TEMA 3.3 Obtención de árboles filogenéticos mediante el método de máxima parsimonia

Contacto: Maite Aguado (maite.aguado@uam.es)

PROGRAMAS NECESARIOS

Para realizar análisis de Máxima Parsimonia con nuestras secuencias una vez alineadas trabajaremos con el programa TNT ("Tree analysis using New Technology").

Con TNT podemos realizar análisis de Máxima Parsimonia en los que incluyamos un número elevado de taxones (ej. 300-500). Una de las grandes ventajas de TNT es la rapidez con la que obtiene los resultados.
El programa TNT es de acceso libre gracias a la Hennig Society y se puede descargar de la red en la página:

http://www.cladistics.com/

En el siguiente enlace puedes encontrar un breve manual para el manejo del TNT:

http://tnt.insectmuseum.org/index.php/Manual

Más información sobre el programa se puede consultar en:

http://www.zmuc.dk/public/phylogeny/tnt/

METODOLOGÍA Y PRÁCTICA

I. Preparando el formato para TNT

TNT trabaja con formato NEXUS con alguna variación respecto al formato general (tema 3.2).

Paso 1. Para preparar la matriz y que sea leída correctamente por el programa abre la el archivo “Glaucoreseda_ITS_LF.nex” con un editor de texto (WordPad para PC, o TextWrangler para MAC; OJO es muy importante que no lo abráis con Word). Al inicio de la matriz sustituye todos los comandos desde #NEXUS hasta MATRIX (ambos incluidos) por los siguientes comandos:

nstates dna;
xread
‘filename’
N M

Escribe estos comandos al inicio de tu matriz sustituyendo:
- ‘filename’ por el nombre que le quieras dar al archivo, pon ‘Glacuoreseda_ITS_LF’
- N por el número de caracteres total de la matriz, 1375 en nuestro caso.

- M por el número de muestras incluidas en la matriz, 17 en nuestro caso.

Si además de secuencias de ADN hubiera datos morfológicos (0/1), entonces en lugar de “nstates dna” habría que escribir “nstates 32;”

A continuación, debes añadir el símbolo > delante del nombre cada secuencia (al igual que en el formato fasta, véase tema 3.2.

Por último elimina el comando end; con el que termina la matriz.

La matriz adaptada debería empezar de la siguiente manera:

```
nstates dna;
xread
'Glaucoreseda_ITS_LF'
1375 17
>O_dre_DQ987166      TCGAAACCTGACCAAGGAGTGCGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>O_lin_FJ212178      TCGAAACCTGACCAAGGAGTGCGACCCGAGAACAAGTATTGTGATGCGGAAGCCGGCAGGCC
>R_alb_DQ987192      TCGAAACCTGACCAAGGAGAGCGACCCGAGAACAAGTATTRTGAYGCGGAAATCGGCAGGCC
>R_bat1_GQ891132     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_bat3_DQ987183     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCRGAAACCGGCAGGCC
>R_com1_DQ987172     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_com1_GQ891136     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_gla1_GQ891137     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_gla3_GQ891139     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_gla4_GQ891140     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGACAKGCC
>R_gla7_DQ987181     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGACAGGCC
>R_gre1_GQ891150     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_gre3_GQ891151     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_luteola_DQ987187  TCGAAACCTGACCAAGGAGAGCGACCCGAGAACAAGTATTGTGATGCGGGAACCGGCAGGCC
>R_vir2_GQ891162     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_vir5_DQ987176     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
>R_vir8_GQ891169     TCGAAACCTGACCAAGGAGTTTGACCCGAGAACAAGTATTGTGATGCGGAAACCGGCAGGCC
```

Una vez que hayas comprobado que la matriz está bien adaptada, guárdala como Glaucoreseda_ITS_LF.

II. El análisis de Máxima Parsimonia en TNT

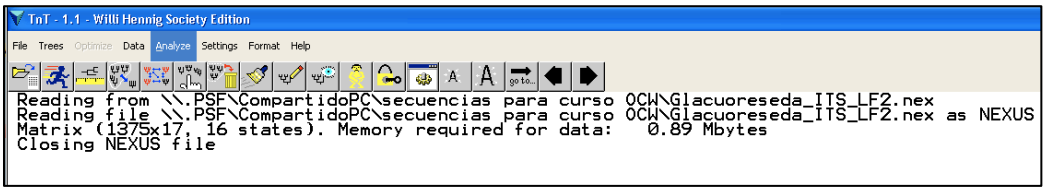
Paso 1. Abre el programa TNT. En primer lugar, debemos aumentar la memoria redirigida al programa. Ve a la función *Settings* en el menú principal y selecciona la opción “Memory”. Se abrirá una ventana en la que debes modificar los siguientes parámetros sustituyendo los valores que aparecen por defecto por los que te indicamos aquí abajo:

Max.Tree: 100000

General RAM: 500 MegaBytes

Paso 2. En la opción *File* del menú principal, ve a la carpeta donde hayas guardado la matriz combinada que has adaptado al formato TNT (Glacuroreseda_ITS_LF_TNT). Inicialmente te aparecerá vacía, para poder visualizar todos los archivos de esa carpeta has de señalar en la

pestaña “tipo de archivo” que aparece abajo la opción “ALL files” (donde por defecto aparece TNT files).



Se creará automáticamente un archivo con este nombre y de extensión “.out” en el que quedará almacenado el histórico de funciones realizadas durante todo el análisis

Paso 3. A continuación debemos crear un archivo donde posteriormente almacenaremos todos los árboles resultado de la búsqueda en formato paréntesis. Para ello, selecciona en *File* “Tree saved file/Open parenthical”

Se creará automáticamente un archivo de extensión “.tree” en el que posteriormente almacenaremos todos los árboles obtenidos en la búsqueda.

Paso 4. Antes de iniciar el análisis debemos especificar algunas opciones. Así, en *Format* selecciona “Use Taxon names”. Esta opción hará que el programa identifique cada secuencia con el nombre que tú le has asignado. La opción que aparece por defecto asigna a cada secuencia un número en vez de el nombre que asignaste.

Paso 5. En *Format* selección “Data Format/Read Gaps as missing”. Esta opción fuerza a considerar todos los *gaps* introducidos fruto del alineamiento como dato perdido. Es importante señalarlo dado que la opción que TNT asume por defecto es considerar estos *gaps* como un quinto carácter (véase tema 3.2). Ello implica que, cada vez que en una muestra aparece un *gap* este es considerado como un cambio respecto a los nucleótidos que tengan el resto de las muestras en esa posición.

Paso 6. Por último, debemos también especificar el *outgroup*. La mayor parte de los programas de filogenias seleccionan por defecto la primera secuencia de la matriz como *outgroup*, en TNT podemos cambiarla por la secuencia donde deseemos enraizar los árboles. Selecciona en *Data* la opción “Data/outgroup taxon” y selecciona la muestra “R. alba” como outgroup.

Paso 7. A continuación podemos proceder ya a establecer los parámetros de la búsqueda. Selecciona en *Analyze* la opción “Tradicional Search”. A continuación se abrirá una ventana en la que aparecen los parámetros basicos de la búsqueda.

En “Starting trees” se define el modo en el que se obtiene el árbol de inicio de la búsqueda en cada réplica (véase tema 2.1). Deja seleccionada la opción que aparece por efecto “Wagner Trees”.

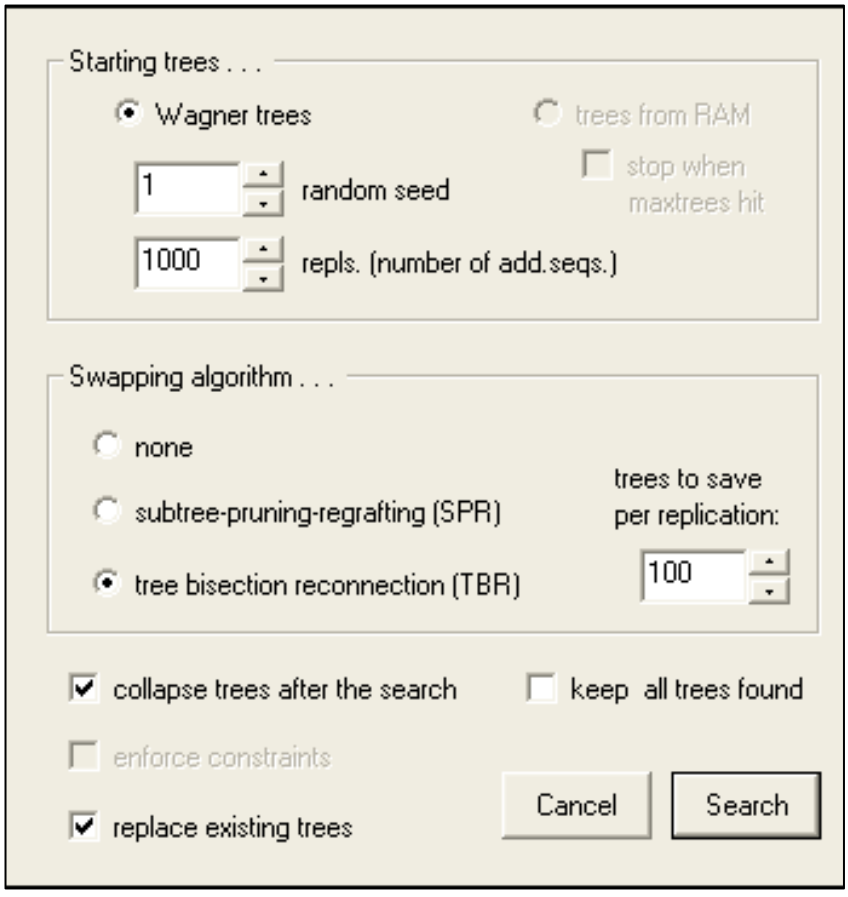
Aquí debes especificar también el número de réplicas. Cada réplica realiza la búsqueda de los árboles más parsimoniosos a partir de un árbol de inicio diferente. Dado que en las búsquedas heurísticas no prospectamos todo el universo de árboles posibles (veáse tema 2.1), cuantas más replicas realicemos menos posibilidades de dejarnos

árboles parsimoniosos no muestreados. Escribe 1000 réplicas. [Ten en cuenta que para una publicación deberías utilizar al menos 1.000.000 de réplicas].

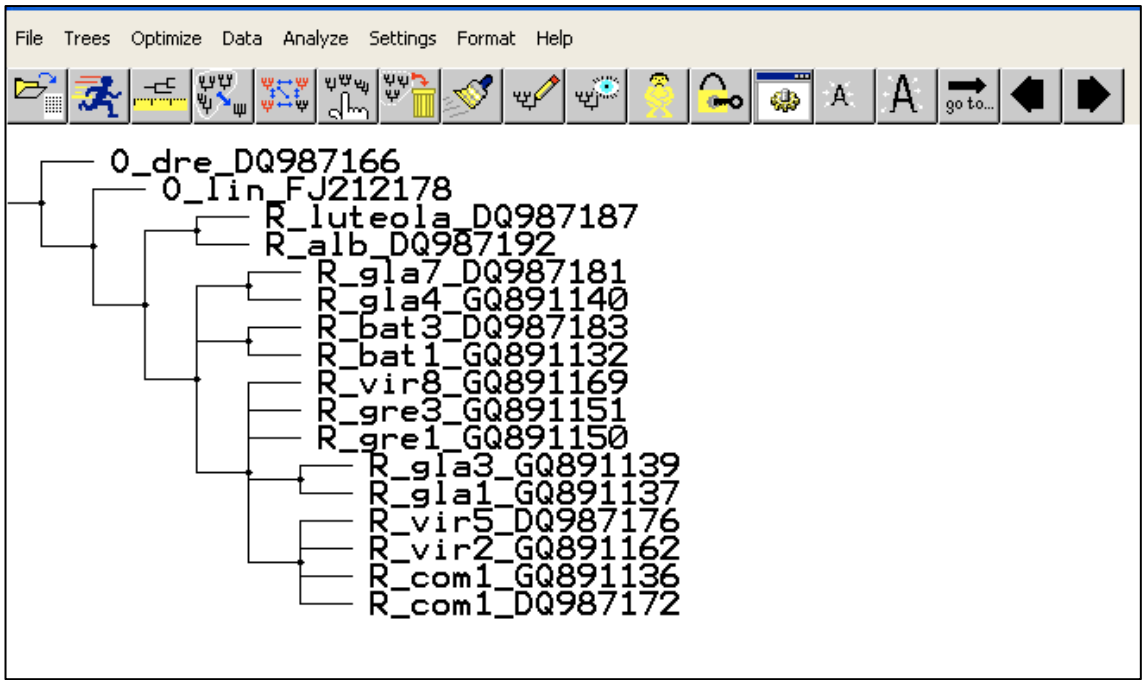
En “Swapping algorthim” podemos seleccionar el algoritmo elegido para realizar el reajuste de las ramas (véase tema 2.1). Selecciona “TBR” (Tree Bisection Reconnection). Una opción interesante para este tipo de búsquedas es la de restringir el número de árboles más parsimoniosos guardados por réplica. En TNT se puede hacer en “Trees to saved per replication”, nosotros especificaremos 100. De este modo, la potencia y tiempo de cálculo se reduce significativamente a la vez evitamos quedar atrapados en mínimos locales.

Selecciona por último la ocpión “collapse trees after the search” (elimina las ambigüedades en el proceso de optimización de caracteres).

Finalmente, para iniciar la búsqueda pulsa el botón “Search”.



Paso 8. Terminada a búsqueda en la pantalla de inicio aparece el número de árboles más parsimoniosos guardados junto con el número de pasos. En nuestro caso, sólo ha encontrado un árbol más parsimonioso que aparecerá automáticamente en la pantalla.



Paso 9. Al haber restringido la búsqueda, limitando el número de árboles parsimoniosos guardados por réplica, debemos completar la búsqueda. Para ello, ve a *Analyzed* y selecciona "Tradicional Search".

En esta ocasión volveremos a establecer las especificaciones del paso 7 salvo en "Starting trees" donde debemos seleccionar "Trees from RAM". Esta opción permite iniciar las búsquedas en cada réplica a partir de cada uno de los árboles más parsimoniosos que hemos encontrado en la búsqueda anterior.

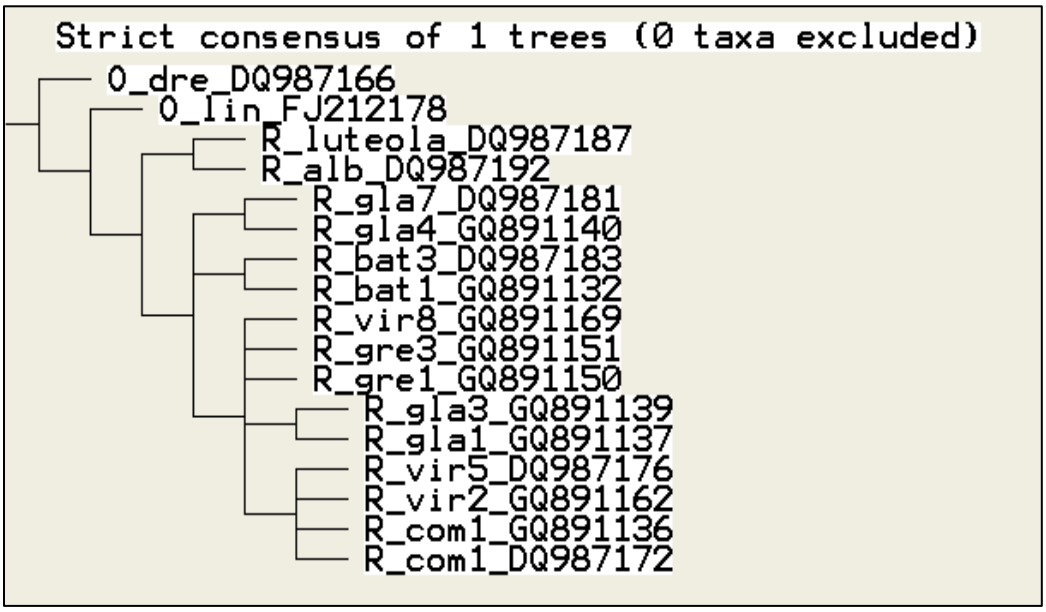
Paso 10. Al terminar la búsqueda, en la pantalla de inicio aparecerán los términos de la búsqueda junto con el número total de árboles más parsimoniosos encontrados y su longitud (número de pasos). Par poder ver todos los pulsa en el icono:

Paso 11. Una vez completada la búsqueda, debemos guardar todos los árboles obtenidos en el archivo ".tree" que hemos creado al incio. Para ello, ve a la opción *File* y en "/Tree saved file/save trees to open file" selecciona "All trees".

En nuestro caso tan sólo hemos encontrado un árbol más parsimonioso, (con una longitud de 194 pasos). Sin embargo, lo más habitual es encontrar numerosos árboles igualmente parsimoniosos. Cada uno de estos árboles refleja una historia filogenética diferente pero todos ellos son igualmente plausibles según el criterio de máxima parsimonia. Por ello, no podemos decantarnos por uno u otro. Sin embargo, para poder reconstruir la historia filogenética del grupo es deseable analizar un único árbol en vez de miles.

Paso 12. Para ello, podemos elaborar un consenso. En *Trees,* señala la opción "consensus" y selecciona "Strict". Inmediatamente aparecerá en otra ventana un árbol de consenso estricto construido a partir de todos los árboles más parsimoniosos encontrados en la búsqueda. Puedes guardar este árbol en formato gráfico para poder editarlo posteriormente en programas de edición de imágines como adobe illustrator o en PowerPoint insertándolo como imagen. Para ello, cuando tengas en la pantalla el árbol, pulsa la tecla "m" y guarda el árbol como "Glaucoreseda_ITS_LF_MPtree". Para volver a la pantalla de inicio pulsa la tecla "esc".

En nuestro caso, al haber obtenido un único árbol, no tiene sentido realizar un consenso.



Ejercicio 3.3.1. Abre el documento "Glacuoreseda_ITS_LF", haz un análisis de máxima parsimonia.

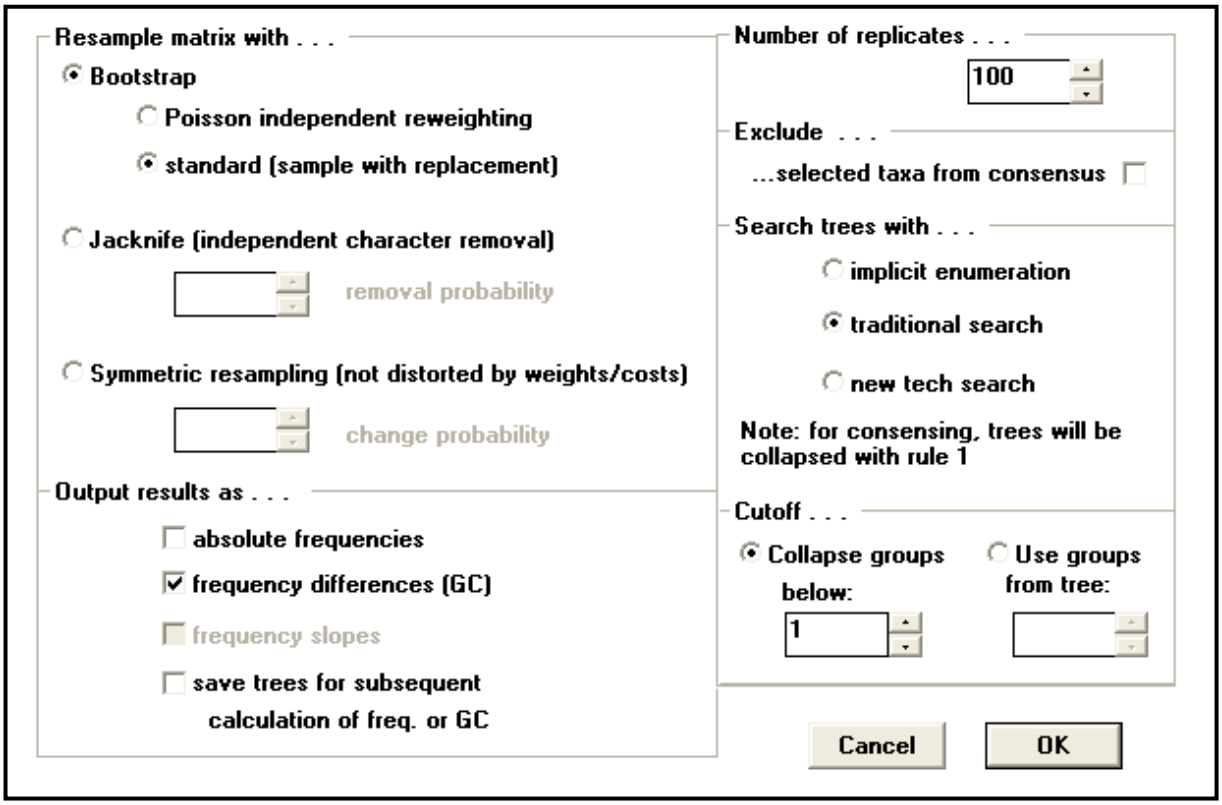
TNT puede trabajar con *nuevos a*lgoritmos "New Technology search" en "Analyze" especialmente diseñados para matrices muy grandes, con un número elevado de terminales (más de 100). Estos algoritmos son el "ratchet", "tree searches" , "drift" y "Tree fusing".

Paso 13. Para obtener los estadísticos descriptivos de la búsqueda -CI (índice de consistencia) y RI (índice de retención)- en TNT necesitáis llamar a un ejecutable ".run" que aparecerá (bajo distintos nombres en función de la versión de TNT que os hayáis descargado) en la misma carpeta donde tenéis el ejecutable del programa. Antes de ejecutarlo, debéis copiar el archivo .run en la misma carpeta donde esté la matriz de datos y, si fuera necesario, cambiarle el nombre a **stats.run**. Una vez allí, escribe en la línea de comando del programa TNT que aparece en la parte inferior de la pantalla el siguiente comando: run stats.run. A continuación te aparecerán dichos valores en la pantalla de inicio. Estos valores permiten estimar la fiabilidad de los resultados de la búsqueda ya que estiman de modo indirecto el nivel de homoplasia reflejado en los árboles (cuanto

más próximo a 1 sea el índice de consistencia, menor será el valor de homoplasia).

III. Cálculo de valores de soporte de ramas en TNT

Paso 1. Los árboles más parsimoniosos encontrados en la búsqueda realizada agrupan las secuencias en clados y a su vez los clados entre sí en función de los caracteres derivados y compartidos (sinapomorfías, véase tema 2.1) de manera que se minimice el número de cambios total. Sin embargo, la aparición de un clado tanto en uno de los árboles más parsimoniosos, como en un árbol de consenso presentan apoyo estadístico. Para poder calcular los valores de apoyo a los clados (Jacknife, Bootstrap y Symmetric Resampling) se utilizan técnicas de remuestreo. Para ello, selecciona en *Analyze* la opción "Resampling".



Se abrirá una nueva ventana en la que debemos especificar los parámetros para el remuestreo. En "resample matrix with…": seleccionar "Bootstrap" y dentro de éste "standard"

En "number of replicates" debemos especificar el número de veces que queremos que se realice el muestreo. Al igual que en la búsqueda de árboles más parsimoniosos, cuanto mayor sea el número de réplicas, mejor. Pon 1000 réplicas. [Ten en cuenta que para una publicación deberías utilizar al menos 10.000 de réplicas].

La opción "Cutoff" permite determinar el colapso de clados por debajo de un valor de apoyo determinado. Por defecto aparece 1, aunque para facilitar la interpretación suele ser útil especificar "50", de modo que

aquellos clados cuyo apoyo sea inferior a 50 no aparecerán resueltos en al árbol.

Hechas las especificaciones, pulsa el botón “OK” para que se inicie el remuestreo.

Paso 2. Finalizado el remuestreo aparecerá una pantalla con un árbol en el que cada clado estará apoyado por un número que representa el apoyo. Por consenso para la interpretación de las relaciones filogenéticas, se suelen reconocer únicamente los clados cuyo apoyo sea al menos un 80 de *boostrap*. Guarda el árbol como “Glaucoreseda_ITS_LF_MPbs” utilizando el comando “m”.

Paso 3. Para volver a la pantalla inicial pulsamos “esc” en el teclado.

En TNT también es posible calcular los valores de Bremer. Este tipo de apoyo de ramas se calcula forzando búsquedas heurísticas en las que los árboles guardados tengan cada vez un número de pasos mayor que el de los árboles más parsimoniosos previamente encontrados. Es decir, forzando que las nuevas búsquedas nos devuelvan cada vez árboles menos parsimoniosos. De esta manera, los clados que habiendo aparecido en la búsqueda inicial desaparecen al aumentar en un paso la longitud final del árbol, se consideran clados poco apoyados (y vendrían determinados por un valor de 1).

Por el contrario, aquellos clados que siguen recuperándose independientemente del número de pasos que añadamos (2, 3, 4 más), se consideran clados robustos (y vendrían apoyados respectivamente por valores de 2, 3 y 4 respectivamente). Para calcular los valores de Bremer, hay que volver a correr la búsqueda heurística después de indicar que retenga los árboles 1, 2, etc.. pasos mas largos. Para ello:

Paso 4. Busca los árboles mas cortos en una búsqueda heurística normal reteniendo como máximo 1000 árboles (setting> memory >max trees).

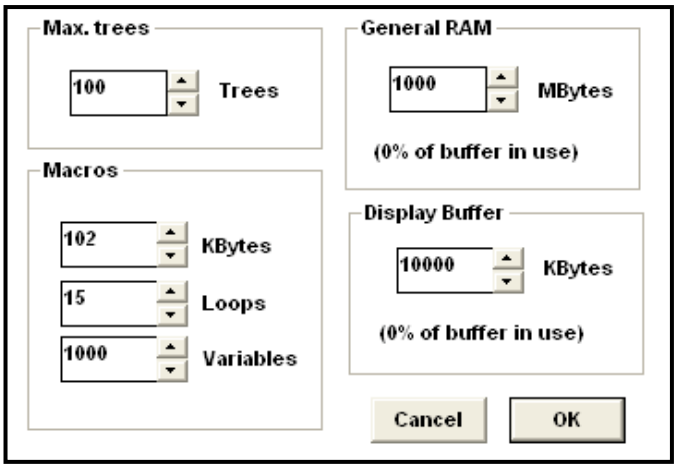
Paso 5. Aumentar el subóptimo en un paso (Analyze/suboptimal) y hacer de nuevo la búsqueda eliminando la opción "replace existing trees". Para evitar que se colapse la memoria muy rápido con árboles subóptimos recomendamos la siguiente secuencia.

Paso 6. Volver a repetir la misma operación aumentando el subóptimo a 3 y el max trees a 2000.

Paso 7. Repetimos de nuevo esta vez son subóptimo a 5 y max trees a 4000. Finalmente utiliza en comando bremer supports en trees para obtener los valores hasta un >5.

Paso 8. Podemos seguir repitiendo las búsquedas aumentando sucesivamente los subóptimos y los max trees hasta alcanzar valores de Bremer altos.

Paso 9. Para ver los valores de bremer seleccionar Trees/Bremer supports. Podemos guardar el gráfico con el comando “m” como “Glaucoreseda_ITS_LF_MPbremer”


Max. trees
100
Trees
General RAM
1000
MBytes
(0% of buffer in use)
Macros
102
KBytes
15
Loops
1000
Variables
Display Buffer
10000
KBytes
(0% of buffer in use)
Cancel
OK

Ejercicio 3.3.2. Calcula los apoyos para las ramas y guarda el archivo en formato metafile.

Ejercicio 3.3.3. Repite ahora cada paso de la práctica con cada una de las matrices por separado; "ITS" y "LF2". ¿Observas alguna diferencia en la topología del árbol que has obtenido al utilizar la matriz combinada con las de los árboles obtenidos al utilizar las matrices de los genes por separado? ¿A qué puede ser debido?