Taller II

Estimación de tiempos de divergencia: BEAST.

Descripción

Este tutorial se propone estimar los tiempos de divergencia, calibrando con fósiles, en un marco Bayesiano. Se utilizará BEAST v1.6, un programa ampliamente utilizado en análisis evolutivos. Además de BEAST, el paquete incluye:

- BEAUti (Bayesian Evolutionary Analysis Utility): es un programa con interfaz gráfica para crear los input para BEAST.
- LogCombiner: puede ser usado para combinar dos análisis independientes, escritos en archivos .log y .trees (output de BEAST), en un solo archivo cada uno, para que sea leído por Tracer (.log) y TreeAnnotator (.trees). NOTA: solo tiene sentido combinar análisis que fueron realizados usando el mismo input .xml.
- TreeAnnotator: es usado para resumir el muestreo de árboles realizado en el análisis de BEAST. El resultado es un árbol que maximiza la credibilidad de los clados y resume las estimaciones posteriores de los parámetros para que sean visualizadas en el árbol.

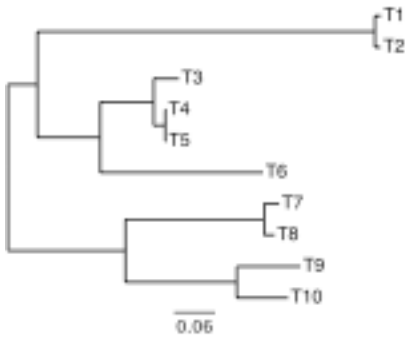
BEAST es un programa libre y puede descargado de su web oficial.

Tutorial

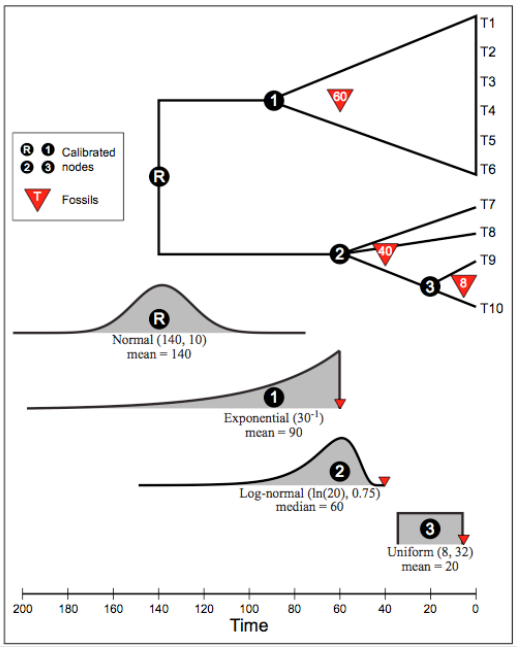
Este tutorial fue modificado del tutorial original de Tracy Heath, disponible en: http://treethinkers.org/divergence-time-estimation-using-beast/

Descripción de la matriz a utilizar

Ubicar el archivo divtime.nex en la carpeta data y abrirlo en un editor de texto. Esta es una matriz nexus sencilla, con 10 individuos y dos genes (gene1 y gene2), cada uno con un largo total de 500 pb (esta información la reconocemos en el bloque final "charset"). Los individuos T1 a T6 corresponden al ingroup, y T7 a T10 son outgroups. Un análisis previo, infirió el siguiente árbol utilizando esta misma matriz:



Supongamos que hay disponibles cuatro puntos de calibración (utilizando información de fósiles) para este set de datos, ilustrados en la figura siguiente. Las distribuciones de los *priors* para esos puntos de calibración, se encuentran representados debajo del árbol.



Estimación de tiempos de divergencia: BEAUti y BEAST.

1. Hacer doble clic sobre el programa BEAUti, dentro del paquete BEAST 1.6.
2. Abrir el archivo nexus divtime.nex en BEAUti, para ello ir a file>import data… o apretar las teclas command+I. También se puede arrastrar divetime.nex hasta la consola de BEAUti usando el mouse. Observar que BEAUti automáticamente reconoce cada gen por separado (esto es posible solo si se define "charset").
3. Seleccionar ambas particiones gene1 y gene2, haciendo clic y manteniendo apretada la tecla command o shift. Hacer clic en los botones "Unlink substitution models". Al hacer esto, estamos pidiendo que se estimen los modelos de substitución por separado para cada gen.
4. Sin soltar la selección anterior, hacer clic en "Unlink clock models". Esto permite estimar un reloj de evolución molecular independiente para cada partición (gene1 y gene2).
5. Ir a la solapa "Taxa", donde se pueden definir los grupos monofiléticos y los cuatro puntos de calibración (aunque el punto de la raíz no es necesario definirlo, ya que está implícito). Hacer clic en el botón + que aparece abajo a la izquierda. Renombrar Taxon Set haciendo clic en "untitled 1" y escribiendo mrca1. Tildar la caja monophyletic?.
6. Seleccionar de la columna "Excluded taxa" y manteniendo apretada la tecla command, los individuos que vamos a incluir en este clado: T1 a T6. Hacer clic en la flecha verde y automáticamente lo individuos pasan a la columna "Included taxa". Entonces, mrca1 define el punto de calibración del ingroup.
7. Repetir los puntos 5 y 6 para definir los puntos de calibración:
   mrca2: incluyendo T7 a T10.
   mrca3: incluyendo T9 y T10.
8. Ir a la solapa "Sites", donde se definen los modelos de substitución. Para ambos genes definir el modelo GTR+G. Notesé que al haber deslinkeado los modelos de substitución en el paso 3, los parámetros del modelo GTR+G de los genes serán estimados por separado.

9. Ir a la solapa "Clocks", donde se definen los modelos de reloj molecular a utilizar. BEAUti permite definir cuatro distribuciones de *priors* para la estimación de relojes moleculares al hacer clic en el menú desplegable "Model". Ellos son:
   Strict clock: asume que la tasa del reloj es estricto entre las ramas del árbol.
   Relaxed Clock Uncorrelated Lognormal: asume que las tasas de substitución asociadas con cada rama es independiente, y la estimación de las tasas se muestrea de una distribución lognormal (Drummond et al., 2006).
   Relaxed Clock Uncorrelated Exponential: asume que las tasas de substitución asociadas con cada rama es independiente, y la estiamación de las tasas se muestrea de una distribución exponencial (Drummond et al., 2006).
   Random local clock: utiliza una búsqueda Bayesiana aleatoria para la selección de promedios de variables sobre relojes locales aleatorios (Drummond and Suchard, 2010).
   Elegir la opción "Lognormal relaxed clock (Uncorrelated)" para ambos genes y tildar la caja "Estimate"
10. Ir a la solapa "Trees". Especificar "Speciation birth-death process" en el menú desplegable de "Tree prior".
    Este modelo y otros nombrados como "Speciation" son apropiados para análisis de relaciones entre especies. Estos modelos involucran procesos estocásticos de ramas con tasa constante de especiación (λ) y de extinción (μ). El caso del modelo Yule, la tasa de extinción es igual a 0. Estos modelos son implementados en BEAST siguiendo a Gernhard (2008), entonces para este caso de modelo birth-death el análisis muestrea meanGrowthRate = $\lambda - \mu$, y tasa relativa de extinción relativeDeathRate = μ / λ).
    Por otro lado, el modelo Coalescent es apropiado para análisis de poblaciones o especies cercanamente relacionadas. Elegir un modelo de este tipo para especies más distantes puede traer problemas por la interacción de *priors* en los nodos y las tasas de las ramas.
11. Ir a la solapa "Priors". En esta ventana se detallan los priors del análisis y sus distribuciones. Hacer clic en gene1.ucld.mean, y definir una distribución exponencial con media 10. Repetir para gene2.ucld.mean.
    En una terminología Bayesiana, un hiperparámetro es aquel parámetro que describe una distribución prior y no un parámetro fijado directamente para el modelo. En una inferencia Bayesiana, una distribución prior puede ser ubicada en un hiperparámetro, y es llamado hiperprior. Al permitir que el valor del hiperparámetro varíe, estamos libres de responsabilidad de fijar una media. Entonces, la cadena de Markov va a muestrear de este hiperparámetro directamente asociado al modelo de los datos, proveyendo una estimación de una distribución posterior.
12. Hacer clic en birthDeath.relativeDeathRate y seleccionar una distribución uniforme con valor upper = 100000 y lower = 0, y valor inicial 0.1.
13. Hacer clic en birthDeath.meanGrowthRate y seleccionar nuevamente una distribución uniforma con valor upper = 1 y lower = 0, y valor inicial 0.5.
14. Hacer clic en TreeModel.rootHeight. Este hiperparámetro permite definir el punto de calibración de la raíz. Siguiendo la figura de más arriba, definimos en este nodo una distribución normal. El valor inicial y la media = 140 y desvío estándar = 10. Las unidades de estos valores corresponden a millones de años, por lo que situamos la media de divergencia de la los linajes más antiguos (esta es la raíz) hace 140 millones de años y lo asociamos a un desvío de 10 millones de años.
15. Hacer clic en tmrca(mrca1). Este punto corresponde al que definimos para mrca1 (conformado por individuos T1 a T6). Elegir una distribución exponencial con media 30 y offset = 60. Esto quiere decir, que ubicamos el ancestro común del ingroup hace 30 millones de años en promedio, y offset de 60 millones. Observar qué valores son más probables de ser muestreados en esta distribución.

16. Hacer clic en tmrca(mrca2). Elegir una distribución lognormal, cuyo valor esperado sea = 20. Entonces, si el valor esperado siguiendo una distribución lognormal, es = 20, se calcula:
$$\mu = \ln(20) - (0.75^2/2)$$
$$= 2.714482$$
donde 0.75 corresponde al valor del desvío estándar que utilizaremos para esta calibración particular.
Entonces, seleccionar como media de la distribución lognormal = 2.714482 y desvío 0.75. Otra opción es tildar la opción "mean in real space" y entonces directamente se puede igualar la media a 20, y el resultado es el mismo, si necesidad de transformar la media como hicimos anteriormente.
17. Hacer clic en tmrca(mrca3). Elegir una distribución uniforme. Para esta distribución se especifica un valor máximo y mínimo y todo el rango tiene exactamente la misma probabilidad de ser muestreado. Por el contrario, ningún valor fuera de este rango puede ser seleccionado, ya que no es abarcado por la distribución del hiperparámetro.
En este caso, seleccionamos como valor máximo = 42, y mínimo = 8.
18. Ir a la solapa MCMC. Definir como largo de la cadena (en Length of chain) = 1,000,000 (este será un análisis corto, para no demorar la clase!, lo recomendable es un largo de cadena > 100,000,000, siendo posible que se requiera un análisis incluso más largo).
19. En general, es conveniente que el análisis cuente con 10,000 árboles y estimaciones de los parámetros muestreados. Entonces, para un análisis de largo 1,000,000, fijemos Log parameters every = 100 (ya que 1,000,000/10,000 = 100). Lo mismo para Each state tos screen every.
20. Hacer clic en Generate BEAST file… (abajo a la derecha) y guardar dentro de la carpeta de este tutorial.
21. Ubicar el archivo .xml que acabamos de generar y abrirlo en un editor de texto. En este archivo están detallados toda la información que fueron provistos a BEAUti en los pasos anteriores. Examinar el contenido.
22. Abrir BEAST, haciendo doble clic sobre el icono BEAST contenido en el paquete del programa.
23. Hacer clic en el botón "Choose file…" y ubicar divtime.xml.
24. Hacer clic en el botón "Run" y esperar a que el análisis se complete.

Validez del análisis: Tracer

1. BEAST generó dos outputs, uno divtime.trees y el otro divtime.log. El primero contiene un total de 10,000 árboles que fueron muestreados a lo largo del análisis. El otro, guarda la estimación de los parámetros para cada generación de MCMC muestreada. Abrir el programa Tracer e importar el archivo generado por BEAST divtime.log.
2. El análisis realizado en este tutorial es corto (por fines prácticos), por lo que la estimación de los parámetros no es buena. Esto se puede ver rápidamente en Tracer, ya que los valores de effective sample size (ESS) son bajos (<200), y se encuentran coloreados en rojo o amarillo. Hacer clic sobre alguno de los parámetros cuyo valor de ESS sea < 200. Explorar la forma de la distribución observada. Cambiar de solapas Estimates y Trace, y observar la información que provee cada uno.
3. Ahora nos enfocaremos en estudiar el resultado de un análisis más largo. Ubicar el archivo divtime.log. Este es un .log que se obtuvo utilizando la misma matriz anterior, pero simplemente realizando un análisis de MCMC de 100 millones de generaciones. Importar este archivo en Tracer.
4. Observar nuevamente la estimación observando diferentes parámetros usando la solapa Estimates y Tracer. Este resultado muestra una análisis mucho más confiable!!!

5. Ubicar los parámetros tmrca(mrca1), tmrca(mrca2) y tmrca(mrca3). Estos parámetros guardan la información de los tiempos de divergencia para los nodos que fueron calibrados.

Resumiendo la información de los árboles: TreeAnnotator y FigTree.

1. Abrir TreeAnnotator ubicado dentro del paquete BEAST. Este programa permite tomar la información de los outputs de BEAST y devolver el árbol resumen y probabilidades posteriores del análisis.
2. Hacer clic en el botón Choose file de la opción Input file y seleccionar el archivo divtime.trees. (usaremos este output, ya que el resultado de tan solo 1 millón de generación demostró un resultado poco confiable).
3. Hacer clic en el botón Choose file de la opción Output file y escribir el nombre del output como divtime.tree.
4. Especificar burnin = 1000 (corresponde a descartar los primero 1000 árboles, de un total de 10,000, entonces 10% burnin); Posterior probability limit = 0.5.
5. Hacer clic en el botón Run.
6. Abrir el programa FigTree (buscar entre los programas dados para el curso, no está incluido en el paquete BEAST). Este programa permite visualizar y modificar árboles.
7. Importar divtime.tree (File>Open).
8. Hacer clic en las opciones Node Labels y Node Bars. Aparecerán las estimaciones de tiempos de divergencia en los nodos y las barras de error correspondientes.
   FigTree permite realizar varios cambios de visualización en los árboles. Probar de cambiar los atributos con las diferentes opciones (cambiar cosas sin miedo!!!).
9. Para exportar el árbol final, ir a File>Export PDF para exportar en pdf, o bien File>Export Graphic para exportar en formato de gráfico (.png, .jpg, .eps, entre otros).

Ejercicios

En la carpeta de datos, esta ubicada otra matriz FPV.nex, la cual corresponde a datos de secuencias de papiloma virus obtenidos de diferentes hospedadores felinos: Felis, Lynx, Puma, Lion, Leopards. Además incluye dos outgroups: Canine y Racoon. NOTA: Esta matriz es parte del tutorial que proveen los creadores de BEAST, el cual se encuentra en la carpeta otros_tutoriales. Para más información ubicar el pdf.

Repetir este tutorial usando está matriz y ajustando según corresponda. Algunos datos necesarios:

Taxon Sets a definir:

Felis/Lynx/Puma (monofilético)

Lion/Leopard (monofilético)

Felinos: Felis-Lynx-Puma-Lion-Leopard (monofilético)

Modelo de substitución molecular: HKY+G.

Clock model: Uncorrelated Lognormal.

Calibración: datar la divergencia de Felis-Lynx-Puma, siguiendo una distribución normal, con media 7.15 y desvío 1.36.

Largo de la cadena: depende del tiempo disponible! Recomendable, al menos 10 millones.

Referencias

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology. 4:e88.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biology. 8:114.

Gernhard T. 2008. The conditioned reconstructed process. Journal of Theoretical Biology. 253:769–778.