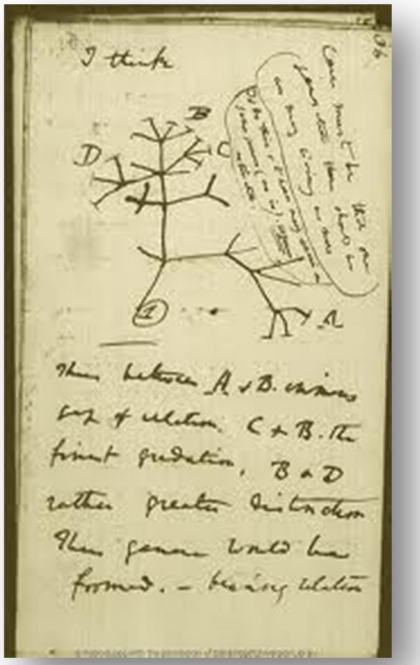
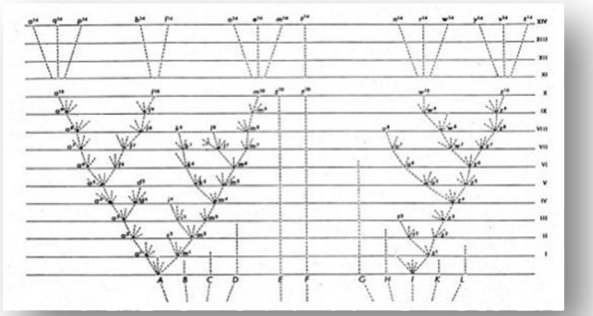
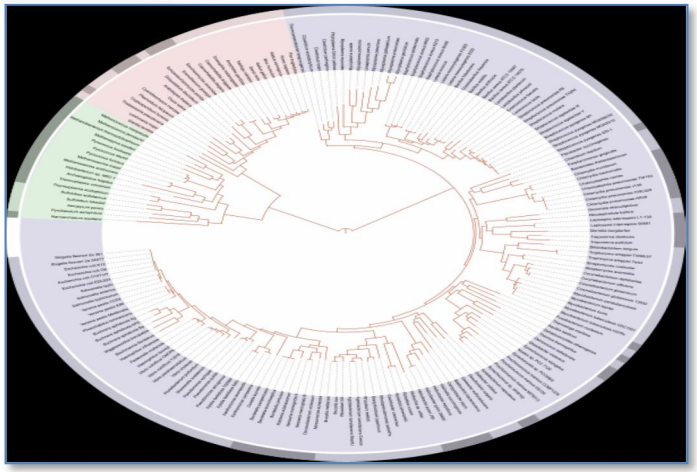
Filogenias





Charles Darwin (1859)

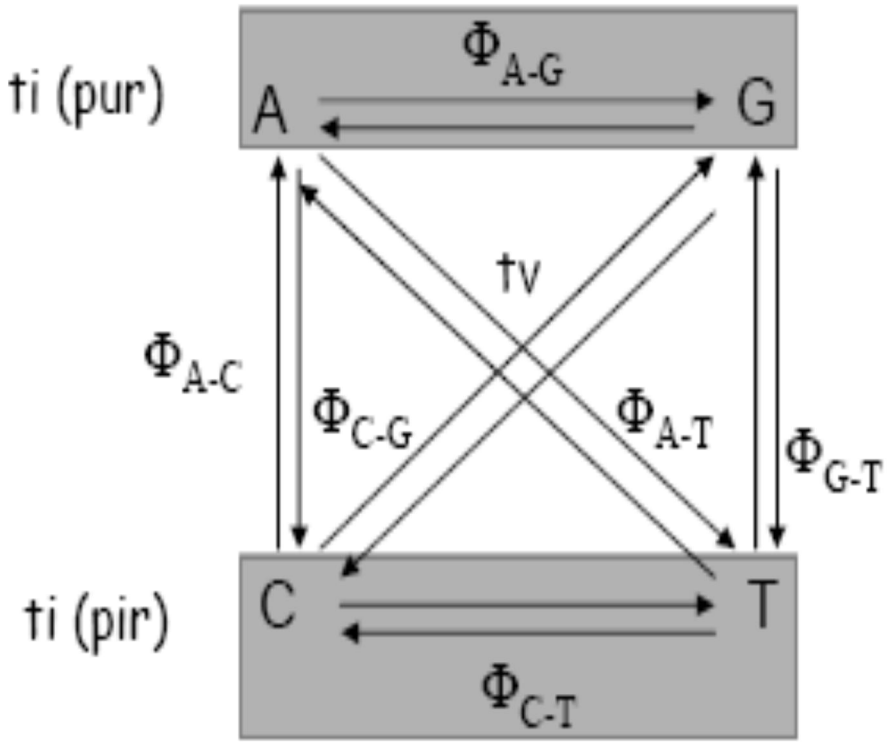


Filogenia

Consiste en el estudio de las relaciones evolutivas entre diferentes grupos de organismos. Su inferencia implica el desarrollo de hipótesis sobre los patrones de relación, buscando aquel que mejor explique las evidencias disponibles, utilizando caracteres fenotípicos o genotípicos.

- Métodos de distancias: La distancia es una medida del grado de divergencia entre dos secuencias. Con apoyo estadístico se generar una matriz de distancias por parejas de secuencias a partir del alineamiento múltiple.
- Máxima parsimonia: donde se selecciona la explicación más simple, esto será, la topología que requiera menor número de cambios para explicar nuestros datos.
- Modelos de evolución sumados a criterios estadísticos, que permiten calcular la probabilidad de obtener la distribución de estados de carácter para cada distinta topología posible

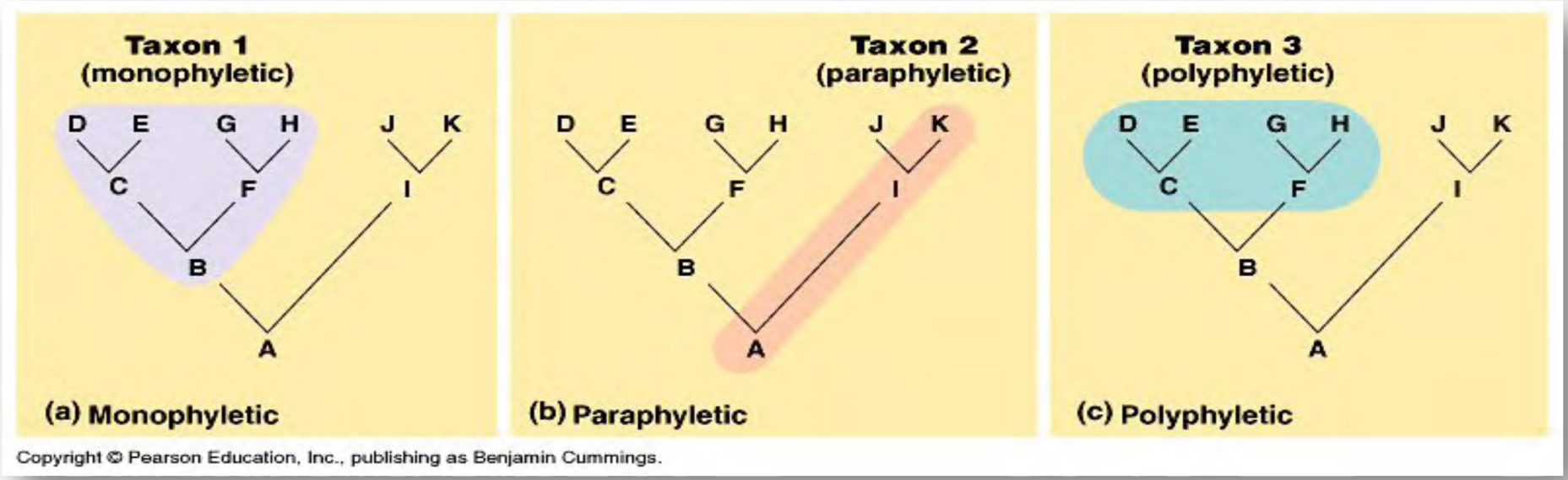
Modelos de sustituciones



- Existen 4 tipos de sustituciones ti y 8 tv; cuando **ti/tv ≠ 0.5 existe un sesgo en sustituciones ti (o tv) en el set de datos. ti generalmente >> 1**

- los modelos evolutivos se diferencian también en la cantidad de parámetros que utilizan para acomodar diversas tasas de sustitución:

tasas	modelo
1	**JC69** (ti=tv)
2	**K2P**, **F84** (ti ≠tv)
3	**TrN** ó **K3P** (2 ti, 1 tv)
6	**GTR** (cada sust. su tasa)



- Un **grupo monofiletico (clado)** es un grupo de taxa incluyendo una especie ancestral y todos sus descendientes, unidos por relaciones que no comparten con otros taxa.
- Un **grupo parafiletico** es un grupo de taxa que incluye a la especie ancestral pero ***no*** a todos sus descendientes.
- Un **grupo polifiletico** es un grupo de taxa que incluye a taxa descendientes de ***más*** de una especie ancestral.

Métodos

Basado en distancia:

- UPGMA, Min. Evolución y WPGMA (en desuso)
- Neightbor Joining (NJ)

Basado en Caracteres:

- Máxima Parsimonia
- Máxima Verosimilitud
- Beyecianos

Table 1 | **Comparison of methods**

Method	Advantages	Disadvantages	Software
Neighbour joining	Fast	Information is lost in compressing sequences into distances; reliable estimates of pairwise distances can be hard to obtain for divergent sequences	PAUP* MEGA PHYLIP
Parsimony	Fast enough for the analysis of hundreds of sequences; robust if branches are short (closely related sequences or dense sampling)	Can perform poorly if there is substantial variation in branch lengths	PAUP* NONA MEGA PHYLIP
Minimum evolution	Uses models to correct for unseen changes	Distance corrections can break down when distances are large	PAUP* MEGA PHYLIP
Maximum likelihood	The likelihood fully captures what the data tell us about the phylogeny under a given model	Can be prohibitively slow (depending on the thoroughness of the search and access to computational resources)	PAUP* PAML PHYLIP
Bayesian	Has a strong connection to the maximum likelihood method; might be a faster way to assess support for treesthan maximum likelihood bootstrapping	The prior distributions for parameters must be specified; it can be difficult to determine whether the Markov chain Monte Carlo (MCMC) approximation has run for long enough	MrBayes BAMBE

For a more complete list of software implementations, see online link to Phylogeny Programs. For software URLs, see online links box.

Basado en distancia

- Evalúa las distancias entre 2 taxa y busca agrupar las más cercanamente relacionadas (Vecinos).
- Los algoritmos difieren en la forma de calcular las distancias para las ramas.

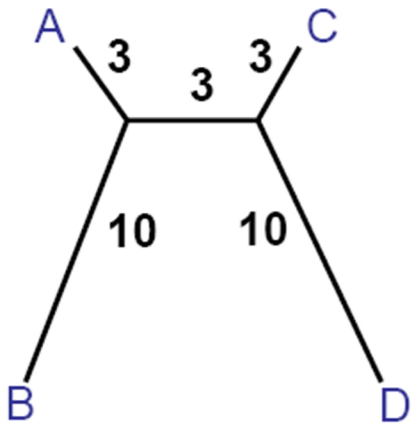
TABLE 5.1 Numbers of possible rooted and unrooted trees for up to 20 OTUs

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025
11	654,729,075	34,459,425
12	13,749,310,575	654,729,075
13	316,234,143,225	13,749,310,575
14	7,905,853,580,625	316,234,143,225
15	213,458,046,676,875	7,905,853,580,625
16	6,190,283,353,629,375	213,458,046,676,875
17	191,898,783,962,510,625	6,190,283,353,629,375
18	6,332,659,870,762,850,625	191,898,783,962,510,625
19	221,643,095,476,699,771,875	6,332,659,870,762,850,625
20	8,200,794,532,637,891,559,375	221,643,095,476,699,771,875

Data from Felsenstein (1978b).

Neighbor-Joining

Consiste en agrupar secuencialmente a los vecinos más próximos para minimizar la longitud total del árbol

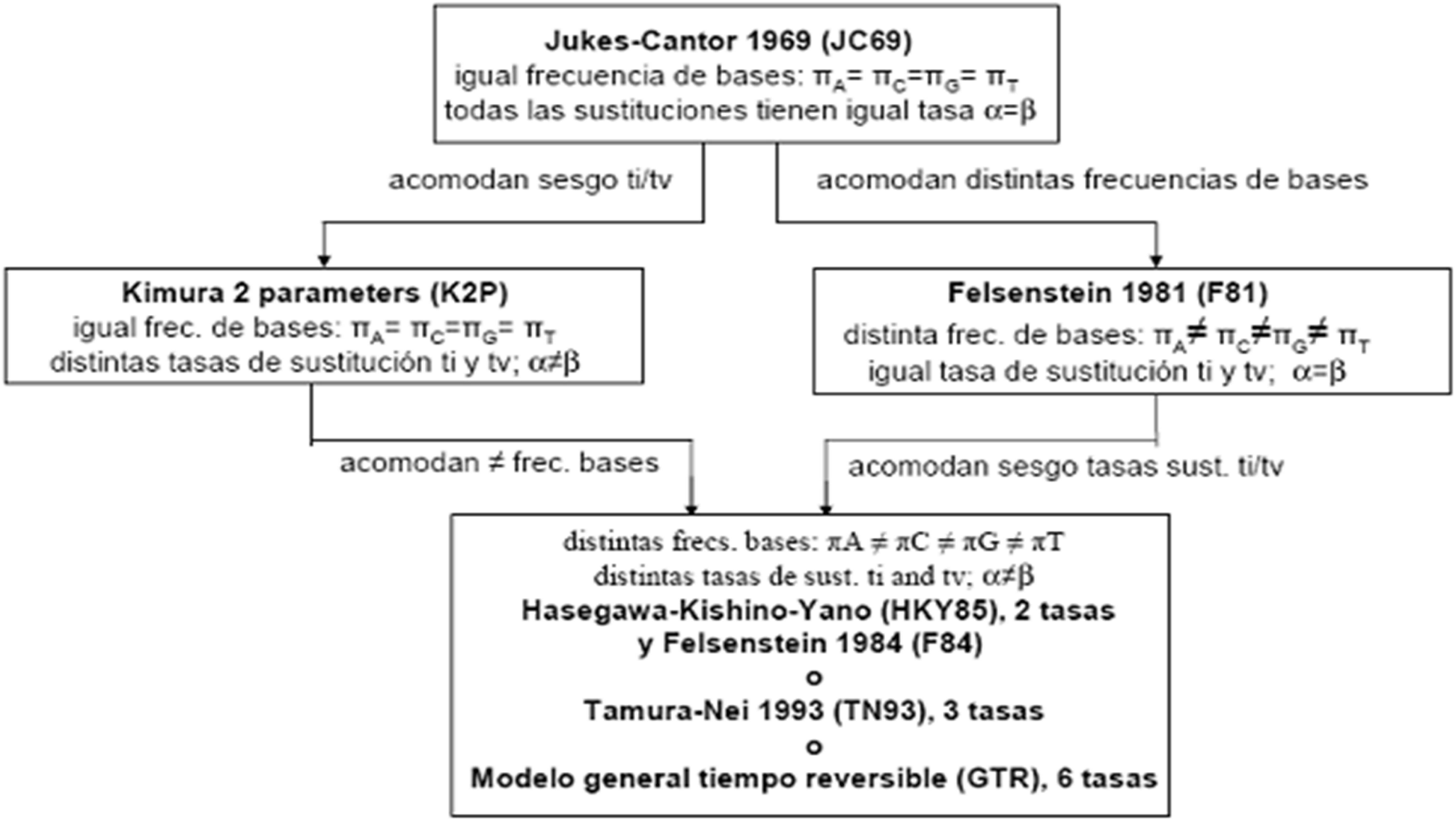


- Se estimanvaloresde Q (Studier & Keppler1988)
- Q12= (N-2)d12-Σd1i-Σd2i
- La matriz de distancias se modifica de forma que la distancia entre cada par depende también de la distancia de ambos respecto al resto de nodos.

$$d_{AB} + d_{CD} < d_{AC} + d_{BD}$$

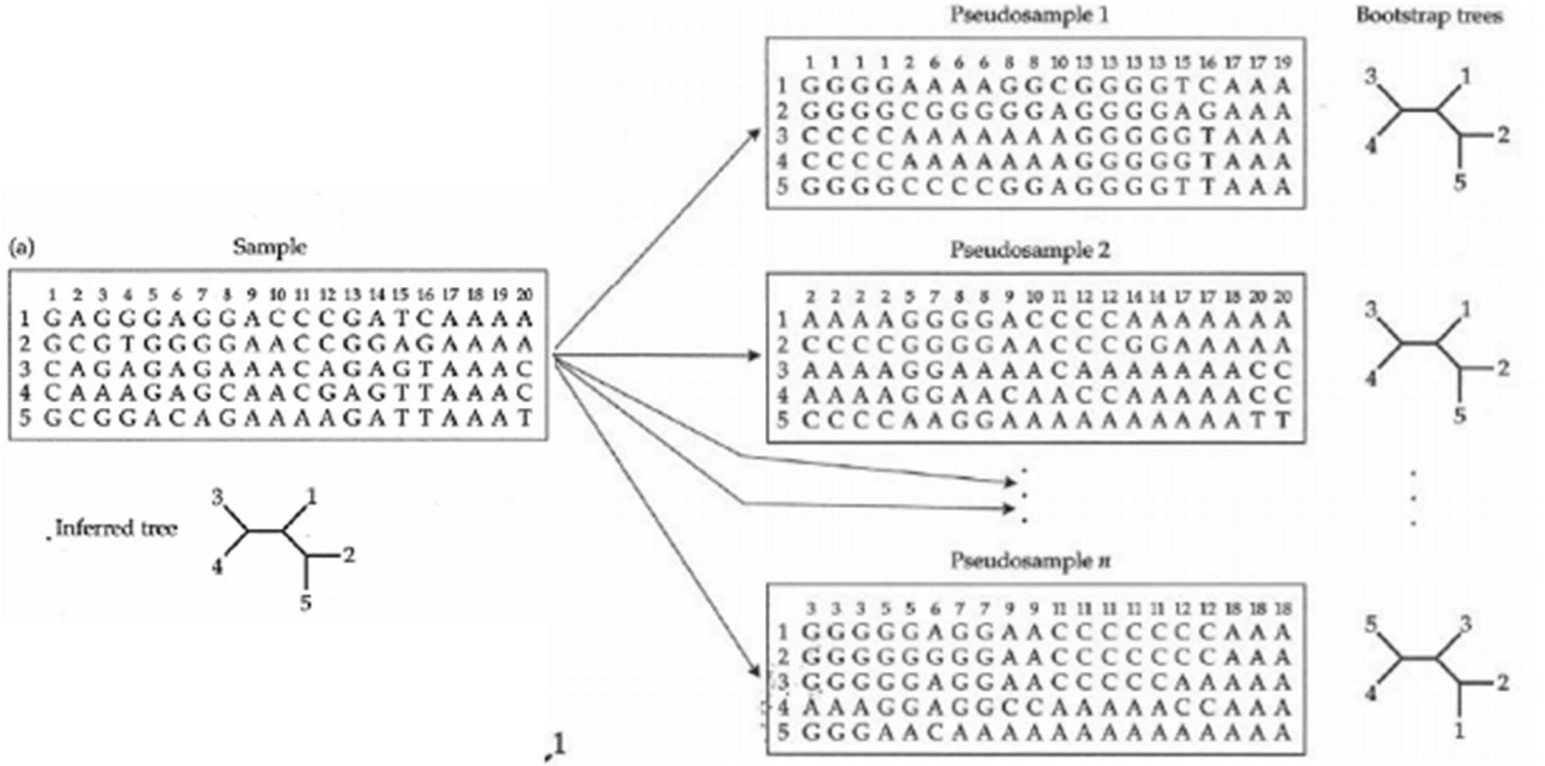
$$d_{AB} + d_{CD} < d_{AD} + d_{BC}$$

neighbors non-neighbors


Jukes-Cantor 1969 (JC69)
igual frecuencia de bases: $\pi_A = \pi_C = \pi_G = \pi_T$
todas las sustituciones tienen igual tasa $\alpha=\beta$
acomodan sesgo ti/tv
acomodan distintas frecuencias de bases
Kimura 2 parameters (K2P)
igual frec. de bases: $\pi_A = \pi_C = \pi_G = \pi_T$
distintas tasas de sustitución ti y tv; $\alpha \neq \beta$
Felsenstein 1981 (F81)
distinta frec. de bases: $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
igual tasa de sustitución ti y tv; $\alpha=\beta$
acomodan ≠ frec. bases
acomodan sesgo tasas sust. ti/tv
distintas frecs. bases: $\pi A \neq \pi C \neq \pi G \neq \pi T$
distintas tasas de sust. ti and tv; $\alpha \neq \beta$
Hasegawa-Kishino-Yano (HKY85), 2 tasas
y Felsenstein 1984 (F84)
o
Tamura-Nei 1993 (TN93), 3 tasas
o
Modelo general tiempo reversible (GTR), 6 tasas

Bootstrap

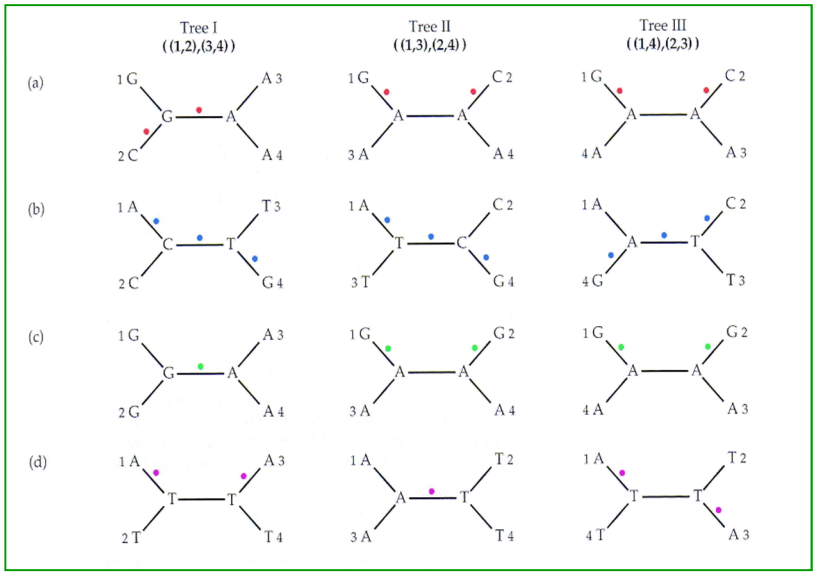
Consiste en crear réplicas de los alineamientos a partir del original, eliminando cierto número de posiciones al azar en cada replica. Para cada una de estas réplicas aplicaremos el método de reconstrucción filogenética y generaremos un árbol.



Basado en Caracteres

Máxima Parsimonia

Se busca el árbol con el menor número de cambios evolutivos para explicar las diferencias entre las Unidades Taxonomicas que lo componen.
No todos los sitios son informativos. Lo son sólo aquéllos que favorecen uno de los posibles árboles frente a los demás



Ventajas

- Fácil interpretación
- Utiliza más información que los métodos de distancias y no requiere un modelo de evolución

Desventajas

- Puede dar resultados erróneos si hay homoplasia
- Se justifica con argumentos filosóficos y no estadísticos.

Máxima Verosimilitud

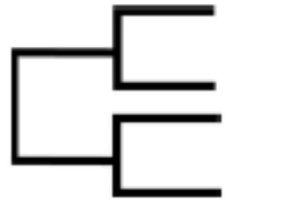
- La verosimilitud (V) de un árbol filogenético es la probabilidad de observar los datos obtenidos bajo cierto árbol y un modelo específico para el cambio de estados de caracteres.
- La idea es encontrar el árbol (entre todos los posibles) con el mayor valor de V

P(D/H)

Probability of

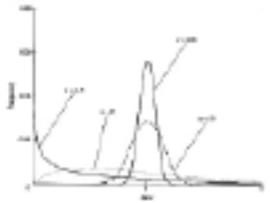
```
Seq1 aacg
seq2 accg
seq3 aaca
seq4 aatg
```

given



$$\pi = [a, c, g, t]$$

$$\begin{bmatrix} a & b & c & d \\ b & a & e & f \\ c & e & a & g \\ d & c & f & a \end{bmatrix}$$



Máxima Verosimilitud

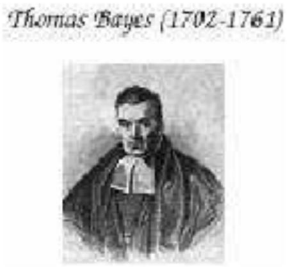
Ventajas

- Estadísticamente fiable y todos los sitios son informativos
- Permite estudiar qué modelo es mejor para los datos
- Permite realizar contrastes estadísticos entre hipótesis alternativas

Desventajas

- Si el modelo es incorrecto, entonces el árbol también será incorrecto

Métodos bayesianos



- **Maximum likelihood**: encuentra el árbol que con mayor probabilidad ha generado las secuencias observadas.

P(D|H)

- **Método bayesiano**: encuentra el árbol (o conjunto de árboles) que son explicados por las secuencias con mayor probabilidad. La p de que la hipótesis sea correcta dados unos datos

P(H|D)

Teorema de Bayes: Se llama *probabilidad a posteriori* porque se refiere a calcular la ***p*** de un modelo (árbol + modelo evolutivo) a partir de los resultados que produciría dicho modelo (el alineamiento de secuencias).

$$P(\theta/D) = \frac{P(\theta)\cdot P(D/\theta)}{P(D)}$$

Métodos bayesianos. MCMC

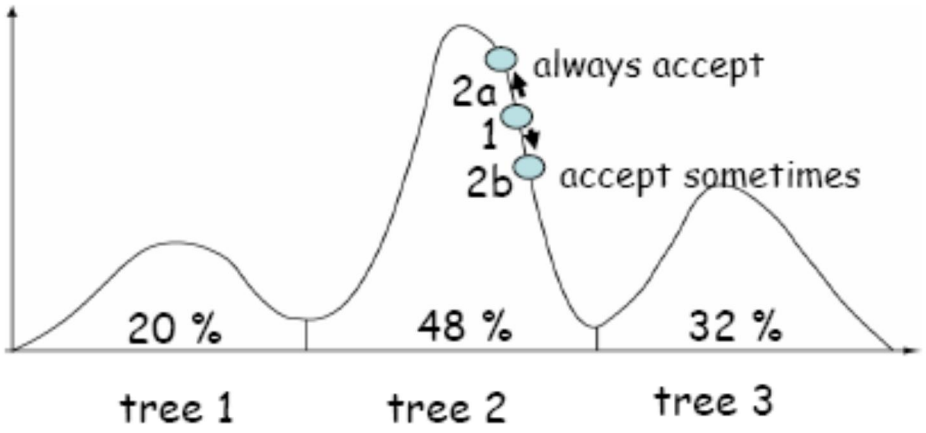
No hay solución analítica al sistema de Bayes.

No obstante, existen formas de estimar la distribución posterior =>

Markov Chain Monte Carlo

1. Comienza en un punto al azar.
2. Se hace un pequeño movimiento aleatorio.
3. Se calcula el radio de altura (*r*) del nuevo y antiguo estados:
   1. r > 1 -> se acepta el nuevo estado
   2. r < 1 -> el nuevo estado se acepta con probabilidad *r*. Si no se acepta permanece en el estado antiguo
4. Vuelve al paso 2.

Se modifica aleatoriamente la topología del árbol, o la longitud de las ramas, o un parámetro del modelo de evolución.



Se asume que la cantidad de veces que el procedimiento MCMC muestrea una región es una estimación de la densidad de probabilidad posterior de dicha región.

- **Pros:**
  - Faster than ML,
  - Accurate branch lengths,
  - There is no need to correct for "anything",
  - The model could include: instantaneous substitution rates, estimated frequencies, among site rate variation and invariable sites,
  - If the dataset is correct, the tree obtained is "correct",
  - All sites are informative,
  - There is no neccesary bootstrap interpretations
- **Cons:**
  - To what extent is the posterior distribution influenced by the prior?
  - How do we know that the chains have converged onto the stationary distribution?
  - **A solution**: Compare independent runs starting from different points in the parameter space

Programas recomendados

- ML: **Phyml**, **Hyphy**, **Paml**, **Tree-Puzzle**
- MB: **MrBayes**
- Selección de modelos: **ModelTest**, **ModelGenerator**, **ProtTest**.
- NJ, MP: **Mega**
- NJ, MP, (ML): **Paup**, **Phylip**
- Interconversión formatos: **ReadSeq**
- Alineamientos múltiples: **clustalw**, **muscle**.
- Visor alineamientos: **jalview**
- Otros**: Mesquite.**

Enlaces de interés

http://evolution.berkeley.edu/evolibrary/home.php

http://bioinformatics.ca/links_directory/

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