

ALINEAMIENTO DE SECUENCIAS

SEMANA 03

25 de Abril del 2020



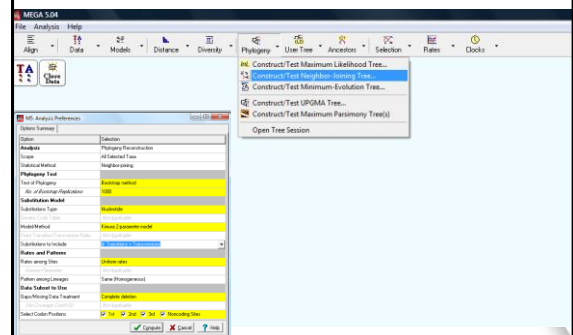
Índice

- Árboles Filogenéticos
- Máxima Parsimonia y Máximo Likelihood
- Practica 1: Clustal
- Practica 2: Mega

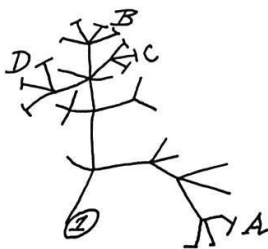


1 ALINEAMIENTO DE SECUENCIAS

Haciendo nuestro primer árbol



Pero qué es un árbol filogenético?

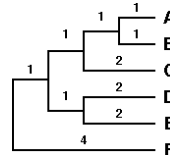


- Un árbol es una **hipótesis** de cómo se relacionan los individuos / grupos.
- Como toda hipótesis necesita ser evaluada.



¿Como agrupar a los individuos?

- **UPGMA:** Unweighted Pair Group Method with Arithmetic Mean



	A	B	C	D	E
A					
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

$$\begin{aligned} \text{dist}(A,B), C &= (\text{dist}AC + \text{dist}BC) / 2 = 4 \\ \text{dist}(A,B), D &= (\text{dist}AD + \text{dist}BD) / 2 = 6 \\ \text{dist}(A,B), E &= (\text{dist}AE + \text{dist}BE) / 2 = 6 \\ \text{dist}(A,B), F &= (\text{dist}AF + \text{dist}BF) / 2 = 8 \end{aligned}$$

http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/lecciones/neighbor_joining.html

¿Como agrupar a los individuos?

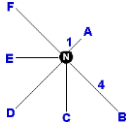
• Neighbour Joining

$$M_{ij} = d_{ij} - \frac{[r(i) + r(j)]}{N - 2}$$

$$S(iu) = \frac{d_{ij}}{2} + \frac{[r(i) - r(j)]}{2(N - 2)}$$

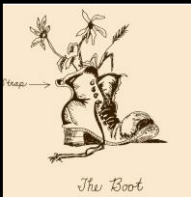
$$d(nk) = \frac{d(in) + d(jn) - d_{ij}}{2}$$

$$S(iu) = d(ij) - S(iu)$$



Editando los árboles

- Es necesario conocer el outgroup (Tpi).
- Se puede enraizar el árbol.
- Se puede modificar el ancho de las ramas, el color de los taxa.
- Aparecen dos árboles: Original tree y Bootstrap consensus.
- Modificar límite de *bootstrap*.



Bootstrap

"... raise oneself by one's own bootstraps..."

Taxa	Characters							
	1	2	3	4	5	6	7	8
A	R	R	Y	Y	Y	Y	Y	Y
B	R	R	Y	Y	Y	Y	Y	Y
C	Y	Y	Y	Y	Y	R	R	R
D	Y	Y	R	R	R	R	R	R
Outgp	R	R	R	R	R	R	R	R

Taxa	Characters							
	1	2	2	5	5	6	6	8
A	R	R	R	Y	Y	Y	Y	Y
B	R	R	R	Y	Y	Y	Y	Y
C	Y	Y	Y	Y	Y	R	R	R
D	Y	Y	Y	R	R	R	R	R
Outgp	R	R	R	R	R	R	R	R



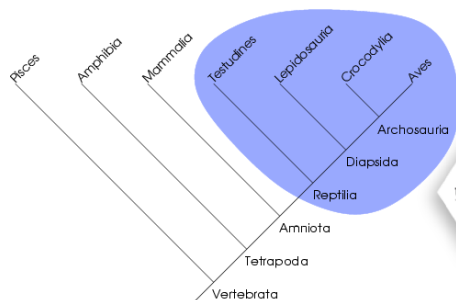
Randomly resample characters from the original data with replacement to build many bootstrap replicate data sets of the same size as the original - analyse each replicate data set

Describiendo el árbol

- Cuáles son los grupos más soportados por los valores de bootstrap?
- Cuántos grupos existen?
- Qué grupos están más relacionados?
- Todos los grupos están resueltos, o existen politomías?
- Los grupos encontrados tienen alguna relación con los lugares de muestreo, procedencia, etc.?
- Coinciden los grupos con los propuestos en el artículo?

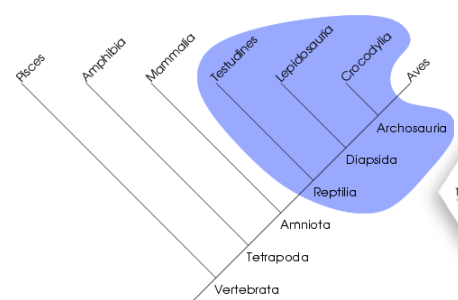
Monofilia

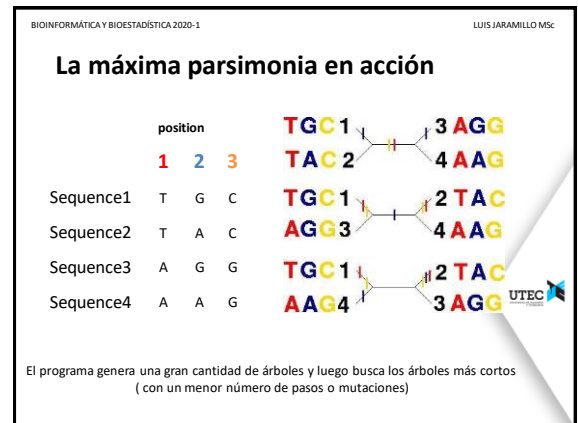
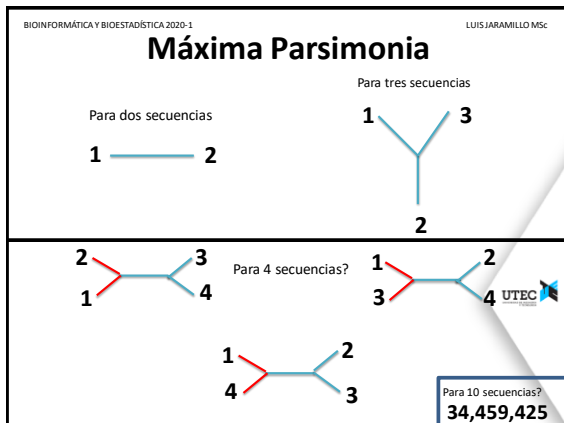
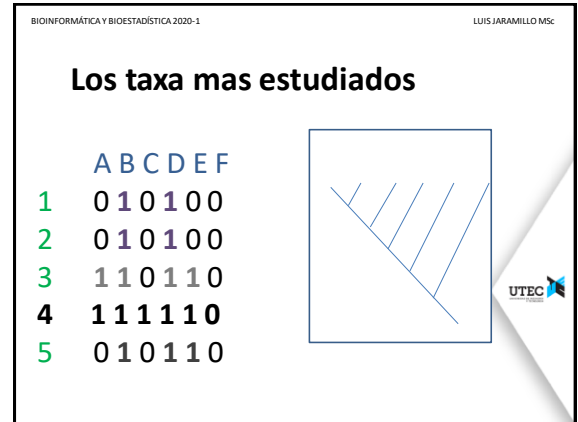
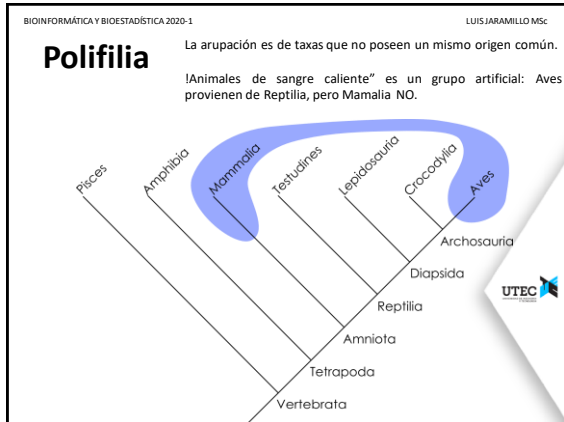
El clado agrupa a TODOS los descendientes de un ancestro común



Parafilia

El clado NO agrupa a todos los descendientes de un ancestro común





BIOINFORMÁTICA Y BIOESTADÍSTICA 2020-1 LUIS JARAMILLO MSc

¿Cuál es la probabilidad de observar un dato?

- Si tiramos una moneda y pensamos que la moneda es normal, entonces podríamos esperar una probabilidad de observar "cara" de 0.5.
- Si creemos que esta "arreglada" y esperamos obtener una "cara" el 80 % de la veces ...luego la probabilidad de observar los datos (una "cara") es 0.8.
- POR LO TANTO:** La "likelihood" de hacer ciertas observaciones es enteramente dependiente de un modelo y de los supuestos que subyacen en éste.

Moraleja: Los datos NO HAN CAMBIADO nuestro modelo SI. Por lo tanto, bajo un nuevo modelo la probabilidad de observar los datos HA

$p = ?$

UTEC

BIOINFORMÁTICA Y BIOESTADÍSTICA 2020-1 LUIS JARAMILLO MSc

Método de Maximum Likelihood :

La likelihood (L) de un árbol filogenético es la probabilidad de observar los datos (secuencia nucleotídica) bajo un árbol dado y un modelo especificado para los cambios en el carácter.

La meta es encontrar un árbol (entre todos los posibles) con el valor más alto de L.


Probabilidad de dado

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


UTEC

BIOINFORMÁTICA Y BIOESTADÍSTICA 2020-1 LUIS JARAMILLO MSc


Parámetros del Modelo de Máxima Probabilidad

TOPOLOGÍA  + La proporción de sitios invariantes (λ).

• La tasa relativa de sustitución en la matriz (TRANS v/s TRANSVER).

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

Las frecuencias de las bases (π).



BIOINFORMÁTICA Y BIOESTADÍSTICA 2020-1 LUIS JARAMILLO MSc

Maximum Likelihood

The Idiot's Guide to the Zen of Likelihood in a Nutshell in Seven Days for Dummies, Unleashed

A gentle introduction, for those of us who are small of brain, to the calculation of the likelihood of molecular sequences

Peter G. Foster*

Likelihood = Probabilidad de los datos dado el modelo

Calcular algo difícil...

a

Modelo:
El 100% de mis resultados es a
El 100% de mis resultados es c
El 50% de mis resultados es a
El 25% de mis resultados es a

BIOINFORMÁTICA Y BIOESTADÍSTICA 2020-1 LUIS JARAMILLO MSc

2

UN POCO DE PRÁCTICA



BIOINFORMÁTICA Y BIOESTADÍSTICA 2020-1 LUIS JARAMILLO MSc

Uso de Clustal / Bioedit

1. Descargar el programa Bioedit en: <https://bioedit.software.informer.com/Download/zip/>
2. Se realizará el alineamiento de cinco secuencias (Cisk).

A. Alineamiento Múltiple

- a. Con este fin vamos seleccionando 5 secuencias de la base de datos en formato txt (Cisk.txt) con las que realizaremos el alineamiento múltiple.
- b. Inmediatamente después de seleccionarlas, las secuencias se disponen en la ventana del programa, pero aún no están alineadas.
- c. Para aplicar el algoritmo ClustalW se va a la barra de herramientas en *Accessory/Align* y se elige *ClustalW/Multiple alignment*.
- d. Inmediatamente se abre una ventana en la cual se indican por defecto algunos parámetros, los cuales pueden ser elegidos según criterio del investigador, entre ellos lo recomendable es seleccionar la opción *Output Clustal format clustal consensus*.
- e. Y luego a *Run ClustalW*, indicando que es el link de la parte inferior del panel.
- f. Para finalmente obtener el alineamiento de las secuencias prometidas de interés.

Observa la línea que dice *Clustal Consensus* (última línea del alineamiento), allí puede ser 3 opciones, que significan:

- Asterisco (*) indican que en dicha posición los nucleótidos son 100% idénticos.
- Dos puntos (:) indican posiciones en las que se han realizado sustituciones conservativas.
- Punto (.) indican sustituciones menos conservativas.

La secuencia consenso nos da una muy buena idea acerca de las características de nuestro alineamiento, en última instancia lo que queremos evaluar con un alineamiento múltiple es el nivel y lugar de conservación de nuestras secuencias.

B. Alineamiento entre dos secuencias

- a. Seleccionar dos secuencias del alineamiento anterior y abrir en *Sequence* y elegir la opción *Compare*.
- b. Luego elegir *Pairwise alignment* y elegir *Align two sequences*.
- c. Se obtendrá los resultados de identidad y similitud entre las dos secuencias comparadas.

BIOINFORMÁTICA Y BIOESTADÍSTICA 2020-1 LUIS JARAMILLO MSc

Uso de Mega

1. Descargar el programa Mega en: <https://www.megasoftware.net/>. Admitir como student of University.
2. Abrir el programa y aplicar *Align*, luego *Edit/Build Alignment*, se abrirá la ventana de *Alignment Editor* y se debe seleccionar *Retrieve a sequence from a file*. Elegir el archivo (Cisk.txt).
3. Se cargarán las secuencias proteicas en la ventana *MX: Alignment Explorer*.
4. Para aplicar el algoritmo ClustalW se va a la barra de herramientas en *Alignment* y se elige *Align by ClustalW*.
5. Para finalmente obtener el alineamiento de las secuencias proteicas de interés. Exportar el alineamiento en formato Mega y guardarlo.
6. Ir en la barra de Herramientas en *Distance* y aplicar *Compute Pairwise Distance* y seleccionar el archivo guardado en Mega.
7. Discutir los parámetros y analizar los resultados.
8. Explorar las herramientas disponibles.

BIOINFORMÁTICA Y BIOESTADÍSTICA 2020-1 LUIS JARAMILLO MSc

Conclusiones

- ✓ Un árbol es una hipótesis de cómo se relacionan los individuos / grupos.
- ✓ Monofilia, Parafilia y Polifilia.
- ✓ Máxima Parsimonia y Máximo Likelihood.
- ✓ Uso de Clustal y Mega

