PRÁCTICAS BIOLOGÍA EVOLUTIVA

Inferencia bayesiana. Guion para el docente

Objetivos

- Introducir o reforzar los conceptos básicos de inferencia bayesiana
- Realizar reconstrucciones bayesianas mediante MrBayes y su línea de comandos
- Interpretar y discutir distintas reconstrucciones filogenéticas tanto cambiado los parámetros del programa como comparándolas con las obtenidas con otros criterios

Preparación

La duración estimada de la práctica es de unas dos horas en laboratorio de informática

Software necesario:

- MrBayes versión 3.2.7 o equivalente instalado en los ordenadores. Funciona con Java, así que basta copiar y pegar las carpetas, pero está disponible libremente (https://nbisweden.github.io/MrBayes/download.html) – pre-compiled version –
- Mesquite (por mantener continuidad con prácticas anteriores, pero valdría FigTree u otro software para visualizar y manipular árboles)
- (Optativo) Tracer para visualizar los resultados de la MCMC con mayor detalle (descargable en https://github.com/beast-dev/tracer/releases/tag/v1.7.2)

Documentación adicional:

- Documento "Implementing Models on MrBayes", una "chuleta" para implementar los modelos en el programa
- Manual de MrBayes en pdf (como material de consulta opcional)
- Alineamiento de los erizos en formato nexus (se les puede pedir que lo transformen ellos de fasta a nexus si se quiere o se les puede dar ya el alineamiento transformado)
- Presentación de diapositivas
- Este guion para el profesor

Desarrollo

<u>Introducción teórica</u>

Se retoma donde concluyó la práctica anterior con el resultado de la reconstrucción de la filogenia de los erizos con MV. La idea es que los estudiantes sean capaces de entender qué hizo MEGA con los datos. Es una buena ocasión para recordar las distintas partes del proceso:

1. El criterio de "optimalidad". ¿Cómo definimos "el mejor árbol" según la MV?
2. El mecanismo de búsqueda (algoritmo heurístico, NNI u otro)
3. La estimación de la solidez del resultado (pseudorréplicas *bootstrap*)

Llegados a este punto podemos exponer alguno de los problemas derivados de usar MV con MEGA, como los máximos locales, o presentar la idea de que la probabilidad

posterior de un árbol es, como concepto, más interesante de alcanzar que la verosimilitud. Estas limitaciones se aplican al análisis básico hecho con MEGA, no quiere decir que no haya otras formas posibles de utilizar MV.

Las siguientes diapositivas son una introducción a la probabilidad bayesiana y al teorema de Bayes. La cuestión clave es que lo que querríamos hacer (obtener una probabilidad posterior a partir de la verosimilitud) se vuelve matemáticamente posible. Se incluye un ejemplo con qué significa "probabilidad de lluvia" desde una óptica frecuentista y desde una óptica bayesiana.

Opcional. A criterio del profesor: si el ejemplo de la probabilidad de lluvia no es suficiente, se incluye un ejemplo inspirado en los de Kahneman y Tversky que salen en este vídeo de Grant Sanderson (3blue1brown) **Bayes theorem, the geometry of changing beliefs** https://www.youtube.com/watch?v=HZGCoVF3YvM&

Supongamos que vemos a un estudiante en el campus, sentado en un banco, dibujando. Dibuja muy bien. ¿Qué probabilidad crees que hay de que este estudiante sea de bellas artes?

Se pregunta a los estudiantes, esperando quizá que digan de forma intuitiva probabilidades altas, 0.8 por ejemplo. El problema aquí es que no están teniendo en cuenta información de contexto (*prior*) que puede modular la estimación. En este caso nuestro prior sería la probabilidad de estudiar bellas artes (una minoría en el conjunto del campus)

Estimaciones de este ejemplo (inventados):
Proporción de estudiantes de Bellas Artes: digamos, uno de cada 100 → **P(BA)=0.01 (prior)**
Probabilidad de dibujar bien entre los que no son de bellas artes → P(dib/¬BA)=0.1
Prob de dibujar bien entre los estudiantes de BA: 0.8 → **P(dib/BA)=0.8 (verosimilitud)**

Aplicando el teorema de Bayes se obtiene que la probabilidad posterior (lo que buscábamos, es decir, la probabilidad de que el estudiante sea de bellas artes sabiendo que dibuja bien) es de: un 0.075, seguramente muy inferior a lo que hubiésemos esperado.

El mensaje clave aquí es que cierta información preliminar (el *prior*), aunque sea vago, puede mejorar sustancialmente el nivel de certidumbre de una estimación y, a menudo, permitirnos alcanzar estimaciones que a primera vista nos hubiesen parecido poco probables o poco intuitivas. Nos da igual que en realidad la proporción de estudiantes de bellas artes sea de 1 de cada 50 y no de 1 de cada 100, lo que importa es que al añadir en nuestro cálculo que son una minoría, se alcanza una estimación más realista.

Después se pasa a explicar cómo usamos inferencia bayesiana y MCMC -y por qué- en filogenia molecular. La presentación de diapositivas está pensada para ser lo más auto-explicativa posible y que se entienda el papel de los distintos elementos.

Análisis del alineamiento de erizos con MrBayes

Para evitar retrasos, lo mejor es colocar el archivo ejecutable de MrBayes en la misma carpeta donde esté el alineamiento nexus.

Las diapositivas muestran paso a paso cómo realizar un análisis básico, pero lo ideal es ir pidiendo a distintos grupos de estudiantes que vayan variando ciertos parámetros como se cuenta a continuación para evaluar su efecto en los resultados. En realidad, el alineamiento del 18S de erizos no va a ofrecer resultados muy distintos ni por cambiar el modelo de nucleótidos, ni los prior, ni por aumentar el número de generaciones o el burnin. Esta también es una oportunidad para que el estudiante practique con una línea de comandos nueva, distinta a la de R, por lo que usando *help* se puede ir explorando la sintaxis de cada comando.

Pasos esenciales

- Se carga el alineamiento (>execute erizos.nex)
- Se implementa el modelo de sustitución K2+G (>lset nst=2 rates=gamma)
- De entre todos los parámetros de prior que se pueden alterar, elegimos el de la longitud de ramas para explicar y discutir con los estudiantes (>prset brlenspr = unconstrained:gammadir(1,0.1,1,1)). Ellos no lo saben, pero en realidad este es el valor por defecto del programa, porque la distribución de las longitudes en filogenia suele acercarse a este tipo de distribución. Por eso podemos estar tranquilos dejando el resto de las opciones por defecto, ya que tienen sentido biológico -los programadores diseñaron MrBayes específicamente para hacer filogenias.
  (ver cuadro en página siguiente para dirigir una discusión sobre con los alumnos)

- Se programa el MCMC con 2000000 de generaciones salvando uno de cada 1000 árboles (>mcmcp Ngen=2000000 samplefreq=1000 printfreq=1000 nruns=2 nchains=4 starttree=random)
- El resultado debería estar listo en unos 5 minutos y con buenas convergencias

- Repaso de parámetros (sump). Se hace primero uno sin burnin (>sump relburnin=no) para que se vea el ascenso inicial. Después se hace uno con burnin de 200 (>sump relburnin=no burnin=200) para comparar (responder “sí” cuando nos pregunten si dejamos que se sobreescriban los ficheros)

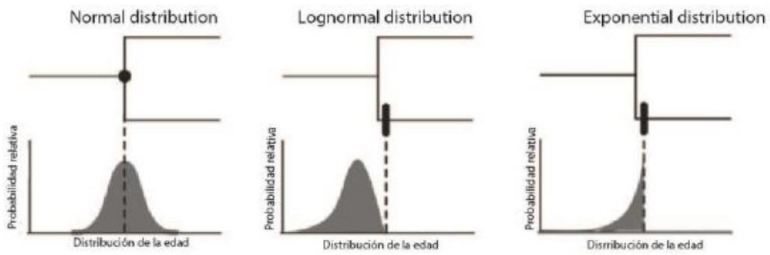
- Generar árbol consenso (>sumt contype=halfcompat burnin=200)

- El archivo del árbol se llamará **erizos.nex.con.tre** Se puede cargar en Mesquite tal y como se explica en la diapositiva

Explicación sobre los prior en filogenia molecular

Al igual que en el ejemplo del estudiante que dibuja bien, un prior es cierto conocimiento o estimación previa que va a ayudarnos a modular el resultado. Hay muchos parámetros que definen el cálculo de un árbol filogenético (topología del árbol; frecuencias de cada nucleótido; longitud de las ramas del árbol; cada una de las tasas de sustitución entre cada par de nucleótidos,...). La cuestión clave aquí: aunque desconozcamos los valores exactos de estos parámetros, hay algunos de ellos que tienen más sentido que otros desde el punto de vista biológico. Todos los valores posibles no son todos necesariamente equiprobables, y de hecho podemos hacernos una idea de qué *distribución* pueden tener esos valores.

Ejemplo de visualización de qué tipo de prior podríamos aplicar al parámetro "antigüedad de un nodo".

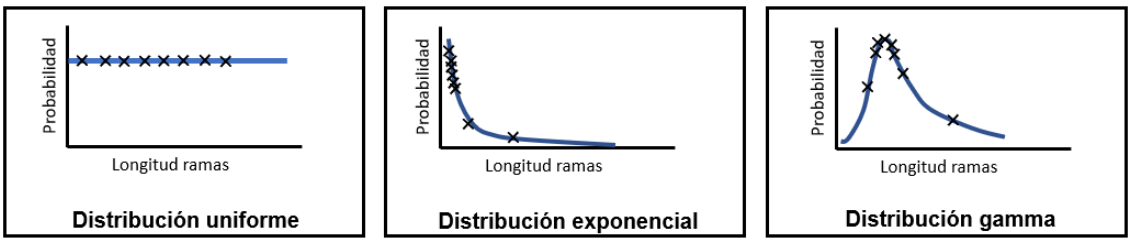


En la práctica no podemos valorar todos estos parámetros, así que usaremos solo uno: **longitud de ramas** las ramas. Comenzaremos preguntando a los estudiantes cuál de las tres distribuciones (equiprobable, exponencial y gamma) deberíamos escoger. Se trata de que entiendan que la elección del prior es importante y que tiene que tener sentido biológico.

- Una distribución **equiprobable** no tendría sentido biológico. Las longitudes de rama están relacionadas con la cantidad de cambios, y todas las secuencias descienden de un ancestro común. Son procesos divergentes, no valores calculados al azar

- Una distribución **exponencial** podría tener sentido, por ejemplo, si fuésemos a estudiar una radiación evolutiva rápida de especies surgidas desde el último máximo glaciar (esperaríamos ramas muy cortas, que han divergido hace poco.

- Pero para nuestros erizos, con linajes que divergieron hace muchos millones de años, lo esperable es que aunque las ramas se muevan alrededor de ciertos valores concretos, reflejando cierta distancia desde que divergieron. Optamos por una distribución **gamma**.



Qué **NO** son los prior

- Los prior no aportan estimaciones numéricas de esos parámetros, sino distribuciones de probabilidad de los mismos

- Los prior no "limitan" el espacio de búsqueda (plano XY de la visualización de la MCMC), sino que afectan al eje Z (es uno de los multiplicandos del numerador de la probabilidad posterior).

Variaciones que pueden explorarse

* Variar los modelos (JC, K2 y GTR), siguiendo la "chuleta" adjunta (esto puede cambiar, por ejemplo, la duración de las carreras).
El resultado variará poco para este set de datos, pero el tiempo de computación será mayor con el modelo más complejo.

* Variar los *prior*. Por ejemplo, poner una distribución uniforme para longitud de ramas: prset brlenspr = unconstrained:uniform(0,1)
Con este set de datos "empeorar" los prior no tiene mucho efecto, lo que nos puede servir como pista de que los resultados no se han visto demasiado afectados por ellos en este caso.

* Variar el número de generaciones. Probar con algún número muy bajo, tal como 20.000 y hasta 10 millones y ver el efecto que tiene en las probabilidades posteriores de cada nodo.
Si las generaciones son muy pocas, la probabilidad posterior de muchos nodos disminuirá ya que la "nube de árboles" no ha llegado a definirse bien. No habrá diferencias muy grandes, sin embargo, entre 2 millones o 10 de generaciones.

* Añadir carreras y cadenas adicionales.
Añadir demasiadas carreras y cadenas tendrá como consecuencia que costará más llegar a la convergencia, porque las cadenas calientes estarán forzando a algunas carreras a "salirse" del pico.

* cambiar el tipo de árbol consenso
(*contype=allcompat* para consenso estricto, por ejemplo). *Servirá de repaso sobre estos conceptos*

Discusión final

Una primera parte de esta puesta en común la constituye la comparación entre los resultados de las variantes de los distintos análisis. En general no diferirán demasiado del que se ve en la diapositiva. Un detalle para comentar es que incluso aunque los parámetros del análisis sean los mismos, el árbol consenso nunca saldrá idéntico en todos sus detalles (por la propia naturaleza estocástica del MCMC. Esto ya refleja esa idea de incertidumbre inherente de la probabilidad bayesiana).

¿Qué significa que haya una probabilidad posterior asociada a cada clado? Con la IB no es necesario ejecutar un análisis paralelo de pseudorréplicas (bootstrap) para tener una idea de la "solidez" de nuestros resultados. La PP de cada rama ya nos sirve para ello. Se puede preguntar a los estudiantes qué valor escogerían como umbral para "creerse" un clado. (Normalmente estos umbrales suelen estar en 0.9 o 0.95, mientras que un nivel de bootstrap de 80 ya se considera "sólido" en muchas publicaciones)

Después se pueden comparar el árbol consenso de la IB y el de MV de la práctica anterior:

- Los clados de *Echinolampas+Echinodiscus* y *Echinocardium+Brissus* se mantienen e incluso "mejoran" su certidumbre con IB, (sobre todo este último)
- El clado *Sphaerechinus+Tripneustes* (que ya era muy endeble con MV) no se mantiene en nuestro árbol consenso de IB, si bien el grado resultante tampoco es que tenga una solidez
- No todos los nodos se resuelven con la IB

Es un buen momento para retomar los lugares comunes de la interpretación de árboles filogenéticos:

¿Qué árbol es "mejor"? ¿Qué sentido tiene esa pregunta?

¿Por qué utilizamos distintos criterios?

¿Qué conclusiones sacamos si distintos criterios resultan en árboles similares? ¿Y si no?

Al final de la presentación hay una tabla que, a modo de resumen, puede servir para comparar los tres criterios explicados en las prácticas (Parsimonia, verosimilitud e inferencia bayesiana). Los alumnos pueden intentar completarla por su cuenta y hacer una puesta en común final.

Referencias y recursos

Hall, B.G. (2005). Phylogenetic Trees Made Easy: A How-To Manual. Sinauer

Lemey, P., Salemi, M., & Vandamme, A. M. (Eds.). (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press.

Yang, Z. (2006). *Computational molecular evolution*. OUP Oxford

Canal de Anders Gorm Pedersen
(Department of Health Technology; Technical University of Denmark)
https://www.youtube.com/channel/UCnEi3WTOcqt7iDTBkHF1LYA/videos

Serie de vídeos de Paul O. Lewis
(Ecology & Evolutionary Biology, University of Connecticut) para phyloseminar
https://www.youtube.com/watch?v=1r4z0YJq580
https://www.youtube.com/watch?v=UsLeY0wZr4Y
https://www.youtube.com/watch?v=4PWlnNsfz90
https://www.youtube.com/watch?v=TLtOS--YwkU