TEMA 2.1. Breve introducción a las técnicas y métodos de reconstrucción filogenética

Contacto: Virginia Valcárcel (virginia.valcarcel@uam.es)

Existen diversos métodos de análisis para estimar reconstrucciones filogenéticas a partir de datos moleculares (Tabla 1). Estos métodos pueden agruparse de diferentes maneras. En esta breve introducción al curso agruparemos los métodos de análisis en dos grandes bloques según el procedimiento seguido: (1) métodos "puramente algorítmicos" [UPGMA, *Neibourgh-Joining* (NJ)] y (2) métodos de búsquedas de árboles basados en criterios de optimización [Máxima Parsimonia (MP), Máxima Verosimilitud (ML; *Maximum Likelihood*), Inferencia Bayesiana (BI**;** *Bayesian Inference*), Mínima Evolución (ME), Mínimos Cuadrados (MC)]. Los primeros incluyen en el proceso de obtención del árbol el criterio de selección y no hacen búsquedas de árboles, por lo que no realizan de manera explícita una optimización de una función de selección con base en el criterio establecido. Los segundos realizan búsquedas de árboles sobre los que se optimiza una función según el criterio bajo el que son evaluados –mínimo número de cambios evolutivos en MP, máxima verosimilitud en ML, máxima probabilidad a posteriori en BI, mínima suma de longitudes de rama (calculadas como *ordinary least square*) en ME, o mejor ajuste entre los pares de distancias estimados y las distancias calculadas a partir del árbol, MC–.

Los métodos basados en distancias –tanto los algorítmicos (UPGMA, NJ) como los basados en búsquedas (ME)–, asumen que la distancia entre táxones es reflejo de su relación filogenética. Esta asunción es únicamente valida en casos de tasas de cambio constantes y ausencia de homoplasia, premisas ambas generalmente vulneradas. Para soslayar ambas premisas, los métodos de distancia asumen también un modelo evolutivo que permite corregir ambas cuestiones (Williams 1992). Las distancias de este modo corregidas son estimas de la distancia evolutiva real, entendida como la media de cambios que se han producido en una posición entre dos pares de secuencias a lo largo de su evolución desde su ancestro común. Así, a partir de los datos y dado un modelo evolutivo (véase tema 3.4), calculan una matriz de distancias. A partir de esa matriz de distancias construyen uno o varios árboles mediante métodos algorítmicos de construcción de árboles (UPGMA, NJ), que pueden ser posteriormente evaluados bajo criterios de optimización (ME, MC).

El método de MP realiza búsquedas de árboles usando como criterio de optimización la máxima parsimonia (Tabla 1). Así, este método optimiza la longitud del árbol calculada como el total de los cambios evolutivos (número de transformaciones de un estado de carácter a otro) necesarios para explicar un árbol a partir de los datos. De esta manera conforme al criterio de MP, el árbol más parsimonioso que conecta cuatro secuencias dos a dos es aquel que precisa del menor número de transformaciones de un estado de carácter a otro para cada una de las posiciones de la matriz. Un punto crítico de este método es la subestimación de la cantidad de cambio evolutivo. Al asumir la explicación más sencilla, la MP no tiene en cuenta la posibilidad de que para una misma secuencia y en una misma posición se hayan producido varios cambios a lo largo del tiempo (t_0 = A, t_1 = T, t_2 = A).

El método de ML en cambio intenta estimar la cantidad de cambio real de acuerdo con un modelo establecido. Este método evalúa la hipótesis (el árbol) mediante una función (verosimilitud) que maximiza la probabilidad de obtener los datos –matriz de secuencias de ADN– dado el árbol y el modelo evolutivo (véase tema 3.4). De esta forma, conforme al criterio de ML el mejor árbol de cuatro secuencias conectadas dos a dos es aquel que presenta el mayor valor de verosimilitud, independientemente del número de transformaciones de estados de carácter que necesite.

El método de BI se basa en la búsqueda de árboles que maximicen la probabilidad a posteriori de los árboles, dados los datos –matriz de secuencias de ADN– y el modelo evolutivo (véase tema 3.4). Este método utiliza el Teorema de Bayes que calcula la probabilidad a posteriori a partir de los valores de probabilidad a priori y versomilitud. La probabilidad a priori de los árboles representa la probabilidad de cada uno de los árboles posibles previa a cualquier observación (datos y modelo). Esto es, si tenemos tres especies, sólo hay tres árboles posibles que las conecten dos a dos, la probabilidad a priori de cada uno de estos tres árboles sería la misma para cada uno. En cambio, la verosimilitud de cada uno de estos tres árboles será distinta al considerar las observaciones (datos y el modelo). Así, la verosimilitud de cada árbol sería proporcional a la probabilidad de los datos –matriz de secuencias de ADN– dado el árbol y el modelo. Por último, la probabilidad a posteriori es proporcional a la probabilidad del árbol dados los datos y el modelo y se calcula combinando la probabilidad a priori y la verosimilitud.

Tabla 1.

Método	Fundamento y asunciones	Ventajas	Inconvenientes	Software
MÁXIMA PARSIMONIA	Busca y selecciona los árboles con menor cantidad de cambios evolutivos Congruencias entre los caracteres son el resultado de relaciones filogenéticas - tipo de problema: no polinomial - método de búsqueda de árboles basado en el criterio de optimización - criterio de optimización: máxima parsimonia - tipo de búsqueda: exhaustiva (*branch and bound*) o heurística - algoritmo de construcción y búsqueda: *star decomposition o stepwise addition*	- minimiza las hipótesis *ad hoc* (reversiones, paralelismos, etc.) - relativamente rápido con grandes matrices de datos - robusto si las longitudes de rama son cortas (amplio muestreo o baja divergencia) - se pueden inferir estados ancestrales	- sensible al orden de entrada de los datos - descarta información potencialmente relevante (autoapomorfías) - posible subestimación del número de sustituciones - altamente afectada por atracción de ramas largas y "zona Felsenstein" (Huelsenbeck 1998, aunque véase Hillis et al. 1996) - ausencia de un modelo evolutivo explícito (Platnick 1985) - alto riesgo de caer en mínimos locales - no asume la superposición de cambios (*multiple hits*) que son tratados como fuente de falsa homología (aunque puede compensarse vía pesado) - múltiples árboles debido al tratamiento de pasos discretos	TNT PAUP MEGA PHYLIP
MÁXIMA VEROSIMILITUD	Selecciona el árbol con mayor probabilidad de explicar los datos dado el árbol y el modelo evolutivo - tipo de problema: no polinomial - método de búsqueda de árboles basado en criterio de optimización - criterio de optimización: máxima verosimilitud - tipo de búsqueda: exhaustiva (branch and bound) o heurística - tipo de algoritmo de construcción y búsqueda: star decomposition o stepwise addition	- los modelos de sustitución nucleotídica se incluyen en el proceso de estima - poco sensible a atracción de ramas largas (Gaut & Lewis 1995) - robusto y poco sensible a la violación de sus asunciones (Huelsenbeck 1995) - método menos afectado por el error de muestreo ya que proporciona las estimas con menor varianza (Hillis et al. 1996) - permite la superposición de múltiples cambios en una misma posición (*multiple hits*)	- fuerte demanda de memoria - fallos cuando hay muchas secuencias y pocos nucleótidos (Piontkivska 2004) - riesgo de caer en mínimos locales (Salter & Pearl 2001) - sensible al modelo de substitución seleccionado	RAxML GARLI PAUP MEGA PHYLIP

Tabla 1. [continuación]

Método	Fundamento y asunciones	Ventajas	Inconvenientes	Software
INFERENCIA BAYESIANA	Selecciona los árboles con mayor probabilidad a posteriori de explicar los árboles, dados los datos y el modelo La distribución *a priori* de los parámetros especificadas - tipo de problema: no polinomial - método de búsqueda de árboles basado en criterio de optimización - criterio de optimización: máxima probabilidad a posteriori - tipo de búsqueda: estocástica - tipo de algoritmo de búsqueda: Metropolis-coupled Markov Chain Monte Carlo	- los modelos de sustitución nucleotídica se incluyen en el proceso de estima - permite la implementación de modelos evolutivos complejos - relativamente rápido con grandes matrices de datos - poco sensible a atracción de ramas largas - permite la superposición de múltiples cambios en una misma posición (*multiple hits*) - proporciona valores de apoyo a las ramas - exploran más espacio al usar MCMC - menor riesgo de caer en mínimos locales al usar la variante Metropolis-coupled de MCMC	- fuerte demanda de memoria - posible sobreestimación de los valores de apoyo de las ramas - sensible al modelo de substitución seleccionado	MrBayes BAMBE BEAST
NEIGHBOUR-JOINING	Calcula distancias entre pares de especies y devuelve el árbol con menor longitud entre pares de especies y nodos Asume modelo evolutivo - tipo de problema: polinomial - método algorítmico basado en coeficientes de distancias - algoritmo de construcción: star decomposition	- rapidez - permite la superposición de múltiples cambios en una misma posición (*multiple hits*)	- sensible al orden de entrada de los datos (Farris et al. 1996) - diferencias entre las secuencias no reflejan fielmente la distancia evolutiva - no se pueden identificar los caracteres que apoyan las ramas - pobre para conjuntos grandes de datos - pérdida de información al convertir las secuencias en distancias (Steel et al. 1988) - poco fiables las distancias calculadas cuando las secuencias son altamente divergentes	PHYLIP PAUP MEGA

Procesos de construcción y búsqueda de los árboles filogenéticos a partir de una matriz de secuencias

Los procesos de construcción y búsqueda se basan en la búsqueda de un árbol a partir de la matriz original de los datos (MP, ML, BI) o a partir de una matriz de distancias calculada a partir de la matriz original de datos (UPGMA, NJ, ME). Las búsquedas de árboles pueden ser exactas, heurísticas o estocásticas.

(1) Las búsquedas exactas prospectan todas las posibilidades garantizando encontrar los árboles óptimos. Por ello, los algortimos que realizan este tipo de búsquedas (algoritmos exhaustivos y *branch-and-bound*, Hendy & Penny 1982) consumen mucho tiempo y sólo se recomiendan para el análisis de matrices pequeñas: máximo 10 táxones para búsquedas exhaustivas y hasta un máximo de 20 para algoritmos *branch-and-bound* (Nei & Kumar, 2000).

Las búsquedas *exhaustivas* comienzan a partir de todas las combinaciones posibles de árboles de tres secuencias que se puedan construir con las secuencias incluidas en la matriz de datos original. Cada uno de estos árboles iniciales cuenta con un nodo del que surgen las tres ramas que conectan las tres secuencias de partida (Fig. 1). A cada uno de estos árboles iniciales de tres secuencias se le conecta una cuarta secuencia generando, mediante la conexión de dicha secuencia a cada una de las tres ramas del árbol inicial, tres nuevos árboles posibles de cuatro secuencias (Fig. 1). A continuación se conectaría la quinta secuencia a cada uno de los tres árboles de cuatro secuencias lo que genera a su vez cinco árboles posibles (Fig. 1). En las búsquedas exactas, el proceso seguiría de esta misma manera hasta que se obtienen todos los árboles posibles que conectan todas las secuencias incluidas en la matriz. Finalmente, se evalúa cada uno de estos árboles conforme al criterio seleccionado y se seleccionan los árboles óptimos.

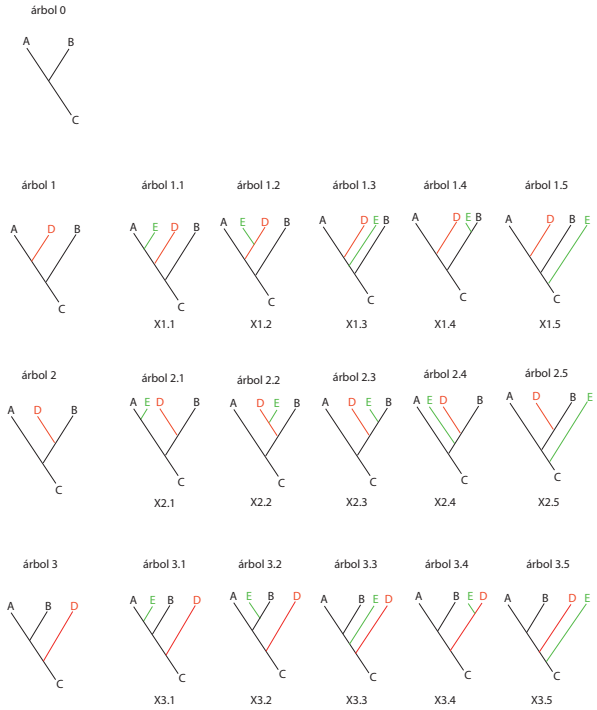

Figura 1. Esquema del proceso de construcción y búsqueda de árboles en un algoritmo exhaustivo. En el ejemplo se muestra un conjunto de datos con cinco muestras (A, B, C, D, E).

Los algoritmos *branch-and-bound* también garantizan encontrar los árboles más óptimos, pero no realizan una búsqueda completa de todos los árboles posibles. Así, estos algoritmos inician la búsqueda construyendo un árbol al azar completamente resuelto que conecte todas las secuencias incluidas en la matriz (árbol 0, Fig. 2) y lo evalúan bajo el criterio de optimización seleccionado. A continuación se vuelve a la matriz original de datos y se construye un árbol que conecte tres secuencias (árbol 1, Fig. 2) que es evaluado conforme al criterio seleccionado. Si el árbol de tres secuencias (árbol 1, Fig. 2) es igual o mejor que el árbol inicial con todas las secuencias (árbol 0, Fig. 2), entonces se le conectaría una cuarta secuencia a una de las cuatro ramas del árbol de tres secuencias. A este árbol de tres secuencias se le conecta una cuarta secuencia a una de las tres ramas posibles generando un árbol de cuatro secuencias (árbol 1.1, Fig. 2) que es evaluado conforme al criterio seleccionado. Si el árbol de cuatro secuencias (árbol 1.1, Fig. 2) es igual o mejor que el árbol inicial con todas las secuencias (árbol 0, Fig. 2), entonces se le conectaría una quinta secuencia a una de las cinco ramas del árbol de cuatro secuencias. Si, por el contrario, al evaluar el árbol de cuatro secuencias (árbol 1.1, Fig. 2) fuese peor que el árbol inicial (árbol 0, Fig. 2), entonces se rechaza este árbol de cuatro secuencias (árbol 1.1, Fig. 2) y todos sus posibles árboles derivados (árbol 1.1.1-1.1.5, Fig. 2) y vuelven al árbol de tres secuencias (árbol 1, Fig. 2) al que le conectarían la misma cuarta secuencia a otra rama diferente y generando un nuevo árbol de cuatro secuencias (árbol 1.2, Fig. 2) y repitiendo el mismo procedimiento (Fig. 2).

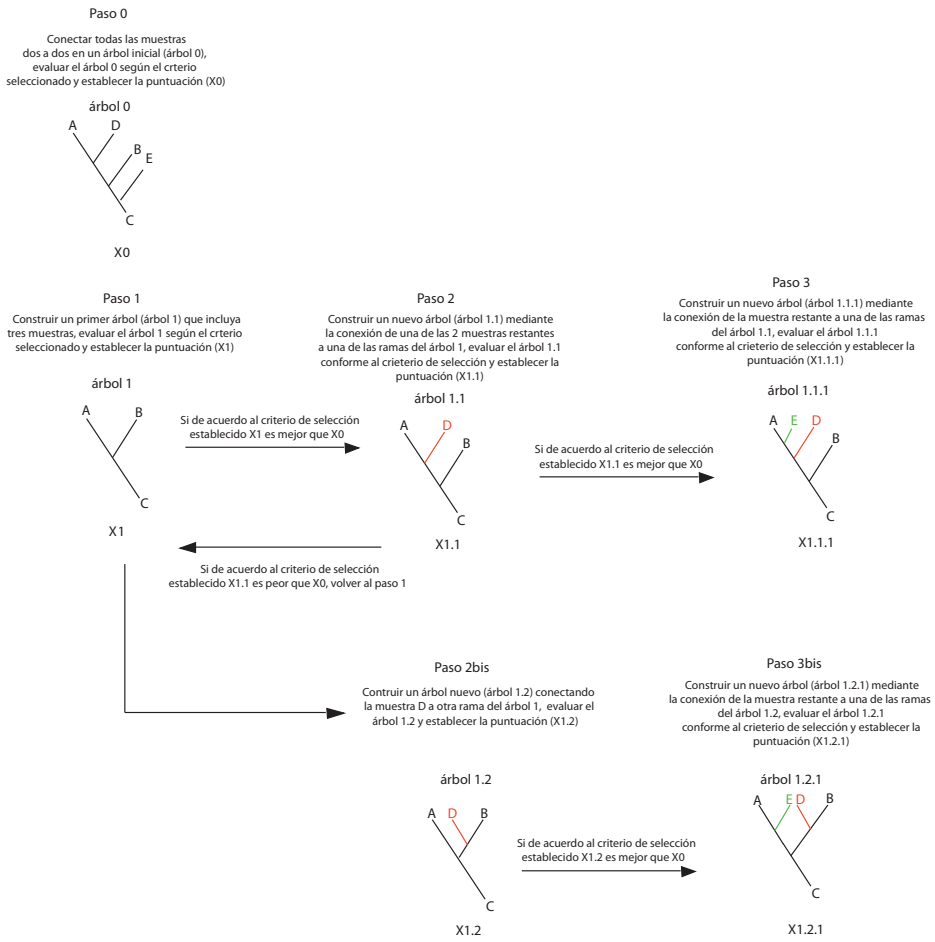

Figura 2. Esquema del proceso de construcción y búsqueda de árboles en un algoritmo *branch-and-bound*. En el ejemplo se muestra un conjunto de datos con cinco muestras (A, B, C, D, E).

(2) Las búsquedas heurísticas (algoritmos *hill-climbing strategies*; *stepwise addition, star decomposition*) prospectan un espacio limitado del universo conforme al criterio seleccionado. Fundamentalmente se utilizan dos tipos de algoritmos para construir los árboles: *star decomposition* o *stepwise addition*.

El algoritmo *star decomposition* (Fig. 3) construye un árbol inicial en forma de estrella que incluye todas las secuencias de la matriz original unidas por un único nodo (Paso 0), a continuación se construyen todos los árboles posibles creando otro nodo que conecte dos secuencias (Paso 1). Se evalúan todos los posibles árboles construidos en el paso 1 bajo el criterio de selección utilizado y se elige el mejor (Paso 2). A partir del árbol seleccionado en el paso 2, se construyen todos los posibles árboles conectando otras dos secuencias, se evalúan y se selecciona el mejor (paso 3). Este proceso se repite hasta que se obtiene un árbol que conecte todas las secuencias dos a dos.

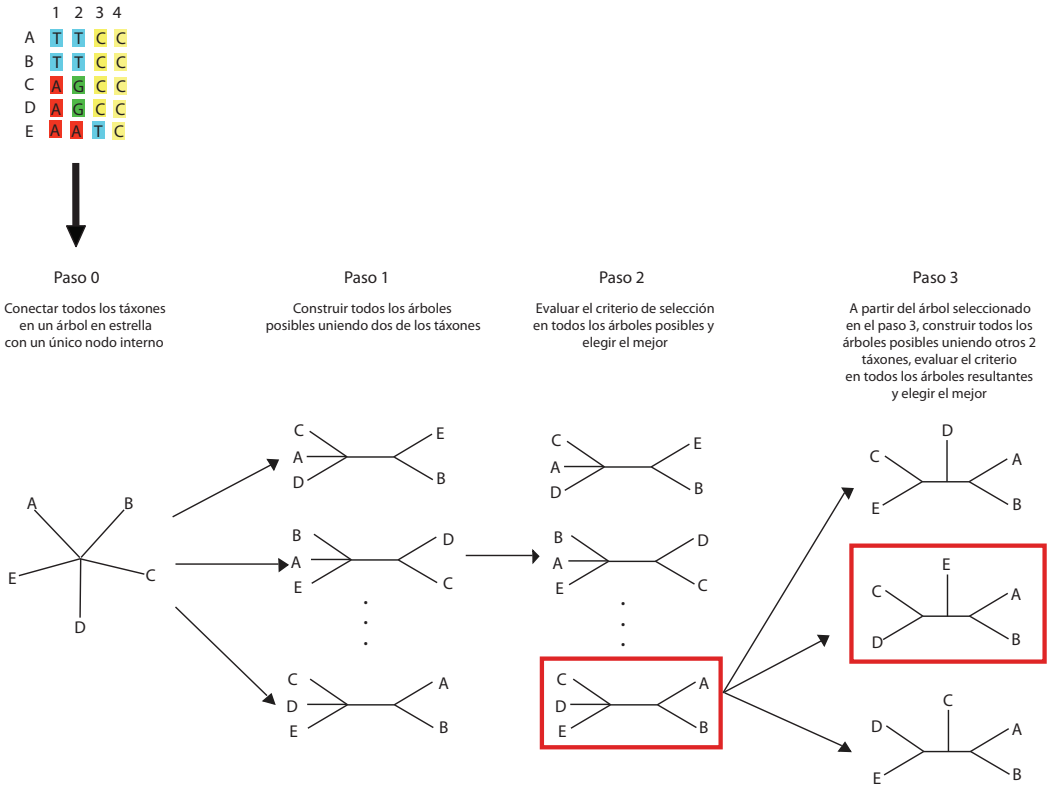

Figura 3. Esquema del proceso de construcción y búsqueda heurísticas de árboles mediante un algoritmo *star decomposition*. En el ejemplo se muestra un conjunto de datos con cinco muestras (A, B, C, D, E).

La alternativa de *stepwise addition* (Fig. 4) construye un árbol inicial que incluye tres secuencias al que se van a ir añadiendo las restantes secuencias una a una. La selección del primer árbol de tres secuencias así como el modo de adición de las restantes secuencias puede realizarse siguiendo distintos criterios —"*as is*" el primer árbol se construye con las tres primeras secuencias de la matriz y la adición de las siguientes secuencias se hace por orden de posición en la matriz; "*random"* las tres primeras secuencias así como la posterior adición de secuencias se hace a partir de una lista de números aleatorios; *"closest"*, de todos los posibles árboles con tres secuencias se selecciona el que presenta menor longitud y las secuencias se van añadiendo siguiendo este mismo criterio, etc.—.

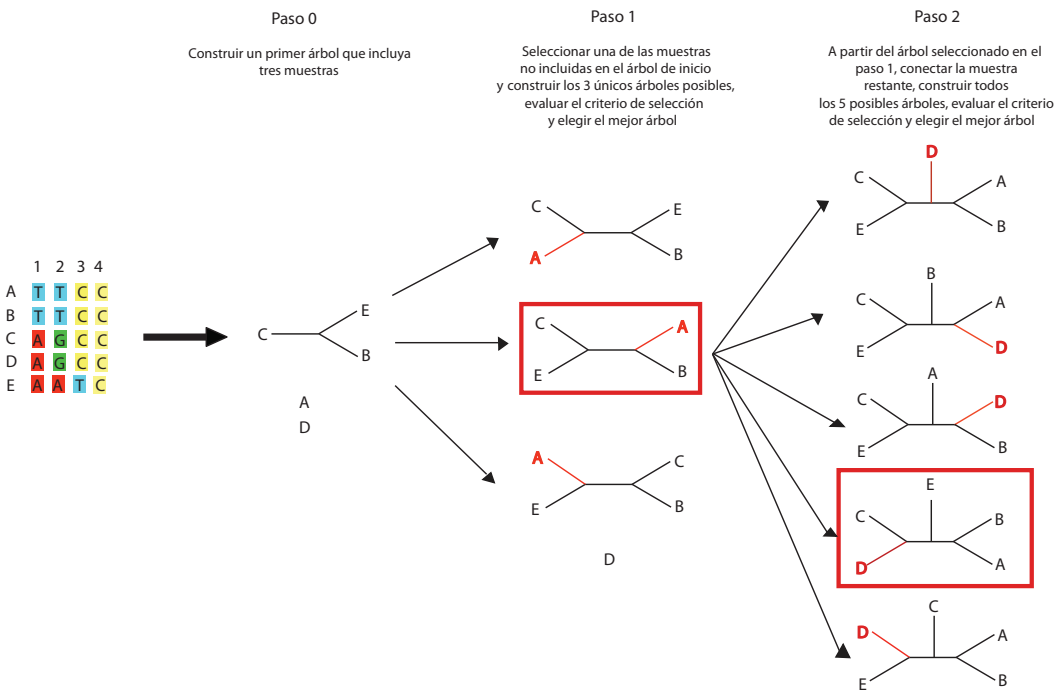

Figura 4. Esquema del proceso de construcción y búsqueda heurísticas de árboles mediante un algoritmo *stepwise addition*. En el ejemplo se muestra un conjunto de datos con cinco muestras (A, B, C, D, E).

Dado que las búsquedas heurísticas no prospectan todo el universo de árboles posibles, estos algoritmos mejoran sus búsquedas garantizando encontrar árboles óptimos mediante procesos de intercambio de ramas "*branch swapping*" (*tree bisection reconection* (TBR), *nearest-neighbour interchange* (NNI), *subtree pruning and regrating* (SPR); Hillis et al. 1996). En cada búsqueda el árbol más óptimo que incluya todos los táxones es mejorado mediante estos procesos de intercambio de ramas. Los árboles construidos en cada paso son evaluados y aceptados o rechazados en función del criterio de optimización utilizado (menor número de pasos en MP, mayor valor de verosimilitud en ML).

(3) Las búsquedas estocásticas que se emplean en filogenia molecular muestrean mediante la técnica de *Markov Chain Monte Carlo* (MCMC) a partir de un árbol inicial construido estocásticamente sobre el que se realizan modificaciones al azar que alteran no sólo la topología del árbol, sino también la longitud de las ramas o los parámetros del modelo de sustitución (véase tema 3.4). El árbol modificado es evaluado, mediante el cálculo de la probabilidad *a posteriori* (PP), y aceptado o rechazado conforme a la probabilidad descrita por el algoritmo de Metropolis & Hastings. Este algoritmo aumentan la probabilidad de encontrar los árboles óptimos mediante la posibilidad de hacer pequeños pasos para atrás. Además, al realizar varias