Filogenética molecular (III)

Bioinformática, 5-4-19

Basado en Kevin Yip-CSE-CUHK
(Universidad china de Hong-Kong)

HOY …

1. Distancia evolutiva y modelos de mutación
2. Árboles: Las estructuras jerárquicas relacionando diferentes objetos biológicos
    1. Formatos de archivo
    2. reconstrucción de árboles filogenéticos
    3. Métodos basados en secuencias
        - máxima parsimonia
        - Máxima verosimilitud
    4. **métodos basados en distancias**
        - UPGMA
        - Unión de vecinos

Motivación

- En los anteriores algoritmos basados en secuencias, las secuencias exactas se utilizan en la reconstrucción de los árboles filogenéticos
- En un método basado en distancia, sólo se consideran las distancias entre pares de secuencias
  - Es bueno si las secuencias son largas, y sólo nos preocupa la estructura de árbol, pero no las secuencias ancestrales
  - Las distancias se pueden calcular por métodos basados en la alineación de secuencias (pero hay distancias calculables sin alineamiento)
  - Una vez que las distancias por pares se han calculado, no se utilizarán las secuencias originales

Distancias basadas en compresión

- Tenemos un algoritmo de compresión, sea |compresión(x)| la longitud del output con entrada x
- El **ratio de compresión** es $C(x)=\frac{|\text{compresión}(x)|}{|x|}$

$$0 \leq C(x) \leq 1$$

- ¿Para saber la distancia de un string x a un string y?
  - Comparamos C(xy) con C(x) y C(y)

Distancias basadas en compresión

- Hay varias opciones
- $NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$
- $CBD(x,y) = \frac{2\,C(xy)}{C(x) + C(y)} - 1$
- Cálculo de la distancia tan eficiente como el algoritmo de compresión

Construcción de árboles filogenéticos

- Dos métodos basados en distancias:
  - UPGMA
  - Unión de vecinos

UPGMA

- Unweighted Pair Group Method with Arithmetic Mean
- Algoritmo:
  1. Calcular la distancia entre cada par de secuencias
  2. Tratar a cada secuencia como un grupo por sí mismo
  3. Combinar los dos grupos más cercanos. La distancia entre dos grupos es la distancia media entre todas sus secuencias (excepto que d (C_i, C_i) = 0):

     $$\mathrm{d}(\mathrm{C_i}, \mathrm{C_j}) = \frac{1}{|\mathrm{C_i}||\mathrm{C_j}|} \sum_{r \in \mathrm{C_i}, s \in \mathrm{C_j}} \mathrm{d}(r, s)$$

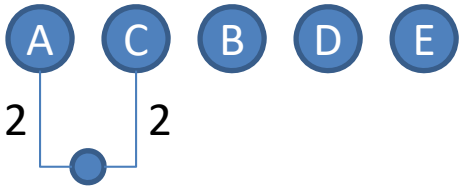
     (Nótese que d (*r*,*s*) es la distancia entre r y s en la matriz de distancia de entrada)
  4. Repetir 2 y 3 hasta que sólo quede un cluster

Ejemplo

	A	B	C	D	E
A	0	8	4	6	8
B	8	0	8	8	4
C	4	8	0	6	8
D	6	8	6	0	8
E	8	4	8	8	0

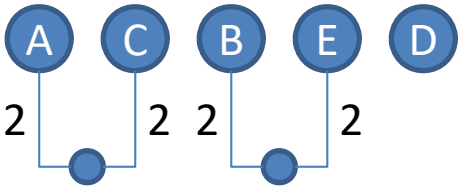
{A},{C}

	A,C	B	D	E
A,C	0	8	6	8
B	8	0	8	4
D	6	8	0	8
E	8	4	8	0



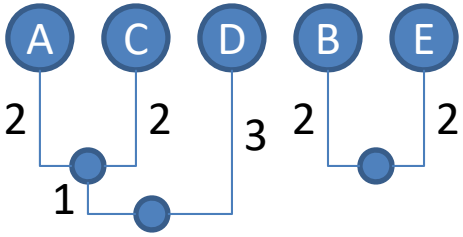
{B},{E}

	A,C	B,E	D
A,C	0	8	6
B,E	8	0	8
D	6	8	0



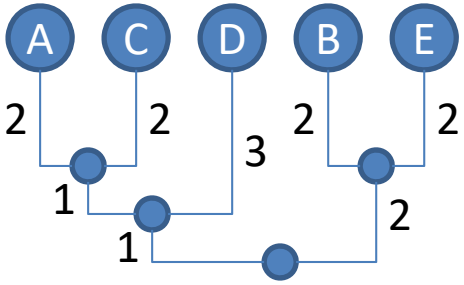
{A,C}, {D}

	A,C,D	B,E
A,C,D	0	8
B,E	8	0



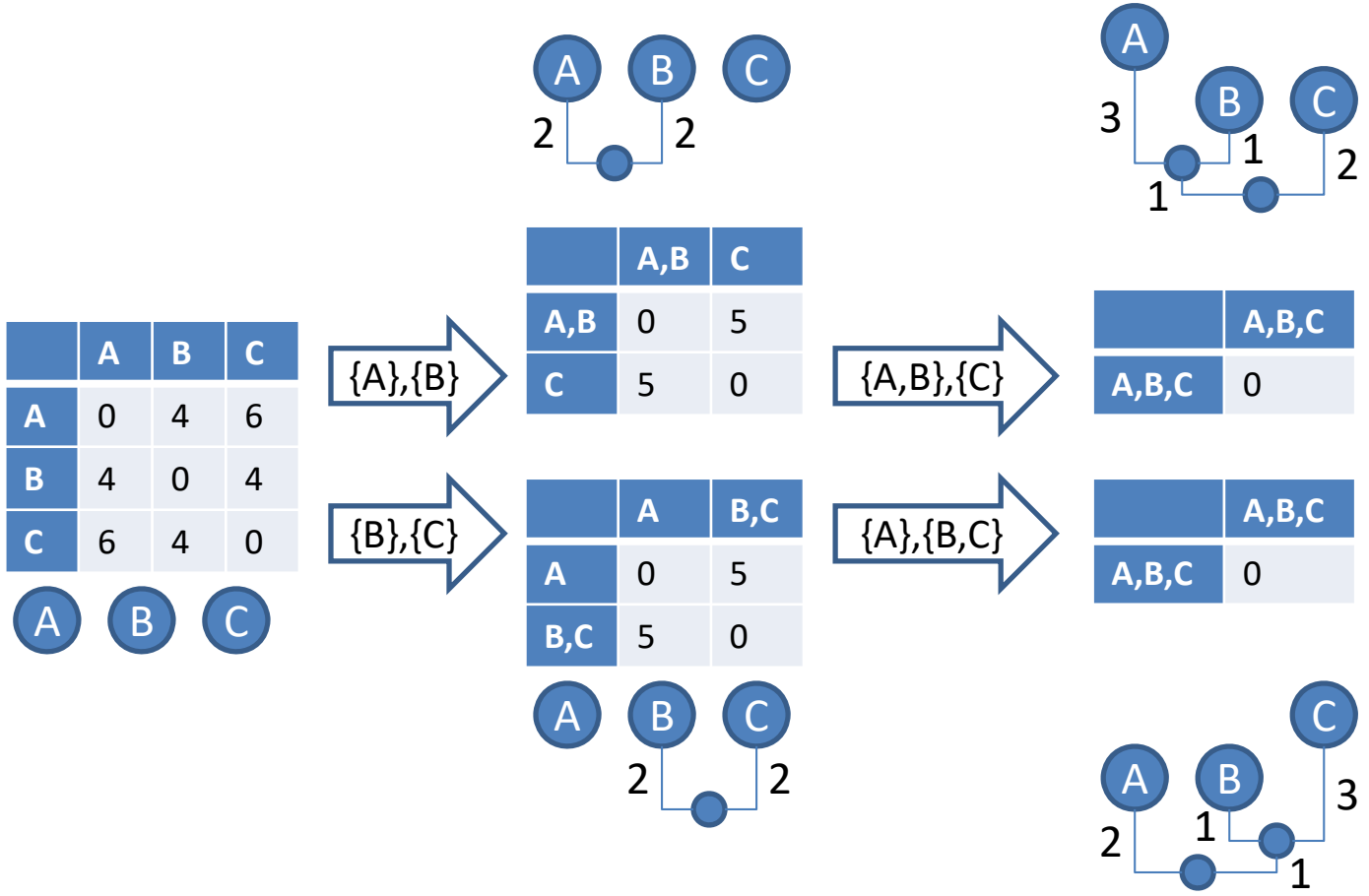
{A,C,D}, {B,E}

	A,B,C,D,E
A,B,C,D,E	0



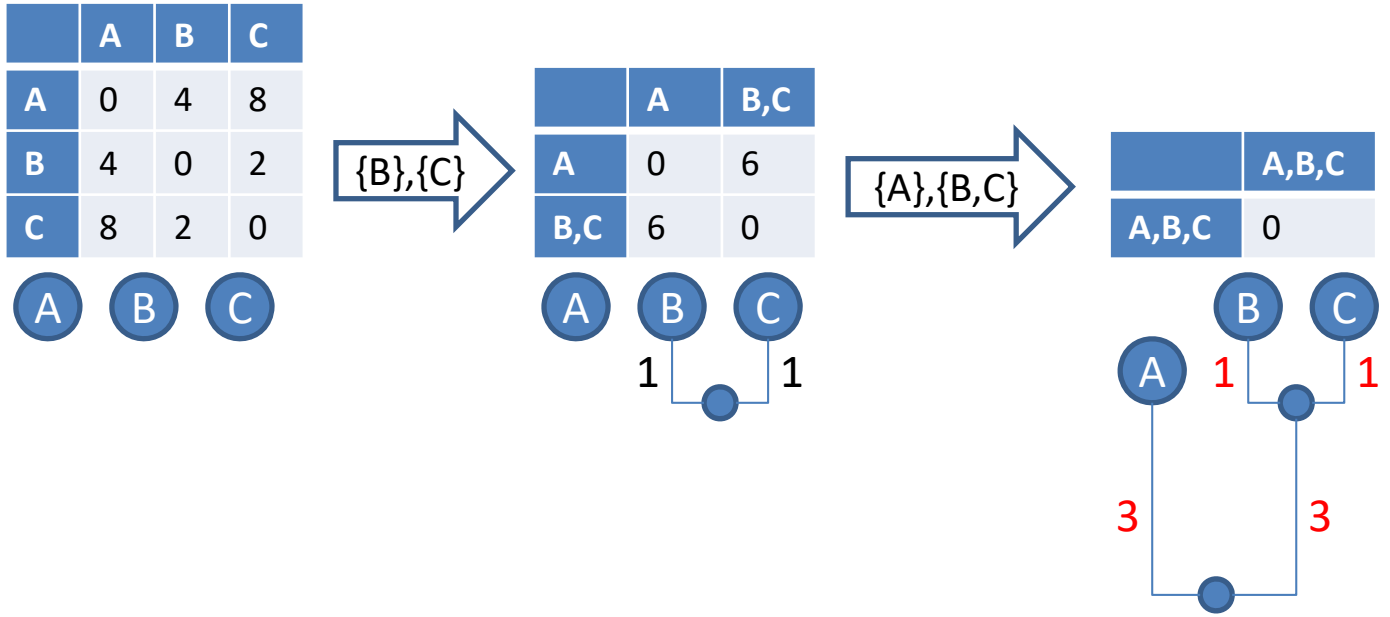
Empates, hojas no alineadas

- No siempre es posible poner todas las hojas en una línea:



Longitudes de rama

- No siempre es posible asignar longitudes de rama de acuerdo a las distancias:



Aquí las brach lengths sólo reflejan las distancias de clúster, no las distancias de secuencia

HOY …

1. Distancia evolutiva y modelos de mutación
2. Árboles: Las estructuras jerárquicas relacionando diferentes objetos biológicos
    1. Formatos de archivo
    2. reconstrucción de árboles filogenéticos
    3. Métodos basados en secuencias
        - máxima parsimonia
        - Máxima verosimilitud
    4. métodos basados en distancia
        - UPGMA
        - **Unión de vecinos**

Unión de vecinos

- En UPGMA, cada vez que se fusionan los dos grupos más cercanos según su distancia (criterio nº 1):

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{r \in C_i, s \in C_j} d(r, s)$$

- Sería bueno elegir el grupo que también está muy lejos de otros grupos (criterio nº 2), medidos por:

$$u(C_i) = \sum_j d(C_i, C_j)$$

Unión de vecinos

- En el algoritmo unión de vecinos, los dos grupos que fusionar es el par que minimiza

$$Q(i,j) = (r-2)d(C_i, C_j) - u(C_i) - u(C_j),$$

donde r es el número actual de grupos

(y Q (i, i) ≡ 0 para todo i)

  - La fórmula considera ambos criterios, mientras que el factor (r-2) es para equilibrar sus pesos relativos

Unión de vecinos

- El algoritmo:
  1. Comienza con cada secuencia como un clúster. Todos ellos están conectados a un centro, formando una estrella.
  2. Encontrar grupos i y j conectados al centro donde Q (i, j) es mínimo entre todos los pares de cluster
  3. Insertar un nuevo nodo interno C_k

     (que minimiza Q desde k a (i,j))

Unión de vecinos

2. Encontrar grupos i y j conectados al centro donde Q (i, j) es mínimo entre todos los pares de cluster
3. Insertar un nuevo nodo interno C_k(que minimiza Q)
   - Conectarlo a C_i, C_j y el centro
   - Asignar la longitud $\frac{d(C_i, C_j)}{2} + \frac{u(C_i) - u(C_j)}{2(r-2)}$ a C_iC_k
   - Asignar la longitud $\frac{d(C_i, C_j)}{2} + \frac{u(C_j) - u(C_i)}{2(r-2)}$ a C_jC_k
   - Para cada nodo C_l, $d(C_k, C_l) = [d(C_i, C_l) + d(C_j, C_l) - d(C_i, C_j)] / 2$ (todos los $d(C_x, C_y)$ son valores ya calculados).
4. Repetir 2 y 3 hasta que todas las longitudes de rama se asignen
   - El resultado final será un árbol sin raíces

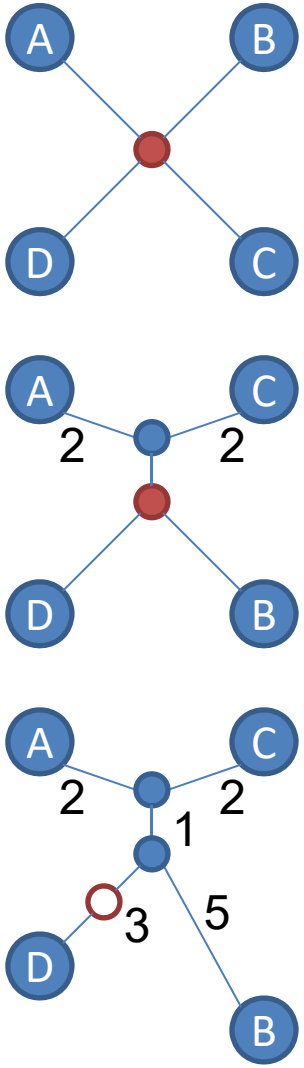
Ejemplo

$Q(i,j) = (r-2)d(C_i,C_j) - u(C_i) - u(C$

Distancia entre A y el nuevo nodo: d(A,C)/2 + [u(A) – u(C)] / [2(r-2)] = 4/2 + (18-18) / [2(2)] = 2

{A}, {C}

{A,C}, {B}



d	A	B	C	D
A	0	8	4	6
B	8	0	8	8
C	4	8	0	6
D	6	8	6	0

u	
A	18
B	24
C	18
D	20

Q	A	B	C	D
A	0	-26	-28	-26
B	-26	0	-26	-28
C	-28	-26	0	-26
D	-26	-28	-26	0

d	A,C	B	D
A,C	0	6	4
B	6	0	8
D	4	8	0

u	
A,C	10
B	14
D	12

Q	A,C	B	D
A,C	0	-18	-18
B	-18	0	-18
D	-18	-18	0

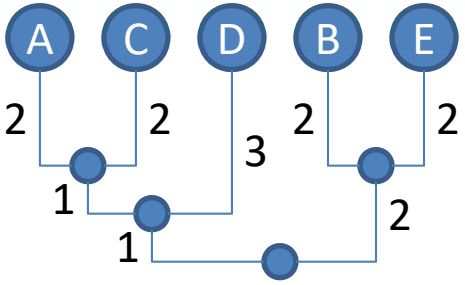
d	A,B,C	D
A,B,C	0	3
D	3	0

u	
A,B,C	3
B	3

En el último paso, quitamos el centro y escribimos la distancia

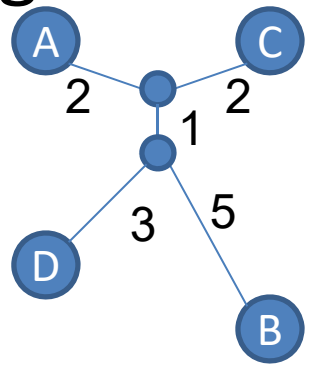
Comparando los resultados

- UPGMA: (con un nodo más, E)



Más sencillo

- Neighbor Joining:



Mayor posibilidad de distancias significativas

Resumen de distancias

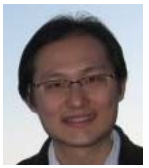
- A partir de una **matriz de distancias** se construye el árbol
- No es necesario conocer la topología
- Tampoco está claro que necesite bootstraps
- La distancia puede ser con o sin alineamiento

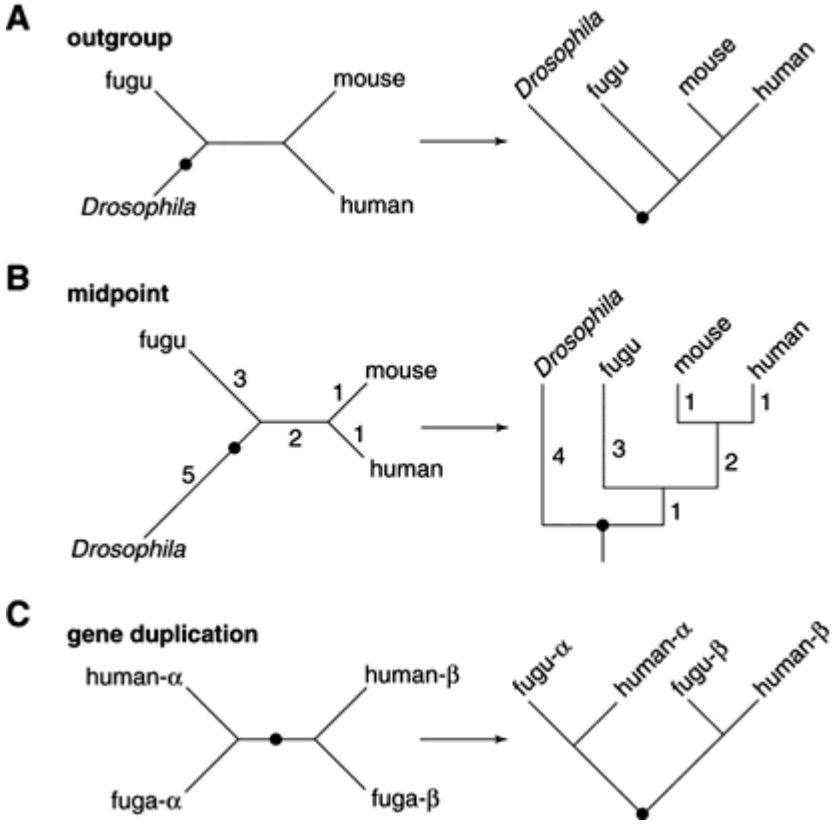
Enraizamiento de un árbol sin raíces

- ¿Cómo encontrar la raíz de un árbol?
  - Por lo general, mediante el uso de un **valor externo**, algo que se ha separado antes
  - Hay algunos otros métodos

outlier





Crédito de la imagen: Wikipedia, http://blog.ohinternet.com/wp-content/uploads/2011/03/fugu.jpg, http://www.currentprotocols.com/protocol/bi0601

observaciones

- Los diferentes tipos de DNA tienen diferentes tasas de mutación
  - árbol de genes frente a árbol de especies
- Algunos DNAs no se heredan según la ruta habitual
  - Por ejemplo, las bacterias pueden adquirir nuevo DNA tomado de plásmido (transferencia horizontal de genes)
  - Necesitamos un grafo filogenético general que permita múltiples padres y bucles
- La reconstrucción de árboles filogenéticos se beneficiaría de tener una alineación de secuencias múltiples precisa, y viceversa
  - Algunos métodos realizan los dos iterativamente

¿Qué método utilizar?

- No hay una respuesta definitiva
  - Hay diferentes campos

Los biólogos prefieren máx., verosimilitud

- En general, es bueno utilizar métodos que
  - No requieran hipótesis fuertes
  - Sean robustos (no producen resultados drásticamente diferentes cuando las entradas se cambian sólo un poco)
    - Construir múltiples árboles utilizando diferentes parámetros, a continuación, combinar
    - Construir árboles con diferentes subconjuntos de secuencias, a continuación, combinar
    - Utilizar métodos probabilísticos
  - computacionalmente eficientes
- Hay muchos otros algoritmos que no cubrimos, incluidos aquellos que consideran modelos de mutación.

Resumen

- Los modelos de mutación nos permiten estimar formalmente el número de mutaciones ocurridas basado en los datos observados
  - modelo Jukes-Cantor de un parámetro
- Los arboles filogenéticos capturan eventos de separación y cuándo sucedieron

Resumen

- Existen dos tipos principales de métodos de reconstrucción del árbol:
  - Basados en secuencia
    - máxima parsimonia
    - máxima verosimilitud
  - basados en distancias
    - UPGMA
    - Unión de vecinos
- Necesitamos métodos escalables, robustos y biológicamente significativos

Problema muy abierto

Epílogo

CASO DE ESTUDIO

Caso de estudio: las clasificaciones inesperadas

- En los viejos tiempos, los biólogos clasificaban las especies en función de sus características de alto nivel
  - Si una especie posee características que la hacen similar a varios tipos de especies, puede ser difícil clasificarla
  - Cuando se disponga de características moleculares (por ejemplo, secuencias de DNA), se pueden utilizar para clasificar las especies de una manera **sistemática**
    - Se ha descubierto que algunas clasificaciones anteriores son incompatibles con la evidencia molecular

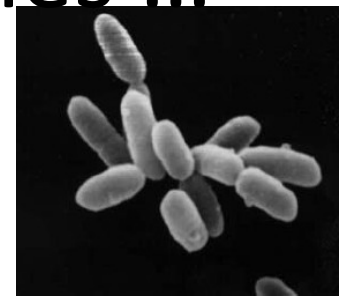
Caso de estudio: clasificaciones …

- Ejemplo 1: Mamíferos
  - Los murciélagos parecen pájaros, los delfines se ven como los peces, pero ambos son en realidad mamíferos
- Reino: Animalia (animales)
  Superphylum: Deuterostomados
  Filo: Chordata
  Subphylum: Vertebrata (animales con columna vertebral)
  Infraphylum: Gnathostomata (vertebrados con mandíbulas)
  Clase: Chondrichthyes (peces cartilaginosos)
  Superclase: Osteichthyes (peces óseos)
  Superclase: Tetrapoda (cuatro ramificac vertebrados)
  Clase: Aves (aves)
  Clase Mammalia (mamíferos)

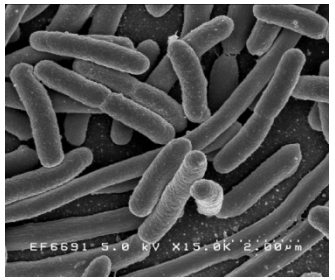
Fuente de la imagen: Wikipedia

Caso de estudio: clasificaciones …

- Ejemplo 2: Los tres dominios
- Todas las especies de la tierra pertenecen a uno de los tres dominios
  - arqueas
    - Unicelulares, sin núcleo
    - Por lo general, viven en lugares con condiciones extremas (por ejemplo, alta temperatura o la salinidad - "extremófilos")
  - Las bacterias
    - Unicelulares, sin núcleo
  - eucariotas
    - Muchos son multicelular, con núcleo



Halobacterias sp. cepa NRC-1, un archaaeon



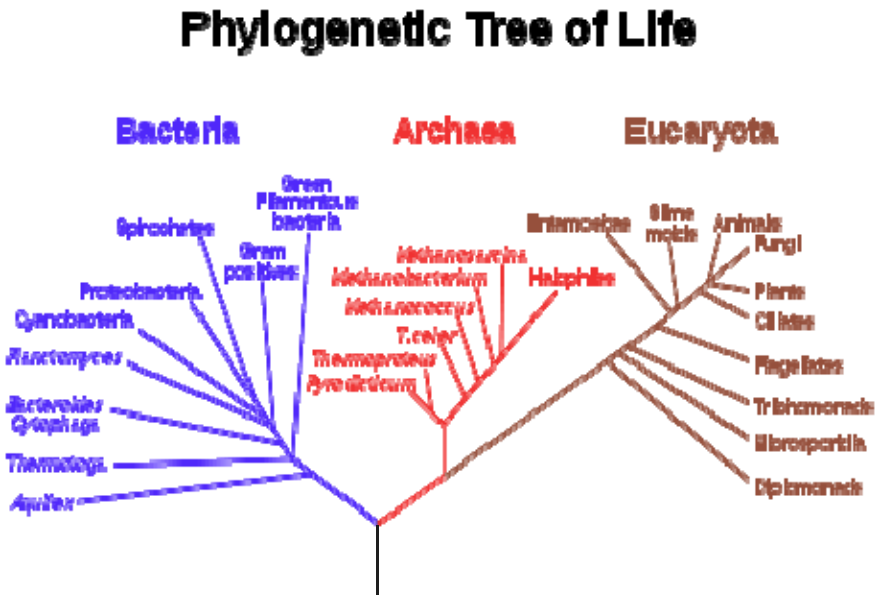
Escherichia coli, una beacterium

Varias especies eucariotas

Fuente de la imagen: Wikipedia

Caso de estudio: clasificaciones …

- Parece razonable suponer que los eucariotas se separaron primero de los otros dos
- Sin embargo, basándose en la secuencia de RNA ribosomal, algo tan importante que evoluciona lentamente, las arqueas están más cerca de los eucariotas que las bacterias



Fuente de la imagen: Wikipedia