Filogenía de Globinas: Construyendo árboles filogenéticos a partir de datos moleculares con MEGA

Adaptado del libro de Johanatan Pevsner y del tutorial de Barry G. Hall

8 de octubre de 2018

Resumen

Vamos a realizar la construcción de árboles filogenéticos con una alineación de secuencias múltiples de 13 proteínas globina, realizadas con MAFFT en entorno linux y su análisis filogenético en windows con la herramienta MEGA. Las secuencias que va a usar corresponden a:

- Tres mioglibinas (Referencia canguro rojo Macropus rufus, P02194)
- Tres globinas alfa (Referencia caballo Equus caballus, P01958)
- Tres globians beta (Referencia Canis lupus familiaris, XP_53790)
- Dos globinas de pescado (Referencia lamprea fluviatilis europea del río, 690951A)
- Una globina de insecto (Referencia midge larva Chironomus thummi thummi, P02229), y
- Una leghemoglobina vegetal (Referencia glicina de soja máx. 711674A)

Para la construcción de los árboles filogenéticos siga el protocolo descrito a continuación.

Introducción

El análisis filogenético a veces se considera un proceso intimidante y complejo que requiere experiencia y años de experiencia. De hecho, es un proceso bastante sencillo que puede aprenderse rápidamente y aplicarse de manera efectiva. Este Protocolo describe los varios pasos necesarios para producir un árbol filogenético a partir de datos moleculares para principiantes. En el ejemplo que se ilustra aquí, el programa MEGA se usa para implementar todos esos pasos, eliminando así la necesidad de aprender varios programas y para manejar múltiples formatos de archivo de un paso a otro. El primer paso, la identificación de un conjunto de secuencias homólogas y la descarga de esas secuencias, es implementado por el propio navegador de MEGA construido sobre el kit de herramientas de Google Chrome. Para el segundo paso, la alineación de esas secuencias, MEGA ofrece dos algoritmos diferentes: ClustalW y MUSCLE. Para el tercer paso, la construcción de un árbol filogenético a partir de las secuencias alineadas, MEGA ofrece varios métodos diferentes. Aquí ilustramos el método de máxima verosimilitud, comenzando con la función Modelos de MEGA, que permite seleccionar el modelo de sustitución más adecuado. Finalmente, MEGA proporciona una interfaz potente y flexible para el paso final, que en realidad dibuja el árbol para su publicación. Aquí se presenta un protocolo paso a paso con suficiente detalle para permitir comenzar con una secuencia de interés y construir un árbol de calidad de publicación que ilustre la evolución de un conjunto apropiado de homólogos de esa secuencia.

Protocolo

Un árbol filogenético es una estimación de las relaciones entre los taxones (o secuencias) y sus hipotéticos ancestros comunes (Nei y Kumar 2000 ; Felsenstein 2004 ; Hall 2011). Hoy en día, la mayoría de los árboles filogenéticos se construyen a partir de datos moleculares: secuencias de ADN o proteínas. Originalmente, el propósito de la mayoría de los árboles filogenéticos moleculares era estimar las relaciones entre las especies representadas por esas secuencias, pero hoy en día los propósitos se han expandido para incluir la comprensión de las relaciones entre las secuencias en sí, sin importar las especies hospedadoras, inferiendo las funciones de los genes que no se han estudiado experimentalmente (Hall et al. 2009), y mecanismos de dilucidación que conducen a brotes microbianos (Hall y Barlow 2006) entre muchos otros. La construcción de un árbol filogenético requiere cuatro pasos distintos:

- (Paso 1) identifica y adquiere un conjunto de secuencias de ADN o proteínas homólogas,
- (Paso 2) alinea esas secuencias,
- (Paso 3) estima un árbol a partir de las secuencias alineadas, y

- (Paso 4) presenta ese árbol de manera que se transmita claramente la información relevante a otros.

Normalmente, podría usar su navegador web favorito para identificar y descargar las secuencias homólogas de una base de datos nacional como GenBank, luego uno de varios programas de alineación para alinear las secuencias, seguido de uno de los muchos programas filogenéticos posibles para estimar el árbol y, finalmente, Un programa para dibujar el árbol para la exploración y publicación. Cada programa tendría su propia interfaz y su propio formato de archivo requerido, lo que obligaría a interconvertir archivos a medida que movía la información de un programa a otro. ¡No es de extrañar que el análisis filogenético a veces se considere intimidante!

MEGA (Tamura et al. 2011) es un programa integrado que realiza los cuatro pasos en un solo entorno, con una única interfaz de usuario que elimina la necesidad de interconvertir formatos de archivos. Al mismo tiempo, MEGA5 es lo suficientemente flexible como para permitir el uso de otros programas para pasos particulares, si así lo desea. MEGA5 es, por lo tanto, particularmente adecuado para aquellos que están menos familiarizados con la estimación de árboles filogenéticos.

Paso 1: Adquirir las secuencias

Irónicamente, el primer paso es el más exigente intelectualmente, pero a menudo recibe la menor atención. Si no se hace bien, el árbol será inválido o imposible de interpretar o ambos. Si se realiza con prudencia, los pasos restantes son operaciones sencillas, esencialmente mecánicas, que darán como resultado un árbol robusto y significativo.

A menudo, el investigador está interesado en un gen o proteína en particular que ha sido objeto de investigación y desea determinar la relación de ese gen o proteína con sus homólogos. La palabra "homólogos" es clave aquí. El supuesto más básico del análisis filogenético es que todas las secuencias en un árbol son homólogas, es decir, que descienden de un ancestro común. Los programas de alineación alinearán secuencias, homólogas o no. Todos los programas de construcción de árboles harán un árbol a partir de esa alineación. Sin embargo, si las secuencias no descienden realmente de un antepasado común, el árbol carecerá de sentido y puede ser bastante engañoso. La forma más confiable de identificar secuencias que son homólogas a la secuencia de interés es hacer una búsqueda de la Herramienta de búsqueda básica de alineación local (BLAST) (Altschul et al. 1997) usando la secuencia de interés como una consulta.

Paso 1.1

Cuando inicia MEGA5, se abre la ventana principal de MEGA5. En el menú Alinear , elija Do Blast Search. MEGA5 abre su propia ventana del navegador para mostrar una página BLAST de nucleótidos del Centro Nacional de Información Biotecnológica (NCBI). Hay un conjunto de cinco pestañas cerca de la parte superior de esa página (blastn, blastp, blastx, tblastn y tblastx). Por defecto, se selecciona la pestaña Blastn (Nucleótido estándar BLAST). Si su secuencia es la de una proteína, haga clic en la pestaña Blastp para mostrar la página BLAST de Proteína estándar.

Tenga en cuenta que NCBI cambia con frecuencia la apariencia de la página BLAST, por lo que puede diferir en algunos detalles de los que se describen aquí.

Hay un cuadro de texto grande (Introduzca el número de acceso ...) donde ingresa la secuencia de interés. Puede pegar la secuencia de consulta directamente en ese cuadro. Sin embargo, si su secuencia de consulta ya está en una de las bases de datos, puede pegar su número de acceso o número de gi. Si su secuencia de ADN es parte de una secuencia del genoma, puede ingresar el número de acceso del genoma y, a continuación, en los cuadros de la derecha (subintervalo de consultas) ingrese el rango de bases que constituyen su secuencia. (¡Realmente no quieres usar una secuencia de varios megabases como tu consulta!)

La sección central de la página le permite elegir las bases de datos que se buscarán y restringir esa búsqueda si así lo desea. El valor predeterminado es la colección de nucleótidos (nr / nt) , pero el cuadro de texto desplegable con triángulo le permite elegir entre un gran número de alternativas, por ejemplo, genomas humanos o genomas NCBI.

El cuadro de texto ***Organismos*** opcional le permite limitar su búsqueda a un organismo en particular o excluir a un organismo en particular. Por ejemplo, si su secuencia es de humanos, es posible que desee excluir a los humanos de la búsqueda, para que no capte muchas variantes humanas cuando esté realmente interesado en homólogos de otras especies. Para incluir más organismos, haga clic en el pequeño signo + junto al cuadro de opciones.

La opción Excluir le permite excluir, por ejemplo, muestras ambientales.

Paso 1.2: ¿Qué algoritmo de BLAST utilizar?

La sección inferior de la página le permite elegir la variante particular de BLAST que mejor se adapte a sus propósitos. Para los nucleótidos, las opciones son *megablast* para secuencias altamente similares, *megablast no contiguo* para secuencias más disímiles, o *blastn* para secuencias algo similares. El valor predeterminado es *blastn*, pero si solo está interesado en identificar homólogos estrechamente relacionados marque *megablast*. Esta es la primera opción que realmente exige un pensamiento. Las secuencias que estarán en tu árbol están muy determinadas por la elección que hagas en este momento.

En la parte inferior de la página, haga clic en el botón BLAST para iniciar la búsqueda; No marque la casilla "Mostrar resultados en una nueva ventana". Aparecerá una ventana de resultados, posiblemente con un gráfico que ilustra los dominios que se han identificado, normalmente con una declaración similar a "esta página se actualizará automáticamente en 5 segundos". Finalmente, aparecerá la ventana de resultados finales. El panel superior resume las propiedades de las secuencias de consulta y una descripción de la base de datos que se buscó. Debajo hay un gráfico que ilustra la alineación de los 100 "hits" principales (secuencias identificadas por la búsqueda). Desplácese hacia abajo para ver la lista de secuencias que producen puntuaciones de alineación significativas. Para cada secuencia, hay un número de acceso (un enlace al que se puede hacer clic), una descripción, un puntaje máximo (también un enlace al que se puede hacer clic), un puntaje total, una cobertura de consulta y un valor E y un ident. Usted usa esa información para decidir cuál de esas secuencias agregar a su alineación y, por lo tanto, incluir en su árbol.

La descripción ayuda a decidir si está interesado en esa secuencia en particular. Puede haber varias secuencias de la misma especie; ¿Desea todos esos o quizás solo un representante de una especie, o incluso de un género? Si está interesado en esa secuencia, consulte la cobertura de consultas. ¿Está interesado en un homólogo que solo se alinea con el 69 % de la consulta? Si no, ignora esa secuencia y continúa. ¿Está interesado en una secuencia que sea 100 % idéntica a su consulta? Si solo estás interesado en homólogos más distantes, puedes no estarlo. Si quieres el árbol más inclusivo posible, puedes ser. Usted debe decidir no hay algoritmo que pueda decirle qué incluir.

Si decides que estás interesado en una secuencia de aciertos, haz clic en el enlace " Puntuación máxima " para ver la serie de alineaciones. Lo que se ve depende de si su consulta fue una secuencia de ADN o una secuencia de proteínas.

Paso 1.3: Secuencias de ADN

La alineación de la consulta con el hit comienza con un enlace al archivo de secuencia a través de sus números de acceso y gi. Si ese enlace es a una secuencia del genoma, o incluso a un archivo grande que incluye secuencias de varios genes, no querrá incluir la secuencia completa en su alineación. Hay dos maneras de lidiar con el problema. 1) Observe la alineación en sí y observe el rango de nucleótidos en el sujeto. Asegúrese de observar si la consulta se alinea con la secuencia del sujeto (Strand = más / más) o con su complemento (Strand = más / menos). Haga clic en el enlace para que aparezca el archivo de secuencia. En la parte superior derecha, haga clic en el triángulo en el cuadro gris Cambiar región mostrada , luego ingrese el primer y último nucleótido del rango, luego haga clic en el botón Actualizar vista . En la región gris de la vista Personalizar , a continuación, marque la casilla Mostrar secuencia , y si Strand = más / menos también marque la casilla Mostrar complemento inverso , luego haga clic en el botón Actualizar vista. Finalmente, haga clic en el botón Agregar a la alineación (una cruz roja) cerca de la parte superior de la ventana. (2) Si su consulta es una secuencia de codificación o es otra característica notable, puede ver las *Features in this part of subject sequence:* justo debajo de la descripción de la secuencia con un enlace a la función. Haga clic en el enlace de la función para que aparezca el archivo de secuencia que ya muestra la región de interés. Verifique para asegurarse de que la secuencia mostrada sea el complemento inverso de la consulta, y si está marcada la casilla ***Mostrar complemento inverso*** en la región de vista Personalizar , actualice la vista, luego haga clic en el botón *Agregar a la alineación* (una cruz roja) cerca de parte superior de la ventana.

Paso 1.31. Cuando hace clic en el botón Agregar a alineación , se abre la ventana del Explorador de alineación de MEGA5 y la secuencia se agrega a esa ventana. Después de agregar una secuencia al Explorador de alineación, use la flecha de retroceso en la ventana de BLAST para regresar a la lista de secuencias homólogas y agregar otra secuencia de interés.

Paso 1.4: Secuencias de proteínas

La principal diferencia con las búsquedas de nucleótidos es que puede ver enlaces a números de acceso a varios archivos de secuencias de proteínas. Todos estos tienen la misma secuencia de aminoácidos, aunque sus secuencias de codificación subyacentes pueden diferir. Haga clic en cualquiera de los enlaces para que aparezca el archivo de secuencia de proteínas, luego haga clic en el botón Agregar a la alineación .

Cosas que pueden salir mal

1. Puede encontrar que todos los resultados que se obtienen de su búsqueda provienen de organismos muy relacionados; es decir, si su consulta era una proteína *E scherichia coli* , todos los impactos pueden ser de *E. coli, Salmonella* y especies estrechamente relacionadas. Si todos los resultados muestran una identidad máxima alta y está bastante seguro de que la secuencia se produce en secuencias más distantes, probablemente se haya topado con el máximo predeterminado de 100 secuencias objetivo. Repita la búsqueda, pero antes de hacer clic en el botón BLAST para iniciar el aviso de búsqueda, justo debajo de ese botón aparece una línea críptica "**+ *Parámetros de algoritmo***". Haga clic en el signo más para revelar otra sección de la página de configuración de BLAST. Establezca las secuencias de

destino máximo en un valor mayor y repita la búsqueda. También es posible que desee excluir algunas especies estrechamente relacionadas en la sección anterior ***Elegir conjunto de búsqueda***. Introduzca un taxón, por ejemplo, *E. Coli* , en la casilla y marque la casilla Excluir . Si desea excluir más de una especie, haga clic en el signo más a la derecha de Excluir para agregar otro campo. Puede excluir hasta 20 especies.

2. Cuando intenta volver a la lista de visitas, puede obtener una página que dice "***¡Sorry! Error: −400 Cache Miss"***. Haga clic en la flecha circular al lado del botón Agregar a la alineación. Serás enviado a la página principal de BLAST pero no te desesperes. En la parte superior derecha de esa página se encuentra la sección Sus resultados recientes. El enlace superior en la lista es su búsqueda más reciente. Simplemente haga clic en ese enlace para volver a sus resultados.

Cuando haya agregado todas las secuencias que desea, simplemente cierre la ventana del navegador MEGA5.

En la ventana del editor de alineación, guarde la alineación seleccionando ***Guardar sesión*** en el menú Datos. Me gusta usar un nombre como *Myfile_unaligned* solo para recordarme que las secuencias no se han alineado. El archivo tendrá la extensión ***.mas.***

Paso 1.5: Alternativas a MEGA5 para identificar y adquirir secuencias

- **Paso 1.51**. Puede acceder a NCBI BLAST a través de cualquier navegador web compatible con NCBI en `http://blast.ncbi.nlm.nih.gov/Blast.cgi`. En la sección BÁSICA de BLAST , haga clic en el enlace de *blast* de nucleótidos o proteína para llegar a la página idéntica a la descrita anteriormente. Todo es igual que cuando se usa el navegador de MEGA5 excepto que no puede hacer clic en un botón conveniente para agregar las secuencias al Editor de Alineación.

- **Paso 1.52**. Abra un nuevo archivo en un editor de texto. Puede usar el editor de texto integrado de MEGA5 seleccionando Editar un archivo de texto en el menú Archivo. Ese editor tiene varias funciones para editar secuencias moleculares, incluida la complementación inversa y la conversión a varios formatos comunes, incluido Fasta. Como alternativa, use un editor de texto. Guarde el archivo con un nombre significativo con la extensión.fasta, por ejemplo, *myfile.fasta*. ¡ No utilice Microsoft Word, Word Pad, TextEdit (Mac) u otro procesador de textos!

- **Paso 1.53**. Cuando haya identificado la secuencia que desea agregar y haga clic en el enlace para ir a la página de ese archivo de secuencia, ajuste la Región que se muestra y Personalice la vista si es necesario. Observe el enlace de Configuración de pantalla cerca de la parte superior izquierda de la página. El ajuste por defecto es GenBank (completo). Cambie eso a Fasta (texto), seleccione todo, cópielo y péguelo en el archivo del editor de texto. A medida que agrega secuencias al archivo, es conveniente, pero no necesario, dejar líneas en blanco entre las secuencias.

La siguiente sección explica cómo importar esas secuencias en el editor de alineación de MEGA5.

Paso 2: Alineando las secuencias

Si la ventana del Explorador de alineación no está abierta, en la ventana principal de MEGA5, elija ***Abrir un archivo/sesión*** en el menú Archivo. Elija el archivo de alineación MEGA5 (.mas) o el archivo de secuencia (.fasta) que guardó en el Paso 1. En el cuadro de diálogo resultante, elija Alinear .

El explorador de alineación muestra un nombre para cada secuencia a la izquierda, seguido de la secuencia, con residuos de color. Típicamente el nombre es muy largo. Ese nombre es lo que eventualmente aparecerá en el árbol, y los nombres largos generalmente son indeseables. Este es el momento para editar esos nombres, de hecho, es el único momento práctico para editar los nombres, así que no pierda la oportunidad. Simplemente haga doble clic en cada nombre y cámbielo a algo más adecuado.

Si tu secuencia es ADN, verás dos pestañas: Secuencias de ADN y Secuencias de proteínas traducidas. La pestaña de secuencias de ADN se elige por defecto. Haga clic en la pestaña Secuencias de proteínas traducidas para ver la secuencia de proteínas correspondiente.

Paso 2.1

Ahora es el momento de alinear las secuencias. Se proporcionan dos métodos de alineación: ClustalW y MUSCLE. Se puede usar cualquiera de las dos, pero en general es preferible MUSCLE. En la barra de herramientas, cerca de la parte superior de la ventana, la alineación de Clustal está simbolizada por el botón W y MUSCLE por un brazo con el puño cerrado. Haga clic en uno de esos botones o elija Clustal o MUSCLE en el menú Alineación. Si su secuencia es ADN, verá dos opciones: Alinear ADN y *Alinear codones*. Si su secuencia es una secuencia de codificación de ADN, es muy importante elegir *Alinear codones*. Eso asegurará que las secuencias estén alineadas por los codones, un enfoque mucho más realista que la alineación directa de las secuencias de ADN porque evita la introducción de espacios en las posiciones

que darían lugar a cambios de marco en las secuencias reales.

Paso 2.2

La elección de un método de alineación abre una ventana de configuración para ese método. Para MUSCLE, te recomiendo que aceptes la configuración predeterminada. Para ClustalW, la configuración predeterminada está bien para el ADN, pero para las proteínas, recomiendo cambiar la penalización de Apertura del intervalo de alineación múltiple a 3 y la penalización de la Extensión del intervalo de alineación múltiple a 1.8.

Paso 2.3

Haga clic en el botón Aceptar para iniciar el proceso de alineación. Dependiendo de la cantidad de secuencias involucradas y del método que elija, la alineación puede tomar desde unos pocos segundos hasta unas pocas horas. Cuando la alineación se haya completado, guardar la sesión. Me gusta guardar las secuencias alineadas con un nombre diferente, por lo tanto, si mi archivo original era *Myfile_unaligned.mas*, guardaría la secuencia alineada como *Myfile.mas*.

Paso 2.4

MEGA5 no puede usar el archivo ***.mas*** directamente para estimar un árbol filogenético, por lo que también debe elegir ***Exportar alineación*** en el menú Datos y exportar el archivo en formato MEGA5, donde obtendrá una extensión ***.meg***. Se le pedirá que ingrese un título para los datos. Puede dejar el título en blanco si lo desea, pero es útil agregar algún tipo de título que sea significativo para usted. Si se trata de una alineación de secuencias de ADN, también se le preguntará si están codificando secuencias.

Paso 2.5: Una alternativa a la alineación con MEGA5

Una vez que se complete la alineación, verá que se han introducido espacios en las secuencias. Esos huecos representan inserciones o eliminaciones históricas, y su propósito es alinear los sitios homólogos en la misma columna. Debe apreciarse que, al igual que un árbol filogenético es una "estimación" de las relaciones entre secuencias, una alineación es solo una estimación de las posiciones de las inserciones y eliminaciones históricas. La calidad de la alineación puede afectar la calidad de un árbol filogenético, pero MEGA5 no ofrece ninguna forma de juzgar la calidad de la alineación. El programa basado en la web *Guidance* (`http://guidance.tau.ac.il/`) proporciona cinco métodos diferentes de alineación, pero lo más importante es que evalúa la calidad de la alineación e identifica regiones y secuencias que contribuyen a reducir la calidad de la alineación.

Guidance requiere que las secuencias no alineadas se proporcionen en un archivo en formato Fasta. Si descargó las secuencias a través de su navegador web favorito y las guardó como un archivo .fasta, ese archivo se puede usar como entrada para ***Guidance***. Si usó MEGA5 para descargar las secuencias en el Explorador de alineación, puede exportar las secuencias no alineadas en formato FASTA seleccionando Exportar alineación en el menú Datos y luego eligiendo el formato FASTA. Si olvidó mantener las secuencias no alineadas, puede seleccionar todas las secuencias (Control-A), luego elija *Eliminar espacios* en el menú Editar antes de exportar las secuencias en formato FASTA.

Paso 3: estimar el árbol

Existen varios métodos ampliamente utilizados para estimar árboles filogenéticos (Neighbor Joining, UPGMA Maximum Parsimony, Bayesian Inference, and Maximum Likelihood [ML]), pero este artículo tratará solo uno: ML.

Paso 3.1

En la ventana principal de MEGA5, elija Abrir un archivo / sesión en el menú Archivo y abra el archivo ***.meg*** que guardó en el Paso 2.

Paso 3.2.

ML utiliza una variedad de modelos de sustitución para corregir múltiples cambios en el mismo sitio durante la historia evolutiva de las secuencias. El número de modelos y sus variantes pueden ser absolutamente desconcertantes, pero MEGA5 proporciona una función que elige el mejor modelo para usted. En el menú *Modelos* , elija ***Encontrar los mejores modelos de ADN / proteína (ML)...*** . Aparecerá un cuadro de diálogo de preferencias, pero está lo suficientemente seguro como para aceptar la configuración predeterminada. Haga clic en el botón Calcular para iniciar la ejecución. Los modelos pueden tomar bastante tiempo para considerar todos los modelos disponibles, pero una barra de progreso muestra cómo van las cosas.

Cuando se completa, aparece una ventana que enumera los modelos en orden de preferencia. Note el modelo preferido, luego estime el árbol usando ese modelo. Para los ejemplos a continuación, el modelo WAG + G + I fue el mejor.

Paso 3.3

En el menú de Filogenia, elija *Construir/Probar el árbol de máxima verosimilitud*... Aparecerá un diálogo de preferencias similar al de la figura 1.

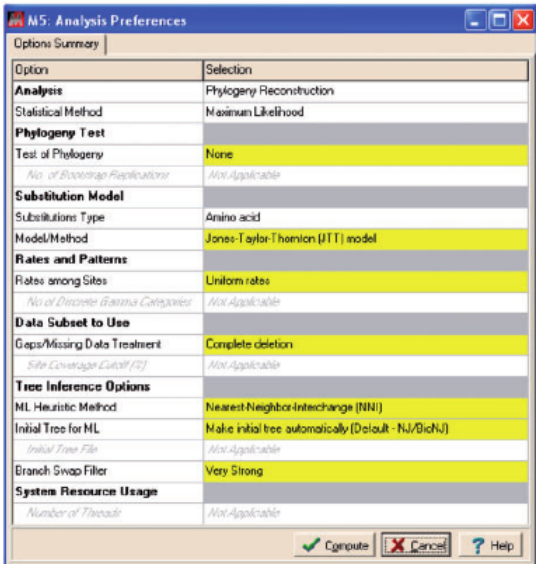

Figura 1: Preferencias de análisis de ML

Las áreas amarillas son parámetros que puedes modificar. Solo debe preocuparse por tres parámetros:

- 1) Modelo / Método ,
- 2) Tasas entre sitios , y
- 3) Tratamiento de datos faltantes / perdidos.

Haga clic en el extremo derecho de un área amarilla para mostrar un menú desplegable.

- Para Modelo / Método , se seleccionaría el modelo WAG.
- Para ***Rates among Sites*** , se seleccionará la opción Gamma distribuida con sitios invariantes (G + I). Junto con la selección de Modelos / Método anterior, esto coincide con el mejor modelo encontrado por la función Modelos.

Gaps/Missing Data Treatment determina cómo se manejan las ***Gaps***. La eliminación completa significa que MEGA5 ignora todas las columnas en las que hay un espacio en cualquier secuencia. A menos que haya muy pocas ***Gaps***, esa opción puede perder mucha información porque elimina muchos sitios de la consideración. Prefiero la opción de eliminación parcial en la que los sitios con datos faltantes se eliminan solo cuando sea necesario porque esa opción conserva más información.

Cuando se establecen las preferencias, haga clic en el botón de cálculo para calcular el árbol. Finalmente, se abrirá una ventana del explorador de árbol que muestra el árbol (fig. 2).

Paso 3.4

Es importante guardar el árbol, para que se pueda modificar más tarde si es necesario. Guarde el árbol desde el menú Archivo. El archivo tendrá la extensión *.mts*.

También puede exportar el árbol para introducir información en otros programas de dibujo de árboles (consulte el Paso 4). En el menú Archivo , elija Exportar árbol actual (Newick).

Paso 3.5: Estimando la Confiabilidad del Árbol

El árbol que estimaste es casi seguro que no es una representación verdadera de las relaciones históricas entre los taxones y sus ancestros. En cambio, es una estimación de esas relaciones. Como con cualquier estimación, es deseable conocer la *confiabilidad* de esa estimación. La forma más común de estimar la confiabilidad de un árbol filogenético es mediante el método **bootstrap**.

Para realizar la prueba de **bootstrap**, vuelva al cuadro de diálogo Preferencias de análisis que se muestra en la figura 1. En Prueba de filogenia, establezca Prueba de filogenia en "Método de **bootstrap**", luego establezca el número de réplicas de arranque a un número entero entre 100 y 2.000. Cuanto mayor sea el número, más tiempo tomará realizar la prueba. Haga clic en calcular. Una ventana con una barra de progreso muestra cómo está avanzando el análisis. Cuando se complete el análisis, aparecerá un árbol con números en cada nodo. Esos números, porcentajes de **bootstrap**, indican la confiabilidad del clúster que desciende de ese nodo; cuanto mayor sea el número, más confiable será la estimación de los taxones que descienden de ese nodo. En general, no tomamos en serio los nodos con una confiabilidad $<70\,\%$. La prueba de **arranque** no estima la confiabilidad general del árbol; en su lugar, estima la fiabilidad de cada nodo. Eso es realmente ventajoso porque le dice qué partes del árbol debe confiar y qué partes no debe tomar en serio.

Paso 3.6: Alternativas a MEGA5 para estimar el árbol

PhyML (`http://www.atgc-montpellier.fr/phyml/binaries.php`) (Guindon et al. 2010)

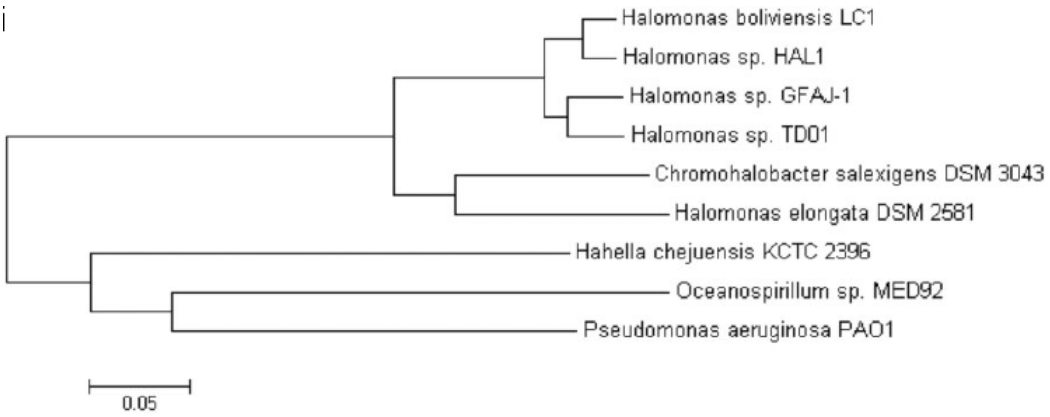

Figura 2: Un árbol ML.

es otro programa que estima árboles de ML, y también se puede usar en la web `http://www.atgc-montpellier.fr/phyml/`. SeaView (`http://pbil.univ-lyon1.fr/software/seaview.html`) (Gouy et al. 2010) es otro programa multipropósito que alinea secuencias, estima árboles mediante varios métodos y dibuja árboles.

Paso 4: Presentar el árbol

Un dibujo de un árbol filogenético transmite mucha información, tanto explícita como implícita. Parte de esa información implícita puede ser engañosa, por lo que es responsabilidad del investigador asegurarse de que la información transmitida, tanto explícita como implícita, sea correcta.

Un árbol filogenético consta de nodos externos (las puntas) que representan las secuencias reales que existen en la actualidad, nodos internos que representan ancestros hipotéticos y ramas que conectan los nodos entre sí. Las longitudes de las ramas representan la cantidad de cambio que se estima que ha ocurrido entre un par de nodos. Esa es la información explícita transmitida por un árbol de dibujo. El árbol en la figura 2 está en el formato de "filograma rectangular" en el que los nodos internos están representados por líneas verticales.

En la figura 2, el nodo más a la izquierda parece representar la raíz, el ancestro común del que descienden todas las secuencias. Esa información implícita es incorrecta y engañosa. De hecho, el método de LD, en común con los métodos de Neighbor Joining, Parsimony, and Bayesian Inference , es incapaz de determinar la raíz de un árbol; todos esos métodos estiman árboles sin raíces.

Paso 4.1

El formato de Radiación o *unrooted* que se muestra en la figura 3 es una mejor manera de dibujar un árbol no arraigado porque no permite que el espectador asuma una raíz desconocida. Para mostrar el árbol en formato de radiación, en la ventana del Explorador de árboles, elija Estilo de árbol/rama en el menú Ver, luego seleccione Radiación en el submenú. Debido a que el formato de Radiación no es familiar para muchos lectores, el formato de Filograma Rectangular predeterminado a menudo se publica, a pesar de que implica erróneamente un árbol arraigado. Un árbol enraizado proporciona dirección al proceso evolutivo, con el orden de descenso desde la raíz hacia las puntas. Suponiendo que la direccionalidad puede conducir fácilmente a suposiciones incorrectas sobre la historia evolutiva de esas secuencias. Para evitar la implicación injustificada de direccionalidad, es importante especificar en la leyenda de la figura o en el texto que el árbol está *unrooted*.

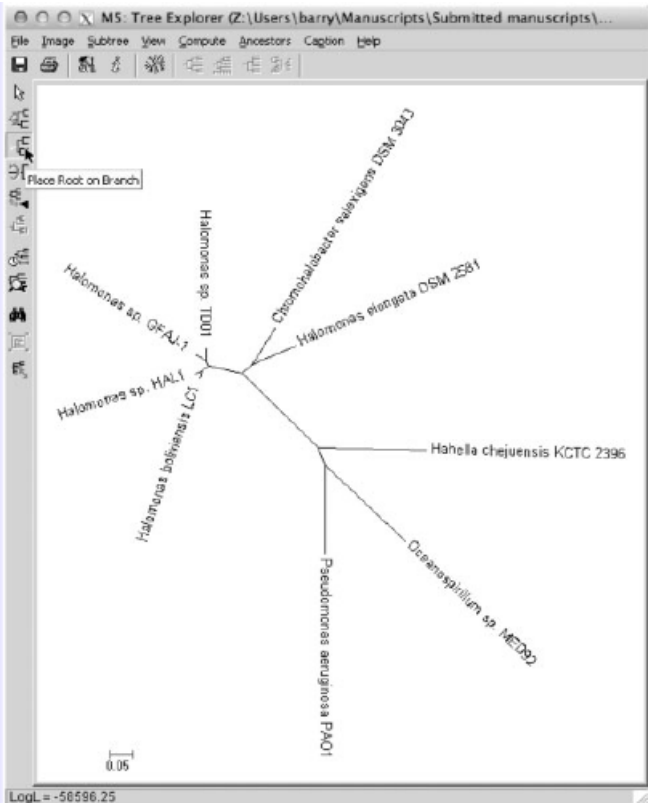

Figura 3: ML en formato de radiación (unrooted).

Paso 4.2

A menudo queremos presentar un árbol arraigado para sacar conclusiones que dependen del orden de descenso. Para hacer eso, necesitamos información adicional sobre las secuencias, información que es externa a las secuencias en sí, es decir, un grupo externo. Un grupo externo es una secuencia que está más distante relacionada con las secuencias restantes (en grupo) que entre sí. No podemos inferir un grupo externo del propio árbol, por lo que recurrimos a otra información. Para las secuencias en la figura 2, sabemos que Pseudomonas aeruginosa pertenece al orden Pseudomonadales, mientras que los organismos restantes pertenecen al orden Oceanospirillales, ambos de la clase Gammaproteobacteria. Así, p. Aeruginosa es un grupo externo legítimo para las secuencias restantes.

Paso 4.21.

Podemos enraizar el árbol en p. aeruginosa mediante la herramienta de enraizamiento que se encuentra en la barra lateral del recuadro del Explorador de árboles (fig. 3). En la vista de Filogramas rectangulares o en la vista de Radiación, mientras se selecciona la herramienta de enraizamiento, haga clic en la rama que lleva a P. aeruginosa para enraizar el árbol en esa secuencia como se muestra en la figura 4.

El árbol enraizado en la figura 4 ahora implica correctamente la dirección de evolución de esas secuencias. Cuando se publique el árbol, sería importante especificar que el árbol estaba enraizado en P. aeruginosa.

Paso 4.3

MEGA5 proporciona una variedad de herramientas para manipular la apariencia del árbol. Ya mencioné los formatos de Filogramas Rectangulares y Radiación. Aunque esos formatos parecen ser muy diferentes, son dibujos de exactamente el mismo árbol. En ambos casos, las ramas se dibujan, de modo que las longitudes de las líneas son proporcionales a las longitudes de las ramas. Esos formatos hacen obvio que ha habido muchos más cambios entre Hahella chejuensis KCTC 2396 y Oceanospirillum sp. MED292 a que ha habido entre Halomonas boliviensis LC1 y Halomonas sp. HAL1.

Paso 4.4

El cladograma, o formato de topología única, es otro formato importante. Seleccione Topología solo en el menú Ver para ver el árbol dibujado, de modo que las longitudes de las líneas de derivación no estén relacionadas con las longitudes de rama. ¿Por qué alguien querría eliminar esa información del dibujo? En algunos árboles, hay algunos nodos que están separados por ramas muy cortas, mientras que otros están separados por ramas muy largas. Cuando las ramas son demasiado cortas, puede ser imposible ver el orden de bifurcación o la topología. El formato Solo topología permite ver el orden de bifurcación de todo el árbol.

Paso 4.4.1.

Pero ¿qué pasa con esas longitudes de rama? ¿Realmente queremos perder esa información? No, no lo hacemos, así que simplemente podemos etiquetar las ramas con sus longitudes de rama. Para hacerlo, elija Mostrar / Ocultar en el menú Ver y seleccione Longitud de rama en el submenú.

Paso 4.5: Publicando el Árbol

Aunque puede imprimir el árbol para sus propios fines, para publicarlo, debe guardarlo en un formato de archivo de gráficos que sea aceptable para la revista. El formato de documento portátil (PDF) es casi universalmente aceptable. Seleccione Guardar como archivo PDF en el menú Imagen .

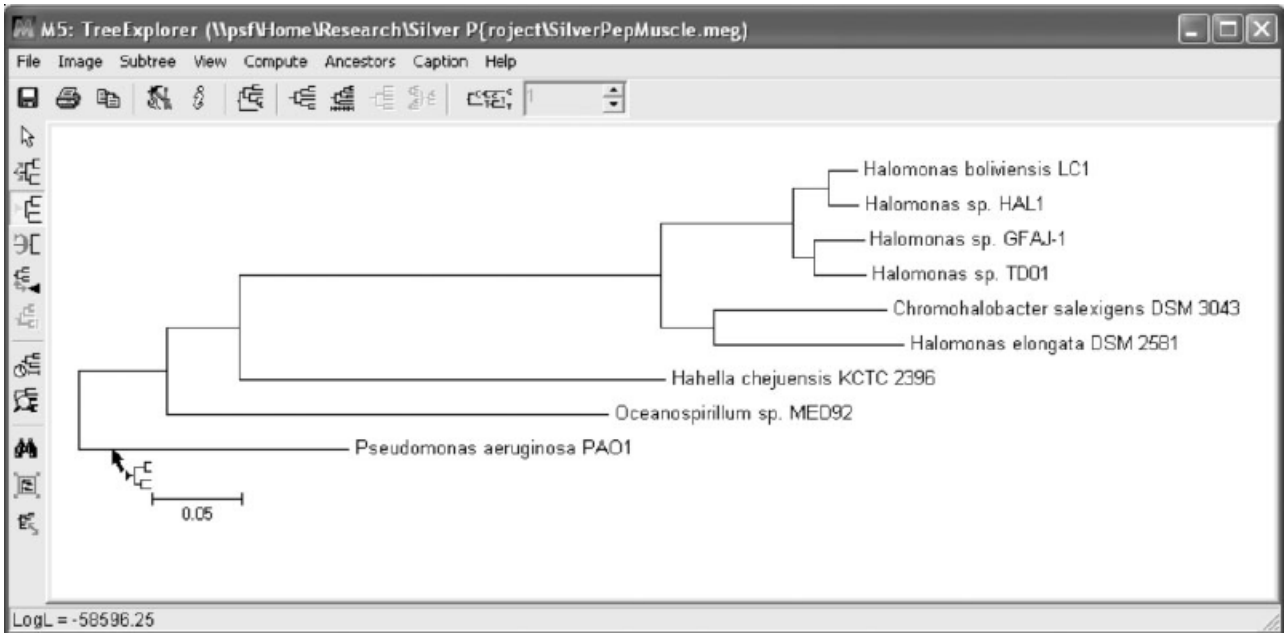

Figura 4: Árbol ML arraigado en Pseudomonas aeruginosa.

Es posible que desee manipular el dibujo de manera que MEGA5 no lo proporcione: poner en negrita algunos nombres de secuencia para llamar la atención, agregar una flecha, etc. Estas manipulaciones se realizan con un programa de dibujo de gráficos. La mayoría de los programas de dibujo aceptan archivos en formato PDF, pero en caso de que no lo hagan, MEGA5 también le permite guardar la imagen en formatos PNG y Meta File mejorados.

Paso 4.6:

Alternativas a dibujar árboles dentro de MEGA5 El programa de dibujo en árbol FigTree (`http://tree.bio.ed.ac.uk/software/figtree/`) es un programa con todas las funciones que ofrece muchas capacidades del sistema MEGA5 y muchas otras capacidades. Está disponible para los sistemas operativos Windows y Mac como un ejecutable Java que se ejecutará en cualquier sistema operativo, incluido Linux. Para importar un árbol a FigTree, expórtelo como un archivo de Newick como se describe en el Paso 3.