I CURSILLO DE FILOGENIA Y FILOGEOGRAFÍA: PARTE 1 RECONSTRUCCIÓN DE FILOGENIAS

ALEJANDRO LÓPEZ LÓPEZ – ÁREA DE BIOLOGÍA ANIMAL – UNIVERSIDAD DE MURCIA

HABLANDO CON PROPIEDAD

LOS ÁRBOLES

El elemento esencial en filogenética es el **árbol**, representación gráfica de las relaciones que unen diversos organismos o taxones. Está compuesto por una serie de **nodos** unidos mediante ramas.

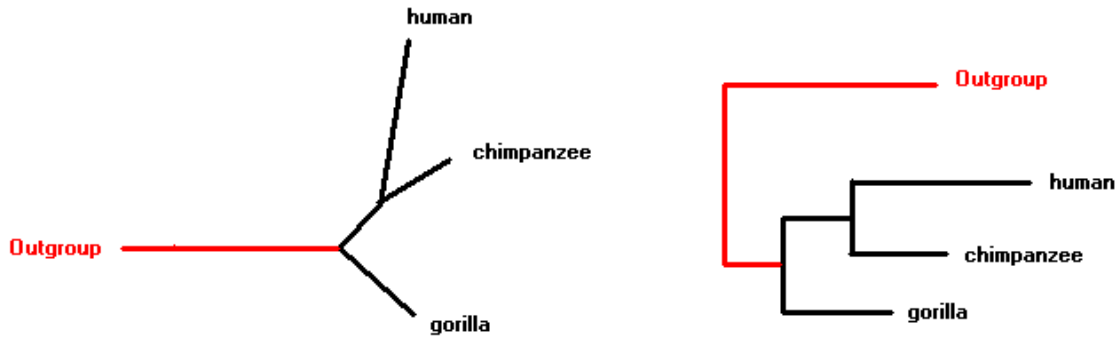


Imagen 1: Árbol sin enraizar (a la izquierda) y enraizado (a la derecha)

Los nodos pueden ser **terminales** (correspondientes a cada uno de los taxones que estamos estudiando) o **internos**. De cada nodo interno salen **tres** ramas, cada una de las cuales puede llevar a un nodo terminal o a otro nodo interno.

EL ENRAIZADO

Los árboles pueden estar **enraizados** o **no enraizados**. Un árbol sin enraizar (Imagen 1, izquierda) se limita a mostrar la distancia que separa a los diferentes taxones, pero no permite determinar cuáles ocupan una posición más **basal** o más **derivada**.

Dos métodos para enraizar, u orientar, un árbol son:

a) **Mid-point rooting**: Se basa en tomar los taxones más alejados y situar la **raíz** en el punto situado a igual distancia de ambos. Al basarse en una premisa tan débil, no es aconsejable.
b) **Uso de un outgroup**: Se basa en incluir al menos un taxón externo al grupo que estamos estudiando, y situar la raíz entre dicho taxón y los organismos de

estudio (**ingroup**). Se aconseja utilizar al menos tres outgroups que posean una posición parafilética con respecto al ingroup (Imagen 2).

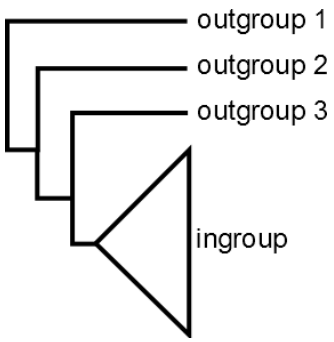


Imagen 2: Árbol indicando las relaciones ideales de los outgroups

TIPOS DE ÁRBOLES

Según sus características, podemos distinguir varios tipos de árboles:

- **Dendrograma:** Esta palabra se utiliza para cualquier tipo de árbol en general, se refiere al árbol como objeto matemático que representa las relaciones entre varios elementos.
- **Cladograma:** Árbol que únicamente muestra información de las relaciones entre los organismos que lo componen (patrón de ramificación). Es el obtenido por métodos cladísticos, como la Máxima Parsimonia.
- **Filograma:** Árbol que muestra información sobre el número de sustituciones nucleotídicas que se han producido en cada linaje, proporcionalmente a la longitud de las ramas.
- **Cronograma:** Árbol en el que la longitud de las ramas es proporcional al tiempo transcurrido entre los diferentes eventos de ramificación.
- **Árbol ultramétrico:** Árbol en el que los nodos terminales se encuentran alineados en vertical.

CLADOS

Los clados, o grupos de organismos, pueden ser varios tipos:

a) **Monofiléticos:** Engloban a *todos* los organismos que descienden de un ancestro común. Es el único tipo de clado considerado como válido. Ej: mamíferos, angiospermas…
b) **Parafiléticos:** Engloban a organismos que descienden de un ancestro común, sin incluir a todos los descendientes de éste. Técnicamente no es correcto definirlos como grupos, aunque algunos se consideran válidos por tradición (ej: reptiles).
c) **Polifiléticos:** Engloban a una serie de organismos sin relación de parentesco. Ej: vertebrados voladores…

Se habla de que dos clados son **clados hermanos** cuando descienden de un ancestro común. En los árboles podemos identificarlos como los dos "descendientes" de un nodo interno. Por ejemplo: humanos y chimpancés, en la Imagen 1 (derecha).

Lo normal es que los nodos se bifurquen en dos clados (**dicotomías**)[1]. En ocasiones podemos encontrar tres (**tritomías**) o más (**politomías**) clados surgiendo de un único nodo. Esto es un artefacto producido porque el algoritmo de construcción del árbol no ha podido decidir cuál es la relación entre dichos clados. Pueden ser debidos a que las secuencias son escasas o de mala calidad, a que tenemos secuencias idénticas, o también pueden surgir a la hora de realizar **árboles consenso**. Son perfectamente normales y publicables, a no ser que sean muy exageradas[2].

LOS DATOS

FORMATO PHYLIP

Fue el primer formato estándar para introducir datos a los programas. Lo desarrolló Joe Felsenstein en 1980, cuando creó su paquete de programas PHYLIP, el cual fue el primer conjunto de programas ampliamente utilizado para reconstruir filogenias. De hecho, la mayoría de los programas actuales utilizan el código de PHYLIP (que aún sigue actualizándose) como base, por lo que pueden ser considerados como "hijos" de éste.

El formato consiste en un archivo de texto. En la primera línea se indican el número de taxones que componen la matriz y el número de caracteres que tenemos para cada uno de ellos. A esto le sigue la matriz en sí: en cada línea se escribe el nombre del taxón (PHYLIP reserva 10 caracteres para ello, así que si el nombre ocupa 6 caracteres hay que añadir 4 espacios en blanco detrás) y la secuencia de caracteres:

```
8    6
Alpha1    AAGAAG
Alpha2    AAGAAG
Beta1     AAGGGG
Beta2     AAGGGG
Gamma1    AGGAAG
Gamma2    AGGAAG
Delta     GGAGGA
Epsilon   GGAAAG
```

El formato PHYLIP requiere que las secuencias tengan la misma longitud y estén alineadas.

---

[1] ¡Cuidado!: cada nodo interno tiene TRES ramas que surgen de él: las dos que lo unen a sus clados "hijos" y la que lo une al resto del árbol. No hay que olvidar a esta última, que a veces tiende a pasarse por alto si el árbol está orientado.

[2] Formadas por muchos clados (las llamamos extraoficialmente "raspogramas" o "peines").

FORMATO FASTA

Este formato fue desarrollado por Lipman y Pearson en 1985. Se trata de un archivo de texto que contiene una colección de secuencias. No tienen por qué estar alineadas, ser de la misma longitud ni tener relación alguna entre ellas, por lo que se usa como "almacén". Cada secuencia ocupa dos líneas: en la primera figura su nombre precedido por el símbolo ">", y en la segunda está la secuencia en sí:

```
>Seq1
ATCGATCGGACGATCGATGCATCGACTG
>Seq2
AGCTAGCTACGATGCATTCGATCGATGCATCGATGC
>Seq3
ACGGACTCGTAGCAGCGACGGAGCATGCATCG
```

Recientemente la empresa Illumina ha desarrollado una versión mejorada, FASTQ, que incluye información acerca de la calidad de cada secuencia. De momento no ha sido implementada en programas filogenéticos, pero podría convertirse en un formato a tener en cuenta en el futuro.

FORMATO MEGA

Es el formato desarrollado por Kumar, Tamura y Nei para ser usado por su programa MEGA. Es similar al FASTA, pero usando el símbolo "#" en vez de ">", añadiendo una línea "#Mega" al principio, y opcionalmente varias líneas de comentarios después de ella:

```
#Mega
Title: prueba

#168
ACTGATCGATCGATCGATGCACGCG
#169
GACTGATCGATGCTAGCTGACGATC
#171
ACGTAGCATGCTAGCTGATCATGCT
#176
GACTGACTGACTACTGCTGATGCTA
#199
ACGTCAGATGATCGATGTAGCATCG
```

El formato MEGA requiere que las secuencias estén alineadas y tengan la misma longitud.

FORMATO NEXUS

Es el más utilizado por su versatilidad y estandarización. En este caso, el archivo de texto comienza con la línea "#NEXUS", y luego vienen una serie de **bloques**.

Cada bloque NEXUS comienza con la palabra `begin` (seguida del nombre del bloque) y acaba con la palabra `end`. Cada línea acaba con un punto y coma. La definición de todos los bloques y comandos posibles se detalla en el artículo que lo definió[3].

El bloque básico es el llamado `data`, diseñado para contener la matriz de datos. Incluye una serie de líneas que definen las características de la matriz (tamaño, tipo de datos, símbolos especiales…), luego una línea que dice simplemente `matrix` y seguidamente la matriz en sí, en la que los nombres de las secuencias están separadas de éstas por un tabulador[4]:

```
begin data;
dimensions ntax=6 nchar=8;
format datatype=dna missing=? gap=-;
matrix
Taxon1     AGCCGTTA
Taxon2     AGTCGT??
Taxon3     AGCCATTA
Taxon4     GGGCGTTA
Taxon5     AGCCG--A
Taxon6     AGCCG--A
;
end;
```

En ocasiones este bloque aparece separado en dos: los bloques `taxa` y `characters`. Personalmente lo encuentro redundante y trato de evitarlo, pero es cuestión de gustos.

Otro tipo de bloque muy útil es `assumptions`, que nos permite definir particiones en nuestra matriz. Esto es necesario si nuestra matriz está formada por varios fragmentos, porque así los programas calcularán los parámetros de forma independiente para cada uno de ellos, produciendo unos resultados más correctos y fiables:

```
begin assumptions;
charset fragmento1 = 1-450;
charset fragmento2 = 451-893;
charset fragmento3 = 894-1432;
end;
```

Este bloque también es útil si queremos que el programa considera de forma diferente a la tercera posición de cada codón en genes codificantes, a los *loops* de los ARN ribosomales, a los intrones, etc.

FORMATO NEWICK

[3] Maddison, D.R., Swofford, D.L., Maddison, W.P. (1997). *NEXUS: An extensible file format for systematic information.* Systematic Biology 46(4):590-621.

[4] Algunos programas usan saltos de línea, lo cual crea a veces problemas cuando otro programa más estricto lee esa matriz.

Es el empleado para almacenar los árboles. Se llama así porque fue inventado por Archie, Day, Felsenstein, Maddison, Meacham, Rohlf y Swofford durante una cena informal en el restaurante Newick de Dover, New Hampshire (Imagen 3).

Imagen 3: Newick Restaurant (¡sigue abierto!) Fotografía de JBTHEMILKER, en Panoramio

Este formato representa cada nodo por medio de unos paréntesis, dentro de los cuales se engloban los clados "hijos" de dicho nodo separados por comas. El final de un árbol viene marcado por un punto y coma obligatorio.

Por ejemplo, un árbol sencillo formado por los taxones A y B figuraría como:

```
(A,B);
```

Los paréntesis se van anidando de la misma forma que los clados se anidan unos dentro los otros:

```
(((A,B),C),(D,E));
```

Se puede añadir información tal como la longitud de las ramas añadiendo dos puntos detrás de cada nodo seguidos por dicha información:

```
(((A:0.01,B:0.02):0.11,C:0.13):0.23,(D:0.09,E:0.07):0.45);
```

DATOS, GAPS Y MISSING DATA

Una de las principales controversias en filogenia es el manejo de los GAPs, es decir, saltos en la secuencia debidos a inserciones o deleciones. Suelen representarse con un guión. Por ejemplo:

```
ATCGA----AGCTAGGTAGCTAGCGATCGA
ATCGA----AGCTAGCTAGCTAGCGATCGA
ACCGA----AGCTAGCTAGCT-GCGATAGA
ATCGA--TGAGCTAGCTAGCT-GCGATCGA
ATCGA--TGATCTAGCTAGCT-GCTATCGA
ATCGATCTAGGCTAGCTAGCT-GAGATCGA
ATCGATCTGAGCTAGACAGCT-GCGATCGA
```

En un principio se trataban como si fueran un quinto tipo de nucleótido (lo que se llamaba "*gaps as 5th state*"), pero surgieron críticos con esta metodología debido a que los gaps en realidad no existen, sino que son un concepto abstracto inventado para hacer que las matrices "cuadren".

Otra forma de tratarlos era codificarlos ocmo "*missing data*", es decir, considerarlos de la misma forma que los fragmentos desconocidos de algunas secuencias incompletas (normalmente se representan mediante un interrogante "?"). Esto también tuvo sus

críticos, ya que ambos conceptos son bastante diferentes y deben ser considerados de diferente manera.

Como alternativa se sugirió eliminar esas posiciones y codificar los gaps como caracteres estándar (con ceros y unos, significando ausencia y presencia respectivamente). Esto es relativamente sencillo para gaps simples, como el que aparece en la posición 22 de la matriz de ejemplo: está claro que está en unas secuencias y en otras no. Pero la cosa se complica cuando miramos, por ejemplo, el gran gap de las posiciones 6 a 9: ¿El gap más grande de las secuencias 1 a 3 es homólogo (tiene el mismo origen) que el más reducido observado en las secuencias 4 y 5? ¿Deriva de él o tiene un origen distinto? ¿O en realidad tenemos dos gaps: uno en las posiciones 6 y 7, y otro en las posiciones 8 y 9?

El debate continúa en el presente mientras los teóricos buscan una solución definitiva a este problema.

MÉTODOS DE DISTANCIAS

UPGMA

Desarrollado por Sokal y Michener en 1958, éste fue el primer método que permitió reconstruir filogenias. Es muy simple, ya que se basa en la reconstrucción jerárquica del árbol a partir de una matriz de distancias sencillas.

Estas distancias pueden calcularse mediante diferentes índices. Uno de los más simples y conocidos es el Índice de Jaccard, el cual divide el número de posiciones diferentes entre ambas secuencias por el número total de posiciones:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Comienza tomando los taxones con menor distancia entre ellos y los dibuja en el árbol. El proceso se va repitiendo hasta que todos los taxones están dibujados. Una explicación detallada y paso a paso del proceso puede encontrarse en el siguiente enlace: http://www.southampton.ac.uk/~re1u06/teaching/upgma/

Este método sería ideal si el grado de diferencia entre dos secuencias fuera proporcional a su tiempo de divergencia, pero en realidad no es así (Imagen 4). Esto es debido a que no siempre una diferencia representa un cambio.

Por ejemplo, si tenemos una A en una secuencia y una C en otra puede haber ocurrido:

a) El ancestro tenía A, ha habido una sustitución en la secuencia 2 (1 diferencia – 1 sustitución).

b) El ancestro tenía A, la secuencia 2 cambió primero a G y luego a C (1 diferencia – 2 sustituciones).
c) El ancestro tenía G, la secuencia 1 cambió a A y la secuencia 2 a C (1 diferencia – 2 sustituciones).
d) El ancestro tenía G, la secuencia 1 cambió primero a T y luego a A, la secuencia 2 cambió primero a C, luego volvió a G y por último cambió a C (1 diferencia – 5 sustituciones).

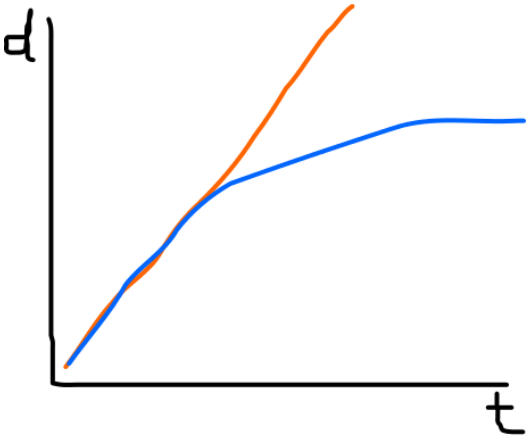


Imagen 4: Representación del número de diferencias entre dos secuencias (d) con respecto al tiempo transcurrido desde su divergencia (t). En naranja la línea esperada. En azul la línea real.

Así mismo, podemos tener casos en los que veamos nucleótidos iguales en ambas secuencias, pero pueden proceder de cambios iguales a partir de un nucleótido distinto en el ancestro. Y muchos más casos diferentes…

Para tratar con esto se inventaron los modelos de sustitución nucleotídica.

MODELOS DE SUSTITUCIÓN NUCLEOTÍDICA

Los modelos constan de una serie de parámetros:

- Por un lado tenemos una matriz que define las tasas de transición de un tipo de nucléotido a otro. Tiene cuatro filas y cuatro columnas, representando a los cuatro nucleótidos en orden alfabético (A, C, G, T). Por ejemplo: el elemento situado en la segunda fila, tercera columna, representa la tasa con la que C cambia a G.
- Por otro lado tenemos una matriz de frecuencias, en la que se representan las frecuencias con las que cada nucleótido puede aparecer en la matriz de datos.

El modelo más sencillo es el propuesto por Jukes y Cantor en 1969 (**Jukes&Cantor**, **JC**). Éste considera que las tasas de sustitución son iguales para todos los nucleótidos, así como sus frecuencias:

$$P_i = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix} \qquad f = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

Posteriormente Kimura propuso una alternativa (**Kimura-2-parámetros**, **K80**, **K2P**) en las que las tasas de **transiciones** (sustituciones entre nucleótidos del mismo tipo) son diferentes a las tasas de transversiones (sustituciones entre nucleótidos de diferente tipo):

$$P_i = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix} \qquad f = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

Felsenstein, por el contrario, propuso una alternativa (**F81**) en la que lo que cambia son las frecuencias de los diferentes nucleótidos (comprobad cómo afectan a la matriz de tasas de sustitución):

$$P_i = \begin{bmatrix} . & \pi_C\alpha & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & . & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & . & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & \pi_G\alpha & . \end{bmatrix} \qquad f = [\pi_A \quad \pi_C \quad \pi_G \quad \pi_T]$$

En definitiva, los modelos pueden ir complicándose añadiendo parámetros a la matriz de sustitución (como hizo Kimura) y/o considerando diferentes frecuencias para los nucleótidos (como hizo Felsenstein). Esta es una tabla que resume todas las combinaciones posibles con las siglas de los diferentes modelos:

Número de tasas de sustitución	Frecuencias iguales	Frecuencias diferentes
1	JC	F81
2	K2P, K80	HKY85
3	K3ST y TrNef	K81uf y TrN
4	TIMef	TIM
5	TVMef	TVM
6	SYM	GTR

El modelo más complejo es el Generalised Time-Reversible (**GTR**), propuesto por Tavaré en 1986:

$$P_i = \begin{bmatrix} . & \pi_C\alpha & \pi_G\beta & \pi_T\gamma \\ \pi_A\alpha & . & \pi_G\delta & \pi_T\varepsilon \\ \pi_A\beta & \pi_C\delta & . & \pi_T\theta \\ \pi_A\gamma & \pi_C\varepsilon & \pi_G\theta & . \end{bmatrix} \qquad f = [\pi_A \quad \pi_C \quad \pi_G \quad \pi_T]$$

NEIGHBOR-JOINING

Esta versión mejorada del UPGMA fue desarrollada por Saitou y Nei en 1987. El método es más complejo, ya que las distancias se corrigen aplicando un modelo de sustitución nucleotídica. Además, el árbol no se reconstruye a partir de la propia matriz de distancias, sino que ésta se transforma en una **matriz Q** mediante un cálculo complejo.

Como ventaja principal cabe destacar que es muy rápido y no requiere gran potencia de cálculo. Por otro lado, es poco fiable y muy sensible a artefactos. Por ello, suele usarse para tener una idea general de cuáles son, a grandes rasgos, las relaciones entre los taxones analizados antes de pasar a realizar análisis más complejos. También es empleado por métodos más complejos, como la Máxima Verosimilitud, para crear el árbol inicial que necesitan.

BOOTSTRAP

El bootstrap nos permite saber cuán fiables son los nodos que obtenemos. Este método sigue este procedimiento:

1. Se realiza el análisis con la matriz original y se obtiene el árbol.
2. Se crean réplicas de la matriz inicial construidas tomando posiciones al azar de dicha matriz.
3. Para cada réplica se repite el análisis con las mismas condiciones y se obtiene su correspondiente árbol.
4. Para cada nodo del árbol inicial, se mira en cuántos árboles de las réplicas aparece. El valor obtenido (en porcentaje) es el valor de bootstrap para ese nodo.

Se considera que un nodo comienza a ser fiable si su valor de bootstrap es superior al 50%. Obviamente, serán mejores valores próximos al 100%.

MÉTODOS CLADÍSTICOS: MÁXIMA PARSIMONIA

BASES TEÓRICAS

El método de Máxima Parsimonia se basa en el principio filosófico conocido como "*la navaja de Occam"* o "*lex parsomoniae*", ya propuesto por el monje inglés Guillermo de Ockham en el siglo XIV (quien tomó la idea de los griegos clásicos). En su formulación moderna se define como:

En igualdad de condiciones, la explicación más sencilla suele ser la correcta.

Es un principio bastante criticado, sobre todo a la hora de aplicarlo a sistemas biológicos (los cuales muestran una tenaz predisposición a ignorar los principios de la lógica[5]), ya que la explicación correcta no tiene por qué ser la más sencilla.

En filogenia fue adoptado con el siguiente aspecto:

El árbol más corto suele ser el correcto.

[5] Hecho conocido como Ley de Malcolm en honor al personaje de Parque Jurásico que lo expresaba como "la vida se abre camino".

En este sentido, la **longitud de un árbol** se define como el número de sustituciones que han debido ocurrir a lo largo de todas sus ramas para producir la matriz de datos original.

Hay que tener en cuenta que la Máxima Parsimonia es muy sensible a un efecto llamado **fenómeno de atracción de ramas largas** (LBA), que se produce cuando linajes con altas tasas de sustituciones nucleotídicas tienden a aparecer juntos en el árbol, aunque no tengan relación real entre ellos. El no tenerlo en cuenta ha llevado en ocasiones a presentar conclusiones taxonómicas erróneas: entre las más famosas encontramos la revelación de que la cobaya no era un roedor (se ha visto que esto es rotundamente falso), la posición del erizo como mamífero ancestral (posteriores análisis lo devuelven a su lugar entre los Insectívoros) o la de los microsporidios como eucariotas basales (realmente son hongos).

CÁLCULO DE LA LONGITUD DE UN ÁRBOL DETERMINADO

Para cada posición en la matriz, se coloca en cada nodo terminal su nucleótido correspondiente. Entonces se van tratando de reconstruir los caracteres ancestrales situados en cada nodo interno del árbol. Por ejemplo:

- En dos nodos terminales hermanos tenemos una A. Entonces en el nodo del que parten lo más probable es que hubiera una A (por la navaja de Occam).
- Si en uno tenemos A y en otro C, se determina el nucleótido ancestral usando uno de los diversos algoritmos de parsimonia que hay definidos. El más común es el conocido como Fitch, aunque existen también otros (Wagner, Dollo, Transversion…).

Cuando se tiene colocado en cada nodo el nucleótido correspondiente, se averigua el número total de sustituciones que han debido ocurrir.

Finalmente se repite el proceso para todas las posiciones de la matriz, y se van sumando las sustituciones para cada una de ellas. El total es la longitud del árbol.

ALGORITMO DE MÁXIMA PARSIMONIA

Lo ideal sería dibujar todos los árboles posibles, medir la longitud de cada uno de ellos y finalmente quedarnos con el más corto.

Sin embargo esto no es casi nunca factible, ya que el número de árboles (enraizados) posibles para un conjunto de *T* taxones viene dado por la siguiente fórmula:

$$N = (2T - 3) \cdot \prod_{i=3}^{T} (2i - 5)$$

Lo cual nos dice que para 5 taxones tenemos 105 árboles, para 10 taxones hay más de 34 millones de árboles posibles, y cuando llegamos a 50 taxones el número de árboles es mayor que el número de átomos que existen en el Universo.

Dado que para calcular todos los árboles haría falta un ordenador más grande que el propio Universo, se ha recurrido a diversos algoritmos para hacer factible la tarea:

a) **Búsqueda exhaustiva**: Consiste en construir todos los posibles árboles y evaluar la longitud de cada uno de ellos. Los árboles se construyen iterativamente: se comienza con tres taxones al azar (el árbol más simple posible) y se van añadiendo taxones (cada uno en todas las posiciones posibles) (Imagen 5). Este método asegura encontrar el árbol más corto, pero sólo es factible para conjuntos de taxones muy pequeños (10 como máximo).

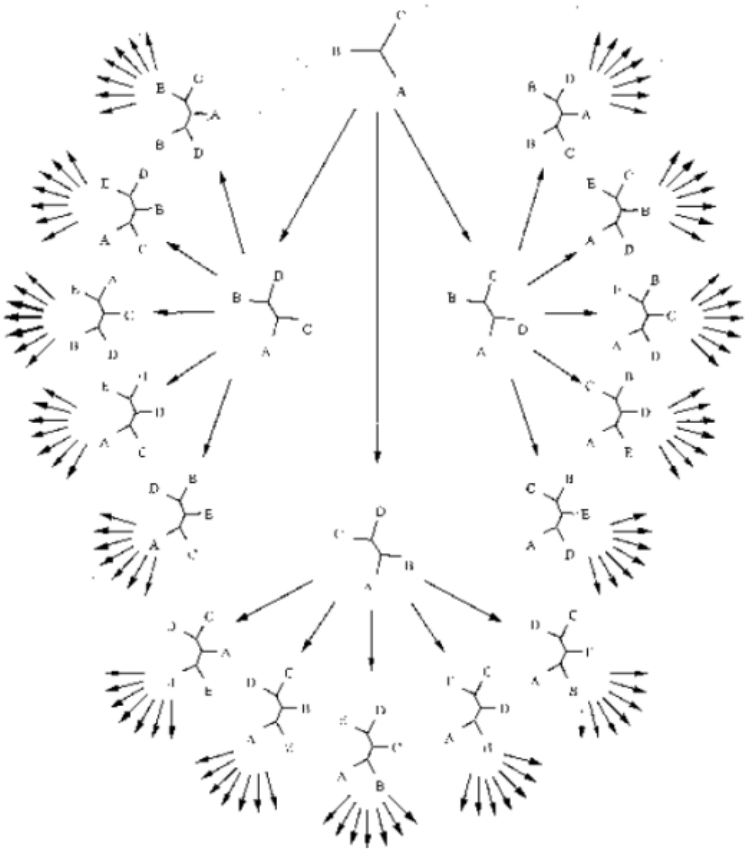


Imagen 5: Algoritmo de construcción de todos los árboles posibles.

b) **Branch and bound**: Es idéntico a la búsqueda exhaustiva, pero hay un umbral de longitud definido. Cada vez que se añade un taxón, se evalúa la longitud de los árboles (incompletos) resultantes. Aquellos que superen el umbral son descartados (y con ellos todos sus "derivados"). Esto reduce enormemente el número de árboles que serán evaluados, y por lo tanto permite realizar análisis imposibles de realizar con el método exhaustivo. Además, sigue garantizando

que se va a encontrar el árbol más corto. No obstante, para grandes cantidades de taxones sigue siendo inaplicable.

c) **Búsqueda heurística**: Hay varios subtipos. El algoritmo llamado **stepwise addition** es idéntico al branch and bound, pero cada vez que añade un taxón evalúa los árboles y sigue únicamente con el más corto. El **star decomposition** es más complejo, ya que implica colapsar el árbol en cada paso y realizar una serie de operaciones complicadas mientras lo va "desplegando". Al mismo tiempo que va realizando esta tarea, en cada paso prueba a reorganizar el árbol y medir la longitud del resultado, ya que esto le permite detectar posibles árboles de longitud menor que se le pueden haber pasado por alto. El algoritmo de reorganización, o *brach-swapping*, más conocido y fiable es el llamado **tree bisection and reconnection (TBR)**. Este tipo de búsqueda es la más rápida, pero no garantiza encontrar el árbol más corto: a pesar del *branch-swapping* puede llegar a no detectar resultados más óptimos que el obtenido.

ÁRBOLES CONSENSO

Cuando un análisis (ya sea de Máxima Parsimonia, Inferencia Bayesiana u otro tipo) nos da una serie de árboles como resultado, es necesario crear un **árbol consenso**. Esto puede suceder, por ejemplo, si una búsqueda de Máxima Parsimonia encuentra que, en vez de haber un árbol mínimo, tenemos una serie de árboles que comparten el valor mínimo de longitud.

Hay varios métodos de construirlo. Algunos de los más frecuentes son:

- **Strict**: El árbol consenso sólo muestra los nodos que aparecen en todos los árboles.
- **Majority Rule**: El árbol consenso solo muestra los nodos que aparecen en al menos un tanto por ciento predefinido de los árboles.

MÁXIMA VEROSIMILITUD

BASES TEÓRICAS

Es un método estadístico, es decir, se basa en el cálculo de probabilidades. En particular, calcula la probabilidad de obtener la matriz de datos (D) dada una determinada hipótesis (H: el árbol y el modelo de sustitución nucleotídica). En otras palabras: ¿cuál es la probabilidad de que, habiendo obtenido un determinado árbol, éste sea el resultado de nuestros datos? En idioma matemático: P(D|H).

Hay que tener en cuenta que la verosimilitud (o *likelihood*) no es la probabilidad de que un árbol sea correcto en términos absolutos, sino de que se ajuste a nuestros datos.

CÁLCULO DE LA VEROSIMILITUD

Veámoslo con un ejemplo. Supongamos que queremos calcular la probabilidad de que la secuencia CCAT haya dado lugar a la secuencia CCGT. Los parámetros de nuestro modelo son los siguientes:

$$P_i = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix} \qquad f = [0.1 \quad 0.4 \quad 0.2 \quad 0.3]$$

Entonces vamos posición por posición calculando la probabilidad de que el nucleótido de la secuencia original haya dado lugar al nucleótido de la secuencia final. En el caso del primero, es un cambio C → C. La probabilidad sería igual a la frecuencia de C por la tasa de sustitución de C a C. Luego multiplicamos la probabilidad para cada una de las posiciones y ese es el valor de verosimilitud.

$$L = (\pi_C \cdot P_{C \to C}) \cdot (\pi_C \cdot P_{C \to C}) \cdot (\pi_A \cdot P_{A \to G}) \cdot (\pi_T \cdot P_{T \to T}) = 0.00003$$

Cuando queremos calcular la verosimilitud de un árbol, creamos todas las posibles combinaciones de estados ancestrales (nucleótidos en los nodos internos). En cada una de ellas calculamos la verosimilitud de cada rama, y luego las multiplicamos para obtener la verosimilitud de esa combinación. Por último, multiplicamos las verosimilitudes de todas las combinaciones para obtener la verosimilitud del árbol.

El resultado final suele ser un número ridículamente pequeño: un cero seguido de una coma, luego miles y miles de ceros y al final diversos decimales. Los ordenadores no son capaces de almacenar números tan pequeños en su memoria, por lo que la verosimilitud suele calcularse en logaritmos neperianos. Si tomamos la verosimilitud obtenida en el ejemplo de arriba:

$$L = 0.00003 \qquad lnL = -10.41431 \ldots$$

Cuanto más pequeño sea el valor de lnL (más cercano a cero), mayor será la verosimilitud.

BÚSQUEDA DE LA MÁXIMA VEROSIMILITUD

Al igual que la Máxima Parsimonia buscaba el árbol más corto, la Máxima Verosimilitud busca el árbol con mayor verosimilitud.

Lo ideal sería calcular la verosimilitud de cada árbol, pero tenemos el mismo problema que con el método anterior: es imposible para más de 10 taxones.

La búsqueda del árbol con mayor verosimilitud sigue un algoritmo de **hill-climbing**: Comienza con un árbol generado, por ejemplo, mediante Neighbor-Joining, y se calcula su verosimilitud. Entonces se van introduciendo cambios en el árbol (se cambia una

rama de sitio, se modifica algún parámetro…) y se mide la verosimilitud del árbol modificado: si es menor que la del árbol original, se descarta la modificación; si es mayor, el árbol modificado pasa a ser el punto de partida para la siguiente modificación. Así, poco a poco, se va "ascendiendo" hasta que se llega a un punto en el que, por más que se modifique el árbol, la verosimilitud no aumenta más. Ese es el **punto de Máxima Verosimilitud**, y el árbol que tenemos es el más probable.

Es un método que se resuelve sorprendentemente rápido en cualquier ordenador actual. De hecho suele durar poco más que un Neighbor-Joining y además es más fiable.

Como inconvenientes podemos destacar que es muy sensible al **fenómeno de atracción de ramas cortas**, completamente opuesto al problema que teníamos con la Máxima Parsimonia. Además sólo considera el árbol más probable: si hay otro casi igual de probable, lo descarta, cuando en realidad debería considerar ambas alternativas. Este último problema lo resuelve la Inferencia Bayesiana, que veremos a continuación.

APLICACIONES DE LA MÁXIMA VEROSIMILITUD

El hecho de que el valor de verosimilitud dependa de los valores de los parámetros de partida hace que este análisis sea útil para poner a prueba diversas hipótesis. Algunas de sus aplicaciones son:

- **Test de modelos**: Hay programas que calculan cuál es el modelo de sustitución nucleotídica más adecuado para nuestra matriz de datos. Para ello hacen un análisis rápido de verosimilitud utilizando cada uno de los modelos: el más adecuado será el que se ha usado en el análisis que ha dado un mayor valor de verosimilitud.
- **Test de topologías**: Si un análisis nos da como resultado un árbol inesperado (las relaciones entre los organismos no son las que esperábamos) podemos repetir el análisis forzando una topología, es decir, obligando a que el árbol generado deba ceñirse a nuestra hipótesis inicial. Luego comparamos los valores de verosimilitud de ambos análisis. Siempre saldrá mayor el del análisis no constreñido, así que para tomar una decisión debemos fijarnos en si la diferencia entre ambos valores es significativa o no. Si lo es, debemos descartar nuestra hipótesis inicial y aceptar los resultados inesperados que hemos obtenido; si no lo es, podemos decir que no podemos descartar nuestra hipótesis inicial.

UN INCISO: ALINEAMIENTOS

A la hora de alinear secuencias para crear nuestra matriz podemos recurrir a diversos tipos de métodos:

a) Métodos de alineamiento clásicos (Clustal, **ClustalW**, T-Coffee): Crean un "árbol guía" con las secuencias a lo bruto y empiezan a construir el alineamiento

añadiendo secuencia tras secuencia según el orden en el que aparecen en ese árbol. Se definen unos valores de "coste de apertura de gap" y "coste de extensión de gap" que modifican el comportamiento del algoritmo cuando tiene que decidir si introduce gaps o no. Son muy rápidos y funcionan perfectamente en fragmentos codificantes, pero suelen dar lugar a errores con cualquier otro tipo de fragmento.

b) Métodos iterativos (PRRN/PRRP, DIALIGN, **MUSCLE**): Son idénticos a los anteriores, pero cada vez que añaden una secuencia al alineamiento echan además una ojeada a las secuencias "en bruto" para detectar posibles artefactos, y van calculando las distancias entre las secuencias como método de optimización.

c) Métodos guiados por filogenias (PRANK, PAGAN): Al mismo tiempo que construye el alineamiento va creando una filogenia rápida (normalmente Máxima Verosimilitud) en cada paso, usándola para ir optimizando el proceso.

OTRO INCISO: ADICIONES A LOS MODELOS (INVARIANTES Y GAMMA)

Los modelos de sustitución nucleotídica pueden complementarse con otros dos parámetros.

El parámetro **invariants** (I) añade información acerca de si hay una serie de nucleótidos que permanecen siempre invariables, es decir, que no cambian.

El parámetro **gamma** (G) informa de cuál es la distribución que sigue la tasa de variabilidad en los nucleótidos variables. Siempre sigue una distribución gamma, la cual tiene un parámetro (normalmente llamado **α** en filogenia) que determina la forma de la curva. Por ejemplo, con un valor α=1 hay muchas posiciones que no varían y pocas que varían mucho; con un valor α=2 hay mucha heterogeneidad; con un valor α=9 la mayoría de las posiciones varían a una velocidad medio-alta…

INFERENCIA BAYESIANA

BASE TEÓRICA

Trata de calcular la probabilidad de una cierta hipótesis (H, el árbol) dados unos ciertos datos de partida (D, la matriz). Es exactamente lo contrario que la verosimilitud, y se calcula mediante el **Teorema de Bayes**:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Es un cálculo de **probabilidad posterior**, mientras que la verosimilitud era probabilidad anterior. Podemos ver la diferencia con un ejemplo sencillo:

- Probabilidad anterior (verosimilitud): Tenemos una urna con dos bolas verdes y tres azules, ¿cuál es la probabilidad de sacar una bola azul?
- Probabilidad posterior (bayesiano): Tenemos una urna y sacamos una bola azul, ¿cuál es la probabilidad de que hubiera dentro dos bolas verdes y tres azules?

Aplicando la fórmula a las filogenias podemos ver a qué equivale cada elemento de la ecuación:

- P(D|H): Es la verosimilitud del árbol, sin más historias. Ya hemos visto antes cómo se calcula.
- P(H): Equivale a la probabilidad de un árbol entre el total de árboles posibles. Normalmente se considera que equivale a $1/N_{árboles}$, aunque en realidad algunos expertos argumentan que eso no es cierto.
- P(D): Es la probabilidad de la matriz, es decir, de que en cada una de las posiciones para cada uno de los taxones encontremos el nucleótido que vemos y no otro. Este valor es absolutamente imposible de calcular. Para solucionar este problema, el algoritmo de Inferencia Bayesiana recurre a un truco que veremos a continuación.

ALGORITMO DE INFERENCIA BAYESIANA

Utiliza las llamadas **cadenas de Markov Monte Carlo**. Este algoritmo es equivalente a "dar un paseo" por un espacio multidimensional formado por todos los árboles posibles. Se empieza en un punto al azar de dicho espacio y se mira qué árbol hay en ese lugar ($T_{anterior}$). Luego damos un "paso", saltando a un lugar cercano (modificando un poco alguno de los parámetros del árbol) y miramos qué árbol tenemos allí (T_{nuevo}). Entonces calculamos el *ratio* entre las probabilidades bayesianas de ambos:

$$R = \frac{Bayes(T_{nuevo})}{Bayes(T_{anterior})}$$

Si desarrollamos la ecuación de Bayes para cada elemento:

$$R = \frac{\frac{P(H_{nuevo}) \cdot P(D|H_{nuevo})}{P(D)}}{\frac{P(H_{anterior}) \cdot P(D|H_{anterior})}{P(D)}}$$

Y aquí viene el truco: Por un lado tenemos el elemento incalculable, P(D), tanto en el numerador como en el denominador, por lo que podemos eliminarlo y nos quitamos el principal escollo de en medio:

$$R = \frac{P(H_{nuevo}) \cdot P(D|H_{nuevo})}{P(H_{anterior}) \cdot P(D|H_{anterior})}$$

Además, ¿no habíamos quedado en que la probabilidad de cada árbol, P(H), es igual en todos los casos? Entonces podemos quitarlas también, y… ¡sorpresa!

$$R = \frac{P(D|H_{nuevo})}{P(D|H_{anterior})}$$

El *ratio* entre las probabilidades bayesianas queda simplificado al *ratio* entre la verosimilitud de cada árbol. En definitiva, hacer Inferencia Bayesiana equivale a hacer Máxima Verosimilitud de una forma más refinada.

Si R > 1, descartamos el árbol inicial y seguimos "paseando" a partir del nuevo árbol. Si R < 1, descartamos el árbol nuevo y volvemos al punto en el que nos encontrábamos.

El "paseo" se realiza de forma paralela por cuatro lados (las llamadas **cadenas**). La cadena que está encontrando los árboles con mayor verosimilitud se llama **cadena fría**, y explora árboles situados en un espacio muy cercano al lugar en el que se encuentra. Las otras tres cadenas son **cadenas calientes**, y se dedican a explorar lugares más alejados por si acaso encuentran árboles con mayor verosimilitud. Si una cadena caliente encuentra un árbol mejor que la cadena fría, se "enfría" y coge el testigo, mientras que la cadena fría se "calienta".

Sin embargo, el objetivo de las cadenas no es encontrar el árbol de Máxima Verosimilitud. Lo que ocurre es lo siguiente: cada cierto número de "pasos", la cadena fría toma el árbol en el que se encuentra y lo añade a un recopilatorio de árboles. Una vez acabado el "paseo", se eliminan (**burn-in**) los primeros árboles, obtenidos cuando las cadenas aún andaban explorando el espacio en busca de lugares óptimos, y con los restantes se crea un árbol consenso.

PROBABILIDAD POSTERIOR

El análisis de Inferencia Bayesiana es lento y requiere una considerable potencia de cálculo. Un análisis típico dura horas, o incluso varios días. Por lo tanto no es factible realizar un test de *bootstrap* para calcular la fiabilidad de los nodos: ¡un análisis normalillo de 8 horas necesitaría 9 años para hacer las 10.000 réplicas aconsejables!

Sin embargo, la Inferencia Bayesiana se encarga de calcular, al mismo tiempo que computa el árbol consenso, la **probabilidad posterior de cada nodo**, y este valor es el que se usa como indicador de soporte.

Al ser de naturaleza distinta al *bootstrap*, no es comparable con éste. Además es más estricto: se considera que un valor de probabilidad posterior comienza a ser aceptable si es superior al 90%.

A pesar de todo, los valores de probabilidad posterior obtenidos con Inferencia Bayesiana son más fiables que los valores de *bootstrap* obtenidos con Máxima Verosimilitud.

DATACIÓN

Hay diversos métodos para **calibrar** un árbol, de forma que la longitud de sus ramas sea proporcional al tiempo transcurrido entre los eventos de divergencia que las delimitan. Se aconseja, siempre que sea posible, utilizar todos los métodos que sean posibles.

FÓSILES Y BIOGEOGRAFÍA

Podemos usar fósiles y eventos biogeográficos para establecer las edades de las diferentes ramas y nodos. No obstante, según el tipo de evento se utiliza un método u otro para establecer su posible edad mínima o máxima:

Método	Base	Cálculo
Fósil	Conocemos un fósil correspondiente al ancestro de uno de los clados.	**Lower bound**: La edad de la rama de la que surge el clado debe ser superior al momento temporal en el que se sitúa el fósil.
Vicarianza	Dos clados están separados geográficamente, y sabemos qué evento produjo su separación.	**Fixed time**: La edad del nodo estará alrededor del momento en el que se produjo la separación.
Archipiélago	Sabemos que los clados han surgido por colonización de nuevas islas conforme éstas iban creándose.	**Upper bound**: La edad del nodo será inferior al momento en el que apareció la isla a la que migró uno de los clados.

RELOJ MOLECULAR

Si desconocemos cualquier fósil o evento biogeográfico, podemos emplear la **tasa de sustitución** del fragmento que estamos usando (la cual viene dada normalmente en sustituciones por sitio por millón de años) para calcular la edad de cada nodo. Este método debe ser usado con cuidado y se debe comprobar concienzudamente si las edades obtenidas tienen sentido.

Hay dos tipos de aproximaciones para calcular las edades:

- **Reloj estricto**: Considera que la tasa es la misma para todas las ramas e invariable. Esto no se ajusta a la realidad, pero hace que el análisis sea más rápido y tenga mayor soporte al considerar menos parámetros.

- **Reloj relajado**: Permite que la tasa varíe ligeramente a lo largo de las ramas y el tiempo. Es más realista pero el análisis requiere mucho más tiempo y pierde fiabilidad.

CONSEJOS PARA EL USO DE LOS PROGRAMAS FILOGENÉTICOS

Sólo uno:

1. **LEED LOS MANUALES**

Repito:

1. **LEED LOS MANUALES**

APÉNDICE 1: CAUSAS DE ERRORES EN FILOGENIAS (SEGÚN ARABI ET AL. 2010)

Elección del ougroup: Un árbol mal enraizado puede llevar a una interpretación errónea de la filogenia al no poder situar correctamente el punto de corte entre outgroup e ingroup. Además puede crear problemas de long-branch attraction.

> Cómo evitarlo: Incluyendo tres outgroups con posición parafilética con respecto al ingroup.

Muestras mal identificadas y contaminación: Errores de secuenciación, identificación incorrecta de la muestra (sobre todo en grupos poco estudiados), contaminación entre muestras al manipularlas, secuencias erróneas en las bases de datos, secuencias sin voucher…

> Cómo evitarlo: Si analizas una matriz concatenada, analiza además separadamente cada gen para detectar quimeras. Al publicar las secuencias, mantenlas enlazadas a su voucher. Usa secuencias del BOLD (Barcode of Life Data System).

Missing data: Se han hecho simulaciones que han dado indicios de que no es un problema mientras haya suficientes caracteres. El problema son las secuencias de baja calidad, que introducen ruido en el análisis reduciendo la resolución e incorporando errores en las secuencias que afectan a la reconstrucción filogenética.

> Cómo evitarlo: Es mejor carecer de algunas secuencias que incorporar secuencias “en mal estado”. Si secuencias varios fragmentos y de una secuencia sólo obtienes uno o dos, descártalos (probablemente se deban a contaminación).

El alineamiento: Hay fragmentos, especialmente los rRNA, que presentan una curiosa relación entre su secuencia y su estructura secundaria, con loops que surgen y desaparecen a lo largo de la evolución y que afectan mucho al alineamiento. Y un alineamiento mal hecho lleva a una reconstrucción filogenética errónea.

> Cómo evitarlo: No usando los métodos tradicionales (primero alineamiento, luego filogenia), sino los nuevos métodos que alinean y reconstruyen el árbol al mismo tiempo (enfoque dinámico).

Orientación de la cadena: El genoma mitocondrial presenta dos cadenas: una positiva con más contenido en AC y otra negativa rica en GT. De vez en cuando un determinado fragmento puede "voltearse", con lo que las proporciones de nucleótidos en cada cadena se han invertido. Entonces se inicia un proceso en el que se tiende a volver a la situación original, lo cual altera las tasas de sustitución nucleotídicas en el linaje en el que se ha dado la inversión.

> Cómo evitarlo: Midiendo las proporciones de los nucleótidos en cada cadena, sobre todo las de la tercera posición de cada codón, para detectar aquellos linajes en los que se ha producido la reorganización.

El marcador: Hay fragmentos que proporcionan más información filogenética que otros. ¿Cuáles? Pues depende del grupo que estemos estudiando. Generalmente en artrópodos el 18S y el 28S tienen mala fama (aunque no en todos los casos). También hay marcadores que "van por su cuenta" y dan resultados que no tienen nada que ver con la realidad, como la histona H3 en algunos casos.

> Cómo evitarlo: Bibliografía, bibliografía… Busca a ver si alguien ha usado ese marcador en tu grupo o uno cercano y mira si le ha salido bien.

APÉNDICE 2: DIRECCIONES ÚLTILES

Descarga de PHYLIP: http://evolution.genetics.washington.edu/phylip.html

ALTER, fantástica aplicación para convertir entre los diferentes formatos de archivos filogenéticos; ideal para solucionar problemas de compatibilidades: http://sing.ei.uvigo.es/ALTER/

Recopilatorio casi completo de todos los programas de filogenia que existen, recopilados por Felsenstein: http://evolution.gs.washington.edu/phylip/software.html

Páginas oficiales de los programas que hemos usado:

Geneious: http://geneious.com/

MEGA: http://www.megasoftware.net/

PAUP: http://paup.csit.fsu.edu/

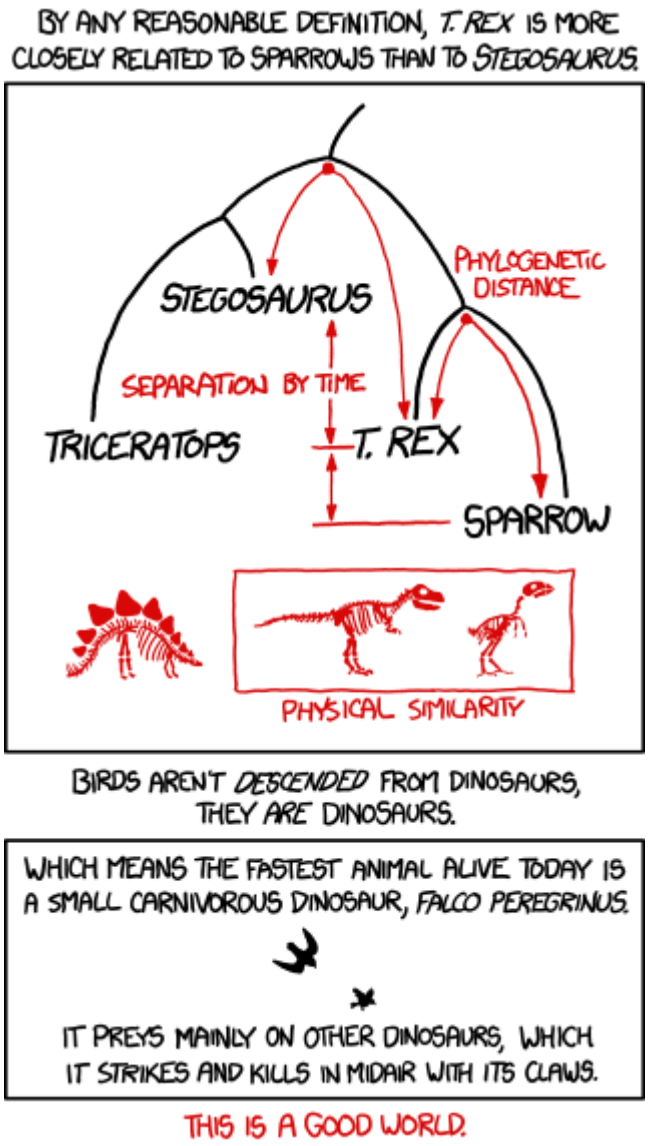
PaupUp: http://www.agro-montpellier.fr/sppe/Recherche/JFM/PaupUp/main.htm

RAxML: https://github.com/stamatak/standard-RAxML

MRBAYES: http://mrbayes.sourceforge.net/

BEAST: http://beast.bio.ed.ac.uk/Main_Page

jMODELTEST: https://code.google.com/p/jmodeltest2/



Crédito: Randall Munroe, XKCD (http://xkcd.com/1211/)

Disclaimer: Este texto puede contener errores, algunos de ellos de bulto, ya que el autor aún no es experto en filogenia (tiene que meter la pata muchas más veces para llegar a serlo).