TEMA 3.4. Selección del modelo evolutivo que mejor se ajuste a nuestros datos

Contacto: Virginia Valcárcel (virginia.valcarcel@uam.es)

INTRODUCCIÓN

Para calcular las distancias entre pares de secuencia, así como para realizar búsquedas con los métodos basados en distancias, máxima verosimilitud e inferencia bayesiana es necesario especificar un modelo de evolución para los datos (véase tema 2.1). Los modelos evolutivos en filogenias moleculares describen el modo y la probabilidad de que una secuencia de nucleótidos cambie a otra secuencia de nucleótidos homóloga a lo largo del tiempo. Es decir, estos modelos describen para cada uno de los sitios de la matriz la probabilidad de que se produzca el cambio de un nucleótido a otro a lo largo de las ramas de un árbol filogenético dado.

Los modelos de evolución de nucleótidos se definen matemáticamente mediante dos clases de parámetros que determinan el cambio:

(1) Frecuencia de cada nucleótido. Parámetro que mide la frecuencia de los nucleótidos en la matriz de datos y que puede tomar los siguientes valores:
   a. En los modelos más sencillos: una misma frecuencia para los cuatro nucleótidos ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$)
   b. En los modelos más complejos: cuatro frecuencias diferentes, una para cada nucleótido ($\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$)

(2) Tipos de sustituciones y sus correspondientes tasas de sustitución (*rate parameters*). Las tasas de sustitución se representan con las tasas relativas de cambio de un nucleótido a otro para una posición de un tiempo t_0 a un tiempo t_1. Así, cada posición de la matriz tendrá una probabilidad asociada de cambio para cada unidad de tiempo (unidad de distancia evolutiva). Así la tasa relativa de sustitución en una posición de un nucleótido A a una C se denota con “a”, de A a G con “b”, de A a T “c” y así hasta “l” (tasa relativa de cambio de T a G). Los modelos más sencillos asumen una misma tasa relativa para todas las sustituciones posibles (en cuyo caso suele denotarse con la letra “α”), mientras que los más complicados asumen una tasa relativa diferente para cada tipo de sustitución. Algunos modelos contemplan también la tasa media de sustitución (μ).

A partir de las combinaciones posibles de estos parámetros se han descrito cerca de 203 modelos se sustitución nucleotídica. Los modelos más sencillos son aquellos que incluyen un menor número de parámetros. Además de los parámetros aquí descritos habría que considerar también el número de longitudes de rama. Así, el modelo más sencillo posible es Jukes and Cantor (JC, Jukes and Cantor 1969; Fig. 1) que asume la misma frecuencia para los cuatro nucleótidos y un único tipo de sustitución, el número de parámetros de JC será el número de longitudes de rama del árbol. El modelo Kimura 2-parámetros (KP2, Kimura 1980) es como JC pero asumiendo la existencia de dos tipos de sustituciones por lo que el modelo tiene como parámetros libres el número de ramas del árbol más uno.

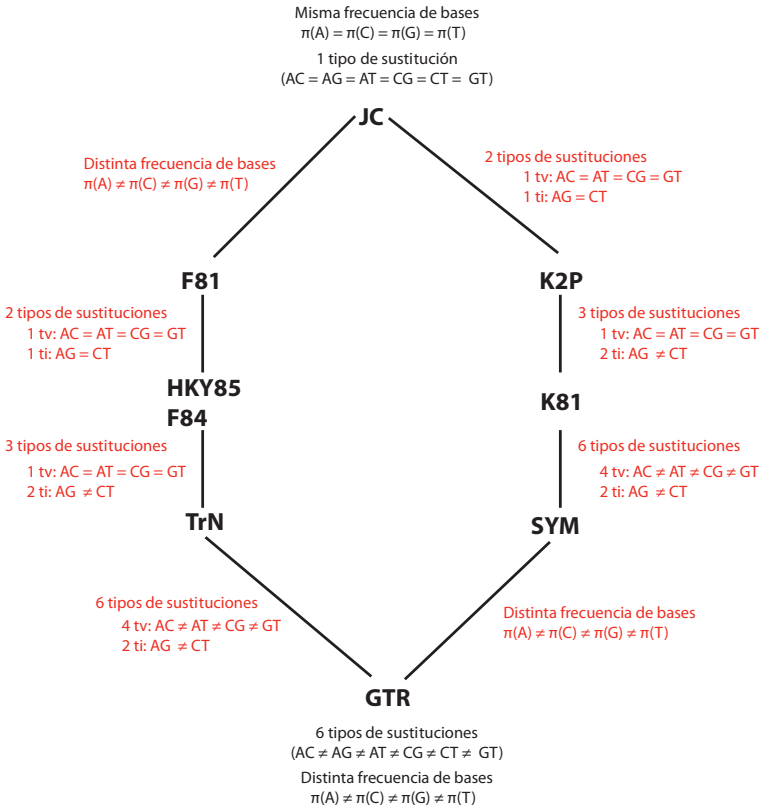


Figura 1. Siete de los c.203 modelos de sustitución nucleotídica de la familia de *General Time-Reversible* (GTR).

A partir de estos parámetros se define cada uno de los modelos posibles. La expresión matemática para los modelos de sustitución es una matriz (*Instantaneous rate matrix* Q; Fig. 2)**.**

$$\begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{pmatrix}$$

Figura 2. Matriz de sustitución (Q) para modelos de sustitución de nucleótidos. Cada uno de los elementos de la primera fila representa la probabilidad de cambio desde una Adenina en un tiempo t_0 a cada uno de los tres nucleótidos transcurrido un tiempo t_1, así como la probabilidad de mantenerse como una Adenina. Los elementos de la segunda fila representan las probabilidades de transición de una Citosina, los de la

tercera fila corresponden a las probabilidades de transición de una Guanina y los de la cuarta fila a las probabilidad de transición de una Timina.

Esta matriz representa la probabilidad de que una posición pase de tener un nucleótido (A, C, G, T) en un tiempo t_0 a no cambiar en un tiempo t_1 (diagonal de la matriz) o bien a ser sustituido por cada uno de los respectivos tres nucleótidos posibles. Así, cada elemento de la matriz representa la probabilidad de cambio desde un nucleótido a otro. Así, el elemento 1.1 $-\mu(a\pi_C + b\pi_G + c\pi_T)$ de la matriz representa la probabilidad de que un sitio que tenga a un tiempo t_0 una A se mantenga como A en el tiempo t_1.

Además, estos modelos pueden incluir parámetros que reflejen la posibilidad de que no todos los sitios de la matriz original de datos evolucionen con la misma tasa, esto es permiten que las tasas sean heterogéneas entre sitios. Para modelar la heterogeneidad de tasas entre sitios, se suele asumir que las tasas relativas cambian entre los sitios según una distribución gamma con media 1 y varianza 1/α. "α" controla la tasa entre sitios, si α < 1 entones hay mucha diferencia entre sitios (esto es, muchos sitios varían poco mientras que unos pocos varían mucho; Fig. 3). A veces también se incluye en el modelo la posibilidad de que exista una proporción de sitios que varían (I).

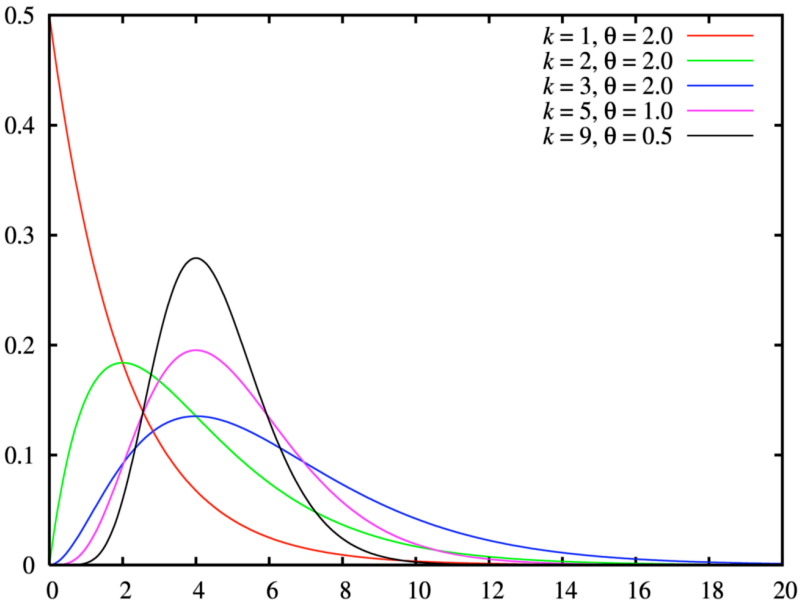


Figura 3. Diferentes distribuciones gamma. Gráfica tomada de Wikipedia.

Ninguno de los modelos de sustitución descritos en la literatura será el modelo real de sustitución que habrán seguido nuestras secuencias. El uso de un modelo de sustitución u otro para una misma matriz de datos puede generar diferentes árboles filogenéticos (Lemmon and Moriarty 2004). Por todo ello, seleccionar el modelo que mejor se ajuste a nuestra matriz de datos es esencial de cara a obtener una reconstrucción filogenética robusta y fiable. Los estadísticos para la selección de modelos se basan casi todos en el principio de Occam por el cuál la mejor hipótesis es la más simple.

¿Cómo se calcula el ajuste del modelo?

El grado de ajuste de un modelo de sustitución a nuestros datos se calcula generalmente mediante la función de verosimilitud (L, *Likelihood function*). En filogenia molecular L es proporcional a la probabilidad de los datos (D: nuestra matriz de ADN) dados: un modelo de evolución (M), un vector con los K parámetros incluidos en el modelo (θ), la topología de un árbol (τ) y un vector de S longitudes de rama (v).

El cálculo del grado de ajuste de los distintos modelos a nuestros datos requiere de una topología y unas longitudes de rama. Por ello, los criterios de selección de modelos suelen comenzar estimando un árbol a partir de los datos y, asumiendo que éste fuera el mejor árbol, se la estiman todos los parámetros incluidos en cada modelo devolviendo un valor final de verosimilitud de nuestros datos para cada uno de los modelos evaluados.

PROGRAMAS NECESARIOS

(1) JModeltest. El programa que vamos a utilizar para realizar esta práctica es jModelTest (Posada 2008). Este programa libre se desarrolló para la selección estadística de modelos de sustitución nucleotídica. Tiene implementados cinco estrategias de selección diferentes (Akaike Information Criterion, AIC; hierarchical Likelihood Ratio Test, hLRT; Dynamical Likelihood Ratio Test, dLRT; Bayesian Information Crierion, BIC; Performance-based selection base don decisión theory, DT).

Accede a la página http://darwin.uvigo.es/software/jmodeltest.html. Rellena las datos solicitados por el autor del programa y descárgatelo en tu ordenador.

Programas alternativos:

Existen muchos otros programas para la selección de modelos (ModelTest, Posada & Crandall 1998; MrModeltest, Nylander 2004a). Sin embargo, estos programas, a diferencia de jModeltest, necesitan del programa con licencia PAUP (Swofford 2002) para poder realizar algunas de las partes del proceso. Una alternativa independiente de PAUP es el programa libre MrAIC.PL (Nyalnder 2004b).

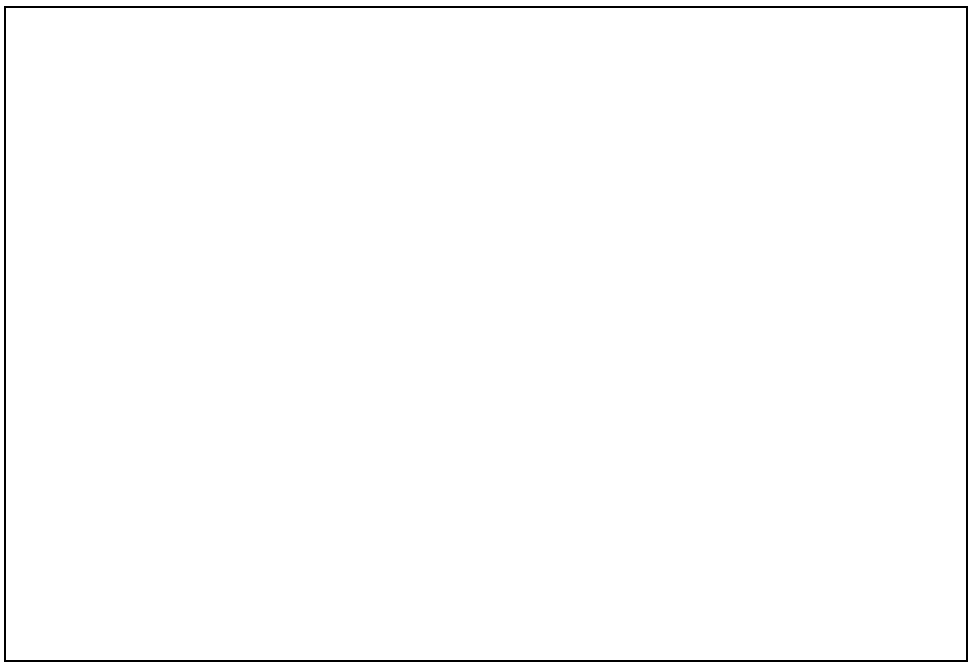
METODOLOGÍA Y PRÁCTICA

Para poder obtener una reconstrucción filogenética de *Reseda* sect. *Glaucoreseda* debemos utilizar la matriz combinada de ITS y *trn*L-F que hemos construido. Para poder llegar a conseguir este objetivo utilizando los métodos basados en distancias, ML y BI necesitamos conocer el modelo evolutivo que mejor se ajusta a cada una de las cuatro regiones de ADN que tenemos representada en dicha matriz. Esto es, necesitamos conocer el modelo que mejor se ajusta a: (1) el espaciador 1 de la región del ADN ribosómico nuclear ITS (ITS-1), (2) el gen 5.8S de la región del ADN ribosómico nuclear ITS, (2) el espaciador 2 de la región del ADN ribosómico nuclear ITS (ITS-2) y (3) el espaciador del ADN plastidial *trn*L-F.

Antes de proceder a los temas 3.5, 3.6 y 3.7, hay que realizar esta práctica y sus ejercicios. Durante el desarrollo de esta práctica aprenderéis el procedimiento para la selección de modelos usando como ejemplo la región plastidial (*trn*L-F). Sin embargo, es imprescindible que repitáis esta práctica para las otras tres regiones.

Paso 1. Abre el programa JModeltest y carga la matriz de *trn*L-F que has construido (Alineamiento_LF_revisado.fasta) en "File/Load DNA alignment". Si el formato de la matriz es correcto aparecerá un mensaje en la pantalla indicando el número de secuencias de la matriz así como el número de posiciones (en nuestro caso: 17 y 736, respectivamente) lo que significa que la matriz se ha cargado correctamente y puedes proceder a realizar los cálculos.

Paso 2. Para realizar los análisis, selecciona "Compute Likelihood Scores". Se abrirá una ventana en la que nos solicitan la especificación de los parámetros necesarios para el cómputo.



En esta ventana hay que seleccionar los modelos que queremos evaluar, esta selección se hace en función del tipo de sustituciones que queramos aceptar. Así, si seleccionamos "3" estaremos incluyendo todos los modelos posibles que admiten desde un sólo tipo de sustitución hasta tres tipos de sustituciones; si seleccionásemos "11" estaríamos incluyendo todos los modelos posibles que admiten desde un sólo tipo de sustitución, hasta 11 tipos de sustitución diferentes. El número de modelos a evaluar en cada una de las opciones va aumentando en consecuencia. Así, si seleccionamos tres tipos de sustituciones evaluaremos el grado de ajuste de 24 modelos diferentes –para "5" serán 40 modelos, para "7" 46 y para "11" 88–. Los programas de análisis filogenéticos con los que contamos en la actualidad no implementan todos los modelos, por lo que seleccionando "3" estaremos prospectando el grado de ajuste de los 24 modelos que implementa en la actualidad MrBayes –programa que utilizaremos en el tema 3.6 para estimar la filogenia del nuestro grupo mediante el método de Inferencia Bayesiana–.

Señala "+F", con ello estamos incluyendo los modelos que asumen misma frecuencia de bases así como los que asumen distinta frecuencia. Señala "+I", con lo que estaremos incluyendo tanto los modelos que asumen la proporción de sitios que varían como los que no. Por último,

selecciona “+G”, de esta manera estaremos incluyendo tanto los modelos que asumen diferente tasa de variación entre sitios como los que no.

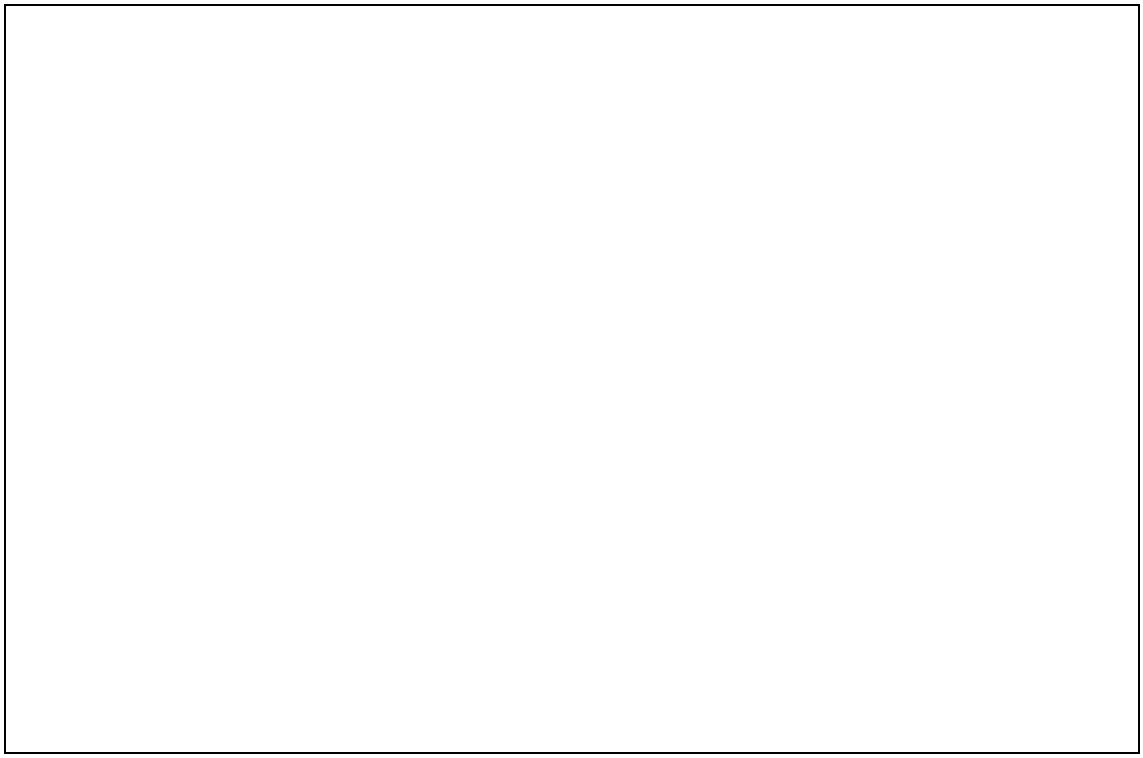
Como necesitamos un árbol inicial para calcular el grado de ajuste, hay que seleccionar de qué manera queremos obtener la topología de dicho árbol. Así, se puede elegir si queremos una misma topología fija para evaluar todos los modelos. En tal caso, dicha topología puede obtenerla el programa mediante un algoritmo basado en NJ que utiliza las distancias de JC (“Fixed BIONJ-JC”) o bien podemos proporcionarla nosotros en caso de que contemos con un árbol y queramos usarlo (“Fixed user topology”. Otra opción es estimar la topología más óptima cada vez que se evalúa un modelo y esto podemos hacerlo mediante una aproximación basada en distancias Neigbourgh joining (“BIONJ”) o de máxima verosimilitud (“ML optimized”). En nuestro caso, dejamos marcada la que aparece por defecto “ML optimized”

Paso 3. Seleccionados los parámetros dale a “*Compute likelihoods*”, con esto estaremos calculando los valores de verosimilitud para cada modelo dada nuestra matriz. Una vez calculados los *Likelihoods* para todos los modelos te aparecerán en la pantalla de inicio los resultados de estos cómputos. En ellos se indica para cada uno de los modelos evaluados: (1) el valor del *Likelihood* estimado para el árbol optimizado y (2) el número de parámetros libres del modelo (K).

Paso 4. Una vez hecho esto puedes proceder a evaluar la bondad de ajuste de los distintos modelos a tus datos según los distintos criterios se selección implementados en este programa (AIC, BIC, DT). Para ello, ve a "Analysis" y selecciona en primer lugar "Do AIC calculations…".

Se abrirá una ventana en la que te preguntará si quieres calcular AICc. Ésta, es una opción que se utiliza cuando el tamaño de la muestra (n: número de posiciones de la matriz) es pequeño en comparación con el número de parámetros de los modelos (K). El AICc realiza una corrección estadística y debe usarse siempre que n/K < 40 (Posada 2004). En tu caso la matriz de *trn*L-F tiene 736 sitios (n) y los 24 modelos necesitan entre K=32 y K=42 parámetros, luego señala la opción AICc.

Una vez terminados los cálculos, en la pantalla te aparecerán los resultados. En primer lugar aparece el mejor modelo seleccionado (Model selected) –en este caso es F81+I– indicando también el valor del *Likelihood* ("-lnL"), el número de parámetros libres ("K"), la frecuencia de bases nucleotídicas ("freqA", "freqC"…) y la proporción de sitios invariables (p-inv).



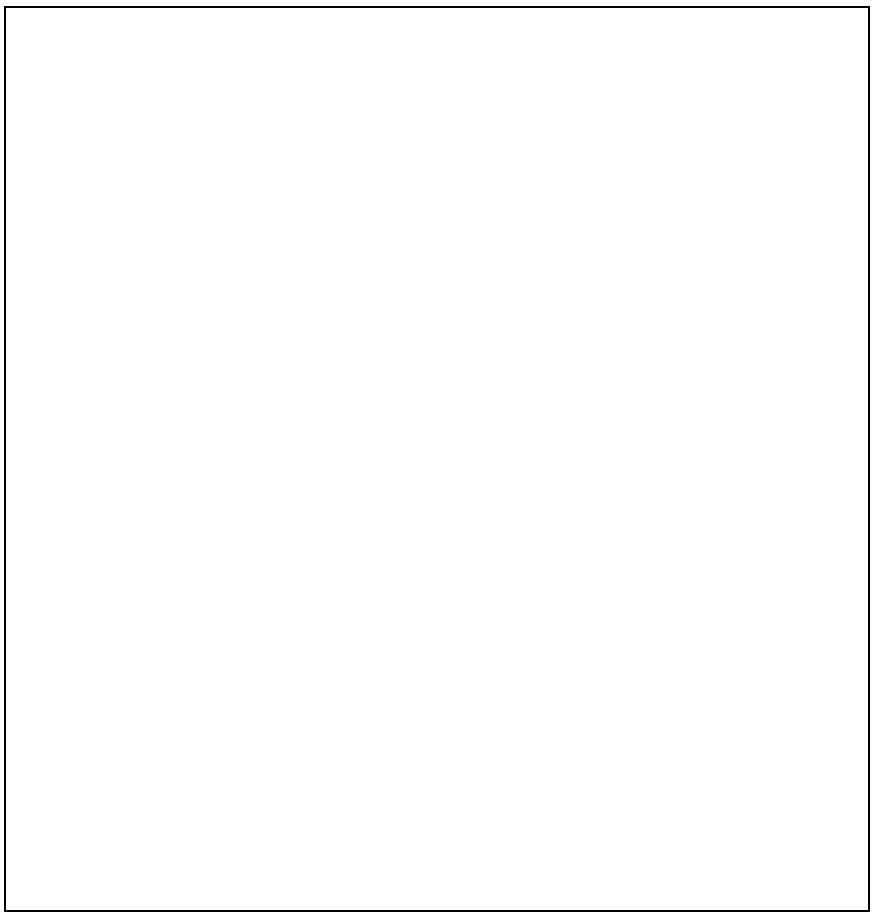
Siempre aparece un modelo seleccionado, pero éste no siempre es significativamente mejor que el segundo mejor modelo. Por ello, debes estudiar con detenimiento la tabla de resultados que aparece a continuación ya que en ella se muestran las estimas de la comparación de todos los modelos. La primera columna indica el modelo, la segunda el *Likelihood* ("-lnL"), la 3ª el número de parámetros del modelo ("K"), la tercera el valor de AICc, la cuarta indica la diferencia entre el valor de AICc del mejor modelo con el modelo especificado ("delta"). Si el valor de

delta es inferior a 3, el mejor modelo (esto es, el modelo seleccionado) no es significativamente mejor que el modelo especificado.

En estos resultados, el modelo F81+I resulta ser el modelo seleccionado al tener el menor valor del *Likelihood* (-lnL = 1219.7011) –su delta, obviamente, es cero–. El siguiente mejor modelo es F81+G con un *Likelihood* ligeramente superior (-lnL = 1219.7236), en este caso el valor de delta es de 0,0450. Esto significa que aunque F81+I es el que presenta el Likelihood menor, F81+I no es significativamente mejor que cuando se utiliza el modelo F81+G.

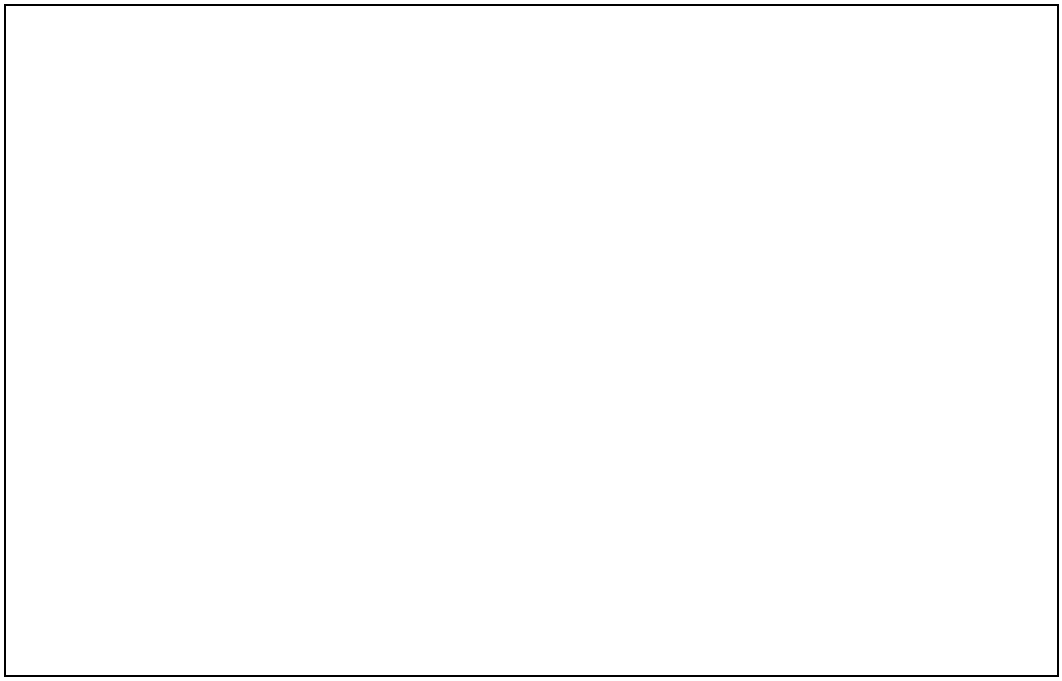
Ejercicio 3.4.1. Teniendo en cuenta estos valores de significación resultantes de aplicar el criterio AIC de selección ¿podrías decir qué modelos son igualmente probables dados tus datos (esto es, dada la matriz de *trn*L-F)?

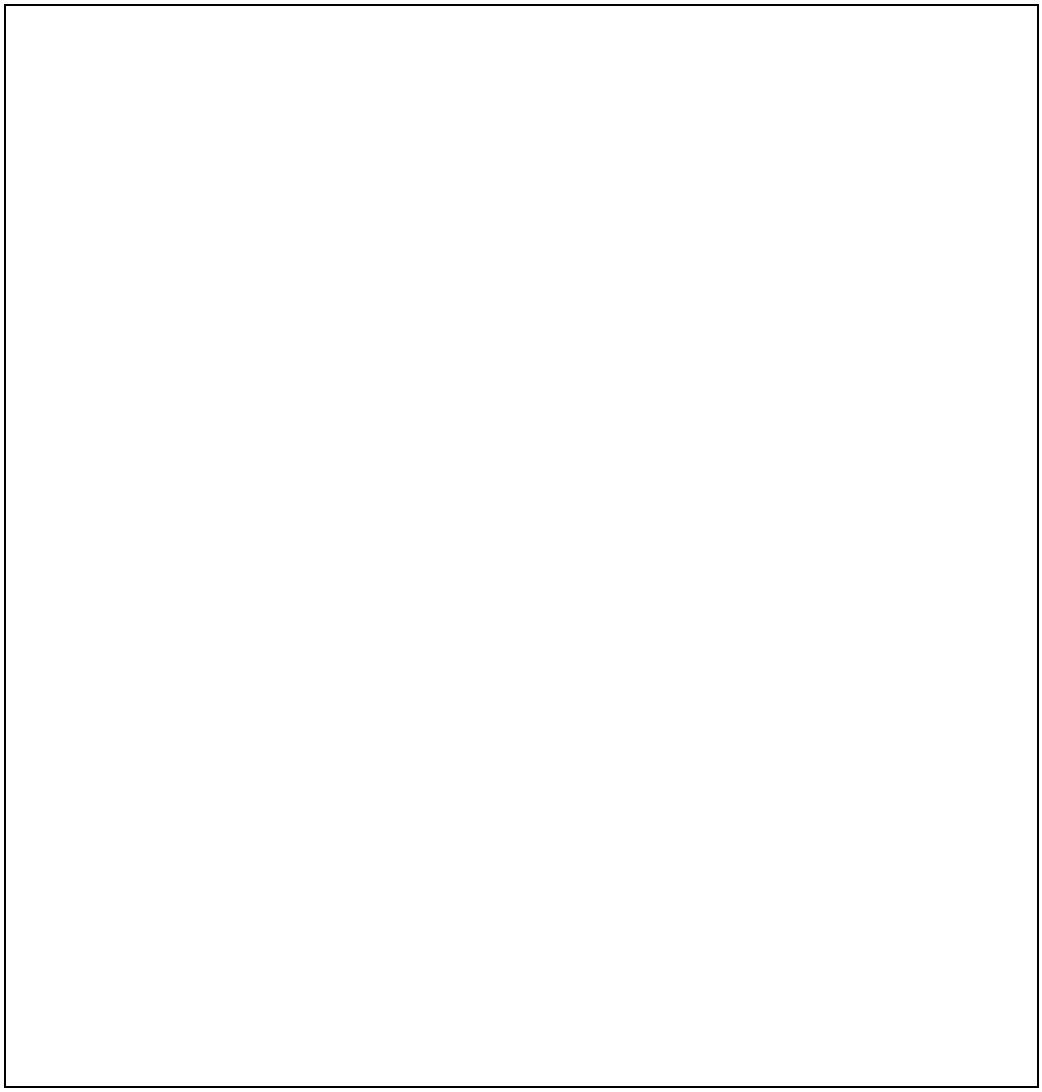
La sexta columna indica el peso de cada modelo ("weight") y la última columna el peso acumulado ("cumWeight"). El intervalo de confianza se alcanza cuando este último toma el valor de 0.95.



Paso 5. Ya tenemos los resultados de la selección del mejor modelo mediante la aproximación que utiliza AICc como criterio de selección.

Ahora vuelve a Anlysis y selecciona "Do BIC calculations…". Después de realizar los cálculos, en la pantalla aparecerán los resultados.





En este caso, el modelo seleccionado es F81.

Ejercicio 3.4.2. ¿Hay algún modelo que se ajuste significativamente mejor a tus datos según el criterio BIC de selección?

Paso 6. Para guardar los resultados del análisis ve a "Edit" y selecciona "Save console". De este modo, se genera un archivo de texto que siempre que quieras podrás consultar para ver los resultados de la búsqueda usando un editor de texto (WordPad para PC, o TextWrangler para MAC; OJO es muy importante que no lo abráis con Word).

Paso 7. Para realizar los análisis de distancias, BI o ML necesitamos especificar qué partes de la matriz utilizada siguen distintos modelos y proporcionar dichos modelos. Por ello, aún no hemos terminado y tenemos que tomar una decisión sobre qué modelo de los seleccionados escogemos. Por un lado está el problema del criterio de selección ¿cuál es mejor BIC o AIC? Habitualmente los resultados bajo ambos criterios son iguales y no existe tal disyuntiva. Pero en el caso de plantearse, como es el nuestro, no existe una respuesta definitiva a la pregunta, ya que cada criterio parece comportarse mejor o peor en función de las características de los datos (Posada & Buckley 2004). Si los modelos seleccionados según los

diferentes criterios son distintos, una aproximación exhaustiva sería realizar todos los análisis por duplicado utilizando uno de los modelos seleccionados por los criterios cada vez. En el caso de que en nuestro caso nos decidiésemos por seguir los resultados del criterio AIC, tendríamos una nueva disyunta y es, qué modelo elegiríamos de entre los nueve modelos que resultan ajustarse significativamente a nuestros datos. En este caso, la aproximación más seguida es la de optar por el mejor modelo, esto es el de mayor *likelihood*.

En cualquier caso, a la hora de seleccionar los mejores modelos en caso de disyuntiva una consideración importante es tener en cuenta que cuanto más sencillo sea el modelo, menor número de parámetro tendrá. Cuanto menor sea el número de parámetros a estimar, menor número de errores asociados incluiremos en nuestra reconstrucción. Por ello, elegir el modelo más sencillo es una buena aproximación. En nuestro caso, y siguiendo este criterio, seleccionaremos el modelo F81 para la región *trn*L-F.

Ejercicio 3.4.3. Teniendo en cuenta que la región ITS incluye dos espaciadores (ITS1, ITS2) separados por el gen 5.8S. Estima el mejor modelo para cada una de las tres regiones del espaciador ITS (para ello utiliza los archivos “Glacuoreseda_ITS_ITS1”, “Glacuoreseda_ITS_58S” y “Glacuoreseda_ITS_ITS2” que te encuentras en la carpeta de inputs) y responde a las siguientes preguntas: (a) ¿Cuál es el mejor modelo para ITS-1 según AIC y BIC?, (b) ¿Cuál es el mejor modelo para 5.8S según AIC y BIC?, (c) ¿Cuál es el mejor modelo para ITS-2 según AIC y BIC?, (d) ¿Qué modelos utilizarías para realizar los análisis de la matriz de ITS?