Estimación de tiempos de divergencia con Beast

Datos Moleculares en Biodiversidad II

Nelson R. Salinas

Octubre 31, 2020

Repositorio sesión:

https://github.com/nrsalinas/datos_moleculares_ii/tree/master/Taller_11.

El presente taller será evaluado. El estudiante debe presentar un reporte a manera de miniartículo sobre la datación molecular de su grupo de interés (punto 14). El miniartículo no debe exceder una página de longitud, excluyendo bibliografía y gráficas. El taller debe ser entregado antes de la sesión del sábado 7 de noviembre.

La primera sección del taller está basado en la clase de datación molecular del *Bodega Applied Phylogenetics Workshop* dictada por Tracy Heath. En dicha sesión, los estudiantes aprendimos la utilización básica de Beast calibrando un dataset simulado, lo cual ahorra bastante tiempo en el momento de ejecutar un análisis bayesiano.

1. En el presente taller se utilizarán varios de los programas instalados en la sesión de inferencia bayesiana (26 de septiembre): Beast, Beauti, TreeAnnotator y Tracer; los tres primeros distribuidos como la suite de análisis filogenético Beast. Todos ellos requieren Java Runtime Environment.

2. Ejecute Beauti y cargue el archivo `divtime.nex` con la opción *Import Alignment*, que contiene la matriz simulada que será analizada.

3. El conjunto de datos contiene datos de dos marcadores, así que se debe sincronizar la estimación del árbol filogenético y del reloj molecular. Para ello seleccione ambas particiones y luego seleccione las opciones *Link Clock Models* y *Link Trees*.

4. Configuración del modelo de substitución. Dado que el dataset fue simulado, conocemos cual es el modelo que originó los datos. En un ejercicio empírico se debe configurar el modelo de acuerdo a una selección previa de modelo.

   (a) En la pestaña *Site model*, seleccione la primera partición y configure el modelo Gamma de heterogeneidad con 4 categorías (*Gamma Category Count*). Asegure que el parámetro *alpha* de la distribución Gamma (*Shape*) es 1 y que la casilla correspondiente de estimación está activada.

(b) Seleccione el modelo GTR de substitución en la opción *Subs Model*.

(c) Ahora seleccione la segunda partición y realice la misma configuración de modelo que realizó para la partición 1.

5. Configuración de los priors que corresponden a los puntos de calibración. En este punto entran en juego las fechas de los fosiles y su respectiva asignación filogenética.

   (a) En la pestaña *Priors*, en la base del listado de priors, añada uno nuevo. Automáticamente emergerá una ventana con los nombres de los terminales. Seleccione los taxones T1, T2, T3, T4, T5 y T6 (grupo interno) e incorpórelos al set (>>). Posteriormente nombre el nodo como "Nodo1" en la casilla *Taxon set label* y acepte con *OK*.

   (b) La monofilia del nodo recién creado será impuesta en el análisis activando la opción *monophyletic*.

   (c) La distribución de este nodo será exponencial, la cual debe ser seleccionada en el menú desplegable. La propiedades de la distribución se configuran en un menú que aparecerá al pinchar el triángulo negro de la margen izquierda. Allí se seleccionará una media (*Mean*) de 30 y un desfase (*Offset*) de 60.

   (d) Añada un nuevo prior que corresponda a los taxa T7, T8, T9 y T10 (el grupo externo) y nombrelos "Nodo2". Imponga la monofília del mismo y seleccione una distribución LogNormal cuya media (M) sea 2.714482, desviación estándar (S) 0.75 y desfase (*Offset*) 40.

   (e) Adicione otro prior para los taxa T9 y T10, cuyo nombre sea "Nodo3". Se seleccionará una distribución uniforme con límite inferior (*Lower*) de 8, un límite superior (*Upper*) de 32 y sin desfase. No imponga la monofília del clado.

   (f) Finalmente se añade un prior para la raíz del árbol, para lo cual se deben incluir todos los taxa en un prior llamado "Nodo4". La calibración será modelada como una distribución normal con media de 140 y desviación estándar de 10. Tampoco se incorporará desfase ni se impondrá la monofília del clado.

6. Ahora se debe seleccionar el modelo de ajuste temporal. Para ello, en la pestaña *Clock Model* seleccione el modelo *Relaxed Clock Log Normal*. De este modelo solo se ajustará el prior de la media. En la pestaña *Priors* busque la sección *ucldMean.c:...* y especifique una distribución exponencial con media de 10.

7. El último paso en Beauti es configurar la longitud de la cadena de Markov Monte Carlo en la pestaña MCMC. La longitud usual de una cadena se ubica en el rango 5-30 millones, pero para este ejercicio solo se correrá por 1.000.000 de generaciones. El muestreo de la cadena se realizará cada 100 generaciones, lo cual se especifica bajo la sección *tracelog*, celda *Log Every*.

8. A continuación guarde el archivo de ejecución con el nombre `divtime0.xml`. Este será el archivo que utilizará Beast para ejecutar el análisis.

9. Antes de cerrar Beauti cree una réplica del análisis, lo cual le ayudará a verificar si la MCMC fue ejecutada correctamente. Para evitar que este segundo archivo de ejecución

sobreescriba los resultados del archivo anterior es necesario que cambie los nombres de de algunos archivos. En la pestaña MCMC, sección *tracelog*, cambie el nombre de la celda *File Name* a `divtime_bis.log`, y en la sección *treelog.t:...* cambie el texto de la celda *File Name* a `$(tree)_bis.trees`. Posteriormente guarde el archivo de ejecución como `divtime1.xml`.

10. Ejecute cada uno de los archivos recién creados con Beast, ya sea a través de la línea de comandos (p.e., `beast divtime0.xml`) o abriendo el ejecutable de Beast y luego seleccionando el archivo xml correspondiente a través de la interfaz gráfica.

11. Después de realizados los análisis verifique que las cadenas MCMC han alcanzado la convergencia. Los detalles de esta rutina están explicados en el taller 6, inferencia bayesiana. Sin embargo, en esta sección del taller no debe realizarse una verificación estricta, ya que se seleccionó un número muy pequeño de generaciones.

12. Realice el sumario de los parámetros de los árboles de acuerdo a lo especificado en el punto 10 del taller de inferencia bayesiana.

13. Visualice el árbol con FigTree. Para desplegar el 95% más probable de la distribución posterior de las edades de los nodos active la sección *Node Bars* y seleccione la opción *height_95%_HPD* en la casilla *Display*. Para mejorar la comprención de las calibraciones puede desplegar la escala temporal de la gráfica activando la sección *Scale Axis*.

14. Ahora realize la datación bayesiana de su grupo de estudio empleando el conocimiento practicado en los pasos anteriores. La principal diferencia será que, antes de resumir las distribuciones de los parámetros en un árbol de máxima credibilidad, debe confirmar que las cadenas MCMC si hayan alcanzado convergencia, para lo cual debe ejecutarlas por un número apropiado de generaciones.