Modelos de mutación y filogenética molecular

Bioinformática, 20-3-19

Basado en

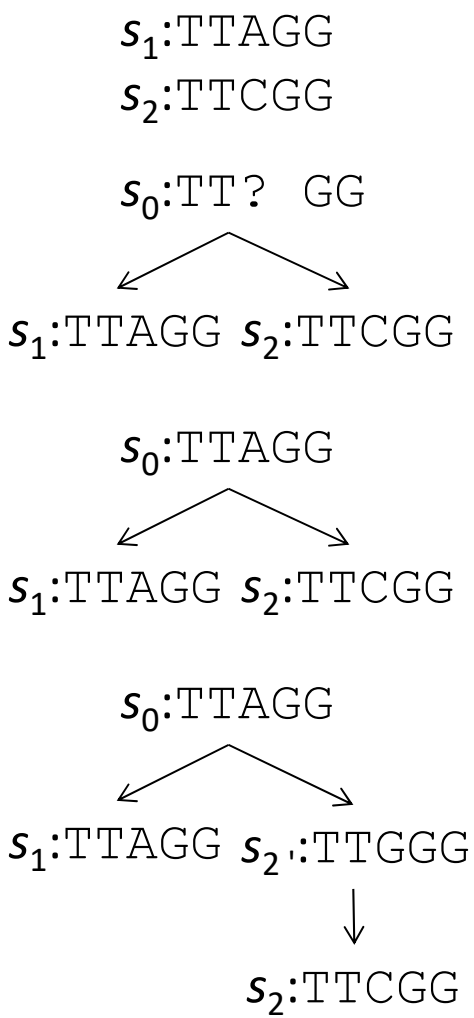
Kevin Yip-CSE-CUHK
(Universidad china de Hong-Kong)

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 distancia
      - UPGMA
      - Unión de vecinos

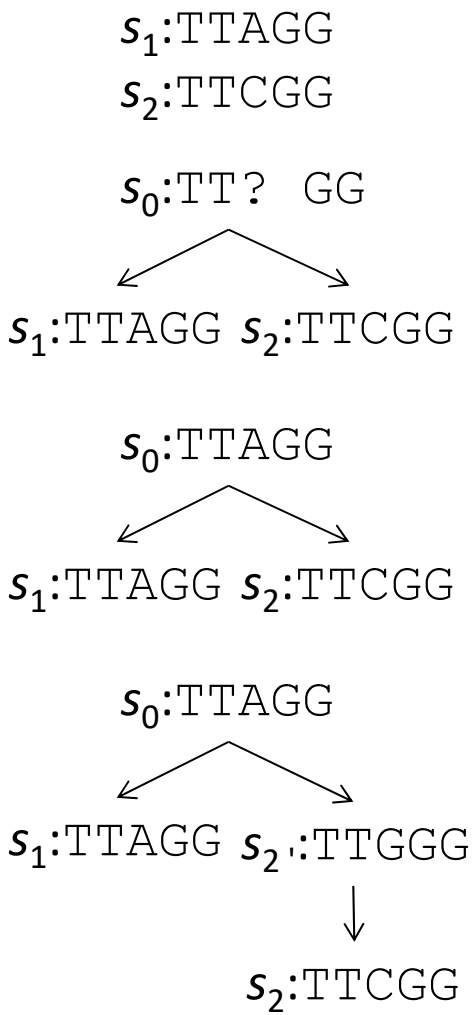
distancia evolutiva

- Supongamos que tenemos una alineación de dos secuencias. En una posición, una secuencia tiene A y otra tiene C.
  - Supongamos que las secuencias tienen un ancestro común (por ejemplo son la misma proteína en ratones y ratas)
  - ¿Qué tenía el antepasado común en esa posición?
    - No sabemos.
  - Digamos A. ¿Cuántas sustituciones han ocurrido desde su divergencia?
    - Podría ser una (s_0[3]:A $\rightarrow$ s_2[3]:C)
    - Podría haber dos (por ejemplo, s_0[3]:A $\rightarrow$ s_2'[3]:G $\rightarrow$ s_2[3]:C)
    - Podrían ser más… ¿tres?



distancia evolutiva

- Queremos una manera de definir la "distancia evolutiva" entre dos secuencias observadas de acuerdo con el número de mutaciones sucedido
  - ¿Cómo podemos definirla sin saber lo que ha sucedido?
  - Necesitamos estimarla usando un *modelo de mutaciones*



modelo de mutación

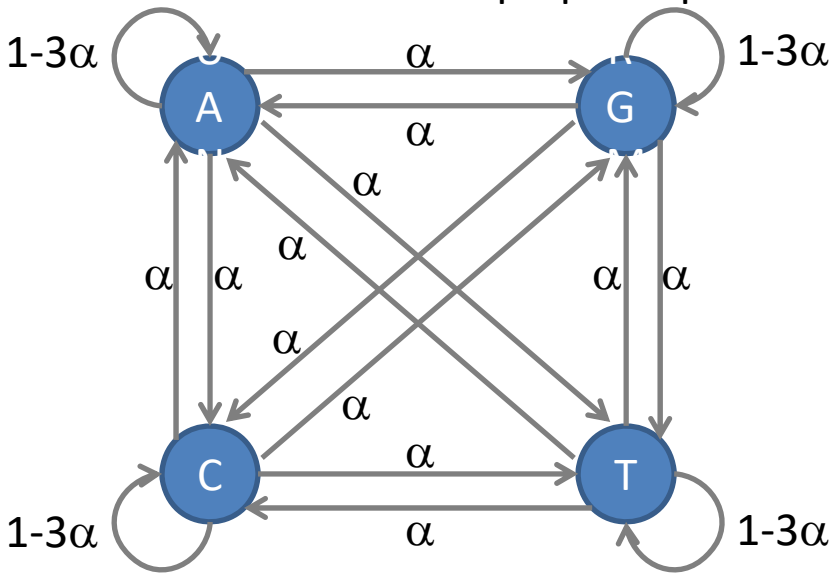
- **Un modelo de mutación (o modelo de evolución)** es un modelo probabilístico que describe cómo ocurren las mutaciones a través del tiempo
  - ¿Con qué frecuencia ocurre una mutación?
  - ¿Qué tipo de mutaciones son más frecuentes?

modelo de mutación

- Para simplificar las cosas, vamos a hacer las siguientes suposiciones:
  1. Las posiciones son independientes
  2. Las tasas de mutación son las mismas para diferentes posiciones y en diferentes momentos de la historia
  3. Teniendo en cuenta el estado actual, los estados futuros no dependen de los estados pasados
- Sabemos que estos supuestos por lo general no son ciertos, pero sin ellos los cálculos pueden ser difíciles
  - Existen modelos más complejos que requieren un menor número de supuestos fuertes. Hoy sólo estudiaremos modelos simples …

El modelo Jukes-Cantor

- Propuesto por Jukes y Cantor en 1969
- Tasa de sustitución, α, idéntica para las otras tres bases **en una unidad de tiempo**
  - Suponemos que hay como mucho una mutación por unidad de tiempo - Siempre podemos hacer la unidad más pequeña para asegurar esto



Nota: Una "sustitución" es una mutación puntual que realmente ocurrió, mientras que una falta de coincidencia en una alineación podría ser causado por una o más sustituciones

Cómo funciona el modelo Jukes-Cantor

- Supongamos que en el momento 0, la posición 1 de una secuencia era A
- En el momento 1:
  - Hay una probabilidad de $1 - 3\alpha$ de que la posición sea A
  - Hay una probabilidad de α de que la posición sea C
  - Hay una probabilidad de α de que la posición sea G
  - Hay una probabilidad de α de que la posición sea T
- En el momento 2, ¿cuál es la probabilidad de que la posición sea A si sólo se sabe que era A en el tiempo 0, pero no se sabe lo que era en el momento 1?
  - Dos posibilidades:
    1. En el momento 1 era A, y no hubo ninguna mutación del instante 1 al 2 [probabilidad: $(1-3\alpha)^2$]
    2. En el momento 1, era C,G ó T, y hubo una mutación a A del instante 1 al 2 [probabilidad: $3\alpha^2$]
  - Por lo tanto, la probabilidad total de que la posición sea A en el momento 2 es $(1-3\alpha)^2 + 3\alpha^2$

fórmulas recursivas

- $P_{X \to Y}(t)$ es la probabilidad de que una base que era X en el instante 0, sea Y en el instante t
  - $P_{A \to A}(1) = 1 - 3\alpha$
  - $P_{A \to A}(2) = (1 - 3\alpha)^2 + 3\alpha^2$
  - En general,
    $P_{A \to A}(t+1) = (1 - 3\alpha)P_{A \to A}(t) + \alpha\,[1 - P_{A \to A}(t)]$
  - En general:
    - $P_{X \to X}(t+1) = (1 - 3\alpha)P_{X \to X}(t) + \alpha\,[1 - P_{X \to X}(t)]$ para cualquier X
    - $P_{X \to Y}(t+1) = [1 - P_{X \to X}(t+1)] / 3$ para cualquier $X \neq Y$

fórmulas recursivas

- En primer lugar, estudiamos cómo calcular $P_{X \to Y}(t)$ para un ratio de mutación α y tiempo de divergencia t
- después estudiamos cómo podemos utilizar $P_{X \to Y}(t)$ para estimar el número de mutaciones que han ocurrido desde la divergencia de las dos secuencias

Solución de $P_{A \to A}(t)$

- $P_{A \to A}(t+1) = (1 - 3\alpha)P_{A \to A}(t) + \alpha\,[1 - P_{A \to A}(t)]$
- $\Delta\, P_{A \to A}(t)$

  $\equiv P_{A \to A}(t+1) - P_{A \to A}(t)$

  $= (1 - 3\alpha)P_{A \to A}(t) + \alpha\,[1 - P_{A \to A}(t)] - P_{A \to A}(t)$

  $= \alpha\,[1-4\,P_{A \to A}(t)]$

Solución de $P_{A \to A}(t)$

- $\Delta P_{A \to A}(t) = \alpha [1\text{-}4 P_{A \to A}(t)]$
- Para una unidad de tiempo infinitamente pequeña, se obtiene una ecuación diferencial de primer orden, que puede ser resuelta mediante el uso de un factor de integración

$$\begin{aligned} \frac{dP_{A \to A}(t)}{dt} &= \alpha[1 - 4P_{A \to A}(t)] \\ \Longrightarrow \quad P_{A \to A}(t) &= \frac{1}{4} + \left(P_{A \to A}(0) - \frac{1}{4}\right) e^{-4\alpha t} \\ &= \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \end{aligned}$$

  - Observación: Cuando *t* es grande, el estado inicial (es decir, nucleótido) ya no importa y las cuatro bases son igualmente probables

fórmulas finales para $\mathrm{P}_{\mathrm{X}\to\mathrm{Y}}(t)$

- Por simetría,

$$\mathrm{P}_{\mathrm{A}\to\mathrm{A}}(t) = \mathrm{P}_{\mathrm{C}\to\mathrm{C}}(t) = \mathrm{P}_{\mathrm{G}\to\mathrm{G}}(t) = \mathrm{P}_{\mathrm{T}\to\mathrm{T}}(t) = \frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4\alpha t}$$

- Del mismo modo, es fácil demostrar que

$$\mathrm{P}_{\mathrm{A}\to\mathrm{C}}(t) = \mathrm{P}_{\mathrm{A}\to\mathrm{G}}(t) = \mathrm{P}_{\mathrm{A}\to\mathrm{T}}(t) = \cdots = \mathrm{P}_{\mathrm{T}\to\mathrm{G}}(t) = \left[1 - \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4\alpha t}\right)\right]/3 = \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4\alpha t}$$

Volviendo al problema

- ¿Qué hemos hecho hasta ahora?
  - Dado
    1. El estado ancestral de un posición
    2. La tasa de sustitución α (probabilidad de cada tipo de mutación en una unidad de tiempo)
    3. El tiempo t
  - Determinar la probabilidad del estado actual, que es *t* unidades de tiempo después del evento de separación, donde *t* también se da

Volviendo al problema

- ¿Qué es lo que realmente queremos?
  - Teniendo en cuenta el estado actual de dos secuencias
    - El estado ancestral se desconoce
  - Determinar el número de sustituciones que sucedió en las dos secuencias desde su divergencia, tanto observadas como no observadas
    - La tasa de sustitución (α) y hace cuánto tiempo que las dos secuencias han divergido (t) son también desconocidos

Las dificultades y las ideas

- Dificultad # 1: No sabemos el estado ancestral, tasa de mutación α ó tiempo de divergencia t
  - soluciones:
    - Debido a la simetría, el estado ancestral no importa si sólo nos preocupa si dos secuencias actuales tienen el mismo nucleótido o no en cada posición
      - $P_{mismo}(t)$
        $= [P_{A\to A}(t)]^2 + [P_{A\to C}(t)]^2 + [P_{A\to G}(t)]^2 + [P_{A\to T}(t)]^2$
        $= [P_{C\to A}(t)]^2 + [P_{C\to C}(t)]^2 + [P_{C\to G}(t)]^2 + [P_{C\to T}(t)]^2$
        $= [P_{G\to A}(t)]^2 + [P_{G\to C}(t)]^2 + [P_{G\to G}(t)]^2 + [P_{G\to T}(t)]^2$
        $= [P_{T\to A}(t)]^2 + [P_{T\to C}(t)]^2 + [P_{T\to G}(t)]^2 + [P_{T\to T}(t)]^2$
    - No conocemos α ó t, pero se puede estimar fácilmente su producto αt y resulta que esto es todo lo que necesitamos

Las dificultades y las ideas

- Dificultad # 2: Aunque conociéramos el estado ancestral, la tasa de mutación α y el tiempo de divergencia *t*, todavía habría un número infinito de posibilidades, cada una con una cierta probabilidad
  - Solución:
    - Vamos a hablar del *número esperado* de mutaciones ocurridas, es decir, el promedio de todos los casos considerando el número de mutaciones de cada uno y la probabilidad de que ocurra

¿Cómo ayudan esperanza y varianza en la solución de nuestro problema?

- No sabemos si una mutación aparece después de una unidad de tiempo, pero si tenemos en cuenta un gran número de unidades de tiempo, el número **esperado** de mutaciones después de t unidades de tiempo es $3\alpha t$
- Si podemos estimar el número de mutaciones que han sucedido, podemos calcular $3\alpha t$ aunque no conozcamos α y t
  - Por eso utilizamos la fórmula de aproximación para $P_{A \to A}(t) = 3/4\ e^{-4\alpha t} + 1/4$ sólo usa el producto αt

Estimando el número de sustituciones

- Para el modelo de Jukes-Cantor:
  - Para una sola posición, la probabilidad de que dos secuencias separadas hace *t* unidades de tiempo tengan el mismo estado es (estado ancestral $\mathtt{A}$, la misma fórmula es válida para otros estados ancestrales):

Idea # 1

$$
\begin{aligned}
& [\mathrm{P}_{\mathtt{A}\to\mathtt{A}}(t)]^2 + [\mathrm{P}_{\mathtt{A}\to\mathtt{C}}(t)]^2 + [\mathrm{P}_{\mathtt{A}\to\mathtt{G}}(t)]^2 + [\mathrm{P}_{\mathtt{A}\to\mathtt{T}}(t)]^2 \\
= \quad & \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4\alpha t}\right)^2 + 3\left(\frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4\alpha t}\right)^2 \\
= \quad & \left(\frac{1}{16} + \frac{6}{16}\mathrm{e}^{-4\alpha t} + \frac{9}{16}\mathrm{e}^{-8\alpha t}\right) + \left(\frac{3}{16} - \frac{6}{16}\mathrm{e}^{-4\alpha t} + \frac{3}{16}\mathrm{e}^{-8\alpha t}\right) \\
= \quad & \frac{1}{4} + \frac{3}{4}\mathrm{e}^{-8\alpha t}
\end{aligned}
$$

Estimando el número de sustituciones

- Para el modelo de Jukes-Cantor:
  - …
  - En consecuencia, la probabilidad de que las dos secuencias tengan diferentes estados en una posición es

    Idea # 2

$$p_{\text{diff}} \equiv 1 - \left(\frac{1}{4} + \frac{3}{4}e^{-8\alpha t}\right) = \frac{3}{4}(1 - e^{-8\alpha t})$$

$$\Rightarrow \quad \alpha t = -\frac{1}{8}\ln\left(1 - \frac{4}{3}p_{\text{diff}}\right)$$

No sabemos el valor de α (tasa de sustitución) o t (tiempo desde la divergencia de las dos secuencias), pero podemos estimar p_{diff}, lo que nos dará una estimación de αt.

Estimando el número de sustituciones

- Cómo estimar p_{diff}, la probabilidad de que dos secuencias aleatorias generadas según el procedimiento anterior tengan diferentes estados en una posición?
  - Estimamos p_{diff} por x/n, Donde x es el número de posiciones diferentes entre las secuencias observadas - nuestra mejor suposición basada en datos observados

Estimando el número de sustituciones

- Poniendo todo junto:
  - Supongamos que tenemos dos secuencias de longitud *n* que divergieron hace tiempo *t* con *x* diferencias
  - Sea K_{sup} el no. de sustituciones por posición ocurridas para las dos secuencias desde su divergencia
  - Según el modelo de Jukes-Cantor, el valor esperado de K_{sup} es (de la página anterior)

$$\mathrm{E}[\mathrm{K_{sup}}] = 2(3\alpha t) = 6\left[-\frac{1}{8}\ln\left(1-\frac{4}{3}\mathrm{p_{diff}}\right)\right] = -\frac{3}{4}\ln\left(1-\frac{4}{3}\mathrm{p_{diff}}\right) = -\frac{3}{4}\ln\left(1-\frac{4x}{3n}\right)$$

  - para *n* grande, la varianza de esta estimación es de aproximadamente

$$\frac{\mathrm{p_{diff}} - (\mathrm{p_{diff}})^2}{n\left(1-\frac{4}{3}\mathrm{p_{diff}}\right)^2} = \frac{x/n - (x/n)^2}{n\left(1-\frac{4x}{3n}\right)^2}$$

Estimando el número de sustituciones

- Vamos a ver cómo podemos aplicar los resultados.
- Ejemplo:
  - Supongamos que dos secuencias, cada una con *n*= 200 nucleótidos tienen *x*= 66 diferencias observadas, luego
    - $p_{diff} = x/n = 66/200 = 0{,}33$
    - $E[K_{sup}] = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_{diff}\right) \approx 0.43$
    - La varianza de esta estimación es $\frac{p_{diff} - (p_{diff})^2}{n\left(1 - \frac{4}{3}p_{diff}\right)^2} \approx 0.0035$
  - observaciones:
    1. El número observado de sustituciones por posición es menor que el número estimado de sustituciones (observadas + no observadas) por posición, como se esperaba
    2. La varianza es bastante grande - el número real puede ser un poco diferente de esta estimación (sería más pequeña para *n* grande)

Comprobación con nuestros datos de simulación

- Suponer α = 0,005, *t*= 50 y *n*= 100
- Dos copias de la misma secuencia aleatoria se generaron en *t*= 0, seguido de mutaciones independientes de acuerdo con el modelo de Jukes-Cantor
- Resultados de la simulación de 10 ensayos:

Trial	0	1	2	3	4	5	6	7	8	9
x	67	56	61	69	61	62	66	66	59	68
pdiff	0.67	0.56	0.61	0.69	0.61	0.62	0.66	0.66	0.59	0.68
E[Ksup] (from formula)	1.68	1.03	1.26	1.89	1.26	1.31	1.59	1.59	1.16	1.78
Variance of E[Ksup] estimation (from formula)	0.19	0.04	0.07	0.33	0.07	0.08	0.16	0.16	0.05	0.25
Actual number of substitutions per site	1.68	1.29	1.23	1.74	1.38	1.43	1.57	1.31	1.47	1.38
Actual variance of 10 trials	0.03									

distancia evolutiva

- Ahora, podemos usar E [K_{sup}] como una medida de la distancia evolutiva entre dos secuencias alineadas sin indeles.
  - Utilizar un modelo de mutación para definir la distancia es más robusto que la definición de la matriz de puntuación de sustitución arbitraria (como coincidencia = 1, diferencia = -1).
  - Por supuesto, tenemos que alinear las secuencias en primer lugar, momento en el que todavía necesitamos una matriz de sustitución para empezar

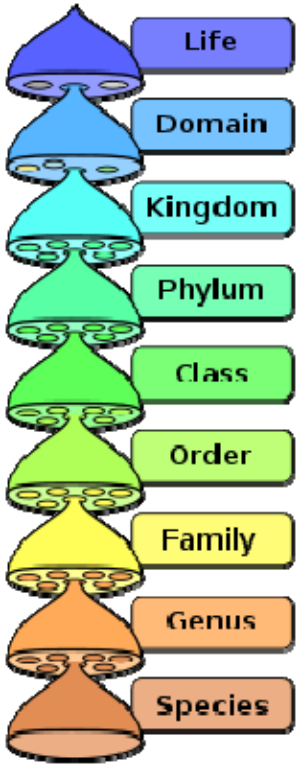
distancia evolutiva

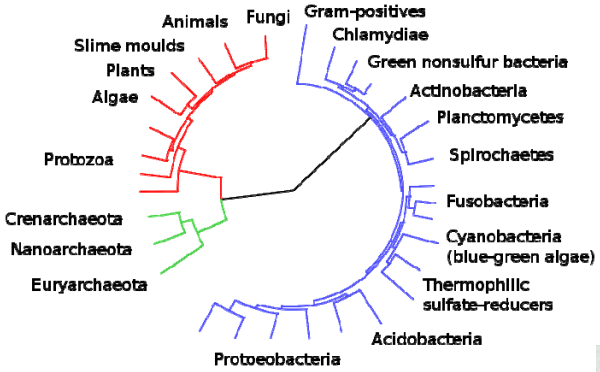
- Hay otros modelos que permiten
  - Más parámetros (por ejemplo, diferentes subtipos de sustituciones)
  - Tasas variables en diferentes posiciones
  - Dependencia entre diferentes posiciones
  - Cambio de las tasas de sustitución con el tiempo
  - indeles

HOY …

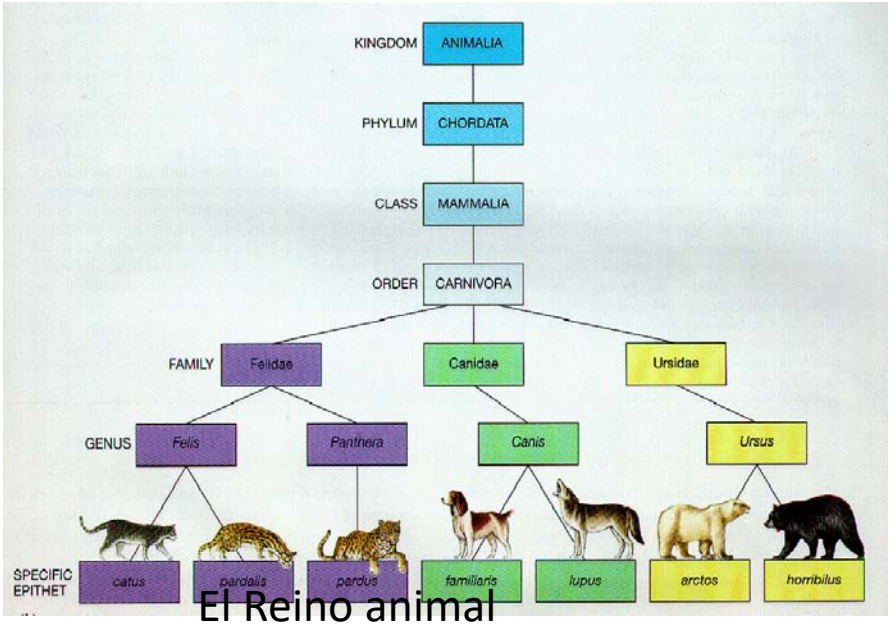
1. Distancia evolutiva y modelos de mutación
2. **Árboles: Las estructuras jerárquicas relacionando diferentes objetos biológicos**
    1. Formatos de archivo
    2. reconstrucción de árboles filogenéticos
    3. Métodos basados en secuencias
        - máxima parsimonia
        - Máxima verosimilitud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

Clasificación de las especies





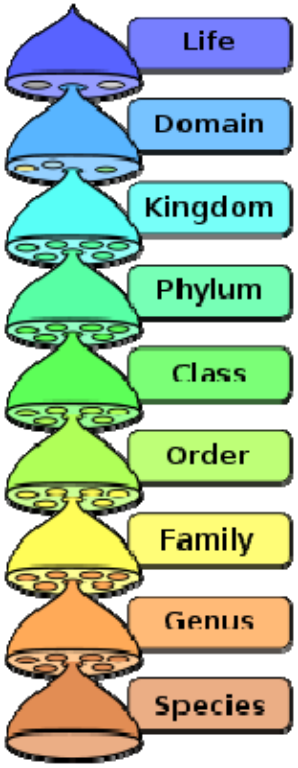
Dominios y reinos

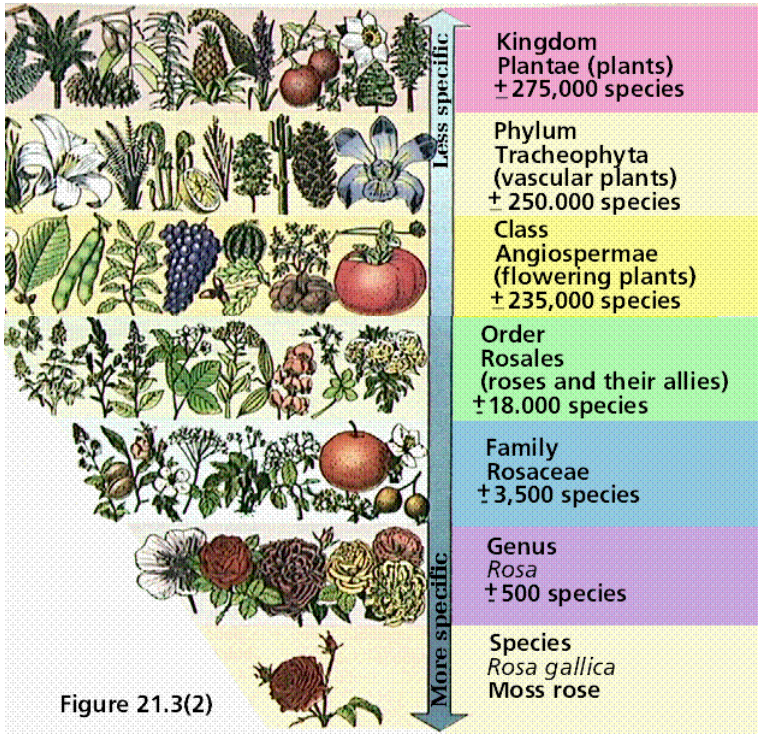


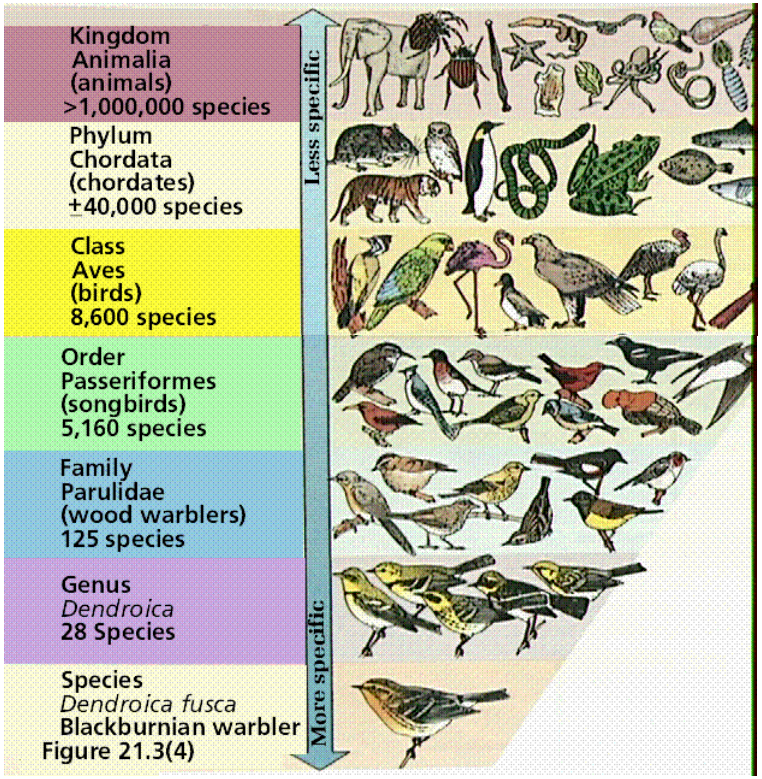
El Reino animal

Crédito de la imagen: Wikipedia, http://ridge.icu.ac.jp/gen-ed/classif-gifs/animal-class-example.gif

taxonomía





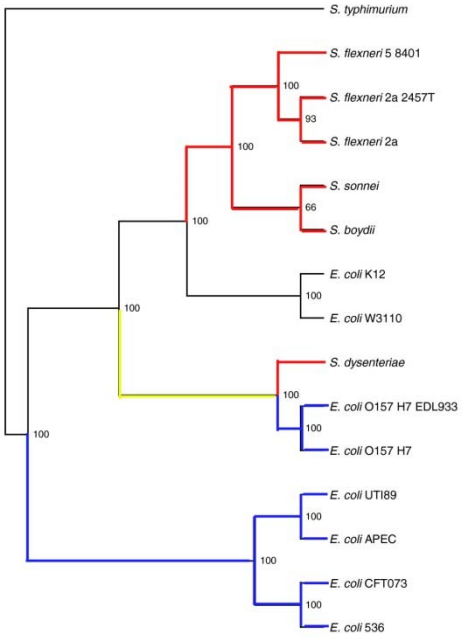


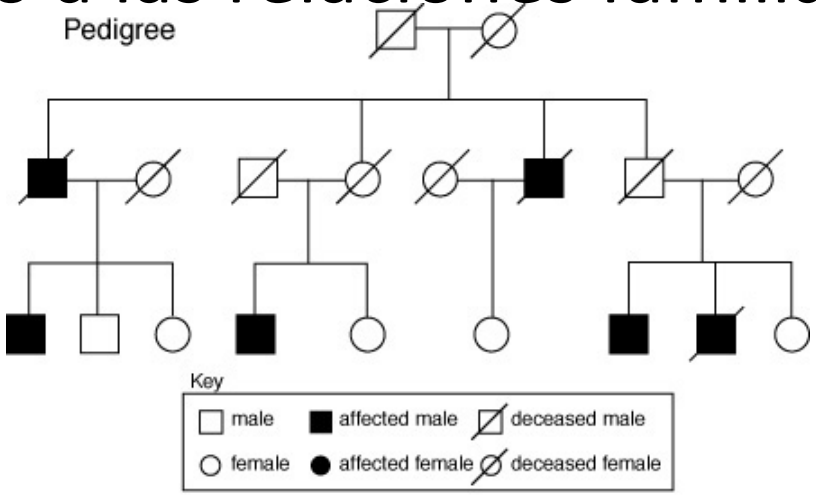
Crédito de la imagen: Wikipedia, http://www2.estrellamountain.edu/faculty/farabee/biobk/BioBookDivers_class.html

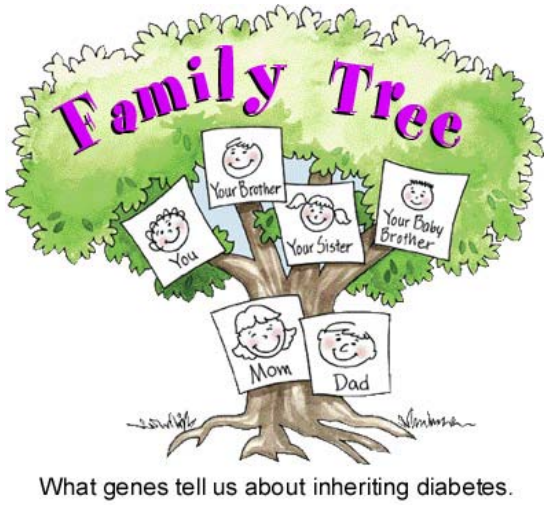
escalas más finas

- La misma idea se puede aplicar a clasificar diferentes cepas de un tipo de bacteria

o a las relaciones familiares



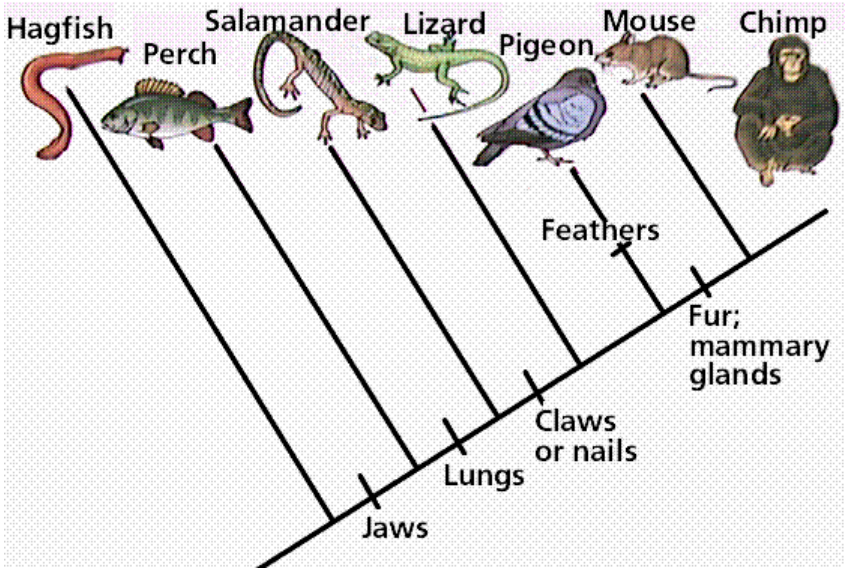




Crédito de la imagen: Hershberg et al, Genome Biology. 8: R164 (2007), http://www.accessexcellence.org/RC/VL/GG/images/pedigree.gif, http://www.jdrf.ca/

Relacionar objetos biológicos

- ¿Cómo se han determinado las jerarquías?
  - Especie: tradicionalmente por las similitudes morfológicas y de comportamiento, o evidencias paleontológicas
  - Las cepas bacterianas: por propiedades biológicas físicas, químicas y
    - Pregunta: ¿Qué características deben ser usadas antes?
  - Miembros de la familia: por pedigrí / genealogía



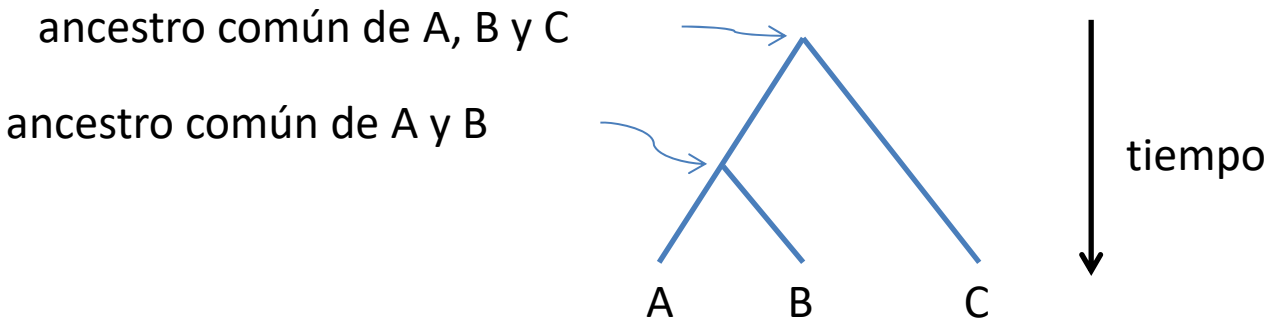
Crédito de la imagen: http://www2.estrellamountain.edu/faculty/farabee/biobk/BioBookDivers_class.html

Filogenia

- Una manera **sistemática y objetiva** para construir estos árboles es mediante la comparación de las secuencias de **DNA / proteínas**
- Estudiamos aquí los árboles que relacionan los objetos **suficientemente diferentes**
  - Especies diferentes
  - Las diferentes cepas / poblaciones de una especie
- Nuestro objetivo es reconstruir las relaciones evolutivas reales en base a secuencias observables

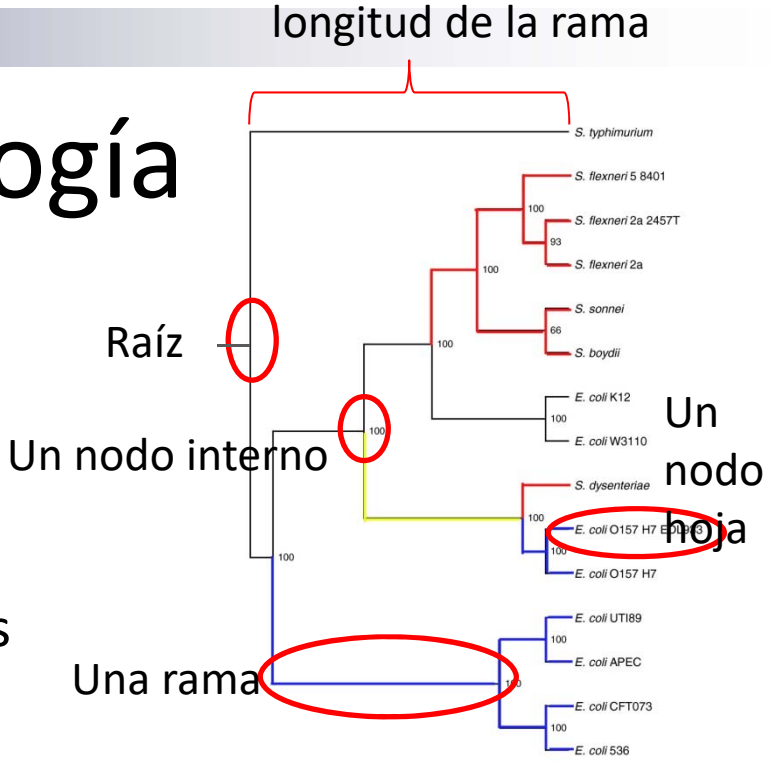
suposiciones

- supuestos básicos detrás de los árboles filogenéticos:
  1. Las secuencias actuales comparten un ancestro común
  2. Todas mutaron a partir del ancestro común
  3. Las mutaciones son raras. Por tanto, si los DNA de A y B son más similares que los de A y C, y que los de B y C, probablemente C se separó de A y B antes de su separación



Terminología

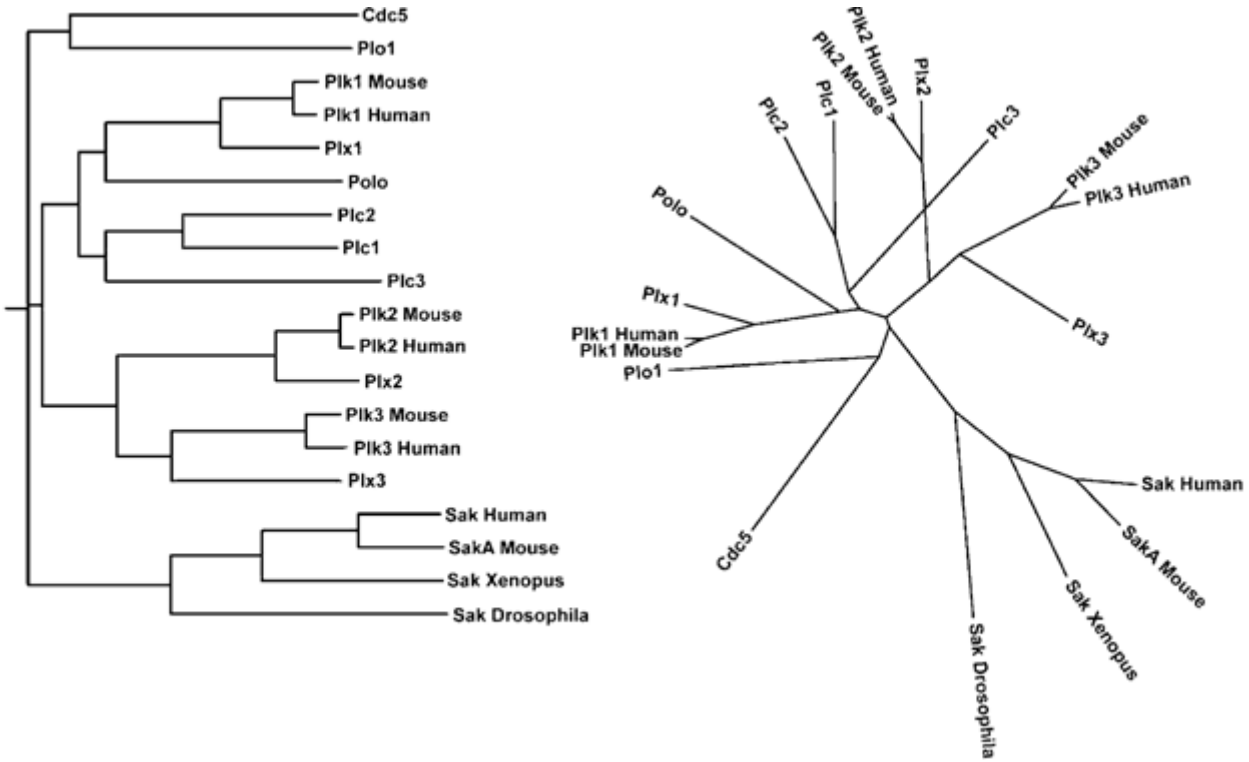
- Un árbol es un grafo acíclico con nodos conectados por aristas
- Un árbol filogenético es un árbol binario con secuencias (nodos) conectados por ramas (aristas)
  - Los **nodos hoja** son las secuencias observadas
  - Los **nodos internos** son las secuencias ancestrales no observadas
  - El **nodo raíz** es el ancestro común de todas las secuencias observadas
  - **Las longitudes de las ramas** pueden representar distancias evolutivas



Crédito de la imagen: Hershberg et al, Genome Biology. 8: R164 (2007), http://www.jdrf.ca/

Árboles con y sin raíz

- A veces no está muy claro dónde incluir el ancestro común
  - Podemos tener un árbol sin raíz



Crédito de la imagen: l

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

formatos de archivo comunes para los árboles filogenéticos

- Newick (paréntesis anidados, con distancias)
- NEXUS (dando identificaciones cortas para secuencias, con más metadatos)
- PhyloXML (utilizando la estructura de XML)

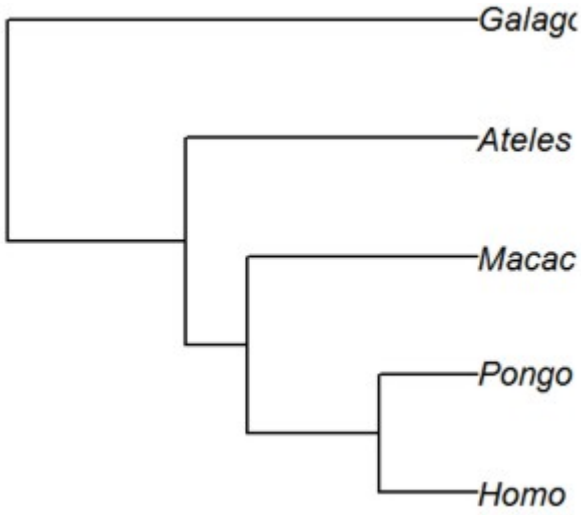
formatos de archivo comunes

- Ejemplo

newick:

((((Homo:0,21, Pongo:0,21):0,28, Macaca:
0,49):0,13, Ateles:0,62):0,38, Galago:1,00);

Representación grafica:



NEXUS:

```
BEGIN TAXA;
	DIMENSIONS NTAX = 5;
	TAXLABELS
		Homo
		Pongo
		Macaca
		Ateles
		Galago
	;
END;
BEGIN TREES;
	TRANSLATE
		1	Homo,
		2	Pongo,
		3	Macaca,
		4	Ateles,
		5	Galago
	;
	TREE * UNTITLED = [&R]
((((1:0.21,2:0.21):0.28,3:0.49):0.13,4:0.62):0.38,5:1
);
END;
```

Crédito de la imagen: http://www.zoology.ubc.ca/~schluter/zoo502stats/Rtips.phylogeny.html

formatos de archivo comunes

- PhyloXML

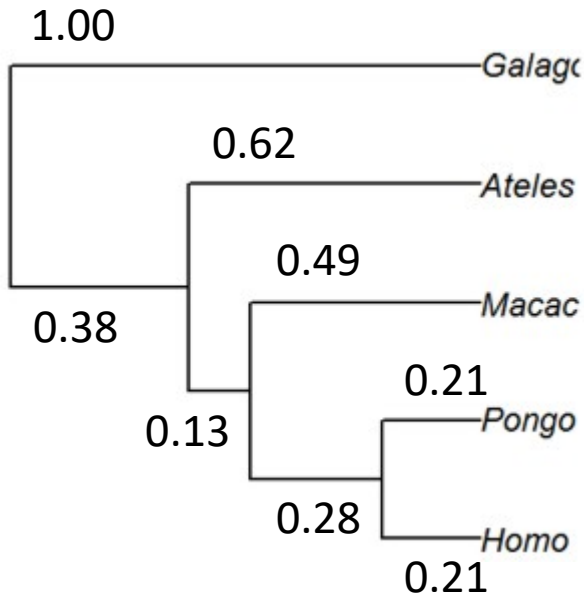
```
<phyloxml xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.phyloxml.org" xsi:schemaLocation="http://www.phyloxml.org
http://www.phyloxml.org/1.10/phyloxml.xsd">
    <phylogeny rooted="true">
        <name>Alcohol dehydrogenases</name>
        <description>contains examples of commonly used elements</description>
        <clade>
            <events>
                <speciations>1</speciations>
            </events>
            <clade>
                <taxonomy>
                    <id provider="ncbi">6645</id>
                    <scientific_name>Octopus vulgaris</scientific_name>
                </taxonomy>
                <sequence>
                    <accession source="UniProtKB">P81431</accession>
                    <name>Alcohol dehydrogenase class-3</name>
                </sequence>
            </clade>
            ...
        </clade>
    </phylogeny>
</phyloxml>
```

Fuente de información: http://www.phyloxml.org/examples_syntax/phyloxml_syntax_example_1.html

El formato Newick

- Utiliza paréntesis y coma para agrupar dos subárboles
- Usa dos puntos para indicar la distancia a los padres, si está disponible
- Termina con un punto y coma

Representación grafica:



newick:

((((Homo: 0,21, Pongo: 0,21): 0,28, Macaca: 0,49): 0,13, Ateles: 0,62): 0,38, Galago: 1,00);

observaciones:

- Para un árbol sin raíz, de una manera sencilla para representarlo utilizando el formato Newick es poniendo una raíz de manera arbitraria
- Se pueden nombrar a los nodos internos, dando la etiqueta después del paréntesis de cierre (por ejemplo, (Homo: 0,21, Pongo: 0,21) HP: 0,28

Crédito de la imagen: http://www.zoology.ubc.ca/~schluter/zoo502stats/Rtips.phylogeny.html

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

reconstrucción de árboles filogenéticos

- problema general:
  - Dado un conjunto de secuencias / proteínas de DNA, encontrar un árbol filogenético de tal manera que probablemente corresponde a los eventos evolutivos históricos reales, describiendo:
    - Cómo se conectan los nodos: Orden de eventos de separación
    - Lo secuencias en los nodos internos: secuencias ancestrales
    - Cuánto tiempo ha pasado desde que la separación: longitud de rama

    Hay varias formas de evaluar la probabilidad de que un árbol sea correcto. Vamos a estudiarlas

reconstrucción de árboles filogenéticos

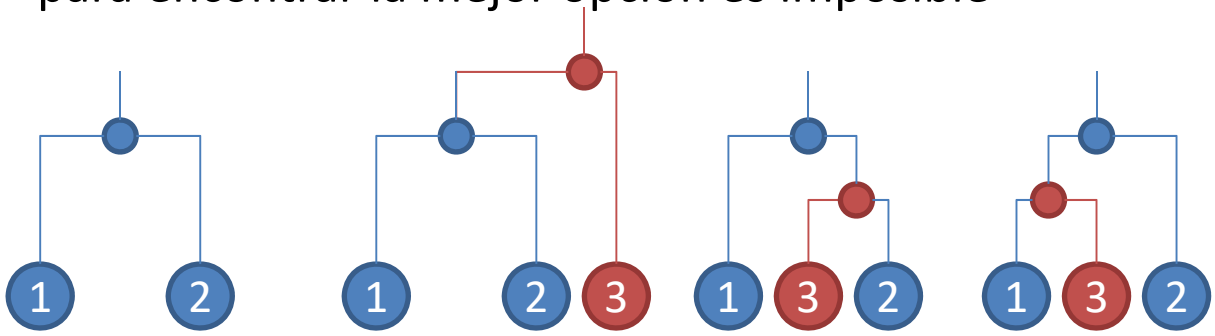
- "Re" -Construcción: El árbol fue definido por la historia. Nosotros sólo tratamos de reconstruirlo desde las secuencias observadas

¿Qué secuencias usar?

- Si estamos estudiando un gen
  - secuencia de DNA / proteína del gen
- Si queremos conocer la relación entre las diferentes especies
  - el genoma completo (puede no ser factible)
  - Algunos genes que evolucionan lentamente
    - RNA ribosomal

La complejidad del problema

- Encontrar el "mejor" árbol es un problema difícil
  - ¿Cuántas (es decir, ignorando longitudes de rama y el orden de izquierda a derecha) son las topologías de árboles para un conjunto de k secuencias?
  - Para los árboles enraizados:
    - k = 2: 1 posible topología de árbol
    - k = 3: 3 ramas posibles para añadir # 3
    - k = 4: 5 posibles para añadir # 4, y así sucesivamente
    - Por lo tanto el número de topologías de árboles es $1\times 3\times 5\times \times \ldots (2k-3)$ - Exponencial
  - Una vez más, la enumeración de todas las topologías para encontrar la mejor opción es imposible

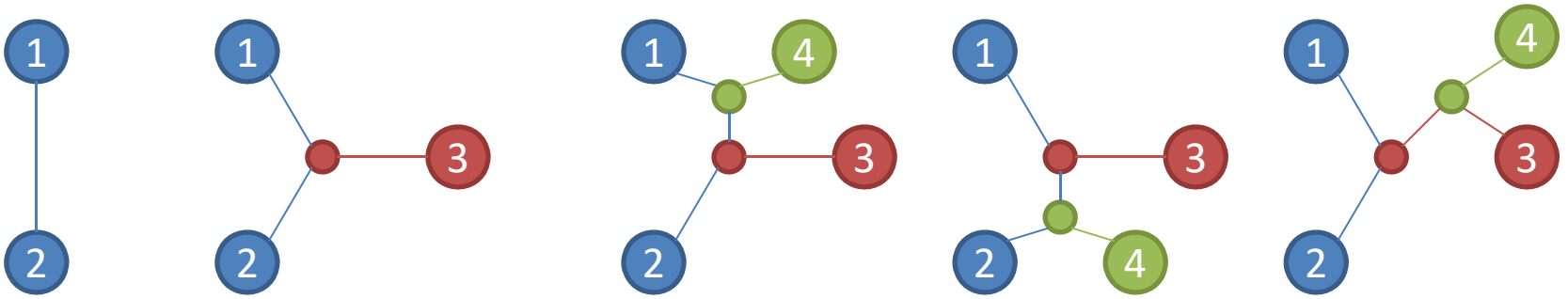


k	Num. of rooted tree topologies
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,370
17	191,898,783,962,511,000
18	6,332,659,870,762,850,000
19	221,643,095,476,700,000,000
20	8,200,794,532,637,890,000,000

La complejidad del problema

- Del mismo modo, para los árboles sin raíces,
  - k = 2: 1 topología
  - k = 3: 1 forma de añadir # 3
  - k = 4: 3 formas de agregar # 4
  - k = 5: 5 formas de añadir # 5
  - El número de topologías es $1 \times 3 \times 5 \times \ldots \times (2k-5)$

k	Num. of rooted tree topologies	Num. of unrooted tree topologies
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	34,459,425	2,027,025
11	654,729,075	34,459,425
12	13,749,310,575	654,729,075
13	316,234,143,225	13,749,310,575
14	7,905,853,580,625	316,234,143,225
15	213,458,046,676,875	7,905,853,580,625
16	6,190,283,353,629,370	213,458,046,676,875
17	191,898,783,962,511,000	6,190,283,353,629,370
18	6,332,659,870,762,850,000	191,898,783,962,511,000
19	221,643,095,476,700,000,000	6,332,659,870,762,850,000
20	8,200,794,532,637,890,000,000	221,643,095,476,700,000,000



La solución del problema: Ideas

- ¿Qué se hace cuando se encuentra un problema difícil computacionalmente?
  - Definir una versión más fácil del problema
    - hacer ciertos supuestos
  - Diseñar algoritmos / estructuras de datos inteligentes para evitar cálculos redundantes
  - Usar heurísticas para resolverlo, no obteniendo necesariamente la solución óptima

métodos de reconstrucción de árboles filogenéticos

- Existen dos tipos principales de métodos:
  - basados en secuencias: necesitan las secuencias
    - métodos de parsimonia (problemas más fáciles, algoritmos inteligentes)
    - métodos probabilísticos (problemas más fáciles, algoritmos inteligentes)
      - Máxima verosimilitud
      - bayesiano
    - ...
  - basados en la distancia: sólo dependen de las distancias entre las secuencias
    - UPGMA (heurística)
    - vecino más próximo (heurística)
    - ...
  - Vamos a estudiar algunos de estos algoritmos

El próximo día …

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