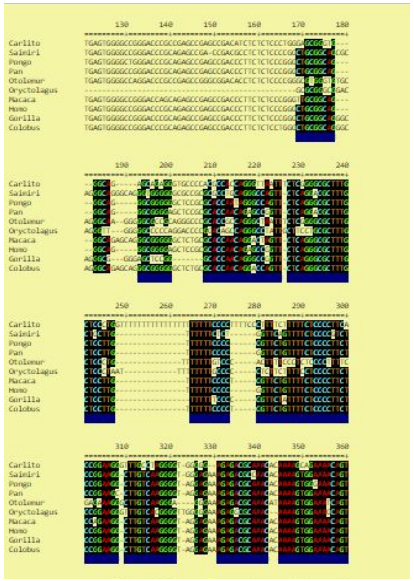
✗ Existen regiones divergentes de difícil alineamiento; ya sea porque no son homólogas o bien porque se han saturado debido a múltiples sustituciones. Estas regiones deben eliminarse del alineamiento antes de construir la filogenia, de modo de asegurarnos la homología entre las posiciones

Reconstrucción filogenética
2. ALINEAMIENTOS

✗ Existen programas que permiten seleccionar los bloques conservados de un alineamiento múltiple (MSA) para realizar filogenias

- Gblocks: se pueden modificar los parámetros de longitud de bloques conservados, regiones flanqueantes, etc
- TrimAl: permite además, modificar las proporciones de gaps en el alineamiento
- Es muy importante visualizar los alineamientos antes de realizar las filogenias
    - aliview
    - mega
    - DNAsp
    - Jalview

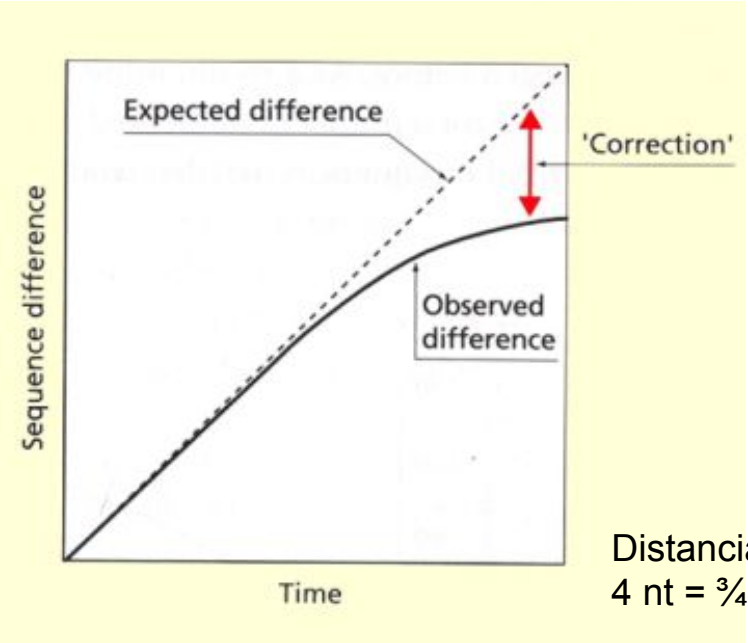




Primera parte de la práctica (A)

Reconstrucción filogenética
3. MODELOS DE SUSTITUCIÓN

Los modelos de sustitución se utilizan para "corregir" (o modelar) el fenómeno de saturación, es decir la subestimación de la distancia/tiempo debida a este fenómeno (llamado también "multiple hits")



Distancia max,
4 nt = ¾ = 0.75

Sustituciones múltiples

T → A → G → C → **T**

ancestral actual

Reconstrucción filogenética
3. MODELOS DE SUSTITUCIÓN

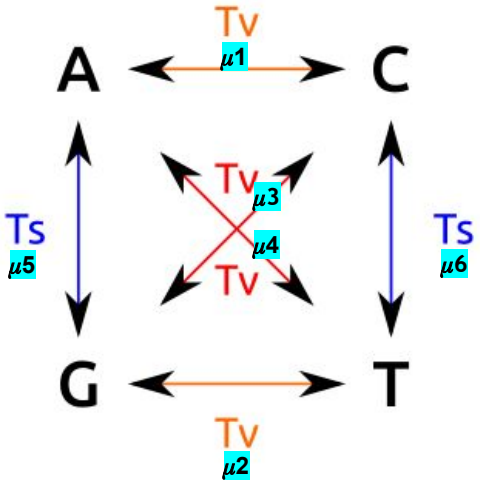
Estos modelos proponen tasas de cambio en función del tiempo, modelan las sustituciones de una secuencia

Consideran diferentes parámetros

- μ: tasa de sustitución (transversiones: A <-> T; A <-> C; G <-> T; G <-> C; transiciones (A <-> G; T <-> C)
- π: frecuencia nucleotídica

Los distintos modelos (y hay muchos) proponen distintas combinaciones de estos parámetros.

Las probabilidades de sustitución de cada nt se calculan a partir de una matriz



Matriz de probabilidades de sustitución

	A	T	C	G
A	-	$\pi_T\mu_4$	$\pi_C\mu_1$	$\pi_G\mu_5$
T	$\pi_A\mu_4$		$\pi_C\mu_6$	$\pi_G\mu_2$
C	$\pi_A\mu_1$	$\pi_T\mu_6$		$\pi_G\mu_3$
G	$\pi_A\mu_5$	$\pi_T\mu_2$	$\pi_C\mu_3$	

Reconstrucción filogenética
3. MODELOS DE SUSTITUCIÓN

Matriz de probabilidades de sustitución

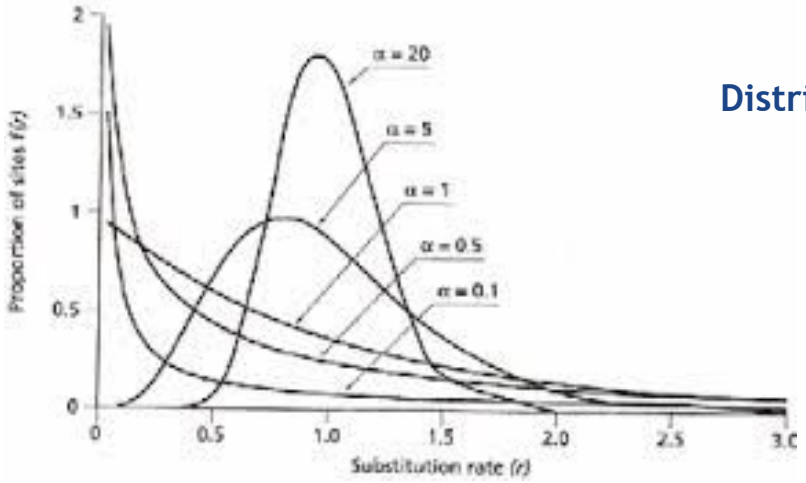
	A	**T**	**C**	**G**
A	-	$\pi_T\mu_4$	$\pi_C\mu_1$	$\pi_G\mu_5$
T	$\pi_A\mu_4$		$\pi_C\mu_6$	$\pi_G\mu_2$
C	$\pi_A\mu_1$	$\pi_T\mu_6$		$\pi_G\mu_3$
G	$\pi_A\mu_5$	$\pi_T\mu_2$	$\pi_C\mu_3$	

Los diferencia entre los modelos existentes es el número de parámetros variables, los hay muy sencillos hasta muy complejos:

- **JC69 (Jukes y Cantor)**: es el más sencillo, la frecuencia es la misma para todos los nucleótidos ($\pi_A=\pi_T=\pi_C=\pi_G=0.25$); existe una única tasa de sustitución μ ($\mu_1=\mu_2=\mu_3=\mu_4=\mu_5=\mu_6$)
- **HKY (Hasegawa, Kishino y Yano)**: las frecuencias de los nt es distinta ($\pi_A\neq\pi_T\neq\pi_C\neq\pi_G$) y la tasa de sustitución de las transversiones ($\mu_1=\mu_2=\mu_3=\mu_4$) es distinta a la de las transiciones ($\mu_5=\mu_6$)
- **GTR (General Time Reversible)**: es el modelo más complejo, todos los parámetros son distintos

Reconstrucción filogenética
3. MODELOS DE SUSTITUCIÓN

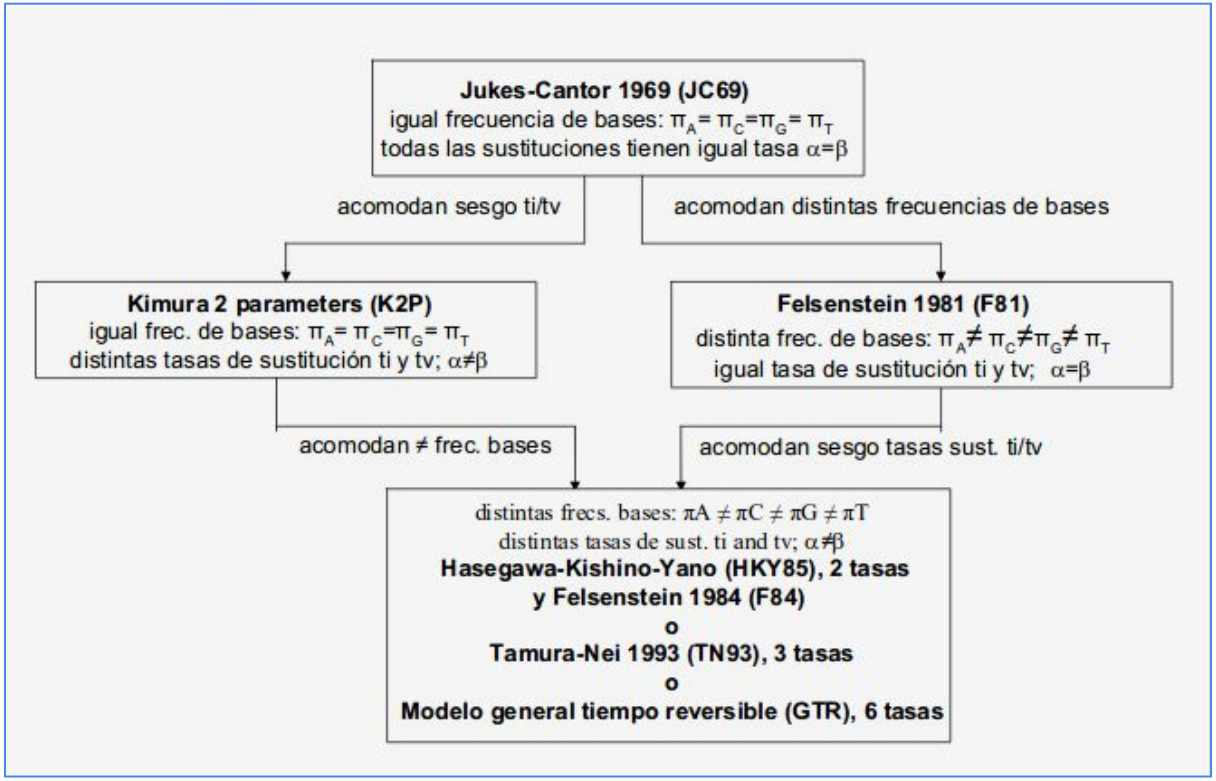
- En todos los modelos se asume que la tasa de sustitución es homogénea entre los distintos sitios de un alineamiento y que no existen sitios invariantes
- Esto no es cierto en los datos biológicos, por tanto distintos modelos pueden incluir estos parámetros
  - gamma (Γ), es la distribución que define que la sustitución es heterogénea entre las distintas posiciones; tiene un parámetro α, un valor de α > 1 supone poca heterogeneidad, por el contrario α < 1 supone elevada heterogeneidad entre sitio (en general se utiliza un valor de 4 o 5)
  - I, es la proporción de sitios invariantes



Distribución gamma

Reconstrucción filogenética
3. MODELOS DE SUSTITUCIÓN

Parámetros de los distintos modelos de sustitución



Reconstrucción filogenética
3. MODELOS DE SUSTITUCIÓN

Existen distintos programas para decidir cuál es el mejor modelo que se ajusta a nuestros datos

Uno de ellos jModelTest

- utiliza máxima verosimilitud (ML), compara los distintos modelos de sustitución y les da una puntutación (score)
- utiliza diferentes criterios o estrategias de selección para evaluar los diferentes modelos AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) y DT (Decision-theoretic performance based)
- Arroja como resultado un score para cada criterio y valores para los distintos parámetros

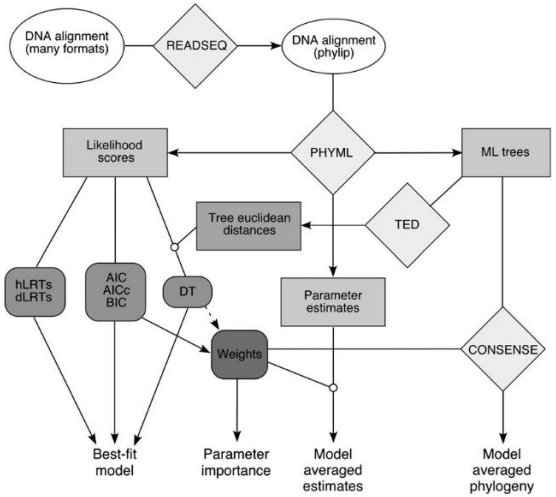


FIG. 1.—jModelTest pipeline. Alignments are loaded using the ReadSeq library (Gilbert 2007). Likelihood calculations, including estimates of model parameters and trees, are carried out with Phyml (Guindon and Gascuel 2003). A custom program called Ted (D. Posada) is used to compute Euclidean distances between trees for performance-based model selection (DT), whereas Consense (Felsenstein 2005) is used to calculate weighted and strict consensus trees representing model-averaged phylogenies.

Reconstrucción filogenética
4. MÉTODOS

✗ **Métodos basados en distancias**

se calculan distancias de a pares entre las secuencias utilizadas, la matriz resultante se utiliza par reconstruir la filogenia. Los más conocidos utilizan algoritmos de clustering para construir el árbol

- Neighbor-Joining (NJ)
- UPGMA (unweighted pair group method with arithmetic mean)

✗ **Métodos basados en caracteres**

Se considera cada posición de la secuencia como un carácter para calcular la puntuación (score) de cada árbol y se comparan todas las secuencias en forma simultánea.

- Máxima Parsimonia (Maximum Parsimony, MP; score = número de pasos)
- Maxima Verosimilitud (Maximum Likelihood, ML; score = log-likelihood)
- Métodos de inferencia Bayesiana (score = posterior probability)

En la actualidad, dado el elevado volumen de datos que se utiliza para reconstruir filogenias, las búsquedas de los mejore árboles son heurísticas (vs exhaustivas)

Todos los métodos comienzan generando uno (o varios) árboles iniciales y hacen rearreglos locales para alcanzar el mejor árbol (aquel con el mejor score)

Reconstrucción filogenética
4. MÉTODOS

✗ **Métodos basados en distancias**

Existen distintos tipos de distancias que son calculadas por lo programas de reconstrucción filogenética, la más simple es el número de diferencias entre dos secuencias.
Mega es uno de los programas que permite obtener filogenias con estos métodos

NJ (Neighbor-Joining)

- es un método muy utilizado, principalmente de manera exploratoria. No asume reloj molecular y utiliza modelos de evolución
- se obtienen resultados rápidos de modo que es útil cuando se tienen datasets muy grandes
- No es recomendable su uso con especies muy divergentes, ya que los errores de muestreo son más importantes
- Este método además es sensible a la presencia de gaps

UPGMA

- utiliza las medias de las distancias entre pares de secuencias
- Asume reloj molecular, generando árboles ultramétricos, por este motivo su uso está mucho menos extendido
- no considera modelos evolutivos

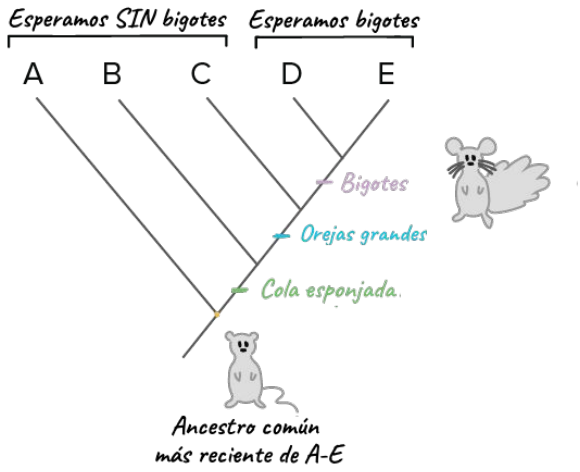
Reconstrucción filogenética
4. MÉTODOS

✗ **Métodos basados en caracteres**

Maxima Parsimonia (MP)

- no utiliza modelos de evolución (sustitución)
- la construcción del árbol se basa en la premisa de que el menor número de eventos evolutivos es lo que mejor explica la historia del grupo
- fue uno de los primeros métodos más ampliamente utilizados, principalmente en filogenias morfológicas
- en filogenias moleculares, la falta de uso de modelos evolutivos, la corrección por saturación (en general tiene problemas de atracción de ramas largas), así como las búsquedas de árboles heurísticas ponen en duda la precisión de este método para obtener buenos resultados

- se utiliza cada vez menos, aunque los resultados se obtienen en forma rápida
- en general deben compararse con otros métodos
- **Mega, PAUP, TNT:** programas que pueden utilizarse para obtener árboles por MP

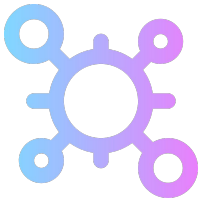


Reconstrucción filogenética
4. MÉTODOS

✗ **Métodos basados en caracteres**

Maximum Likelihood (ML)

- método ampliamente utilizado, se han desarrollado numerosos modelos evolutivos para reconstrucciones con ML
- considera cada topología (árbol) un modelo (hipótesis), mientras que los parámetros son la longitud de las ramas más los parámetros propios del modelo de sustitución elegido
- los parámetros son valores fijos y desconocidos
- la puntuación (*score*) de cada árbol es una comparación estadística entre varios modelos posibles (árboles), para obtener el mejor
- método robusto, aunque computacionalmente demandante
- permite evaluar hipótesis de reloj molecular así como ajustes a los modelos evolutivos
- muchos programas que reconstruyen filogenias ML: **PAUP, Phylip** son los más antiguos, actualmente **RAxML, PhyML** (más rápidos y permiten el análisis de datasets grandes) y **MEGA**



Reconstrucción filogenética
4. MÉTODOS

✗ **Métodos basados en caracteres**

Inferencia Bayesiana

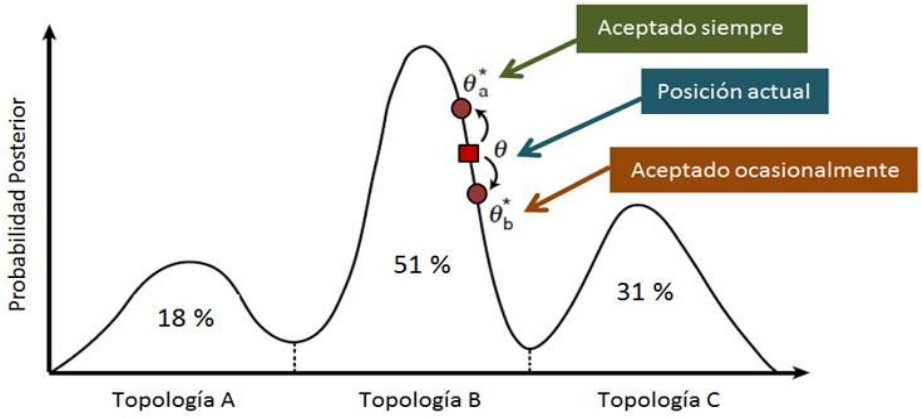
- método estadístico de reconstrucción más complejo
- comenzó a utilizarse en 1990 y se popularizó gracias al uso de MCMC (Cadenas de Markov de Monte Carlo)
- todos los parámetros que incluye son variables con distribuciones estadísticas asociadas (mientras que en ML son constantes fijas desconocidas)
- antes de comenzar el análisis a todas las variables se les asigna un "*prior*" o distribución *a priori* que describe los valores que puede tomar el parámetro
- los "*posteriors*" o **probabilidades *a posteriori* son proporcionales a su prior multiplicado por su verosimilitud (o probabilidad de los datos dados un árbol y modelo evolutivo)** (Teorema de Bayes)
- son muy complejos, es difícil definir los *priors* y requieren muchos recursos computacionales
- programas disponibles para Inferencia Bayesiana **MrBayes** y **BEAST**

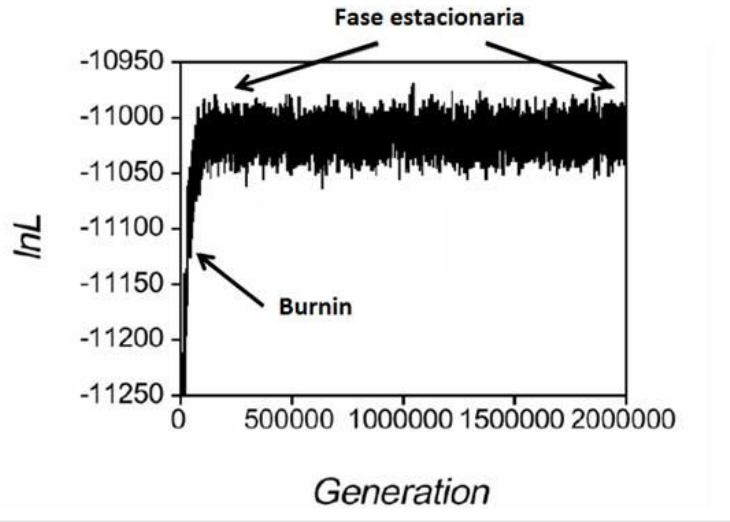
Reconstrucción filogenética
4. MÉTODOS

✗ **Métodos basados en caracteres**

Inferencia Bayesiana

- para obtener las probabilidades a posteriori se utilizan las MCMC
- las MCMC son algoritmos de simulación que permiten calcular los *posteriors* cambiando los valores de los *priors* según la distribución propuesta
- las cadenas comienzan en un árbol al azar, realizan el cálculo de los *posteriors* y se mueven hacia un árbol ligeramente diferente, si los resultados son mejores sigue desde ahí, sino vuelve al punto inicial; cada cambio es una generación. Se desea alcanzar el estado estacionario





Reconstrucción filogenética
4. MÉTODOS

✗ **Métodos basados en caracteres**

Inferencia Bayesiana

Este método permite estimar diversos parámetros evolutivos

- la tasa de sustitución dentro de un árbol
- el tamaño poblacional efectivo
- se utilizan para estimar cambios en el tamaño de la población en el tiempo
- datar filogenias
- realizar análisis epidemiológicos

Beast2

Bayesian evolutionary analysis by sampling trees

Reconstrucción filogenética
4. MÉTODOS

Table 2 | **A summary of strengths and weaknesses of different tree reconstruction methods**

Strengths	Weaknesses
Parsimony methods	
• Simplicity and intuitive appeal • The only framework appropriate for some data (such as SINES and LINES)	• Assumptions are implicit and poorly understood • Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution • Branch lengths are substantially underestimated when substitution rates are high • Maximum parsimony may suffer from long-branch attraction
Distance methods	
• Fast computational speed • Can be applied to any type of data as long as a genetic distance can be defined • Models for distance calculation can be chosen to fit data	• Most distance methods, such as neighbour joining, do not consider variances of distance estimates • Distance calculation is problematic when sequences are divergent and involve many alignment gaps • Negative branch lengths are not meaningful
Likelihood methods	
• Can use complex substitution models to approach biological reality • Powerful framework for estimating parameters and testing hypotheses	• Maximum likelihood iteration involves heavy computation • The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret
Bayesian methods	
• Can use realistic substitution models, as in maximum likelihood • Prior probability allows the incorporation of information or expert knowledge • Posterior probabilities for trees and clades have easy interpretations	• Markov chain Monte Carlo (MCMC) involves heavy computation • In large data sets, MCMC convergence and mixing problems can be hard to identify or rectify • Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge • Posterior probabilities often appear too high • Model selection involves challenging computation[138,139]

Reconstrucción filogenética
5. SOPORTE DE ÁRBOLES

La reconstrucción filogenética es, en realidad, una inferencia estadística; de modo que es esencial evaluar la fiabilidad tanto del árbol como de los distintos nodos dentro del mismo

Estas medidas o valores proporcionan fiabilidad y robustez a la filogenia (hipótesis evolutiva)

Existen varias medidas de soporte de árbol, en general están integradas a los programas de reconstrucción filogenética:

Bootstrap

- es la más conocida y consiste en un remuestreo de los datos (secuencias) con reemplazo, es decir se generan pseudoréplicas en las cuales se muestrea una parte de los sitios, se modifica el orden de las posiciones y se repiten algunas
- Con estas réplicas se realiza una nueva filogenias, este paso se repite tantas veces como el usuario determine (número de réplicas, 100 o 1000)
- Se obtiene un valor que indica el número de veces que ese nodo se repite entre las réplicas. i.e un bootstrap de 100 (o 1) significa que ese nodo es exactamente el mismo para todas las réplicas generadas, por tanto es “confiable”
- Este método se utiliza tanto en NJ, como ML y MP

Reconstrucción filogenética
5. SOPORTE DE ÁRBOLES

Bootstrap

Constructing bootstrap data sets. The original data set of 4 taxa (A–D) each with 10 nucleotide characters is bootstrapped across characters (with replacement) to produce bootstrap pseudoreplicates. Each pseudoreplicate contains each of the 4 original taxa, but some original characters are represented more than once and some not at all.

Original Data Set

Taxa	Characters
	1 2 3 4 5 6 7 8 9 10
A	C G A A C C A C T T
B	C G A A C C G G T T
C	G G T A C C G G A T
D	G C T A G C G C A T

Bootstrap Data Sets

Bootstrap Pseudoreplicate 1:

Taxa	Characters
	8 10 7 4 1 10 2 8 5 3
A	C T A A C T G C C A
B	G T G A C T G G C A
C	G T G A G T G G C T
D	C T G A G T C C G T

Bootstrap Pseudoreplicate 2:

Taxa	Characters
	1 8 10 4 2 9 2 8 5 6
A	C C T A G T G C C C
B	C G T A G T G G C C
C	G G T A G A G G C C
D	G C T A C A C C G C

Bootstrap Pseudoreplicate 3:

Taxa	Characters
	3 2 5 7 1 6 9 4 4 10
A	A G C A C C T A A T
B	A G C G C C T A A T
C	T G C G G C A A A T
D	T C G G G C A A A T

Bootstrap Pseudoreplicate 4:

Taxa	Characters
	7 8 5 8 9 6 4 10 1 5
A	A C C C T C A T C C
B	G G C G T C A T C C
C	G C G A C C A T G C
D	G C G C A C A T G C

Reconstrucción filogenética
5. SOPORTE DE ÁRBOLES

Existen varias medidas de soporte de árbol:

Jacknife

- Se puede usar en MP y ML
- Este método realiza pseudoréplicas de las filogenias, pero eliminando un taxon en diferente en cada una
- La premisa de este método es la ausencia de inconsistencias dentro de la base de datos a estudiar, la extracción de un taxon no debe modificar la topografía del árbol

aLRT

- *Approximate Likelihood-Ratio Test* es un análisis alternativo al *bootstrap* mucho más rápido y se basa el test de *Likelihood-Ratio Test*
- LRT evalúa para cada rama la diferencia entre los logaritmos de máxima verosimilitud del mejor reordenamiento y el logaritmo de máxima verosimilitud considerando que la rama de interés tiene longitud cero

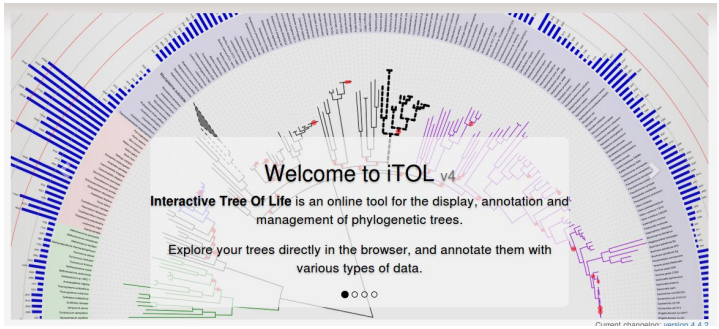
Reconstrucción filogenética
6. VISUALIZACION DE ÁRBOLES

Existen numerosos programas de visualización de árboles

- ITOL es una herramienta web https://itol.embl.de/
  que permite modificar el árbol, agregar gran cantidad de información y obtener árboles completos y listos para publicar

- FigTree, no tiene tantas posibilidades de modificar los gráficos pero es la herramienta que se utiliza para visualizar los árboles anotados de BEAST

- Mega, es muy útil porque permite manipular el árbol en forma sencilla, aunque el gráfico no suele ser tan “fashion” para publicar



Aplicaciones útiles

phylemon

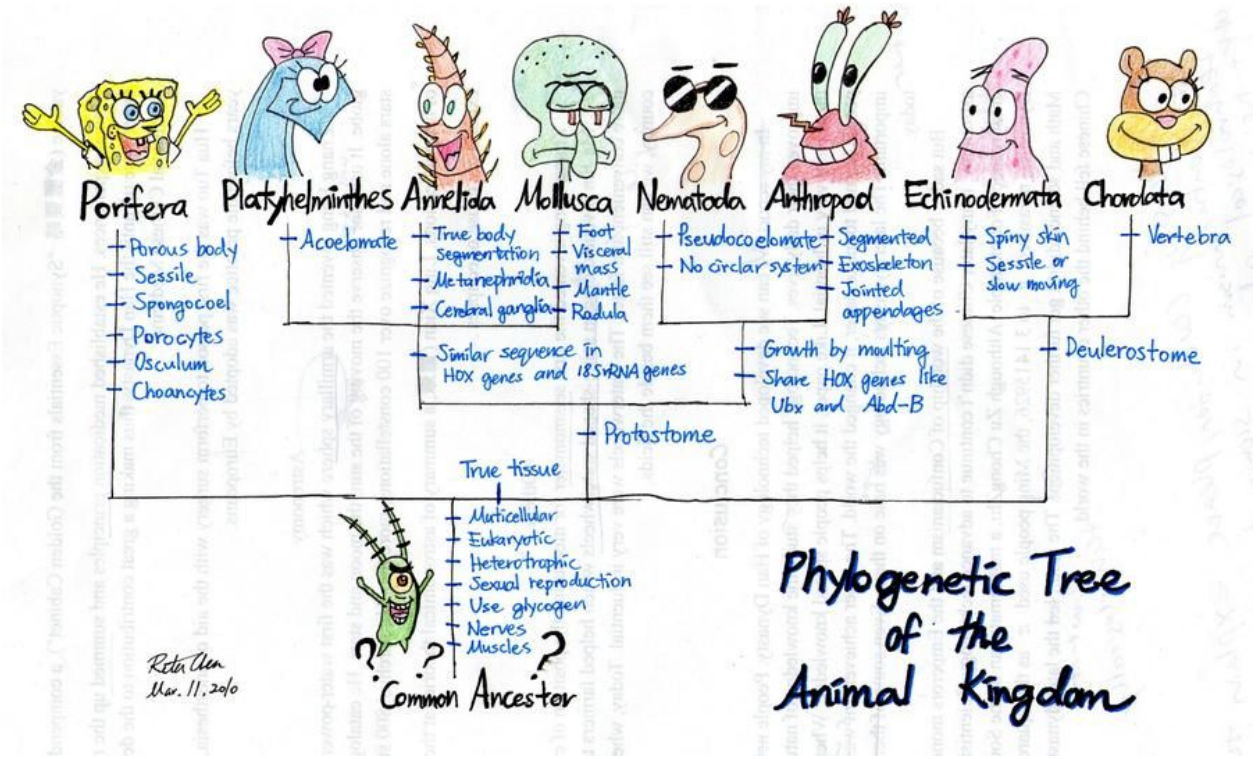
Table 1. Programs available in Phylemon 2.0 web server

Program[a]			Version	Function	Output to program	Pplin[b]
Alignment						
1	T	ClustalW	2.0.10	Multiple alignments. DNA and protein sequences	5, 8, 9, 11–16, 21–23, 26, 28	Y
2	T	Muscle	3.7	Multiple alignments. DNA and protein sequences	5, 8, 9, 11–16, 21–23, 26, 28	Y
3	T	Lagan	2.0	Pairwise alignment. Long and distant genomic sequences	5, 8, 9, 11	–
4	T	M-Lagan	2.0	Multiple alignments. Long and distant genomic sequences	5, 8, 9, 11–16, 21–23, 26, 28	–
5	U	TrimAl	1.3	Automated trimming of MSAs	8, 9, 11–16, 21–23, 26, 28	Y
6	U	CDS-ProtAl	1.0	Alignment of DNA coding sequence using protein template	8, 11–16, 22, 28–30	Y
7	U	ConcatenAl	1.0	Concatenation of MSAs	8, 9, 11–16, 21–23, 25, 26	–
8	U	ReadAl	1.3	File format conversion	1–5, 8, 9, 11–16, 21–23, 25–30	–
Phylogeny reconstruction						
9	T	Seqboot	Phylip 3.68	Bootstrap, jackknife or permutation resampling methods	11–16	Y
10	T	Consense	Phylip 3.68	Consensus tree reconstruction	20	Y
11	T	Dnadist	Phylip 3.68	DNA pairwise distances computation	17, 18	Y
12	T	Protdist	Phylip 3.68	Protein pairwise distances computation	17, 18	Y
13	T	DnaML	Phylip 3.68	ML tree reconstruction from DNA data	10, 20	–
14	T	ProML	Phylip 3.68	ML tree reconstruction from protein data	10, 20	–
15	T	DnaPars	Phylip 3.68	Maximum parsimony tree reconstruction from DNA data	10, 20	–
16	T	ProtPars	Phylip 3.68	Maximum parsimony tree reconstruction from protein data	10, 20	–
17	T	Neighbor	Phylip 3.68	Tree reconstruction using UPGMA and NJ methods	10, 20	Y
18	T	Fitch	Phylip 3.68	Tree reconstruction using LS and ME methods	10, 20	Y
19	U	TreeDist	Phylip 3.68	Distance computation among tree topologies	–	–
20	U	ETE	2.1 beta	Tree visualization	–	Y
21	T	PhyML-Best-AIC-Tree	1.0	ML tree with the best model fitting data under AIC estimation	20	Y
22	T	PhyML	3.00	Maximum likelihood analysis (MLA) of DNA & protein data	20	Y
23	T	Tree-Puzzle	5.2	MLA of DNA & protein sequences using quartets	20	–
24	T	MrBayes	3.1.2	Bayesian phylogenetic analysis of DNA and protein sequences	20	–
Evolutionary tests						
25	T	ProtTest	1.4	ML fitting of protein sequences to evolutionary models	–	–
26	T	jModelTest	0.1.0	Model testing and phylogeny averaging	–	–
27	T	RRTree	1.1.11	Relative rate test	–	–
28	T	SLR	1.3	Site-wise analysis of positive and negative selection	–	–
29	T	YN00	PAML 4.4c	Pairwise analysis of positive selection (PS) with counting methods	–	–
30	T	CodeML	PAML 4.4c	MLA of PS using sites, branch and branch-site models	–	–

Programs are assembled in three main blocks: (i) alignment and files format conversion; (ii) phylogenetic reconstruction; and (iii) evolutionary tests. New resources in this version are shown in cursive.
[a]T-U: tools/utilities.
[b]Pplin: programs able to run in the Pipeliner.

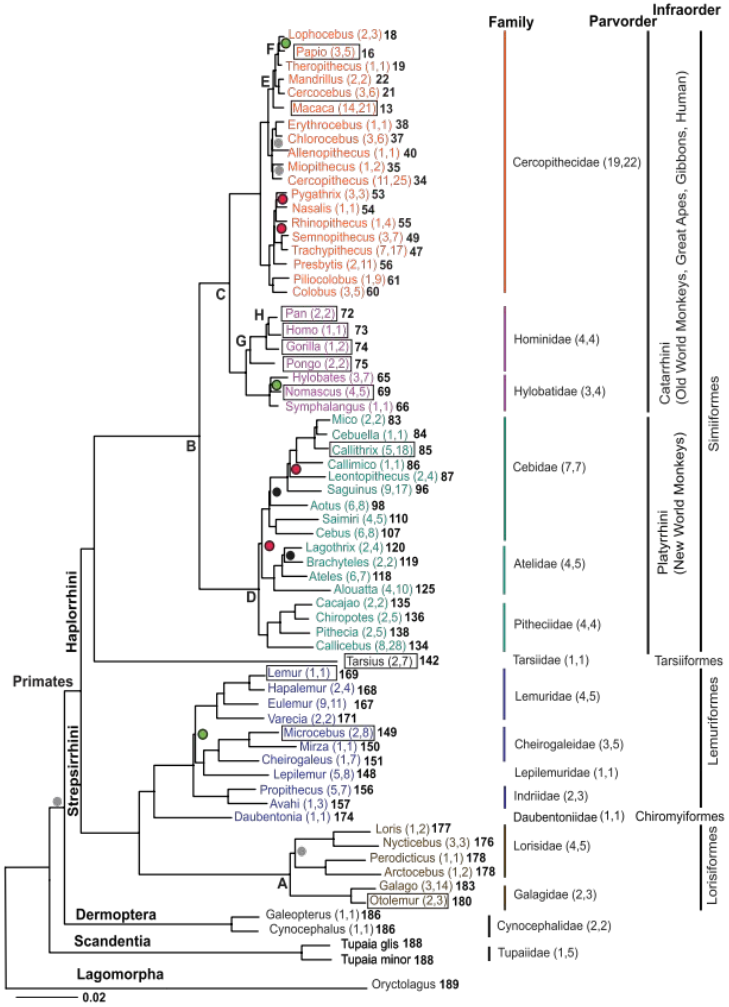
http://evolution.gs.washington.edu/phylip/software.html

Gracias!!!

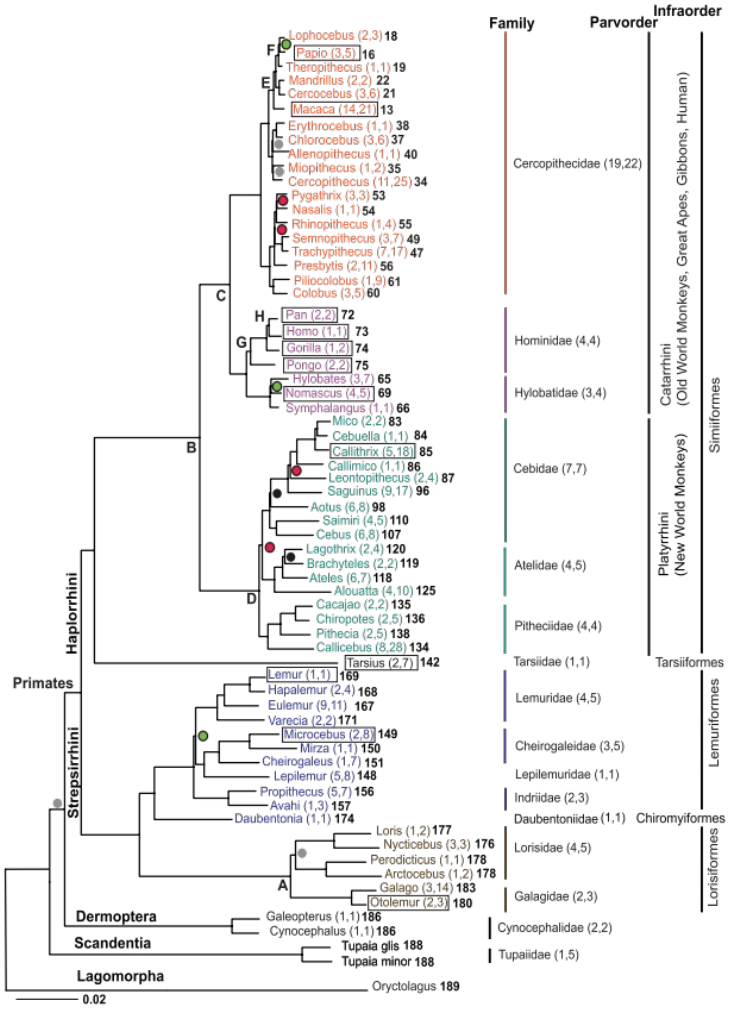

Porifera
Platyhelminthes
Annelida
Mollusca
Nematoda
Arthropod
Echinodermata
Chordata
Porous body
Sessile
Spongocoel
Porocytes
Osculum
Choanocytes
Acoelomate
True body segmentation
Metanephridia
Cerebral ganglia
Foot
Visceral mass
Mantle
Radula
Pseudocoelomate
No circlar system
Segmented
Exoskeleton
Jointed appendages
Spiny skin
Sessile or slow moving
Vertebra
Similar sequence in HOX genes and 18SrRNA genes
Growth by moulting
Share HOX genes like Ubx and Abd-B
Deulerostome
Protostome
True tissue
Multicellular
Eukaryotic
Heterotrophic
Sexual reproduction
Use glycogen
Nerves
Muscles
Common Ancestor
Phylogenetic Tree of the Animal Kingdom
Mar. 11, 2010

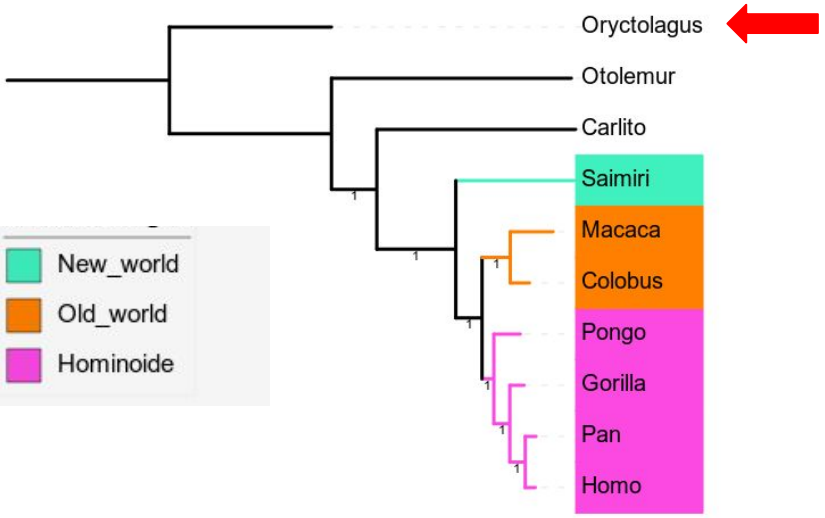
Práctica

Perelman et al. 2011. A Molecular Phylogeny of Living Primates. PLoS Genetics Vol 7 Issue 3 e1001342

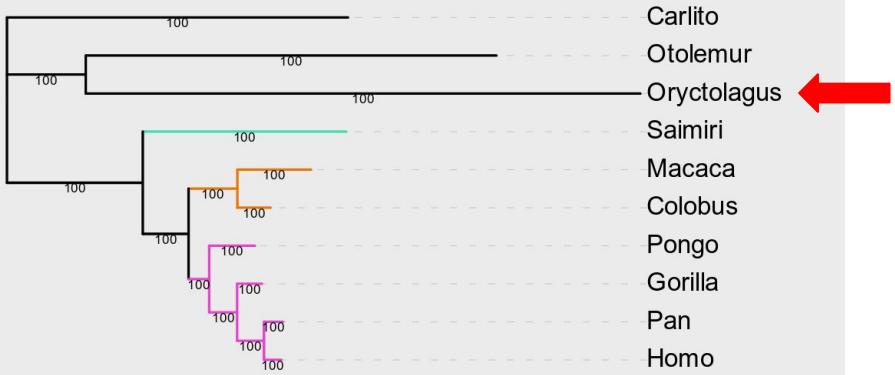


Práctica





MrBayes



Publicaciones

Yang et al. 2012. Molecular phylogenetics: principles and practiceNature Reviews doi:10.1038/nrg3186

Felsenstein J. 2004. Inferring Phylogenies. Sinauer Associates, Inc

Capella-Gutierrez et al. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15): 1972–1973 10.1093/bioinformatics/btp348

Bergsten, J. 2005. A review of long-branch attraction. *Cladistics* **21**(2): 163-193

Posada D. 2009. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008 Jul;25(7):1253-6. doi: 10.1093/molbev/msn083

Soltis & Soltis. 2003. Applying the Bootstrap in Phylogeny Reconstruction. Statistical Science 18(2). DOI:10.1214/ss/1063994980

Koonin E. 2005. Orthologs, Paralogs, and Evolutionary Genomics. Ann. Rev. Annu. Rev. Genet. 39:309–38. doi: 10.1146/annurev.genet.39.073003.114725

Sánchez R et al. 2011. Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. Nucleic Acids Research 39. doi:10.1093/nar/gkr408

Talavera & Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56, 564-57

Darriba et al. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9(8), 772