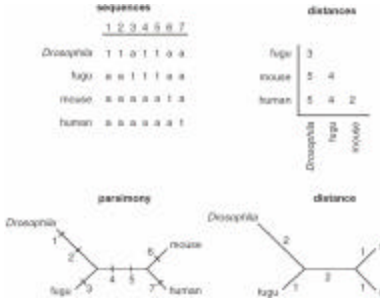
## Criterios de optimización – Máxima parsimonia

• Los métodos de distancia primero convierten los alineamientos de secuencias en una matriz de distancias genéticas en base al modelo evolutivo seleccionado, la cual es usada por el método de reconstrucción para calcular el árbol (LS y ME; UPGMA y NJ)

• Los métodos discretos basados en crit. de opt. (MP y ML) consideran cada sitio del alineamiento (o una función probabilística para cada sitio) directamente




• Un set de 4 seqs. y la matriz de distancias correspondiente

• Un árbol de parsimonia y uno de distancias (ME) para el mismo set de datos produce topologías y longitudes de ramas idénticas

• La diferencia radica en que el árbol de parsimonia identifica qué sitio del alineamiento contribuye cada paso mutacional en la longitud de cada rama

## Criterios de optimización – Máxima parsimonia

**máxima parsimonia**: involucra la **identificación de la(s) topología(s) con la menor longitud total del árbol, es decir, que requiere(n) el menor número de cambios evolutivos (transformaciones en estados de caracter) para explicar las diferencias observadas entre OTUs** (Kluge & Farris 1969; Farris, 1970; Fitch, 1971)

• Justificación filosófica - La "**cuchilla de Ockham**": la mejor hipótesis es aquella que requiere el menor número de suposiciones ("elimínese todo lo prescindible"), es decir, favorecemos a la hipótesis más simple

• Se ha sugerido en un marco conceptual Popperiano que la parsimonia es el único método consistente con un marco hipotético-deductivo de contraste de hipótesis

• Estudios recientes muestran en cambio que la relación entre MP y simplicidad no es obvia: se ha demostrado que la ML bajo modelos muy parametrizados que asignan un parámetro individual para cada caracter (posición) y rama del árbol, se hace equivalente a la MP. ¿Indica esto una clara relación entre MP y simplicidad?

(Tuffley & Steel 1997. Bull. Math. Biol. 59:581-607; Queiroz & Poe 2001. Syst. Biol. 50:305-321)

## Criterios de optimización – Máxima parsimonia

• El modelo de MP se justifica en filogenética dado que

1) se asume que los cambios de estado de caracter (mutaciones) son poco frecuentes y

2) no se puede conocer con exactitud el camino evolutivo de dichos cambios, por lo que se busca **maximizar la similitud evolutiva** que se puede explicar como **homóloga** (por ancestría compartida). De esta manera se busca de **minimizar la homoplasia** (similitud no heredada directamente del ancestro), ya que las **hipótesis** de homoplasia (convergencia, evolución paralela ...) pueden ser juzgadas como intentos ***ad hoc*** de explicar porqué determinados datos no encajan en una hipótesis evolutiva (árbol filogenético) particular

## Criterios de optimización – Máxima parsimonia

• Cualquier discusión sobre métodos de MP debe **distinguir entre el criterio de optimización** (árbol de longitud mínima bajo una serie de restricciones impuestas a los cambios posibles entre estados de caracter) **y el algoritmo** empleado para **para buscar estos árboles óptimos** en el espacio de topologías posibles.

• Los algoritmos de búsqueda se van mejorando con el tiempo y algunos pueden quedar obsoletos, mientras que el criterio de MP está claramente establecido en ciencia desde hace mucho tiempo y ha perdurado en filogenética desde su implementación en esta disciplina por Edwards y Cavalli-Sforza en 1963 (ver aspectos históricos tratados en el tema I).

• Por lo tanto **vamos a tratar dos puntos en este tema**:

1.- El **criterio de optimización de máxima parsimonia (MP)**

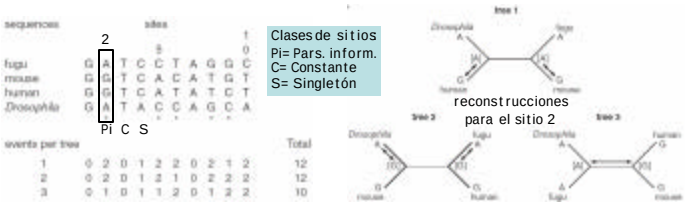
2.- las **estrategias de búsqueda** exhaustivas y heurísticas empleadas en la actualidad por paquetes de inferencia filogenética tales como Phylip y PAUP*.

## Criterios de optimización – Máxima parsimonia

- El **árbol de máxima parsimonia** representa a la hipótesis evolutiva consistente con el camino evolutivo más corto que explica o conduce a los caracteres observados
- Para sets de datos complejos y con homoplasias se encuentra generalmente más de una topología de igual longitud (número de cambios en estado de caracter); estos **árboles** son **igualmente parsimoniosos y tienen el mismo score (*L*)**
- Se han desarrollado diversos métodos de MP para inferencia filogenética con el fin de poder analizar diferentes tipos de datos:
  - **Parsimonia de Wagner**: trabaja sobre **caracteres multiestado ordenados** A <-> B <-> C (cambio de A a C require 2 pasos)
  - **Parsimonia (estándar) de Fitch**: trabaja sobre **caracteres multiestado desordenados** (nt)
  - **Parsiminia (ponderada) generalizada**: usa una **matriz de pasos** para dar mayor peso a *tv* que a *ti*
  - **Parsimonia de Dollo**: se emplea cuando existe **asimetría en la probabilidad de evolución de estados de caracter** (p. ej. caracteres de sitios de restricción: la pérdida es más probable que la ganancia paralela de un sitio de restricción)

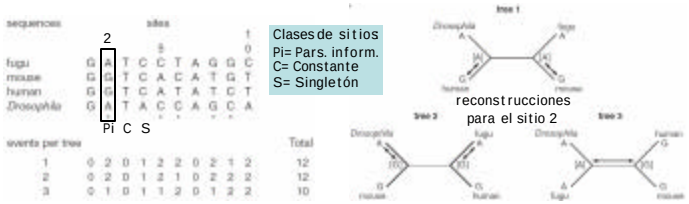
## Máxima parsimonia estándar (de Fitch)

- clasificación de caracteres:
  - sitios (C) **invariantes** o constantes
  - sitios (V) **variables**: (**informativos** (Pi) vs. **no informativos o Singletones** (S)




- Un sitio es *Pi* sólo si existen al menos 2 est. car. (nts) y cada uno de ellos es compartido al menos por 2 de la secuencias a analizar (marcados con *). Sólo así son filogenet. informat.
- Para encontrar el árbol de MP se identifican primero los *Pi*. Para cada topología posible se calcula el número min. de sust. de cada *Pi*. Sobre la(s) topología(s) más parsimoniosas se mapean finalmente todas las sustituciones (informativas o no) para calcular las long. de rama
- Nótese que los resíduos en los nodos internos de cada árbol representan sólo una de las diversas reconstrucciones posibles. Por ej. podemos sutituír las [As] por [ Gs] para el sitio 2 en el árbol 1 y no cambia su puntuación; si ponemos una [T] ó [C] implicaría 4 sust., etc.

## Máxima parsimonia estándar (de Fitch)




- En nuestro caso la **topología # 3 es la más parsimoniosa**, puesto que demanda 2 pasos menos que las topologías #1 y #2
- Para cada sitio var. del alineamiento el objetivo es reconstruir su evolución bajo la constricción de invocar el número mínimo de pasos evolutivos. El número total de cambios evolutivos sobre un árbol (**longitud en pasos evolutivos del árbol**) es simplemente la suma de cambios de estados de caracter (p. ej. mutaciones) en cada sitio var. de la matriz o alineamiento

$$L = \sum_{i=1}^{k} l_i \qquad K = \text{no. de sitios; } l = \text{longitud de cada sitio}$$

## Ejercicio – MP estándar (FITCH)

Para el siguiente alineamiento:

A) haz una clasificación de caracteres según el criterio de máxima parsimonia estándar (Fitch Parsimony)

**1. Alineamiento:** No. sitios : 15; OTUs (taxa) = 4

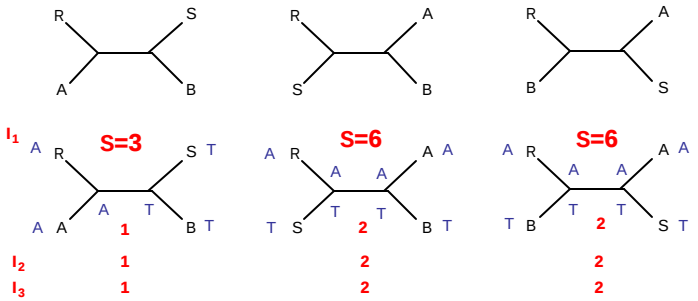
| | | |
|---|---|---|
| *Rhizobium* | GGA GGG AGG AGG CCT | C = 6 |
| *Agrobacterium* | GGC GGG AGG AGG CCT | V = 9 |
| *Sinorhizobium* | GGG GGA AGG TGT CCG | S = 6 |
| *Bradyrhizobium* | GGT CGT AGC TGT GTG | Pi = 3 |
| | CCS SCS CCS ICI SSI | S 15 |

Caracteres
- Constantes (C)
- Variables
  - Singletones (S)
  - Informativos (I)

## Ejercicio – MP estándar (FITCH)

C) Dibuja las toplogías posibles para los 4 OTUs, indica cual es la topología más parsimoniosa de ellas y calcula la longitud de la misma

| | |
|---|---|
| *Rhizobium* | GGA GGG AGG AGG CCT |
| *Agrobacterium* | GGC GGG AGG AGG CCT |
| *Sinorhizobium* | GGG GGA AGG TGT CCG |
| *Bradyrhizobium* | GGT CGT AGC TGT GTG |

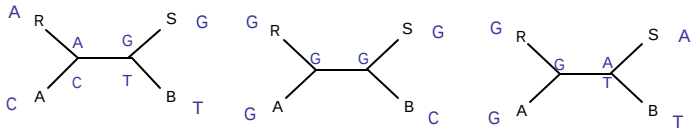



## Ejercicio – MP estándar (FITCH)

C) Dibuja las toplogías posibles para los 4 OTUs e indica cual es la topología más parsimoniosa de ellas y calcula la longitud de la misma

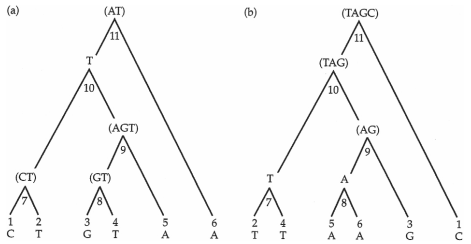
| | |
|---|---|
| *Rhizobium* | GGA GGG AGG AGG CCT |
| *Agrobacterium* | GGC GGG AGG AGG CCT |
| *Sinorhizobium* | GGG GGA AGG TGT CCG |
| *Bradyrhizobium* | GGT CGT AGC TGT GTG |
| | CCS SCS CCS ICI SSI |

**3 1 2 1 1 1 111 S = 12 =TL**



## Máxima parsimonia estándar (de Fitch)

- **Reconstrucción de estados de caracter ancestrales**




- El set en un nodo interno es la intersección ( ∩ ) de los dos sets en los dos nodos inmediatamente descendientes siempre que la intersección no esté vacía; de ser así, es la unión (U)
- Nótese que la inferencia de los caracteres ancestrales es dependiente de la topología
- Cuando se requiere una U para definir el set nodal, tuvo que haber acontecido una sustitución en dicho sitio durante su evolución. Por tanto el número de Us = no. mínimo de sust. que se requieren para explicar el estado de caracter de un nodo descendiente de otro ancestral
- El no. de sust. en un sitio no Pi es igual al no. de nts diferentes en dicho sitio -1

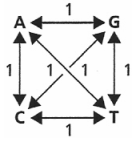
## Máxima parsimonia generalizada (ponderada)

- Para compensar la pérdida de señal filogenética que se produce más rápidamente para *ti* que *tv*, se puede dar mayor peso a estas últimas, ya que suelen ser un mejor indicador filogenético. En el caso más extremo, a las *tis* se les da un peso = 0, habándose entonces de "**transversion parsimony**".

Modelo de sustitución

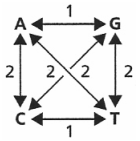
**Matriz de pasos** (ponderación)

MP no ponderada



| De \ Hacia | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 1 | 1 | 1 |
| C | 1 | 0 | 1 | 1 |
| G | 1 | 1 | 0 | 1 |
| T | 1 | 1 | 1 | 0 |

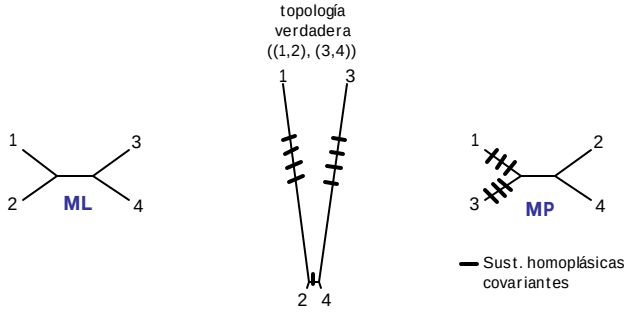
MP ponderada



| De \ Hacia | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 2 | 1 | 2 |
| C | 2 | 0 | 2 | 1 |
| G | 1 | 2 | 0 | 2 |
| T | 2 | 1 | 2 | 0 |

## Máxima parsimonia - objeciones

• **Inconsistencia** bajo ciertos modelos de evolución: **atracción de ramas largas** ("**zona de Felsenstein**": Felsenstein 1978. Syst. Zool. 27:401-410)

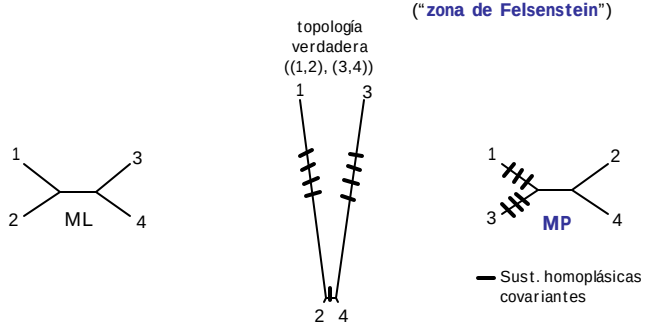



• **MI es estadísticamente consistente**: converge con la topología verdadera con mayor frecuencia a medida que incrementa el no. de datos (sitios)

• **MP es estadísticamente inconsistente**: converge con la topología incorrecta con mayor frecuencia a medida que incrementa el no. de datos (sitios)

## Máxima parsimonia - objeciones

• **Inconsistencia** bajo ciertos modelos de evolución: **atracción de ramas largas** ("**zona de Felsenstein**")



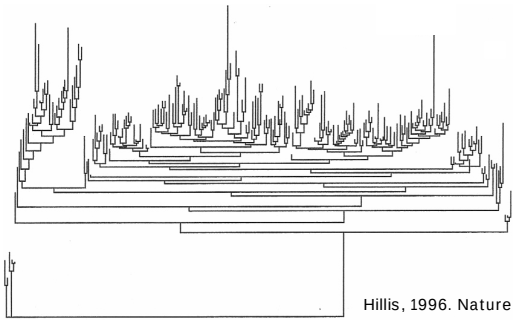

• La MP requiere que existan más sitios soportando la topología ((1,2), (3,4)) que ((1,3), (2,4)) para que la primera sea la recuperada en un análisis

• Si la rama central es muy corta, OTUs 1 y 3 pueden adquirir las mismas sustituciones convergentes (homoplásicas) por azar, las cuales pueden llegar a pesar más que las pocas sust. homólogas que se acumulan en la rama interna

## Máxima parsimonia - objeciones

• El efecto de atracción de ramas largas se encuentra en datos verdaderos cuando:
- tenemos pocas secuencias (cuartetos) y algunas de ellas presentan tasas de sustitución mucho mayor que otras o 2) éstas son muy divergentes

• **La consistencia de la MP incrementa drásticamente cuando los árboles tienen muchas ramas (OTUs) que "rompen" las ramas largas**. Esto ha sido demostrado mediante estudios de simulación de secuencias de distinta long. a lo largo de filogenias como la mostrada

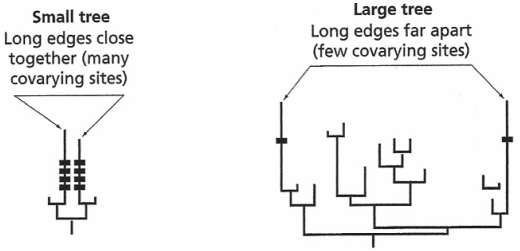


Hillis, 1996. Nature 383:130-131

## Máxima parsimonia - objeciones

• Más que la presencia de ramas largas lo que afecta a la consistencia de la MP es que existan sustituciones convergentes (covariantes) a lo largo de las ramas largas

• La probabilidad de que existan dichas sustituciones homoplásicas covariantes decrece mucho si las ramas largas están muy separadas en la topología, dado que sus caracteres ancestrales por lo tanto también son muy distintos. Lo contrario sucede para ramas largas próximas sobre topologías con pocos OTUs



## ¿Es el ML siempre consistente?

**NO** ML tiendea ser inconsistente cuando el modelo seleccionadoes incorrecto (presenta muy mal ajuste). La presencia de ramas largas puedeser un síntoma de un modelo con pobre ajuste

- fuertevariación de tasas de sustitución entre sitios

Gaut & Lewis 1995. Mol. Biol. Evol. 12:152-162

- cuando los sitios no evolucionan independientemente

Schöniger & von Haeseler 1995. Syst. Biol. 44:533-547

• En general ML es bastante robusto a violaciones de los supuestos

- cada vez se tiene más claro qué factores evolutivos son los relevantes en distintos tipos de secuencias y se continúan desarrollando más y mejores modelos que consideran dichos factores para hacer la reconstrucción filogenética

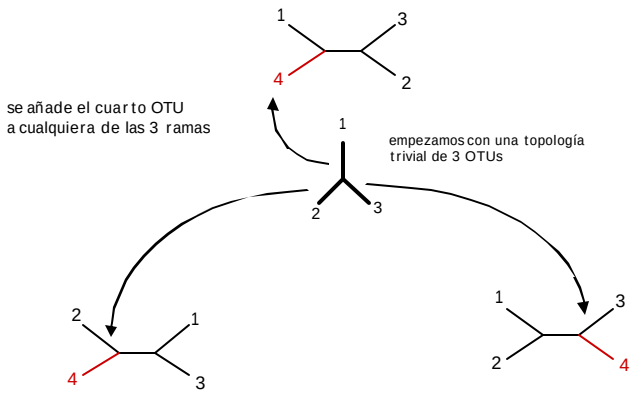
## Métodos de búsqueda de árboles

• Pasos lógicos de los métodos filogenéticos basados en criterios de optimización (MP, ML ...)

1. definir el criterio de optimización (descrito formalmente en una **función objetiva**)
2. Construir un árbol de partida que contenga todos los OTUs
3. Emplar **algoritmos de búsqueda** que tratan de encontrar árboles mejores bajo el criterio de optimizaci'on escogidoque el árbol actual o de partida.

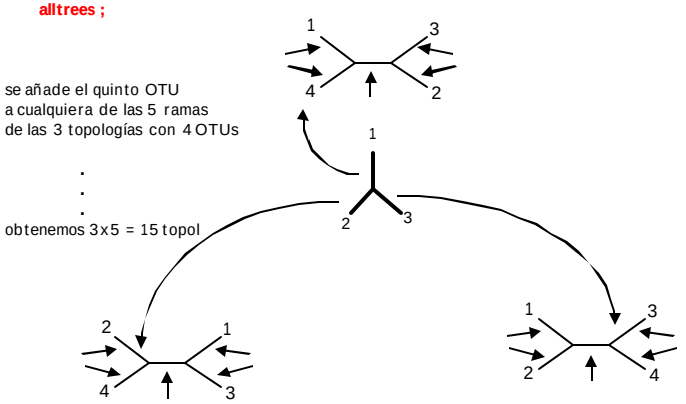
| 1. Criterios de optimización | 2. Estrategias de búsqueda | |
|---|---|---|
| Máxima parsimonia | Enumeración exhaustiva (n = 12) (exhaustive enumeration) | **Métodos exactos**: garantizan encontrar la topología óptima |
| Máxima verosimilitud | Ramificación y límite (n = 25) (branch-and-bound) | |
| Evolución Mínima | Decomposición en estrella (star decomposition) | **Métodos heurísticos**: no garantizan encontrar la topología óptima |
| Mínimos cuadrados | Adición secuencial (stepwise addition) | |
| | (Inter-)cambio de rama (branch swapping) | |

## Métodos de búsqueda de árboles -enumeración exhaustiva (n = 12)

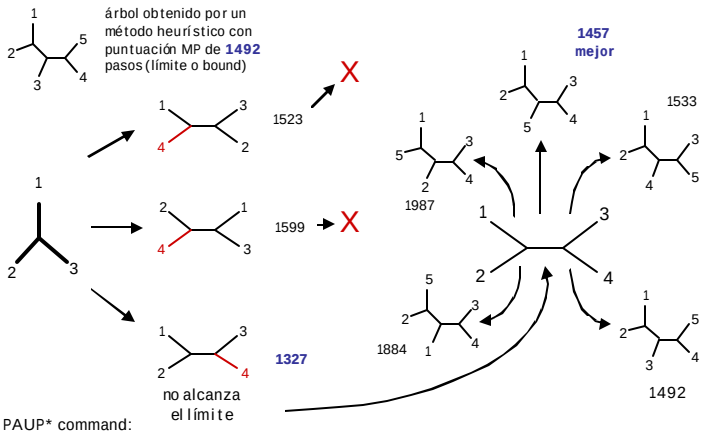



## Métodos exactos de búsqueda de árboles -enumeración exhaustiva (n = 12)

PAUP* command:
**alltrees ;**



## Métodos exactos de búsqueda de árboles - "branch and bound" (n = 25)




• PAUP* command: **bandb;**

• Al igual que la búsqueda exhaustiva, garantiza encontrar el árbol óptimo

## Métodos de búsqueda de árboles

**I.- el problema del número de topologías**

El número de topologías posibles incrementa factorialmentecon cada nuevo taxon o secuencia que se añade al análisis

No. de árboles no enraizados = $(2n-5)!/2^{n-3}(n-3)$

No. de árboles enraizados = $(2n-3)!/2^{n-2}(n-2)$

| Taxa | árboles no enraiz*. | árb. enraiz. |
|---|---|---|
| 4 | 3 | 15 |
| 8 | 10,395 | 135,135 |
| 10 | 2,027,025 | 34,459,425 |
| 22 | $3x10^{23}$ | ... |
| 50 | $3x10^{74}$ | ... |

*por ej. para sólo 15 OTUs tenemos 213,458,046,676,875 topologías

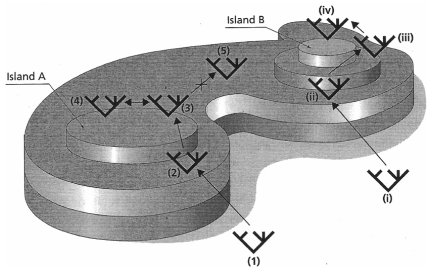
- ¡ si pudiésemos evaluar $1x10^6$ topol./seg. necesitaríamos 6 años y 9 meses para completar la búsqueda! El no. de Avogadro es ~ 6 $x10^{23}$ (átomos/mol). Según la teor. de la relatividad de la estructura del universo de Einstein, existen $10^{80}$ átomos en el universo ...

Por tanto se requieren de **estrategias heurísticas de búsqueda** árboles cuando se emplean métodos basados en criterios de optimización y $n > \sim 25$

## Métodos heurísticos de búsqueda de árboles - islas de árboles

• En la mayor parte de los casos se emplean métodos heurísticos;
- éstos comienzan con un árbol (aleatorio, NJ o de adición secuencial) para realizar intercambios de ramas (**branch swappig**) sobre esta topología inicial con el propósito de encontrar topologías de mejor puntuación (según la func. de objetividad) que la de partida

• estos métodos heurísticos no garantizan encontrar la topología óptima pero trabajan muy bien cuando se comparan con sets de datos de = 25 secs. analizados mediante B&B

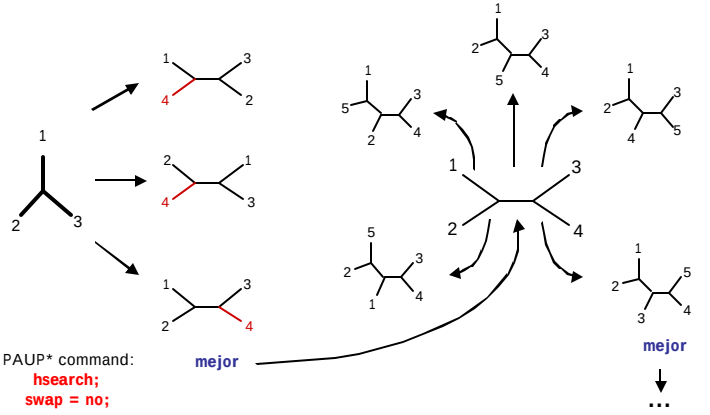



• El espacio de árboles puede visualizarse como un paisaje con colinas de diversas alturas; cada pico representa un máximo local de score o puntuación (**isla de árboles**)

• Es recomendable hacer múltiples búsqudeas heuríst. comenzando cada una desde una topología distinta para minimizar el riesgo de obtener un árbol ubicado en una isla topológica subóptima

## Métodos heurísticos de búsqueda de árboles - adición secuencial (aleatorizada)

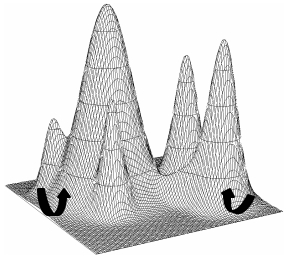
Este método se usa con frecuencia para generar distintos "**árboles semilla**" a partir de los cuales comenzar búsquedas heurísticas, partiendo de "distintos puntos del espacio de árboles



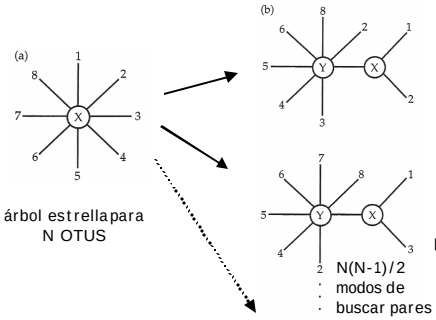

PAUP* command:
**hsearch;**
**swap = no;**

## Métodos heurísticos de búsqueda de árboles - adición secuencial (aleatorizada)

- El órden en el que se añaden los OTUs puede cambiar los resultados
- Por ello suele repetirse varias veces, añadiendo OTUs en cada ciclo de manera aleatorizada
- Sirve por lo tanto **para iniciar distintas búsquedas heurísticas** partiendo de topologías potencialmente diferences para una eficiente exploración del espacio de topologías (pero **no adecuado como hipótesis evolutiva en sí misma**)



## Métodos heurísticos de búsqueda de árboles - decomposición de estrella



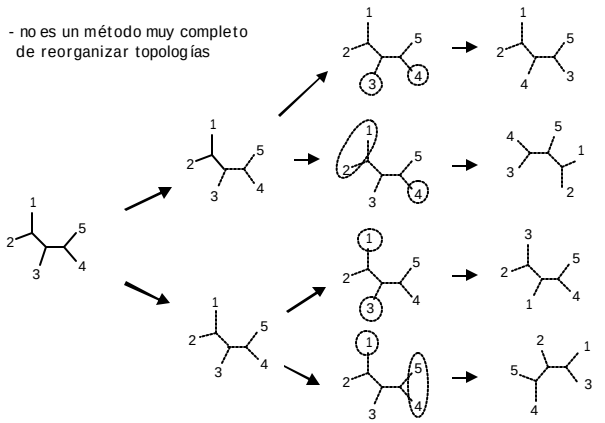

PAUP* command: **stardecomp;**

- NJ usa este método junto al criterio de evolución mínima
- una vez que 2 OTUs han sido unidos ya no pueden ser desacoplados más adelante; en esto difiere del algoritmo de adición secuencial
- sensible al orden en que se van uniendo los OTUs; problema incrementa con el no. de OTUs
- no debe ser por tanto usado como método de búsqueda definitivo
- buena estrategia para producir árboles iniciales que sean mejorados mediante otras estrategias heurísticas

## Métodos heurísticos de búsqueda de árboles - intercambio de ramas (branch swapping)

- **Intercambio entre vecinos más próximos (Nearest Neighbor Interchange, NNI)**

- no es un método muy completo de reorganizar topologías



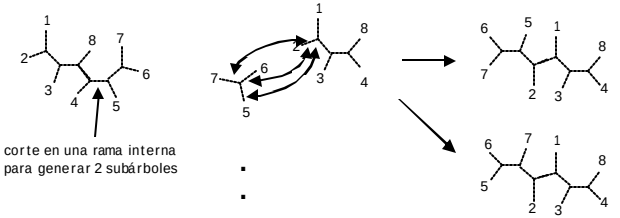
PAUP* cmmd : **hsearch swap=nni start=stepwise addseq=random;**

## Métodos heurísticos de búsqueda de árboles - intercambio de ramas (branch swapping)

- **Bisección-reconexión de árboles (Tree Bisection-Reconection, TBR)**

-Este método evalúa muchas más topols. que el NNI

se reconectan los dos subárboles en todas las posiciones posibles (ej: 3x5 =15 subarreglos en nuestro ejemplo



corte en una rama interna para generar 2 subárboles

se repite esta operación para reconectar el subárbol chico en las ramas terminales 1, 8, 4 y 3 del subárbol grande

PAUP* cmmd : **hsearch swap=tbr start=stepwise addseq=random;**

## Métodos heurísticos de búsqueda de árboles - estrategias de búsqueda para muchos OTUs *n* > 25

- Generalmente se combinan distintos tipos de búsquedas
  - es frecuente comenzar con (una o varias) topología generada por adición secuencial aleatorizada y mejorarla mediante un TBR
  - a veces se intercala una búsqueda NNI
- Una vez encontrada una topología mejor en una ronda de "branch-swapping", ésta sirve como topología de partida para nuevos rearreglos. Por tanto es conveniente partir de árboles "buenos" para minimizar el número de ciclos de branch swapping que se han de realizar para encontrar la topología localmente óptima. Las topologías generadas por adición secuencial aleatorizada son generalmente suficientemente "buenas" para iniciar los ciclos de branch-swapping que permiten una exploración eficiente del espacio de topologías.