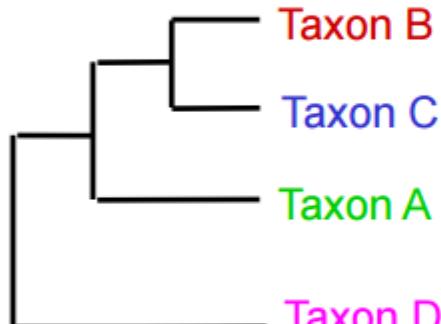


Filogenia: Tipos de Arboles

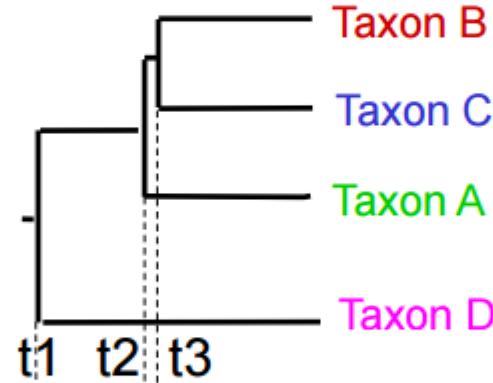
Three types of trees

Cladogram



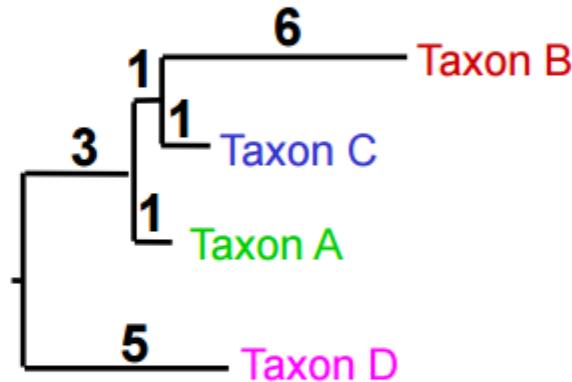
Topology only

Chronogram



Topology + Divergence times

Phylogram

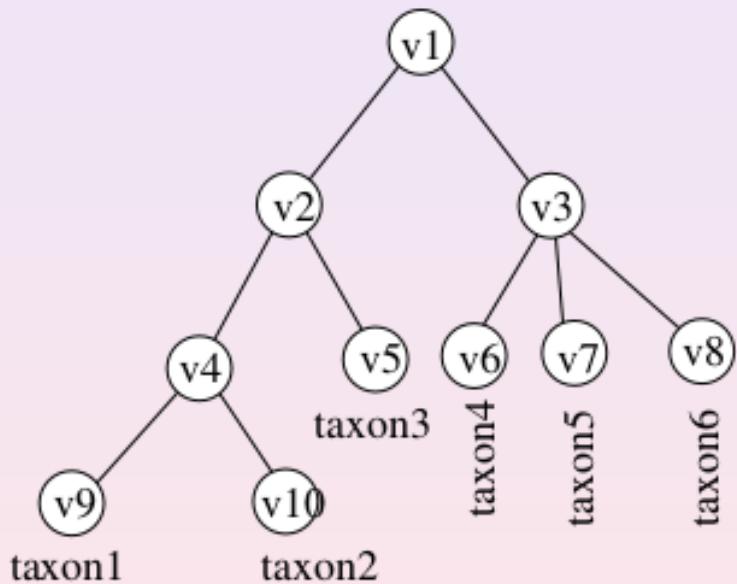


**Topology +
Divergence times +
Divergence rates**

Filogenia: Representación de Arboles

Such a nested structure can be written using nested brackets:

Phylogenetic tree



nested description

v_1
↓
 (v_2, v_3)
↓
 $((v_4, v_5), (v_6, v_7, v_8))$
↓
 $(((v_9, v_{10}), v_5), (v_6, v_7, v_8))$

Description:

$((\text{taxon}_1, \text{taxon}_2), \text{taxon}_3), (\text{taxon}_4, \text{taxon}_5, \text{taxon}_6))$

Filogenia: Representación de Arboles

Formato Newick

- El Arbol termina con un punto y coma
- Los nodos interiores se representan por un par de parentesis “(“ y “)”
- Entre los parentesis, existen representaciones de los nodos que son descendientes inmediatos del nodo anterior, separados por comas
- Las hojas se representan por sus nombres
- Los arboles se pueden dividir (generar multiples ramas) a cualquier nivel
- Las longitudes de las ramas se pueden incorporar al arbol colocando un numero real despues del nodo y precedido por dos puntos. Esto representa la longitud de la rama que se encuentra debajo de ese nodo.

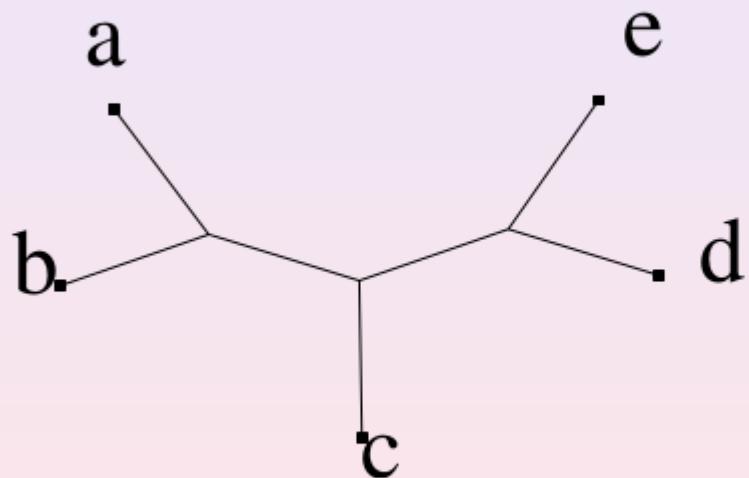
Filogenia: Representación de Arboles

Formato Newick

Printing a phylogenetic tree

Here are two examples:

unrooted tree



Newick string:

`((a,b), c, (d,e));`

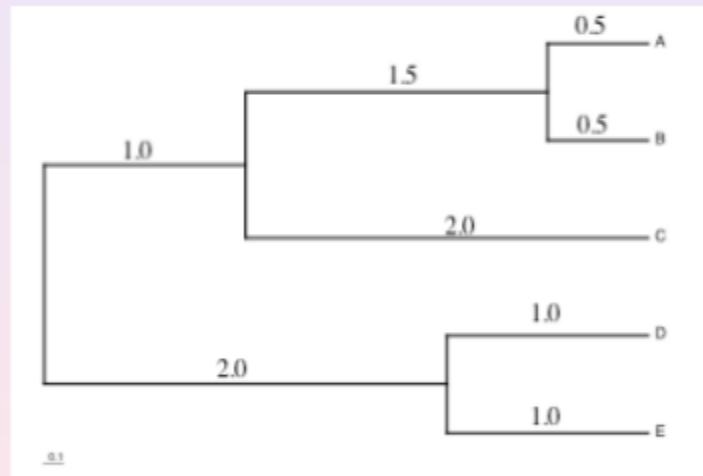
Filogenia: Representación de Arboles

Formato Newick

Printing a phylogenetic tree

Here are two examples:

rooted tree with edge lengths



Newick string:

```
(( ((A:0.5,B:0.5):1.5,C:2.0):1.0, (D:1.0,E:1.0):2.0);
```

Filogenia: Representación de Arboles

Formato Phylo

```
<phyloxml xmlns="http://www.phyloxml.org" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.phyloxml.org
http://www.phyloxml.org/1.10/phyloxml.xsd">
  <phylogeny rooted="false">
    <clade>
      <color>
        <red>128</red>
        <green>128</green>
        <blue>128</blue>
      </color>
    <clade>
      <clade>
        <name>A</name>
      </clade>
      <clade>
        <name>B</name>
      </clade>
    </clade>
    <clade>
      <color>
        <red>0</red>
        <green>0</green>
        <blue>255</blue>
      </color>
      <clade>
        <name>C</name>
      </clade>
      <clade>
        <name>D</name>
      </clade>
    </clade>
    <clade>
      <color>
        <red>250</red>
        <green>128</green>
        <blue>114</blue>
      </color>
      <clade>
        <name>E</name>
      </clade>
      <clade>
        <name>F</name>
      </clade>
      <clade>
        <name>G</name>
      </clade>
    </clade>
  </phylogeny>
</phyloxml>
```

Filogenia: Representación de Arboles – Posibles topologías

- N° de posibles topologías para n nodos finales:
 - Para árboles enraizados:
 - $N_r = (2n-5)! / 2^{n-3}(n-3)$
 - Para árboles no enraizados:
 - $N_u = (2n-3)! / 2^{n-2}(n-2)$
- A partir de n=12, es obligatorio usar heurísticas
 - Imposible calcular todos los árboles posibles

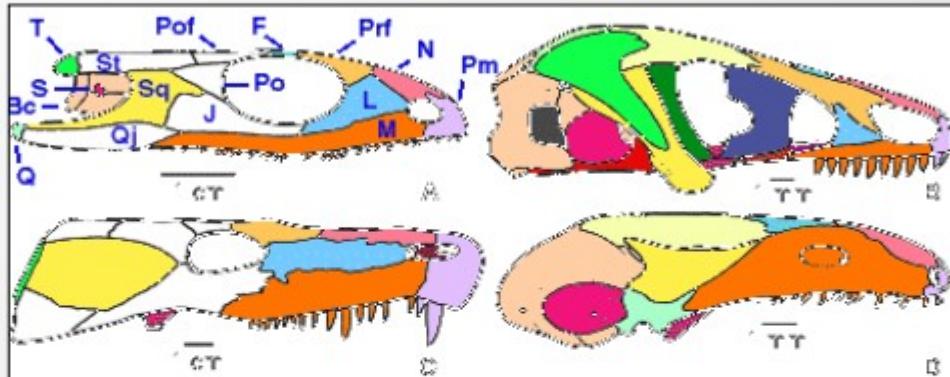
| No. of OTUs | No. of Rooted Trees | No. of Unrooted Trees |
|-------------|----------------------|-----------------------|
| 2 | 1 | 1 |
| 3 | 3 | 1 |
| 4 | 15 | 3 |
| 5 | 105 | 15 |
| 6 | 945 | 105 |
| 7 | 10,395 | 945 |
| 8 | 135,135 | 10,395 |
| 9 | 2,027,025 | 135,135 |
| 10 | 34,489,707 | 2,027,025 |
| 15 | 213,458,046,676,875 | 8×10^{12} |
| 20 | 8×10^{21} | 2×10^{20} |
| 50 | 2.8×10^{76} | 3×10^{74} |

OTU – Operational Taxonomic Unit –
Hojas.

Filogenia:

Trees can be inferred by several criteria:

- Traditional traits: Morphology data



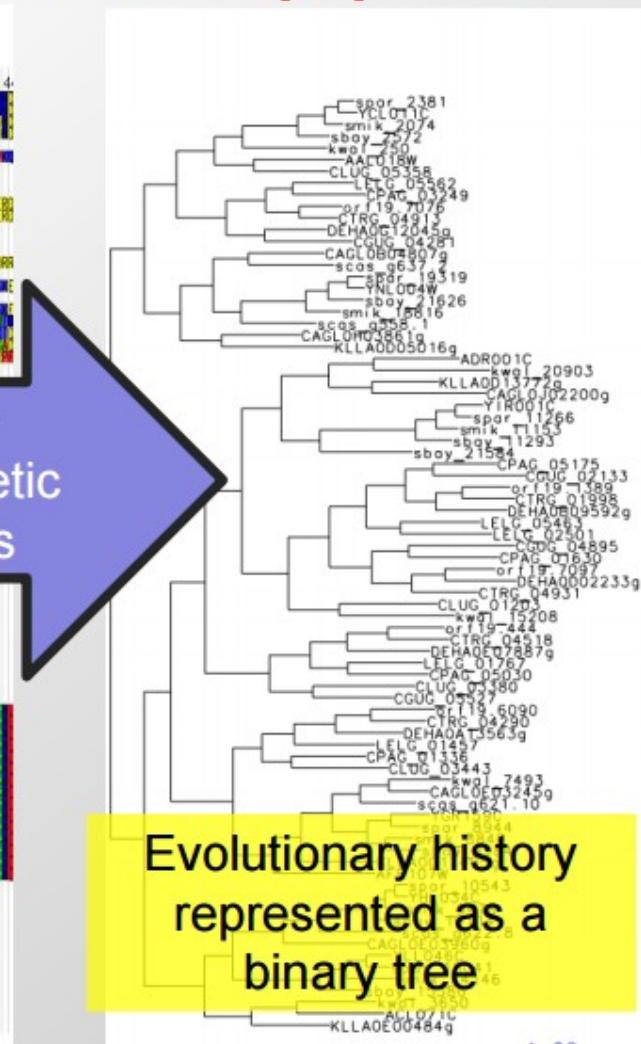
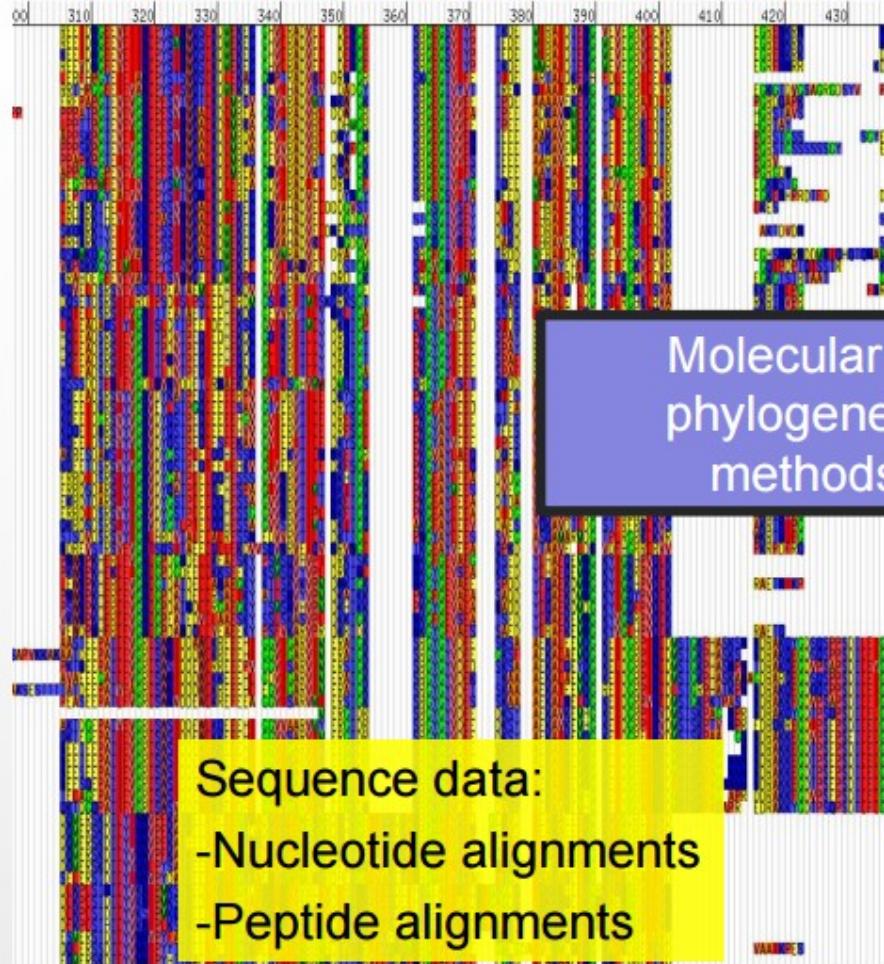
© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

- Modern traits: Molecular data

| | |
|----------|--|
| Kangaroo | A CAGT GACGCCCCAAACGT |
| Elephant | A CAGT GACGCTACAAACGT |
| Dog | CCTG TGACGTAAACAAACGA |
| Mouse | CCTG TGACGTAGCAAACGA |
| Human | CCTG TGACGTAGCAAACGA |

Filogenia:

Inferring a tree from nucleotides/peptides



Filogenia:

- 1 Procedimiento para construir árboles filogenéticos
- 2 Métodos basados en distancias
- 3 Métodos basados en caracteres

Filogenia:

Procedimiento para construir árboles filogenéticos

- Recordemos que el procedimiento para construir árboles filogenéticos se divide en 5 pasos:
 - 1 Elección de los marcadores moleculares
 - 2 Alineamiento múltiple de secuencias
 - 3 Elección de un modelo de evolución
 - 4 Determinación de un método de construcción de árboles
 - 5 Verificación de la fiabilidad del árbol construido

Filogenia:

Procedimiento para construir árboles filogenéticos

Elección de los marcadores moleculares

1 Procedimiento para construir árboles filogenéticos

- Elección de los marcadores moleculares
- Alineamiento
- Modelos de evolución

Filogenia:

Elección de los marcadores moleculares

- Para la construcción de árboles filogenéticos moleculares, se pueden utilizar secuencias de nucleótidos o de proteínas
- La elección de los marcadores moleculares es una cuestión importante porque puede hacer una gran diferencia en la obtención de un árbol correcto
- La decisión de utilizar las secuencias de nucleótidos o proteínas depende de las propiedades de las secuencias y los propósitos del estudio
- Es recomendable utilizar secuencias de nucleótidos, que evolucionan más rápidamente que las proteínas, cuando se estudian organismos estrechamente relacionados (e.g. regiones no codificantes de ADN mitocondrial)



Filogenia:

Elección de los marcadores moleculares

- Por otra parte para estudiar la evolución de grupos de organismos más ampliamente divergentes es aconsejable utilizar secuencias de nucleótidos con lenta evolución (e.g. ARN ribosomal o secuencias de proteínas)
- Si las relaciones filogenéticas que se están analizando están en el nivel más profundo, por ejemplo entre bacterias y eucariotas, lo adecuado es usar secuencias de proteínas conservadas

Filogenia:

Procedimiento para construir árboles filogenéticos

Alineamiento

1 Procedimiento para construir árboles filogenéticos

- Elección de los marcadores moleculares
- **Alineamiento**
- Modelos de evolución

Filogenia:

Procedimiento para construir árboles filogenéticos

Alineamiento

Alineamiento

- El segundo paso en el análisis filogenético es construir el alineamiento de secuencias
- Es probablemente el paso más crítico del procedimiento debido a que éste establece las correspondencias posicionales en la evolución
- Sólo el alineamiento correcto produce inferencias filogenéticas correctas

Filogenia:

Alineamiento

- Por esta razón es importante utilizar los métodos del estado del arte para alineamiento múltiple de secuencias
- Se recomienda obtener el resultado del alineamiento de varias fuentes y compararlos cuidadosamente para identificar el mejor

Filogenia:

Procedimiento para construir árboles filogenéticos

Modelos de evolución

1 Procedimiento para construir árboles filogenéticos

- Elección de los marcadores moleculares
- Alineamiento
- Modelos de evolución

Filogenia:

Modelos de evolución

- Una forma simple de medir la divergencia entre 2 secuencias es contar el número de substituciones en un alineamiento (distancia entre secuencias)
- Sin embargo, el número de substituciones observadas pueden no reflejar los verdaderos eventos evolutivos que ocurrieron
- Cuando una mutación es observada, e.g. que *A* sea reemplazado con *C*, el nucleótido pudo haber tenido en realidad varios pasos intermedios: $A \rightarrow T \rightarrow G \rightarrow C$
- Del mismo modo, podría haber ocurrido una mutación inversa, donde se dan cambios como $G \rightarrow C \rightarrow G$
- Además, un nucleótido idéntico observado en el alineamiento puede deberse a mutaciones en paralelo en ambas secuencias

Filogenia:

Modelos de evolución

- Todo lo anterior dificulta la estimación de las verdaderas distancias evolutivas entre las secuencias estudiadas
- Este efecto es conocido con el nombre de *homoplasia*, la cual si no es corregida puede llevar a la construcción de árboles incorrectos
- Para corregir la homoplasia se requieren modelos estadísticos (*modelos de evolución*) para inferir las verdaderas distancias evolutivas entre secuencias

Filogenia: Modelos de sustitucion

- Definición matemática de la distancia entre dos secuencias de longitud N
- Distancia de Hamming: cuenta el número de cambios (p)
- Distancia de Hamming normalizada: $p' = p/N$

A. Sequences

sequence A AC_GC_GT_TGGGCGA_TGGCAAC
sequence B AC_GC_GT_TGGGCGA_CGGTAAT
sequence C AC_GCATTGA_ATGATGA_TAAAT
sequence D ACACATTGA_GTGATA_ATAAT

B. Distances between sequences, the number of steps required to change one sequence into the other.

n_{AB} 3
 n_{AC} 7
 n_{AD} 8
 n_{BC} 6
 n_{BD} 7
 n_{CD} 3

C. Distance table

| | A | B | C | D |
|---|---|---|---|---|
| A | - | 3 | 7 | 8 |
| B | - | - | 6 | 7 |
| C | - | - | - | 3 |
| D | - | - | - | - |

D. The assumed phylogenetic tree for the sequences A-D showing branch lengths. The sum of the branch lengths between any two sequences on the trees has the same value as the distance between the sequences.



Filogenia: Modelos de sustitucion

- Problema: sustituciones observadas vs reales

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | A | C | C | T | T | C | A | A | T | C | A | C | G | G | G | A | C | T |
| T | T | C | C | T | T | C | A | A | T | C | A | C | G | G | G | A | C | T |
| T | T | C | C | T | T | C | A | A | T | C | A | C | G | G | G | A | C | T |
| T | T | C | C | T | T | C | A | A | T | C | A | C | G | G | G | A | C | T |
| T | T | C | C | T | T | C | A | A | T | C | A | C | C | G | G | A | C | T |
| T | T | C | C | T | T | C | A | A | T | C | T | C | C | G | G | A | C | T |
| C | A | C | C | T | T | C | A | A | T | C | T | C | C | G | G | A | C | T |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

Observada: 3

Real: 6

Filogenia:

Modelos de evolución

- Algunos de los modelos de evolución más conocidos son:
 - El modelo Jukes–Cantor
 - El modelo Kimura
- El **modelo Jukes–Cantor** es el más simple de los dos y asume que todos los nucleótidos son substituidos con igual probabilidad

Filogenia:

Modelos de evolución

- Este modelo emplea una función logarítmica para derivar las distancias evolutivas que incluyen cambios ocultos:

$$d_{AB} = -(3/4) \ln[1 - (4/3)p_{AB}]$$

- Donde d_{AB} es la distancia evolutiva entre las secuencias A y B , y p_{AB} es la distancia observada medida como la proporción de substituciones sobre la toda la longitud del alineamiento
- Por ejemplo, si un alineamiento de las secuencias A y B tiene 20 nucleótidos de largo y 6 pares son diferentes, la secuencia difiere en 30 %, i.e., tienen una distancia observada de 0.3:

$$d_{AB} = -(3/4) \ln[1 - (4/3 \times 0.3)] = 0.38$$

Filogenia: Modelos de sustitucion

- ◆ Corrección de Jukes-Cantor (1969)
 - ◆ Sea p' la distancia de Hamming normalizada y s el número de residuos distintos (4 para nucleótidos, 20 para aminoácidos)

$$d = -\frac{s-1}{s} \ln\left(1 - \frac{s}{s-1} p'\right)$$

- ◆ d es una estimación del número de cambios reales
 - ◆ Considera que la probabilidad de sustitución es igual para todas las combinaciones de nucleótidos/aminoácidos

Filogenia:

Modelos de evolución

- El **modelo Kimura** es más sofisticado (realista) ya que considera diferentes las tasas de mutación para las *transiciones* (substitución de una purina por otra o una pirimidinas por otra) y para las *transversiones* (substitución de una purina por una pirimidina o vice versa)
- De acuerdo a este modelo las transiciones ocurren más frecuentemente que las transversiones, lo cual provee mejores estimaciones de la distancia evolutiva

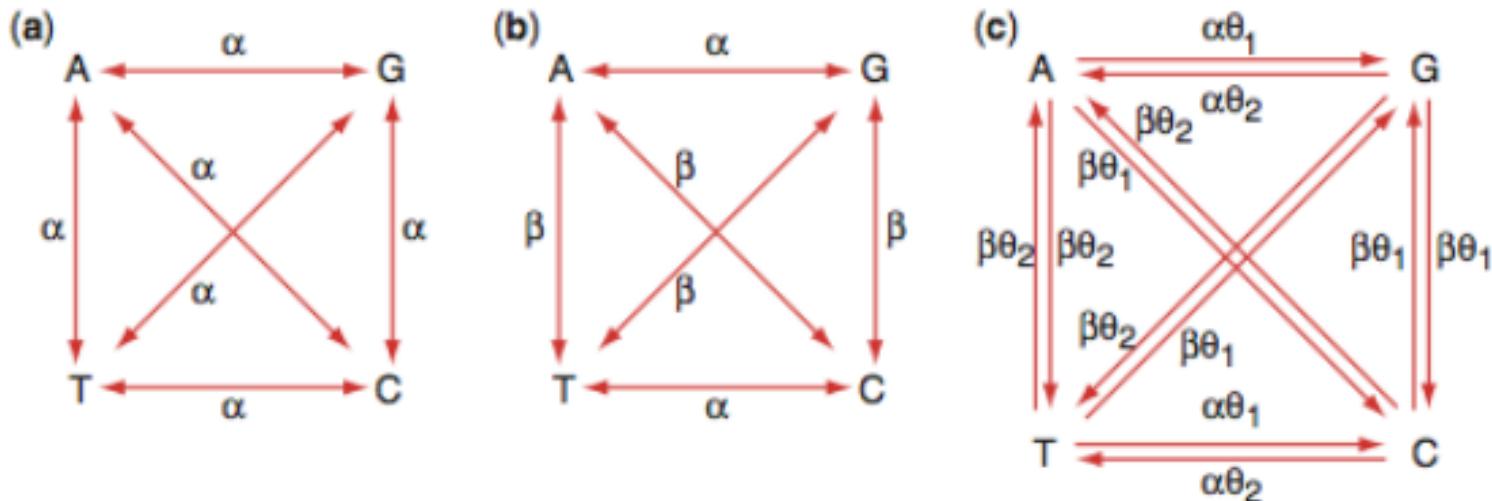
$$d_{AB} = -(1/2) \ln(1 - 2p_{ti} - p_{tv}) - (1/4) \ln(1 - 2p_{tv})$$

- Donde p_{ti} es la frecuencia observada de transición y p_{tv} la frecuencia de transversión



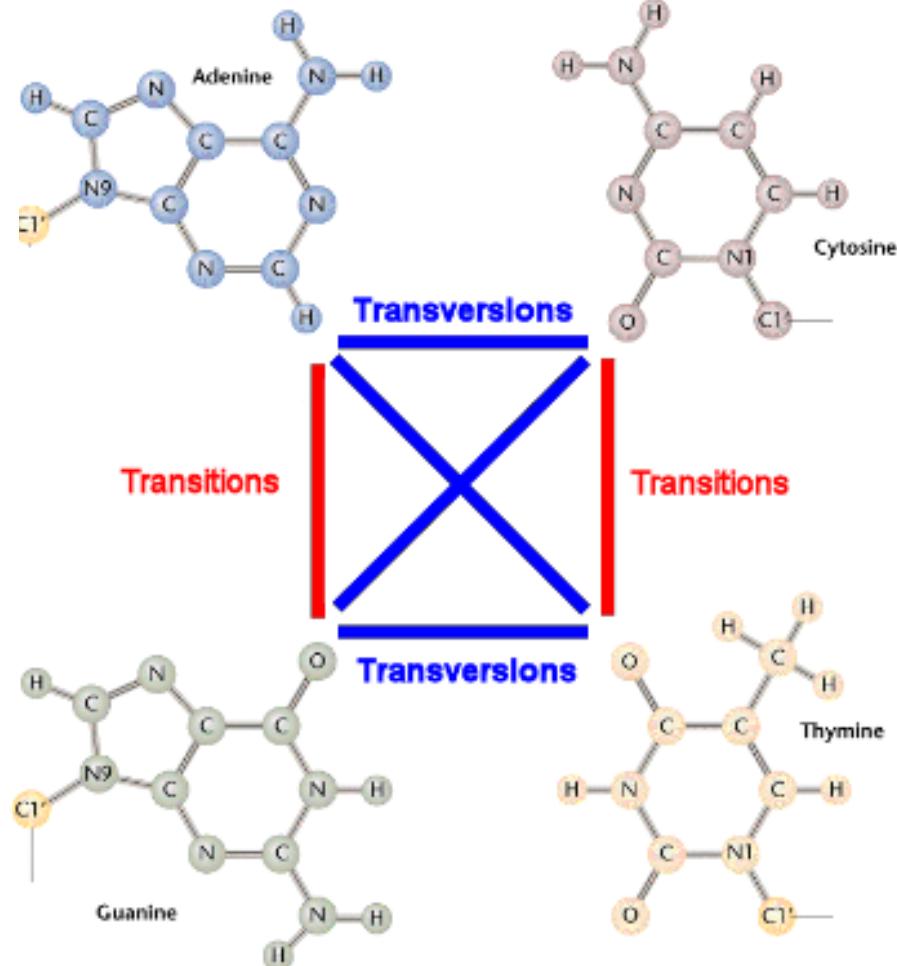
Filogenia: Modelos de sustitucion

- Modelos de Kimura (1980) para nucleótidos
 - Asigna distintas probabilidades de sustitución
 - Modelo de dos parámetros (b): distinta probabilidad a transversión que a transición
 - Transición (α): cambio de purina a purina (o de pirimidina a pirimidina)
 - Transversión (β): cambio de purina a pirimidina (o viceversa)
 - Modelos más complejos (c): distintas probabilidades para cada sustitución



Filogenia: Modelos de sustitucion

- Aunque hay cuatro tipos de transversiones y sólo dos de transiciones, por las propiedades químicas de las bases, la transición es mucho más común
 - Debido a la diferencia en anillos



Filogenia:

Modelos de evolución

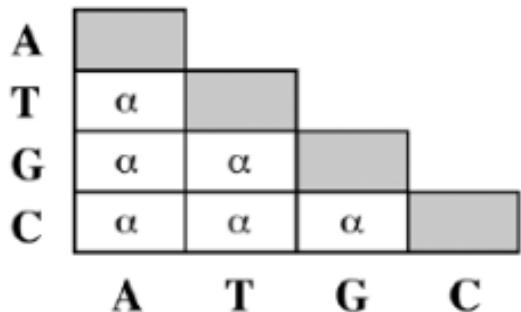
- Por ejemplo, supongamos que las secuencias A y B difiere en 30 %, donde 20 % de los cambios corresponden a transiciones y 10 % a transversiones
- Usando el modelo Kimura tenemos que la distancia evolutiva d_{AB} entre las secuencias A y B puede ser calculado así:

$$d_{AB} = -(1/2) \ln(1 - 2 \times 0.2 - 0.1) - (1/4) \ln(1 - 2 \times 0.1) = 0.40$$

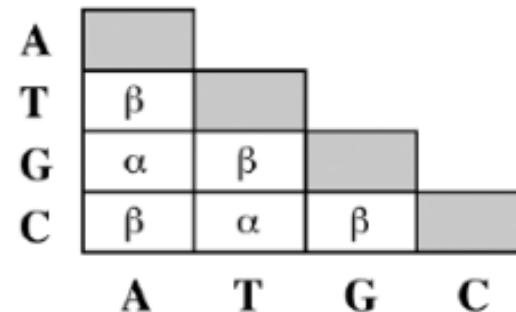
Filogenia:

Modelos de evolución

- Comparación entre los modelos Jukes–Cantor y Kimura



Jukes-Cantor model



Kimura model

Filogenia:

Modelos de evolución

- Algunos otros modelos evolutivos más complejos: TN93, HKY, y GTR
- Toman en cuenta más parámetros para realizar los cálculos
- Sin embargo, normalmente no son usados en la práctica (cálculos complicados, alta variabilidad del resultado)

Modelos de evolución

- Para secuencias de proteínas, se emplean las matrices de substitución de aminoácidos: PAM o JTT
- También existen variantes de los modelos Jukes–Cantor y Kimura para proteínas
- Por ejemplo, el modelo Kimura utiliza la siguiente fórmula:

$$d = -\ln(1 - p - 0.2p^2)$$

- Donde p es la distancia observada entre dos secuencias

2

Métodos basados en distancias

- Introducción
- Métodos basados en agrupamiento
- Métodos basados en optimalidad

Filogenia:

- Como hemos visto las verdaderas distancias de evolución entre secuencias pueden ser calculadas a partir de las distancias observadas después de una corrección con algún modelo evolucionario
- Las distancias evolutivas calculadas pueden ser usadas para construir una matriz de distancias entre todos los pares de taxones
- Basado en los puntajes de distancias entre pares de la matriz, es posible construir un árbol filogenético para todos los taxones involucrados

Filogenia:

Introducción

- Los algoritmos basados en distancias para construir árboles filogenéticos pueden ser subdivididos:
 - Métodos basados en agrupamiento
 - Métodos basados en optimalidad
- Los **algoritmos basados en agrupamiento** calculan el árbol usando una matriz de distancias e iniciando por los pares de secuencias más similares
- El método de Pares No Ponderados Utilizando Media Aritmética (*unweighted pair group method using arithmetic average*, UPGMA) y de Unión de Vecinos son ejemplos de este tipo de algoritmos

Introducción

- Los **algoritmos basados en optimalidad** comparan muchas topologías alternativas de árboles y seleccionan el que tenga el mejor ajuste entre las distancias estimadas en el árbol y las distancias evolutivas reales
- Esta categoría incluye los algoritmos Fitch-Margoliash y de Evolución Mínima

Filogenia:

2

Métodos basados en distancias

- Introducción
- **Métodos basados en agrupamiento**
- Métodos basados en optimalidad

Filogenia:

Métodos basados en agrupamiento

- El método más simple **basado en agrupamiento** es **UPGMA** (*unweighted pair group method using arithmetic average*)
- Construye un árbol por un método de agrupamiento secuencial
- Dada una matriz de distancias, éste inicia mediante la agrupación de los dos taxones con la menor distancia
- Un nodo interior es colocado en el punto medio entre ellos y se crea una matriz reducida al considerar el nuevo grupo como un único taxón

Filogenia:

Métodos basados en agrupamiento

- Las distancias entre este nuevo taxón compuesto y el resto de los taxones se calculan para crear dicha matriz
- El mismo proceso de agrupamiento se repite y otra nueva matriz reducida se crea
- La iteración continúa hasta que todos los taxones se colocan en el árbol
- El último taxón añadido se considera como el grupo fuera lo que produce un árbol con raíz

Filogenia: UPGMA Unweighted Pair Group Method with Arithmetic Mean

El Algoritmo UPGMA construye un dendograma que refleja la estructura de una matriz de similitud. A cada paso los clusters mas cercanos se combinan en un cluster de mas alto nivel. La distancia entre dos clusters \mathcal{A} y \mathcal{B} , es el promedio de todas las distancias $d(x,y)$ entre pares de objetos x en \mathcal{A} e y en \mathcal{B} , de manera que el promedio de la distancia entre elementos de cada cluster es:

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y)$$

Es decir en cada paso del algoritmo la distancia recalculada entre los clusters unidos $\mathcal{A} \cup \mathcal{B}$ y un nuevo cluster X esta dado por el promedio proporcional de las distancias $d_{\mathcal{A},x}$ y $d_{\mathcal{B},x}$

$$d_{(\mathcal{A} \cup \mathcal{B}),X} = \frac{|\mathcal{A}| \cdot d_{\mathcal{A},X} + |\mathcal{B}| \cdot d_{\mathcal{B},X}}{|\mathcal{A}| + |\mathcal{B}|}$$

Filogenia: UPGMA Unweighted Pair Group Method with Arithmetic Mean

(a)

| | 1 | 2 | 3 | 4 | 5 |
|---|-----|-----|-----|-----|---|
| 1 | — | | | | |
| 2 | 0.1 | — | | | |
| 3 | 0.8 | 0.8 | — | | |
| 4 | 0.8 | 1 | 0.3 | — | |
| 5 | 0.9 | 0.9 | 0.3 | 0.2 | — |

(b)

| | (1,2) | 3 | 4 | 5 |
|-------|-------|-----|-----|---|
| (1,2) | — | | | |
| 3 | 0.8 | — | | |
| 4 | 0.9 | 0.3 | — | |
| 5 | 0.9 | 0.3 | 0.2 | — |

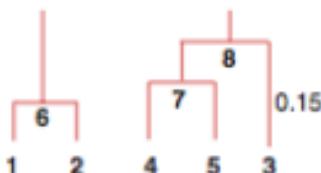
(c)

| | (1,2) | 3 | (4,5) |
|-------|-------|-----|-------|
| (1,2) | — | | |
| 3 | 0.8 | — | |
| (4,5) | 0.9 | 0.3 | — |



(d)

| | (1,2) | [3,(4,5)] |
|-----------|-------|-----------|
| (1,2) | — | |
| [3,(4,5)] | 0.85 | — |



- Es un método sencillo que se basa en agrupar las secuencias más cercanas en base a su distancia
- El proceso es el siguiente:
 - a) Calculamos la matriz de distancias, elegimos la menor distancia: $d_{1,2}$
 - b) Unimos las secuencias 1 y 2, siendo la longitud de la rama la $0.5 \cdot d_{1,2}$. Calculamos las distancias al nuevo nodo (1,2) y seleccionamos la menor distancia ahora: $d_{4,5}$
 - c) Calculamos las distancias al nuevo nodo (4,5), la longitud de las ramas, y elegimos de nuevo la menor: $d_{3,45}$
 - d) Continuamos hasta terminar de unir nodos
- UPGMA asume que el reloj molecular de todos los nodos es igual
- Es un método muy utilizado en análisis de microarrays, pero para análisis filogenéticos suele ser bastante menos preciso que el método de Neighbor-Joining

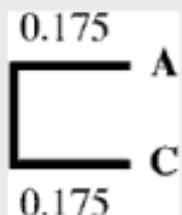
Filogenia:

Métodos basados en agrupamiento

- Ejemplo de construcción de un árbol filogenético usando el método UPGMA (tomado del libro de Xiong)

| | A | B | C |
|---|-------------|-------------|-------------|
| B | 0.40 | | |
| C | 0.35 | 0.45 | |
| D | 0.60 | 0.70 | 0.55 |

1. Using a distance matrix involving four taxa, A, B, C, and D, the UPGMA method first joins two closest taxa together which are A and C (0.35 in grey). Because all taxa are equidistant from the node, the branch length for A to the node is $AC/2 = 0.35/2 = 0.175$.



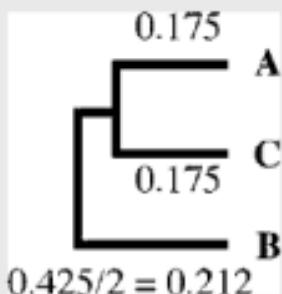
Filogenia:

Métodos basados en agrupamiento

2. Because A and C are joined into a cluster, they are treated as one new composite taxon, which is used to create a reduced matrix. The distance of A-C cluster to every other taxa is one half of a taxon to A and C, respectively. That means that the distance of B to A-C is $(AB + BC)/2$; and that of D to A-C is $(AD + CD)/2$.

| | A-C | B |
|---|--------------------------------|------|
| B | $\frac{0.4 + 0.45}{2} = 0.425$ | |
| D | $\frac{0.55 + 0.6}{2} = 0.575$ | 0.70 |

3. In the newly reduced-distance matrix, the smallest distance is between B and A-C (in grey), which allows the grouping of B and A-C to create a three-taxon cluster. The branch length for the B is one half of B to the A-C cluster.



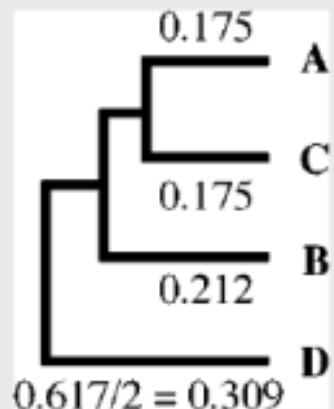
Filogenia:

Métodos basados en agrupamiento

4. When B and A-C are grouped and treated as a single taxon, this allows the matrix to reduce further into only two taxa, D and B-A-C. The distance of D to the composite taxon is the average of D to every single component which is $(BD + AD + CD)/3$.

| | B-A-C |
|---|--------------------------------------|
| D | $\frac{0.7 + 0.6 + 0.55}{3} = 0.617$ |

5. D is the last branch to add to the tree, whose branch length is one half of D to B-A-C.



Filogenia:

Métodos basados en agrupamiento

6. Because distance trees allow branches to be additive, the resulting distances between taxa from the tree path can be used to create a distance matrix. Obviously, the estimated distances do not match the actual evolutionary distances shown, which illustrates the failure of UPGMA to precisely reflect the experimental observation.

| | A | B | C |
|---|------|------|------|
| B | 0.42 | | |
| C | 0.35 | 0.42 | |
| D | 0.62 | 0.62 | 0.62 |

Filogenia:

2

Métodos basados en distancias

- Introducción
- Métodos basados en agrupamiento
- **Métodos basados en optimalidad**

Filogenia:

Métodos basados en optimalidad

- Estos métodos, a diferencia de los basados en agrupamiento, tienen un algoritmo bien definido para comparar todas las posibles topologías de árboles a fin de seleccionar la que mejor se ajuste a la matriz de distancias evolutivas real
- Basados en los diferentes criterios de optimalidad, hay dos tipos de algoritmos: Fitch-Margoliash y de Evolución Mínima
- Una clara desventaja de este tipo de algoritmos son los altos tiempos de cómputo que demandan debido a la búsqueda exhaustiva que realizan

Filogenia:

Métodos basados en optimalidad

- El método **Fitch–Margoliash** (FM) selecciona el mejor árbol entre todos los posibles basándose en la mínima desviación entre las distancias calculadas en la totalidad de las ramas del árbol y las distancias del conjunto de datos original
- Inicia por agrupar aleatoriamente 2 taxones en un nodo y crear 3 ecuaciones para describir las distancias
- Despues resuelve algebraicamente las 3 ecuaciones para longitudes de rama desconocidas
- Con ayuda de este grupo de 2 taxones se crea una nueva matriz reducida

Filogenia:

Métodos basados en optimalidad

- Este proceso itera hasta que el árbol se forma completamente
- El método busca todas las posibles topologías y selecciona aquella que tiene la menor desviación cuadrática entre las distancias reales y las longitudes calculadas de las ramas
- El criterio de optimalidad es expresado con la fórmula:

$$E = \sum_{i=1}^{T-1} \sum_{j=j+1}^T \frac{(d_{ij} - p_{ij})^2}{d_{ij}^2} \quad (1)$$

- Donde E es el error del árbol estimado, T es el número de taxones, d_{ij} es la distancia en el conjunto de datos original entre los taxones i, j y p_{ij} es la longitud de la rama correspondiente



Filogenia:

Métodos basados en optimalidad

- Métodos basados en distancias
 - Ventaja: habilidad para hacer uso de diferentes modelos de substitución para corregir las distancias evolutivas
 - Desventaja: La información real de las secuencia se pierde cuando todas las variaciones son reducidas a un único valor, impidiendo la inferencia de secuencias ancestro en los nodos internos

Filogenia:

3

Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- Construcción del árbol filogenético con MP
- Planteamiento formal del problema de MP
- Trabajo relacionado
- Parsimonia ponderada
- Métodos de búsqueda en árboles
- Ventajas y desventajas
- Atracción de ramas largas
- Máxima verosimilitud
- Construcción del árbol filogenético con MV

Filogenia:

Introducción

- Los métodos basados en caracteres (también llamados métodos discretos) están basados directamente en el análisis de los caracteres que forman las secuencias y no de las distancias entre pares de éstas
- Estos métodos cuentan los eventos de mutación acumulados en las secuencias y pueden por lo tanto eliminar la perdida de información que se da cuando los caracteres son transformados a distancias
- Esta preservación de información de los caracteres significa que la dinámica evolutiva de cada uno de ellos puede ser estudiada

Filogenia:

Introducción

- Adicionalmente, también es posible inferir secuencias de ancestros
- Los dos métodos basados en caracteres más populares son:
 - Máxima Parsimonia (MP)
 - Máxima Verosimilitud (MV)

Filogenia:

3

Métodos basados en caracteres

- Introducción
- **Máxima parsimonia**
- Construcción del árbol filogenético con MP
- Planteamiento formal del problema de MP
- Trabajo relacionado
- Parsimonia ponderada
- Métodos de búsqueda en árboles
- Ventajas y desventajas
- Atracción de ramas largas
- Máxima verosimilitud
- Construcción del árbol filogenético con MV

Filogenia:

Máxima parsimonia

- El método de MP selecciona el árbol que tiene el mínimo número de cambios evolutivos, i.e., el árbol cuyas ramas tengan promedio la mínima longitud
- Se basa en el principio conocido como **Navaja de Occam** (*Occam's razor*) formulado por William Ockham en el siglo XIV
- Este principio hace referencia a un tipo de razonamiento basado en una premisa muy simple: en igualdad de condiciones la solución más sencilla es probablemente la correcta
- Esto es porque la solución más simple requiere el menor número de suposiciones y de operaciones lógicas

Filogenia:

Máxima parsimonia

- Para el análisis filogenético, la parsimonia es una buena suposición
- Siguiendo este principio, un árbol con el menor número de substituciones es probablemente la mejor opción para explicar las diferencias entre los taxones estudiados
- Esta perspectiva se justifica por el hecho de que los cambios evolutivos que suceden dentro de lapsos de tiempo cortos son relativamente raros

Filogenia:

Máxima parsimonia

- Esto implica que un árbol con cambios mínimos es muy probable que sea una buena estimación del verdadero árbol
- Al minimizar los cambios, el método minimiza el ruido filogenético debido a la homoplasia (cambio evolutivo paralelo que hace que dos organismos presenten un mismo carácter adquirido independientemente) y a la evolución independiente

Filogenia:

3

Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- **Construcción del árbol filogenético con MP**
- Planteamiento formal del problema de MP
- Trabajo relacionado
- Parsimonia ponderada
- Métodos de búsqueda en árboles
- Ventajas y desventajas
- Atracción de ramas largas
- Máxima verosimilitud
- Construcción del árbol filogenético con MV

Filogenia:

Construcción del árbol filogenético con MP

- La construcción del árbol filogenético de MP funciona buscando todas las posibles topologías de árboles y reconstruyendo secuencias de ancestros que requieren el mínimo número de cambios evolutivos a las secuencias actuales
- Para ahorrar tiempo de cómputo, sólo un pequeño número de sitios, que tienen información filogenética importante, son usados en la determinación del árbol
- Estos sitios son llamados *sitios informativos*, los cuales son definidos como sitios que tienen al menos dos tipos diferentes de caracteres, cada uno ocurriendo al menos dos veces

Filogenia:

Construcción del árbol filogenético con MP

- Ejemplo de extracción de sitios informativos

| | sites | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|-----|-------|---|---|---|---|---|---|---|---|
| | taxa | | | | | | | | |
| I | | A | A | T | T | A | G | C | T |
| II | | G | G | T | C | G | T | A | G |
| III | | A | A | T | G | C | G | C | T |
| IV | | A | G | T | A | A | G | C | A |
| V | | A | C | T | T | C | G | C | G |
| VI | | A | C | A | T | G | G | C | A |

Filogenia:

Construcción del árbol filogenético con MP

- Los sitios informativos son los que pueden a menudo ser explicados median una topología de árbol única
- Los sitios no-informativos son constantes o tienen cambios que ocurren una sola vez
- Los sitios constantes obviamente no son útiles para evaluar diferentes topologías
- Los sitios con cambios ocurriendo una sola vez tampoco son útiles porque pueden ser explicados por múltiples topologías
- Por esta razón los sitios no-informativos son desechados en el proceso de construcción de un árbol filogenético de MP

Filogenia:

Construcción del árbol filogenético con MP

- Una vez que los sitios informativos son identificados y los no-informativos son descartados, el mínimo número de substituciones en cada sitio informativo es calculado para una topología dada
- El número total de cambios en todos los sitios informativos son sumados para cada posible topología
- Y el árbol con el más pequeño número de cambios es elegido como el mejor

Filogenia:

Construcción del árbol filogenético con MP

- La clave para contar un número mínimo de sustituciones para un sitio particular es determinar los estados del carácter ancestral en los nodos internos
- Debido a que estos estados de caracteres ancestrales no se conocen directamente, pueden existir múltiples soluciones posibles
- En este caso, el principio de parsimonia se aplica para elegir los estados de los caracteres que resultan en un mínimo número de sustituciones

Filogenia:

Construcción del árbol filogenético con MP

- La inferencia de una secuencia ancestral se realiza en dos pasos:
 - ① Se recorre el árbol de las hojas hacia la raíz para determinar todos los posibles estados de los caracteres ancestrales
 - ② Se recorre el árbol de la raíz hacia las hojas para asignar secuencias ancestrales que exigen el número mínimo de sustituciones (*puntaje de parsimonia*)

| | | |
|-----------------------------|-------------|---|
| ✓ myoglobin kangaroo | L F K G H | E T L E K F D K F K M L K S E D E M K A S E D L L K K H D I T V L T A L O N I L K K E |
| ✓ myoglobin harbor porpoise | L F K G H | E T L E K F D K F K M L K T E A E M K A S E D L L K K H D I T V L T A L O O I L K K E |
| ✓ myoglobin gray seal | L F K S H | E T L E K F D K F K M L K S E D D M R R S E D L R K H D I T V L T A L O O I L K K E |
| ✓ alpha globin horse | M F L O F | T T K T Y F P H F - D L S H O - - - S A Q V K A H D E K V D A L T L A V O H L |
| ✓ alpha globin kangaroo | T F H S F | T T K T Y F P H F - D L S H O - - - S A Q I O A M D K K I A D A L G D A V E H I |
| ✓ alpha globin dog | T F G S F | T T K T Y F P H F - D L E P G - - - S A Q V K A H D E K V D A L T T A V A H L |
| ✓ beta globin dog | L L I V Y | P W T Q R F F D S F G D L S T P D A V M S N A K V K A H D K K V L V A F G D A I K N L |
| ✓ beta globin rabbit | L L V V Y | P W T Q R F F E S F G D L S A N A V M N N P K V K A H D K K V L V A F G D A I K N L |
| ✓ beta globin kangaroo | L L I V Y | P W T S R F F D H F G D L S N A K A V M A N P K V L A N D A K V L V A F G D A I K N L |
| ✓ globin river lamprey | F F T S T | F A A Q E F F F P K F K O M T S A D E L K K S A D V R W H A E R I I I N A V N D A V A S M |
| ✓ globin sea lamprey | F F T S T | F A A Q E F F F P K F K O L T T A D O L K K S A D V R W H A E R I I I N A V N D A V A S M |
| ✓ globin insect | V F K A S | F S I M A K F T Q F A G E D L E S - I K G T A P F E I H A N R I I V G F F E K I I G E L |
| ✓ globin zebrafish | I L E K A F | F A A K D L F S F L A N P T D G - - - V N P K L T O H A E K L F A L V R D S A G O L |

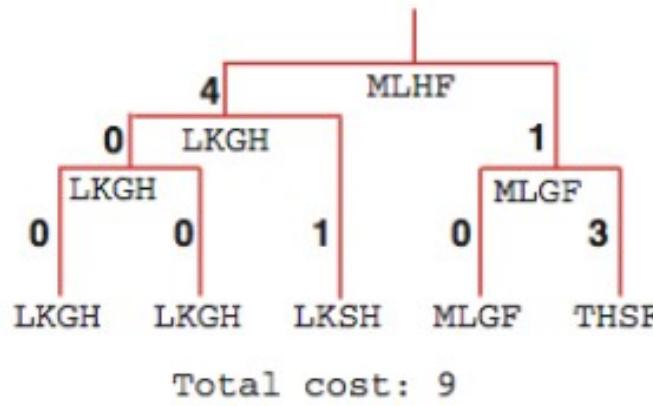
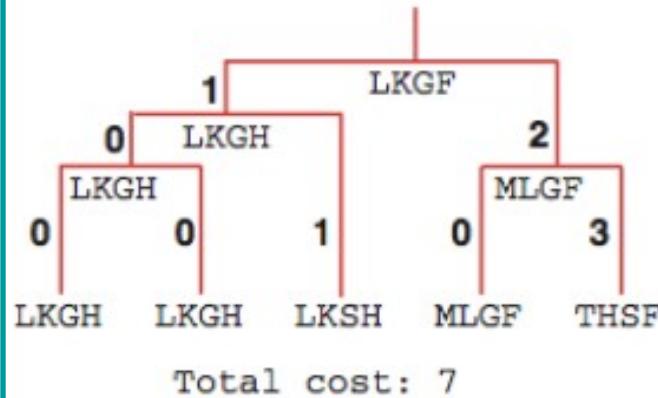
▲ ▲ △ △ △

▲ ▲

a) Los residuos con flecha se descartan por ser poco informativos (demasiado consenso)

| | |
|-------------------|------|
| kangaroo | LKGH |
| porpoise | LKGH |
| gray seal | LKSH |
| horse α | MLGF |
| kangaroo α | THSF |

b) Tomemos un ejemplo con 4 de los 5 primeros aminoácidos para 5 secuencias



d) Construimos árboles a partir de posibles secuencias ancestrales, contando el número de cambios en cada rama. En este caso elegiríamos el de la izquierda

Filogenia:

Construcción del árbol filogenético con MP

- Es necesario subrayar que en realidad la secuencia de nodos ancestrales no siempre puede ser determinada sin ambigüedad
- A veces, puede haber varios caracteres que resultan en un mismo puntaje de parsimonia para un determinado número de topologías
- También es posible que haya dos o más topologías que tienen el mismo puntaje de parsimonia
- En estos casos se tiene que construir un árbol de consenso que representa a todos los árboles parsimoniosos

Filogenia:

Construcción del árbol filogenético con MP

Importancia del problema

Ciencias biológicas

- Desarrollo de nuevas vacunas
- Estudio de la dinámica de comunidades microbianas
- Estudio de antibacteriales y herbicidas
- Desarrollo inteligente de nuevos fármacos

Ciencias de la computación

- El problema de MP es NP-completo
- Equivale al problema del árbol de Steiner en hipercubos [Garey and Johnson, 1977]
- El número de árboles con raíz para n secuencias es:

$$|\mathcal{T}| = (2n - 3)! / 2^{n-2} (n - 2)!$$

- Para $n = 30$ hay 4.95×10^{38} árboles (100 millones de sol./seg $\approx 1.57 \times 10^{21}$ siglos)

Filogenia:

3

Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- Construcción del árbol filogenético con MP
- **Planteamiento formal del problema de MP**
- Trabajo relacionado
- Parsimonia ponderada
- Métodos de búsqueda en árboles
- Ventajas y desventajas
- Atracción de ramas largas
- Máxima verosimilitud
- Construcción del árbol filogenético con MV

Filogenia:

Planteamiento formal del problema de MP

- Dado un conjunto $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$ de n secuencias de longitud k , previamente alineadas, sobre un alfabeto \mathcal{A} ;
- Un árbol binario con raíz $T = (V, E)$, también llamado árbol filogenético, está compuesto por los conjuntos V y E que corresponden a sus nodos y aristas
- $|V| = (2n - 1)$ se encuentra dividido en dos subconjuntos:
 - I , que contiene $n - 1$ *nodos internos* (ancestros hipotéticos) cada uno con 2 descendientes;
 - L , compuesto de n *hojas*, i.e., nodos sin descendientes.

Filogenia:

Planteamiento formal del problema de MP

- La secuencia de parsimonia P_w para cada nodo interno $w \in I$ cuyos descendientes son $S_u = \{x_1, \dots, x_k\}$ y $S_v = \{y_1, \dots, y_k\}$ se calcula con la siguiente relación:

$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases} \quad \text{para } 1 \leq i \leq k,$$

- El costo de parsimonia (mutaciones) de la secuencia P_w está definido por:

$$\phi(P_w) = \sum_{i=1}^k C_i \quad \text{donde} \quad C_i = \begin{cases} 1, & \text{si } x_i \cap y_i = \emptyset \\ 0, & \text{sino} \end{cases}$$

- El costo de parsimonia para el árbol T se obtiene de la siguiente manera:

$$\phi(T) = \sum_{w \in I} \phi(P_w) \tag{2}$$

Filogenia:

Planteamiento formal del problema de MP

- El problema de MP consiste entonces en encontrar una topología de árbol T^* para la cual $\phi(T)$ sea mínimo, *i.e.*,

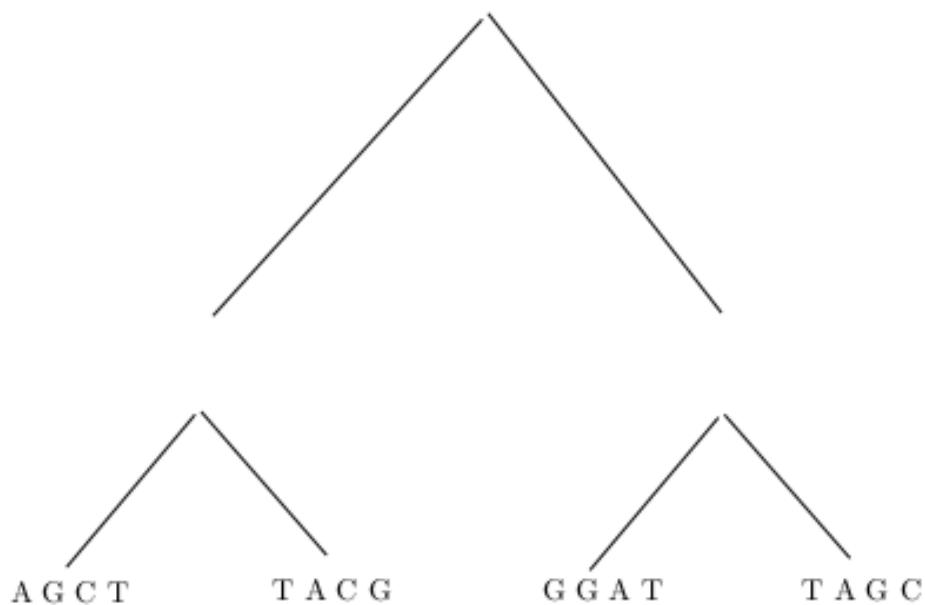
$$\phi(T^*) = \min\{\phi(T) : T \in \mathcal{T}\}$$

donde \mathcal{T} es el conjunto de todas las posibles topologías de árbol (espacio de búsqueda).

Filogenia:

Planteamiento formal del problema de MP

- Dadas $n = 4$ secuencias de longitud $k = 4$ y la siguiente topología de árbol, calculamos la secuencia de parsimonia P_w para cada nodo interno y sumamos el número total de mutaciones



$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases}$$

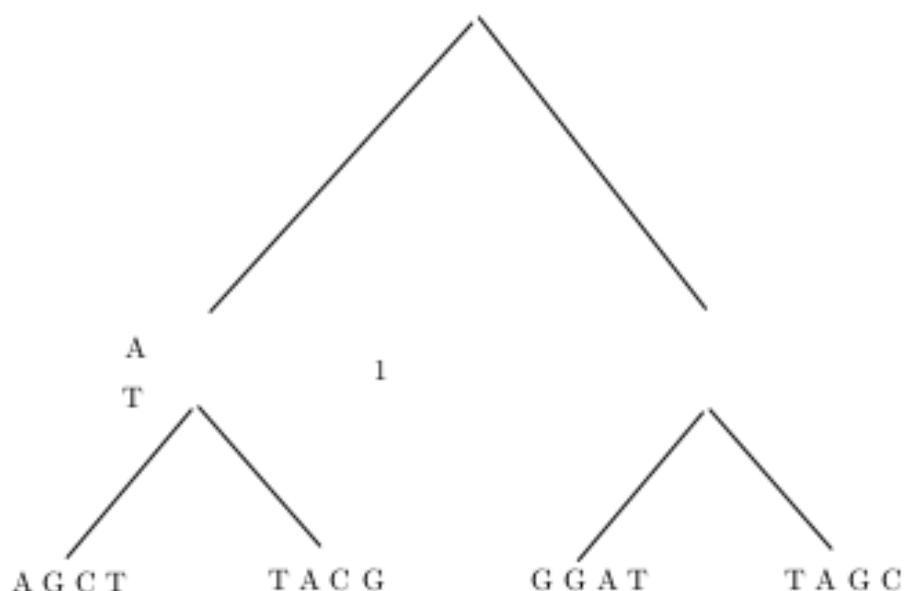
$$\phi(P_w) = \sum_{i=1}^k C_i$$

$$C_i = 1 \text{ si } x_i \cap y_i = \emptyset$$

Filogenia:

Planteamiento formal del problema de MP

- Dadas $n = 4$ secuencias de longitud $k = 4$ y la siguiente topología de árbol, calculamos la secuencia de parsimonia P_w para cada nodo interno y sumamos el número total de mutaciones



$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases}$$

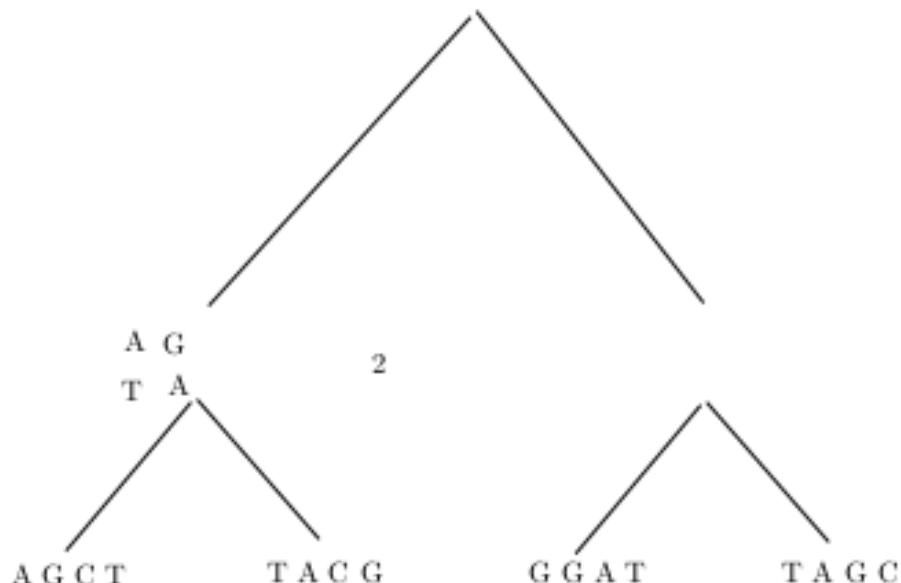
$$\phi(P_w) = \sum_{i=1}^k C_i$$

$$C_i = 1 \text{ si } x_i \cap y_i = \emptyset$$

Filogenia:

Planteamiento formal del problema de MP

- Dadas $n = 4$ secuencias de longitud $k = 4$ y la siguiente topología de árbol, calculamos la secuencia de parsimonia P_w para cada nodo interno y sumamos el número total de mutaciones



$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases}$$

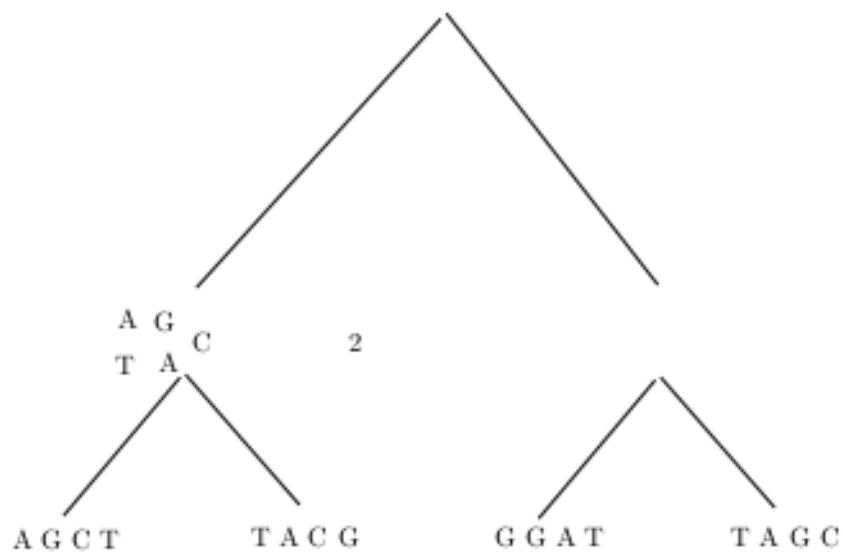
$$\phi(P_w) = \sum_{i=1}^k C_i$$

$$C_i = 1 \text{ si } x_i \cap y_i = \emptyset$$

Filogenia:

Planteamiento formal del problema de MP

- Dadas $n = 4$ secuencias de longitud $k = 4$ y la siguiente topología de árbol, calculamos la secuencia de parsimonia P_w para cada nodo interno y sumamos el número total de mutaciones



$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases}$$

$$\phi(P_w) = \sum_{i=1}^k C_i$$

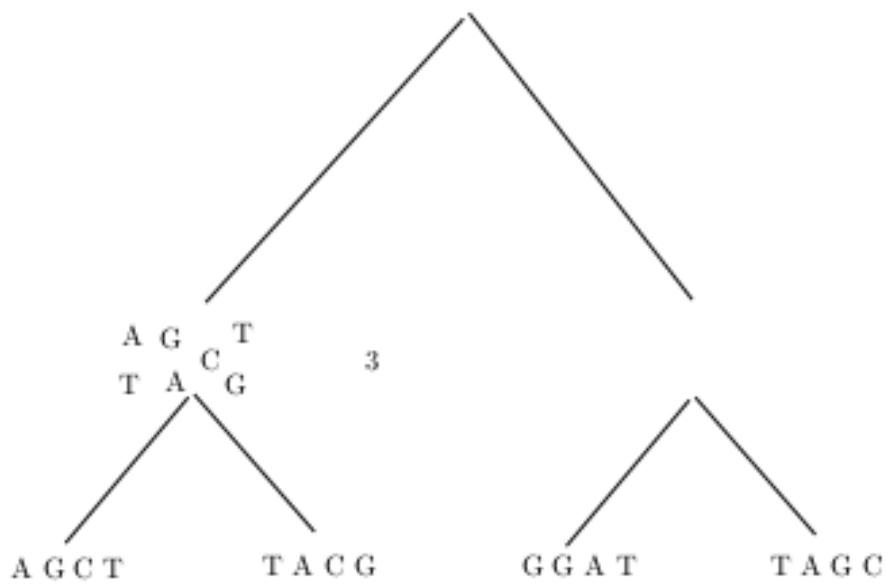
$$C_i = 1 \text{ si } x_i \cap y_i = \emptyset$$



Filogenia:

Planteamiento formal del problema de MP

- Dadas $n = 4$ secuencias de longitud $k = 4$ y la siguiente topología de árbol, calculamos la secuencia de parsimonia P_w para cada nodo interno y sumamos el número total de mutaciones



$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases}$$

$$\phi(P_w) = \sum_{i=1}^k C_i$$

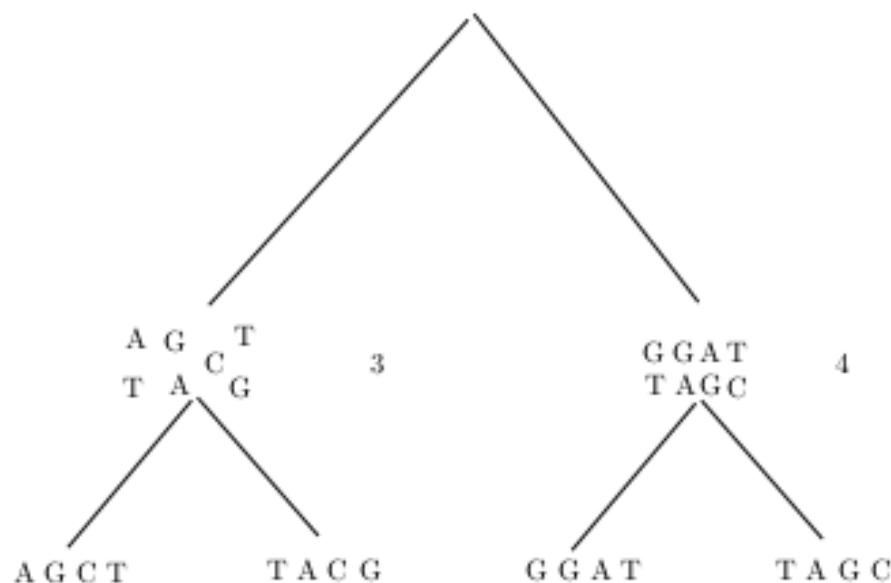
$$C_i = 1 \text{ si } x_i \cap y_i = \emptyset$$



Filogenia:

Planteamiento formal del problema de MP

- Dadas $n = 4$ secuencias de longitud $k = 4$ y la siguiente topología de árbol, calculamos la secuencia de parsimonia P_w para cada nodo interno y sumamos el número total de mutaciones



$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases}$$

$$\phi(P_w) = \sum_{i=1}^k C_i$$

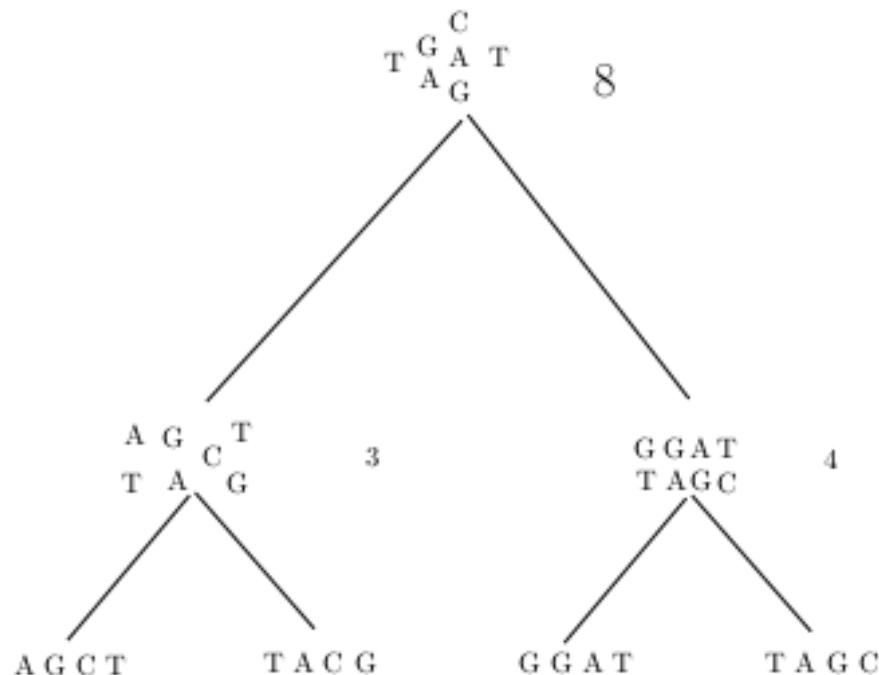
$$C_i = 1 \text{ si } x_i \cap y_i = \emptyset$$



Filogenia:

Planteamiento formal del problema de MP

- Dadas $n = 4$ secuencias de longitud $k = 4$ y la siguiente topología de árbol, calculamos la secuencia de parsimonia P_w para cada nodo interno y sumamos el número total de mutaciones



$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases}$$

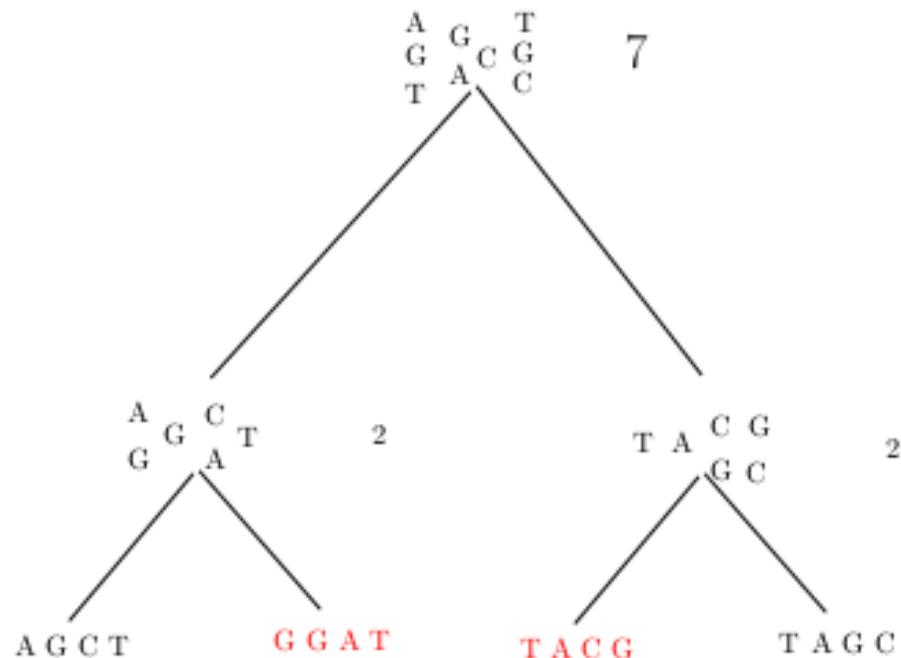
$$\phi(P_w) = \sum_{i=1}^k C_i$$

$$C_i = 1 \text{ si } x_i \cap y_i = \emptyset$$

Filogenia:

Planteamiento formal del problema de MP

- Dadas $n = 4$ secuencias de longitud $k = 4$ y la siguiente topología de árbol, calculamos la secuencia de parsimonia P_w para cada nodo interno y sumamos el número total de mutaciones



$$z_i = \begin{cases} x_i \cup y_i, & \text{si } x_i \cap y_i = \emptyset \\ x_i \cap y_i, & \text{sino} \end{cases}$$

$$\phi(P_w) = \sum_{i=1}^k C_i$$

$$C_i = 1 \text{ si } x_i \cap y_i = \emptyset$$



Filogenia:

Trabajo relacionado

- **Algoritmo exacto**

- Branch & bound (B&B) [Hendy and Penny, 1982] ($n \leq 10$)

- **Algoritmos aproximados**

- Algoritmos voraces [Andreatta and Ribeiro, 2002] (resultados lejanos al óptimo)
- Recocido simulado multiarranque (LVB) [Barker, 2003, Barker, 2012].
- GRASP (*greedy randomized adaptive search procedure*) [Ribeiro and Vianna, 2005]
- GA+PR+LS [Ribeiro and Vianna, 2009]
- Hydra, algoritmo memético [Richer et al., 2009]
- SAMPARS, recocido simulado [Richer et al., 2012] (mejor conocido)

Filogenia:

3 Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- Construcción del árbol filogenético con MP
- Planteamiento formal del problema de MP
- Trabajo relacionado
- **Parsimonia ponderada**
- Métodos de búsqueda en árboles
- Ventajas y desventajas
- Atracción de ramas largas
- Máxima verosimilitud
- Construcción del árbol filogenético con MV

Filogenia:

Parsimonia ponderada

- El método que venimos de describir es no ponderado porque trata todas las mutaciones como equivalentes
- Este método es una sobresimplificación ya que se sabe que las mutaciones de algunos sitios ocurren menos frecuentemente que en otros
- Por ejemplo
 - Las transversiones con respecto a las transiciones
 - Los sitios funcionalmente importantes con respecto a los neutrales

Filogenia:

Inferencia basada en caracteres Parsimonia ponderada

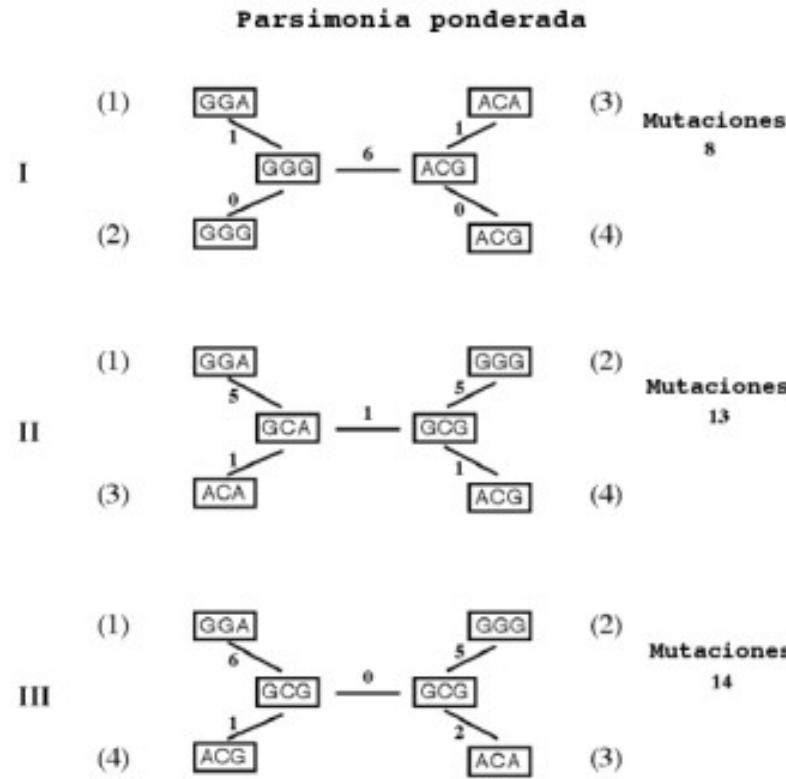
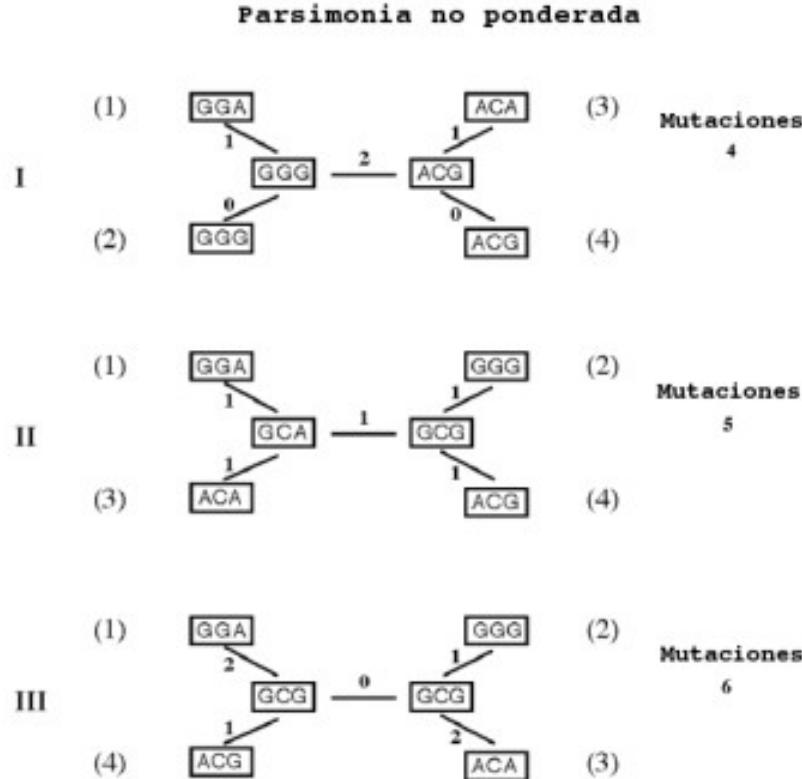
Parsimonia ponderada

- Por lo tanto, un esquema ponderado que tome en cuenta los diferentes tipos de mutaciones ayudaría a seleccionar las topologías de árboles más precisamente
- Este tipo de esquema recibe el nombre de *Parsimonia Ponderada*

Filogenia:

Parsimonia ponderada

- Parsimonia no ponderada y ponderada (transiciones 1, transversiones 5)



Filogenia:

3

Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- Construcción del árbol filogenético con MP
- Planteamiento formal del problema de MP
- Trabajo relacionado
- Parsimonia ponderada
- **Métodos de búsqueda en árboles**
- Ventajas y desventajas
- Atracción de ramas largas
- Máxima verosimilitud
- Construcción del árbol filogenético con MV

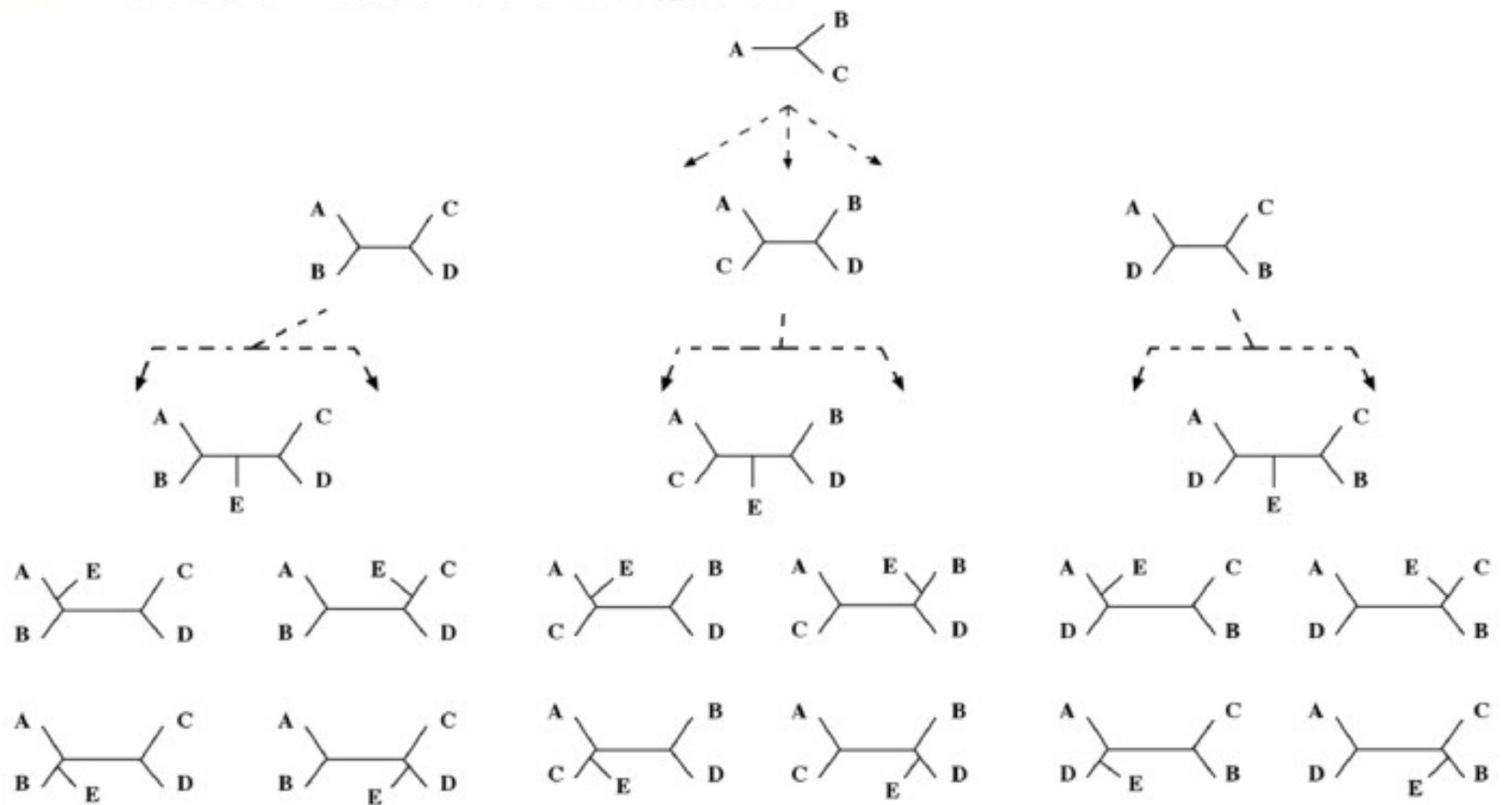
Métodos de búsqueda en árboles

- Como mencionamos el método de parsimonia examina todas las posibles topologías de árboles para encontrar el árbol con máxima parsimonia
- Este es un método exhaustivo de búsqueda que sigue los pasos siguientes:
 - ① Construye un árbol sin raíz con tres taxones tomados aleatoriamente (sólo existe una topología)
 - ② Agrega un cuarto taxón al árbol, produciendo 3 topologías posibles
 - ③ Agrega los taxones restantes progresivamente para formar todas las posibles topologías

Filogenia:

Métodos de búsqueda en árboles

- Método exhaustivo de búsqueda



Filogenia:

Métodos de búsqueda en árboles

- Obviamente este algoritmo de fuerza bruta sólo funciona para casos con pocas secuencias (menos de 10)
- La razón es que el número potencial de topologías de árboles puede enorme aún con un número moderado de taxones
- Recordemos que el número de árboles con raíz (N_R) para n taxones está dado por la siguiente fórmula:

$$N_R = (2n - 3)! / 2^{n-2}(n - 2)! \quad (3)$$

- Y el número de topologías para árboles sin raíz (N_U) es:

$$N_U = (2n - 5)! / 2^{n-3}(n - 3)! \quad (4)$$

Filogenia:

Métodos de búsqueda en árboles

- Para intentar solucionar esta problemática se han desarrollado algunas técnicas para reducir la complejidad de la búsqueda
- Un ejemplo es la técnica de *Branch & Bound* (B&B), la cual comienza construyendo un árbol basado en distancias con todos los taxones usando Unión de Vecinos o UPGMA
- Despues calcula el mínimo número de sustituciones para este árbol para usarlo como *cota superior* contra la cual son comparados todos los árboles
- La idea es que el árbol con máxima parsimonia debe ser igual o más pequeño que el árbol basado en distancias

Filogenia:

Métodos de búsqueda en árboles

- B&B construye árboles en una forma similar a la búsqueda exhaustiva
- La diferencia es que la cota superior precalculada es usada para limitar el crecimiento del espacio de búsqueda
- Cuando la longitud total de una topología parcial excede la cota superior, la búsqueda hacia esa dirección particular se aborta
- Esto reduce dramáticamente el número de árboles considerados (menos tiempo de cpu) mientras continua garantizando encontrar el árbol con MP

Filogenia:

Métodos de búsqueda en árboles

- Cuando el número de taxones excede 20, aún el método B&B se vuelve computacionalmente inviable
- La solución es usar *métodos heurísticos* (aproximados)
- En un método heurístico de búsqueda en árboles, sólo un pequeño subconjunto de todas las posibles topologías es examinado
- Comienza por calcular un árbol inicial mediante el método de Unión de Vecinos
- Continua modificandolo ligeramente para formar otra topología y analizar si este cambio lleva a un árbol con mayor parsimonia (más pequeño)

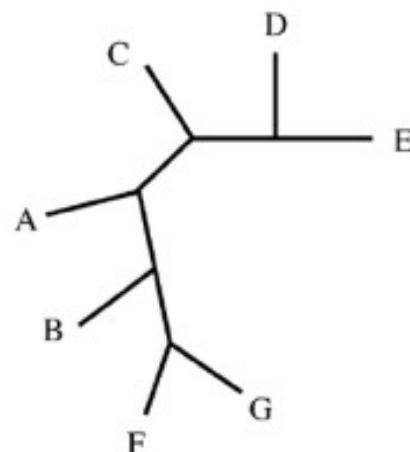
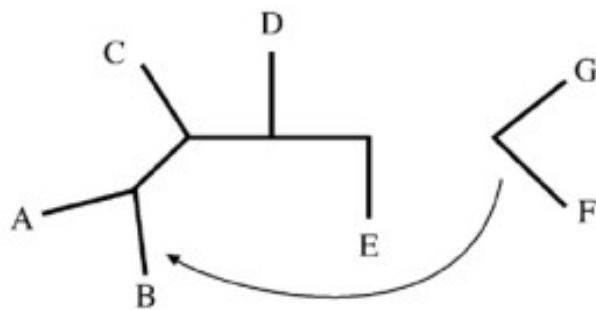
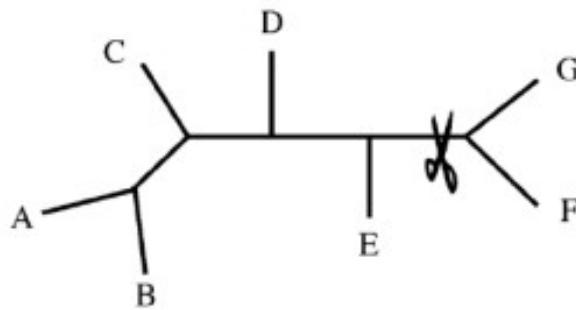
Métodos de búsqueda en árboles

- Los cambios ligeros aplicados al árbol inicial incluyen podar una rama o subárbol y pegarla en otra parte del árbol
- La longitud total del nuevo árbol es recalculada, si esta es más pequeña que la del árbol inicial entonces el nuevo árbol es usado como punto de partida para otra iteración
- Este proceso continua hasta que ningún árbol más pequeño es encontrado

Filogenia:

Métodos de búsqueda en árboles

- Método de heurísticos de intercambio de ramas



Filogenia:

Métodos de búsqueda en árboles

- Este método es muy rápido, pero no garantiza encontrar el árbol con MP
- Los algoritmos heurísticos de intercambio de ramas (*branch-swapping*) más comunes son:
 - Intercambio del vecino más cercano
 - Biseción del árbol y reconexión
 - Poda de un subárbol y pegado

Filogenia:

Métodos de búsqueda en árboles

- La desventaja de los algoritmos heurísticos de intercambio de ramas es que los nuevos árboles generados con cambios ligeros tienden a enfocarse en una área local
- Esto provoca que este tipo de algoritmos se estanquen cuando la longitud mínima de una rama local es alcanzada
- Para evitar que queden estancados en un mínimo local, una opción de búsqueda global es implementada en ciertos programas

Filogenia:

Métodos de búsqueda en árboles

- Esto permite remover todos los posibles subárboles y pegarlos en todas las posibles formas, para incrementar la oportunidad de encontrar el árbol con MP
- Este enfoque incrementa considerablemente el tiempo de cómputo y por lo tanto compromete el compromiso entre obtener un árbol óptimo y hacerlo en un tiempo razonable

Filogenia:

3

Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- Construcción del árbol filogenético con MP
- Planteamiento formal del problema de MP
- Trabajo relacionado
- Parsimonia ponderada
- Métodos de búsqueda en árboles
- **Ventajas y desventajas**
- Atracción de ramas largas
- Máxima verosimilitud
- Construcción del árbol filogenético con MV

Filogenia:

Ventajas y desventajas

Ventajas:

- El método de MP es intuitivo (fácil de comprender)
- Provee información evolutiva acerca de los caracteres en la secuencia (homoplasia y estados ancestros)
- Tiende a producir árboles más confiables que aquellos producidos con métodos basados en distancias cuando la divergencia entre secuencias es baja (suposición de parsimonia es cierta)

Filogenia:

Ventajas y desventajas

Desventajas:

- Cuando la divergencia entre secuencias es alta, o la cantidad de homoplasia es grande, la estimación de un árbol por MP puede ser menos efectiva (suposición de parsimonia no es cierta)
- La estimación de la longitud de las ramas puede también ser errónea porque MP no usa modelos de substitución para corregir substituciones múltiples (aumenta cuando las secuencias son demasiado divergentes)
- MP sólo considera sitios informativos, e ignora otros sitios con lo cual ciertas señales filogenéticas pueden perderse
- MP es lento comparado con los métodos basados en distancias y muy sensible a la **atracción de ramas largas** (LBA, *long-branch attraction*)

Filogenia:

3

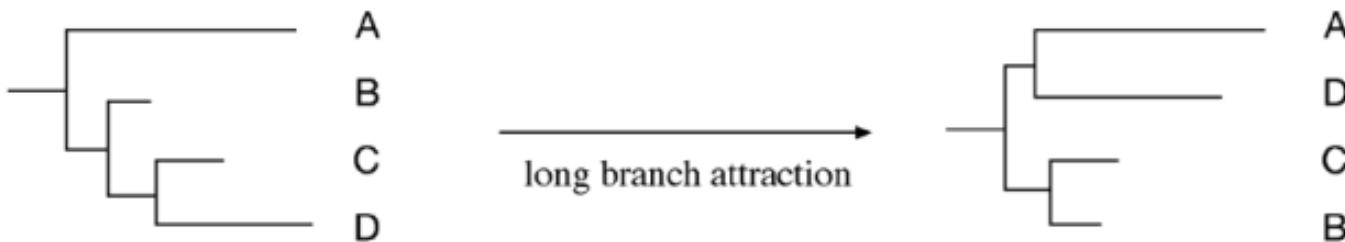
Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- Construcción del árbol filogenético con MP
- Planteamiento formal del problema de MP
- Trabajo relacionado
- Parsimonia ponderada
- Métodos de búsqueda en árboles
- Ventajas y desventajas
- **Atracción de ramas largas**
- Máxima verosimilitud
- Construcción del árbol filogenético con MV

Filogenia:

Atracción de ramas largas

- La atracción de ramas largas (LBA, *long-branch attraction*) es un problema particular asociado con los métodos de MP
- Se refiere a un fenómeno filogenético por el cual los taxones que evolucionan rápidamente con ramas largas son colocados juntos en un árbol, sin importar sus posiciones verdaderas en el árbol



Filogenia:

Atracción de ramas largas

- Esto se debe parcialmente a la suposición en parsimonia que indica que todos los linajes evolucionan a la misma tasa y que todas las mutaciones (transiciones y transversiones) contribuyen de igual manera a la longitud de las ramas
- También se debe a que las substituciones múltiples en los sitios individuales y entre sitios tienen una tasa de heterogeneidad que MP no es capaz de corregir

Filogenia:

Atracción de ramas largas

- Existen diversas posibles soluciones al problema de atracción de ramas largas:
 - Para homoplasias utilizar métodos basados en distancias y en Máxima Verosimilitud (próxima clase) que emplean modelos de substitución y modelos de tasa de heterogeneidad
 - Usar parsimonia ponderada que permite luchar contra las desviaciones de las transiciones cuando se producen las transiciones más a menudo que las transversiones
 - Aumentar el tamaño de muestreo de taxones también puede ayudar, porque la introducción de taxones intermedios rompe las ramas largas

Filogenia:

3

Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- Construcción del árbol filogenético con MP
- Planteamiento formal del problema de MP
- Trabajo relacionado
- Parsimonia ponderada
- Métodos de búsqueda en árboles
- Ventajas y desventajas
- Atracción de ramas largas
- Máxima verosimilitud**
- Construcción del árbol filogenético con MV

Filogenia:

Máxima verosimilitud

- El método de *Máxima Verosimilitud* (MV) emplea modelos probabilísticos para seleccionar el mejor árbol, i.e., aquel que tenga la más alta probabilidad (verosimilitud) de reflejar el proceso evolutivo real
- MV es un método exhaustivo que busca todas las posibles topologías y considera cada posición en un alineamiento (no sólo sitios informativos)

Filogenia:

Máxima verosimilitud

- Empleando un modelo particular de substitución de residuos ML calcula la verosimilitud total de las secuencias ancestro que evolucionan en nodos internos y eventualmente a las secuencias existentes (nodos hoja)
- En ocasiones también incorpora parámetros que consideran las tasas de variación entre sitios

Filogenia:

3

Métodos basados en caracteres

- Introducción
- Máxima parsimonia
- Construcción del árbol filogenético con MP
- Planteamiento formal del problema de MP
- Trabajo relacionado
- Parsimonia ponderada
- Métodos de búsqueda en árboles
- Ventajas y desventajas
- Atracción de ramas largas
- Máxima verosimilitud
- **Construcción del árbol filogenético con MV**

Filogenia:

Construcción del árbol filogenético con MV

- MV trabaja calculando la probabilidad de un determinado camino evolutivo para una secuencia particular existente
- Los valores de probabilidad son determinados por un modelo de substitución
- Por ejemplo, para secuencias de ADN usando el modelo Jukes-Cantor, la probabilidad P de que un nucleótido permanezca igual después de un tiempo t es:

$$P(t) = 1/4 + 3/4e^{-\alpha t} \quad (5)$$

- donde α es la tasa de substitución del nucleótido en el modelo Jukes-Cantor (asignada empíricamente o estimada experimentalmente)



Filogenia:

Construcción del árbol filogenético con MV

- Por el contrario para un nucleótido que cambia a un residuo diferente después de un tiempo t , la probabilidad P es:

$$P(t) = 1/4 + 1/4e^{-\alpha t} \quad (6)$$

- Para otros modelos de substitución, las fórmulas son mucho más complejas

Filogenia: Maxima Similitud

Dado un modelo probabilistico para la sustitucion de nucleotidos (o proteinas) p.ej. Jukes & Cantor, escoger el árbol que tiene la más alta probabilidad de generar la data observada.

Ej: Dada la Data D y el modelo M, encontrar e árbol T que cumple la probabilidad $P(D| T,M)$ es la máxima.

El Modelo da los valores de $p_{ij}(t)$, es decir , la probabilidad de pasar del nucleotido (i) a (j) en el tiempo (t)

Filogenia: Maxima Similitud

Se hacen 2 asunciones de independencia:

- Diferentes sitios evolucionan en forma independiente
- Secuencias o especies divergentes evolucionan independientemente despues de divergir.

Ej: Sea D_i la data para el sitio (i), entonces

$$P(D| T,M) = \prod_i P(D_i | T,M)$$

Filogenia: Maxima Similitud

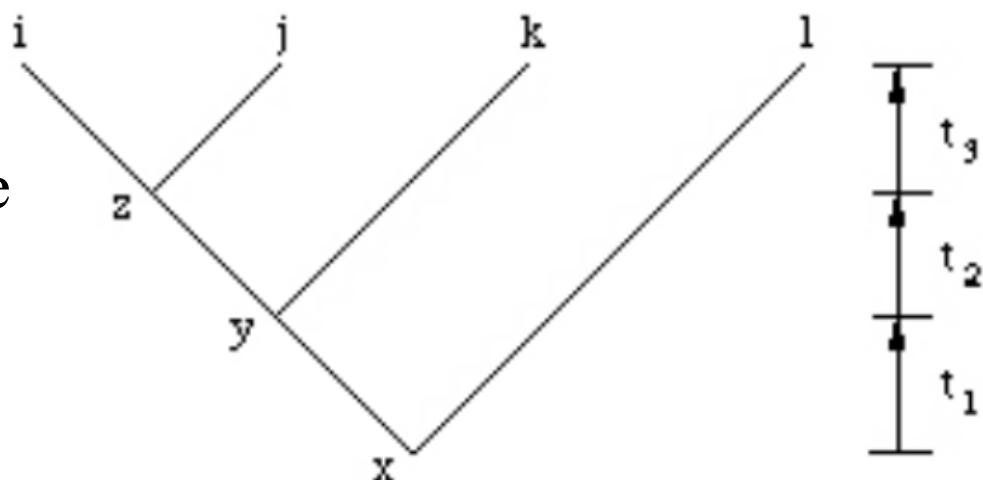
Como se calcula $P(D_i | T, M)$?

$p_{xy}(t)$ – es la probabilidad de pasar de x a y en el tiempo t

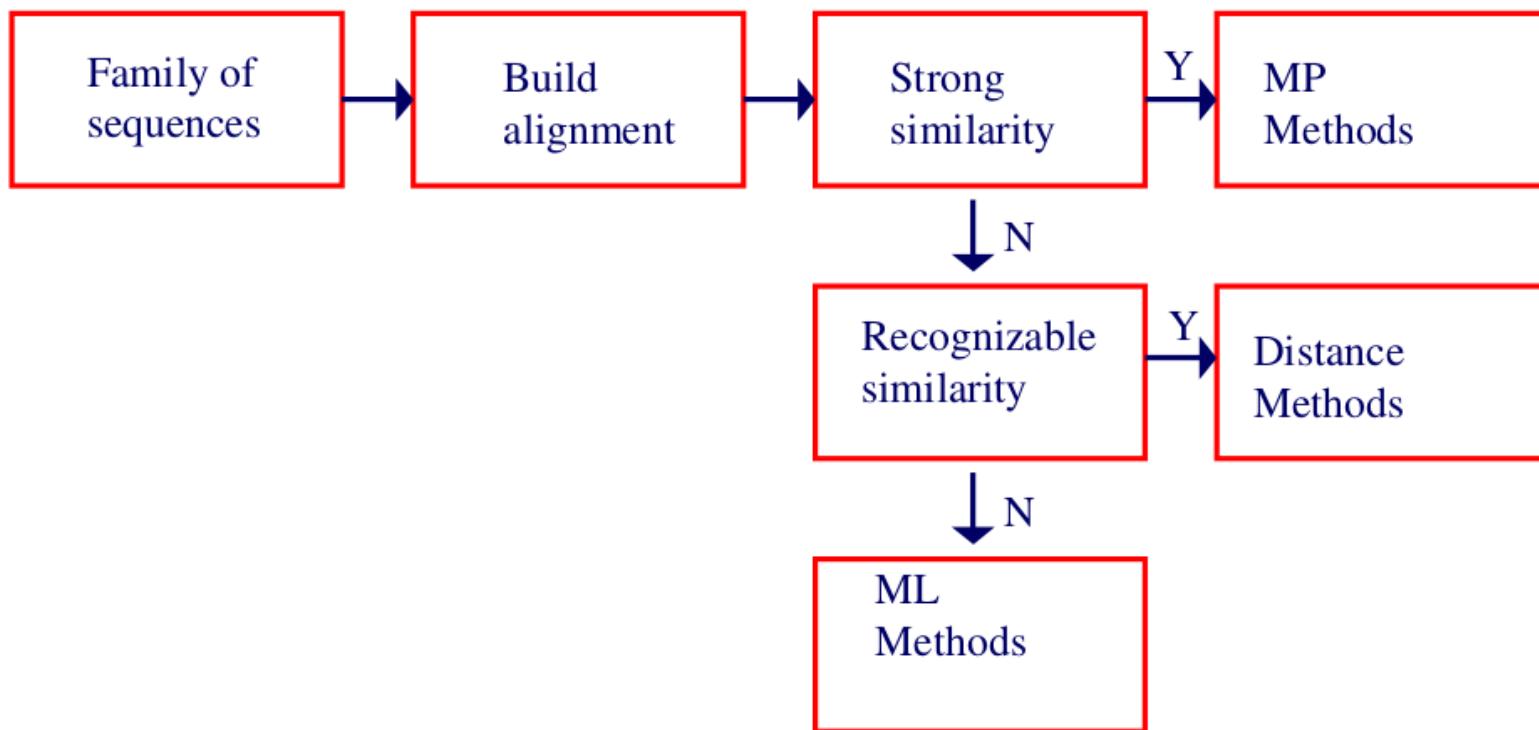
$$P(i,j,k,l | T, M) =$$

$$\sum_x \sum_y \sum_z p(x) (p_{xl}(t_1 + t_2 + t_3) p_{xy}(t_1) p_{yk}(t_2 + t_3))$$

$$p_{yz}(t_2) p_{zi}(t_3) p_{zj}(t_3))$$



Filogenia: Diagrama de Flujo



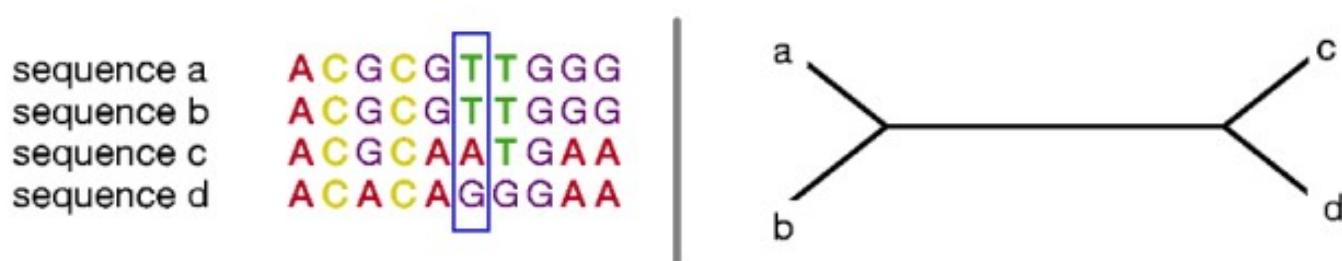
MP – Maximum parsimony

ML - Maximum Likelihood

Filogenia: Diagrama de Flujo

Construcción del árbol filogenético con MV

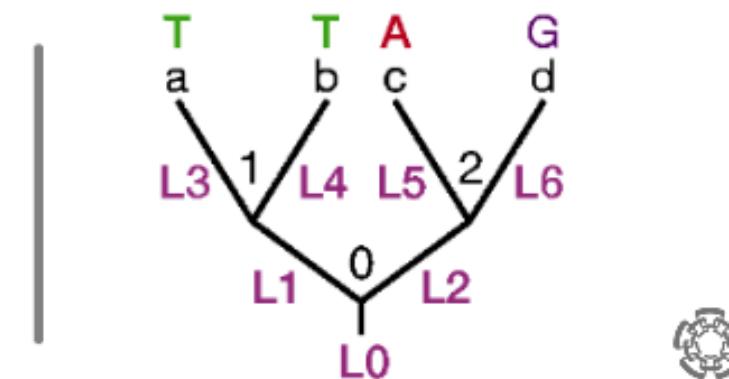
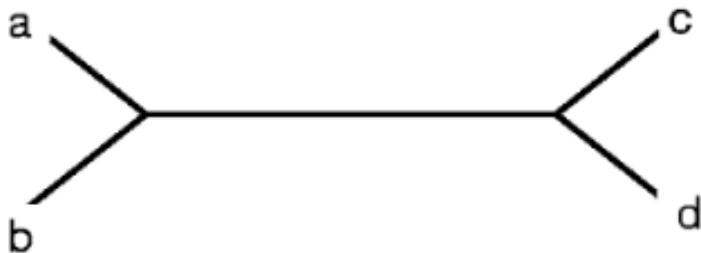
- Veamos un ejemplo más detallado con 4 secuencias hipotéticas
- Para cuatro taxones existen 3 posibles árboles sin raíz (se muestra uno)
- Se toma una columna para analizarse



Filogenia: Diagrama de Flujo

Construcción del árbol filogenético con MV

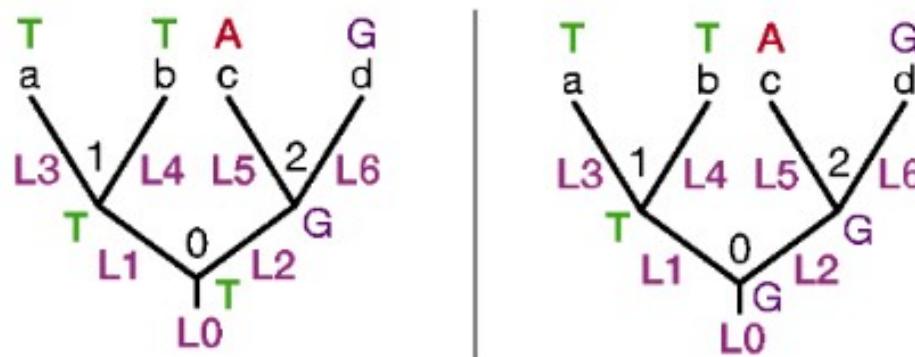
- Una de las 5 posibles topologías con raíz derivadas del árbol anterior (contiene 3 nodos internos 0, 1, 2)
- Se consideran todas las posibles asignaciones de bases para los nodos internos ($4 * 4 * 4 = 64$) y se calcula la verosimilitud para cada una ($L_1 - L_6$)
- La verosimilitud de esta topología es: $V(T_1) = \prod_{i=1}^6 L_i$ (se pueden usar sumas de logaritmos \ln)



Filogenia: Diagrama de Flujo

Construcción del árbol filogenético con MV

- 2 de las 64 posibles asignaciones de bases para los nodos internos
- $V(6) = V(T_1) + V(T_2) + \cdots + V(T_{64})$
- Estos cálculos son repetidos para todas las columnas del alineamiento, $V(1) \dots V(10)$



Filogenia: Diagrama de Flujo

Construcción del árbol filogenético con MV

- La verosimilitud de la topología es la suma de las verosimilitudes calculadas para cada columna, $\sum_{i=1}^{10} V(i)$
- Cada una de las 3 posibles topologías (para 4 taxones) es evaluada de manera similar y se identifica aquella con la máxima verosimilitud
- Como puede verse este proceso es muy demandante en tiempo de cómputo

Filogenia: Diagrama de Flujo

Referencias bibliográficas I

-  Andreatta, A. and Ribeiro, C. C. (2002).
Heuristics for the phylogeny problem.
Journal of Heuristics, 8(4):429–447.
-  Barker, D. (2003).
LVB: parsimony and simulated annealing in the search for phylogenetic trees.
Bioinformatics, 20(2):274–275.
-  Barker, D. (2012).
LVB homepage.
-  Garey, M. R. and Johnson, D. S. (1977).
The rectilinear Steiner tree problem is NP-Complete.
SIAM Journal on Applied Mathematics, 32(4):826–834.
-  Hendy, M. D. and Penny, D. (1982).
Branch and bound algorithms to determine minimal evolutionary trees.
Mathematical Biosciences, 59(2):277–290.
-  Ribeiro, C. C. and Vianna, D. S. (2005).
A GRASP/VND heuristic for the phylogeny problem using a new neighborhood structure.
International Transactions in Operational Research, 12(3):325–338.
-  Ribeiro, C. C. and Vianna, D. S. (2009).
A hybrid genetic algorithm for the phylogeny problem using path-relinking as a progressive crossover strategy.
International Transactions in Operational Research, 16(5):641–657.

Filogenia: Diagrama de Flujo

Referencias bibliográficas II



Richer, J. M., Goëffon, A., and Hao, J. K. (2009).
A memetic algorithm for phylogenetic reconstruction with maximum parsimony.
Lecture Notes in Computer Science, 5483:164–175.



Richer, J. M., Rodriguez-Tello, E., and Vazquez-Ortiz, K. E. (2012).
Maximum parsimony phylogenetic inference using simulated annealing.
Advances in Intelligent and Soft Computing, 175:189–203.

Filogenia: Software adicional.

Software adicional:

http://king2.sc.fsu.edu/CEBProjects/awty/awty_start.php

AWTY es un sistema para la exploracion gráfica de convergencias de cadenas de Markov Montecarlo (MCMC) en métodos Bayesianos de inferencia filogenética.

<http://www.megasoftware.net/>

Provee herramientas para explorar, descubrir y analizar secuencias de ADN y proteínas desde una perspectiva evolutiva.

Filogenia:

EJEMPLO

Filogenia: Ejemplo UPGMA

Primer paso:

Dados 5 elementos (a,b,c,d,e) y la siguiente matriz de distancias entre ellos D_1

| | a | b | c | d | e |
|---|----|----|----|----|----|
| a | 0 | 17 | 21 | 31 | 23 |
| b | 17 | 0 | 30 | 34 | 21 |
| c | 21 | 30 | 0 | 28 | 39 |
| d | 31 | 34 | 28 | 0 | 43 |
| e | 23 | 21 | 39 | 43 | 0 |

En este ejemplo $D_1(a,b)=17$ es el menor valor de D_1 , así que unimos los elementos a y b.

Sea u el nodo al cual a y b están ahora conectados. Sea $\delta(a,u) = \delta(b,u) = D_1(a,b)/2$ que asegura que los elementos a y b son equidistantes de u. Esto corresponde a la expectación de que se cumple la hipótesis de ultrametricidad.. Las ramas que unen a y b a u entonces tienen la longitud $\delta(a,u) = \delta(b,u) = 17/2=8.5$

Filogenia: Ejemplo UPGMA

Entonces procedemos a actualizar la matriz de distancias inicial D_1 en una nueva matriz de distancias D_2 , reducida en tamaño en una fila y en una columna a causa del clustering de a y b. Los valores en negrita en D_2 corresponden a las nuevas distancias calculadas al promediar las distancias entre el primer cluster (a,b) y cada uno de los elementos resultantes:

$$D_2((a,b),c) = (D_1(a,c) \times 1 + D_1(b,c) \times 1) / (1+1) = (21+30)/2 = 25.5$$

$$D_2((a,b),d) = (D_1(a,d) + D_1(b,d)) / 2 = (31+34)/2 = 32.5$$

$$D_2((a,b),e) = (D_1(a,e) + D_1(b,e)) / 2 = (23+21)/2 = 22$$

| | (a,b) | c | d | e |
|-------|-------------|-------------|-------------|-----------|
| (a,b) | 0 | 25.5 | 32.5 | 22 |
| c | 25.5 | 0 | 28 | 39 |
| d | 32.5 | 28 | 0 | 43 |
| e | 22 | 39 | 43 | 0 |

Filogenia: Ejemplo UPGMA

Segundo paso: Repetir el procedimiento con D_2

Aquí, $D_2((a,b), e) = 22$ es el menor valor de D_2 , así que unimos el cluster (a,b) y el elemento e .

| | (a,b) | c | d | e |
|-------|-------|------|------|----|
| (a,b) | 0 | 25.5 | 32.5 | 22 |
| c | 25.5 | 0 | 28 | 39 |
| d | 32.5 | 28 | 0 | 43 |
| e | 22 | 39 | 43 | 0 |

Sea v el nodo al cual (a,b) y e están ahora conectados. Sea $\delta(a,v) = \delta(b,v) = \delta(e,v) = 22/2 = 11$ (por la restricción de ultrametricidad)

Deducimos la longitud de la rama que falta:

$$\delta(u,v) = \delta(e,v) - \delta(a,u) = \delta(e,v) - \delta(b,u) = 11 - 8.5 = 2.5$$

Procedemos a actualizar la matriz de distancias

Filogenia: Ejemplo UPGMA

| | (a,b) | c | d | e |
|-------|-------|------|------|----|
| (a,b) | 0 | 25.5 | 32.5 | 22 |
| c | 25.5 | 0 | 28 | 39 |
| d | 32.5 | 28 | 0 | 43 |
| e | 22 | 39 | 43 | 0 |

$$D_3(((a,b),e),c) = (D_2((a,b),c) \times 2 + D_2(e,c) \times 1) / (2 + 1) = (25.5 \times 2 + 39 \times 1) / 3 = 30$$

$$D_3(((a,b),e),d) = (D_2((a,b),d) \times 2 + D_2(e,d) \times 1) / (2 + 1) = (32.5 \times 2 + 43 \times 1) / 3 = 36$$

| | ((a,b),e) | c | d |
|-----------|-----------|----|----|
| ((a,b),e) | 0 | 30 | 36 |
| c | 30 | 0 | 28 |
| d | 36 | 28 | 0 |

Filogenia: Ejemplo UPGMA

| | ((a,b),e) | c | d |
|-----------|-----------|----|----|
| ((a,b),e) | 0 | 30 | 36 |
| c | 30 | 0 | 28 |
| d | 36 | 28 | 0 |

$$D_4((c,d), ((a,b), e)) = (D_3(c, ((a,b), e)) \times 1 + D_3(d, ((a,b), e)) \times 1) / (1 + 1) = (30 \times 1 + 36 \times 1) / 2 = 33$$

$$\delta(((a,b), e), r) = \delta((c,d), r) = 33/2 = 16.5$$

| | ((a,b),e) | (c,d) |
|-----------|-----------|-------|
| ((a,b),e) | 0 | 33 |
| (c,d) | 33 | 0 |

$$\delta(v, r) = \delta(((a,b), e), r) - \delta(e, v) = 16.5 - 11 = 5.5$$

$$\delta(w, r) = \delta((c,d), r) - \delta(c, w) = 16.5 - 14 = 2.5$$

