

Inferencia bayesiana de filogenias moleculares

Pablo Vinuesa

Centro de Ciencias Genómicas

UNAM

vinuesa@ccg.unam.mx

<http://www.ccg.unam.mx/~vinuesa/>

Inferencia filogenética molecular - clasificación de métodos

Podemos clasificar a los métodos de reconstrucción filogenética en base a:

- el tipo de **datos** que emplean (**caracteres discretos vs. distancias**)
- En base al **método** de reconstrucción de la topología, **método algorítmico vs. un criterio de optimización**

		Tipo de datos	
		distancias	caracteres discretos
Método de reconstrucción	algoritmo de agrupamiento	UPGMA Neighbour joining	
	criterio de optimización	Evolución mínima Mínimos cuadrados	MP ML (MV) bayesiana

Métodos de reconstrucción filogenética - La alternativa bayesiana

• Aproximaciones tradicionales (ML, MP ...)

1.- la búsqueda tiene por objetivo encontrar la topología óptima (**estima puntual**)

El cálculo de la función global de verosimilitud L_H puede tardar mucho si el problema de inferencia es muy complejo: (muchos OTUs, patrón complejo de sustituciones, modelo rico en parámetros, y datos con mucha homoplasia)

2. no pueden establecer el soporte relativo de las biparticiones a partir de una única búsqueda (ni medidas del error de estima de valores de parámetros)

- requerimos hacer análisis de bootstrap o jackknifing para obtener una medida de soporte de los clados
- aunque recientemente se han implementado métodos de aLRTs

Métodos de reconstrucción filogenética - La alternativa bayesiana

• Aproximación bayesiana

- muestrea una **población de árboles en función de su probabilidad posterior (pP)**

Probabilidad anterior (**prior**) incondicional de la hipótesis o parámetro

verosimilitud

$$pP(\tau_i | X) = \frac{\Pr(\tau_i) \Pr(X|\tau_i)}{\sum_{j=1}^{B(s)} \Pr(\tau_j) \Pr(X|\tau_j)}$$

Probabilidad (incondicional) de los datos (= constante normalizadora)

- la muestra de árboles obtenidos (τ_i) en una sola sesión de "muestreo" es usada para valorar el soporte de cada split en términos de pP

Inferencia bayesiana - el teorema de Bayes

• Thomas Bayes (1702-1761)
• Teorema de Bayes generalizado por Laplace (1763)



- La **inferencia bayesiana** se basa la relación cuantitativa existente entre la **función de verosimilitud** y las **distribuciones anteriores y posteriores** de probabilidad

Teorema de Bayes:
$$Pr(H|D) = \frac{Pr(H) Pr(D|H)}{Pr(D)}$$
 A = datos = D
B = hipótesis o parámetro = H

- $Pr(H|D)$ = **prob. posterior**; probabilidad de H (o valor del parámetro), dados D
- $Pr(D|H)$ = Esto es la **VEROSIMILITUD** DE LOS DATOS dada la hipótesis
- $Pr(H)$ = **probabilidad anterior** ("prior"); es la prob. incondicional de H
- $Pr(D)$ = **prob. incondicional de los D**, que puede ser obtenida usando la ley de la prob. total, calculando $\sum_H Pr(H) Pr(D|H)$. Funciona como constante normalizadora, asegurando que la sumatoria de $pP = 1$

Perspectivas frecuentistas vs. bayesianas en estadística -un ejemplo sencillo en un marco de datos discretos (tomado de P. Lewis 2001)

Urna A 40% bolas negras	Urna B 80% bolas negras
----------------------------------	----------------------------------

$$Pr(\text{urna A} | \text{sacamos bola negra}) = \frac{(0.5)(0.4)}{(0.6)} = 1/3 = 0.33$$
$$Pr(\text{urna B} | \text{sacamos bola negra}) = \frac{(0.5)(0.8)}{(0.6)} = 2/3 = 0.67$$

- La **distribución posterior** se puede concebir como una **versión actualizada de la distribución anterior** (después de haber visto los datos)
- En nuestro caso la distribución posterior (0.33, 0.67) es la versión actualizada de la distribución anterior (0.5, 0.5). La evidencia para actualizar la distribución fue el evento de haber sacado una bola negra
- Una de las grandes ventajas de la aproximación bayesiana sobre la frecuentista radica en el hecho de calcular probabilidades para las hipótesis (o valores de los parámetros) de interés. Las verosimilitudes son útiles, pero difíciles de interpretar, ya que representan la probabilidad de los datos dada la hipótesis. **La aproximación bayesiana permite estimar la probabilidad de la hipótesis dados los datos**, que es lo que queremos por lo general

Perspectivas frecuentistas vs. bayesianas en estadística -un ejemplo sencillo en un marco de datos discretos (tomado de P. Lewis 2001)

Urna A 40% bolas negras	Urna B 80% bolas negras
----------------------------------	----------------------------------

Cada urna cuenta con millones de bolas blancas y negras

D = 1 bola negra; H1: proviene de A; H2: proviene de B

¿Cuál es la probabilidad de que la bola haya salido de la urna A ó B?

- **Aproximación frecuentista (máxima verosimilitud):** $P_A = 0.4$; $P_B = 0.8$
- **Aproximación bayesiana:** nos permite seleccionar una prob. anterior que refleje por ejemplo nuestra ignorancia acerca de la distrib. de bolas blancas y negras en las dos urnas: la prob. anterior de cada urna = 0.5

$$Pr(D) = Pr(\text{bola negra}) = (0.5)(0.4) + (0.5)(0.8) = 0.6$$

$$Pr(H|D) = \frac{Pr(H) Pr(D|H)}{Pr(D)}$$
$$Pr(\text{urna A} | \text{sacamos bola negra}) = \frac{(0.5)(0.4)}{(0.6)} = 1/3 = 0.33$$

$$Pr(\text{urna B} | \text{sacamos bola negra}) = \frac{(0.5)(0.8)}{(0.6)} = 2/3 = 0.67$$

Perspectivas frecuentistas vs. bayesianas - conclusiones estadísticas

- Una crítica a las aproximaciones bayesianas radica en la subjetividad de los priors
- nótese que en el ejemplo de las 2 urnas y la bola negra (con un solo dato) sería necesario estipular un probabilidad anterior > 2/3 para la urna A para hacer "empatar o revertir" el resultado anterior (0.33, 0.67)

Urna A 40% bolas negras	Urna B 80% bolas negras
----------------------------------	----------------------------------

$$Pr(D) = Pr(\text{bola negra}) = (0.7)(0.4) + (0.3)(0.8) = 0.52$$

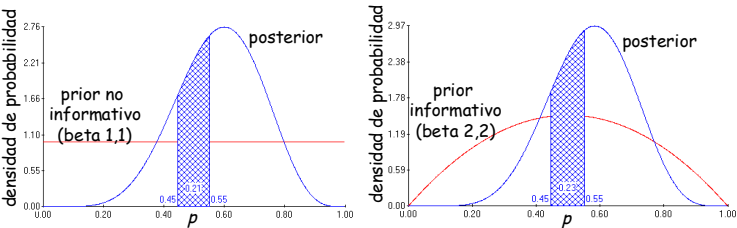
$$Pr(\text{urna A} | \text{sacamos bola negra}) = \frac{(0.7)(0.4)}{(0.52)} = 0.54$$

$$Pr(\text{urna B} | \text{sacamos bola negra}) = \frac{(0.3)(0.8)}{(0.52)} = 0.46$$

- Aunque la distribución posterior siempre cambia cuando lo hace la probabilidad anterior, las conclusiones no son generalmente muy sensibles al prior; de hecho **el efecto del prior decrece a medida que incrementa la cantidad de datos** (la función de verosimilitud "pesa más" en el análisis.
- De ahí que los análisis bayesianos generalmente comiencen con priors vagos o planos
- Además, la subjetividad inherente al prior es explícita y por tanto ha de ser defendible

Inferencia bayesiana con parámetros (hipótesis) continuos - funciones de densidad probabilística

En filogenética las topologías y caracteres ancestrales representan caracteres discretos, mientras que muchos de los parámetros de interés ($v, \alpha, K \dots$) son continuos. Para ellos las **funciones de densidad probabilística** reemplazan las probabilidades de las hipótesis discretas, pero el teorema de Bayes sigue siendo aplicable



Distribuciones de densidad anteriores y posteriores del parámetro p (probabilidad de soles) de un experimento de lanzado de monedas (10 repeticiones) con 6 soles como resultado. Se utilizaron dos priors, mostrándose el efecto que tienen sobre la probabilidad posterior de que obtengamos valores entre 0.45 y 0.55 de obtener soles (Ejemplo adaptado de Lewis, 2001; se usó el programa Bayesian coin-tosser de P.O. Lewis)

Reconstrucción filogenética bayesiana - MCMC

MCMC crea una **cadena de Markov** cuyas dimensiones corresponden a la hipótesis de interés (τ, v, θ) y cuya distribución estacionaria (equilibrio) es la distribución deseada (pP).

Pasos de la cadena según el algoritmo de Metropolis-Hastings (M-H):

1. Parte de un estado inicial aleatorio ($T_i = pP$ del árbol i)
 2. Se propone un nuevo estado próximo al anterior (T_j) - (i.e. puede explorar todo el espacio del parámetro).
 3. Se calcula el **cociente de probabilidades R** (o funciones de densidad probabilística) entre T_j y T_i
$$R = f(T_j)/f(T_i)$$
 4. Si $R \geq 1$ se acepta el nuevo estado (es decir, valores mejores de pP siempre se aceptan)
 5. Si $R < 1$ se toma un número aleatorio ξ entre 0 y 1, si $\xi < R$ se acepta el nuevo estado
 6. Si $\xi \geq R$ se rechaza T_j como nuevo estado y se continúa con T_i
 7. Se vuelve al paso 2 \Rightarrow la cadena se repite o corre cuantas más iteraciones mejor (∞)
- se trata de una cadena de Markov ya que se trata de un proceso estocástico en el que el siguiente estado depende sólo del estado actual y no del anterior
- la cadena visita estados (árboles y params. mod. sust.) proporcionalmente a su pP

Estima bayesiana de filogenias - cadenas markovianas de Monte Carlo

la probabilidad posterior de un árbol puede interpretarse como la probabilidad de que dicho árbol o clado sea correcto

$$Pr(\text{Arbol} | D) = \frac{Pr(\text{Arbol}) Pr(D | \text{Arbol})}{Pr(D)}$$

la probabilidad posterior de un árbol, aunque fácil de formular, implica la sumatoria sobre todos los árboles (τ) y, para cada árbol, la integración sobre todas las posibles combinaciones de longitudes de rama (v) y parámetros (θ) del modelo de sustitución

$$f(\tau_i | X) = \frac{f(\tau_i) f(X | \tau_i)}{\sum_{j=1}^{B(s)} f(\tau_j) f(X | \tau_j)} \quad f(X | \tau_i) = \int_v \int_\theta f(\tau_i, v_i, \theta) f(X | \tau_i, v_i, \theta) dv d\theta$$

- es imposible estimar dicha pP analíticamente ni siquiera para el caso más simple de 4 OTUs ($(2s - 3)!/2s - 2 (s - 2)!$ topologías y $2n - 3$ long. de rama, para arb. no enraiz.)
- existen **métodos numéricos** que permiten **aproximar la probabilidad posterior** de un árbol (o de cualquier otra hipótesis compleja). El más útil es el de las **cadenas markovianas de Monte Carlo** (MCMC), implementado en algoritmos como el de **Metropolis-Hastings**
- **MCMC se basa en el muestreo de una distribución simulada** en vez de calcular dicha distribución mediante integración. Así es posible aproximar el área bajo la curva que representa la distribución de densidad probabilística posterior para inferencias complejas

Estima bayesiana de filogenias - cadenas markovianas de Monte Carlo

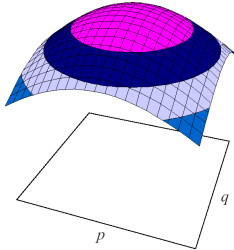
Para el experimento de lanzado de una moneda se podía representar la densidad posterior como una curva en un espacio 2-dimensional. **Con más de 1 parámetro, la densidad de P posterior se torna en una superficie en un espacio multidimensional** (una dimensión más que parámetros a estimar)

así p. ej. un problema sencillo de 2 parámetros, como el de hacer inferencias sobre la altura de mujeres y hombres en una población, implica una superficie de densidad 3-dimensional, en la que p = altura de las mujeres y q = altura de los hombres. Asumiendo una distribución normal, con varianzas conocidas, tendríamos una distribución posterior normal bivariada

para problemas filogenéticos tendríamos muchísimas más dimensiones que explorar y no se pueden representar gráficamente

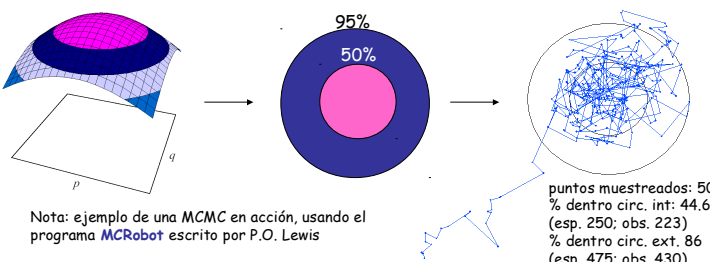
esto no representa un problema ya que no hace falta visualizar la distribución posterior para hacer inferencias a partir de ella

lo que necesitamos es poder calcular los volúmenes bajo ella, y es en ello en lo que MCMC nos ayuda



Estima bayesiana de filogenias

- cadenas markovianas de Monte Carlo



Nota: ejemplo de una MCMC en acción, usando el programa **MCRobot** escrito por P.O. Lewis

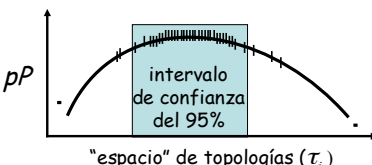
puntos muestreados: 500
% dentro circ. int: 44.6 (esp. 250; obs. 223)
% dentro circ. ext. 86 (esp. 475; obs. 430)

- Dado el **algoritmo de MH** se demuestra que el "robot" visita puntos proporcionalmente a su altura, que equivale a la func. densidad de probabilidad posterior de un problema bayesiano
- por lo tanto, para aproximar el volumen dentro del círculo interno, sólo hay que contar el número de pasos dados dentro de dicho círculo y dividirlos por el no. total de pasos dados. Esto vale para estimar el volumen bajo cualquier porción especificada del espacio
- cuanto más tiempo se le dé al robot para "pasear" por el espacio paramétrico, mejor será la aproximación al volumen o espacio real
- si se descartan los pto. fuera de los círculos ("burnin"), la cadena está visitando estados en proporción a su densidad probabilística

Métodos de reconstrucción filogenética

- la alteranativa bayesiana

- Aproximación bayesiana
 - se muestrea una **población de árboles en función de su probabilidad posterior (pP)**

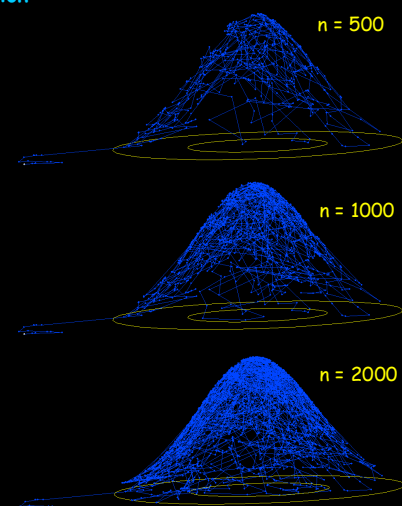

$$pP(\tau_i | X) = \frac{\Pr(\tau_i) \Pr(X|\tau_i)}{\sum_{j=1}^{B(s)} \Pr(\tau_j) \Pr(X|\tau_j)}$$

"espacio" de topologías (τ_i)

La proporción de veces que la cadena visita un cierto estado es una aproximación válida de la pP de ese estado (e.g. árbol filogenético). Así si de 10^6 muestras (τ_i) un clado es recuperado en 975676 de ellas \Rightarrow que dicho clado tiene una $pP \approx 0.98$

Para modelos de múltiples parámetros (filogenia) MCMC puede actualizar los estados de esos parámetros simultánea o individualmente.

cadenas markovianas de Monte Carlo - exploración de una distribución posterior normal bivariada



para las simulaciones se usó el programa **MCRobot** de P.O. Lewis



Estima bayesiana de filogenias usando MrBayes 3.2 (Ronquist et al., 2012)

- Los métodos bayesianos fueron introducidos al campo de la filogenética en 1996
 - Li, S. PhD thesis, Ohio State University, Columbus
 - Mau, B. PhD thesis, University of Wisconsin, Madison
 - Rannala, B. and Yang, Z. 1996. J. Molec. Evol. 43:304-311
- El primer programa que implementaba eficientemente algoritmos de MCMC para la inferencia bayesiana de filogenias fue puesto en el dominio público en 1998
 - Simon, D. and Larget, B. 1998. BAMBE, Duquesne Univ., Pittsburgh
 - Larget, B. and Simon, D. L. 1999. Mol. Biol. Evol. 16:750-759

<http://www.maths.duq.edu/larget/bambe.html>
- Junto a BEAST, es actualmente el programa más versátil y completo para la inferencia bayesiana de filogenias: **MrBayes 3.2.6 (GNU license, 25 Nov.2015)**
 - Huelsenbeck, J. P., and Ronquist, F. 2001. Bioinformatics 17:754-755
 - Ronquist F. et al. (2012). Syst Biol. 61(3):539-542

<http://mrbayes.sourceforge.net/>
<https://github.com/NBISweden/MrBayes>



Estima bayesiana de filogenias usando MrBayes 3.2 (Ronquist et al., 2012)

- MrBayes3 está escrito para diversas plataformas (UNIX, Windows y Macintosh) y se maneja de igual manera en las tres plataformas a nivel de línea de comandos
- Para correrlo eficientemente se necesita una computadora razonablemente rápida (CPU de alto rendimiento, P-IV) y con al menos 512 o 1 GB de RAM) si se pretenden analizar matrices de datos > 20 OTUs
- Además de la computadora y el programa necesitamos:
 - 1.- Los **datos** (X): secuencias (nt y aa) u otros caracteres discretos (morfológicos sitios de restricción etc.).
 - 2.- **Un modelo probab. de evolución** (modelos de la familia GTR (1, 2, y 6 tasas de sust.) o modelos de sust. de aa. basados en matrices empíricas (JTT, WAG, BLOSUM ...) distribución gamma de variación de tasas entre sitios y prop. sitios invariantes o modelos morfológicos
 - 3.- **Probabilidades anteriores para todos los parámetros del modelo:**
 - topología, longitudes de rama (2n-3)
 - freq. de nt o aas; tasas relativas de sust.
 - heterogeneidad de tasas (I, α)
- Para ver las opciones de comandos en MrBayes usar el comando <help> ó <help comando>

Estima bayesiana de filogenias usando MrBayes 3.2 - los datos y un bloque de comandos sencillo

• Los **datos** (X): se presentan en una variante del **formato NEXUS** (como el que usan PAUP* y MacClade)

```
#NEXUS
[En corchetes puede ir cualquier comentario, que es ignorado por MrBayes]

begin data;
dimensions ntax=38 nchar=1104; [dimensiones de la matriz]
format datatype=DNA interleave=yes gap=- missing=?; [formato de la matriz]
matrix
  BC_C1 CCGACTCCGAACTTGC GCGG CAAAACTCAGATCAAGGAAT ...
  BC_C2 CCGACTCCGAACTTGC GCGG CAAAACTCAGATCAAGGAAT ...
  BC_P6 CCGACACCGAATTTGC GCGG CAAAACTCAGATCAAGGAAT ...
  BC_P14 ??GACTCCGAACTTGC GCGG CAAACACAGATCAAGGAAT ...
;
end;
```

bloque de datos

```
begin mrbayes;
Lset nst=6 rates=invgamma Ngammacat=6; [modelo GTR+I+G]
mcmc ngen=3000000 printfreq=5000 samplefreq=100
nchains=4 temp=0.2 savebrlens=yes; [detalles de la cadena de MCMC]
end;
quit
```

bloque de comandos de MrBayes

Estima bayesiana de filogenias usando MrBayes 3.2 - un bloque de comandos para modelos particionados por codones y 2 genes

```
BEGIN mrbayes;
log start filename=38UglnII_recA_bycod_def01_log1;
delete Sm1021;
outgroup Rho_palustris;
charset glnII = 1-594;
charset glnII_1st = 1-594\3;
charset glnII_2cnd = 2-594\3;
charset glnII_3rd = 3-594\3;
charset recA = 595-1104;
charset recA_1st = 595-1104\3;
charset recA_2cnd = 596-1104\3;
charset recA_3rd = 597-1104\3;
partition by_gene = 2: glnII, recA;
partition by_codon = 6:glnII_1st, glnII_2cnd, glnII_3rd, recA_1st, recA_2cnd, recA_3rd;
end;
```

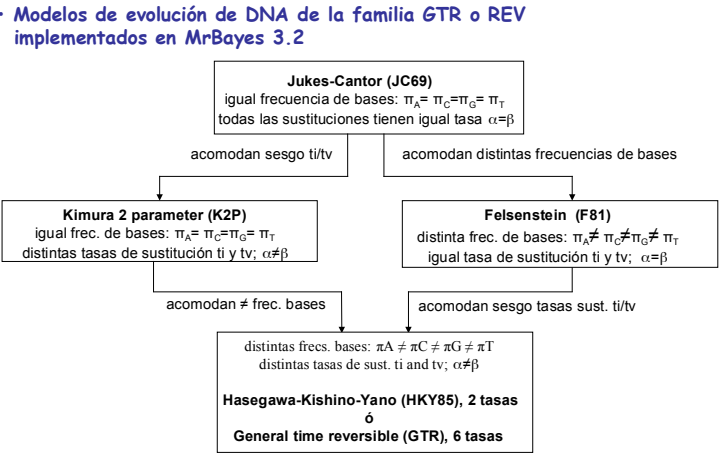
Tomado de Vinuesa et al. 2005 Mol. Phylogenet. Evol. 34:29-54

Comandos relativos a la manipulación de OTUs y caracteres

```
begin mrbayes;
set partition=by_codon;
Prset applyto=(all) ratepr=variable;
Lset applyto=(1,3,4,6) nst=2 rates=gamma Ngammacat=8;
Lset applyto=(2,5) nst=2 rates=propinv Ngammacat=8;
unlink shape=(all) pinvar=(2,5) tratio=(all) statefreq=(all) revmat=(all);
set autoclose=yes;
mcmc ngen=3000000 printfreq=5000 samplefreq=100
nchains=4 temp=0.15 savebrlens=yes;
end;
;
```

Comandos relativos a la especificación del modelo de sustitución propabilidades anteriores y parámetros de las cadenas estocásticas de Markov

Estima bayesiana de filogenias usando MrBayes 3.2 - modelos de sust. de nt (4X4)



Estima bayesiana de filogenias usando MrBayes 3.2
- modelos de sust. de nt (4X4)

• Lset

Es el comando que activa los parámetros del modelo de verosimilitud. Su uso es:

```
lset <parameter>=<option> ... <parameter>=<option>
```

- por ejemplo, "lset nst=6 rates=gamma" para especificar el modelo GTR+G

Default model settings:

Parameter	Options	Current Setting
Nucmodel	4by4/Doublet/Codon	4by4
Nst	1/2/6	1
Code	Universal/Vertmt/Mycoplasma/ Yeast/Ciliates/Metmt	Universal
Floidy	Haploid/Diploid	Diploid
Rates	Equal/Gamma/Propinv/Invgamma/Adgamma	Equal
Ngammacat	<number>	4
Nbetacat	<number>	5
Omegavar	Equal/Ny98/M3	Equal
Covarion	No/Yes	No
Coding	All/Variable/Noabsencesites/ Nopresencesites	All
Parsmodel	No/Yes	No

Estima bayesiana de filogenias usando MrBayes 3.2
- análisis de MCMC y MC³

• MCMC

Este comando inicia el análisis de MCMC para **aproximar la probabilidad posterior del árbol filogenético** (y parámetros del modelo de sustitución) mediante el muestreo de árboles de la distribución posterior. Además se puede correr un análisis de **Metropolis-coupled Markov chain Monte Carlo**, o **MCMCMC** o **MC³** en el que se corren N cadenas, N-1 de las cuales son cadenas "calentadas" por un factor $B = 1 / (1 + \text{temp } Xi)$. B es la potencia a la que se eleva la probabilidad posterior. Cuando $B = 0$, todas las topologías tienen igual probabilidad y la cadena visita los árboles libremente. $B = 1$ es la **cadena "fría"** (la distribución de interés). Se emplea generalmente **MC³** ya que se produce un mejor "mezclado" que el obtenido mediante MCMC. Después de que todas las cadenas han terminado un ciclo, se seleccionan dos cadenas al azar y se intenta cambiar los estados (la prob. del cambio viene determinada por la ecuación de Metropolis et al.). Esto permite a la cadena "saltar" valles profundos. Las cadenas secuencialmente calentadas "ven" un espacio paramétrico proporcionalmente más suave (valles menos profundos entre picos).

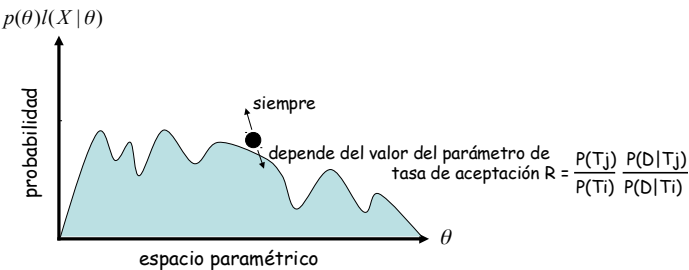
el uso correcto del comando es:

```
mcmc <parameter> = <value> ... <parameter> = <value>
```

por ejemplo: `mcmc ngen=100000 nchains=4 temp=0.2`

que ejecuta un análisis de MCMCMC con 4 cadenas y la temperatura (factor de calentamiento puesto en 0.2. Las cadenas se corren por 100000 de generaciones

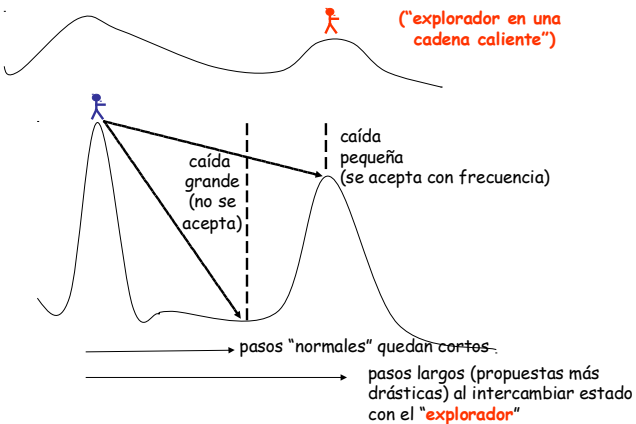
Estima bayesiana de filogenias usando MrBayes 3.2
- exploración del espacio paramétrico y de árboles mediante MCMC



- Se construye una **cadena estocástica de Markov** que tiene por su estado espacial los **parámetros del modelo estadístico** y una **distribución estacionaria** que representa la **distribución posterior** de probabilidad de los parámetros
- Para una cadena de Markov adecuadamente construida y corrida durante suficientes ciclos resulta que **la proporción de tiempo que cualquier topología particular es visitada representa una buena aproximación a la probabilidad posterior de dicho árbol**

Estima bayesiana de filogenias usando MrBayes 3.2
- el principio de Metropolis-coupled MCMC (MC³)

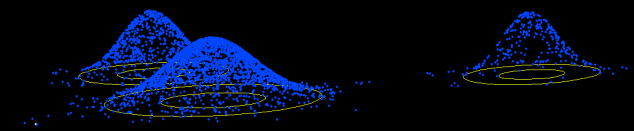
- las **cadenas calientes** hacen las veces de exploradores del espacio de parámetros para la **cadena fría**



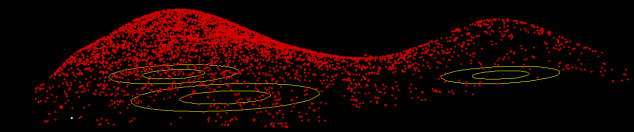
Estima bayesiana de filogenias usando MrBayes 3.2
- exploración del espacio paramétrico mediante Metropolis-coupled MCMC (MC³)

- las cadenas calientes hacen las veces de exploradores del espacio de parámetros para la cadena fría

paisaje frío: picos separados por valles profundos

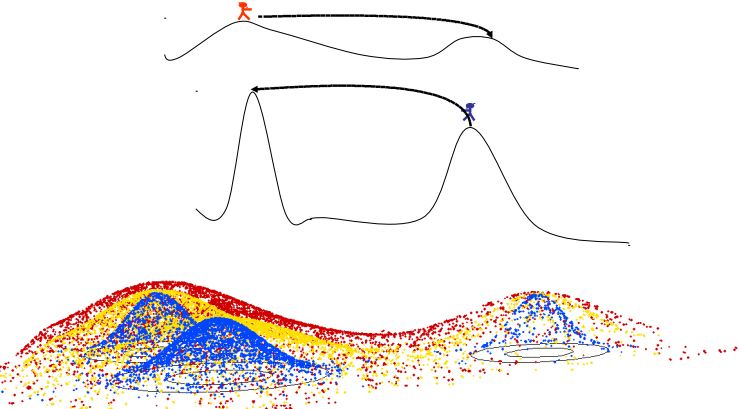


paisaje caliente: picos separados por valles poco profundos



Estima bayesiana de filogenias usando MrBayes 3.2
- el principio de Metropolis-coupled MCMC (MC³)

- las cadenas fría y caliente intercambian sus estados ("chain swapping")



MrBayes 3.2
- MCMC

- Opciones por defecto del comando MCMC (o MCMCP) en MrBayes 3.2

Parameter	Options	Current Setting
Seed	<number>	1116367232
Swapseed	<number>	1116367232
Ngen	<number>	1000000
Nruns	<number>	2
Nchains	<number>	4
Temp	<number>	0.200000
Reweight	<number>, <number>	0.00 v 0.00 ^
Swapfreq	<number>	1
Nswap	<number>	1
Samplefreq	<number>	100
Printfreq	<number>	100
Printall	Yes/No	Yes
Printmax	<number>	8
Mcmcdiagn	Yes/No	Yes
Diagnfreq	<number>	1000
Minpartfreq	<number>	0.10
Allchains	Yes/No	No
Allcomps	Yes/No	No
Relburnin	Yes/No	Yes
Burnin	<number>	0
Burninfrac	<number>	0.25
Stoprule	Yes/No	No
Stopval	<number>	0.01
Filename	<name>	temp.out.<p/t>
Startingtree	Random/User	Random
Nperts	<number>	0
Saveburnins	Yes/No	Yes
Ordertaxa	Yes/No	No

Estima bayesiana de filogenias usando MrBayes 3.2
- definición de distribuciones de probabilidad anterior (priors)

- Prset

Este comando **especifica los priors** del modelo filogenético. Recuerden que en un análisis bayesiano hay que especificar una distribución de probabilidad anterior para cada parámetro del modelo de verosimilitud (topología, long. de rama, parámetros del modelo de sustitución). Estos priors representan las ideas o hipótesis sobre la distribución de los parámetros previas a la observación de los datos. Este comando permite manipular los supuestos sobre los priors.

-En muchos casos se usan priors no informativos para que sea la función de verosimilitud la que determine de manera decisiva el resultado de un análisis

- prset applyto

En el caso de un análisis complejo con múltiples particiones, podemos definir distintos settings de priors para cada partición.

prset applyto=(1,2) statefreqs=fixed (equal, para JC y K2P)

prset applyto=(3) statefreqs=dirichlet(1,1,1,1) (para F81, HKY y GTR)

Estima bayesiana de filogenias usando MrBayes 3.2.2
- definición de distribuciones de probabilidad anterior (priors)

• En la inferencia bayesiana de filogenias se usan principalmente las siguientes distribuciones de probabilidad para definir las probabilidades (o funciones de densidad probabilística) de los parámetros del modelo filogenético, tal y como se resume en la siguiente tabla.

Parámetro	Distribución	Comentario
Topología	Uniforme discreta	define un prior no informativo
Long. de ramas	Exponencial	Prior no informativo
Var. tasas sust. entre sitios	Gamma	Forma muy flexible
Frecuencia de bases;	Dirichlet	Para proporciones del total
Tasas de ti/tv	Beta(a1,a2)	Probabilidad de 2 proporciones

Estima bayesiana de filogenias usando MrBayes 3.2
- definición de distribuciones de probabilidad anterior (priors)

• Opciones por defecto del comando Prset

Parameter	Options	Current Setting
Tratioopr	Beta/Fixed	Beta(1.0,1.0)
Revmatpr	Dirichlet/Fixed	Dirichlet(1.0,1.0,1.0,1.0,1.0,1.0)
Aamodelpr	Fixed/Mixed	Fixed(Poisson)
Aarevmatpr	Dirichlet/Fixed	Dirichlet(1.0,1.0,...)
Omegapr	Dirichlet/Fixed	Dirichlet(1.0,1.0)
Ny98omegalpr	Beta/Fixed	Beta(1.0,1.0)
Ny98omega3pr	Uniform/Exponential/Fixed	Exponential(1.0)
M3omegapr	Exponential/Fixed	Exponential
Codoncatfreqs	Dirichlet/Fixed	Dirichlet(1.0,1.0,1.0)
Statefreqpr	Dirichlet/Fixed	Dirichlet
Treeheightpr	Exponential/Gamma	Exponential(1.0)
Ratepr	Fixed/Variable=Dirichlet	Fixed
Shapepr	Uniform/Exponential/Fixed	Uniform(0.0,50.0)
Ratecorrpr	Uniform/Fixed	Uniform(-1.0,1.0)
Pinvarpr	Uniform/Fixed	Uniform(0.0,1.0)
Covswitchpr	Uniform/Exponential/Fixed	Uniform(0.0,100.0)
Symdirihyperpr	Uniform/Exponential/Fixed	Fixed(Infinity)
Topologypr	Uniform/Constraints	Uniform
Brlenspr	Unconstrained/Clock	Unconstrained:Exp(10.0)
Speciationpr	Uniform/Exponential/Fixed	Uniform(0.0,10.0)
Extinctionpr	Uniform/Exponential/Fixed	Uniform(0.0,10.0)
Sampleprob	<number>	1.00
Thetapr	Uniform/Exponential/Fixed	Uniform(0.0,10.0)

Estima bayesiana de filogenias usando MrBayes 3.2
- muestreo de la cadena estocástica de Markov

- Tomar un árbol (posición del robot) cada 100-1000 ciclos de MCMC (adelgazamiento)
Esto se controla con el el parámetro **samplefreq** del comando mcmc
- Conviene **adelgazar la cadena para reducir el nivel de autocorrelación** de las muestras
- Si se usa MC³, **sólo la cadena fría es muestreada**
- La distribución marginal de cualquier parámetro se puede obtener de esta muestra

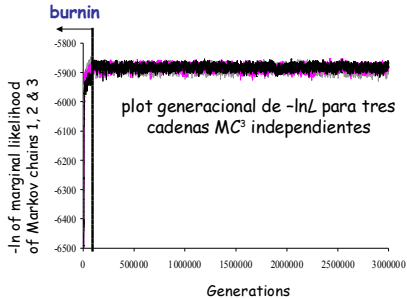
```
begin mrbayes;  
  lset nst=6 rates=invgamma Ngammacat=6;  
  mcmc ngen=3000000 printfreq=5000 samplefreq=300  
  nchains=4 temp=0.2 savebrlens=yes;  
end;  
;
```

- los comandos **sump** y **sumt** nos dan un resumen del muestreo de parámetros y árboles de un análisis

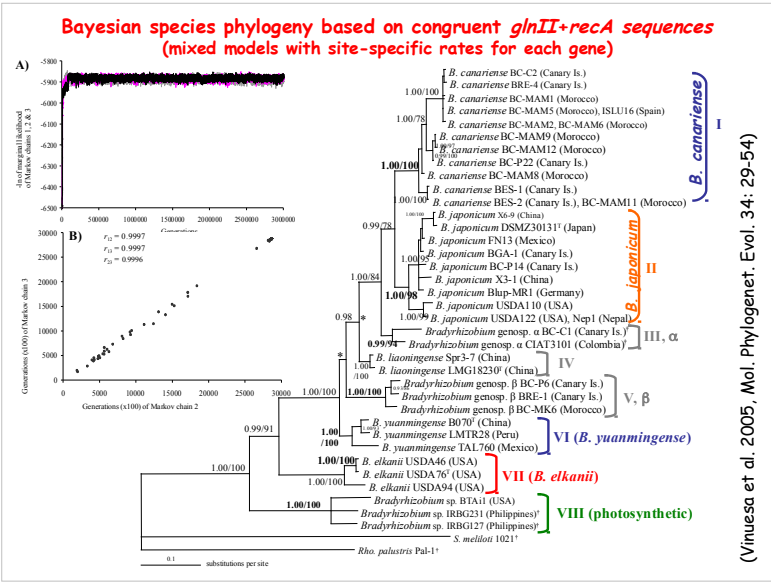
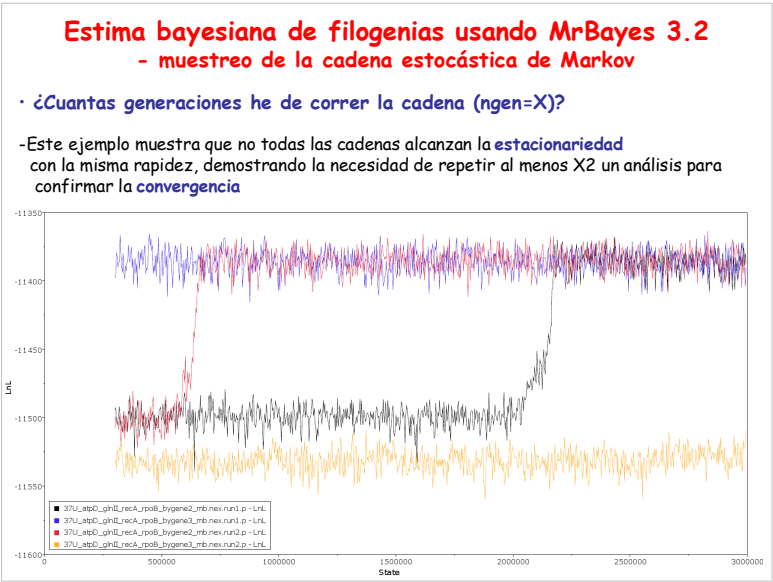
Estima bayesiana de filogenias usando MrBayes 3.2
- muestreo de la cadena estocástica de Markov

• ¿Cuántas generaciones he de correr la cadena (ngen=X)?

- básicamente hasta que se alcance la **estacionariedad** y un **mezclado adecuado** de la cadena y se hayan colectado suficientes muestras
- idealmente debemos repetir al menos X2 un análisis para confirmar la **convergencia**



- el parámetro **burnin** de los comandos mcmc o sump y sumt nos permiten determinar la cantidad de muestras (no generaciones!) a desechar (**burnin = 300**)



Métodos de reconstrucción filogenética
- la **alternativa bayesiana**

Recapitulación:

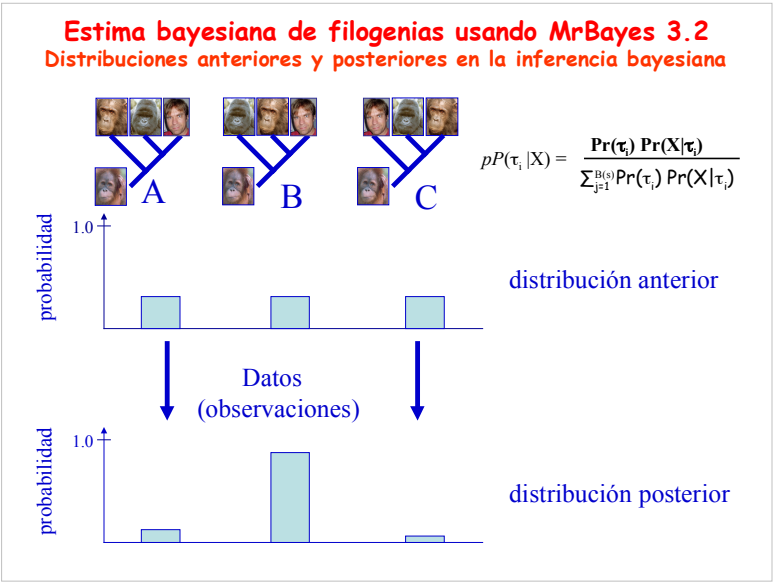
- Hemos visto que la **verosimilitud de un set de datos** de secuencias depende de los siguientes parámetros desconocidos: topología, longitudes de rama y parámetros del modelo de sustitución:
$$L_H = \Pr(D|H) = \Pr(D|\tau, \psi)$$
- El método de ML estima estos parámetros buscando los valores que maximizan la función de verosimilitud (**estima puntual**) dada una topología y un modelo estocástico de sustitución (PAUP*, PHYLIP, PAML, etc.).
- Inferencia Bayesiana** en cambio se basa en la **P posterior de la hipótesis** => calcular **la distribución de la pP conjunta** de todos los parámetros del modelo filogenético, dado un set de datos:
$$f(\tau_i|X) = \frac{f(\tau_i) f(X|\tau_i)}{\sum_{j=1}^{B(\Theta)} f(\tau_j) f(X|\tau_j)}$$
- la **pP para esta compleja distribución posterior conjunta** no se puede resolver analíticamente: hemos de **aproximarla mediante MCMC**

Reconstrucción filogenética bayesiana - MCMC

- **Markov Chain Monte Carlo (MCMC)**: toma muestras *dependientes* de la distribución de interés, de modo que el estado muestreado en el próximo intervalo (Xt+1) depende sólo del estado muestreado en el intervalo presente (Xt), y no del anterior (cadena de Markov)
- Aunque las muestras son dependientes se puede demostrar que si el número de muestras es elevado (cadena larga), la distrib. de *pP* se puede aproximar adecuadamente (Ley de los grandes números).

MCMC ha revolucionado la IB → permite tratar complejos problemas estadísticos de otro modo intratables (p. ej. grandes filogenias) y de un modo mucho más eficiente y tan preciso como ML.

Algoritmos MCMC
Habitualmente la distribución de la *pP* no puede ser analíticamente calculada porque no se puede integrar. El objetivo es aproximar la distrib. de *pP* usando un método MCMC como **Metropolis-Hastings (MH)** o **MH-Green (MHG)**



Estima bayesiana de filogenias
- referencias recomendadas

1. Alfaro ME, Zoller S, Lutzoni F (2003) Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* 20:255-266
2. Buckley TR (2002) Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51:509-523
3. Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJ (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20:248-254
4. Erixon P, Svennblad B, Britton T, Oxelman B (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52:665-673
5. Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* 4:275-284
6. Huelsenbeck JP, Larget B, Miller RE, Ronquist F (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51:673-688
7. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310-2314
8. Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53:47-67
9. Ronquist F, Huelsenbeck JP (2003) MrBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-1574
10. Ronquist, F and Deans, A. R. (2010). Bayesian phylogenetics and its influence on insect systematics. *Annu. Rev. Entomology.* 189-206
11. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61(3):539-42.