Práctica. Introducción al análisis filogenético

(*texto en itálica corresponde a comandos que hay que tipear*, "texto entre comillas, corresponde a opciones que hay que seleccionar",$indica el prompt de la terminal)

A) SELECCIÓN DE SECUENCIAS

1. Búsqueda de secuencias de proteínas

En primer lugar realizaremos la búsqueda de proteínas por similitud utilizando la herramienta Blast (*Basic Local Alignment Search Tool*) del ncbi. Puede elegir cualquier proteína de su interés o buscar la **α-amilasa** de ratón (**alpha-amylase, Mus musculus,** CAA24099.1)
Haga una búsqueda de esta proteína en ncbi para conocer lo que se sabe sobre la misma. En la página del ncbi puede buscar en la base de datos de proteínas con el ID indicado o bien tipear el nombre de la proteína de interés y el organismo. Descargue la secuencia de interés en formato fasta; para ello elija la opción "send to: File" "Format:FASTA"

Para realizar la búsqueda de secuencias similares, ir a la página https://blast.ncbi.nlm.nih.gov/Blast.cgi, seleccionar la opción "Protein Blast" y hacer una búsqueda utilizando el ID de la secuencia de interés.
-utilice la base de datos "non-redundant protein sequences"
-realice un blast estándar ("protein-protein BLAST)

Elija cuatro secuencias con distinto porcentaje de identidad y guarde las secuencias en formato fasta. Ahora vamos a crear un archivo multifasta utilizando el comando "cat", este comando concatena archivos, es decir pega los archivos uno debajo de otro;

$*cat file1.fasta file2.fasta …. File5.fasta*
en este caso el archivo saldrá por pantalla muy rápido, para guardarlo en un archivo multifasta;

$*cat file1.fasta file2.fasta …. File5.fasta* > proteinas.fasta

🔍Ahora podemos discutir sobre un poco sobre similitud y homología

2. Visualización de archivos

Vamos a comprobar cómo se ve nuestro archivo multifasta, vamos a utilizar el programa aliview que podemos abrir desde el icono superior de la barra izquierda (a mi me recuerda a un remolino morado…) allí se pueden buscar los programas instalados en el equipo que tienen interfaz gráfica, tipear *aliview* si no vemos el icono y hacer doble click (como en windows). Desde la consola podemos escribir;

$*aliview*

Para ver el archivo podemos usar el menú “File, open” o bien arrastrar el multifasta el programa.

🔍¿El archivo está 'ordenado’?

3. Alineamiento de secuencias

Si la elección de proteínas fue variada, es bastante probable que las secuencias tengan distinto tamaño y por ello no estén 'alineadas' como corresponde, por ello vamos a realizar el alineamiento con el programa clustal omega
(siempre que uses un programa en linux que no controles, comienza mirando la ayuda … esto es útil para saber cómo escribir el comando, las opciones que tiene, etc; al principio resulta difícil de entender pero poco a poco se adquiere la costumbre a leer el lenguaje y el trabajo se simplifica muchísimo)

$*clustalo -h #a veces es --h o -help, hay que probar*
$*clustalo -i PATH/input_file.fas --infmt=fa -t DNA -o PATH/output_file.fas --outfmt=fa -v --force*

-i: archivo de entrada (con su ruta -*PATH*- si el archivo no se encuentra en el directorio donde estás trabajando)
--infm/--outfmt: formato del archivo de entrada/salida (fa=fasta)
-t: tipo de dato, DNA/RNA/Proteina
-o: nombre del archivo de salida
-v: muestra el proceso en pantalla (opcional)
--force: fuerza a sobreescribir el archivo de salida (opcional)

🔍¿Cómo se ven las secuencias ahora? ¿Por qué es necesario alinear las secuencias?

B) FILOGENIA DE PRIMATES

Realizaremos una filogenia de primates basada en los datos publicados por Perelman et al (2011). Los datos corresponden a un concatenado de 5 genes en 9 grupos de primates del viejo y nuevo mundo, así como homínidos. Las secuencias fueron descargadas del ncbi. Se utilizaron secuencias de conejo como *outgroup* (Oryctolagus).El objetivo de este ejercicio es familiarizarse con distintos métodos (NJ, ML e inferencia Bayesiana) y herramientas comúnmente utilizadas para reconstruir filogenias.

1. Preparación de los datos (secuencias)

-Observar el archivo multifasta primates.fas

🔍¿Qué deberías hacer antes de continuar el análisis?

-Alinear las secuencias del archivo primates.fas. Observar el archivo alineado

🔍¿Qué puedes decir de este alineamiento?

2. **Seleccionar los bloques de secuencias conservadas con Gblocks** (http://molevol.cmima.csic.es/castresana/Gblocks_server.html)

Este programa nos permite seleccionar los bloques más conservados y 'limpiar' nuestro alineamiento con el fin de eliminar secuencias de dudosa homología, o saturadas de sustituciones que nos lleven a inferencias filogenéticas incorrectas.

-Para abrir el programa puedes cambiar al directorio donde se encuentra el programa y ejecutarlo, o ejecutarlo donde estés indicando el directorio donde se encuentra;

$./*PATH/Gblocks*

Este herramienta funciona con un desplegable de parámetros que hay que ir eligiendo, en el primer paso, hay que abrir el archivo:

```
*****************************************************
                  GBLOCKS 0.91b
SELECTION OF CONSERVED BLOCKS FROM MULTIPLE ALIGNMENTS
        FOR THEIR USE IN PHYLOGENETIC ANALYSIS
*****************************************************

o. Open File

b. Block Parameters

s. Saving Options

g. (Get Blocks)

q. Quit


Your Choice:
```

Elegir la opción o, luego pedirá el nombre del archivo de entrada /PATH/input_file.fasta

Luego seleccionamos t (Type of sequence), hay que repetir la opción t hasta que veamos la opción adecuada (DNA, Protein…)

Luego elegimos la opción b para seleccionar los parámetros de selección de bloques; en esta ocasión dejaremos los parámetros por defecto, para ver las opciones disponibles y el significado de los distintos parámetros es aconsejable consultar la documentación del programa (http://molevol.cmima.csic.es/castresana/Gblocks/Gblocks_documentation.html)
Por último elegimos la opción g, nos arrojará el resultado inmediatamente
Tipear m, luego q

El programa genera dos archivos de salida que se encuentran en el mismo directorio del archivo de entrada, se llaman input_file.fas-gb e input_file.fas-gb.htm.
El archivo .htm nos permite visualizar los bloques que fueron seleccionados, haciendo doble click sobre el archivo se abre en el navegador. O bien El archivo .fas-gb es el que utilizaremos para los siguientes análisis, vamos a renombrarlo:

$*mv input_file.fas-gb file_gblocks.fas*

-Observar el archivo final con aliview

-Guardar el archivo final en formato nexus (.nex) y phylip (.phy), que usaremos en los análisis subsiguientes. Esto vamos a hacerlo desde aliview en el menú "File" se pueden guardar los archivos con ambos formatos. También se puede utilizar un programa por línea de comandos, esta opción es útil cuando necesitamos convertir archivos muy grandes o muchos archivos y lo hacemos con un bucle. Uno de los programas que se pueden usar es seqmagick, se pueden seleccionar varios formatos de salida, además de elegir si se trata de secuencias de ADN o proteínas

🔍¿Por qué es importante 'limpiar' el alineamiento?

3. **Selección del modelo evolutivo (o modelo de sustitución)**

Para ello vamos a utilizar la herramienta jModelTest2, que se utiliza para evaluar estadísticamente qué modelo se ajusta mejor a nuestros datos. Utiliza diferentes test o criterios para poder realizar esta selección
-Para abrir el programa
$runjmodeltest-gui.sh

El mismo tiene una interfaz gráfica, en el menú "File" elegir "Load DNA Alignment" y seleccionar el archivo .phy. Este programa debe funcionar también con archivos .fasta, aunque originalmente utilizaba solo formatos phy.
En la consola del programa deben verse el número de caracteres del alineamiento y debe indicar si se ha cargado el archivo OK!

Luego en el menú "Analysis" vamos a realizar todos los análisis disponibles (Likelihood, AIC, BIC y DT). Finalmente en el menú "Results" seleccionamos "Show results table"

🔍Qué modelo ajusta mejor a nuestros dato? ¿Los test arrojan el mismo resultado?

(El modelo TPM3 es igual a Kimura 3 parameters (K3P, K81))

4. Filogenia NJ (Mega)

Este método utiliza matrices de distancia; si bien no se los acepta como las mejores filogenias para publicar, es un método muy rápido y nos permite una primera visualización en formato de árbol de nuestros datos. Utilizaremos el programa Mega para realizar este análisis

$*megaproto*

El programa tiene una interfaz gráfica que nos permite seleccionar el análisis a realizar

- En el cuadro "Data Type to Analize" seleccionar *Nucleotide alignment*
- en el menú desplegable "Phylogeny", seleccionar *Construct/Test Neighbor-Joining Tree*
- seleccionar los siguientes parámetros:
    Test of Phylogeny "boostrap"
    No of Boostrap Replications "100"
    Model/method "no. differences"
    Substitutions "d:Transitions + Transversions"
    Rates among sites "Gamma distributed with invariant sites"
    Gamma parameter "4"

- "Save settings" guardar el archivo con el nombre por defecto en la carpeta donde se encuentra el alineamiento que generamos con gblocks (file_gblocks.fas)

- Ahora realizamos la filogenia

$*megacc -d file_gblocks.fas -a infer_NJ_coding.mao -o output_file*

5. Visualización de árboles (ITOL)

Existen muchas herramientas que permiten ver la filogenia resultante en forma de árbol, una muy cómoda y versátil es itol (https://itol.embl.de/). Es una herramienta web que permite modificar muy fácilmente el árbol y agregar información de interés en forma sencilla.

-Crear una cuenta (si no tienen una previa)
-Ir al menú "My Trees" y crear un "Project" con el nombre que prefieran
-Arrastrar el archivo .nwk que obtuvimos (no el consenso), enraizar utilizando el outgroup y hacer visibles los valores de boostrap…. Aquí nos toca jugar un poco para que nuestro árbol quede genial!

6. Filogenia ML (raxml)

Vamos a utilizar el mismo dataset para realizar una filogenia basada en Maximum Likelihood o máxima verosimilitud. Hay muchos programas utilizan este método, incluso Mega lo hace; sin

embargo con raxml es posible reconstruir filogenias de datasets muy grandes, ya sea con muchas secuencias o con pocas pero muy largas, como es el caso de filogenias basadas en genomas completos. Además tiene la ventaja de que el proceso se puede paralelizar y tiene una opción de boostrap rápido.

-Este programa no se abre, simplemente se ejecuta el comando con los parámetros adecuados.

$raxml -f a -s input_file.fas -n output_file -m GTRCATI -x 1232154 -N 10 -p 123540

$/PATH/raxmlHPC-PTHREADS -f a -s input_file.fas -n output_file -m GTRCATI -x 1232154 -N 10 -p 123540. Si nosotros los instalamos debe ejecutarse con este comando

- f a: esta opción permite realizar un boostrap rápido y seleccionar el árbol con el mejor score en una corrida
-s: nombre del archivo de entrada
-n: nombre del archivo de salida
-m GTRCAT: modelo GTR + gamma + i
-x: número entero al azar (random seed) para iniciar el rapid bootstrapping
-N: el número de árboles que se utilizan para empezar el análisis
-p: número al azar para las inferencias, hace que los resultados sean más reproducibles

-Visualizar el árbol con itol

7. Filogenia Bayesiana (MrBayes)
Finalmente, realizaremos una filogenia basada en métodos bayesianos. Actualmente la mayoría de las publicaciones incluyen este método y lo comparan con los resultados obtenidos con ML. La estadística bayesiana aparece como la opción más robusta al momento de realizar reconstrucciones filogenéticas.

-Comenzar abriendo el programa (*mb* o *./mb*)
Se desplegarán los créditos y a continuación aparecerá el prompt, aquí es cuando podemos comenzar a escribir
MrBayes>

Primero es necesario cargar nuestro alineamiento, este programa lee secuencias en formato nexus (.nex)
MrBayes> */PATH/execute input_file.nex*

-En segundo lugar vamos a seleccionar el tipo de datos, el modelo sustitución (modelo evolutivo) que hemos definido previamente, en este caso el ejemplo está dado para GTR + I + G
MrBayes> *lset Nucmodel=4by4 nst=6 rates=invgamma*

Ahora podemos ver cómo ha quedado nuestro modelo...
MrBayes> showmodel

-Ahora hay que definir los parámetros de la corrida

MrBayes> *mcmc ngen=100000 samplefreq=100 printfreq=100 diagnfreq=1000*

ngen= número de generaciones o longitud de las cadenas
samplefreq= número de árboles o muestras que guarda y utiliza para calcular las probabilidades a posteriori, en este caso es 1000 (=100000/100)
printfreq= la frecuencia de impresión de los datos en pantalla esto es cada 100 generaciones
diagnfreq: la evaluación del modelo la hace cada 1000 generaciones

- Vamos a utilizar los "piors" por defecto, si quieres ver todos los que podrías modificar
$ *help prset*

- El programa va a realizar el análisis hasta completar las 100000 generaciones y va a imprimir los resultados de las probabilidades a posteriori. Una vez complete el análisis nos preguntará si queremos continuar. Por regla general si el *desvío estándar de las frecuencias divididas* es menor a 0.01, entonces podemos asumir que se ha alcanzado la convergencia y no necesitamos seguir agregando generaciones, indicar *no* y evaluar los resultados

- Hay dos tipos de resultados que debemos evaluar, en primer lugar veremos el resultado de los parámetros del modelo de sustitución, se presentan la media, moda y el intervalo del 95% de credibilidad. Los resultados son válidos si todos los parámetros tienen un PSRF 'potential scale reduction factor' cercano a 1.0 (si esto no ocurre hay que aumentar el número de generaciones). Lo mismo sucede con el ESS 'effective sample size, si los valores son menores a 100 es necesario realizar una corrida con mayor número de generaciones. Para el resumen de los resultados vamos a eliminar el porcentaje de burning (25% de las 1000 muestras que guardamos =250), ya que corresponden al previo a alcanzar el estado estacionario.

MrBayes> *sump burnin=250*

- El segundo resultado corresponde al resumen de los árboles registrados, también descartamos los primeros 250 (25% burning)

MrBayes>*sumt burnin=250*
- Aquí podemos observar los valores de soporte de rama y la topología del árbol

8. Visualización de árboles (FigTree)

Luego visualizamos el archivo .tre, que estará guardado en en la carpeta donde tenemos el archivo de entrada. En esta ocasión vamos a utilizar el programa FigTree.

- Para abrir el programa;
$*figtree*
(*$java -jar /PATH/lib/figtree.jar)*

- Se abre la interfaz gráfica
- En el menú “File” seleccionar “Open” y buscar el archivo .tre
- Observar la topología del árbol
- Enraizar utilizando nuestro outgroup utilizando el icono “Reroot” y seleccionar en las opciones de la izquierda el menú “Node labels” y “Display, pob” para observar los valores de soporte de las ramas.

🔍Compara las filogenias obtenidas por los distintos métodos
¿Las filogenias obtenidas reflejan las mismas relaciones evolutivas que se observan en la publicación original?
¿Hay algo que llame especialmente su atención?