Clase 5

Análisis filogenético Bayesiano

El marco Bayesiano

Análisis filogenético Bayesiano

- El análisis filogenético Bayesiano se desarolló a mediados de los 90s
- Ahora uno de los métodos más ampliamente usados

MrBayes 

 *BEAST* 1

RevBayes 

 *BEAST* 2

3

Análisis filogenético Bayesiano

Probabilidad	Modelo 1
Árbol 1	0.1
Árbol 2	0.7
Árbol 3	0.15
Árbol 4	0.05
Suma	1

Máxima verosimilitud = $\Pr(D \mid \theta)$

Inferencia Bayesiana = $\Pr(\theta \mid D)$

- Para tener un marco estadístico formal, la inferencia Bayesiana incluye un concepto **previo** de los parámteros, el **Prior**

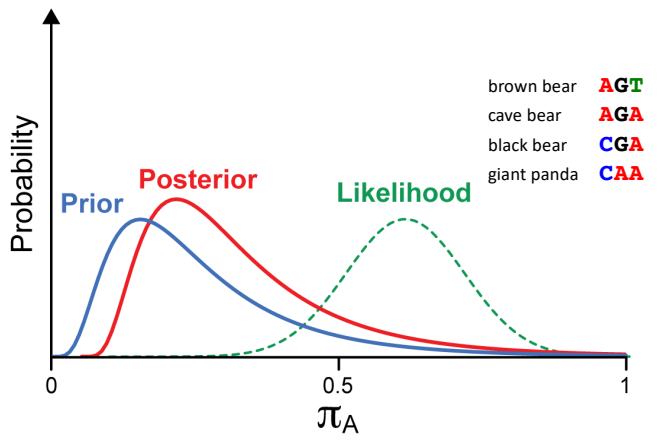
4

El paradigma Bayesiano

- Contrasta con la estadística frecuentista (verosimilitud)
- Los parametros vienen de **distribuciones**
- Antes de que los datos sean observados, cada parámetro tiene una **distribución como prior**
- Se calcula la **verosimilitud** de los datos
- Se combina (actualiza) la distribución del prior con la versimilitud para llegar a la **distribución a posterior**

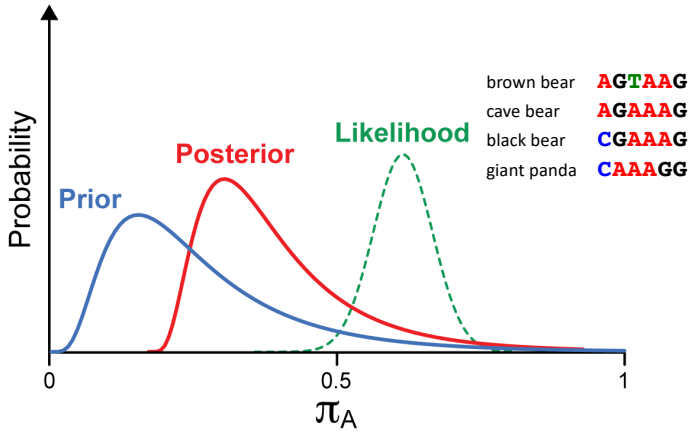
5

Ejemplo simple



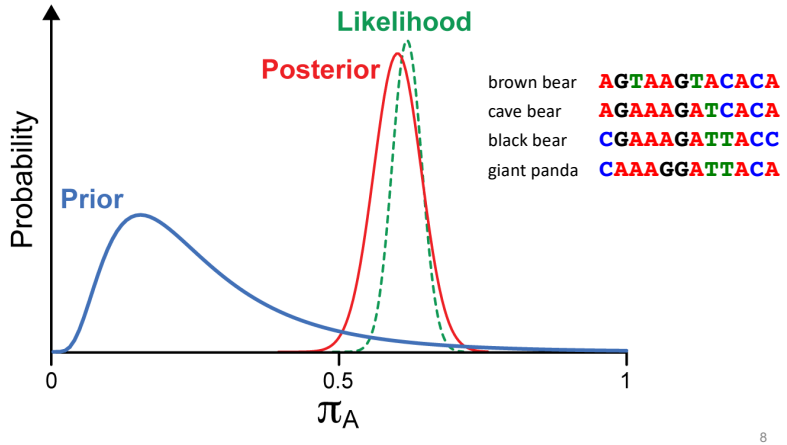
6

Ejemplo simple



7

Ejemplo simple



8

Inferencia Bayesiana

Prior
Especificado por el usuario, independiente de los datos

Verosimilitud
Calculada de los datos

$$\Pr(\theta \mid D) = \frac{\Pr(\theta)\Pr(D \mid \theta)}{\Pr(D)}$$

Posterior
Lo que queremos estimar

Constante normalizadora
La verosimilitud marginal de los datos dado el modelo **(no la tenemos)**

9

Inferencia Bayesiana

Prob a priori del árbol
La topología y las longitudes de ramas

Prob a prior de los parámetros del modelo de sustitución
Parámetros de tasas, frecuencias de bases

$$\Pr(\tau, M \mid D) = \frac{\Pr(\tau)\Pr(M)\Pr(D \mid \tau, M)}{\Pr(D)}$$

Posterior
Lo que queremos estimar

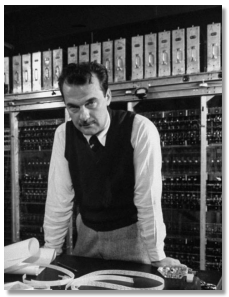
Verosimilitud
Calculada de los datos

10

Muestreo usando una cadena markoviana de monte carlo (MCMC)

Estimar el posterior

- No podemos obtener el posterior directamente
- Podemos estimarlo usando **simulaciones en una cadena markoviana de monte carlo**
- Esto se suele hacer usando el algorítmo **Metropolis-Hastings**



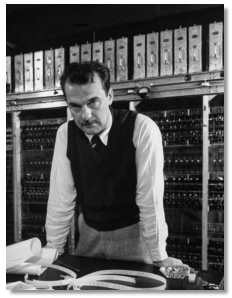
Nicholas Metropolis
Los Alamos, 1953

12

Estimar el posterior

- No podemos obtener el posterior directamente
- Podemos estimarlo usando **simulaciones en una cadena markoviana de monte carlo**
- Esto se suele hacer usando el algorítmo **Metropolis-Hastings**

$$\frac{Pr(\theta)\ \ Pr(D \mid \theta)}{Pr(D)}$$



Nicholas Metropolis
Los Alamos, 1953

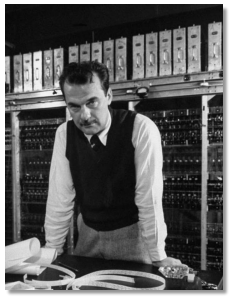
13

Estimar el posterior

- No podemos obtener el posterior directamente
- Podemos estimarlo usando **simulaciones en una cadena markoviana de monte carlo**
- Esto se suele hacer usando el algorítmo **Metropolis-Hastings**

$$\frac{Pr(\theta)\ \ Pr(D \mid \theta)}{Pr(D)}$$

$$\frac{Pr(\theta)\ \ Pr(D \mid \theta)}{Pr(D)}$$



Nicholas Metropolis
Los Alamos, 1953

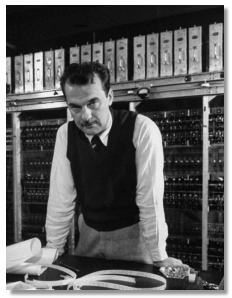
14

Estimar el posterior

- No podemos obtener el posterior directamente
- Podemos estimarlo usando **simulaciones en una cadena markoviana de monte carlo**
- Esto se suele hacer usando el algorítmo **Metropolis-Hastings**

$$\frac{Pr(\theta)\ \ Pr(D \mid \theta)}{\cancel{Pr(D)}}$$

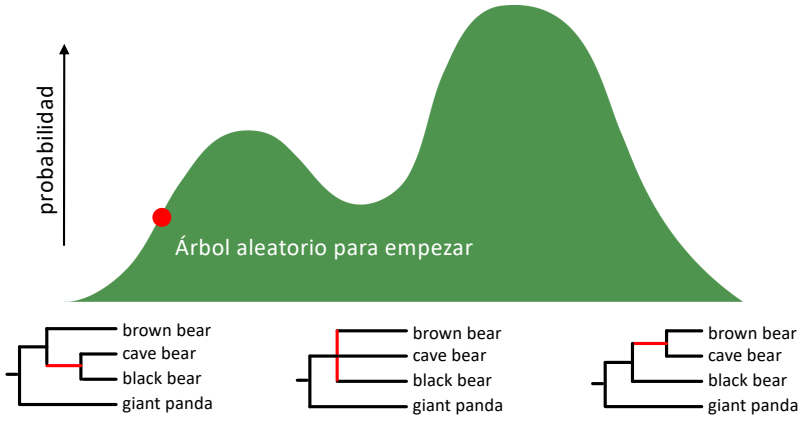
$$\frac{Pr(\theta)\ \ Pr(D \mid \theta)}{\cancel{Pr(D)}}$$



Nicholas Metropolis
Los Alamos, 1953

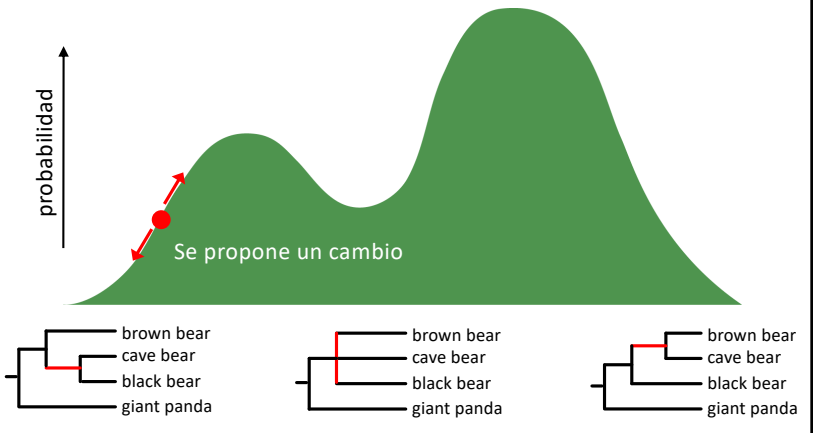
15

Simulación MCMC



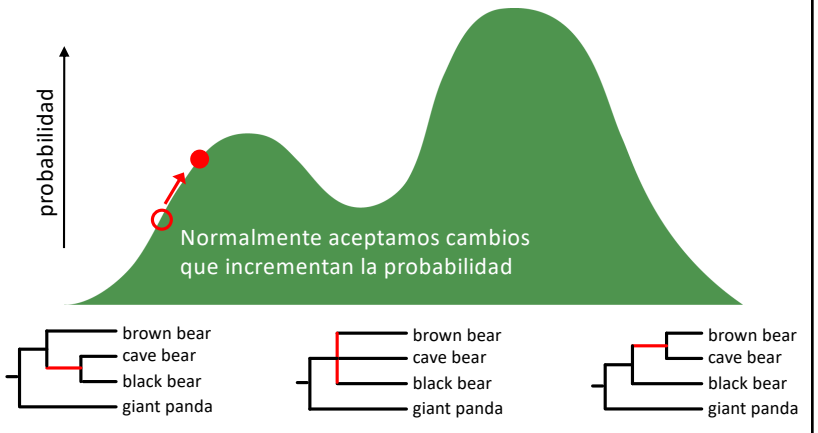
16

Simulación MCMC



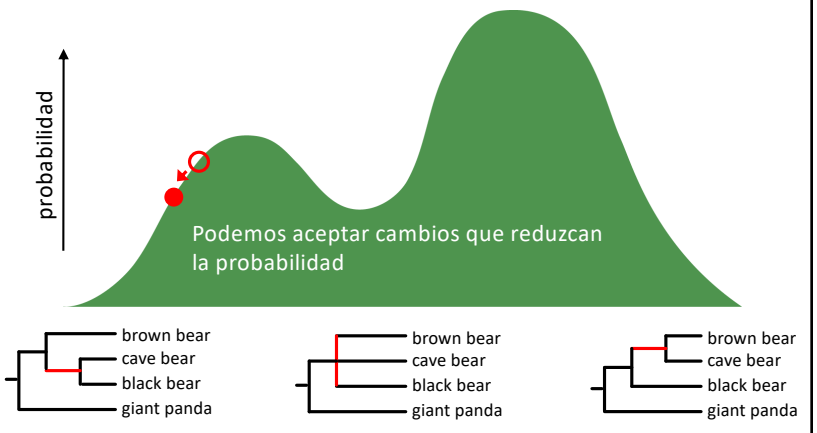
17

Simulación MCMC



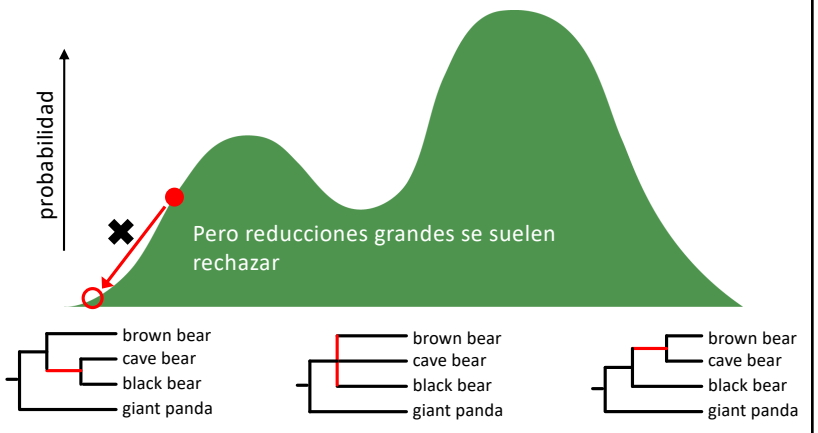
18

Simulación MCMC



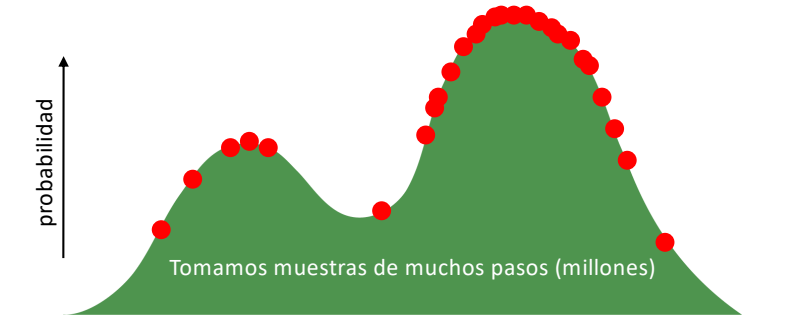
19

Simulación MCMC



20

Simulación MCMC



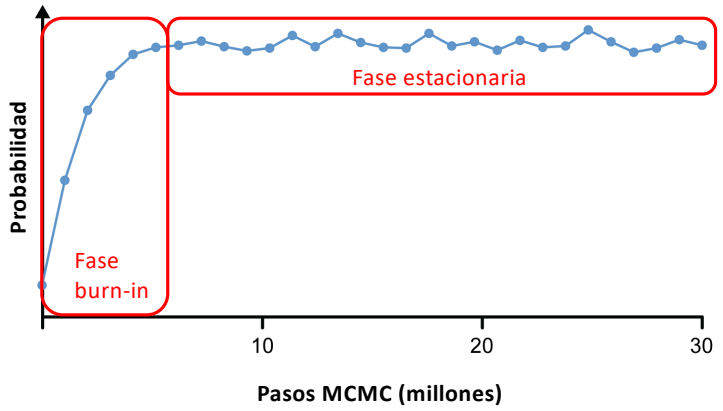
21

Muestras del MCMC

- El resultado de un análisis filogenético Bayesiano:
  - Una lista de **valores de los parametros** que visitamos en la cadena markoviana
    (archivo terminado en .p file usando *MrBayes*, y terminado en .log usando *BEAST)*
  - Una lista de árboles visitados por la cadena markoviana (terminado en .t usando *MrBayes*, y terminado en .trees cuando usasmos *BEAST*)

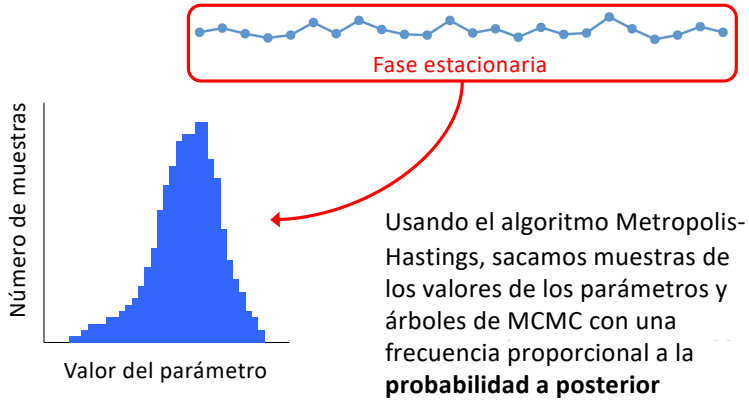
22

Muestras del MCMC



23

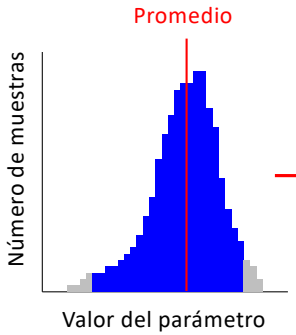
Muestras del MCMC



Usando el algoritmo Metropolis-Hastings, sacamos muestras de los valores de los parámetros y árboles de MCMC con una frecuencia proporcional a la **probabilidad a posterior**

24

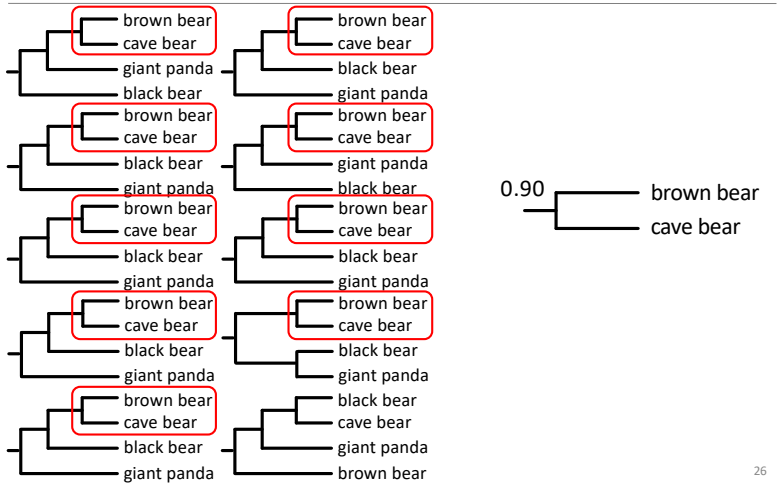
Muestras del MCMC



- Podemos tomar el promedio de los valores muestreados
  Estimativo promedio a posterior
- Y tomamos el 95% 'central' de los valores muestreados
  el intervalo de credibilidad del 95%

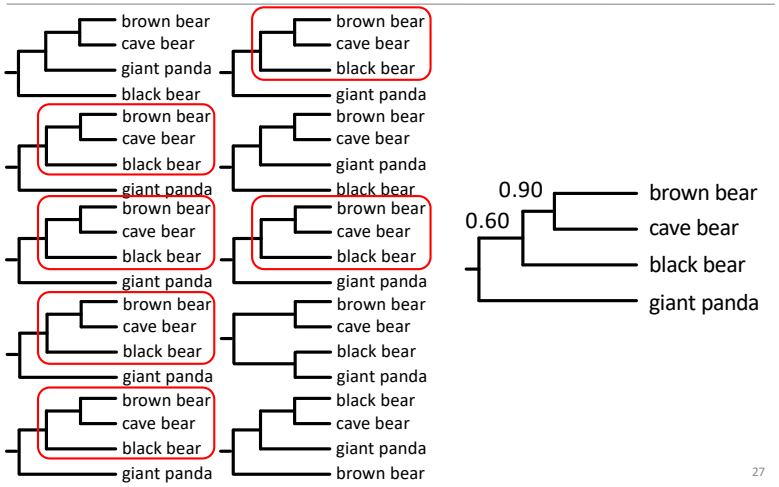
25

Muestras del MCMC



26

Muestras del MCMC



27

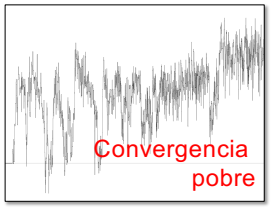
Muestras del MCMC

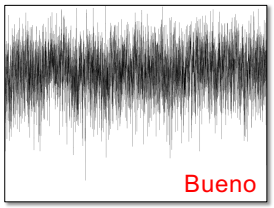
- **Árbol consenso de la mayoría (Majority-rule consensus; *MrBayes*)**
  Muestra todos los nodos que tengan credibilidad a posterior >0.50
- **Árbol con el posterior máximo (MAP)**
  El árbol muestreado con la mayor probabilidad a posterior
- **Árbol con clados cn máxima credibilidad (MCC; *BEAST/TreeAnnotator*)**
  El árbol muestreado con la mayor suma o producto de las probabilidades a posterior en los nodos

28

Diagnósticos

1. **Convergencia**
   ¿Estamos tomando muestras de la distribución estacionaria?
2. **Muestreo suficiente**
   ¿Ya tomamos suficientes muestras para permitirnos hacer estimativos confiables del posterior?





29

Convergencia

- Es recomendable llevar a cabo al menos 2 análisis independientes
- Las versimilitudes deberían ser similares
- Los estimativos de los parámetros del modelo deberían ser similares

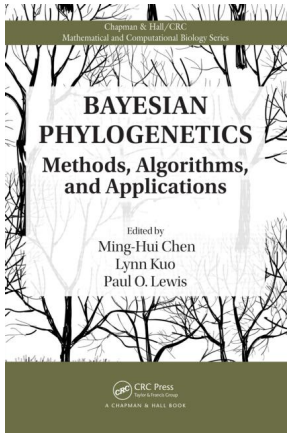
30

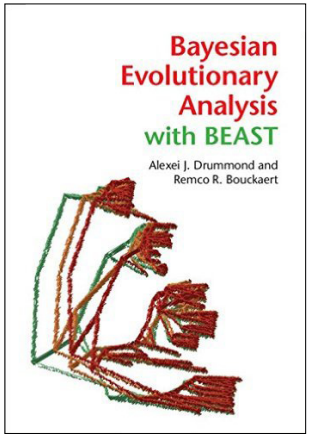
Muestreo suficiente

- **El número efectivo de muestras (*Effective sample size*; ESS)**
  ¿Ya tomamos suficientes muestras para permitirnos hacer estimativos confiables del posterior?
- El ESS debería ser preferiblemente **>200** para cada parámetro
- El ESS se puede incrementar:
  - Llevando a cabo MCMC por más tiempo, sacando más muestras (y reduciendo la frecuencia con la que se guardan las muestras)
  - Modificando que se hacen en MCMC

31

Referencias útiles





32