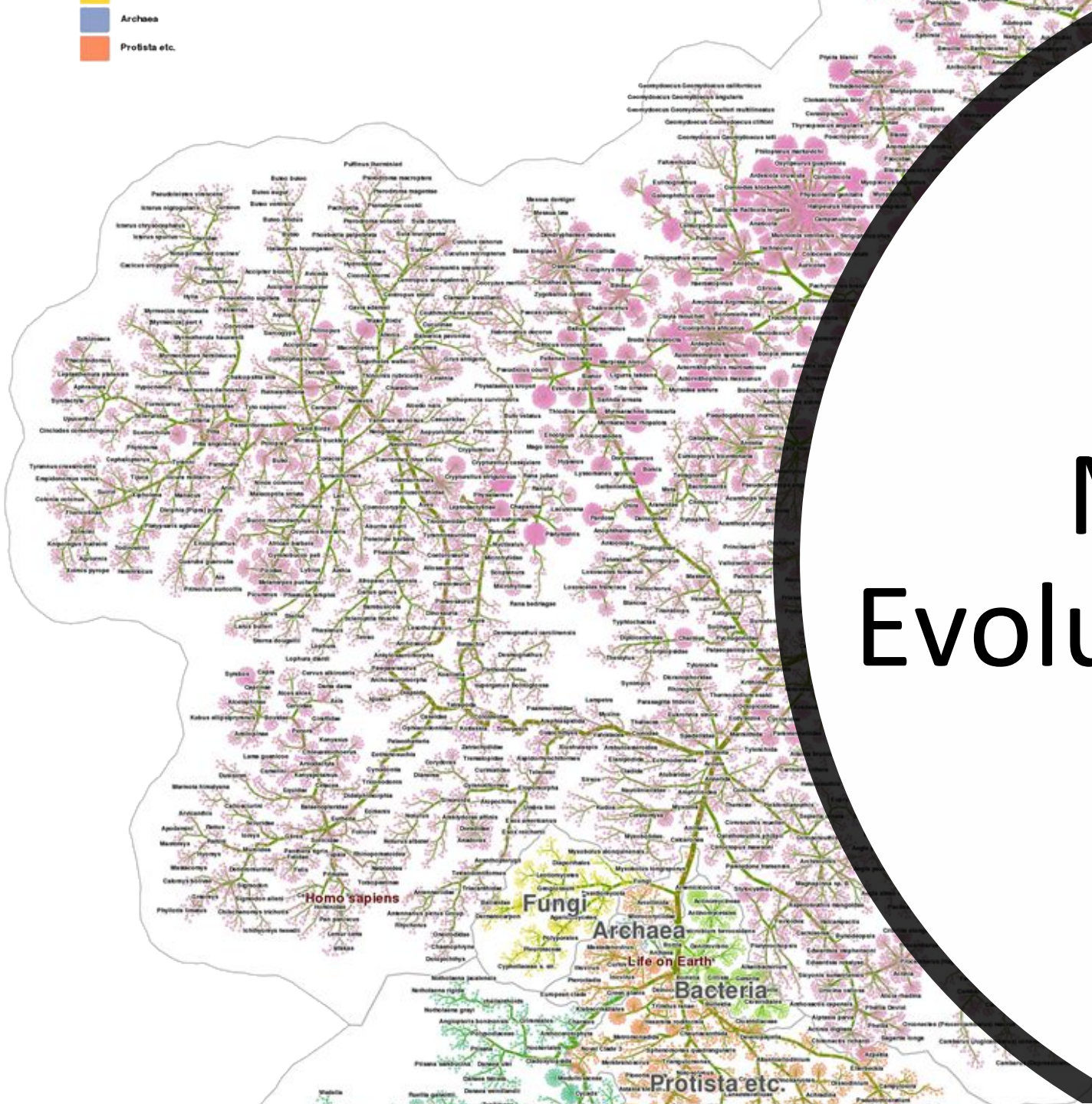
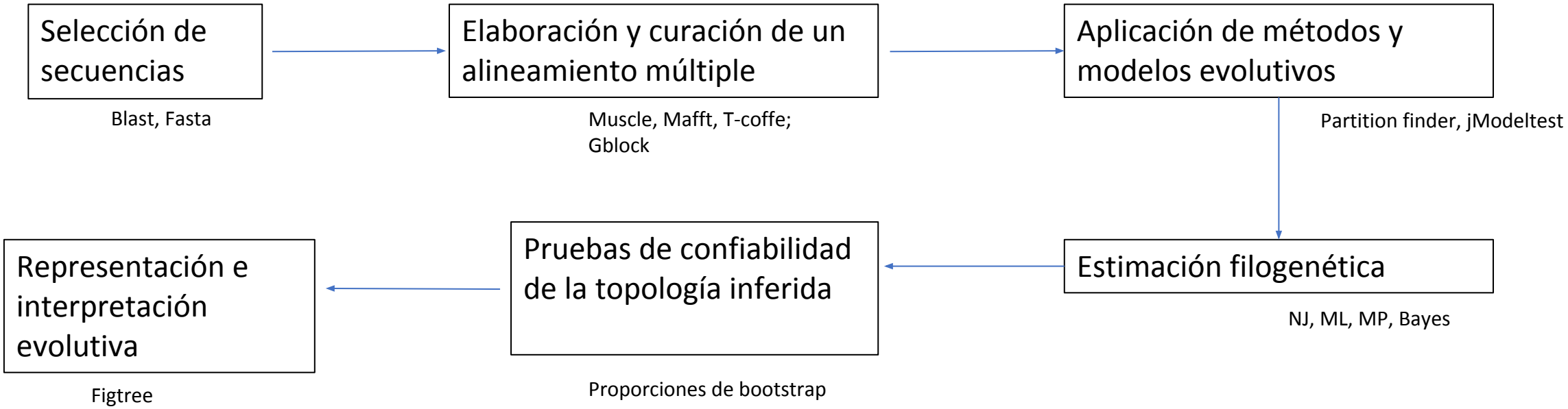
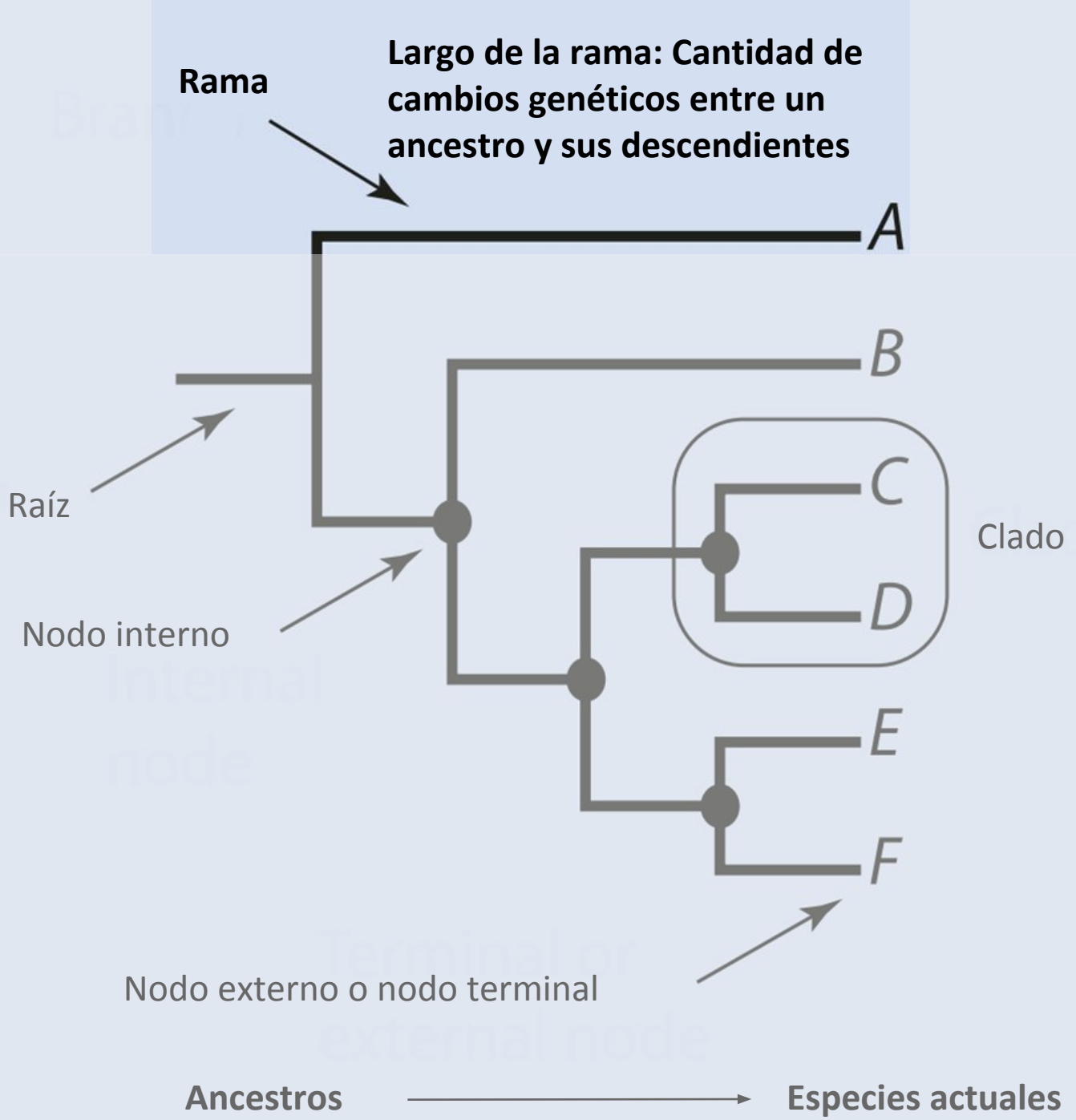


Modelos de Evolución Molecular

- Aunque puede parecer ventajoso tener un solo árbol a considerar, esa comodidad puede ser engañosa porque da la falsa impresión de que el árbol que ve es el "correcto". **Es esencial comprender que el árbol "correcto" no existe**. Estamos tratando de deducir el orden en que los taxones existentes (con lo que aquí nos referimos a secuencias) divergieron de un ancestro común hipotético, y calcular la cantidad de cambio a lo largo de las ramas entre los eventos divergentes**. Es extremadamente improbable que esas deducciones sean correctas en cada detalle, por lo que el árbol que vemos no será una descripción precisa de los acontecimientos históricos.** Incluso si solo nos interesa la topología de los árboles, nunca podemos estar seguros de que la topología del árbol refleje con precisión el orden histórico de ramificación. **Cualquiera sea el método que elijamos, lo único de lo que podemos estar seguros es que el árbol resultante está equivocado.** Lo mejor que podemos esperar es un árbol que se acerque más a lo que sucedió en el pasado. En resumen, debemos reconocer que, dado que no sabemos lo que sucedió en el pasado, nunca podemos estar completamente seguros de la precisión del árbol. Los métodos de búsqueda de árboles pueden producir uno o varios árboles, pero todos los métodos reconocen implícitamente que los árboles producidos son solo un subconjunto de los posibles árboles que son consistentes con los datos.

Cómo se construye un árbol filogenético




Rama
Largo de la rama: Cantidad de cambios genéticos entre un ancestro y sus descendientes
A
B
C
D
E
F
Raíz
Nodo interno
Clado
Nodo externo o nodo terminal
Ancestros
Especies actuales

Sustitución nucleotídica

1. Número de diferencias nucleotídicas

2. p-distance

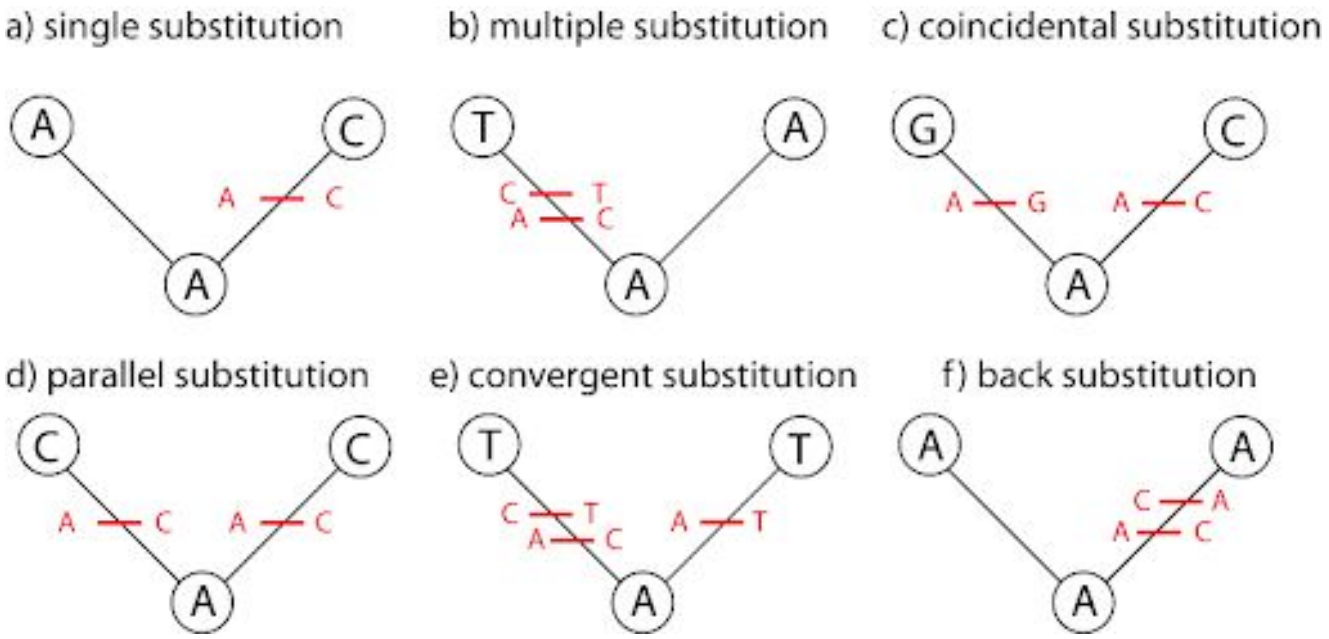
Taxa	1	2	3*	4	5	6	7
1	T	G	C	G	T	A	T
2	T	G	G	G	T	A	T

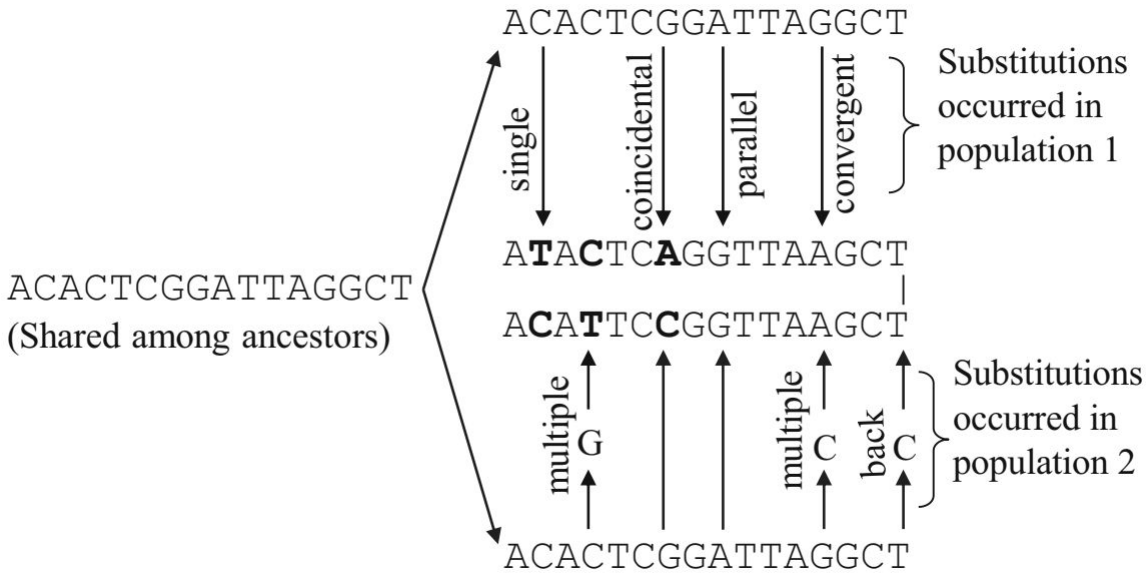
$$p - distance = \frac{\#\ of\ nucleotide\ difference}{Total\ \#\ of\ nucleotide\ compared}$$

$$p - distance = \frac{1}{7} = 0{,}143$$

Debido a sustituciones posteriores o paralelas, la distancia p a menudo subestima el número de sustituciones que se han producido (la distancia p funciona bien para secuencias muy similares, digamos, con p <5%).

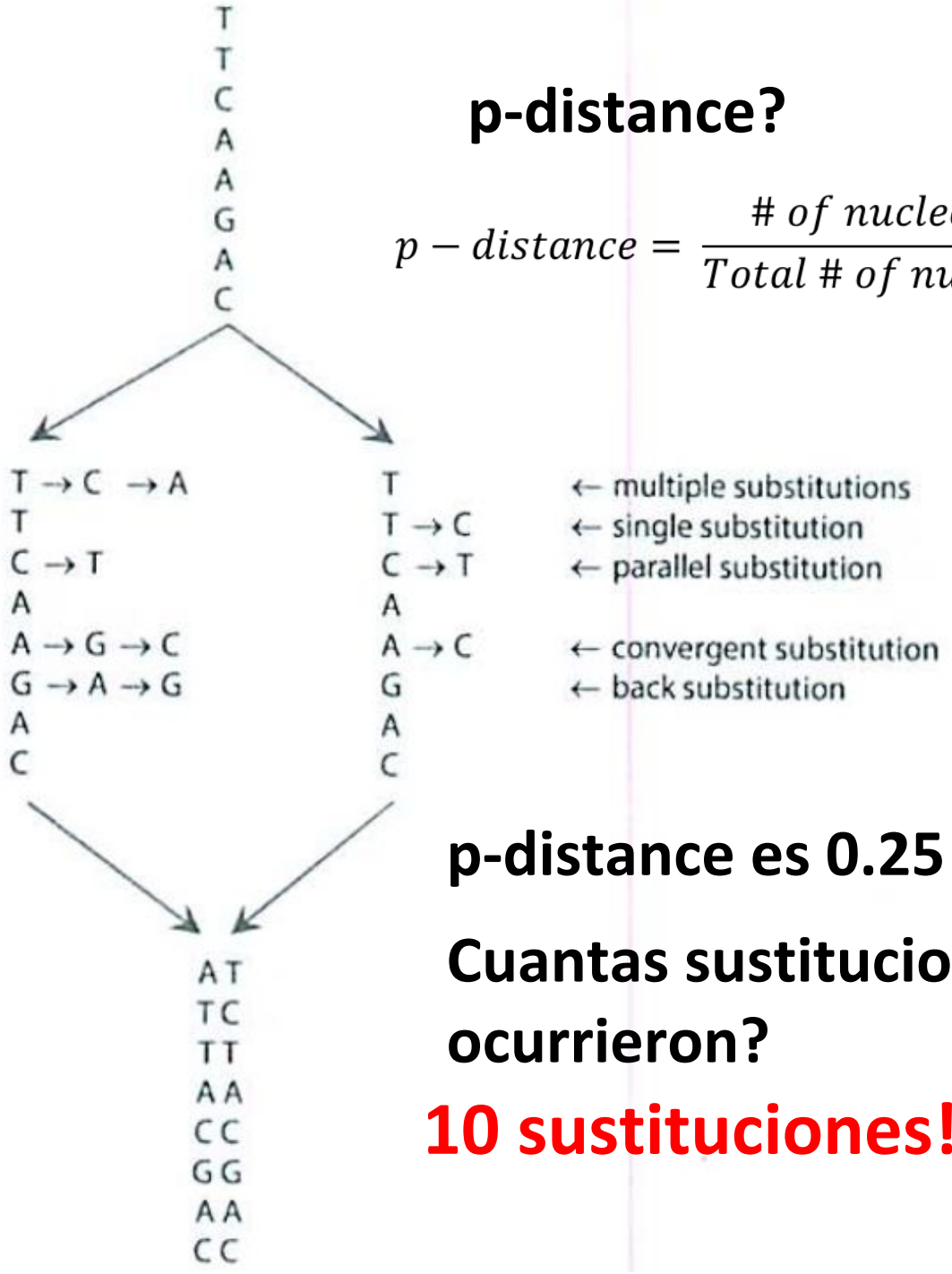
Sustitución nucleotídica





p-distance?

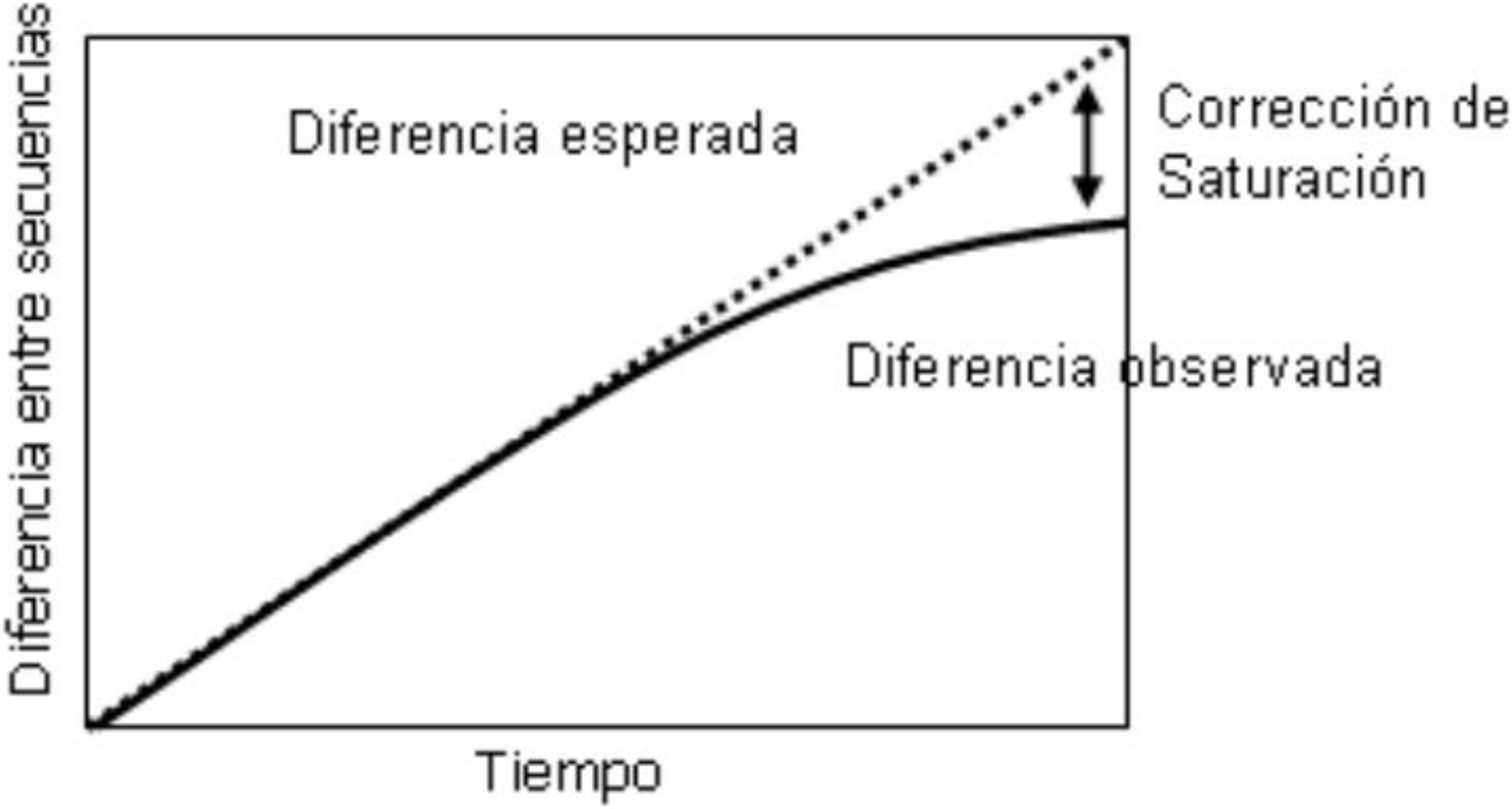
$$p - distance = \frac{\#\ of\ nucleotide\ difference}{Total\ \#\ of\ nucleotide\ compared}$$



p-distance es 0.25 (2/8)

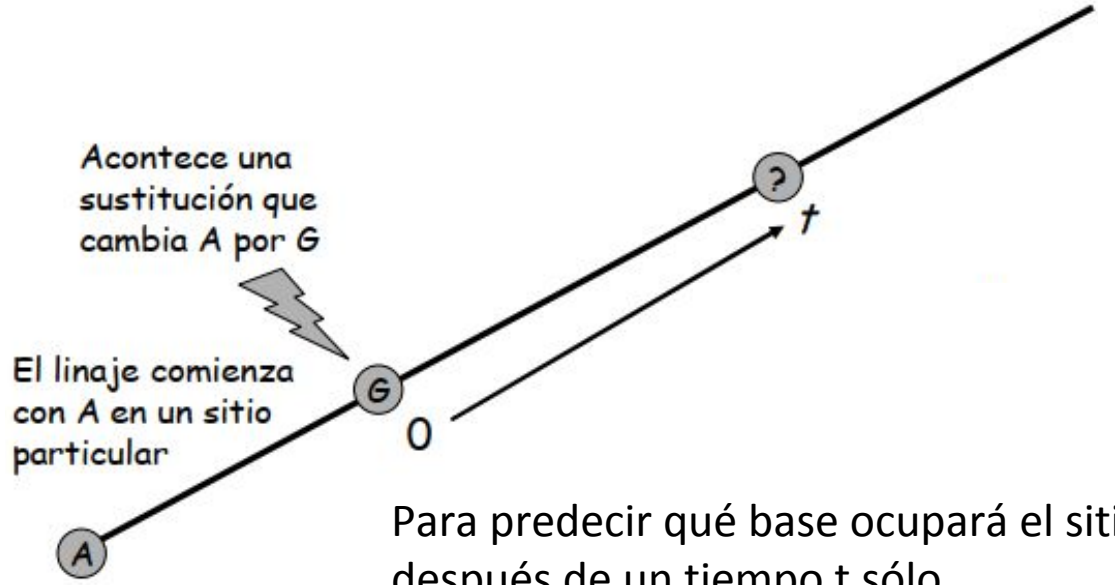
Cuantas sustituciones ocurrieron?

10 sustituciones!!


Diferencia entre secuencias
Diferencia esperada
Corrección de Saturación
Diferencia observada
Tiempo

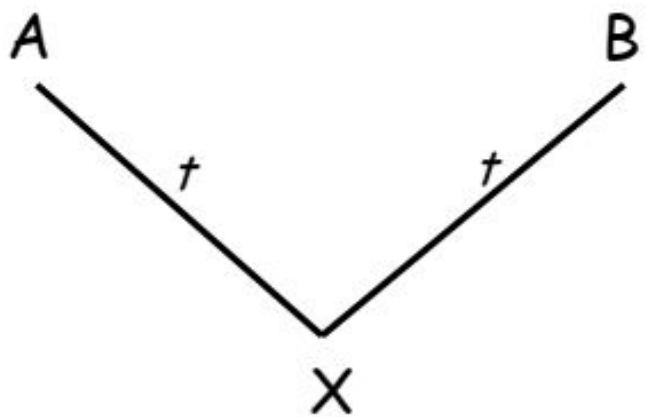
Modelos de sustitución nucleotídica

- Consideraciones generales:
  - **Modelos de Márkov** (la probabilidad de que un evento ocurra solo depende del estado anterior).
  - **Homogeneidad** (las probabilidades de sustitución no varían a lo largo del tiempo y de linajes)
  - **Estacionalidad** (equilibro de base de frecuencia)



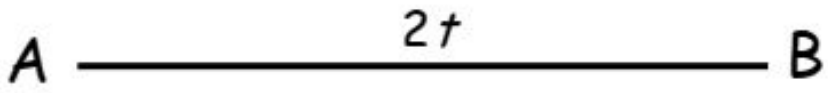
Para predecir qué base ocupará el sitio después de un tiempo t sólo necesitamos conocer qué base lo ocupó en el tiempo 0 (G en este caso). El hecho que A lo ocupó anteriormente es irrelevante en un modelo de Markov

Time-homogeneous time-continuous stationary Markov Models



- Los **modelos** de sustitución empleados para hacer reconstrucciones filogenéticas mediante métodos de distancia o de máxima verosimilitud tienen la propiedad de **reversibilidad temporal**, lo que quiere decir que la probabilidad de cambio es independiente del lugar donde esté la raíz.
Ello es muy conveniente, ya que permite trabajar:

1) con árboles no enraizados y
2) con caracteres no ordenados



- Así la línea que conecta a A y B puede ser considerada como la única rama de un árbol no enraizado que contiene sólo estos dos nodos

Matriz Q o matriz de tasas

$$Q = \begin{array}{c} \begin{matrix} \text{A} & \text{C} & \text{G} & \text{T} \end{matrix} \\ \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu i\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu j\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{bmatrix} \end{array} \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{matrix}$$

μ es la tasa de sustitución general

π_X frecuencias de bases ($_X$: A, C, G o T)

Las letras “a” a “l” representan la frecuencia de posibles sustituciones.

Matriz Q o matriz de tasas

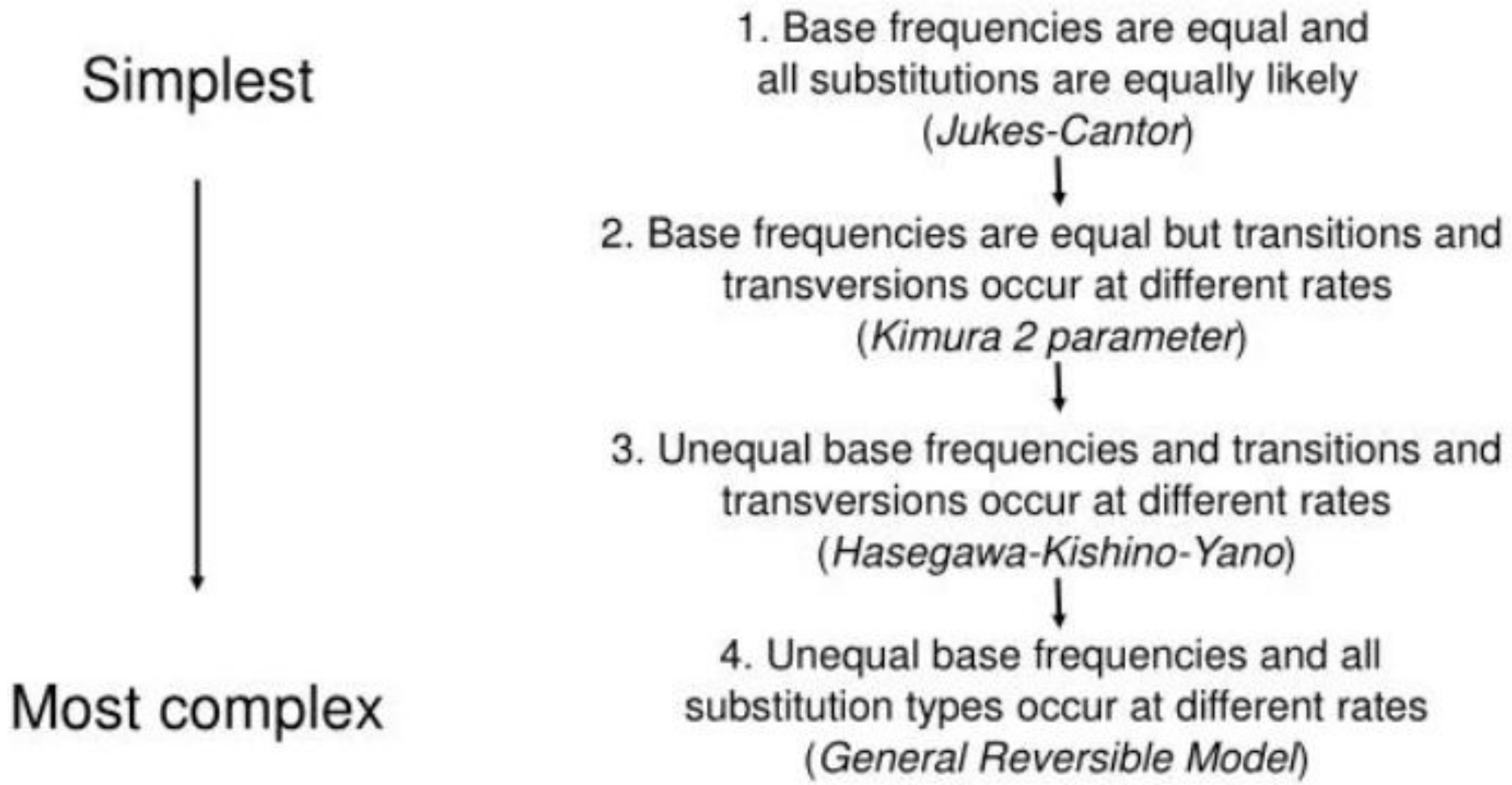
$$Q = \begin{array}{c} \begin{matrix} \text{A} & \text{C} & \text{G} & \text{T} \end{matrix} \\ \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{bmatrix} \end{array} \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{matrix}$$

μ es la tasa de sustitución general

π_X frecuencias de bases ($_X$: A, C, G o T)

Las letras “a” a “f” representan la frecuencia de posibles sustituciones. Se pueden obtener hasta 6 tipos de sustitución diferentes (A – C, A – G, A – T, C – G, C – T, G – T) (**Reversibles**)

Modelos de sustitución nucleotídica



Acomodan sesgo ti/tv

JC69 Jukes-Cantor 1969

	A	G	C	T
A	–	α	α	α
G	α	–	α	α
C	α	α	–	α
T	α	α	α	–

$\pi_A = \pi_C = \pi_G = \pi_T$

todas las sustituciones tienen igual tasa

$\alpha = \beta$

Acomodan distintas frecuencias de base

K80 ó Kimura 2 parámetros (K2P)

	A	G	C	T
A	–	α	β	β
G	α	–	β	β
C	β	β	–	α
T	β	β	α	–

$\pi_A = \pi_C = \pi_G = \pi_T$

Distintas tasas de sustitución ti y tv

$\alpha \neq \beta$

F81/TN84 Felsenstein 1981 Tamura-Nei 1984

	A	G	C	T
A	–	π_G	π_C	π_T
G	π_A	–	π_C	π_T
C	π_A	π_G	–	π_T
T	π_A	π_G	π_C	–

$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

Igual tasas de sustitución ti y tv

$\alpha = \beta$

$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

$\alpha \neq \beta$

HKY85 Hasegawa-Kishino-Yano

	A	G	C	T
A	–	$\alpha\pi_G$	$\beta\pi_C$	$\beta\pi_T$
G	$\alpha\pi_A$	–	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	–	$\alpha\pi_T$
T	$\beta\pi_A$	$\beta\pi_G$	$\alpha\pi_C$	–

2 tasas de sustitución

F84 Felsenstein 1984

	A	G	C	T
A	–	$(1+\frac{\kappa}{\pi_R})\beta\pi_G$	$\beta\pi_C$	$\beta\pi_T$
G	$(1+\frac{\kappa}{\pi_R})\beta\pi_A$	–	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	–	$(1+\frac{\kappa}{\pi_Y})\beta\pi_T$
T	$\beta\pi_A$	$\beta\pi_G$	$(1+\frac{\kappa}{\pi_Y})\beta\pi_C$	–

3 tasas de sustitución

TN93 Tamura-Nei 1993

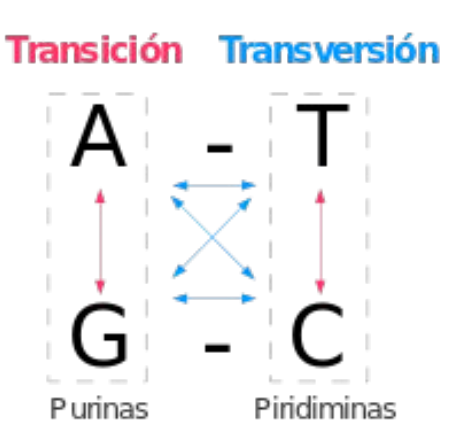
	A	G	C	T
A	–	$\alpha_1\pi_G$	$\beta\pi_C$	$\beta\pi_T$
G	$\alpha_1\pi_A$	–	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	–	$\alpha_2\pi_T$
T	$\beta\pi_A$	$\beta\pi_G$	$\alpha_2\pi_C$	–

3 tasas de sustitución

GTR Modelo general tiempo reversible

	A	G	C	T
A	–	$a_1\pi_G$	$a_2\pi_C$	$a_3\pi_T$
G	$a_1\pi_A$	–	$a_4\pi_C$	$a_6\pi_T$
C	$a_2\pi_A$	$a_4\pi_G$	–	$a_5\pi_T$
T	$a_3\pi_A$	$a_6\pi_G$	$a_5\pi_C$	–

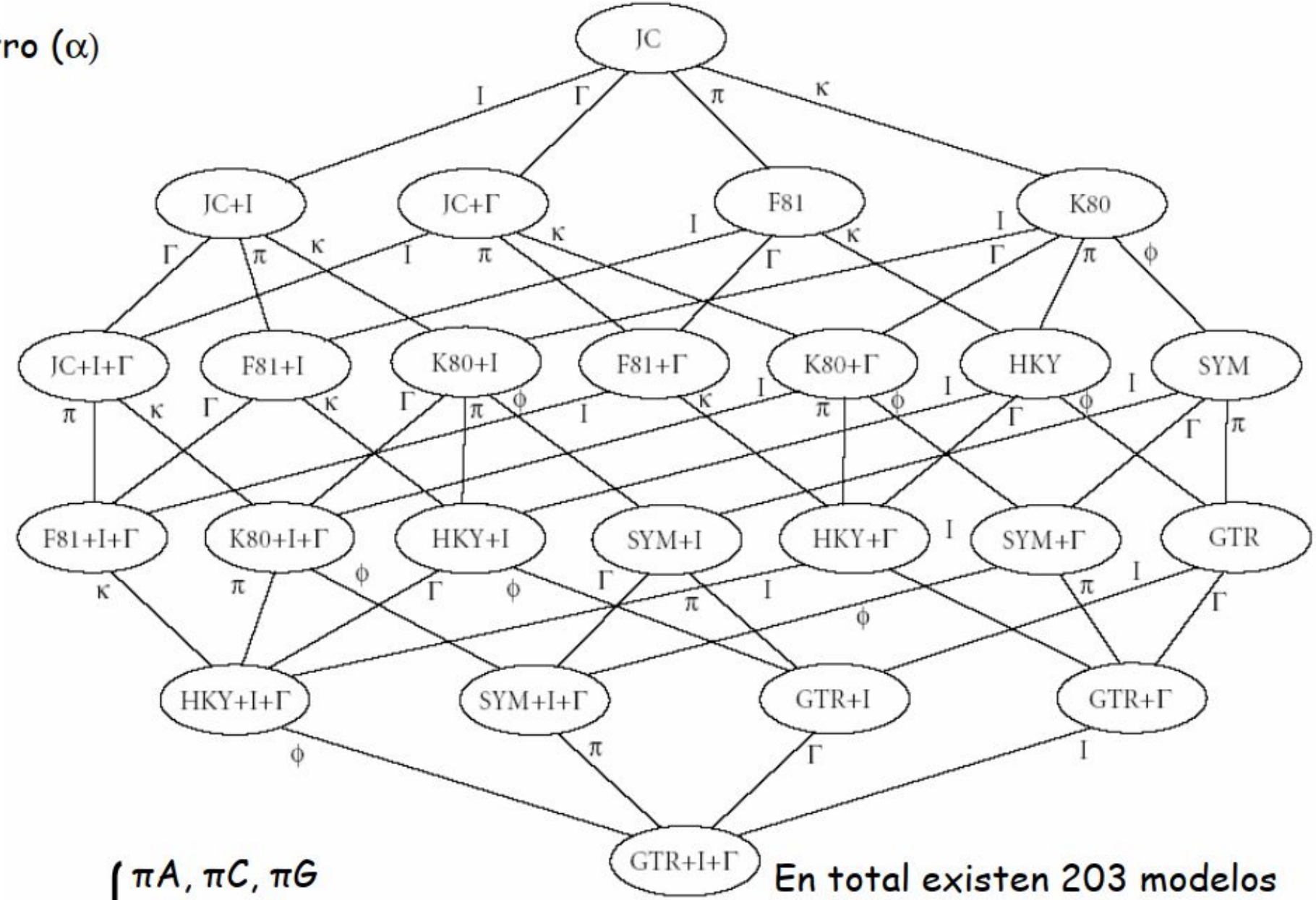
6 tasas de sustitución



1 parámetro (α)

Incremento en el número de parámetros

modelos más generales



11 parámetros libres a estimar

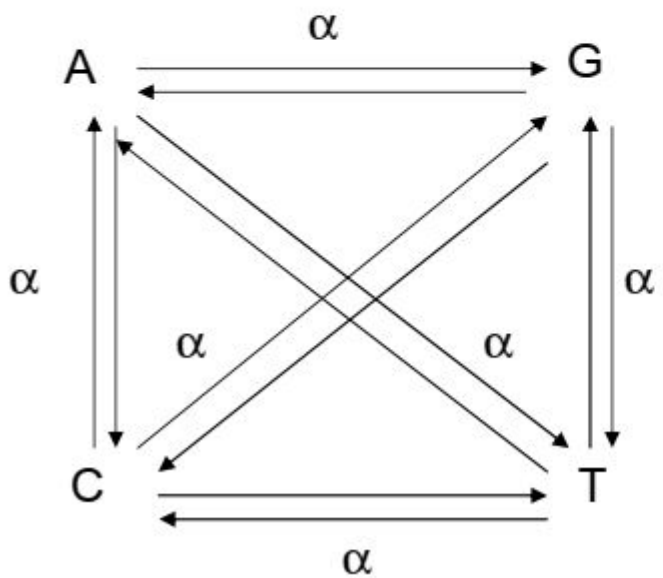
- πA, πC, πG
- a, b, c, d, e
- μ,
- I, T

En total existen 203 modelos posibles en la familia GTR al combinar params. de frec., tasa, G e I
La mayoría de ellos carecen de nombre

Modelos de sustitución y distancia evolutiva

Distancia de Jukes-Cantor

- La distancia evolutiva más sencilla es la basada en el modelo de Jukes y Cantor (JC69)
  - asume que todas las sustituciones ocurren con la misma tasa α
  - asume que los nts tienen la misma frecuencia 0.25



- El único parámetro del modelo JC69 es α

(tasa de sustitución instantánea)

Modelos de sustitución y distancia evolutiva

Distancia de Jukes-Cantor

- La **tasa** de sustitución ha de ser multiplicada **por** un intervalo de **tiempo** para poder obtener el **no. de sustituciones** acontecidas en dicho intervalo. Dado que tenemos que estimar la tasa, nunca podemos saber con exactitud el no. de sust. que han ocurrido entre 2 secuencias. De ahí que se habla del **no. esperado de sustituciones**, que a su vez es dependiente del modelo empleado para hacer la inferencia
- Si contamos con una estima de la tasa de sustitución y conocemos el intervalo de tiempo (t), entonces podemos obtener la **distancia evolutiva** simplemente multiplicando la **tasa x tiempo**

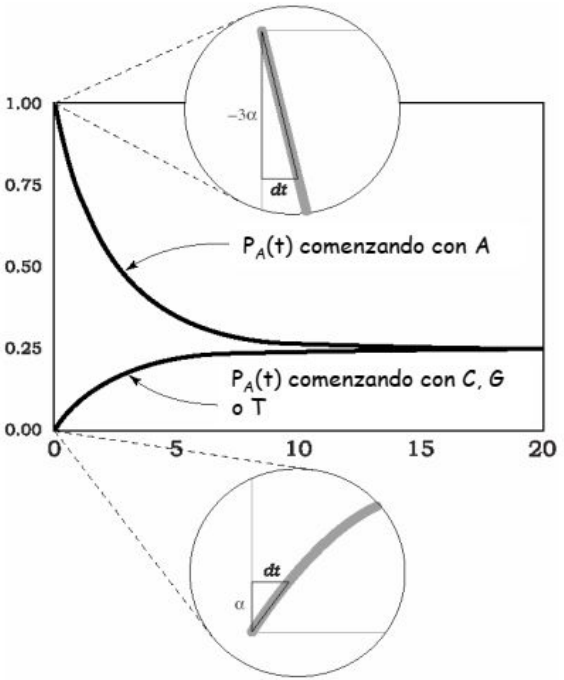
estado "de" \ estado "hacia"	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

Matriz de sustitución instantánea para el modelo JC69

Modelos de sustitución y distancia evolutiva
Distancia de Jukes-Cantor

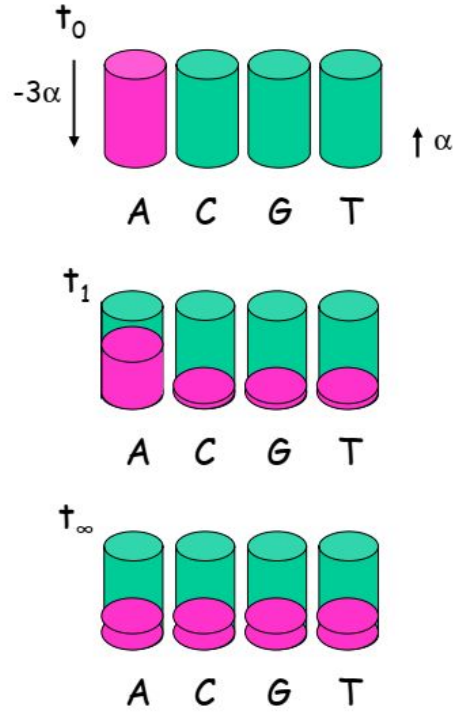
estado "de" \ estado "hacia"	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

- Gráfica de la **función de probabilidad de transición** (de la "presencia de A") en función del tiempo

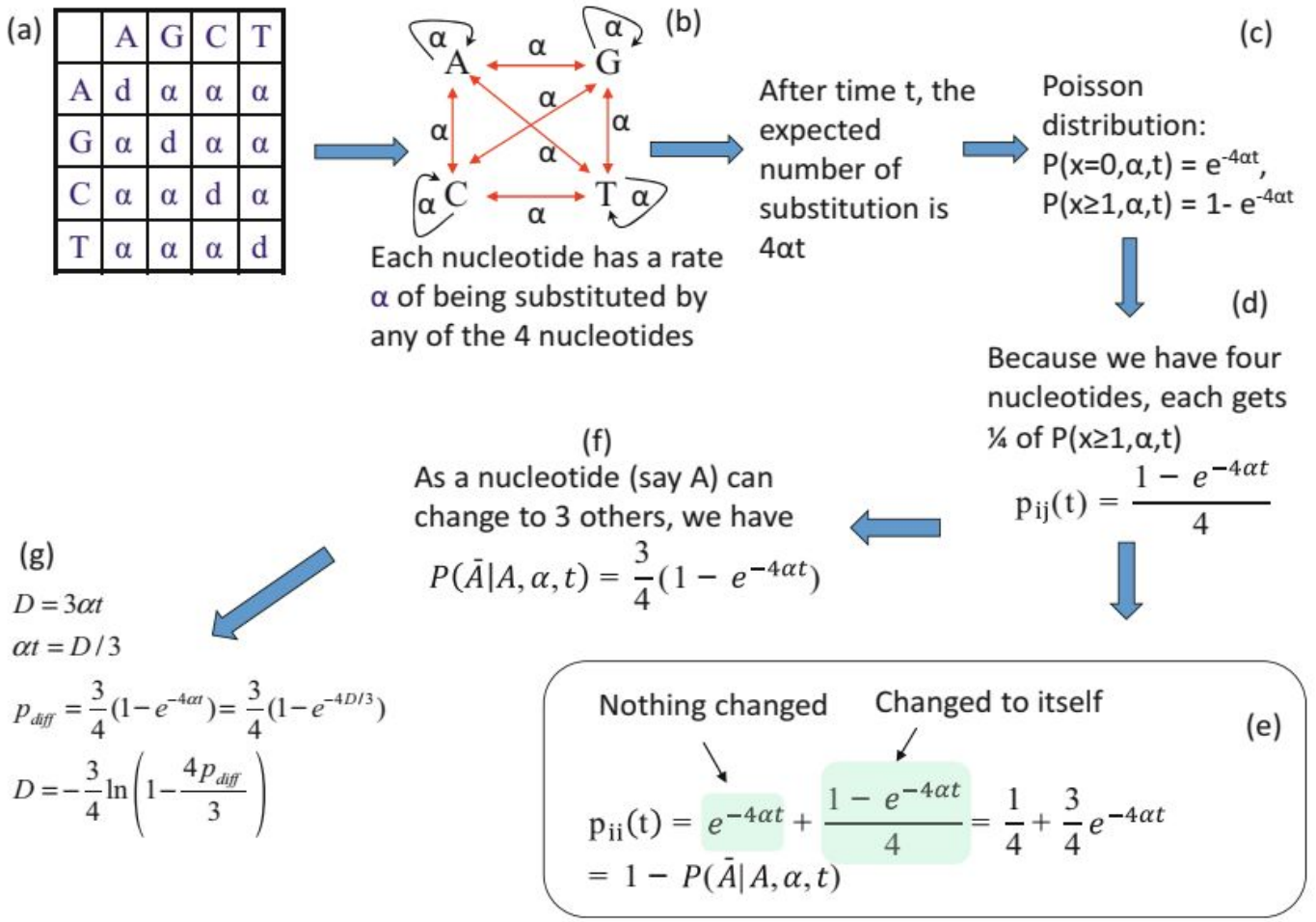


- la curva superior asume que partimos de A ocupando una posición particular en el momento 0. Con el tiempo la probabilidad de continuar viendo una A en ese sitio **decae**, ya que su tasa de sustitución a otro resíduo es α (y de quedar igual es -3α).

- A partir de las tasas instantáneas de transic. podemos por tanto calcular la probabilidad de que podamos observar A en un sitio particular después de un tiempo *t*



Modelos de sustitución y distancia evolutiva
Distancia de Jukes-Cantor



Distribución de Poisson: distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo

Modelos de sustitución y distancia evolutiva
Distancia de Jukes-Cantor

Thus, the transition probability matrix for the JC69 model is (in the order of A, G, C, and T)

$$P_{\text{JC69}} = \begin{bmatrix} \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} \end{bmatrix} \quad (12.4)$$

$$p_{\text{diff}} = 3p_{ij}(t) = \frac{3}{4} - \frac{3}{4}e^{-4\alpha t}$$

$$p_{diff} = \frac{3}{4}(1 - e^{-4\alpha t})$$

- Una estimación de p_{diff} (p_{ij}) es la proporción observada de diferencia de sitios entre dos sitios: ***p-distance***
- Sustituyendo αt con d, nosotros finalmente obtenemos la formula de corrección de Jukes-Cantor entre dos secuencias:

$$d = -\,3/4\ ln\ (1-\,4/3\ p)$$

Calcular distancia JC

Fórmula de corrección de Jukes-Cantor para la distancia genética d entre dos secuencias:

$$d = -\,3/4\ ln\ (1-\,4/3\ p)$$

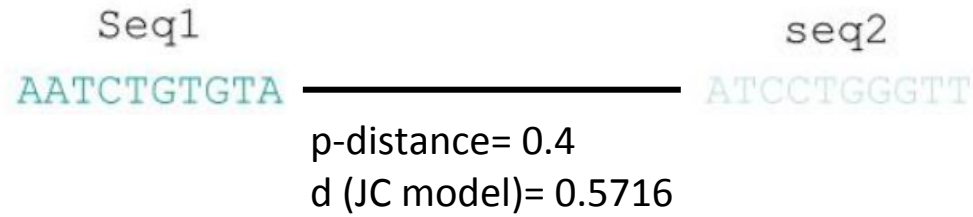
Ejemplo

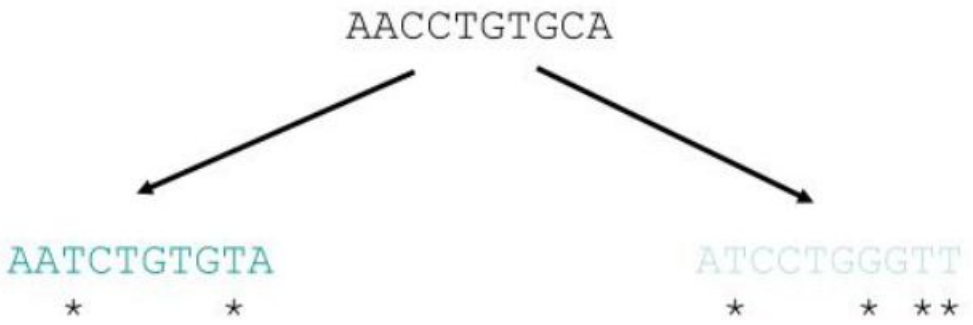
```
Seq1    AATCTGTGTA
seq2    ATCCTGGGTT
         **   *  *
```

$$p - distance = \frac{\#\ of\ nucleotide\ difference}{Total\ \#\ of\ nucleotide\ compared}$$

$$p - distance = \frac{4}{10} = 0.4$$

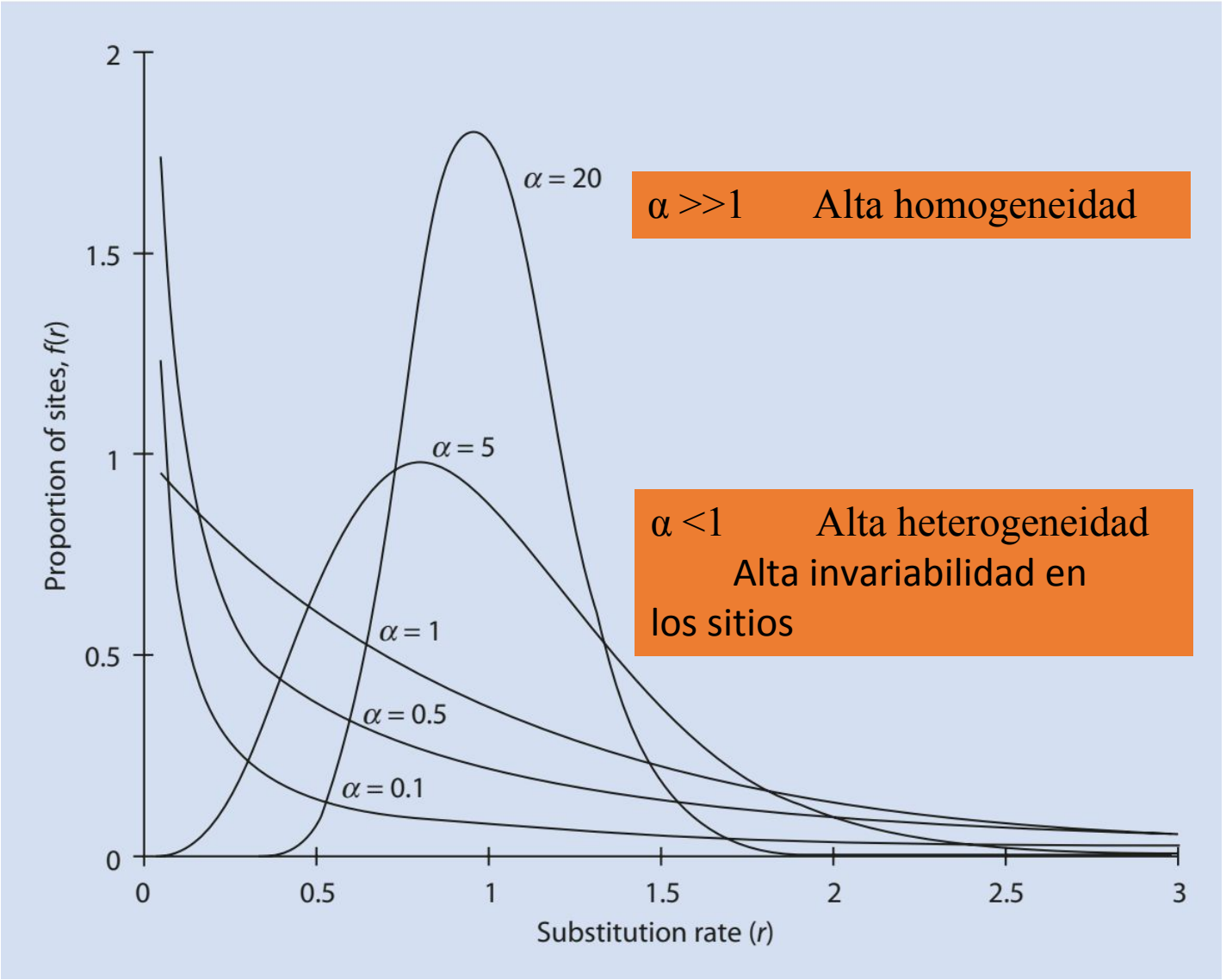
d (JC model) = - 3/4 ln [1- 4/3 (0.4)] = **0.5716**





Heterogeneidad de los datos

- Uno de los supuestos fundamentales de los análisis filogenéticos es la independencia entre los caracteres empleados
- Entonces, ¿cada columna tiene una tasa de cambio diferente?
- Distribución tipo gamma (+G ó +Γ). Se utiliza una distribución estadística para permitir que diferentes sitios caigan en categorías de diferentes tasas de sustitución.
- Proporción de sitios invariables (+I)



α >>1 Alta homogeneidad

α <1 Alta heterogeneidad
Alta invariabilidad en los sitios

¿Cómo seleccionar el modelo adecuado?