Inferencia Bayesiana

Datos Moleculares en Biodiversidad II

Nelson R. Salinas

Septiembre 26, 2020

Repositorio sesión:

https://github.com/nrsalinas/datos_moleculares_ii/tree/master/Taller_6.

El presente taller será evaluado. El estudiante debe presentar un reporte a manera de miniartículo resumiendo las actividades desarrolladas. El taller debe ser entregado antes de la sesión del sábado 3 de octubre.

1. Si aún no lo ha hecho, instale Java Runtime Environment.
2. Descargue la suite de programas de inferencia phylogenética bayesiana Beast, versión 2.
3. Descargue el programa Tracer, utilizado para verificar la eficacia de muestreo de ejecuciones de cadenas de MCMC.
4. Descargue el archivo `ML_bootstrap.tree`, que contiene el mejor árbol de la búsqueda de máxima verosimilitud (ML) en RAxML con un valor de soporte clados: frecuencia de bootstrap.
5. Visualice el mejor árbol de ML en Figtree. Ejecute la aplicación y con ella abra el archivo. Figtree le pedirá que asigne un nombre a las etiquetas de los nodos; nómbrelos "bootstrap". Luego habilite la opción para ver las etiquetas de los nodos en el menú de la izquierda (*Node Labels*) y en la dentro de esta pestaña indique que las etiquetas a desplegar son los valores de soporte (*Display* $\rightarrow$ *bootstrap*). Si los valores de bootstrap se ven muy sobrepuestos utilize las opciones *Zoom* y *Expansion* para mejorar la visualización. El valor de soporte siempre aparecerá a la derecha del nodo correspondiente.
6. Utilice Tracer para confirmar que las cadenas MCMC convergieron. Primero abra al mismo tiempo los archivos `params_0.log` y `params_1.log`. Los dos archivos son MCMC independientes que corrieron por 6 millones de generaciones, pero fueron muestreadas al archivo de reporte cada 1000 generaciones. Ajuste el *burnin* de ambas cadenas al 20% de las muestras (1200000), para lo cual simplemente debe digitar el número deseado en la celda apropiada de la sección superior izquierda.

7. Diagnostique las cadenas, es decir, defina si muestrearon apropiadamente el espacio paramétrico. Para ello observe lo siguiente:

   (a) Seleccione las filas de los dos archivos en la tabla superior izquierda.

   (b) En la tabla inferior izquierda observará listado de parámetros reportados por la MCMC: verosimilitud del modelo (`LnL`), prior (`LnPr`) y los todos los parámetros relacionados con los modelos de substitución de nucleótidos, como son las tasas individiales de substitución (e.g., `r(A<->C){1}`), frecuencia estacionaria de cada nucleótido (e.g., `pi(A){1}`), el parámetro alpha de la distribución gamma de heterogeneidad de substituciones (e.g., `alpha{1}`) y la proporción de sitios invariables (e.g., `pinvar{1}`).

   (c) En la misma tabla observe la columna denominada ESS, que corresponde al tamaño poblacional estimado del parámetro. De acuerdo a lo mencionado en la clase, ¿son valores altos que reflejan un buen muestreo de la MCMC?

   (d) Seleccione el parámetro de la verosimilitud (LnL) y señale en el panel de la derecha la pestaña de *Trace*. La gráfica que observará indicará cómo fue la dispersión del muestreo a lo largo de las generaciones de la MCMC. Usualmente se debe observar una gran variación, pero la media debe permanecer más o menos invariable. En términos prácticos, debe observar el patrón de "oruga peluda". ¿Es ese el caso para los parámetros de los archivos?

   (e) Otra forma de verificar la convergencia es asegurarse que las distribuciones de cada parámetro en dos cadenas independientes tengan más o menos las mismas propiedades: misma media y mismo patrón de dispersión. Esto se puede verificar en Tracer seleccionando los dos archivos y un parámetro, luego en el panel de la derecha seleccionando la pestaña *Marginal Density*. Si son distribuciones parecidas las gráficas de ambos archivos deben estar sobrepuestos casi en su totalidad. ¿Observa este patrón en los archivos analizados?

8. Diagnostique las cadenas producidas por un análisis mucho más corto (400000 generaciones de MCMC). El reporte de dicho análisis está contenido en los archivos `params_0_sub.log` y `params_1_sub.log`. Abralos en Tracer y repita las observaciones del punto 7. ¿El muestreo de este experimento es apropiado? ¿Por qué?

9. Descargue del repositorio de la clase los arboles producidos por el experimento MCMC de 6 millones de generaciones, los cuales están depositados en el archivo `chain_0.trees`.

10. Anote las distribuciones de los parámetro en el árbol de mayor probabilidad. Abra el programa TreeAnnotator, que se encuentra en el folder de ejecutables de Beast (descargado en el punto 2). Como porcentaje de burnin seleccione 20%. Como árbol plantilla seleccione el árbol de máxima credibilidad (*Maximum clade credibility tree*) en la opción *Target tree*. Para la altura de los nodos (*Node heights*) se conservarán las longitudes del arbol de máxima credibilidad (*Keep target heights*). A continuación seleccione el archivo descargado en el punto anterior como *Input tree file*. Como nombre del archivo de salida digite "bayes.tree" en el campo *Output file*.

11. Visualize el arbol creado en el punto anterior en FigTree. Active la pestaña *Node labels*. En la opción *Display* de la misma seleccione *posterior*, que es la frequencia con la cual dicho nodo se encuentra en la distribución posterior de árboles, por lo cual presenta valores entre 0

y 1. Compare las medidas de soporte de los clados de interés (los clados que contienen taxa parásitos) con los encontrados en el análisis de Máxima Verosimilitud. ¿La topología del árbol de máxima credibilidad es igual a la del árbol de máxima verosimilitud? ¿Observa diferencias en la magnitud relativa de los soportes encontrados por cada método de inferencia? Recuerde que la frecuencia de bootstrap se encuentra en el rango 0-100 y la probabilidad posterior en el rango 0-1.