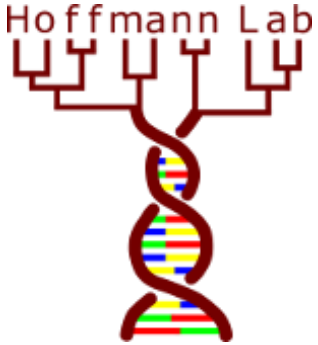
# Evolución de familias multigénicas 2020

Introducción al análisis filogenético

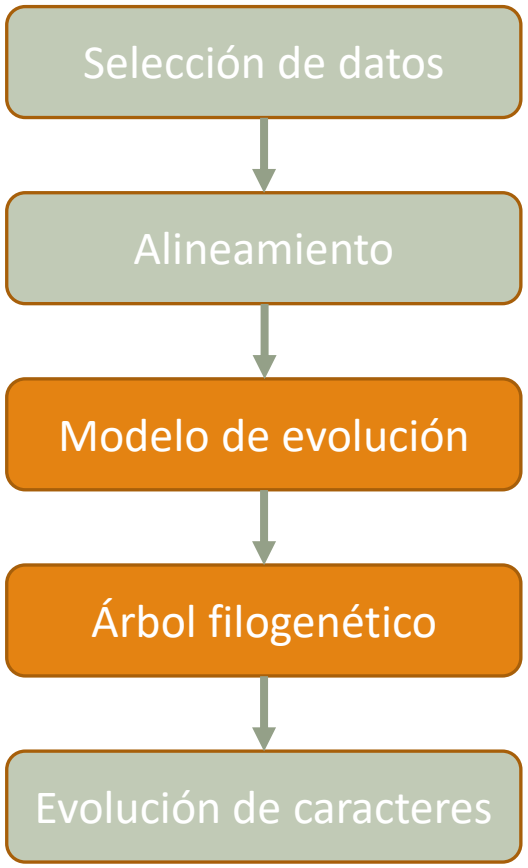
(4)

FACULTAD DE CIENCIAS

UDELAR | fcien.edu.uy



# Modelos, distancias, y árboles



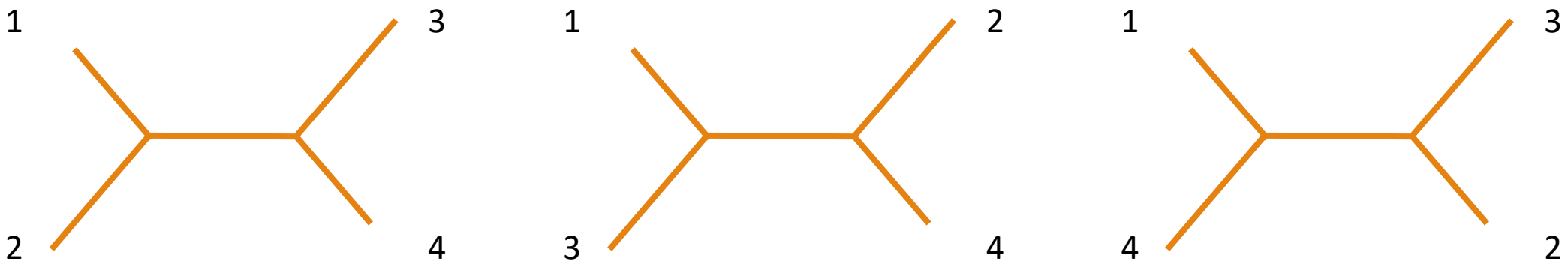

La selección de un modelo de evolución molecular es un paso necesario para:

- Estimar distancias moleculares.
- Obtener árboles a partir de dichas distancias (por ejemplo, usando Unión de Vecinos [Neighbor Joining]).
- Obtener árboles a partir de métodos de inferencia estadística:
  - Máxima verosimilitud (máximum likelihood).
  - Inferencia bayesiana.

# Algunos métodos de inferencia filogenética

| Método | Variantes | Criterio de optimización | Uso de variación no observada |
|---|---|---|---|
| Parsimonia | máxima parsimonia | minimizar el número de pasos requeridos para obtener los datos | No |
| Distancias | - evolución mínima | minimizar los cambios requeridos para obtener las distancias estimadas entre taxa | Sí (incorporadas en las distancias) |
| | - unión de vecinos (neighbor joining) | una aproximación al árbol de evolución mínima | ídem |
| Inferencia estadística | Máxima verosimilitud | maximizar la verosimilitud de observar los datos, dado un árbol y un modelo de evolución molecular. | Sí (considerando todos los estados posibles en los nodos). |
| | Inferencia bayesiana | | ídem |

# 4 taxones, 3 árboles (topologías) sin raíz

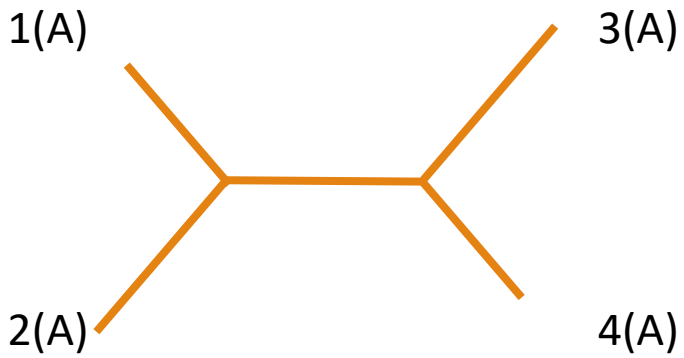


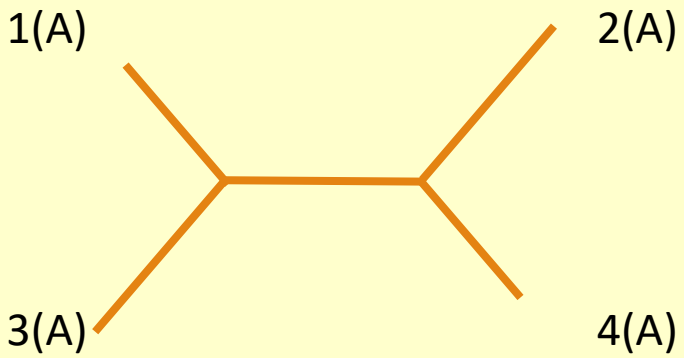
# Tres miradas sobre los datos

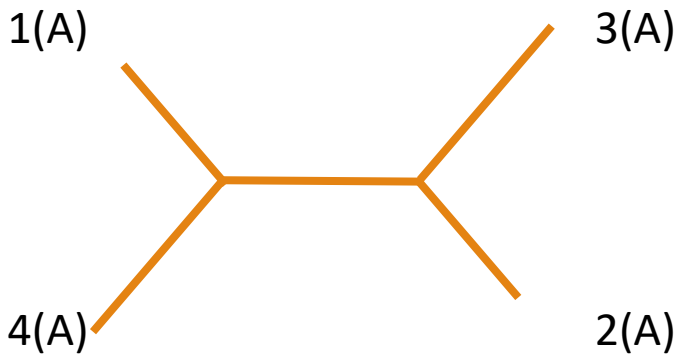
| Taxón/sitio | | | | |
|---|---|---|---|---|
| 1 | A | C | A | T |
| 2 | A | C | A | C |
| 3 | A | C | G | T |
| 4 | A | T | G | C |
| Parsimonia | no informativo | No informativo | [(1,2),(3,4)] | [(1,3),(2,4)] |
| Distancias | Usar datos y modelo de sustitución para estimar distancias pareadas corregidas para sustituciones múltiples (1,2); (1,3); (1,4);(2,3);(2,4);(3,4) | | | |
| Inf. estadística | L = P(D\|T,M); D:datos; T: árbol (tree); M: modelo (incluye modelo de sustitución y árbol) | | | |

# Parsimonia

| Taxón/sitio | | | | |
|---|---|---|---|---|
| 1 | A | C | A | T |
| 2 | A | C | A | C |
| 3 | A | C | G | T |
| 4 | A | T | G | C |
| Parsimonia | no informativo | No informativo | [(1,2),(3,4)] | [(1,3),(2,4)] |

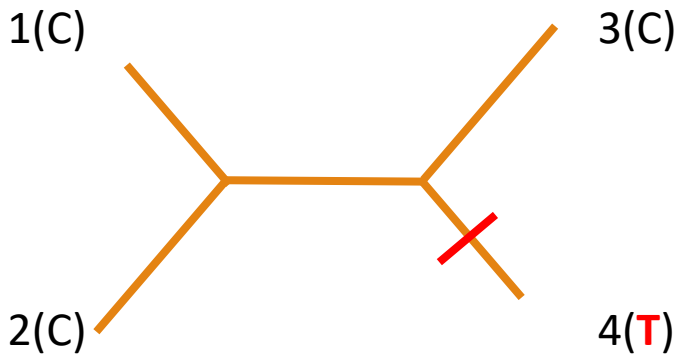


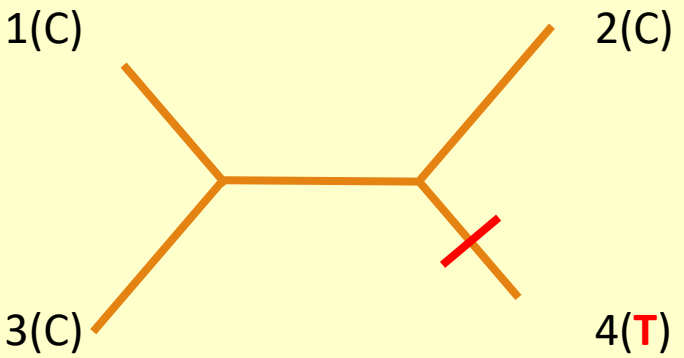



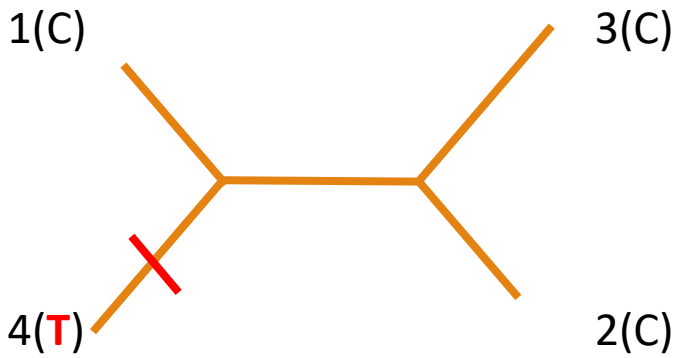



# Parsimonia

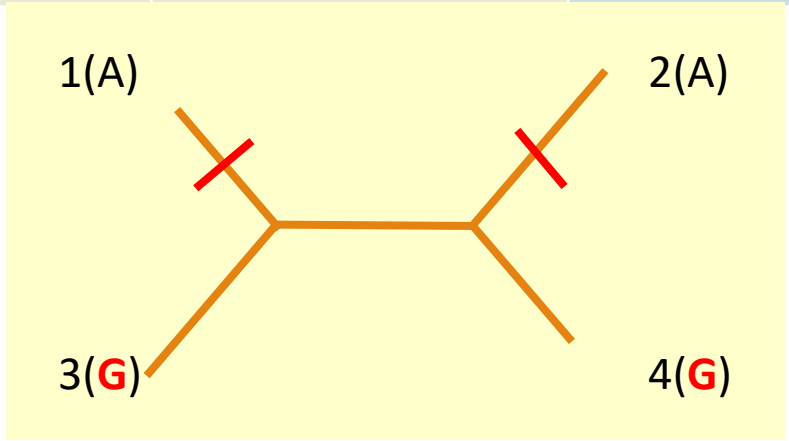
| Taxón/sitio | | | | |
|---|---|---|---|---|
| 1 | A | C | A | T |
| 2 | A | C | A | C |
| 3 | A | C | G | T |
| 4 | A | T | G | C |
| Parsimonia | no informativo | No informativo | [(1,2),(3,4)] | [(1,3),(2,4)] |

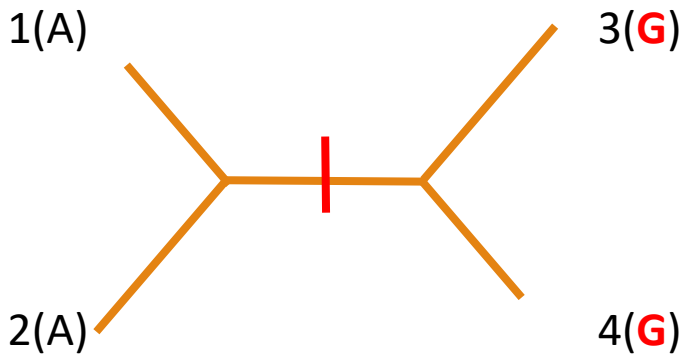


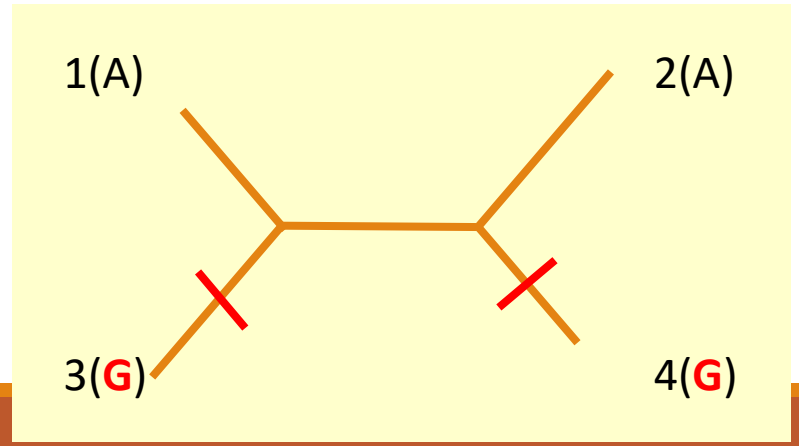



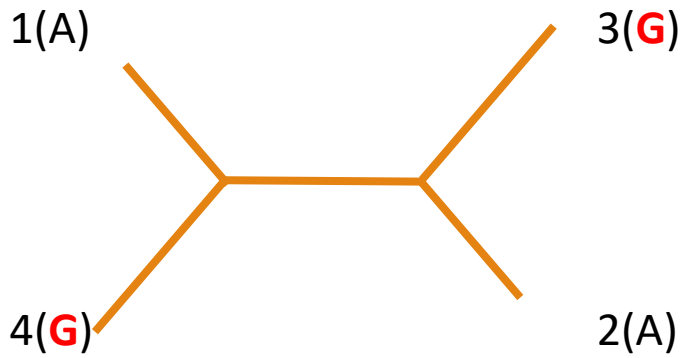



| Taxón/sitio | | | | |
|---|---|---|---|---|
| **1** | A | C | A | T |
| **2** | A | C | A | C |
| **3** | A | C | G | T |
| **4** | A | T | G | C |
| **Parsimonia** | **no informativo** | | **No informativo** | **[(1,2),(3,4)]** | **[(1,3),(2,4)]** |

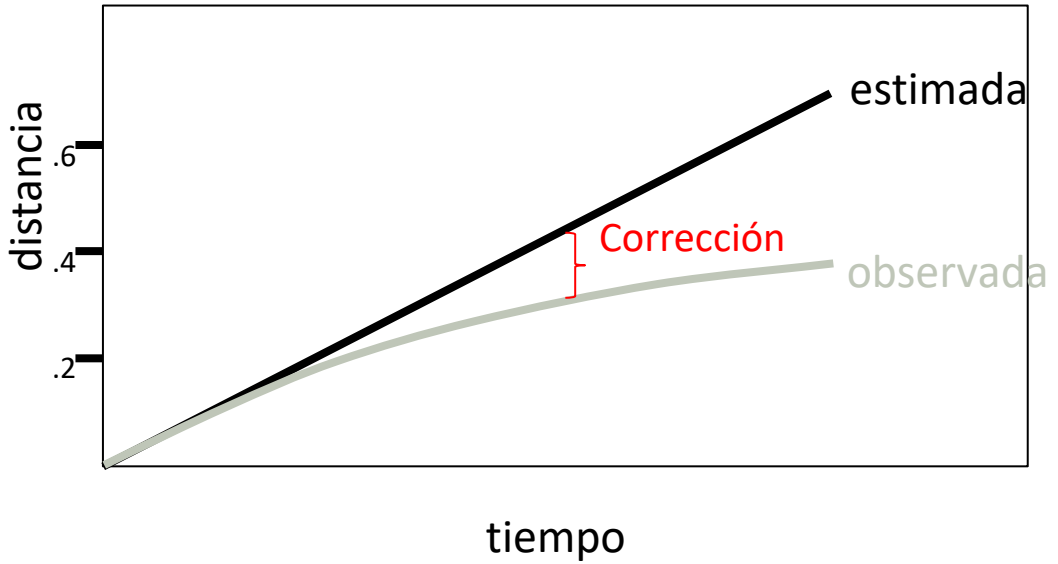


# Distancias




| Taxón/sitio | | | | |
|---|---|---|---|---|
| 1 | A | C | A | T |
| 2 | A | C | A | C |
| 3 | A | C | G | T |
| 4 | A | T | G | C |
| Distancias | Usar todos los datos y modelo de sustitución para estimar distancias pareadas corregidas para sustituciones múltiples (1,2); (1,3); (1,4);(2,3);(2,4);(3,4) | | | |

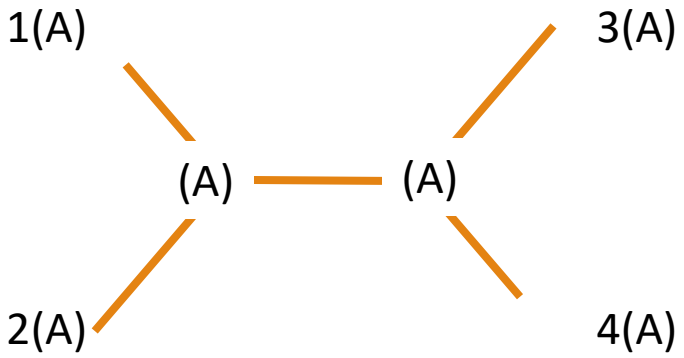
# Inferencia estadística

| Taxón/sitio | | | | |
|---|---|---|---|---|
| 1 | A | C | A | T |
| 2 | A | C | A | C |
| 3 | A | C | G | T |
| 4 | A | T | G | C |
| Distancias | Usar datos y modelo de sustitución para estimar distancias pareadas corregidas para sustituciones múltiples (1,2); (1,3); (1,4);(2,3);(2,4);(3,4) | | | |
| Inf. estadística | L = P(D\|T,M); D:datos; T: tree; M: modelo (incluye modelo de sustitución y árbol) | | | |

# Algunos métodos de inferencia filogenética

| Método | Variantes | Criterio de optimización | Uso de variación no observada |
|---|---|---|---|
| Parsimonia | máxima parsimonia | minimizar el número de pasos requeridos para obtener los datos | No |
| Distancias | - evolución mínima | minimizar los cambios requeridos para obtener las distancias estimadas entre taxa | Sí (incorporadas en las distancias) |
| | - unión de vecinos (neighbor joining) | una aproximación al árbol de evolución mínima | ídem |
| Inferencia estadística | Máxima verosimilitud | maximizar la verosimilitud de observar los datos, dado un árbol y un modelo de evolución molecular. | Sí (considerando todos los estados posibles en los nodos). |
| | Inferencia bayesiana | | ídem |

# Ejemplo: sitio sin variación observada




Parsimonia: el único estado aceptable en todo el árbol es el observado (A)

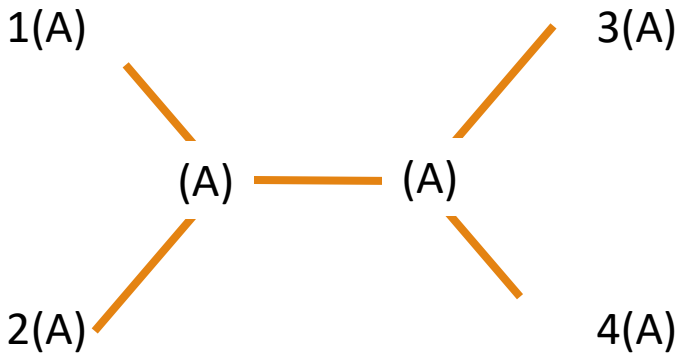
Distancias: corrección global (no sitio a sitio) que procura estimar la fracción de cambios no observados.

Máxima verosimilitud: solución máximo verosímil es aquella que maximiza L (likekihood)

$$L = P(D|T,M)$$

Debemos calcular P(D) sumando las probabilidades parciales de que los nodos sean (A,A) [como en el diagrama], (A,C), (A,G), (A,T), (C,A)….. etc.

# Ejemplo: sitio sin variación observada




Máxima verosimilitud: solución máximo verosímil es aquella que maximiza L (likekihood)

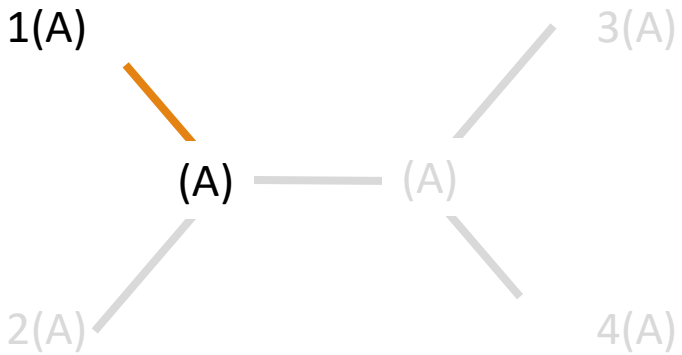
$$L = P(D|T,M)$$

Debemos calcular P(D) sumando las probabilidades parciales de que los nodos sean (A,A) [como en el diagrama], (A,C), (A,G), (A,T), (C,A)….. etc.

Modelos reversibles en el tiempo: nos permiten calcular a lo largo de **cada rama**, independientemente del sentido del tiempo.

Fijamos T (la topología y las longitudes de las ramas) y M para calcular lo largo de cada rama.

# Ejemplo: sitio sin variación observada




Máxima verosimilitud: solución máximo verosímil es aquella que maximiza L (likekihood)

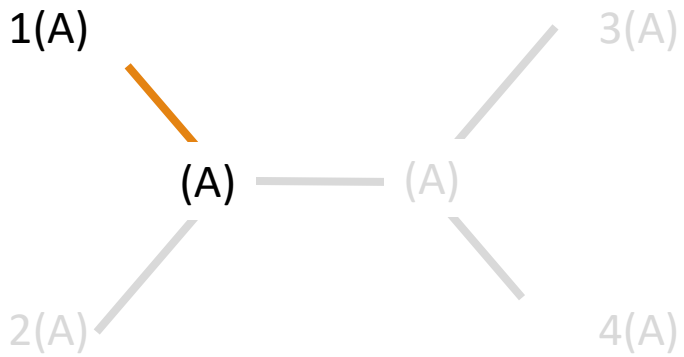
$$L = P(D|T,M)$$

Debemos calcular P(D) sumando las probabilidades parciales de que los nodos sean (A,A) [como en el diagrama], (A,C), (A,G), (A,T), (C,A)….. etc.

Modelos reversibles en el tiempo: nos permiten calcular a lo largo de **cada rama**, independientemente del sentido del tiempo.

Fijamos T (la topología y las longitudes de las ramas) y M para calcular a lo largo de cada rama.

# Cálculos a lo largo de una rama: 1 sitio




Comienzo en el estado observado en el taxón 1.

Modelo Jukes-Cantor: primera generación

$$P(x_1 = A \mid x_0 = A) = 1 - 3\mu$$

Segunda generación:

Probabilidad de no estar en A en $t_1$

$$P(x_2 = A) = (1 - 3\mu)P(x_1 = A) + \mu[1 - (P(x_1 = A)]$$

Probabilidad de estar en A en $t_1$

| | G | A | T | C |
|---|---|---|---|---|
| G | 1-3μ | μ | μ | μ |
| A | μ | 1-3 μ | μ | μ |
| T | μ | μ | 1-3 μ | μ |
| C | μ | μ | μ | 1-3 μ |

# Cálculos a lo largo de una rama: 1 sitio

Primera generación

$$P(x_1 = A \,|\, x_0 = A) = 1 - 3\mu$$

Segunda generación:

Probabilidad de estar en A en $t_1$ por la probabilidad de no cambiar de $t_1$ a $t_2$

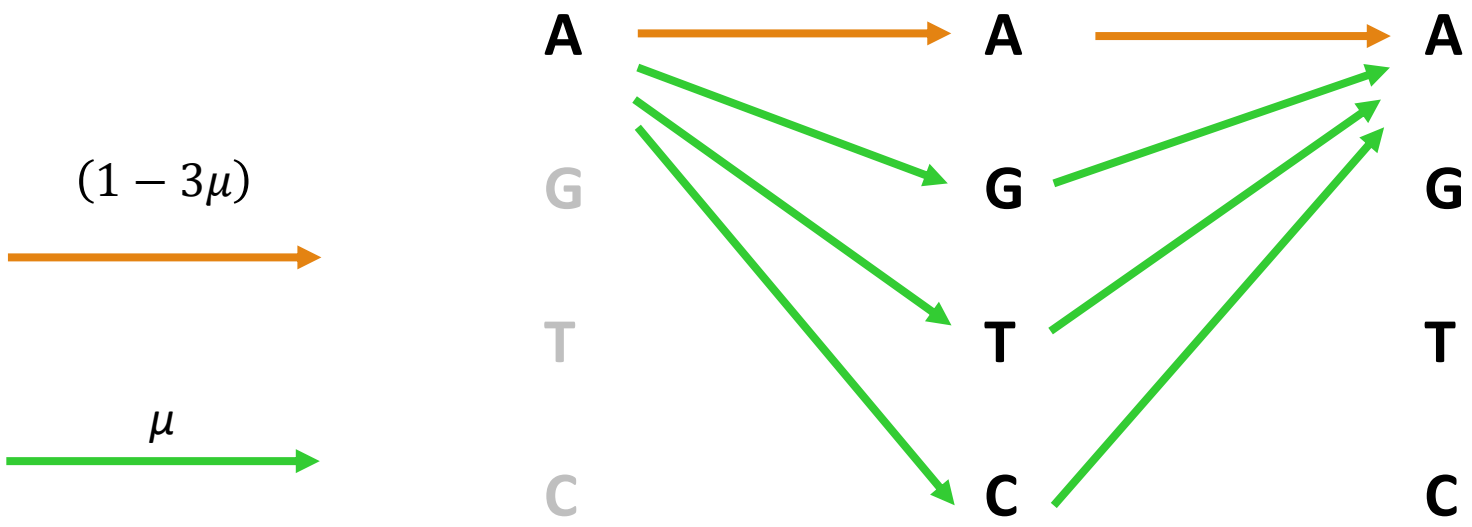
$$P(x_2 = A) = (1 - 3\mu)P(x_1 = A) + \mu[1 - (P(x_1 = A)]$$

Probabilidad de no estar en A en $t_1$ por la probabilidad de cambiar a A de $t_1$ a $t_2$

# Cálculos a lo largo de una rama: 1 sitio

Primera generación $P(x_1 = A | x_0 = A) = 1 - 3\mu$

Segunda generación: $P(x_2 = A) = (1 - 3\mu)P(x_1 = A) + \mu[1 - (P(x_1 = A)]$




Conceptualmente: sumamos probabilidades para trayectorias mutuamente excluyentes.

# Cálculos a lo largo de una rama: 1 sitio

Observaciones:

- Necesitamos la longitud de dicha rama.  La probabilidad de pasar de un estado en un extremo a otro en el otro extremo depende del modelo mutacional y de dicha longitud.
- Lo anterior vale para la probabilidad de observar un mismo estado en los dos extremos.
- Notamos (sin mostrarlo) que, conociendo la topología, el modelo y los datos se pueden obtener las longitudes de las ramas.
- Notamos que nuestra ecuación lleva naturalmente a una recurrencia, puesto que pasamos de $t_2$ a $t_3$ con las mismas reglas que aplicamos previamente para pasar de $t_1$ a $t_2$.
- Normalmente pasamos a tiempo continuo para simplificar los cálculos.

# Verosimilitud de un árbol: 1 sitio

El árbol es el conjunto de las ramas (con sus longitudes).

Por tanto, “sabemos” cómo acumular los cálculos para todas las ramas para nuestro sitio de interés.  Fijamos los estados en los taxones terminales y calculamos sobre todas las combinaciones posibles de estados en todos los demás.

El árbol máximo verosímil es aquel para el que obtengo un valor de L mayor, aplicando los conceptos y cálculos que acabamos de esbozar.

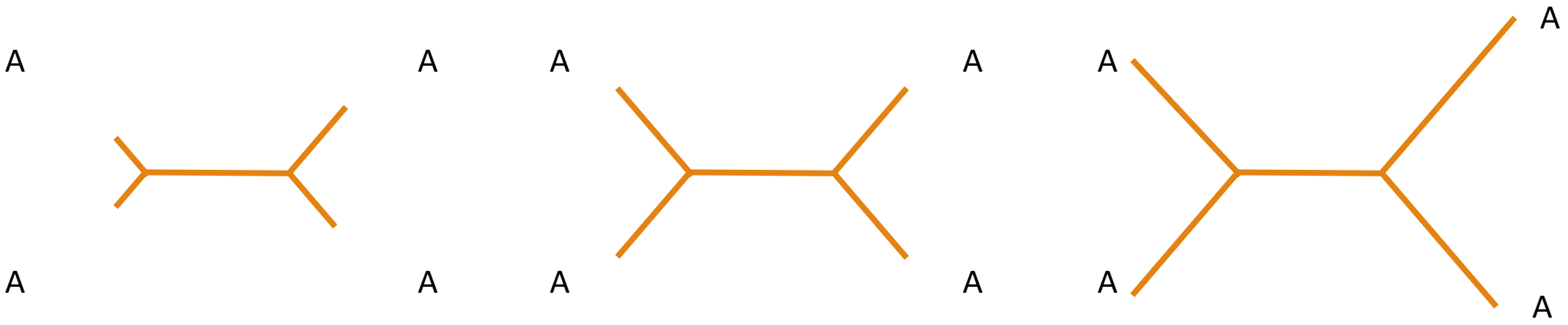
Notar que la longitud de las ramas (tiempos disponibles para sustituciones) se fija a partir de cada topología y los datos para todos los sitios.

# Verosimilitud de un árbol: 1 sitio

El árbol máximo verosímil es aquel para el que obtengo un valor de L mayor, aplicando los conceptos y cálculos que acabamos de esbozar.

Notar que la longitud de las ramas (tiempos disponibles para sustituciones) se fija a partir de cada topología y los datos para todos los sitios.

Por tanto, aún para un sitio sin variación observada puede haber un árbol más verosímil.



# Verosimilitud de un árbol: todos los sitios

Ya vimos que, para cada sitio, **sumamos** la probabilidad de observar los datos (dada la topología y las longitudes de las ramas) acumulando entre trayectorias mutuamente excluyentes.

Repetimos el ejercicio sitio a sitio.

Si los sitios evolucionan en forma independiente, La verosimilitud para todos los sitios es el **producto** de las verosimilitudes de los sitios.

El árbol máximo verosímil es aquel para el cual dicho producto es mayor.

Ese es nuestro criterio de optimización.

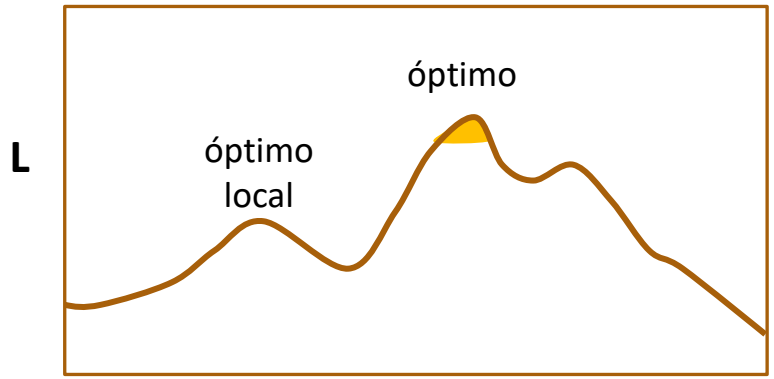
# Selección de árboles

Al igual que en parsimonia, el único método seguro de encontrar el árbol óptimo es:

1) Calcular el valor de cada árbol (L, número de pasos)

2) Ordenar según dicho valor y elegir el óptimo.

Como el número de árboles posibles es grande, hay algoritmos de búsqueda para aproximarse con razonable probabilidad al resultado deseado.



# Máxima verosimilitud: comentarios adicionales

La verosimilitud se obtiene sitio a sitio.  Por tanto, es posible aplicar modelos distintos a distintos sitios, o clases de sitios (posiciones de codón, dominios, genes diferentes).

El límite práctico a dicho enfoque no es conceptual sino computacional.

La razón de verosimilitudes (LRT) ofrece un estadístico aproximado (hay otros) para comparar árboles e identificar conjuntos de árboles que no difieren significativamente del más verosímil. Es la “zona de confianza” de la estimación filogenética.

Examinamos nuestro criterio de optimización: L(D l T, M):

- D y M son fijos (los datos provienen de observaciones y el modelo de una selección previa).
- Aspiramos, por tanto, a evaluar posibles “valores” de T, que son hipótesis filogenéticas.

Estamos a “un paso” del enfoque bayesiano, que parte de la aspiración de estimar L(T l D, M)