Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Intoducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS 2021
https://github.com/vinuesa/intro-filoinfo_UAS

Introducción a la Filoinformática: Pan-genómica y Filogenómica – Diplomado en Bioinformática, Univ. Autónoma de Sinaloa, Oct. 2021
http://econtinua.uas.edu.mx/diplomados/2510-2700-18_001.htm

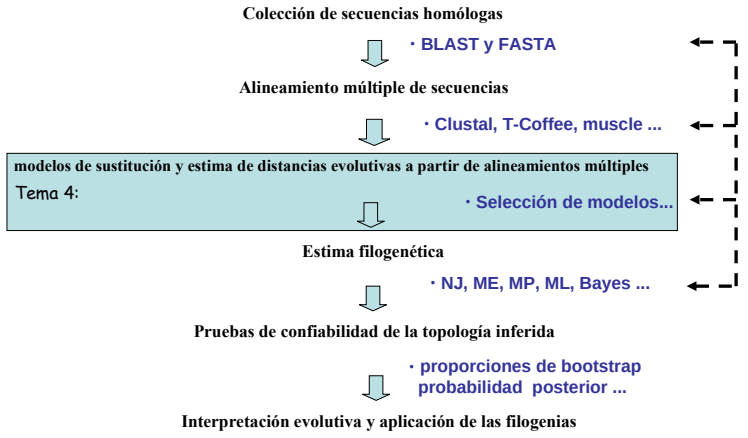
Pablo Vinuesa (vinuesa[at]ccg . unam . mx; @pvinmex)
Centro de Ciencias Genómicas, (CCG-UNAM), Campus Morelos, Cuernavaca, México **http://www.ccg.unam.mx/~vinuesa/**

Todo el material del curso (presentaciones, tutorales y datos) lo encontrarás en:
https://github.com/vinuesa/intro-filoinfo_UAS

- **Tema 4: Introducción a la filogenética y modelos de evolución de secuencias**

1. Para qué sirven los modelos en ciencia
2. Modelos paramétricos vs. empíricos de evolución de secuencias
3. Parametrización de los modelos de sustitución de DNA
4. Condiciones (supuestos) de aplicabilidad de los modelos
5. Modelos de sustitución y distancias evolutivas
6. La familia GTR(+I+G) de modelos de sustitución para secuencias nucleotídicas
7. Bootstrapping: remuestreo con reemplazo de caracteres y soporte de biparticiones

Protocolo básico para un análisis filogenético de secuencias moleculares



Inferencia filogenética molecular – clasificacón de métodos

- Podemos clasificar a los métodos de reconstrucción filogenética en base a:

1 . el tipo de datos que emplean (**caracteres discretos vs. distancias**)

2. uso de un **método algorítmico o un criterio de optimización** para encontrar la topología

Tipo de datos

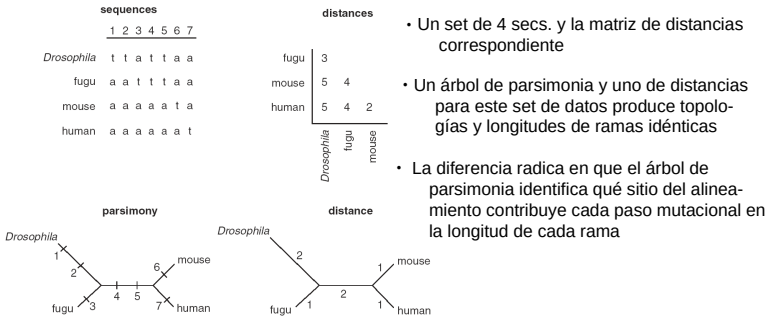
Método de reconstrucción	distancias	caracteres discretos
algoritmo de agrupamiento	UPGMA Neighbour joining	X
criterio de optimización	Evolución mínima	Máxima parsimonia Máxima verosimilitud

Métodos de reconstrucción filogenética – una clasificación

I.- Tipos de datos: distancias vs. caracteres discretos

- Los **métodos de distancia** primero convierten los alineamientos de secuencias en una matriz de distancias genéticas en base al modelo evolutivo seleccionado, la cual es usada por el método algorítmico de reconstrucción para calcular el árbol (**UPGMA y NJ**)

- Los **métodos discretos** (**MP, ML, Bayesianos**) consideran cada sitio del alineamiento (o una función probabilística para cada sitio) directamente



- Un set de 4 secs. y la matriz de distancias correspondiente
- Un árbol de parsimonia y uno de distancias para este set de datos produce topologías y longitudes de ramas idénticas
- La diferencia radica en que el árbol de parsimonia identifica qué sitio del alineamiento contribuye cada paso mutacional en la longitud de cada rama

Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Inferencia filogenética molecular – métodos basados en matrices de distancias

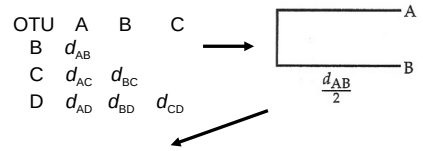
• Unweighted pair group method with arithmetic means (**UPGMA**)

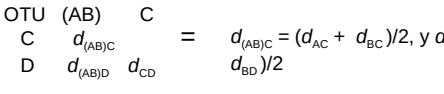
- este es uno de los pocos métodos que construye **árboles ultramétricos** (todas las hojas equidistantes de la raíz), es decir **asume un reloj molecular** perfecto a lo largo de toda la topología, lo que resulta en una **topología enraizada**.
  Además se obtienen las longitudes de rama simultáneamente con la topología

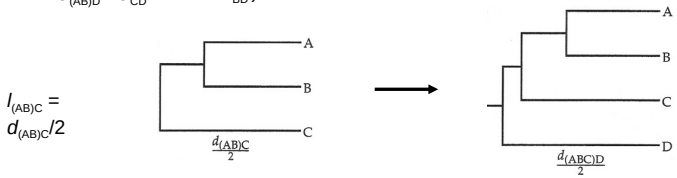
- se puede concebir como un método heurístico para encontrar la topología ultramétrica de mínimos cuadrados para una matriz de distancias pareadas

Inferencia filogenética molecular – métodos basados en matrices de distancias

• Unweighted pair group method with arithmetic means (**UPGMA**)







• UPGMA, por construir un **árbol ultramétrico**, resulta en una **topología enraizada**.
Además se obtienen las longitudes de rama simultáneamente con la topología

Ejercicio:

Calcula una matriz de distancias pareadas en base al número observado de diferencias entre OTUs, y en base a ella dibuja un árbol de UPGMA, indicando las longitudEes de cada rama

1. **Alineamiento:** No. sitios : 15; OTUs (taxa) = 4

```
Rhizobium       GGA GGG AGG AGG CCT
Agrobacterium   GGC GGG AGG AGG CCT
Sinorhizobium   GGG GGA AGG TGT CCG
Bradyrhizobium  GGT CGT AGC TGT GTG
```

2. **Matriz de distancias:** *d* : distancia (no. de diferencias observadas)

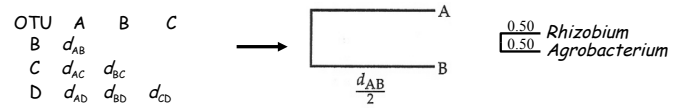
[	A	B	C	D]
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	

Inferencia de un árbol UPGMA usando el no. de dif. obs. como medida de la distancia genética entre OTUs

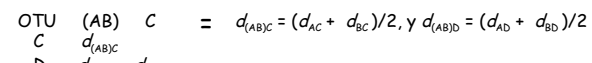
Matriz de distancias:

[	A	B	C	D]
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	

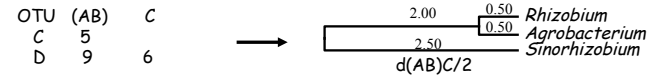
1.



2.



3.



Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Inferencia de un árbol UPGMA usando el no. de dif. obs. como medida de la distancia genética entre OTUs

Matriz de distancias:

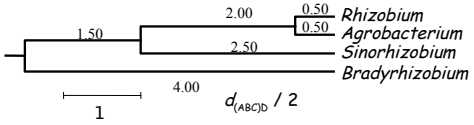
	A	B	C	D
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	

4. OTU (ABC) D

$d_{(ABC)D} = (d_{AD} + d_{BD} + d_{CD}) / 3$

$= d_{(AB)C} = (9 + 9 + 6) / 3 = 8$

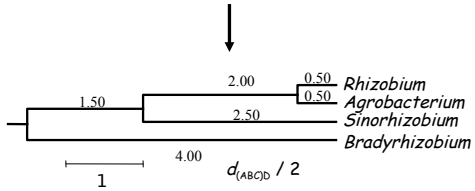
5.



Inferencia de un árbol UPGMA usando el no. de dif. obs. como medida de la distancia genética entre OTUs

Matriz de distancias:

	A	B	C	D
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	



- ¿Notan alguna inconsistencia entre las distancias topológicas y observadas?

- La distancia entre C y D no es aditiva y no queda adecuadamente reflejada en la correspondiente longitud de rama

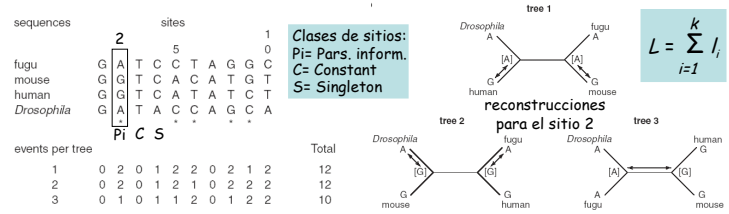
Métodos de reconstrucción filogenética – una clasificación

III. máxima parsimonia: dados dos árboles, se prefiere el que requiere menos cambios en estados de caracter

- El **método de máxima parsimonia** (**MP**) considera cada sitio filogenéticamente informativo (**Pi**) el alineamiento (al menos 2 pares de secuencias que compartan un polimorfismo distinto). Los sitios constantes (**C**) no son considerados y los singletones (**S**) no son Pars. informativos

- El supuesto teórico (modelo de evolución) implícito al método es que el árbol más verosímil es aquel que requiere el mínimo número de sustituciones para explicar los datos del alineamiento. **El criterio de optimización de la MP es el de cambio o evolución mínima.**

- Para cada sitio del alineamiento el objetivo es reconstruir su evolución bajo la constricción de **invocar el número mínimo de pasos evolutivos**. El número total de cambios evolutivos sobre un árbol de MP (longitud en pasos evolutivos del árbol) es simplemente la suma de cambios de estados de caracter (p. ej. sustituciones) de cada sitio variable



Métodos de búsqueda de árboles

I.- el problema del número de topologías

El número de topologías posibles incrementa factorialmente con cada nuevo taxon o secuencia que se añade al análisis

No. de árboles no enraizados = $(2n-5)!/2^{n-3}(n-3)$

No. de árboles enraizados = $(2n-3)!/2^{n-2}(n-2)$

Taxa	árboles no enraiz*.	árb. enraiz.
4	3	15
8	10,395	135,135
10	2,027,025	34,459,425
22	$3x10^{23}$	...
50	$3x10^{74}$	...

*por ej. para sólo 15 OTUs tenemos 213,458,046,676,875 topologías

- ¡ si pudiésemos evaluar $1x10^6$ topol./seg. necesitaríamos 6 años y 9 meses para completar la búsqueda! El no. de Avogadro es ~ $6 x10^{23}$ (átomos/mol). Según la teor. de la relatividad de la estructura del universo de Einstein, existen 10^{80} átomos de H_2 en el universo ...

http://en.wikipedia.org/wiki/Observable_universe

Por tanto se requieren **estrategias heurísticas de búsqueda** árboles cuando se emplean métodos basados en criterios de optimización y n > ~25

vinuesa{at}ccg{dot}unam{dot}mx
https://www.ccg.unam.mx/~vinuesa/

Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

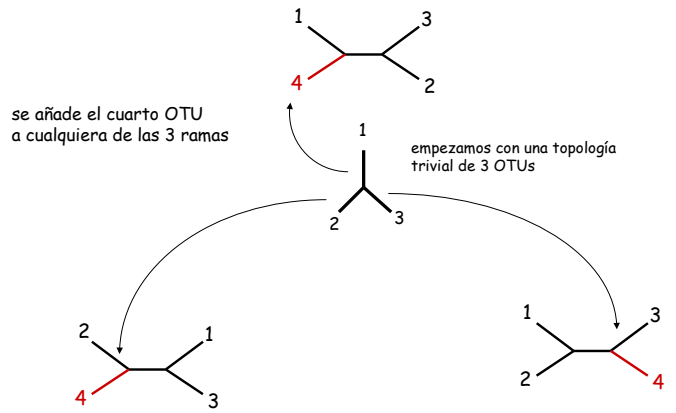
Métodos de búsqueda de árboles

- Pasos lógicos de los métodos filogenéticos basados en criterios de optimización (MP, ML y By)

1. definir el criterio de optimización (descrito formalmente en una **función objetiva**)
2. Construir un árbol de partida que contenga todos los OTUs
3. Emplar **algoritmos de búsqueda** que tratan de encontrar árboles mejores bajo el criterio de optimización escogido que el árbol actual o de partida.

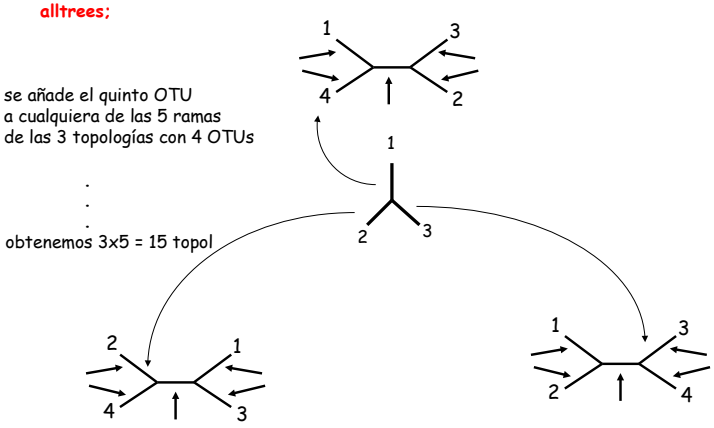
1. Criterios de optimización	2. Estrategias de búsqueda	
Parsimonia	Enumeración exhaustiva (n ≤ 12) (exhaustive enumeration)	**Métodos exactos**: garantizan encontrar la topología óptima
Máxima verosimilitud	Ramificación y límite (n ≤ 25) (branch-and-bound)	
Bayesiana	Decomposición en estrella (star decomposition)	**Métodos heurísticos**: no garantizan encontrar la topología óptima
	Adición secuencial (stepwise addition)	
	(Inter-)cambio de rama (branch swapping)	

Métodos de búsqueda de árboles -enumeración exhaustiva (n ≤ 12)

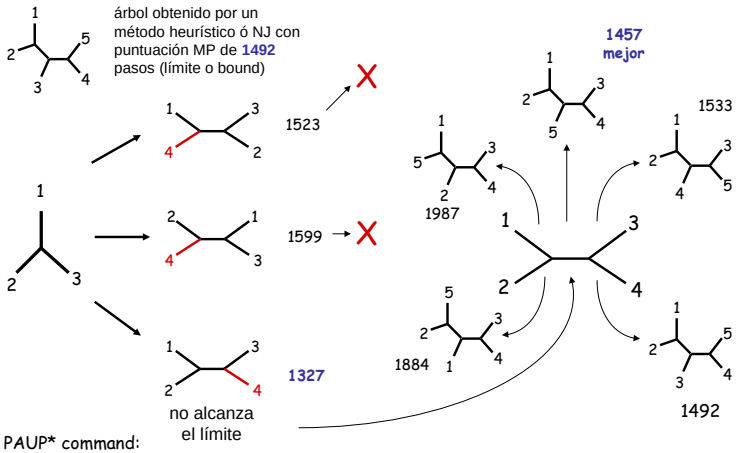


Métodos exactos de búsqueda de árboles -enumeración exhaustiva (n ≤ 12)

PAUP* command:
alltrees;



Métodos exactos de búsqueda de árboles - "branch and bound" (n ≤ 25)



- PAUP* command:
bandb;

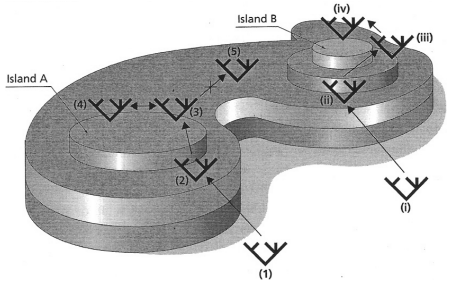
- Al igual que la búsqueda exhaustiva, **garantiza encontrar el árbol óptimo**

Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Métodos heurísticos de búsqueda de árboles - islas de árboles

- En la mayor parte de los análisis emplearán métodos heurísticos;
- éstos comienzan con un árbol (aleatorio, NJ o de adición secuencial) para realizar intercambios de ramas (**branch swappig**) sobre esta topología inicial con el propósito de encontrar topologías de mejor puntuación (según la func. de objetividad) que la de partida

- estos métodos heurísticos no garantizan encontrar la topología óptima pero trabajan muy bien cuando se comparan con sets de datos de ≤ 25 secs. analizados mediante B&B

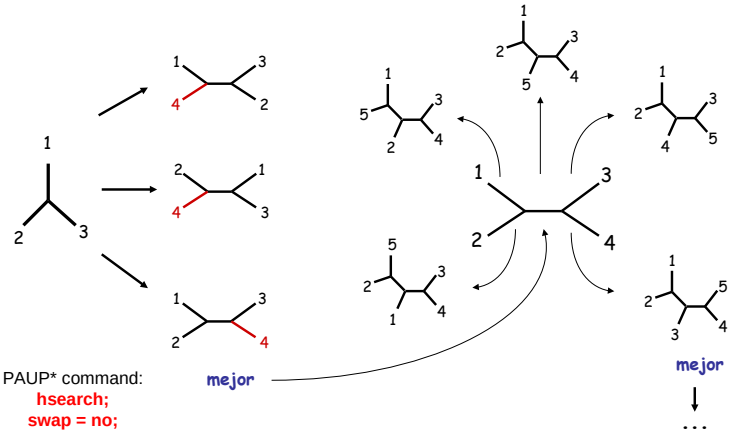


- El espacio de árboles puede visualizarse como un paisaje con colinas de diversas alturas; cada pico representa un máximo local de score o puntuación (**isla de árboles igualmente parsim.**)

- Es recomendable hacer múltiples búsqudeas heuríst. comenzando cada una desde una topología distinta para minimizar el riesgo de obtener un árbol ubicado en una isla topológica subóptima

Métodos heurísticos de búsqueda de árboles - adición secuencial (aleatorizada)

Este método se usa con frecuencia para generar distintos "**árboles semilla**" a partir de los cuales comenzar búsquedas heurísticas, partiendo de "distintos puntos del espacio de árboles



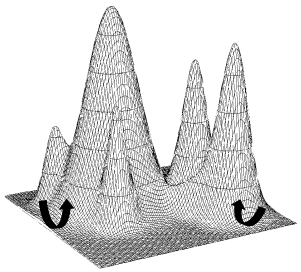
PAUP* command:
hsearch;
swap = no;

Métodos heurísticos de búsqueda de árboles - adición secuencial (aleatorizada)

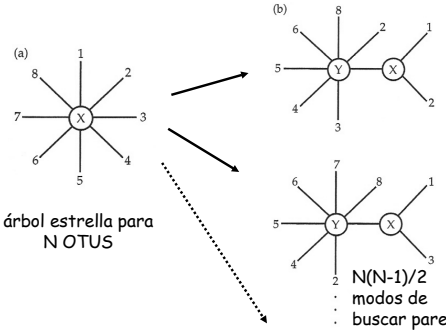
- El órden en el que se añaden los OTUs puede cambiar los resultados

- Por ello suele repetirse varias veces, añadiendo OTUs en cada ciclo de manera aleatorizada

- Sirven por lo tanto como **árboles semilla para iniciar distintas búsquedas heurísticas** partiendo de topologías potencialmente diferences para eficientizar la exploración del espacio de topologías (pero **no adecuados como hipótesis filogenética en sí mismos**)



Métodos heurísticos de búsqueda de árboles - decomposición de estrella



PAUP* command:
stardecomp;

- NJ usa este método junto al criterio de evolución mínima
- una vez que 2 OTUs han sido unidos ya no pueden ser desacoplados más adelante; en esto difiere del algoritmo de adición secuencial
- sensible al orden en que se van uniendo los OTUs; problema incrementa con el no. de OTUs
- no debe ser por tanto usado como método de búsqueda definitivo
- buena estrategia para producir árboles iniciales que sean mejorados mediante otras estrategias heurísticas

vinuesa{at}ccg{dot}unam{dot}mx
https://www.ccg.unam.mx/~vinuesa/

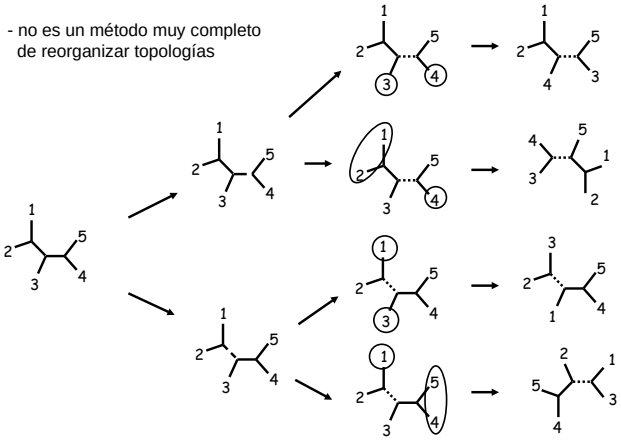
Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Intoducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS 2021
https://github.com/vinuesa/intro-filoinfo_UAS

Métodos heurísticos de búsqueda de árboles - intercambio de ramas (branch swapping)

- **Intercambio entre vecinos más próximos (Nearest Neighbor Interchange, NNI)**

- no es un método muy completo de reorganizar topologías



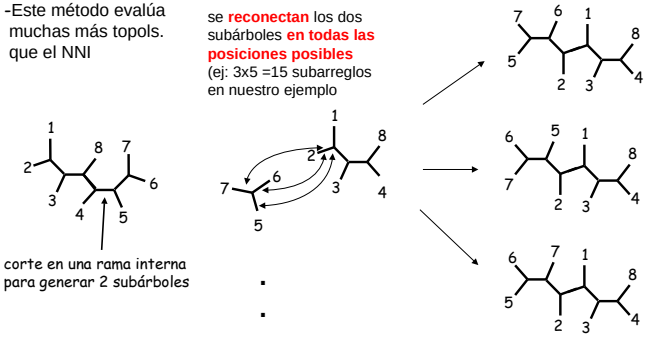
PAUP* cmmd:
hsearch swap=nni start=stepwise addseq=random;

Métodos heurísticos de búsqueda de árboles - intercambio de ramas (branch swapping)

- **Bisección-reconexión de árboles (Tree Bisection-Reconection, TBR)**

-Este método evalúa muchas más topols. que el NNI

se **reconectan** los dos subárboles **en todas las posiciones posibles** (ej: 3x5 =15 subarreglos en nuestro ejemplo



corte en una rama interna para generar 2 subárboles

se repite esta operación para reconectar el subárbol chico en las ramas terminales 1, 8, 4 y 3 del subárbol grande

PAUP* cmmd:
hsearch swap=tbr start=stepwise addseq=random;

Modelos de evolución de secuencias -introducción

- Para el análisis filogenético de secuencias alineadas virtualmente todos los métodos describen la evolución de las secuencias usando un modelo que consta de dos componentes:

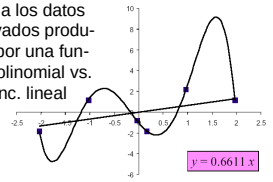
1. un árbol filogenético
2. una descripción de las probabilidades con las que se dan las sustituciones de aa o nts a lo largo de las ramas del árbol

- ¿Porqué necesitamos modelos y para qué sirven?

- Los modelos nos sirven para interpolar adecuadamente entre nuestras observaciones con el fin de poder hacer predicciones inteligentes sobre observaciones futuras

$y = -1.5972x^5 + 23.167x^4 - 126.18x^3 + 319.17x^2 - 369.22x + 155.67$

ajuste a los datos observados producidos por una función polinomial vs. una func. lineal

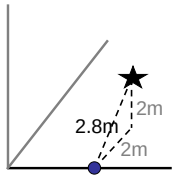


- **añadir parámetros** a un modelo generalmente mejora su ajuste a los datos observados
- **modelos infra-parametrizados** conducen a un pobre ajuste a los datos observados
- **modelos supra-parametrizados** conducen a una pobre predicción de eventos futuros
- existen métodos estadísticos para **seleccionar modelos ajustados** a cada set de datos

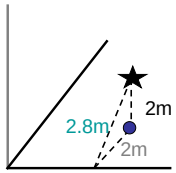
Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Modelos de evolución de secuencias -introducción

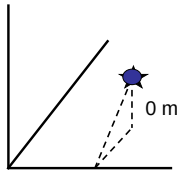
- **Dimensiones de un modelo**: cada parámetro en un modelo estadístico puede ser concebido como la adición de una nueva dimensión, tal y como se ilustra en el ejemplo siguiente:



- En este **modelo 1D** lo más cerca que podemos llegar al pto. marcado en un espacio 3D es 2.8 m



- En este **modelo 2D** lo más cerca que podemos llegar al pto. marcado en un espacio 3D es 2 m

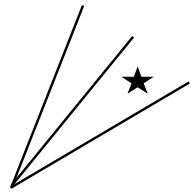


- En este **modelo 3D** podemos aproximar el punto exactamente.
  El modelo 3D se ajusta 100% a la realidad espacial.

- **En el caso de modelos de sustitución nunca obtendremos un ajuste del 100% entre el modelo y la realidad**. Todos los modelos son sólo aproximaciones de la realidad, pero algunos modelos son útiles para describir el proceso de sustitución (y otros mucho menos)

Modelos de evolución de secuencias -introducción

- **Dimensiones de un modelo**: en realidad, los **parámetros** de un modelo complejo **no** son siempre **independientes**, existiendo diversos grados de **colinearidad**. En el peor de los casos, dos parámetros pueden ser totalmente colineares, en cuyo caso uno de ellos es 100% redundante, por lo que no aporta nada a la fuerza del modelo para explicar los datos observados



- En este **modelo 3D** existe un nivel significativo de **colinearidad** entre sus dimensiones (o parámetros)

- uno de los objetivos primordiales de los modelos de sustitución de nt y aa es el de **incorporar los parámetros más relevantes, que expliquen características fundamentales de las secuencias** cuya evolución tratan de modelar de la manera más realista posible

Modelos de evolución de secuencias -introducción

- **Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales**

Corolario:

1. El grado de confianza que tengamos en una filogenia particular realmente depende de la que tengamos en el modelo subyacente

2. Por lo tanto, siempre que usemos un método basado en un modelo explícito de evolución (NJ, ML, By) es necesario usar rigurosas pruebas estadísticas para seleccionar el modelo y el valor de sus parámetros que mejor se ajusten a la matriz de datos a analizar

Modelos de evolución de secuencias -introducción

- **Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales**

- Existen dos aproximaciones para construir modelos de evolución de secuencias.

1. construcción de **modelos empíricos** basados en propiedades del proceso de sustitución calculadas a partir de comparaciones de un gran número de secuencias. Los modelos empíricos **resultan en valores fijos de los parámetros**, los cuales son estimados sólo una vez, suponiéndose que son adecuados para el análisis de otros sets de datos. Esto los hace fácil de usar e implementar en términos computacionales, pero su utilidad real para cada caso particular ha de ser evaluada críticamente

2. construcción de **modelos paramétricos** basado en el modelaje de propiedades químicas o genéticas del aas y nts. Los modelos paramétricos tienen la ventaja de que los **valores de los parámetros pueden ser derivados de cada set de datos** al hacer un análisis de los mismos usando métodos de ML o By, por tanto ajustándolos a cada matriz de datos particular

© Pablo Vinuesa 2021. @pvinmex
vinuesa{at}ccg{dot}unam{dot}mx
https://www.ccg.unam.mx/~vinuesa/

Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Modelos de evolución de secuencias -DNA

· Modelos de sustitución de nucleótidos

- El modelaje de la evolución a nivel del DNA se ha concentrado en la aproximación paramétrica. Se manejan **tres tipos principales de parámetros** en estos modelos:

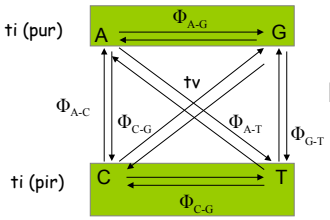
1. parámetros de **frecuencia**
2. parámetros de **tasas de intercambio**
3. parámetros de **heterogeneidad de tasas de sustitución** entre sitios

Modelos de evolución de sustitución de nucleótidos -modelos paramétricos

- **los diversos modelos evolutivos se distinguen por su grado de parametrización**

I. Frecuencias de nt : $\pi A = \pi C = \pi G = \pi T = 0.25$ ó $\pi A \neq \pi C \neq \pi G \neq \pi T$
- modelos de = frecuencia: JC69; K2P, K3P ...
- modelos de ≠ frecuencia: F81, HKY85, TrN93, GTR ...

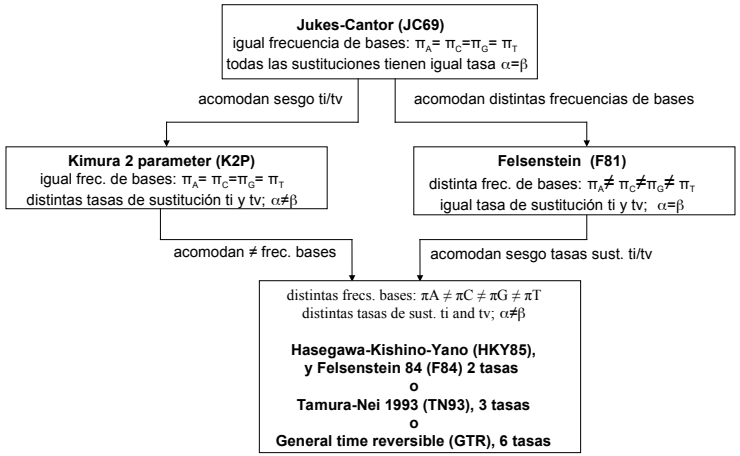
II. Tasas de sustitución transicionales/transversionales



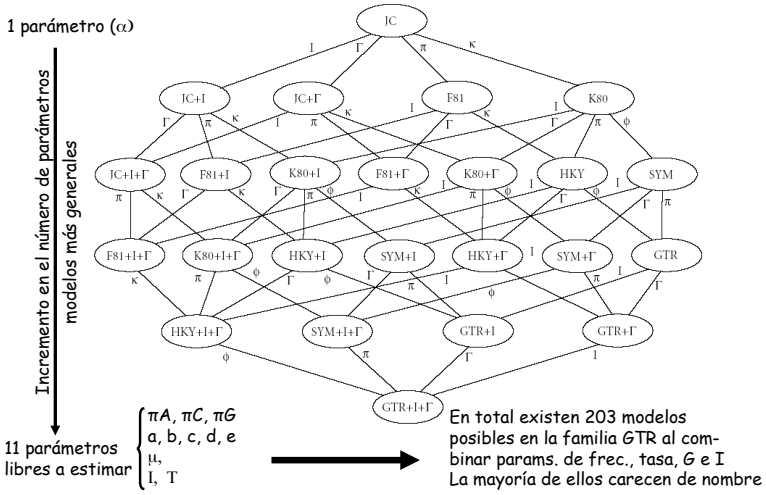
- Existen 4 tipos de sustituciones ti y 8 tv; cuando **ti/tv ≠ 0.5 existe un sesgo en sustituciones ti (o tv) en el set de datos. ti generalmente >> 1**
- los modelos evolutivos se diferencian también en la cantidad de parámetros que utilizan para acomodar diversas tasas de sustitución:

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

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV



Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV



Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Modelos de evolución de sustitución de nucleótidos -modelos paramétricos

Condiciones de aplicabilidad de los modelos (supuestos)

1.- **Supuesto de independencia**: las mutaciones en un sito no afectan a otros en la secuencia. Violado por ej. en el caso de rRNAs suelen seleccionarse mutaciones compensatorias (evolución covariariada entre sitios)

2.- **Supuesto de homogeneidad de tasas de sustitución a lo largo del tiempo y entre linajes**: en este supuesto se basa el reloj molecular y de su cumplimiento depende la posibilidad de poder utilizar un "**reloj molecular**" para datar clados

3.- **Las frecuencias de nucleótidos son homogéneas entre linajes**: este supuesto es frecuentemente violado cuando usamos secuencias de linajes muy distantes, particularmente en procariontes, ya que los contenidos de G+C de distintos grupos microbianos varía mucho, del 22 % (*Wigglesworthia, gamma-Proteobacteria*) al 75 % (*Anaeromyxobacter, delta-Proteobacteria)*

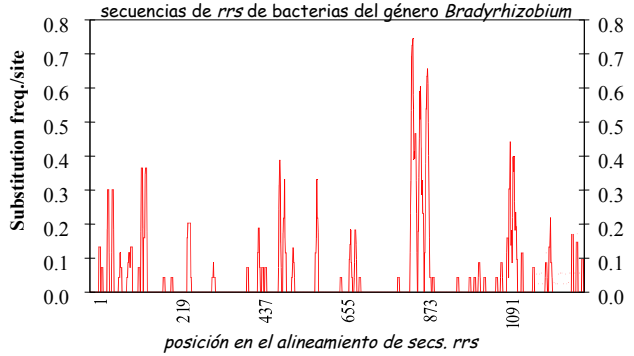
4.- **Las probabilidades de sustitución son las mismas para cada sitio**: este supuesto es violado casi sin excepción. Así por ejemplo, las 3as. pos. de los codones acumulan mutaciones mucho más rápidadmente que la 2a y 1a. Los distintos dominios de una proteína o rRNA también evolucionan con tasas distintas. **Distribución Gamma (Γ)**

Condiciones de aplicabilidad de los modelos (supuestos)

Acomodo de la heterogeneidad de tasas de sustitución entre sitios

(**I**) acomoda las posiciones invariables (**proporción de sitios invariantes**)
(**Γ**) acomoda la **heterogeneidad de tasas de sust.** entre las posiciones variables

Heterogeneidad de tasas de sust. a lo largo de un aln. múltiple de 29 secuencias de *rrs* de bacterias del género *Bradyrhizobium*

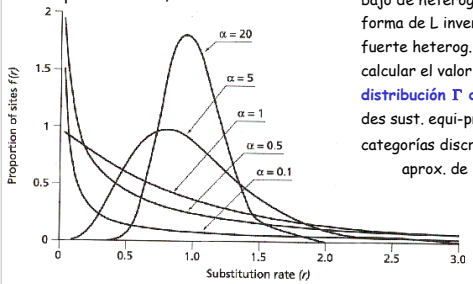


Condiciones de aplicabilidad de los modelos (supuestos)

2.- **Distribución gamma y heterogeneidad de tasas de sust. entre sitios**: Para modelar con cierto realismo el proceso de sustitución es esencial acomodar adecuadamente la heterogeneidad de tasas de sustitución entre sitios de un alineamiento

$$Pdf(r) = \alpha^{\alpha} r^{\alpha-1} / \exp(\alpha r)\, \Gamma(\alpha)$$

Diversas formas de la distribución gamma (Γ) para un rango de valores del parámetro $\alpha = 1/CV^2$, donde CV = coef. de var. de las tasas.
Así para CV = 0.3, $\alpha = 1/0.09 = 11.1111$



Para ello se asume generalmente una **distribución gamma (Γ)** de las tasas y que cada sitio tiene una tasa tomada aleatoriamente de dicha distribución, e independientemente de los demás sitios. El parámetro α controla la forma de la distribución. Para $\alpha > 1$ la distribución tiene forma de campana y modela un nivel bajo de heterogeneidad. Para $\alpha < 1$ la distribución toma forma de L invertida, des-cribiendo una situación de fuerte heterog. de tasas de sust. entre sitios. Para calcular el valor de α se emplea generalmente una **distribución Γ discreta** con un número c finito de tasas des sust. equi-probables (q_1, q_2 ..., q_c). El uso de 4 a 8 categorías discretas permite obtener una buena aprox. de la distrib. contínua.

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

- El método de momentos es de utilidad limitada en estadística (y filogenética) ya que no permite obtener una fórmula explícita para calcular la distancia entre secuencias usando modelos más complejos como el HKY85 o GTR

- La fórmula explícita de distancia para el modelo K2P es:

$$d = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

-este modelo tiene 2 parámetros, ***P*** y ***Q*** (proporción de *ti* y *tv* en que difieren 2 secuencias, donde $p = P + Q$

Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

- Comparación de los modelos de JC69 y K2P en su capacidad de corregir distancias observadas (*p*) entre pares de secuencias según su grado de divergencia

$$d_{JC69} = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right) \quad \text{vs.} \quad d_{K2P} = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right)+\frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

- Escenario I:

  - sean 2 secs. de long. = 200 nt, que difieren en 20 *ti* y 4 *tv*

  por lo tanto L = 200, *P* = 20/200 = 0.1 y *Q* = 4/200 = 0.02

p = 24/200 = 0.12

$d_{JC69} \approx 0.13$ (sust./sitio)

$d_{K2P} \approx 0.13$ (sust./sitio)

no. de sust. esperadas = 0.13 X 200 ≈ 26

no. de sust. esperadas = 0.13 X 200 ≈ 26

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

- Comparación de los modelos de JC69 y K2P en su capacidad de corregir distancias observadas (*p*) entre pares de secuencias según su grado de divergencia

$$d_{JC69} = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right) \quad \text{vs.} \quad d_{K2P} = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right)+\frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

- Escenario II:

  - sean 2 secs. de long. = 200 nt, que difieren en 50 *ti* y 16 *tv*

  por lo tanto L = 200, *P* = 50/200 = 0.25 y *Q* = 16/200 = 0.08

p = 66/200 = 0.33

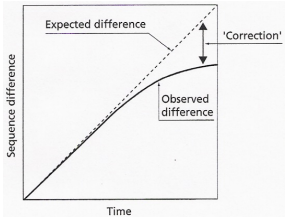
$d_{JC69} \approx 0.43$ (sust./sitio)

$d_{K2P} \approx 0.48$ (sust./sitio)

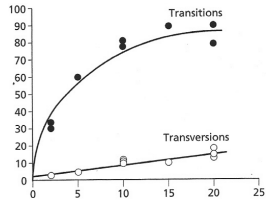
no. de sust. esperadas = 0.43 X 200 ≈ 86

no. de sust. esperadas = 0.48 X 200 ≈ 96

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV



- El objetivo de los modelos de sustitución es el de **compensar para los eventos homoplásicos de múltiples sustituciones**, y así obtener estimas de distancias evolutivas corregidas



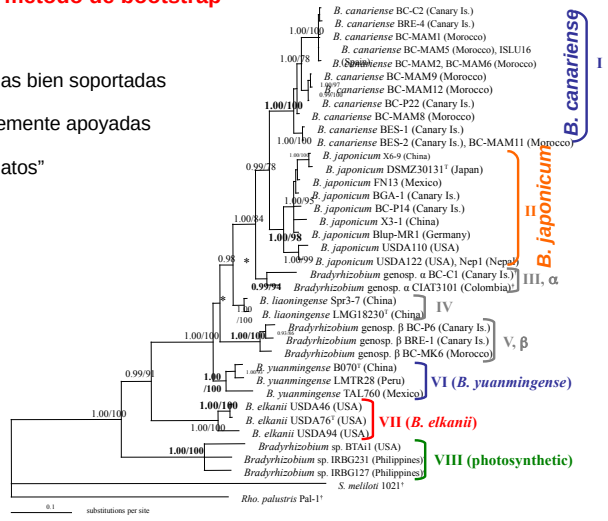
- El número de *ti* es generalmente > que el de *tv*, fenómeno que se acentúa cuanto mayor es la divergencia entre las secuencias a comparar. De ahí que en nuestro ejemplo las diferencias entre los escenarios I y II sólo se hicieron notar en el caso en el que la divergencia entre las secuencias era mayor (escenario II)

Estima de la confianza que podemos tener en distintas partes de una filogenia: el método de bootstrap

"Filogenias bien soportadas

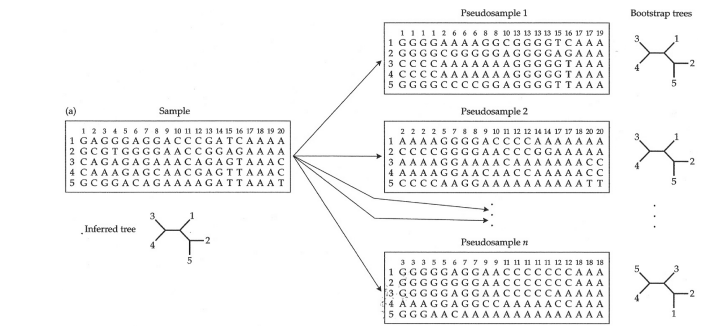
vs. pobremente apoyadas

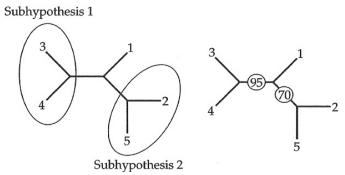
por los datos"



Tema 4: Métodos filogenéticos y modelos de sustitución nucleotídica

Estima del error de muestreo mediante el método de bootstrap





BOOTSTRAPPING

- generación de n (100-1000) **pseudoréplicas** (muestreo aleatorio con reemplazo)
2. estima de la filogenia para cada pseudo-réplica
3. cálculo de un **árbol consenso**
4. mapeo de las proporciones de bootstrap sobre la topología inferida de los datos originales

vinuesa{at}ccg{dot}unam{dot}mx
https://www.ccg.unam.mx/~vinuesa/