TEMA 3.5. Obtención de árboles mediante el método de inferencia bayesiana

Contacto: Isabel Draper (Isabel.draper@uam.es)

PROGRAMAS NECESARIOS

(1) Para la obtención de árboles mediante Inferencia Bayesiana vamos a utilizar el programa MrBayes (Huelsenbeck & Ronquist, 2001), que puede ser descargado gratuitamente desde:

http://mrbayes.sourceforge.net/

En esa misma página se explica cómo instalar el programa y se puede acceder al manual de instrucciones. Para realizar la siguiente práctica ten en cuenta que el formato de archivo necesario para este programa es el de tipo nexus. Además, no olvides que este archivo debe estar en la misma carpeta en la que está el ejecutable de MrBayes.

Programas alternativos:

BEAST (Drummond & Rambaunt 2007) es otro programa gratuito que utiliza cadenas de Markov para la aplicación de inferencia bayesiana en el análisis de secuencias moleculares. Este programa está orientado a filogenias con tiempos de divergencia.

En el siguiente enlace puedes encontrar el link a las descargas además de un manual para su uso y algunos otros enlaces de interés.

http://beast.bio.ed.ac.uk/Main_Page

METODOLOGÍA Y PRÁCTICA

I. Preparación del archivo nexus para MrBayes

Paso 1. MrBayes trabaja con formato nexus. Para preparar tu matriz de datos, comprueba que el archivo tiene la siguiente estructura inicial, teniendo especial cuidado de respetar los puntos y coma finales [OJO, no olvidar ningún ";"]:

```
# nexus
begin data;
dimensions ntax=N nchar=M
format datatype=X
interleave=Y
gap= — missing=?;
matrix
```

Abre la matriz “Glaucoreseda_ITS_LF” y escribe estos comandos al inicio de tu matriz sustituyendo:

N por el número de muestras
M por el número de caracteres de cada muestra
X por el tipo de datos:

dna (como en nuestro caso)
standard (si son datos morfológicos)
restriction (si son gaps codificados)
mixed (si se trata de una matriz mixta en la que se combinan diferentes tipos de datos). En este caso habría que indicar a continuación y entre paréntesis qué caracteres son de cada tipo, por ejemplo:
format datatype = mixed (dna:1-25, restriction:26-30)

Y por NO (ya que las secuencias en nuestra matriz no están divididas en bloques, sino que se encuentran como una cadena seguida).

Recuerda también que las secuencias no pueden empezar por el símbolo > y que después de la matriz se debe indicar que acaba el archivo mediante (véase tema 3.2):

;end; [end of file]

Con este tipo de información en el archivo, MrBayes va a tratar por defecto todos los datos de ADN como una única partición. Sin embargo, si hemos combinado varios genes suele ser conveniente dividir los datos en particiones (una para cada región genética). De este modo, podemos analizarlas por separado (por ejemplo si cada región presenta un modelo evolutivo diferente). Para separar los datos en particiones, lo más conveniente es añadir, utilizando un editor de texto (WordPad para PC, o TextWrangler para MAC; OJO es muy importante que no lo abráis con Word), un bloque informativo más en el archivo que va a analizar MrBayes (a continuación de la matriz de datos). En concreto se tiene que añadir la siguiente información (cada comando seguido de un signo de igualdad y la información que se quiera asociar al comando, y cada línea siempre terminada en punto y coma):

begin mrbayes
[para indicar que se trata de un bloque informativo para ese programa]
charset
[para asociar un nombre con un conjunto de caracteres]
partition favored
[para definir una partición asociada a cada nombre dado]
set partition
[para indicar a MrBayes que queremos trabajar con nuestras particiones en lugar de con la partición por defecto; las particiones a las que hace referencia pueden ser las nombradas con *charset* o indicarse con los números de caracteres: por ejemplo *partition favored = 4: 1-275, 276-427, 428-639, 640-1375;*]
end
[para cerrar el bloque]

En nuestra matriz combinada de ITS y trnLF el bloque debería ser así:

```
begin mrbayes;
charset ITS1 = 1-275;
charset 5.8S = 276-427;
charset ITS2 = 428-639;
charset trnLF = 640-1375;
partition favored = 4: ITS1, 5.8S, ITS2, trnLF;
set partition = favored;
end;
```

Ejercicio 3.5.1. Abre la matriz "Glaucoreseda_ITS_LF" y prepárala para su análisis en MrBayes.

Ejercicio 3.5.2. De acuerdo a estas directrices, abre la matriz combinada de ITS y LF (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), establece las particiones y guarda el archivo bajo el nombre

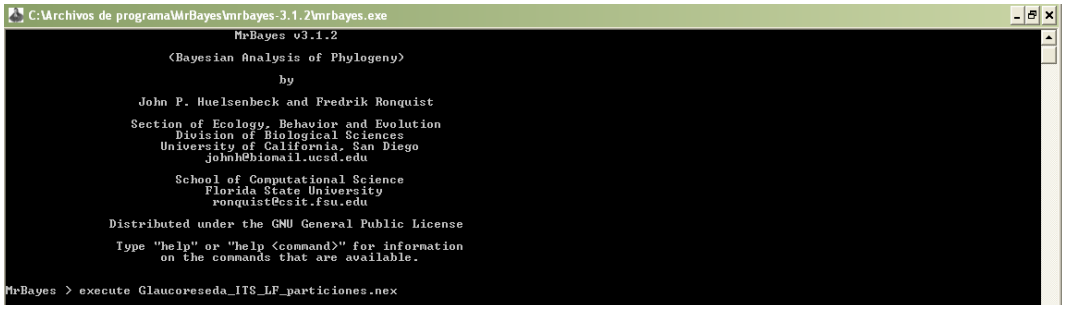
"Glaucoreseda_ITS_LF_particiones.nex".

II. Analásis de Inferencia Bayesiana con MrBayes

A continuación hacemos un breve resumen de los comandos básicos del programa. Te proponemos que realices una reconstrucción filogenética utilizando para ello A) la matriz combinada de ITS y LF con las particiones especificadas (Glaucoreseda_ITS_LF_particiones.nex), y B) la misma matriz combinada pero incluyendo al final los gaps codificados según Simmons y Ochoterena (2000) (Glaucoreseda_ITS_LF_gaps_particiones.nex). De esta manera podrás comprobar la influencia que tiene considerar o no la información de los gaps a la hora de reconstruir la filogenia.

Guarda el archivo que vayas a utilizar (Glaucoreseda_ITS_LF_particiones.nex y Glaucoreseda_ITS_LF_gaps_particiones.nex) en la carpeta raíz del programa de MrBayes. Ten en cuenta que el nombre del archivo no puede contener espacios.

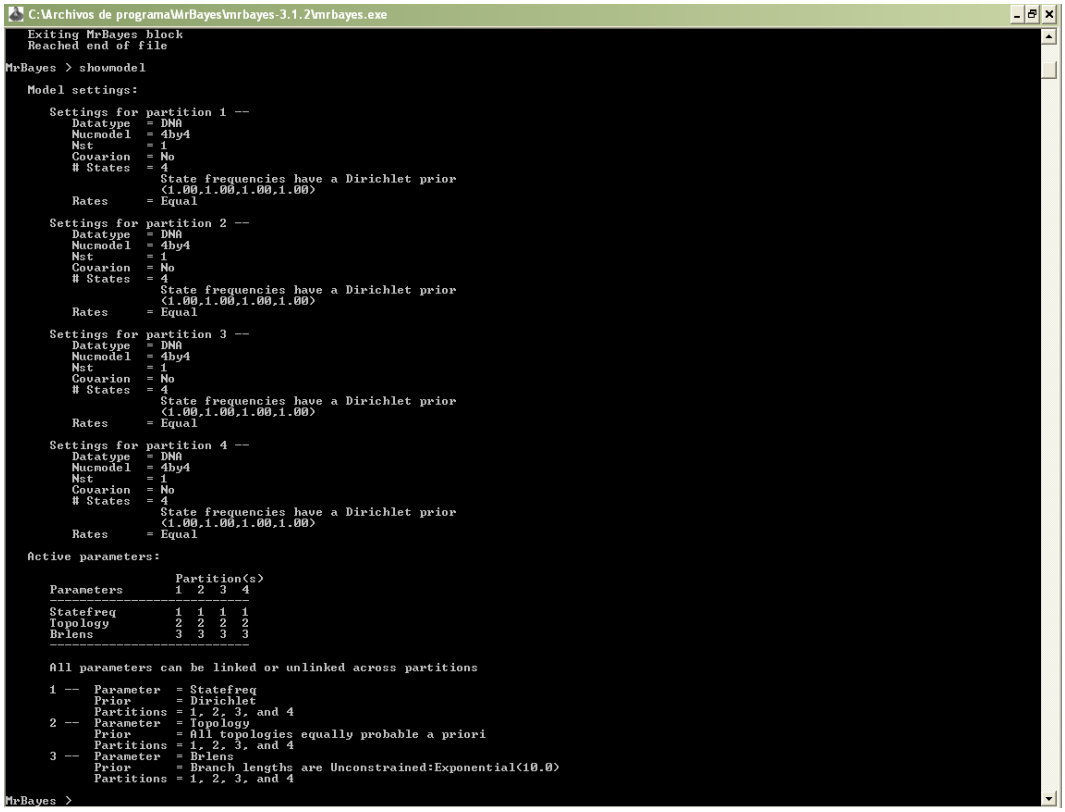
Paso 1. Abre el programa MrBayes y ejecuta el archivo con el comando *execute* Glaucoreseda_ITS_LF_particiones.*nex*



Ten en cuenta que MrBayes por defecto considera como outgroup el taxon que aparece en primer lugar en la matriz. Si quieres que se considere otro se puede indicar utilizando el comando *outgroup* N [donde N es el número de orden en la matriz del taxon que quieres que sea el *outgroup*]

Paso 2. Especifica el modelo evolutivo que mejor se ajusta a tus datos (véase tema 3.4):

El comando *showmodel* te permite ver el tipo de modelo que MrBayes aplica por defecto.



Por defecto, se aplica el modelo F81 a todas las particiones. Por lo que, aquellas particiones para las que el modelo que mejor se haya ajustado a tus datos no sea este (véase tema 3.4), debes cambiarlo. El tipo de modelo se puede cambiar utilizando los siguientes comandos:

lset Nst=**X** *Rates*=**Y**

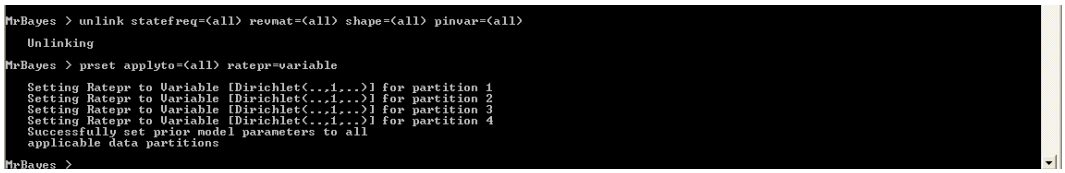
Donde *lset* sirve para cambiar el modelo
Nst sirve para indicar qué modelo se quiere seleccionar (sustituir X por 1 si el modelo es JC o F81, 2 si es HKY o K80, y 6 si es GTR o SYM)
rates indica la tasa de sustitución (sustituir Y por equal, gamma [+G], propinv [+I] o invgamma [+I+G])

Además hay que especificar a qué particiones queremos aplicar el cambio utilizando el comando *applyto*. Por ejemplo, para indicar que ITS1 (nuestra primera partición) se ajusta a un modelo evolutivo K80, teclea *lset applyto=(1) nst=2 statefreqpr=fixed(equal)*

Repite este comando para ajustar a cada partición el modelo que hemos seleccionado como en el tema 3.4. Con ello conseguirás que a cada región genética se le aplique su modelo evolutivo.

Hace falta además que cada partición tenga independientes el resto de los parámetros (los *priors*: frecuencia estacionaria de los nucleótidos o *statefreq*, tasa de sustitución de los nucleótidos o *revmat*, proporción de sitios invariables o *pinvar* y forma de la distribución gamma o *shape*). Para ello, hay dos maneras: (1) especificar para cada partición los valores utilizando el comando *statefreqpr*= [en nuestra partición 1 sería *=fixed(equal)*] o (2) que lo haga automáticamente el programa para cada partición tecleando el comando *unlink statefreq=(all) revmat=(all) shape=(all) pinvar=(all)*.

Finalmente, hay que indicar que la tasa de variación general puede ser variable entre las particiones. Esto se consigue con el parámetro *ratepr* del comando *prset*. Teclea *prset applyto=(all) ratepr=variable*.



Paso 3. Selecciona el número de generaciones que quieres correr utilizando el comando *Ngen*= y empieza el análisis:

El comando *mcmcp* te permite cambiar los parámetros de análisis sin empezarlo. Teclea *mcmcp Ngen=1000000*, con ello estarás especificando que cuando se haga el análisis se realicen 1000000 de generaciones.

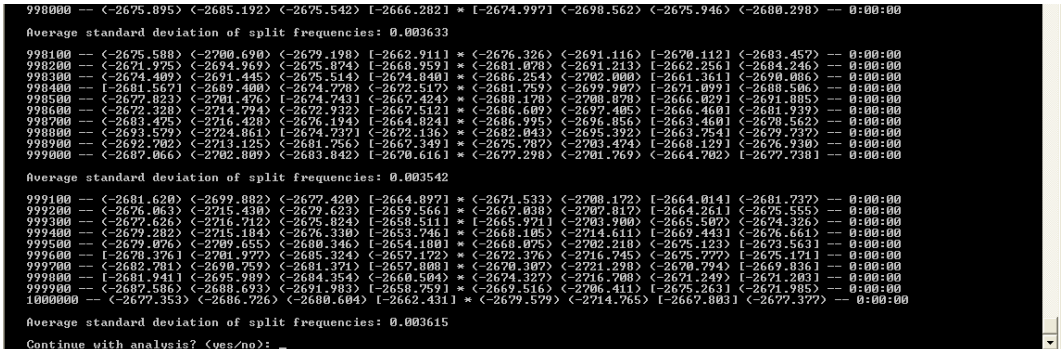
Paso 4. Ya hemos establecido todas las especificaciones necesarias y podemos proceder al análisis. El comando *mcmc* comienza el análisis con los parámetros previamente seleccionados o, en caso de no haberlo establecido, con los que se indiquen en el momento. Por ejemplo si tecleas *mcmc Ngen=1000000* se inicial el análisis con 1000000 de generaciones [Ten en cuenta que para una publicación deberías utilizar al menos 30.000.000 de generaciones].

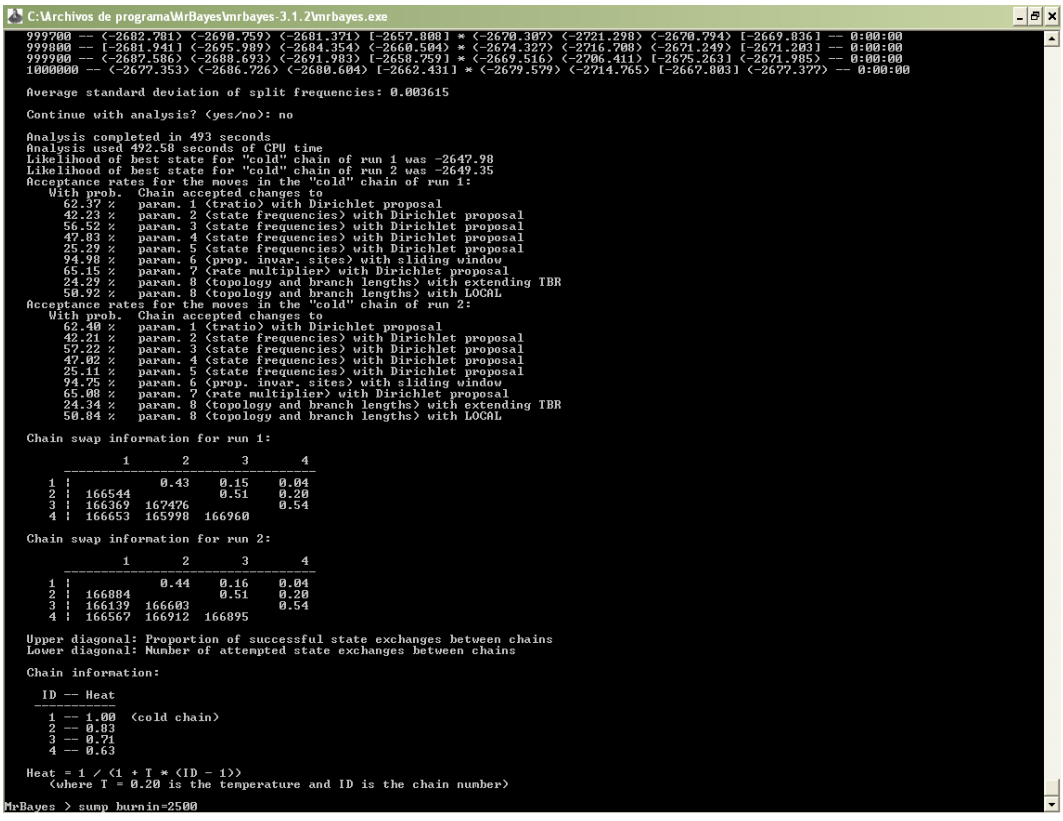
El comando *mcmcmc* incrementa la velocidad a la que se estabiliza la varianza porque implica que se salte de un pico a otro.

Paso 5. El análisis debe continuar hasta que la varianza se estabilice por debajo de 0.01. En la pantalla se va mostrando el número de generaciones que

lleva junto con los valores del *likelihood* obtenidos en cada generación para cada una de las cuatro cadenas. Además, se indica la varianza alcanzada y el tiempo que falta para que se termine el análisis. Cuando se alcanza el número de generaciones solicitado el programa pregunta si se quiere continuar con el análisis (*yes*) o no (*no*). Si el número de generaciones solicitado no ha sido suficiente como para estabilizar la varianza por debajo de 0.01 contestaremos que sí queremos continuar con el análisis, indicando cuántas generaciones adicionales queremos hacer hasta obtener la varianza requerida.

Ejercicio 3.5.3. ¿Debes realizar más generaciones o con 1.000.000 es suficiente?

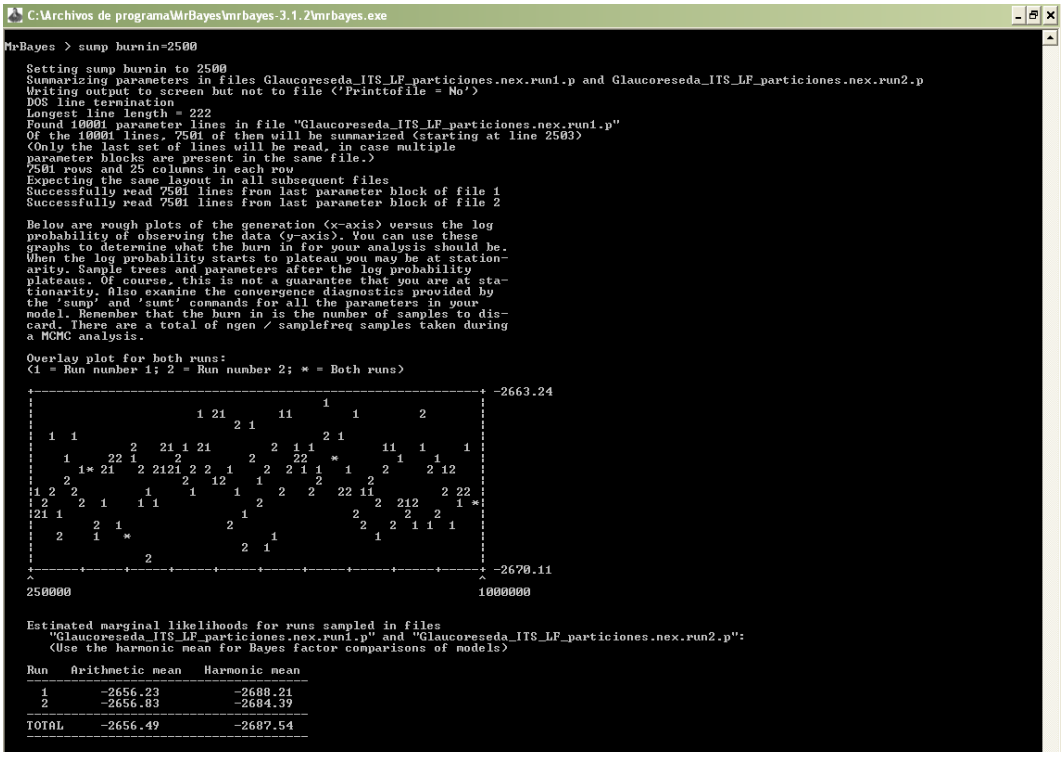


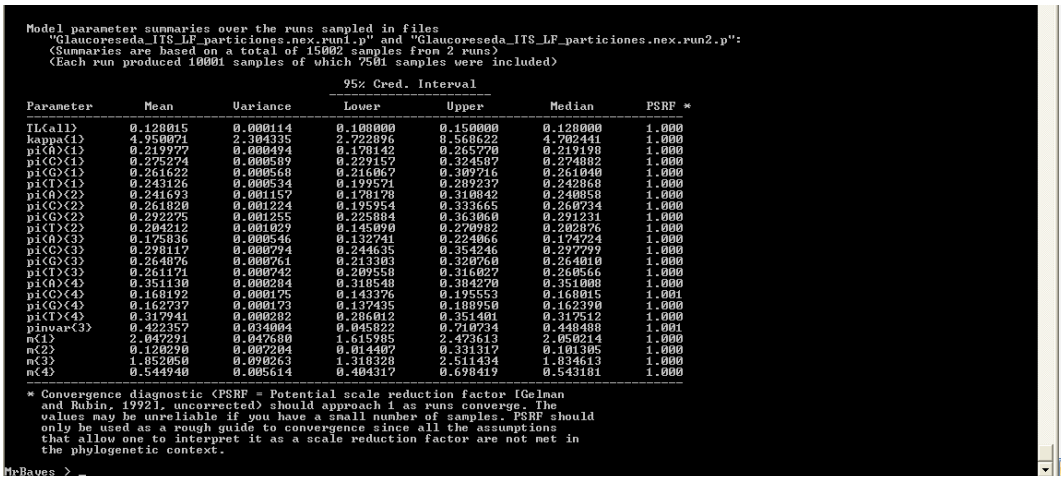


Paso 6. Una vez estabilizada la varianza y finalizado el análisis debemos visualizar los resultados. Para ello, teclea el comando *sump*. Lo que nos proporciona este comando es una tabla (en la que debemos comprobar que el parámetro PSRF es próximo a 1), y un gráfico de dispersión (que no debe mostrar tendencias si la varianza estaba estabilizada). Sin

embargo, estos resultados no nos interesan ya que incluyen también resultados previos a la estabilización. Para matrices no muy complejas, como la nuestra, se puede asumir que eliminando el 25% inicial de los árboles construidos estaremos eliminando aquellos árboles obtenidos antes de la estabilización. Para ello necesitamos el comando *burnin*. Por defecto, MrBayes guarda los resultados de la búsqueda cada 100 generaciones (un árbol y los valores de los parámetros asociados), por lo que si hemos realizado un análisis de 1000000 de generaciones, debemos eliminar el burnin tecleando la línea de comando:

sump burnin=2500





Paso 7. Para obtener un resumen de los árboles el comando que se utiliza es *sumt*, también descartando el 25% inicial de muestras. En el ejemplo anterior, la línea de comando sería:

sumt burnin=2500

El comando *sumt* crea tres archivos en la carpeta raíz: uno con extensión .parts, en el que se incluyen las particiones (clados), su probabilidad a posteriori (pp) y la longitud de rama asociada; uno con extensión .con, en el que se incluyen dos árboles consenso, uno con la longitud de las ramas y otro con la probabilidad; y uno con extensión .trprobs, en el que se guardan los árboles obtenidos ordenados por su probabilidad posterior.

Por defecto, para la construcción del árbol de consenso, MrBayes condensa los clados con menos de 50 de probabilidad posterior (*halfcompat*). Esto se puede cambiar con el comando *contype*= [Por ejemplo, *contype=allcompat* no condensa ningún clado]

Paso 8. Una vez obtenidos los árboles salimos del programa con el comando *quit*.

Otra opción alternativa consiste en incluir todos los comandos del análisis en un bloque de comandos a continuación del que hemos añadido especificando las particiones al final del archivo de la matriz. Para ello, puedes abrir la matriz con un editor de texto (WordPad para PC, o TextWrangler para MAC; OJO es muy importante que no lo abráis con Word) y a continuación del bloque donde has definido las particiones escribir lo siguiente:

```
lset applyto=(1) nst=2 rates=equal;
prset applyto=(1) statefreqpr=fixed(equal);
lset applyto=(2) nst=1 rates=equal;
prset applyto=(2) statefreqpr=fixed(equal);
lset applyto=(3) nst=1 rates=propinv;
prset applyto=(3) statefreqpr=fixed(equal);
lset applyto=(4) nst=1 rates=equal;
prset applyto=(4) statefreqpr=dirichlet(1,1,1,1);
```

[hasta aquí le hemos indicado a MrBayes que al ejecutar esta matriz aplique el modelo evolutivo K80 a la primera partición (ITS1), JC a la 2ª (5.8S), JC+I a la 3ª (ITS2) y F81 a la 4ª (*trn*L-F)]

```
set autoclose=yes;
```

[Con este comando le indicamos al programa que se cierre una vez terminado el análisis]

```
mcmc ngen=1000000 printfreq=100 samplefreq=100 nruns=2
nchains=4 savebrlens=yes burninfrac=0.25;
```

[Aquí le estamos indicando los parámetros para realizar la búsqueda: mcmc -que inicie las cadenas-, ngen=1000000 -que realice un millón de generaciones-, printfreq=100 –que nos muestre en la pantalla los resultados cada 100 generaciones-, samplefreq=100 –que guarde los

resultados de la búsqueda (árbol, más parámetros) una vez cada 100 generaciones-, nruns=2 –que inicie dos análisis al mismo tiempo-, nchains=4 –que corra cuatro cadenas de Markov por cada análisis-, savebrlens=yes –que guarde la longitud de las ramas-, burninfrac=0.25 –que aplique un burnin del 25%-]

```
sump;
sumt contype=halfcompat;
end;
```

[Estos últimos comandos le indican que realizado el análisis: sump; -nos compile los resultados-, sumt contype=halfcompat; -que compile los árboles y compute el árbol de consenso-, end; -que termine el análisis-]

Ejercicio 3.5.4. Escribe los comandos para el análisis dentro del archivo de la matriz Glaucoreseda_ITS_LF_particiones.nex.