Análisis de la Evolución mediante Herramientas Moleculares:

Búsqueda en Bases de Datos

© Grupo de Innovación sobre la Docencia en Diversidad Biológica
BIOINNOVA :: www.innovabiologia.com

ANÁLISIS FILOGENÉTICO BASADO EN SECUENCIAS DE ADN

FILOGENIA

La filogenia estudia el desarrollo de las líneas evolutivas y la formación de especies a partir de otras preexistentes, e intenta establecer el árbol genealógico de todos los organismos, actuales o extintos (Haeckel, 1866).

La necesidad por parte del ser humano de elaborar una clasificación biológica es muy antigua. Inicialmente se basaba en agrupar los organismos en taxones con base en su morfología y anatomía (el fenotipo). Actualmente, y con el desarrollo de las nuevas técnicas, no se concibe una clasificación filogenética sin incluir información molecular. Es por ello que en esta práctica, queremos ilustrar la elaboración básica de árboles filogenéticos basados en secuencias de ADN.

En primer lugar, se buscarán las secuencias de interés en una base de datos *online* del **NCBI** (*National Center for Biotechnology Information*), y a continuación se procederá al alineamiento de dichas secuencias y la elaboración de árboles filogenéticos mediante el software **MEGA** (*Molecular Evolutionary Genetics Anslysis*).

La versión más reciente de este programa es MEGA 7, que se encuentra disponible para distintos sistemas operativos y se puede descargar gratuitamente desde la página web:

MEGA Molecular Evolutionary Genetics Analysis

http://www.megasoftware.net/

1. Búsqueda de secuencias en NCBI

Nuestros análisis están limitados por la información contenida en las bases de datos *online*. Se utilizarán secuencias de **beta-actina** para que los resultados sean comparables a los del análisis de proteínas en laboratorio (ver práctica de laboratorio). Sin embargo, no todas las especies utilizadas disponen de secuencias en la base de datos, por lo que este análisis se limitará a aquellas especies cuya beta-actina haya sido secuenciada previamente por otros autores.

Existe la posibilidad de utilizar otros genes. Por ejemplo, los genes mitocondriales son generalmente buenos marcadores filogenéticos, dado que al contener cada célula varias mitocondrias el ADN mitocondrial, haploide, se extrae en grandes cantidades. Además, no se recombina, ya que se transmite de madres a hijos, pasando casi intacto de generación en generación y siendo su única fuente de variabilidad las mutaciones. Por último, su evolución es relativamente rápida en comparación con el ADN nuclear. Todo ello facilita la reconstrucción de las relaciones evolutivas dentro de una especie y entre especies. En este caso, sin embargo, nos ceñiremos a la beta-actina para ser consistentes con la práctica de laboratorio.

En primer lugar, es necesario obtener las secuencias de interés de la base de datos. Para ello, abrir en un navegador la página del NCBI:

NCBI
National Center for
Biotechnology Information

https://www.ncbi.nlm.nih.gov

Seleccionar "*Nucleotide*" en el desplegable señalado con una flecha en la Figura 1, para buscar las secuencias de ADN disponibles, y escribir la especie de interés y el nombre del gen en el cuadro de búsqueda. En el ejemplo de la Figura 1 se incluyó el rodaballo: "*Psetta máxima beta actin*". Clicar en "*Search*".

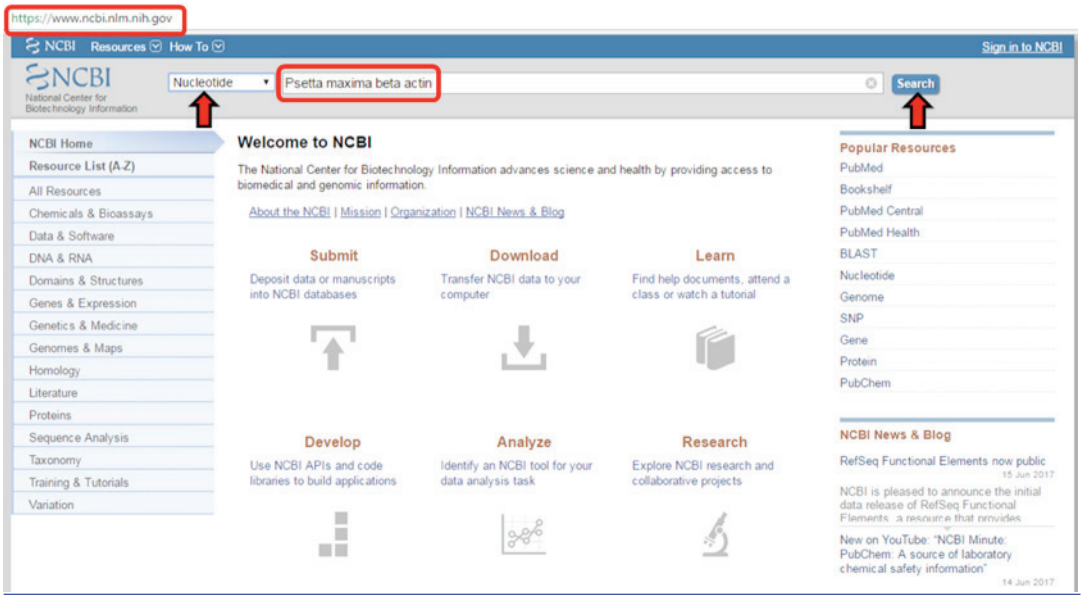

Figura 1

Aparecerán todas las entradas disponibles para esa especie y ese gen. Es conveniente utilizar la secuencia más larga. La longitud de la secuencia está expresada en pares de bases (bp) debajo del enlace con el nombre (recuadro en Figura 2). En este caso la primera entrada es válida. Es conveniente saber que podría presentarse la misma especie con dos nombres válidos (sinonimia): en este caso aparecen *Psetta máxima* y *Scophthalmus maximus*. Tras hacer click en el nombre (flecha en Figura 2) se verá la secuencia y toda la información asociada (Figura 3), incluido el número de acceso. Este número, si se conoce, se puede utilizar para buscar las secuencias (en lugar de usar el nombre de la especie).

Con el objeto de uniformar los resultados y asegurar la correcta selección de secuencias, se proporciona a continuación un listado con los **números de acceso** de las secuencias que se deben descargar para la práctica. De este modo, la búsqueda en la página del NCBI se realiza igual salvo que, en lugar de escribir el nombre de la especie y el gen de interés, se escribe directamente el número de acceso (en el caso de *Psetta maxima*: EU686692.1). De esta forma aparecerá directamente la información de la secuencia (Figura 3).

Los números de las secuencias para descargar en esta práctica son:

- *Psetta maxima* (rodaballo): EU686692.1
- *Merluccius merluccius* (merluza): EU022566.1
- *Salmo salar* (salmón): AF012125.1
- *Solea solea* (lenguado): KJ510524.1
- *Thunnus albacares* (atún): KF536727.1
- *Petromyzon marinus* (lamprea): KM925135.1

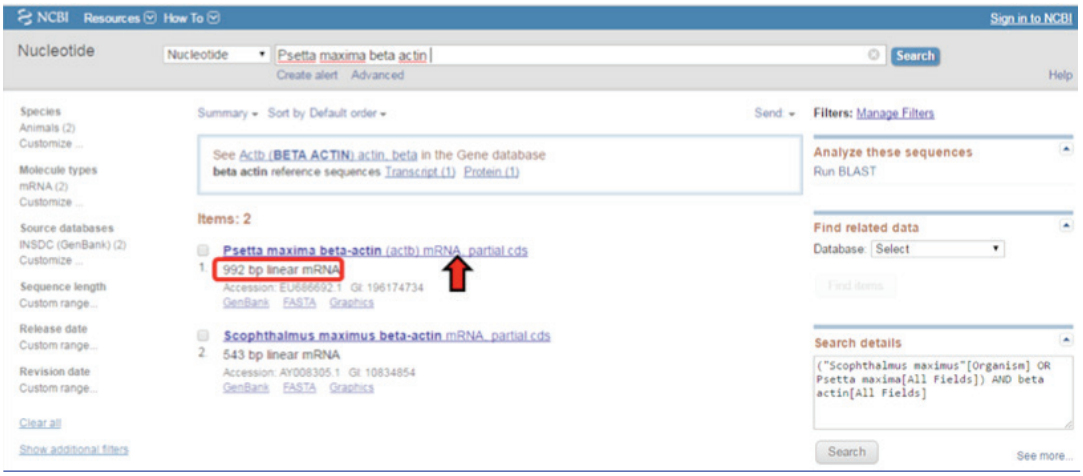

Figura 2

En el momento de elaboración de este dossier no estaban disponibles las secuencias de otras especies usadas en la práctica de laboratorio: *Conger conger* (Congrio), *Engraulis encrasicolus* (Boquerón), *Lepidorhombus whiffiagonis* (Gallo) y Raja pulchra (Raya). Sin embargo, es probable que en un futuro las secuencias para algunas de estas especies aparezcan en las bases de datos. En ese caso, se invita al alumno a que explore la posibilidad de incluirlas en el análisis siguiendo el mismo procedimiento.

Se ha incluido la secuencia de la lamprea con objeto de enraizar el árbol. Se trata de un **grupo externo** que, por ser más antiguo que el resto, permitirá contextualizar los resultados del árbol y colocar correctamente las ramas del mismo.

GRUPO EXTERNO (*OUTGROUP*)

Es un organismo (o grupo de organismos) utilizado en un análisis filogenético y que no se incluye dentro del taxón de interés (grupo interno), aunque es cercano a éste. Se incluye con fines comparativos, para comprender la evolución de los caracteres en una filogenia, ya que permite conocer la dirección de los cambios.

2. Descarga de secuencias en formato FASTA

Una vez dentro de la página que incluye la información de una secuencia, se debe clicar en "*Send*" y se mostrará un desplegable. En el mismo, seleccionar "*File*" como destino, "FASTA" como formato y la opción "*Show GI*" (que mostrará el número de acceso en la descarga). Una vez marcadas estas opciones, clicar en "*Create File*" (Figura 3). Esto provocará la descarga de un archivo de texto que contiene la secuencia. Se recomienda renombrar el archivo con el nombre de la especie.

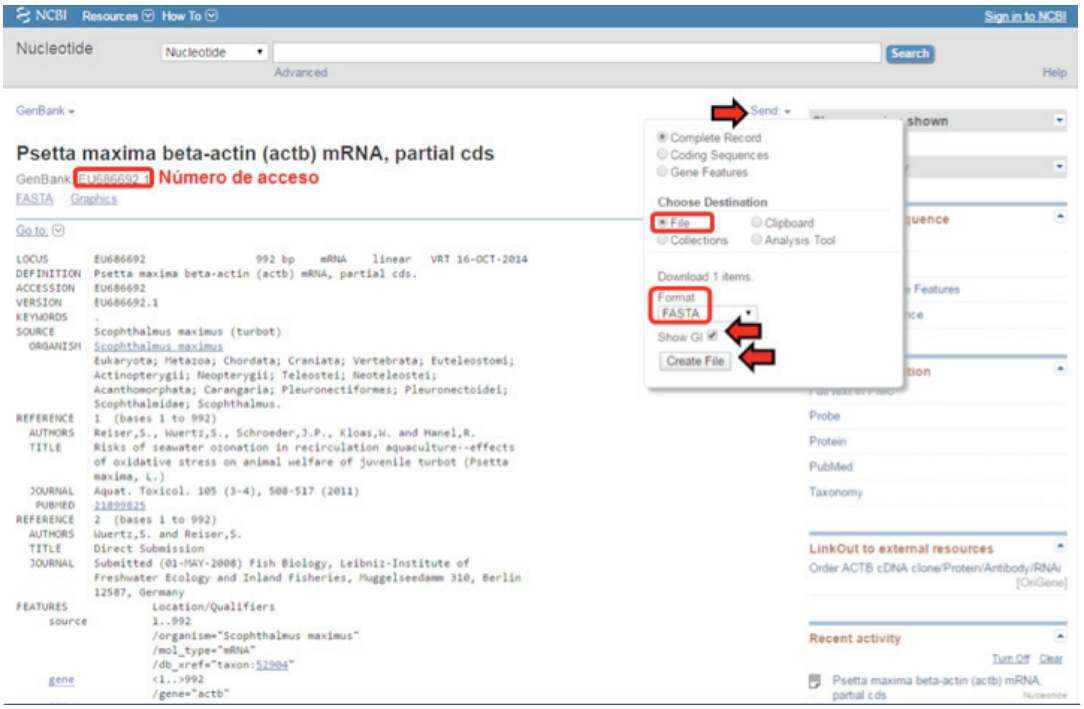

Figura 3

Al abrir el archivo de texto descargado se puede comprobar en qué consiste el formato **FASTA** (Figura 4). Básicamente, es un archivo de texto plano que consta de una línea inicial que comienza con el símbolo ">" y a continuación la información sobre la secuencia. En muchos programas utilizados para filogenia sólo aparecerá la información hasta el primer espacio, de tal forma que si se quiere conservar la información se deben sustituir los espacios por guiones bajos (_).

En cualquier caso, según el programa, puede existir limitación de caracteres para los nombres, así que igualmente la información puede ser reducida por el propio programa. Por ello, se editarán las secuencias dejando el número de acceso y nombre de la especie para evitar problemas (Figura 5). Tras un salto de línea, comienza la secuencia en sí.

Hay muchos otros formatos (NEXUS, PHYLIP, CLUSTAL, MEGA, etc.), adecuados según el programa que se utilice. Existen herramientas *online* para realizar la conversión. Por ejemplo: http://sequenceconversion.bugaco.com/converter/biology/sequences/.

Figura 4

En este caso, con el formato FASTA será suficiente, ya que el programa MEGA lo abre sin problema y lo transforma. Pero antes de llevar a cabo el análisis de las secuencias, es preciso juntar todas las especies en un único archivo de texto (*.txt). Para ello, simplemente se deben copiar y pegar **todas las secuencias dentro del mismo archivo**, usando el bloc de notas (Windows) o algún programa equivalente (Figura 5). Usando la opción de "Guardar como", añadiremos al nombre del archivo la extensión ".fasta". Por ejemplo, el archivo "Todos_BetaActin. fasta" contendrá las secuencias del gen beta-actina en formato FASTA de todas las especies a analizar. Esto permitirá que el programa MEGA reconozca el formato.

Figura 5

3. Alineamiento de secuencias en MEGA

Abrir el programa MEGA y seleccionar "*File*" y "*Open A File/Session*". Seleccionar el archivo fasta que hemos creado ("Todos_BetaActin.fasta") y después clicar en "*Align*" (Figura 6). Esto abrirá automáticamente nuestras secuencias sin alinear en otra ventana (Figura 7). Para proceder con el **alineamiento**, desplegar el menú "*Alignment*" y clicar en la opción "*Align by ClustalW*". Tras seleccionar ClustalW, MEGA pregunta: "*Nothing selected for Alignment, select all?*" Clicar en OK. Aparecerá una nueva ventana para modificar los parámetros de alineamiento (*ClustalW parameters*). Dejamos las opciones que vienen por defecto y clicamos en "OK".

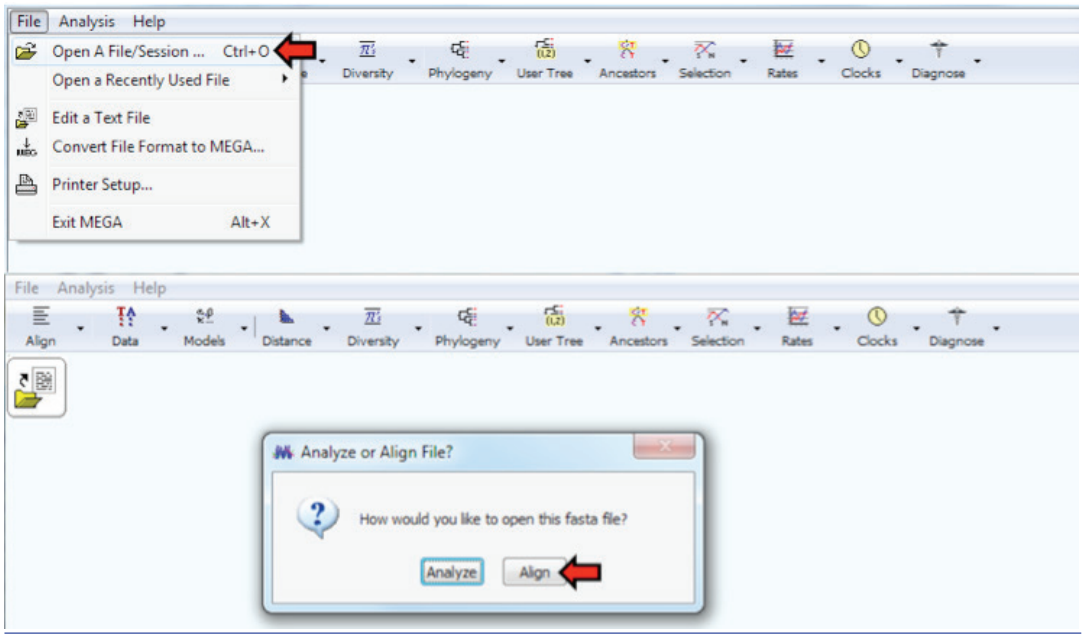

Figura 6

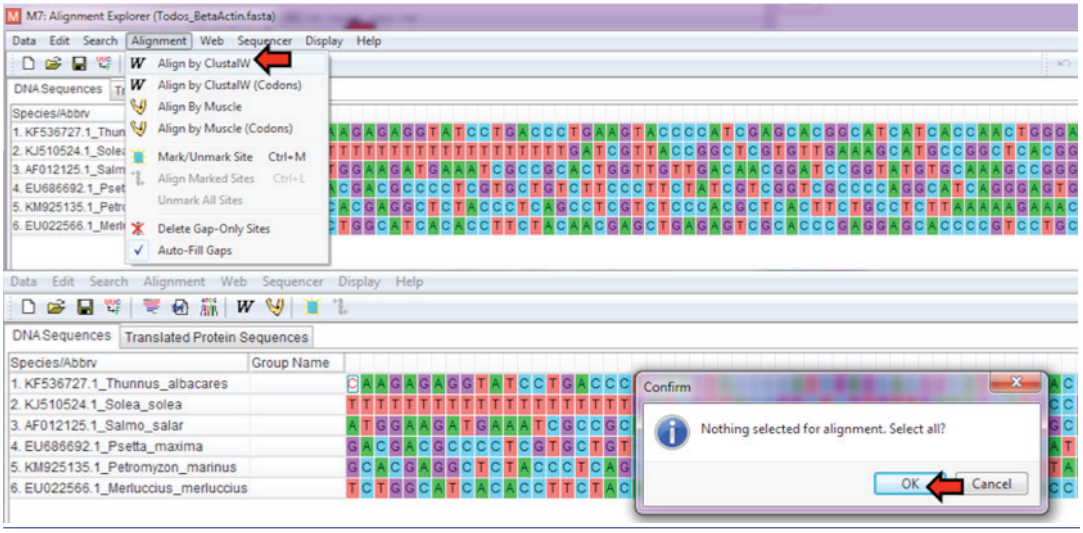

Figura 7

El programa coloca alineadas las posiciones de la secuencia que considera homólogas (**homología**: semejanza derivada de un ancestro común). Usa el criterio de **parsimonia**: busca el máximo número de correspondencias o similitudes (se espera evolutivamente el menos número de cambios). En un alineamiento múltiple como este, primero se comparan las secuencias por pares y se calcula la **distancia genética** (número de diferencias).

A continuación aparecerán las secuencias alineadas. Es conveniente cortarlas para que todas sean del **mismo tamaño** (en este caso la secuencia de *Solea* es mucho más larga que el resto). Para ello, seleccionar las posiciones de todas las secuencias que queremos borrar (el rectángulo rojo de la Figura 8 indica dónde hay que pinchar y arrastrar para la selección). A continuación se debe clicar en "*Edit*" y "*Delete*" para borrar las posiciones cuya información no está disponible en todas las secuencias (Figura 8). La primera posición será la 2086, para la cual todas las secuencias tienen información (nucleótido T). Hacer lo mismo con la parte final de las secuencias de tal manera que la longitud final del alineamiento sea 371 bp. Las secuencias acaban con la secuencia CACCA. MEGA marca las **posiciones del alineamiento que están conservadas** (son iguales) en todas las secuencias con un **asterisco** en la línea superior (a la derecha de "*Group Name*").

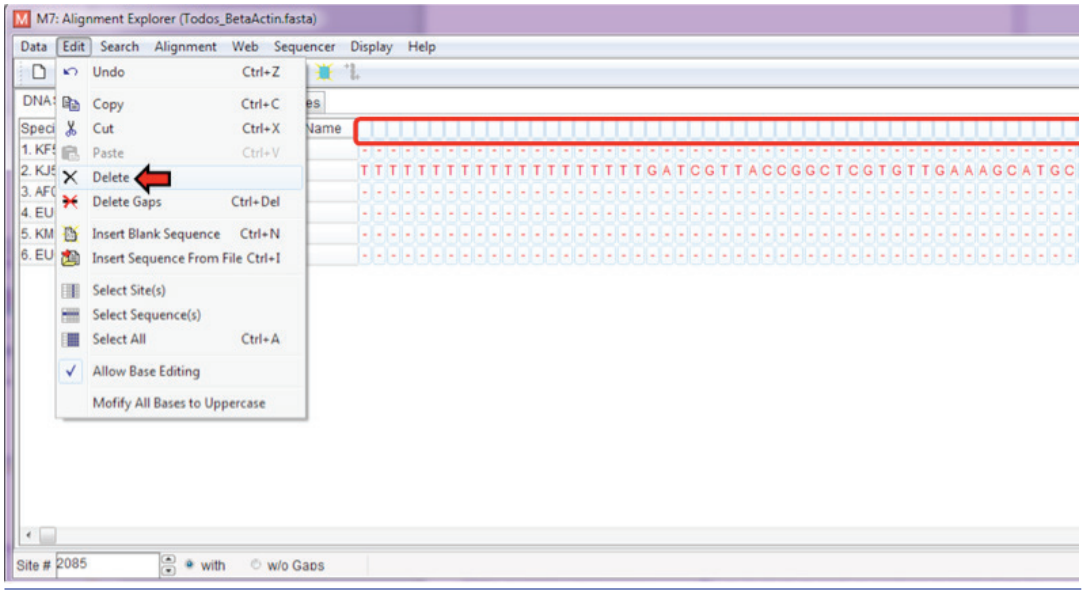

Figura 8

4. Elaboración de árboles filogenéticos

En la misma ventana del alineamiento, en el menú "*Data*" seleccionar "*Phylogenetic analyses*". MEGA preguntará si las secuencias son codificantes: "*Protein-coding nucleotide-sequence data?*". Clicar en "*Yes*" (Figura 9). El programa creará el formato necesario para proceder a los **análisis filogenéticos**. Una vez hecho esto, hay que volver a la ventana principal y en el menú "*Phylogeny*" seleccionar "*Construct/Test Maximum likelihood Tree*". El programa pregunta si queremos usar la hoja de datos activa ("*Would you like to use the currently active data?*"). Seleccionar "*Yes*" (Figura 10), ya que se trata del set de datos para filogenia que se ha creado tras el alineamiento.

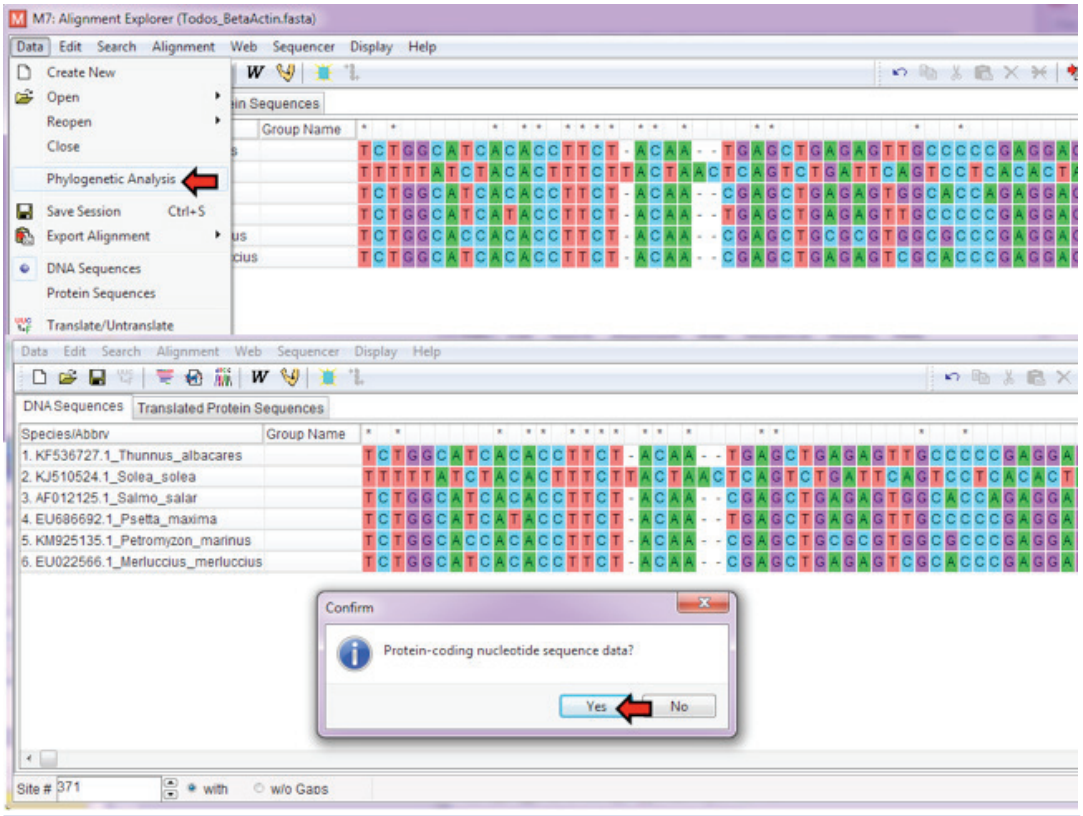

Figura 9

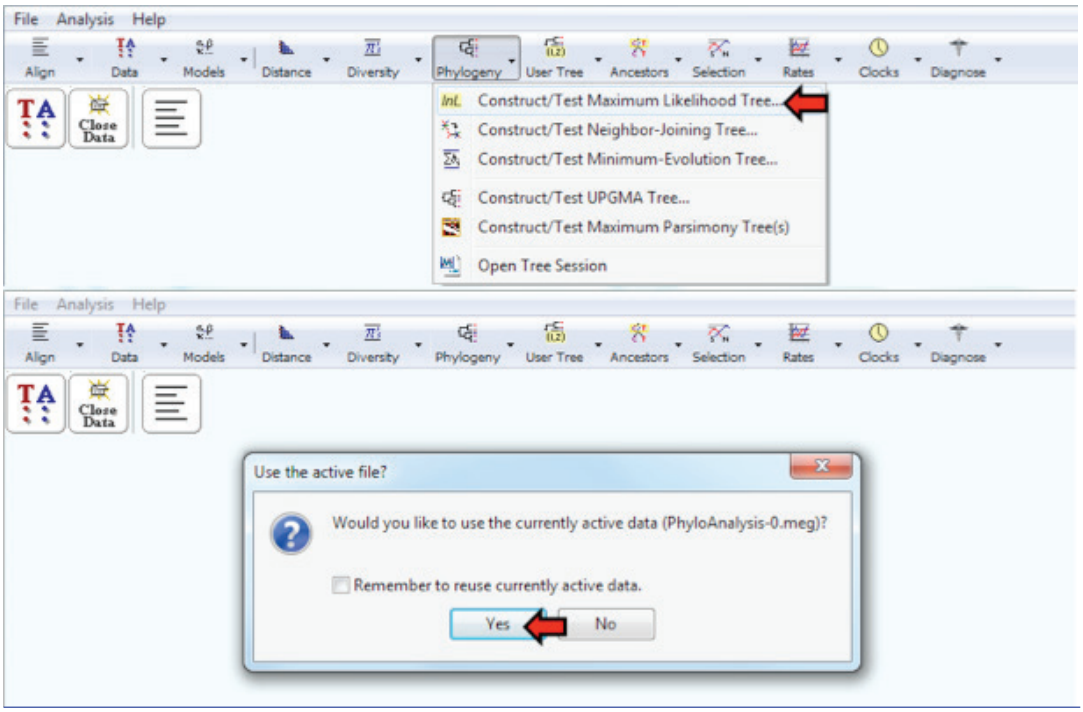

Figura 10

Aparecerá una nueva ventana para seleccionar las opciones del análisis filogenético. Se modificará "*Test of phylogeny*" para incluir "*Bootstrap method*" (500 réplicas es suficiente), en "*Model/Method*" se incluirá el "*General Time Reversible model*" (GTR), y en "*Rates among Sites*" especificaremos "*Gamma Distributed with Invariant Sites (G+I)*" (Figura 11). En principio, el resto de opciones se dejarían por defecto. En cualquier caso, para esta práctica comprobar que se dejan los mismos parámetros que aparecen en la Figura 1. Finalmente, ejecutar el análisis clicando en "*Compute*".

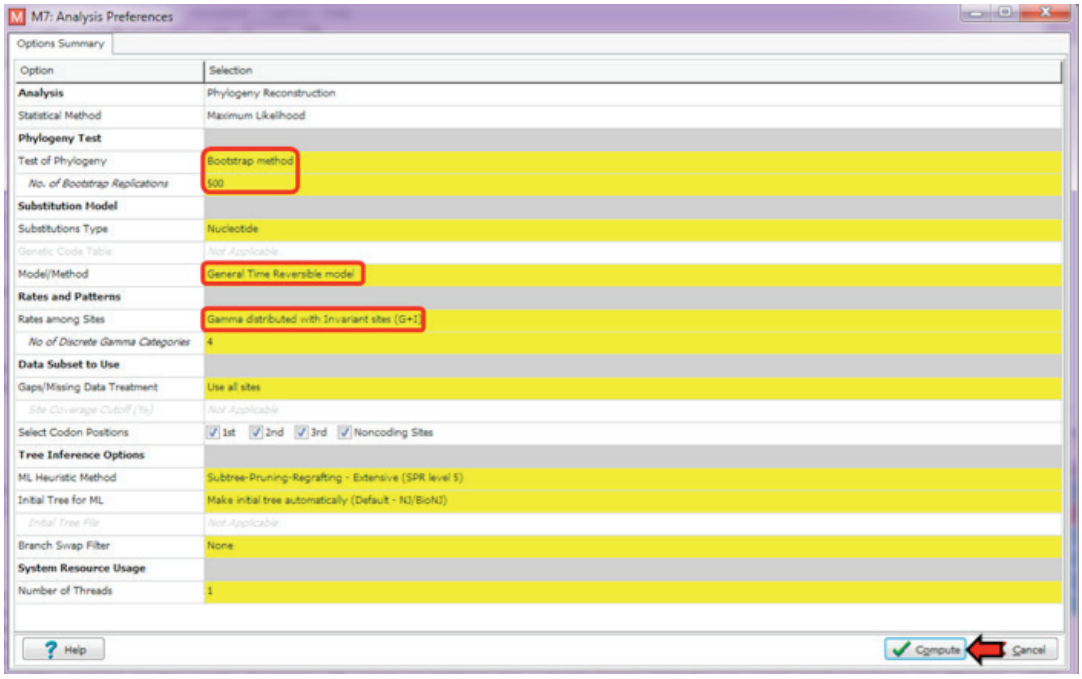

Figura 11

Existen varios métodos para elaborar reconstrucciones filogenéticas:

- **DISTANCIAS (*neighbor joining* o del vecino más próximo):**
  El árbol se calcula a partir de matrices de distancias genéticas entre las secuencias (o número de caracteres o nucleótidos en los que difieren). Este método ya no es muy utilizado en filogenia.

- **MÁXIMA PARSIMONIA (o del camino más corto):**
  El método evalúa los árboles posibles dándoles una puntuación en función del número de mutaciones que serían necesarias para explicar sus topologías. El árbol que precise el menor número de mutaciones será el seleccionado. En dicha puntuación se pueden penalizar procesos evolutivos más difíciles, como inserciones o deleciones. Este método es frecuentemente utilizado para reconstrucciones basadas en caracteres morfológicos.

- **PROBABILÍSTICOS (máxima verosimilitud y método bayesiano):** Utilizan técnicas estadísticas para inferir el árbol con mayor probabilidad de haber ocurrido. Requieren un modelo de sustitución (modelo evolutivo) para establecer la probabilidad de cada tipo de mutación en las secuencias. Estos son los métodos más utilizados actualmente para filogenia basada en secuencias de ADN y es por ello que en esta práctica se ha optado por la elaboración de un árbol de máxima verosimilitud (Maximum likelihood).

Los modelos de sustitución, o modelos evolutivos, se utilizan porque el número de diferencias observadas entre las secuencias no tiene por qué ser igual que el número de sustituciones que se han producido a lo largo de la historia evolutiva. Se podrían haber producido sustituciones múltiples en una misma posición (hay posiciones más variables que otras, como por ejemplo la tercera posición del codon que codifica un aminoácido), por lo que generalmente el número de sustituciones observadas subestima las sustituciones reales. Partiendo de esta base, se han ido implementando modelos evolutivos cada vez más complejos.

Hay programas que calculan el modelo al que se ajustan nuestras secuencias con más probabilidad. Sin embargo, y para no complicar la práctica, aquí hemos seleccionado el modelo que más comúnmente siguen las secuencias de ADN: *General Time Reversible model* (GTR) con *Gamma Distributed with Invariant Sites* (G+I). Este modelo permite aplicar distintas tasas de sustitución a lo largo de la secuencia, considera que los nucleótidos no son equifrecuentes (que hay distinta cantidad de A, T, G y C) y tiene en cuenta que hay posiciones que se saturan antes que otras (G), es decir en las que hay múltiples sustituciones con más probabilidad, y posiciones invariables (I). Por último, mencionar el test de soporte filogenético (**Bootstrap**), un método de remuestreo que permite asignar un valor de soporte a cada rama del árbol filogenético inferido.

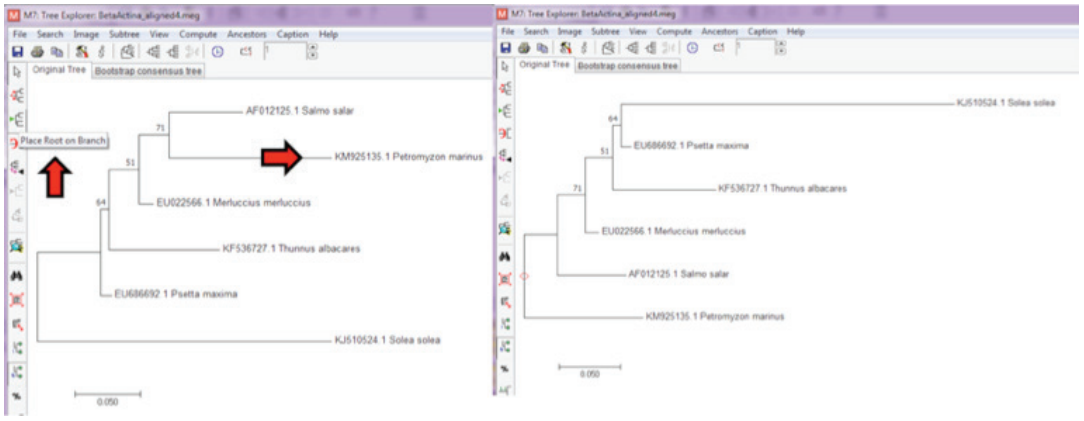

Figura 12

En la Figura 12 se muestra el árbol inferido por MEGA. Para enraizar el árbol, debemos asignar a la lamprea la posición de grupo externo. Para ello, clicamos en la rama de la lamprea (*Petromyzon*) y en el icono que indica "*Place root on branch*". Este sería nuestro árbol final, en cuyas ramas se muestran los valores de soporte sobre 100. No son valores de soporte demasiado elevados, probablemente por el reducido número de especies incluidas en el análisis y la gran separación evolutiva entre algunas de ellas.

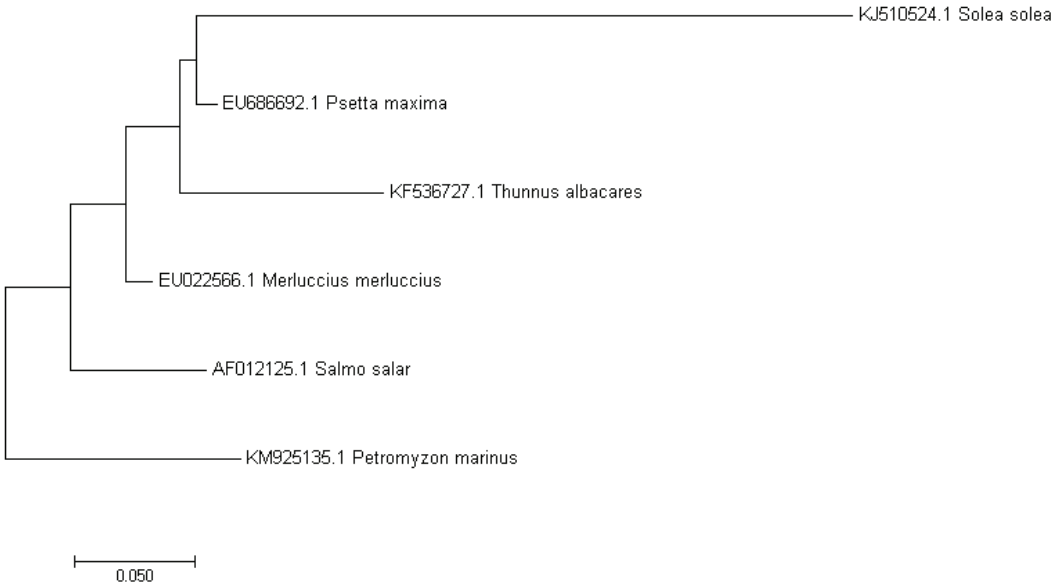

Figura 13

Como resultado, en la Figura 13 se muestra en detalle el árbol filogenético obtenido para las secuencias de beta-actina de seis especies de peces evolutivamente diversas: *Psetta maxima* (rodaballo), *Merluccius merluccius* (merluza), *Salmo salar* (salmón), *Solea solea* (lenguado), *Thunnus albacares* (atún) y *Petromyzon marinus* (lamprea).