

# Evolución de familias multigénicas 2020

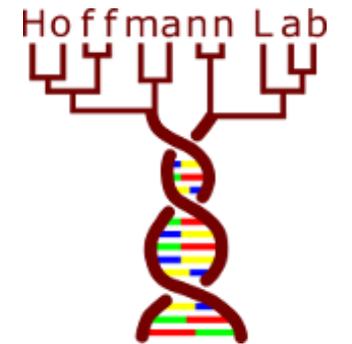
---

## Introducción al análisis filogenético (2)



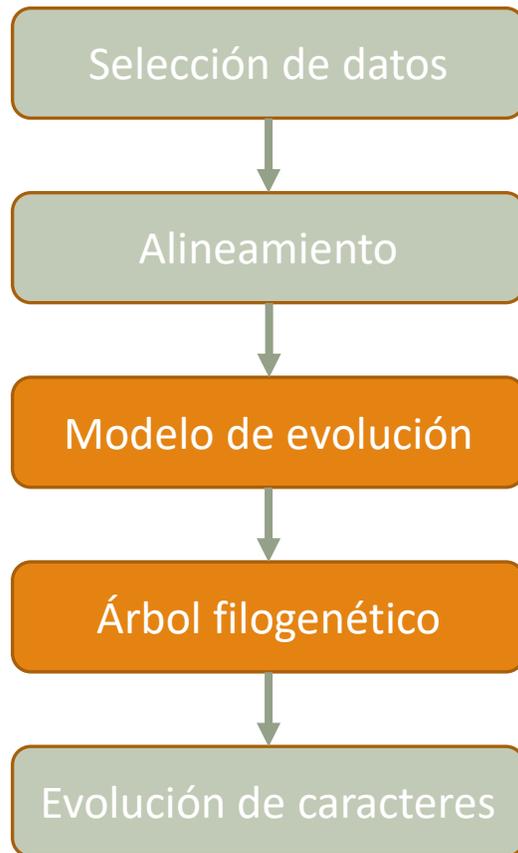
FACULTAD DE  
**CIENCIAS**

UDELAR | [fcien.edu.uy](http://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 (maximum likelihood).
  - Inferencia bayesiana.

# Modelos anidados y no anidados

---

El origen está en los modelos lineales y el trabajo pionero de Ronald Fisher.

Vamos a considerar:

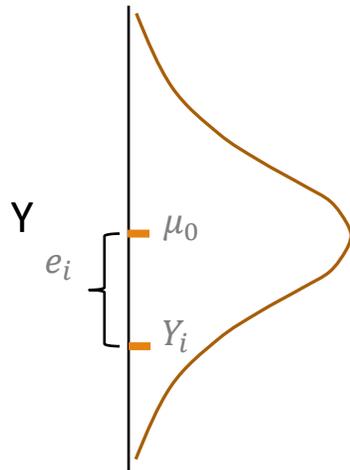
- una variable de interés  $Y$  (variable dependiente)
- Variables potencialmente explicativas (independientes):
  - Continuas:  $X_1, X_2, \dots$
  - Discreta:  $\mu_0, \mu_1, \mu_2, \dots$

Un ejemplo:

- $Y$ : longitud del húmero
- ( $X_1$ : longitud del cuerpo)
- $\mu$ : constante(s) para representar media de  $Y$

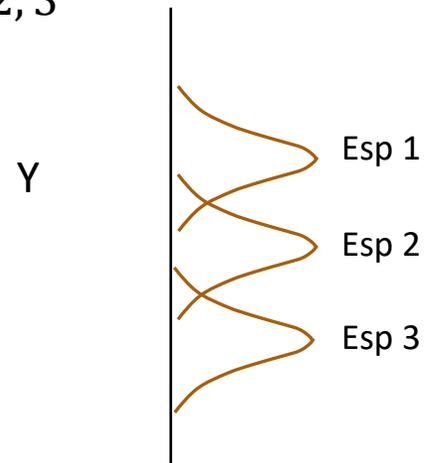
# Modelos lineales clásicos

$$Y_i = \mu_0 + e_i$$

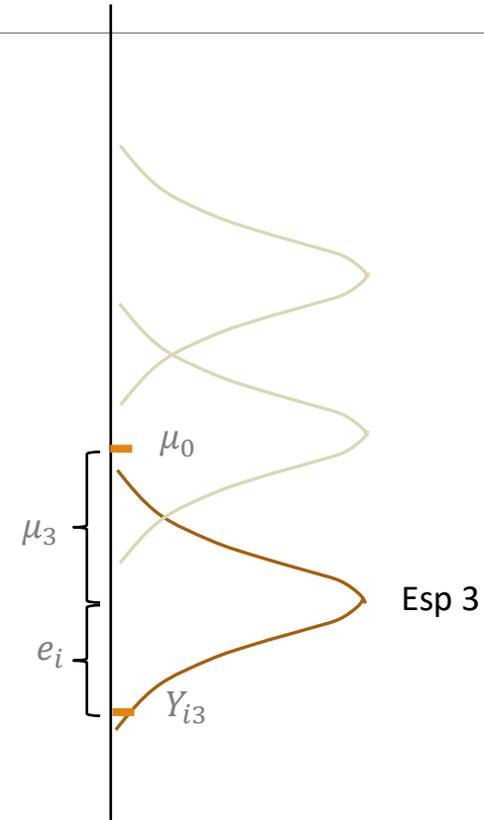


Modelo simple ( $H_0$ )

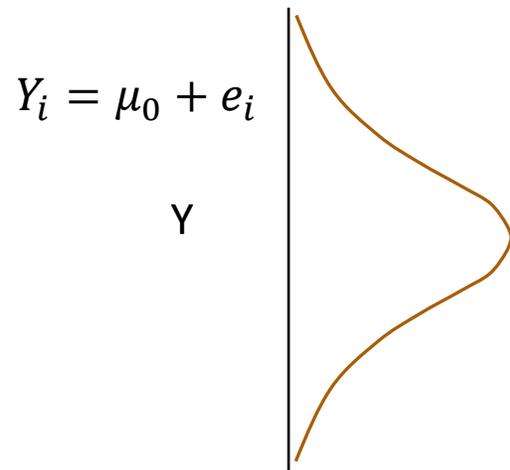
$$Y_{ij} = \mu_0 + \mu_j + e_i$$
$$j = 1, 2, 3$$



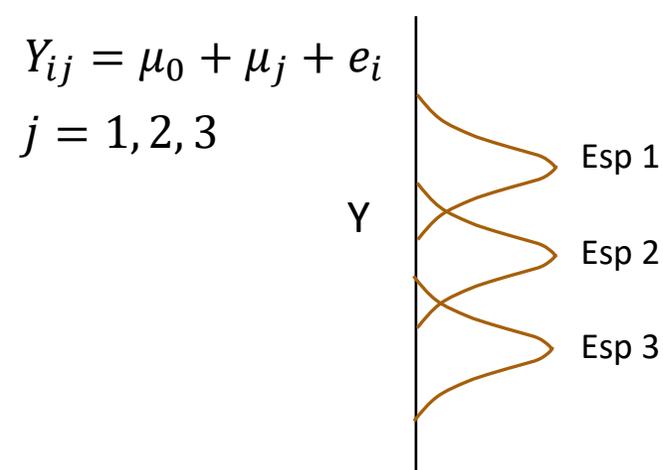
Modelo "completo"



# Modelos lineales clásicos



Modelo simple ( $H_0$ )



Modelo "completo"

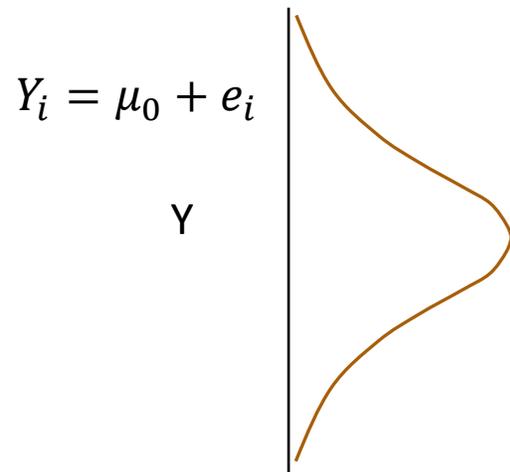
El modelo simple está anidado en el completo, del que es un caso particular ( $\mu_j = 0$ ).

Corresponde a la hipótesis nula de que no hay diferencias de medias entre poblaciones.

La selección entre estos dos modelos implica resolver si se justifica estadísticamente rechazar el modelo simple en favor del completo.

Prueba de hipótesis clásica: ANOVA

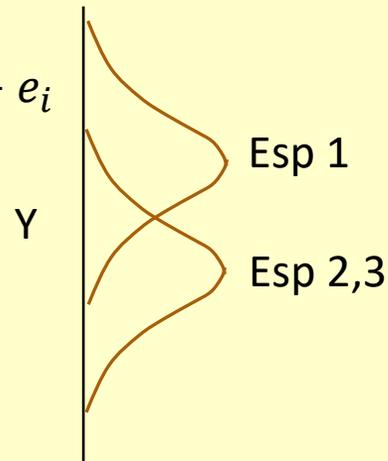
# Modelos lineales clásicos



Modelo simple ( $H_0$ )

$$Y_{ij} = \mu_0 + \mu_j + e_i$$

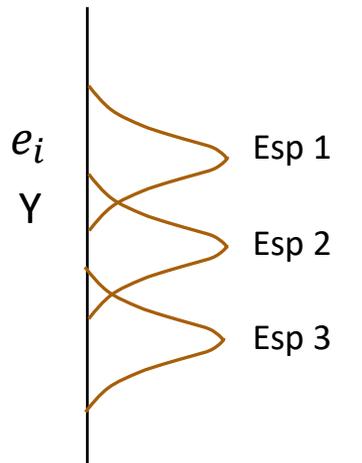
$j = 1, (23)$



Modelo intermedio

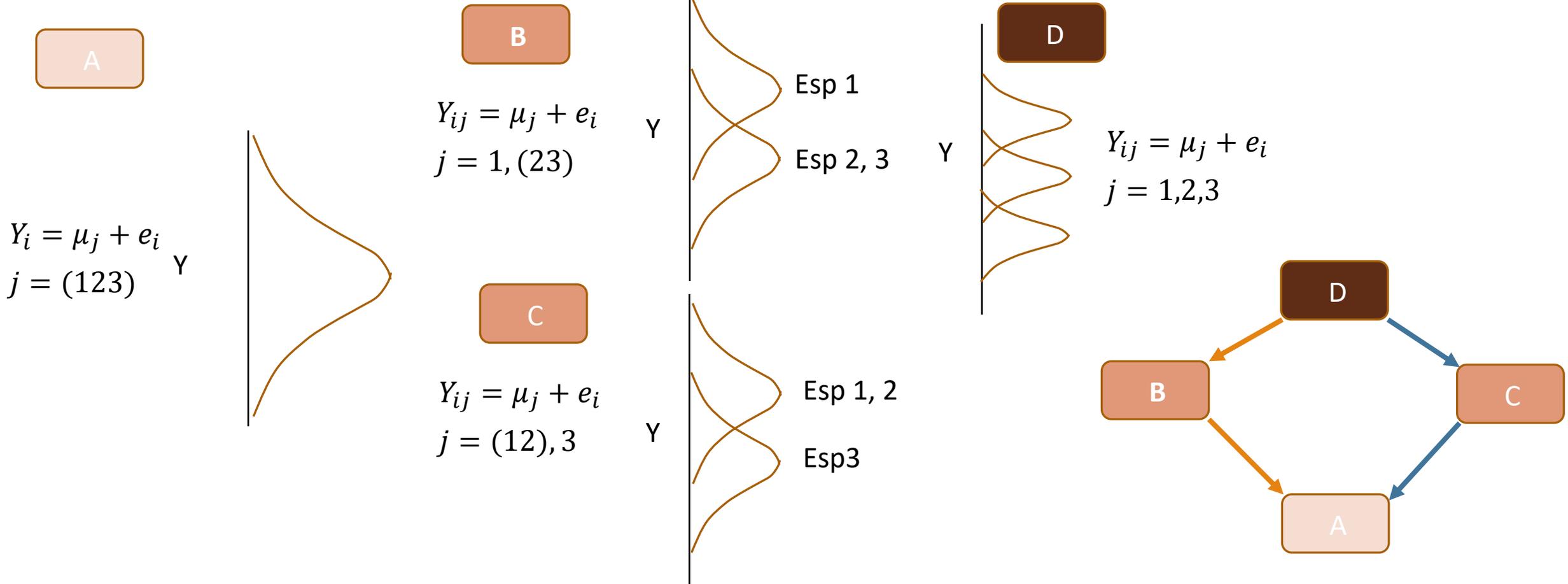
$$Y_{ij} = \mu_0 + \mu_j + e_i$$

$j = 1, 2, 3$



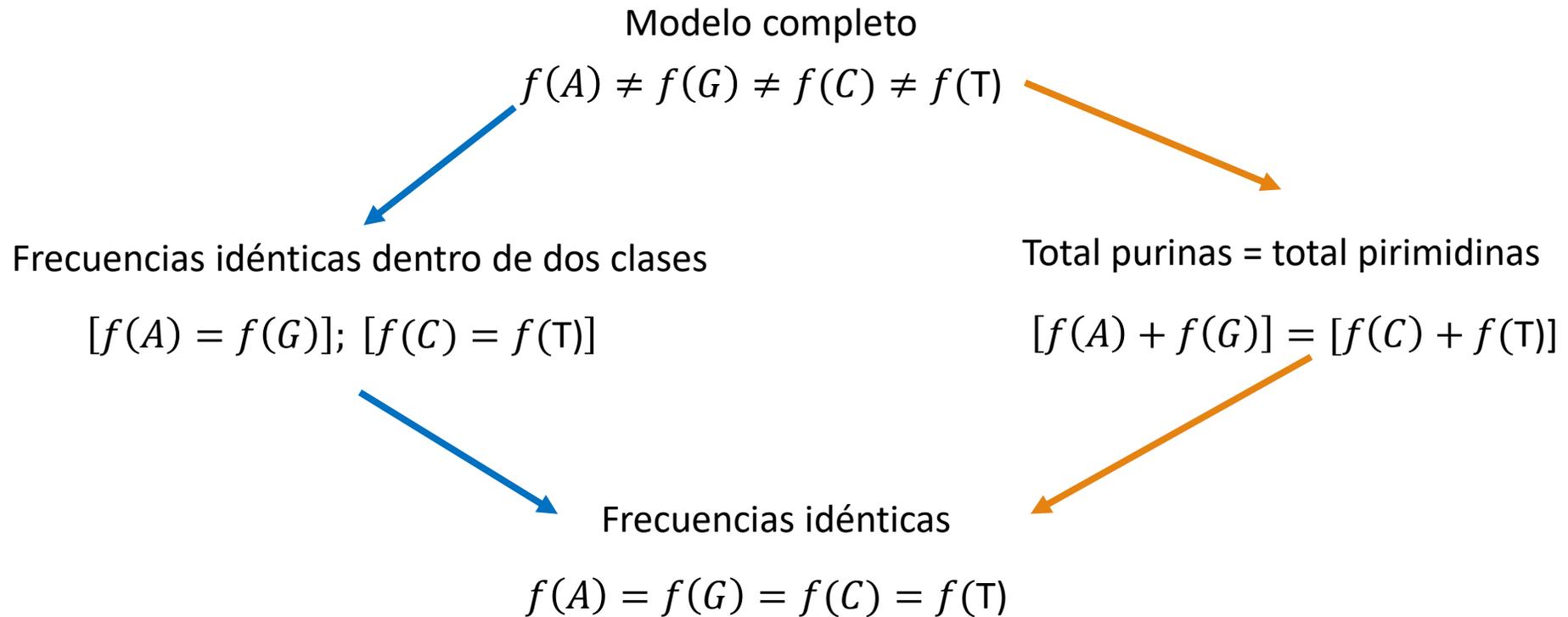
Modelo "completo"

# Modelos lineales clásicos



# Modelos anidados y no anidados

Frecuencias nucleotídicas en secuencias de ADN (consideramos una hebra):



# Selección de modelos

---

## Modelos anidados:

- Agregar parámetros permite un mejor ajuste a los datos (se reduce el término  $e_i$  de error o incertidumbre)
- Múltiples pruebas de hipótesis (desde el ANOVA hasta el presente) orientadas a evaluar si el rechazo de un modelo simple (por ejemplo,  $H_0$ ) en favor de uno más complejo está justificado.
- Una clásica aplicada en selección de modelos moleculares es el LRT (likelihood ratio test).

## Modelos no anidados:

- No tan fáciles de poner a prueba, pero existen pruebas aproximadas.
- Una aproximación se basa en un trabajo de Akaike, que encuentra una función de cómo crece el ajuste del modelo con el número de parámetros bajo una hipótesis nula, y por tanto permite ordenar modelos por su ajuste aunque no estén anidados.
- Una prueba muy usada en esta dirección es el llamado AIC (Akaike information content).

# Dos modelos anidados de sustitución nucleotídica

El modelo de Jukes y Cantor tiene un solo parámetro ( $\mu$ ).

	G	A	T	C
G	$1-3\mu$	$\mu$	$\mu$	$\mu$
A	$\mu$	$1-3\mu$	$\mu$	$\mu$
T	$\mu$	$\mu$	$1-3\mu$	$\mu$
C	$\mu$	$\mu$	$\mu$	$1-3\mu$

Un modelo popular de dos parámetros es el de Kimura, que distingue transiciones y transversiones:

	G	A	T	C
G	$1-\alpha-2\beta$	$\alpha$	$\beta$	$\beta$
A	$\alpha$	$1-\alpha-2\beta$	$\beta$	$\beta$
T	$\beta$	$\beta$	$1-\alpha-2\beta$	$\alpha$
C	$\beta$	$\beta$	$\alpha$	$1-\alpha-2\beta$