

# **Análisis Estadístico Básico de Secuencias de DNA**

**Informática Aplicada a la Bioquímica  
2º curso del Grado en Bioquímica.**

**Dpto. Ciencias de la Computación e Inteligencia Artificial**

**Profesor: Francisco J. Romero Campero**

**Universidad de Sevilla**



# Guión de la Unidad

- Introducción histórica
- Definiciones Básicas
- Modelos de Secuencias de DNA
  - Modelos multinomiales:
    - Estimación de modelos multinomiales
    - Composición de bases
    - Contenido en GC
  - Modelos markovianos:
    - Estimación de modelos markovianos
    - Frecuencia de k-meros
- Sesgos en el Uso de Codones

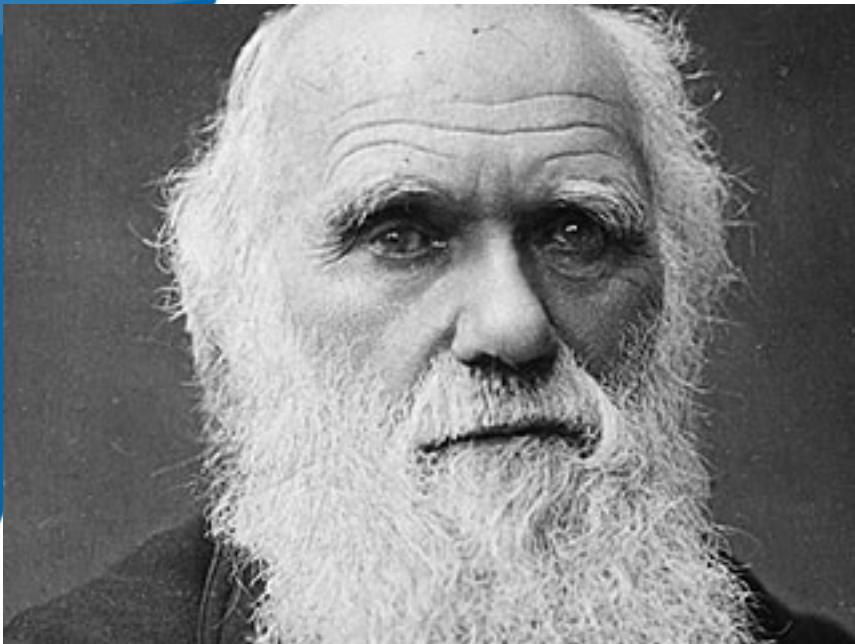


# Guión de la Unidad

- **Introducción histórica**
- **Definiciones Básicas**
- **Modelos de Secuencias de DNA**
  - Modelos multinomiales:
    - Estimación de modelos multinomiales
    - Composición de bases
    - Contenido en GC
  - Modelos markovianos:
    - Estimación de modelos markovianos
    - Frecuencia de k-meros
- **Sesgos en el Uso de Codones**



# Antes de la Era Genómica



**Charles Darwin publica en 1866 “On the Origin of Species”**

**Mendel publica en 1859 su tratado sobre la herencia de caracteres, leyes de Mendel**

# Antes de la Era Genómica

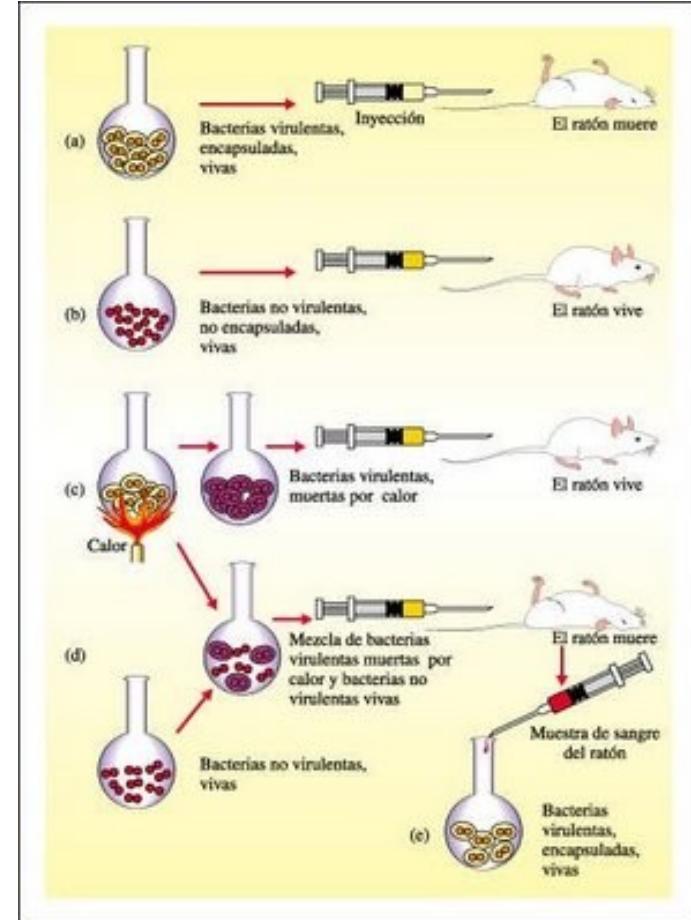


Nuclein



**Johannes Friedrich Miescher descubrió el DNA en 1869**

# Antes de la Era Genómica



**Frederick Griffith descubrió la existencia de un factor transformador en 1928.**

# Antes de la Era Genómica



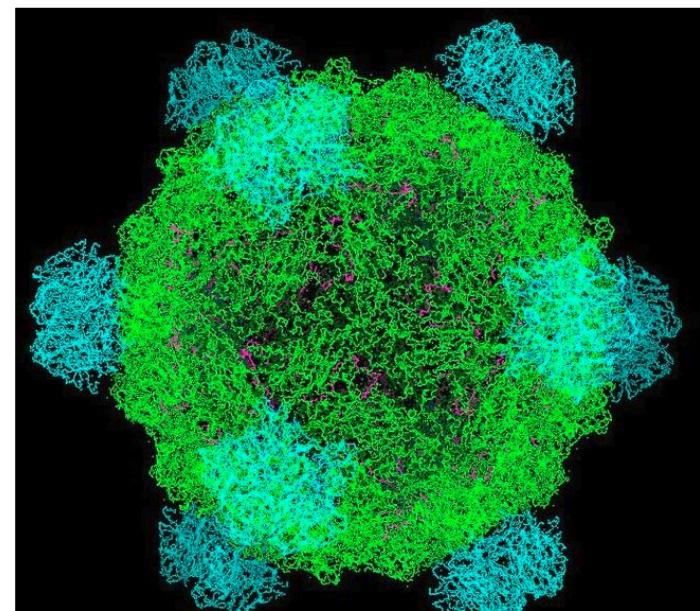
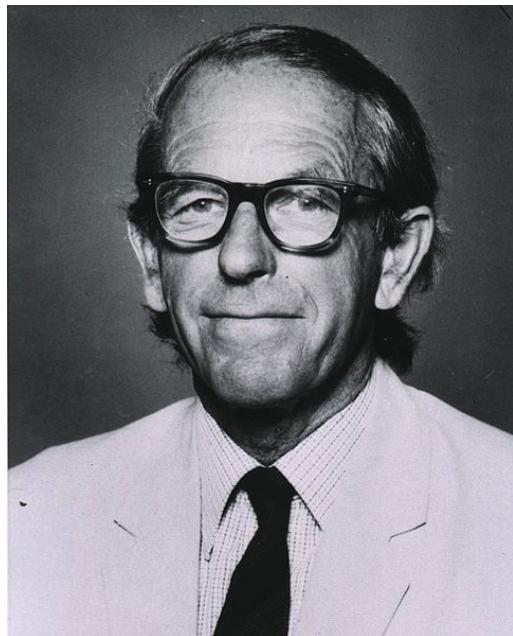
**La estructura en doble hélice del DNA les valió el premio nobel a Watson y Crick en 1962.**

**Rosalind Franklin**

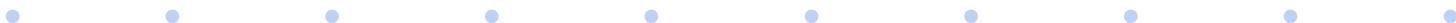


# Los Albores de la Era Genómica

- En 1977 se secuencia el primer genoma, el bacterio fago *phiX174*.
- Frederick Sanger recibió en 1980 el premio nobel por el *método de terminación de la cadena*.

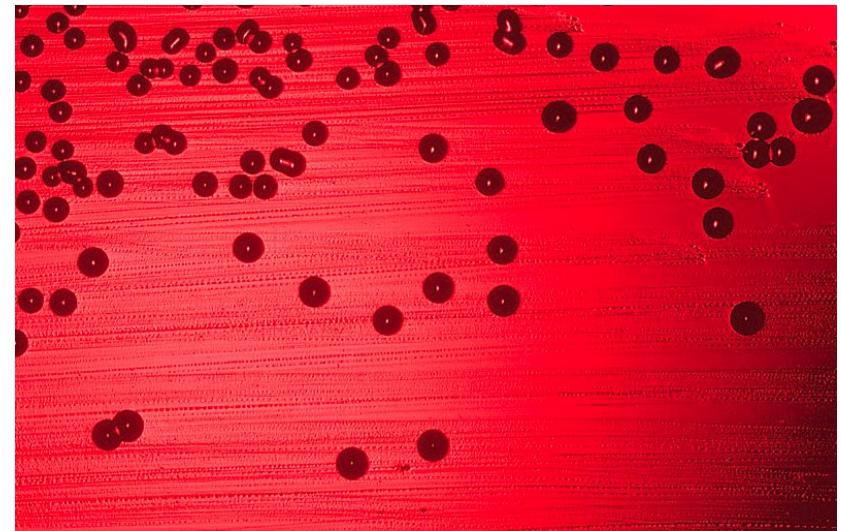
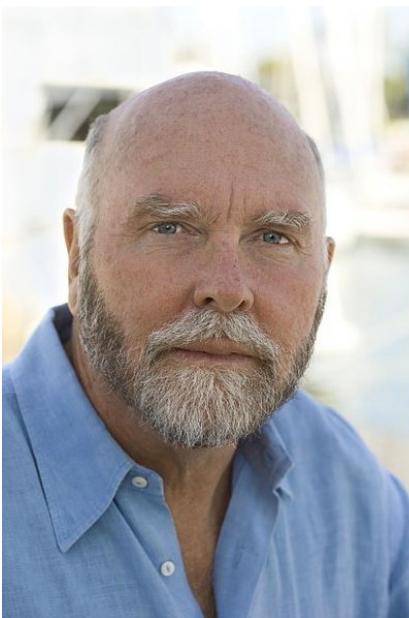


- Con este método se secunciaron genomas pequeños tales como la *mitocondria humana*, el *bacterio fago lambda* y el *virus del HIV*.



# La Era Genómica

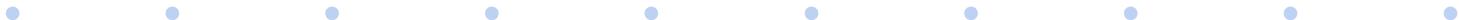
- Craig Venter revolucionó la secuenciación de genomas de cualquier tamaño.
- *Haemophilus influenzae* fue la primera bacteria en secuenciarse en 1995.



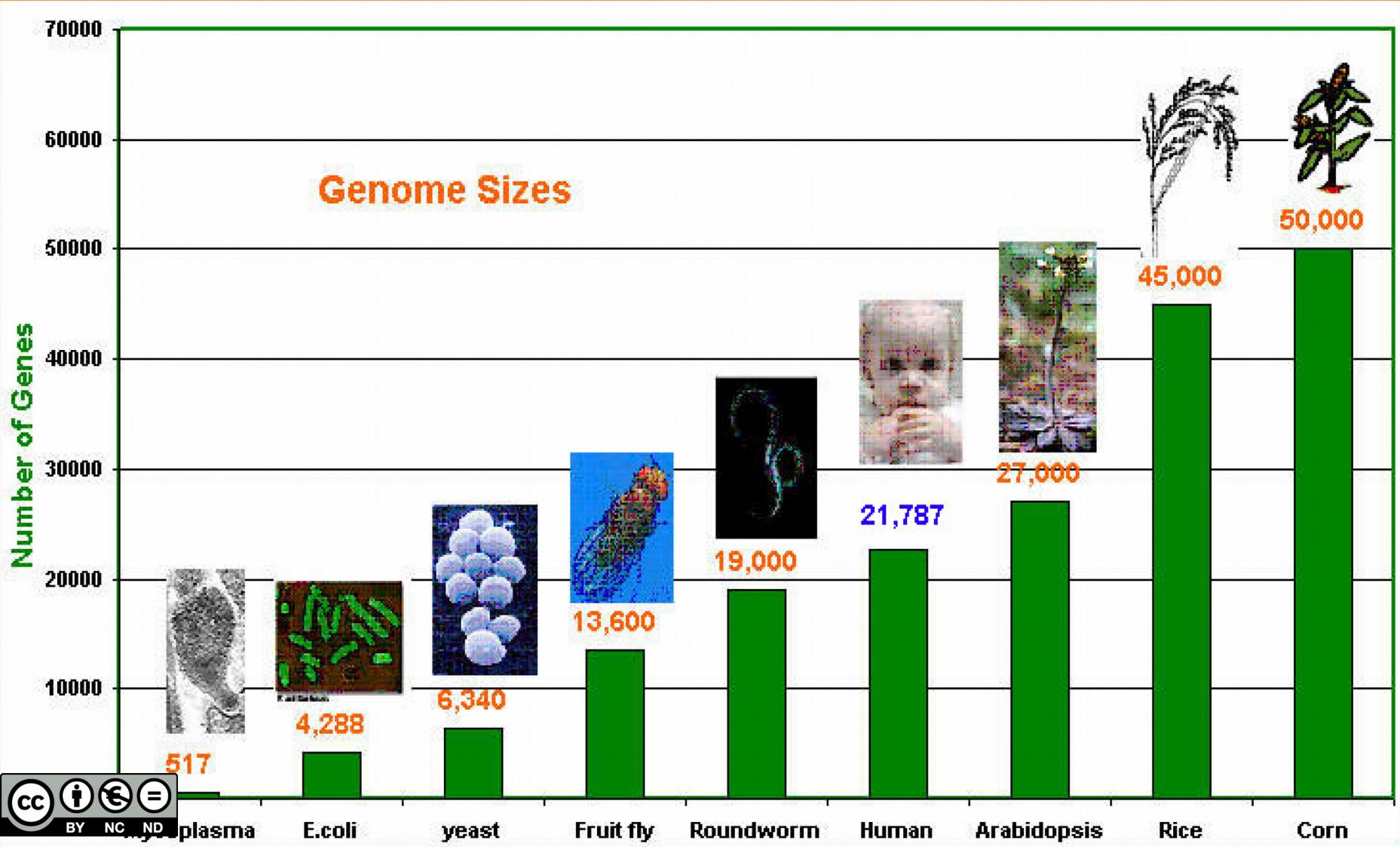
- Le siguieron *Mycoplasma genitalium* (1995) y *Escherichia coli* (1997).
- El primer genoma eucariota en secuenciarse fue el de *Saccharomyces cerevisiae* (1996).

# La Era Genómica

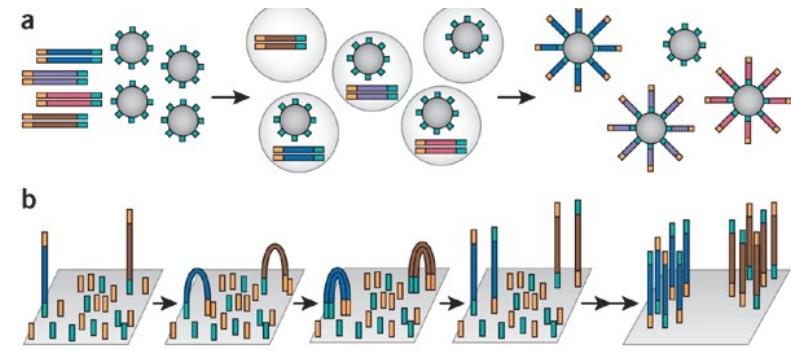
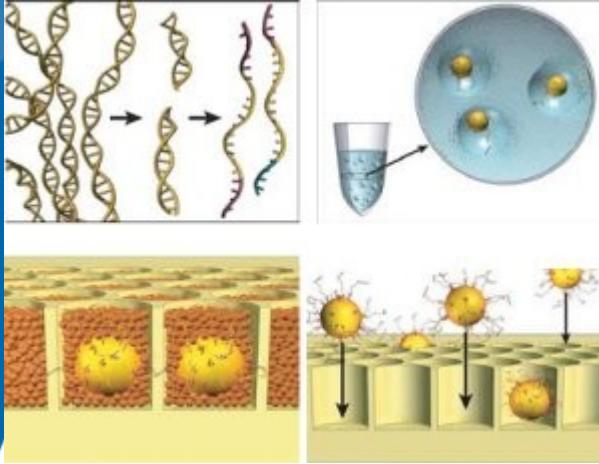
- Los primeros organismos multicelulares en secuenciarse fueron *Caenorhabditis elegans* (1998), *Drosophila melanogaster* (2000) y *Arabidopsis thaliana* (2000).
- El primer borrador de la secuencia de *Homo sapiens* se obtuvo en 2001.



# Tamaños Genómicos



# Secuenciación de Nueva Generación



- Las técnicas de secuenciación de nueva generación como la pirosecuenciación de 454 Roche, Illuminar y Ion protón permiten obtener la secuencia de genomas completos de plantas y animales en menos de una semana.

# Cantidad Masiva de Datos Genómicos

## Una pregunta común

>*Synechocystis* sp. strain PCC6803

CAGAATTAAACACGGATTTAACCTTGGAAATCGTCGGACTCAATTGGGTGTA  
GTCTTGATTCTTCTACTTGCTTCGTTGCCTAACCGGTGCTTGCGCTCCCCC  
GATGTGGCGTGGGTAAAAAAGGTGACATGGGAAGCATTAACTCCCGCAGAAA  
GGGAAAAGTTCCCCCGCTCTGTCCAGATTGTGTTGGAACTACTTCCCCC  
AGTGATGCCCTAGCTGCAACTCAAGAAAAATGCGTGAATATTAAAGCTGTGG  
AACGGAGTTAGGCTGGTTAATTAACTCCGATTGATCAACAAGTGGAAATATATC  
GCCCTGATTCCCCGATAGAAATATTAAACAAACCCCCAAACACTCAATGGCGAT  
CGCCTGTTGCCGGACTTGGAACTCAATGTGGCTTGGTTATGGGGGGCAAAC  
AATTCAAGAATCTCCATGACACTTCAAAGACTTGCCTATTACCTTCATAAGAAAA  
TAGGGATATTAAATAATTGCCGAGTAACAAAGGCACATATGGGTATGATAGCGCG  
AATTTCACCCTAGGACAATTTTTAATAATTAAACACTAGATTAAAGAAAAC  
CTATCAATTATAAAATAATTACCAACAAAATAAGTTACTATATCAACCTCGAT  
AAAATGAAGGTA  
AAAACCTCTGTAC  
GGGCACCCATGA  
ATGATTCAAGCAGTTGACCAACAAAGGGGTGAATGTACCTCTGGTTTGCAC  
CACTGCCTATGCTTACCGCTATTATCCAAGAGGGCGGGTCTGGAAACAAAAAC  
TGCAGAGATCTCTTACAGATCTCGATGTGAATGACATGGCCAATCTCCAGGAG  
CGGGGTCAATTAGCTCGGCAATTAAATCTTAGACACCCCCCTCCCCAAAATTAA  
CAAACGGCGATGCCGAAGCCTATGGTGCCTGTGAACGCTATGGCCA  
AAATGGGTCGTACCGGGGTGGACGTGGCGGTGCGCTCCAGTGCTACAGCGG  
AGGATTGCCAGAGGGCTAGTTGCAGGGACAACAGGAAACCTACCTTAATGTC  
CTTCCCTGTGTTGGAAATCTGCCATAAAATGCTTGCTTCGTTATT  
GGGCTATTCCATGCCATCACAAATGGTTT

¿Qué es esto?

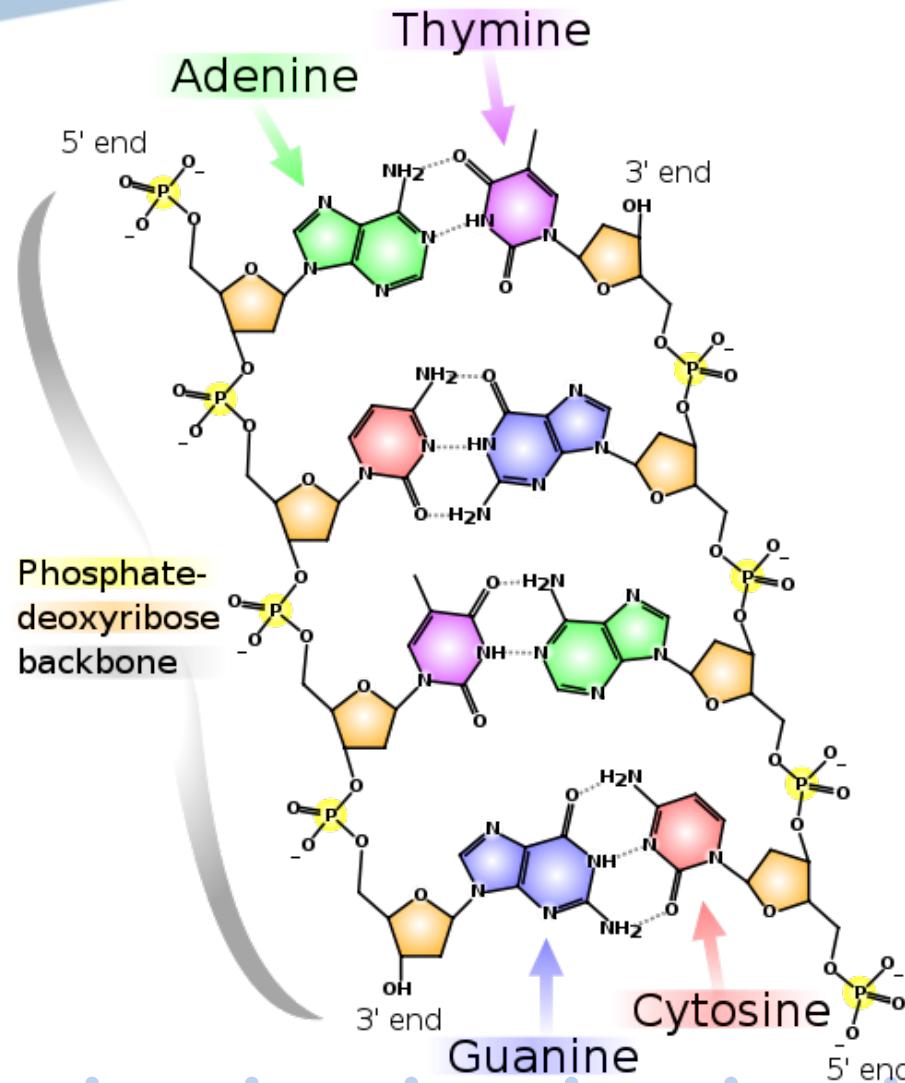


# Guión de la Unidad

- Introducción histórica
- **Definiciones Básicas**
- Modelos de Secuencias de DNA
  - Modelos multinomiales:
    - Estimación de modelos multinomiales
    - Composición de bases
    - Contenido en GC
  - Modelos markovianos:
    - Estimación de modelos markovianos
    - Frecuencia de k-meros
- Sesgos en el Uso de Codones



# Ácido Desoxiribonucleico (DNA)



# Genomas y su representación

- El **genoma** de un organismo es el conjunto de todo el DNA presente en cada una de sus células. Esto incluye los cromosomas nucleares y los cromosomas de las organelas tales como las **mitocondrias** y **cloroplastos**.

## Alfabetos y secuencias.

- Un **alfabeto** es un conjunto finito de símbolos:

$$S_{DNA} = \{ a, c, g, t \} \quad \bar{S}_{DNA} = \{ a, c, g, t, n, r, y \}$$

$$S_{AA} = \{ A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V \}$$

- Las **secuencias biológicas** se representan como sucesiones de los símbolos que representan sus bloques básicos según el alfabeto correspondiente.
- Para esta representación utilizaremos **vectores** en R.



# Guión de la Unidad

- Introducción histórica
- Definiciones Básicas
- **Modelos de Secuencias de DNA**
  - Modelos multinomiales:
    - Estimación de modelos multinomiales
    - Composición de bases
    - Contenido en GC
  - Modelos markovianos:
    - Estimación de modelos markovianos
    - Frecuencia de k-meros
- Sesgos en el Uso de Codones



# Modelos estadísticos básicos de DNA

- La cantidad masiva de datos genómicos disponibles hace necesario el desarrollo de modelos.
- Modelos como representación que no de lugar a dudas y que permitan compartir descubrimientos.
- Modelos que resumen toda la información de un **genoma** eliminando las características superfluas y resaltando las relevantes.
- Modelos básicos de secuencias de DNA:
  - Modelos multinomiales (independientes del contexto)
  - Modelos markovianos (dependientes del contexto)



# Guión de la Unidad

- Introducción histórica
- Definiciones Básicas
- Modelos de Secuencias de DNA
  - **Modelos multinomiales:**
    - Estimación de modelos multinomiales
    - Composición de bases
    - Contenido en GC
  - Modelos markovianos:
    - Estimación de modelos markovianos
    - Frecuencia de k-meros
- Sesgos en el Uso de Codones

# Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g

Los modelos multinomiales asumen que los nucleótidos aparecen en la secuencia de DNA de los genomas de forma independiente a su contexto y con igual probabilidad a lo largo de toda la secuencia.

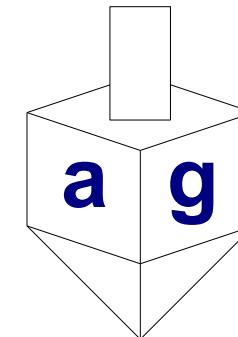


# Modelos Multinomiales

g



$$\begin{aligned}P[s_1 = a] &= p_a \\P[s_1 = c] &= p_c \\P[s_1 = g] &= p_g \\P[s_1 = t] &= p_t\end{aligned}$$



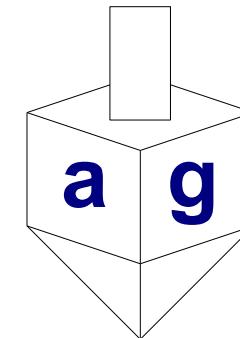
# Modelos Multinomiales

**g**

$$\begin{aligned}P[s_1 = a] &= p_a \\P[s_1 = c] &= p_c \\P[s_1 = g] &= p_g \\P[s_1 = t] &= p_t\end{aligned}$$



a g t t t a t c g c t t c c a t g a c g c a g a a g

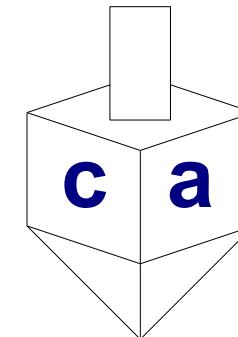


# Modelos Multinomiales

g **a** g t t t a t c g c t t c c a t g a c g c a g a a g



$$\begin{aligned} P[s_2 = a] &= p_a \\ P[s_2 = c] &= p_c \\ P[s_2 = g] &= p_g \\ P[s_2 = t] &= p_t \end{aligned}$$

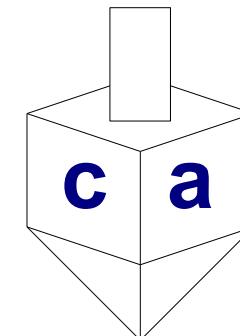


# Modelos Multinomiales

g **a** g t t t a t c g c t t c c a t g a c g c a g a a g



$$\begin{aligned} P[s_2 = a] &= p_a \\ P[s_2 = c] &= p_c \\ P[s_2 = g] &= p_g \\ P[s_2 = t] &= p_t \end{aligned}$$

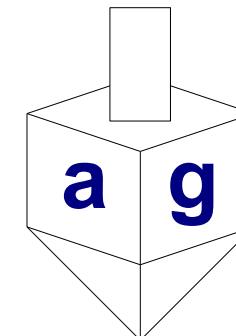


# Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$\begin{aligned} P[s_3 = a] &= p_a \\ P[s_3 = c] &= p_c \\ P[s_3 = g] &= p_g \\ P[s_3 = t] &= p_t \end{aligned}$$

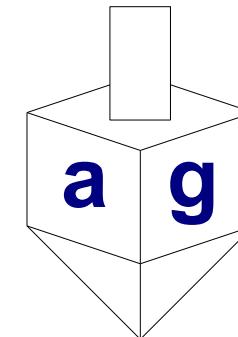


# Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$\begin{aligned} P[s_3 = a] &= p_a \\ P[s_3 = c] &= p_c \\ P[s_3 = g] &= p_g \\ P[s_3 = t] &= p_t \end{aligned}$$

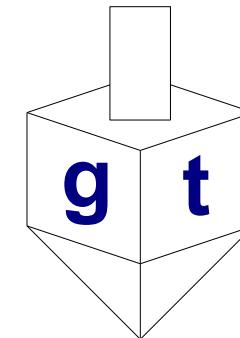


# Modelos Multinomiales

g a g **t** t t a t c g c t t c c a t g a c g c a g a a g



$$\begin{aligned} P[s_4 = a] &= p_a \\ P[s_4 = c] &= p_c \\ P[s_4 = g] &= p_g \\ P[s_4 = t] &= p_t \end{aligned}$$

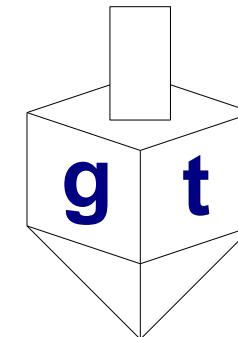


# Modelos Multinomiales

g a g **t** t t a t c g c t t c c a t g a c g c a g a a g



$$\begin{aligned} P[s_4 = a] &= p_a \\ P[s_4 = c] &= p_c \\ P[s_4 = g] &= p_g \\ P[s_4 = t] &= p_t \end{aligned}$$



# Modelos Multinomiales

g a g t t t t a t c g c t t c c a t g a c g c a g a a g



$$\begin{aligned} P[s_{17} = a] &= p_a \\ P[s_{17} = c] &= p_c \\ P[s_{17} = g] &= p_g \\ P[s_{17} = t] &= p_t \end{aligned}$$

- La probabilidad de que aparezca un nucleótido en una posición en particular es independiente de la posición y no depende del contexto.
- Los nucleótidos aparecen de forma idenpendiente e identicamente distribuidos (iid).
- El modelo multinomial de una secuencia de DNA queda unívocamente definido por cuatro valores, las probabilidades de obtener cada nucleótido:  
**multinomial.model(s) =  $(p_a, p_c, p_g, p_t)$      $p_a, p_c, p_g, p_t \geq 0$     y     $p_a + p_c + p_g + p_t = 1$**



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g

```
freq["a"] ← 7
```



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g

```
freq["a"] ← 7
```



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g

**freq[“a”] ← 7**

**freq[“c”] ← 6**



# Estimación de Modelos Multinomiales

g a g t t t t a t c g c t t c c a t g a c g c a g a a g

**freq[“a”] ← 7**

**freq[“c”] ← 6**

**freq[“g”] ← 7**

**freq[“t”] ← 8**



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g

**freq[“a”] ← 7**

**freq[“c”] ← 6**

**freq[“g”] ← 7**

**freq[“t”] ← 8**



**freq / 28**



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g

**freq[“a”] ← 7**

**freq[“c”] ← 6**

**freq[“g”] ← 7**

**freq[“t”] ← 8**



**freq / 28**

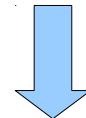


$(p_a, p_c, p_g, p_t)$



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



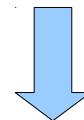
$$(p_a, p_c, p_g, p_t)$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$(p_a, p_c, p_g, p_t)$$

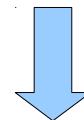
Entrada: Un vector de caracteres que representa una secuencia de DNA.

Paso 1: Calcular las frecuencias absolutas de a, c, g y t con table.



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$(p_a, p_c, p_g, p_t)$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

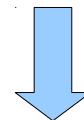
Paso 1: Calcular las frecuencias absolutas de a, c, g y t con table.

Paso 2: Dividir por la suma total de las frecuencias absolutas.



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$(p_a, p_c, p_g, p_t)$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

Paso 1: Calcular las frecuencias absolutas de a, c, g y t con table.

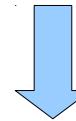
Paso 2: Dividir por la suma total de las frecuencias absolutas.

Paso 3: Nombrar cada elemento  $p_a$ ,  $p_c$ ,  $p_g$  y  $p_t$  respectivamente.



# Estimación de Modelos Multinomiales

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$(p_a, p_c, p_g, p_t)$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

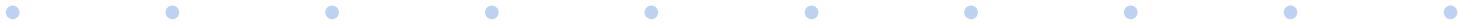
Paso 1: Calcular las frecuencias absolutas de a, c, g y t con table.

Paso 2: Dividir por la suma total de las frecuencias absolutas.

Paso 3: Nombrar cada elemento  $p_a$ ,  $p_c$ ,  $p_g$  y  $p_t$  respectivamente.

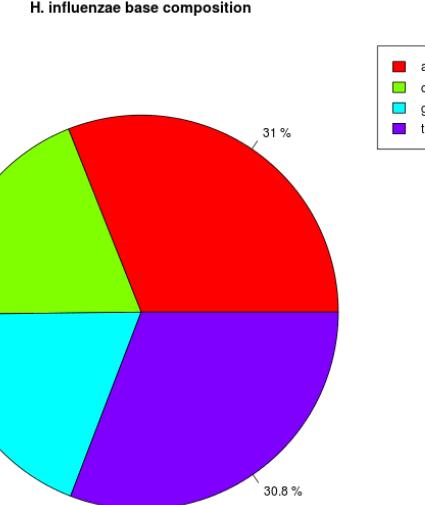
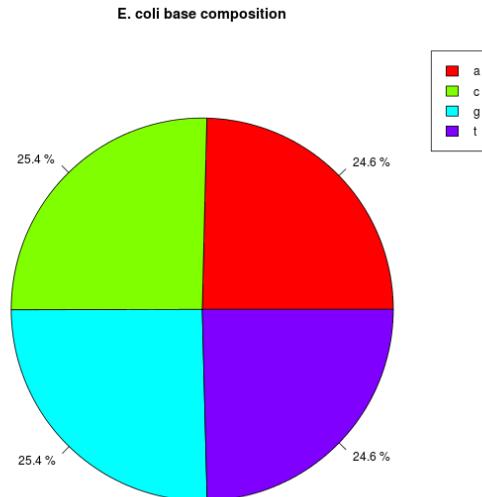
Paso 4: Devolver el modelo multinomial.

Salida: Un vector formado por cuatro elementos que representan  $p_a$ ,  $p_c$ ,  $p_g$  y  $p_t$



# Composición de Bases Global

- La **composición de bases** de un genoma se define como la frecuencia con la que aparece cada base en su secuencia completa.

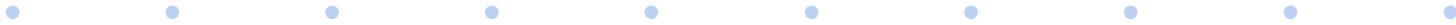


- La composición de bases constituye una **característica específica** de cada organismo.

# Composición de Bases Local

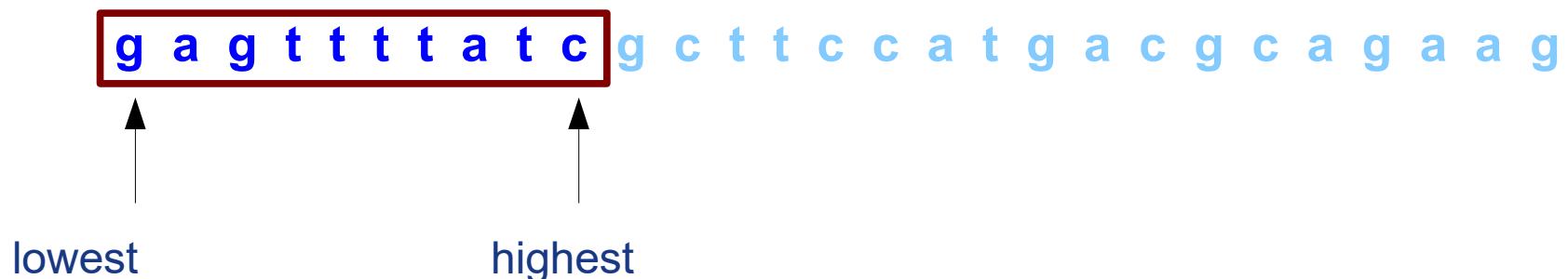
- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.

g a g t t t a t c g c t t c c a t g a c g c a g a a g



# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.



# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.

The diagram shows a sequence of DNA bases: g a g t t t a t c g c t t c c a t g a c g c a g a a g. A red box highlights a window of length `window.length` starting at index 1. Two arrows point upwards from the labels `lowest ← 1` and `highest ← window.length` to the start and end of the highlighted window respectively.

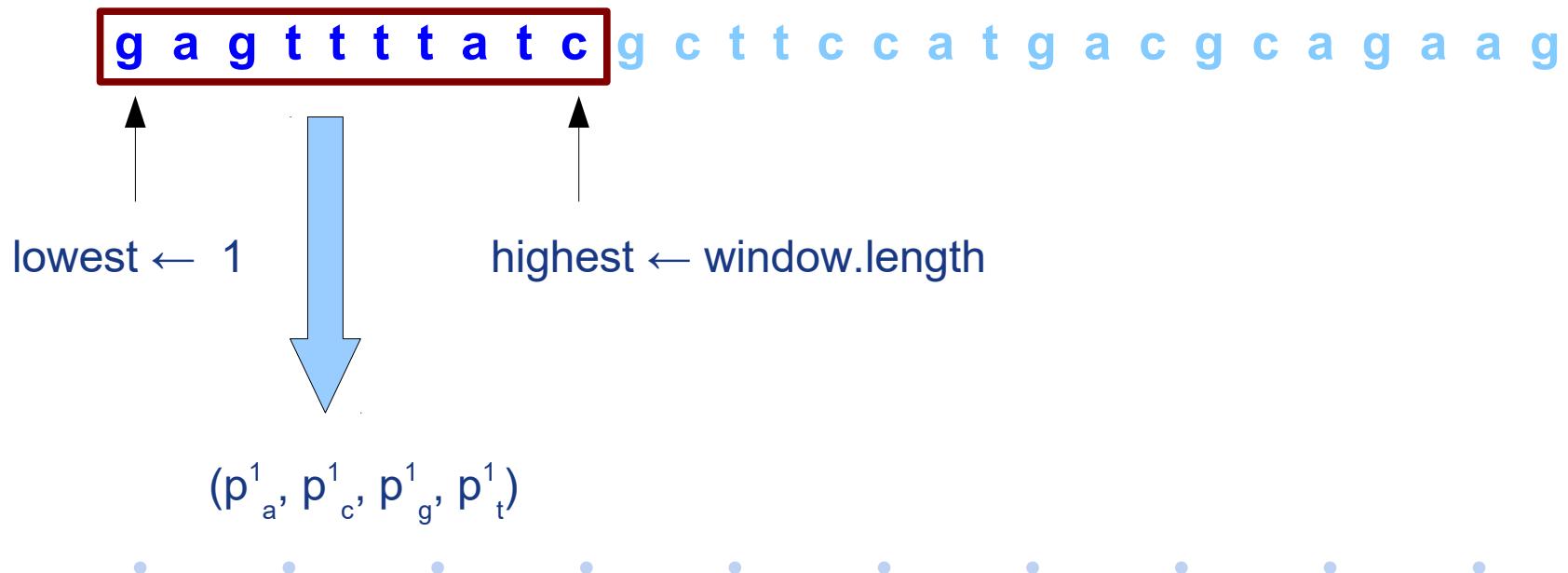
g a g t t t a t c g c t t c c a t g a c g c a g a a g

↑                              ↑

lowest ← 1                      highest ← window.length

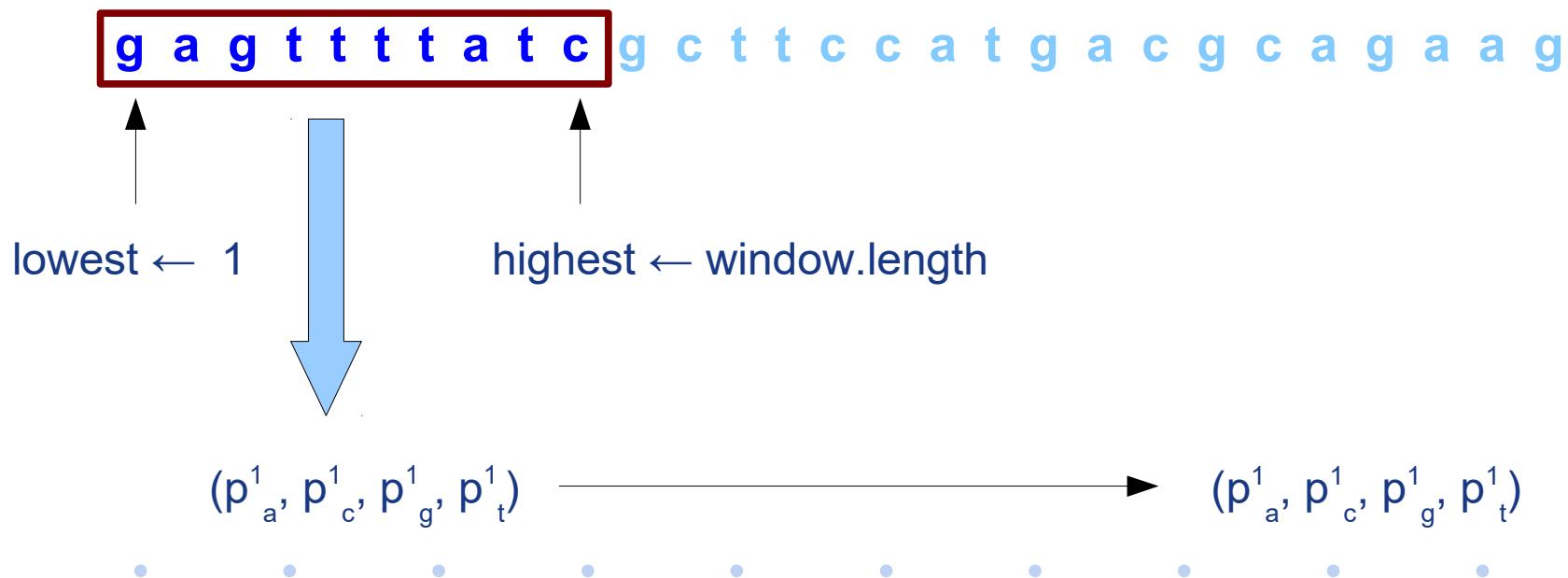
# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.



# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.



# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.

g a g t t t t a t c g c t t c c a t g a c g c a g a a g



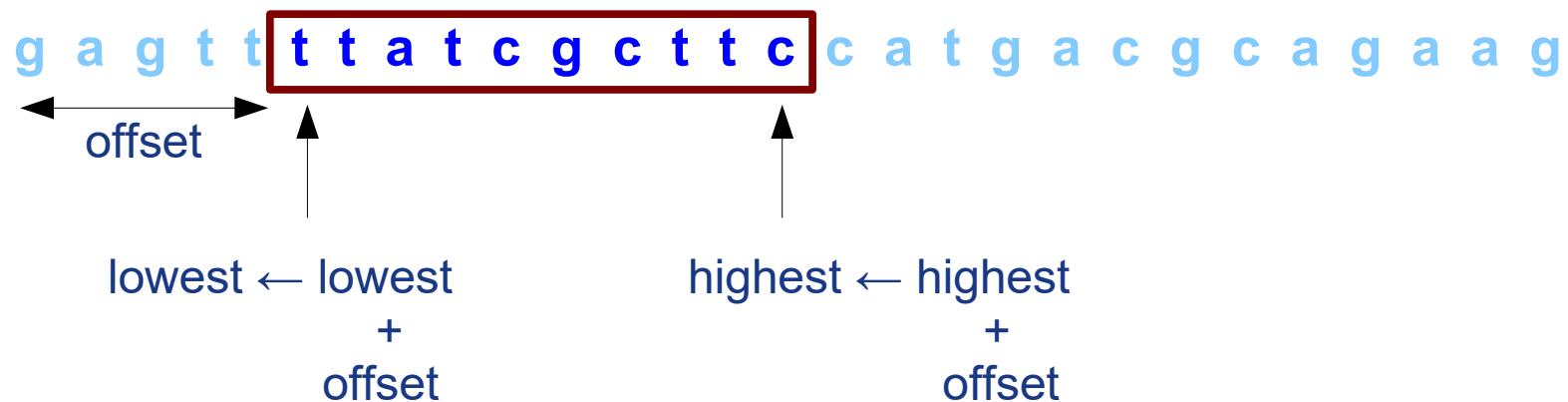
# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.

The diagram shows a sequence of DNA bases: g a g t t t t a t c g c t t c c a t g a c g c a g a a g. A red rectangular box highlights a window of length 11 bases: t t a t c g c t t c. Below the sequence, a double-headed arrow labeled "offset" indicates the distance between the start of the current window and the start of the previous window. Ellipses below the sequence indicate that this pattern repeats along the genome.

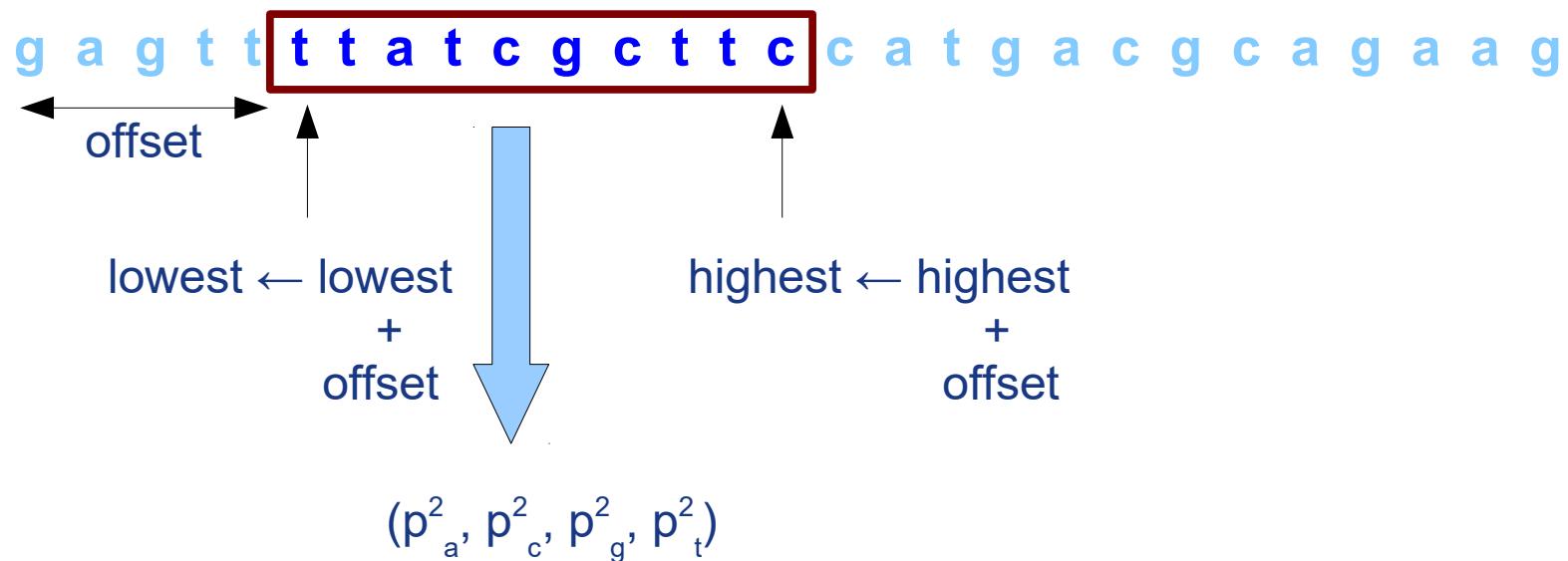
# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.



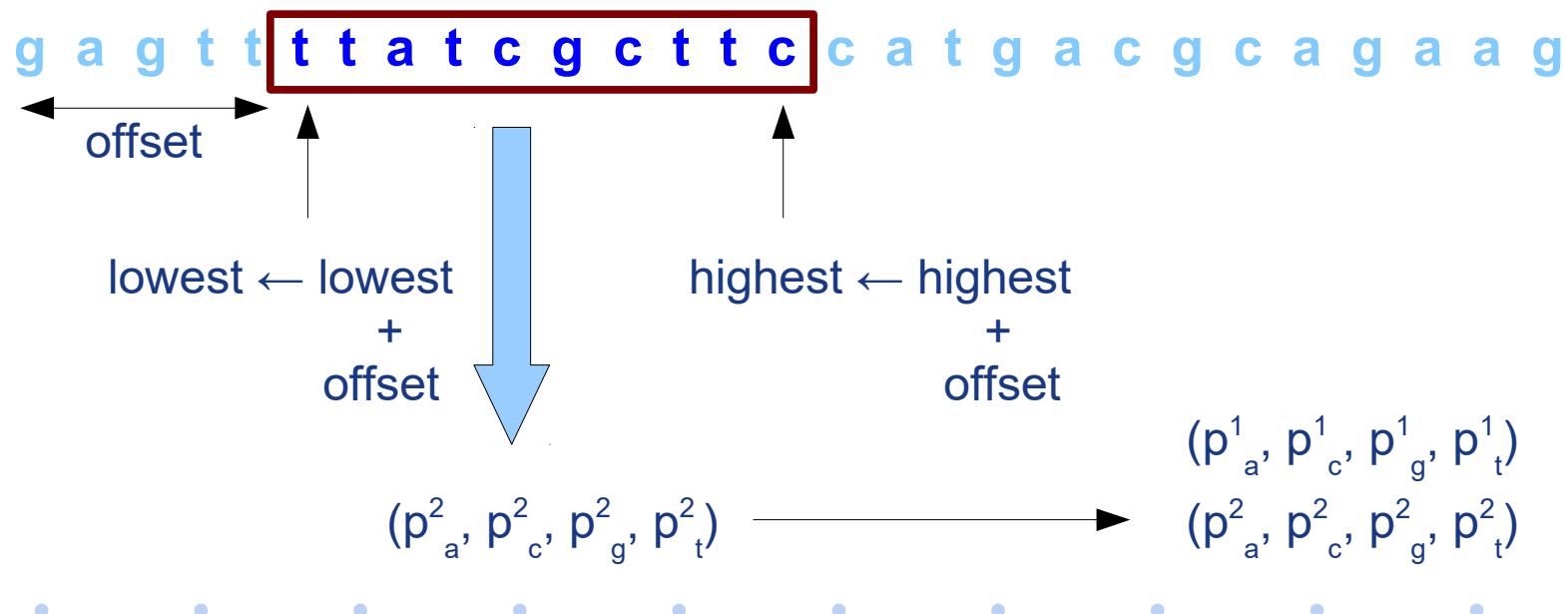
# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.



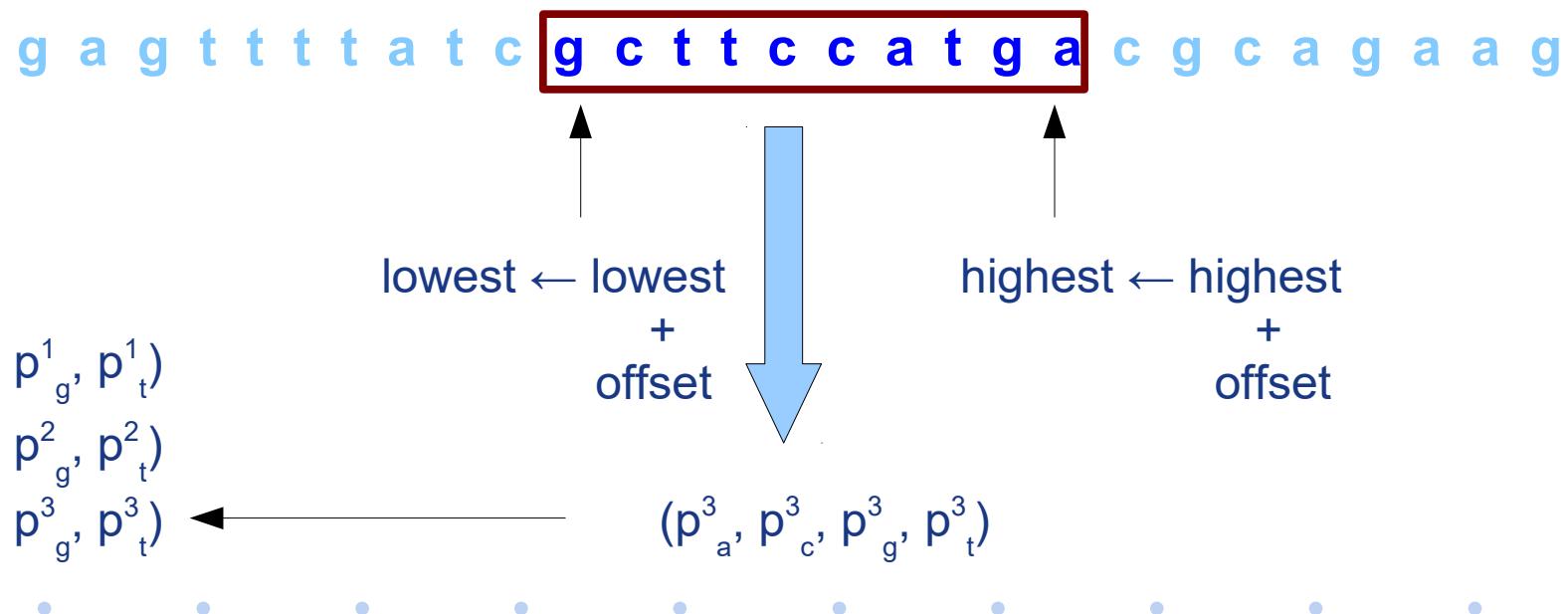
# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.



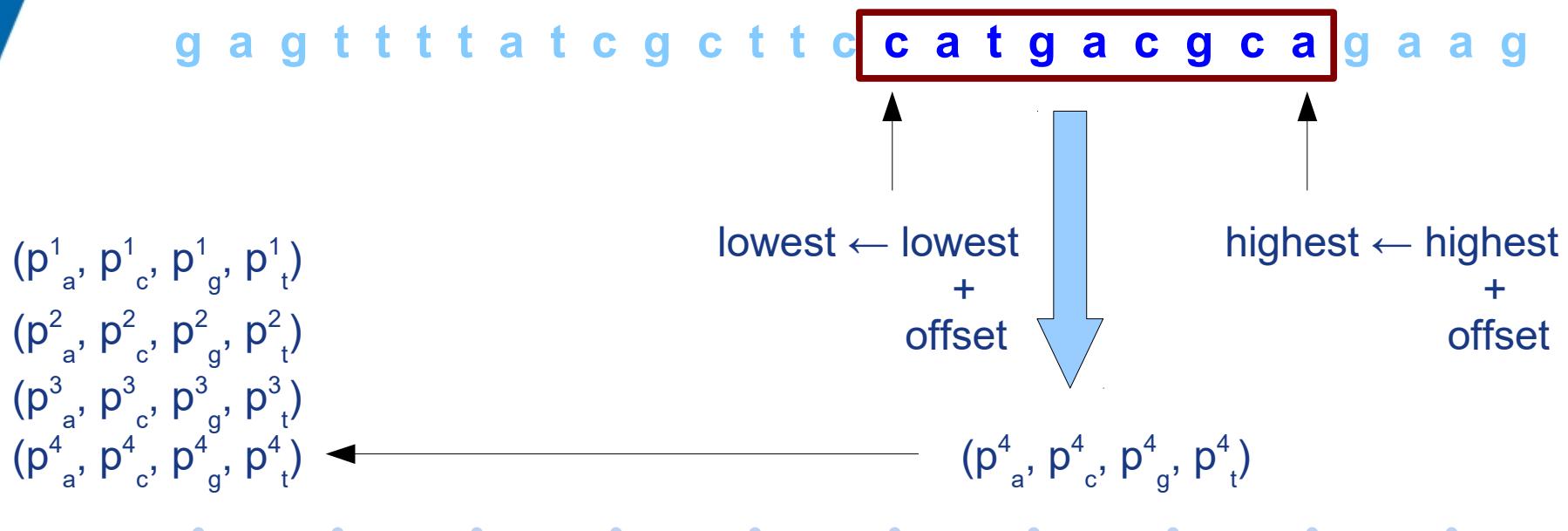
# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.



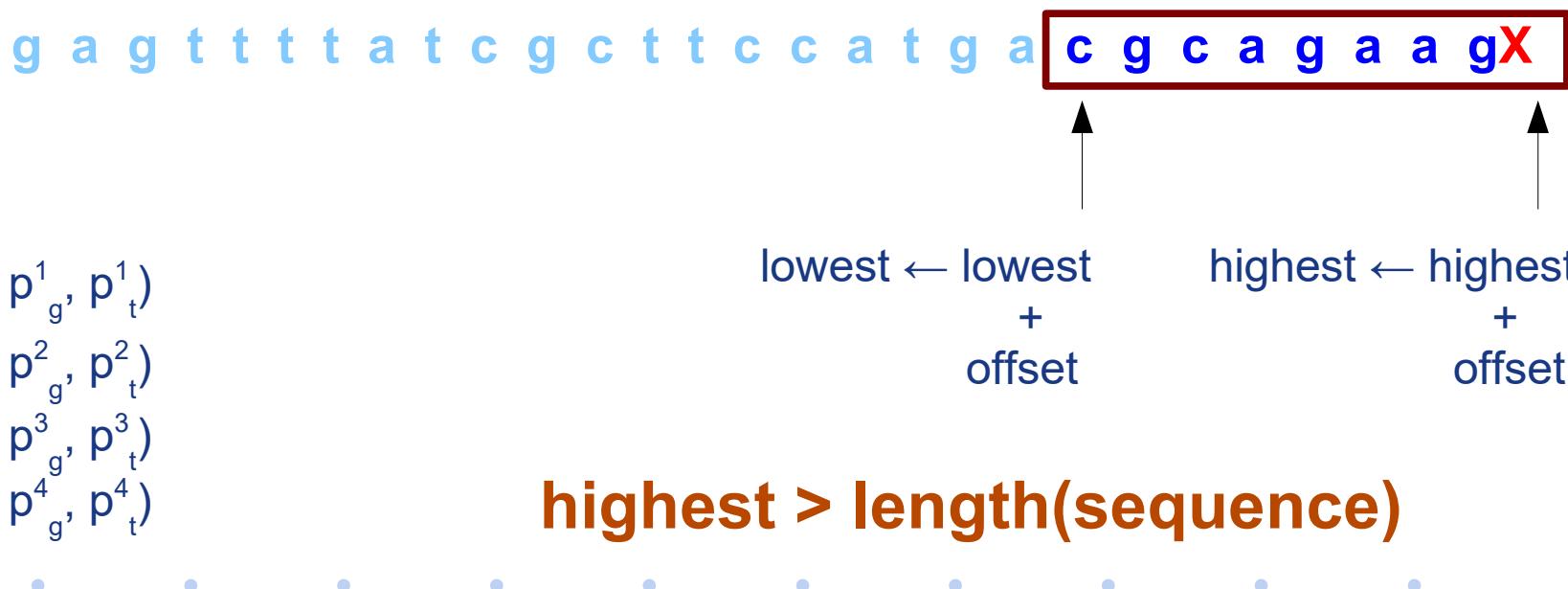
# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.

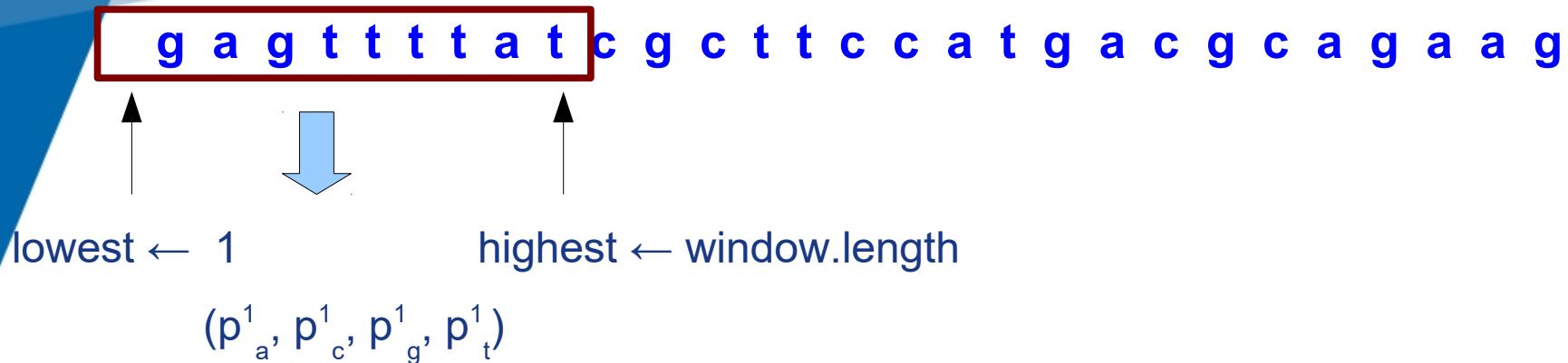


# Composición de Bases Local

- La búsqueda de regiones cuya composición de bases difiere de la composición de bases global puede identificar estructuras relevantes en el genoma.
- Para esto se calcula la composición de bases a través de una ventana de longitud `window.length` que se desliza a lo largo del genoma con un desplazamiento de longitud `offset`.



# Algoritmo: Composición de Bases Local



Entrada: Un vector que representa una secuencia de DNA (dna.sequence), la longitud de la ventana(window.length) y el desplazamiento (offset)

Paso 1: Inicialización de variables antes del bucle:

```
lowest <- 1  
highest <- window.length  
result <- numeric()
```

Paso 2: Mientras  $\text{highest} \leq \text{length}(\text{dna.sequence})$

Modelo multinomial local:  $\text{l.multi.model} \leftarrow \text{multinomial.model}(\text{dna.sequence}[lowest:highest])$

Actualizar resultado:  $\text{result} \leftarrow \text{rbind}(\text{result}, \text{l.multi.model})$

Actualizar ventana:  $\text{lowest} \leftarrow \text{lowest} + \text{offset}$

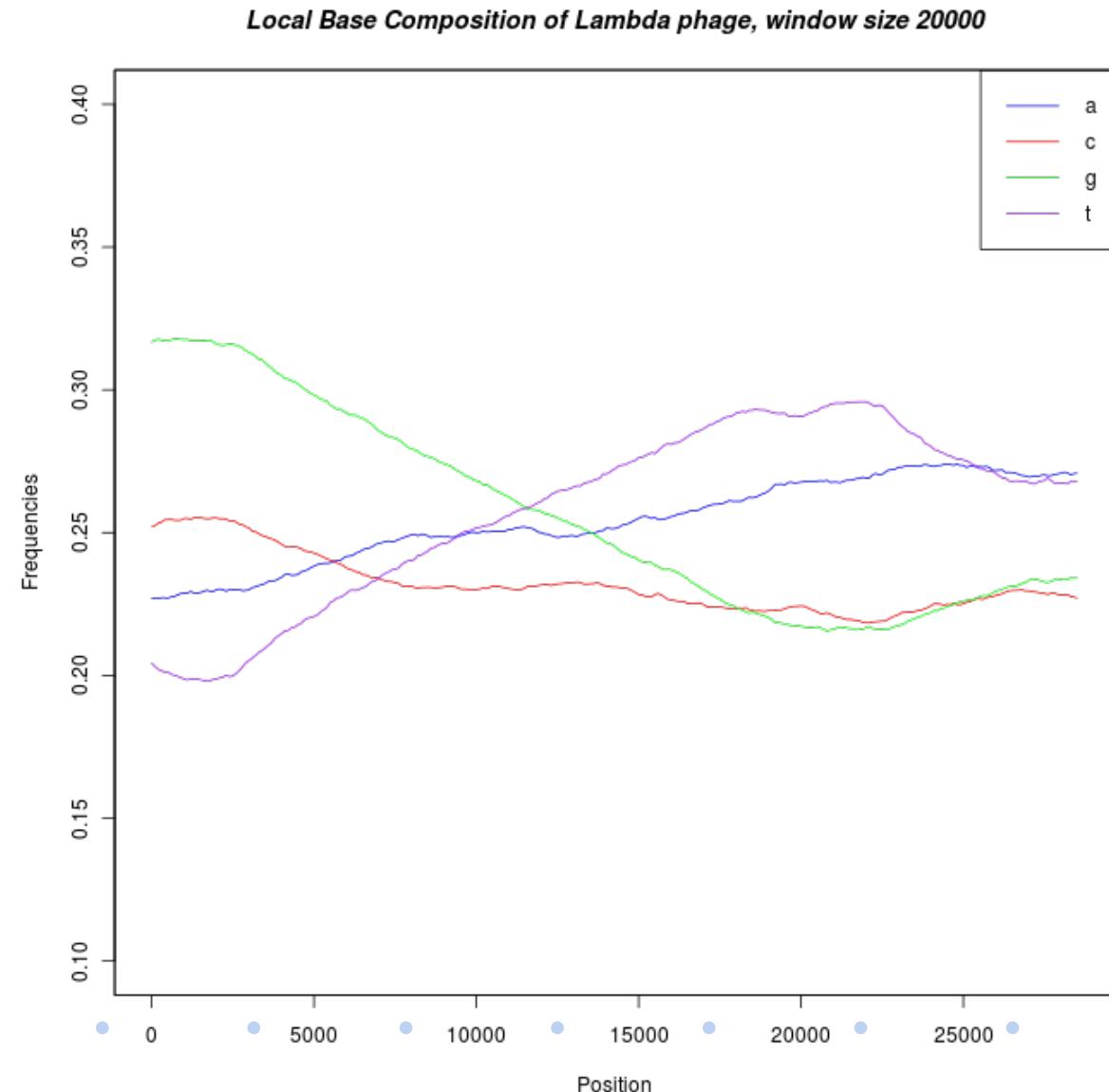
$\text{highest} \leftarrow \text{highest} + \text{offset}$

Paso 3: Devolver result.

# Algoritmo: Composición de Bases Local

El tamaño de la ventana se determina usualmente de forma empírica.

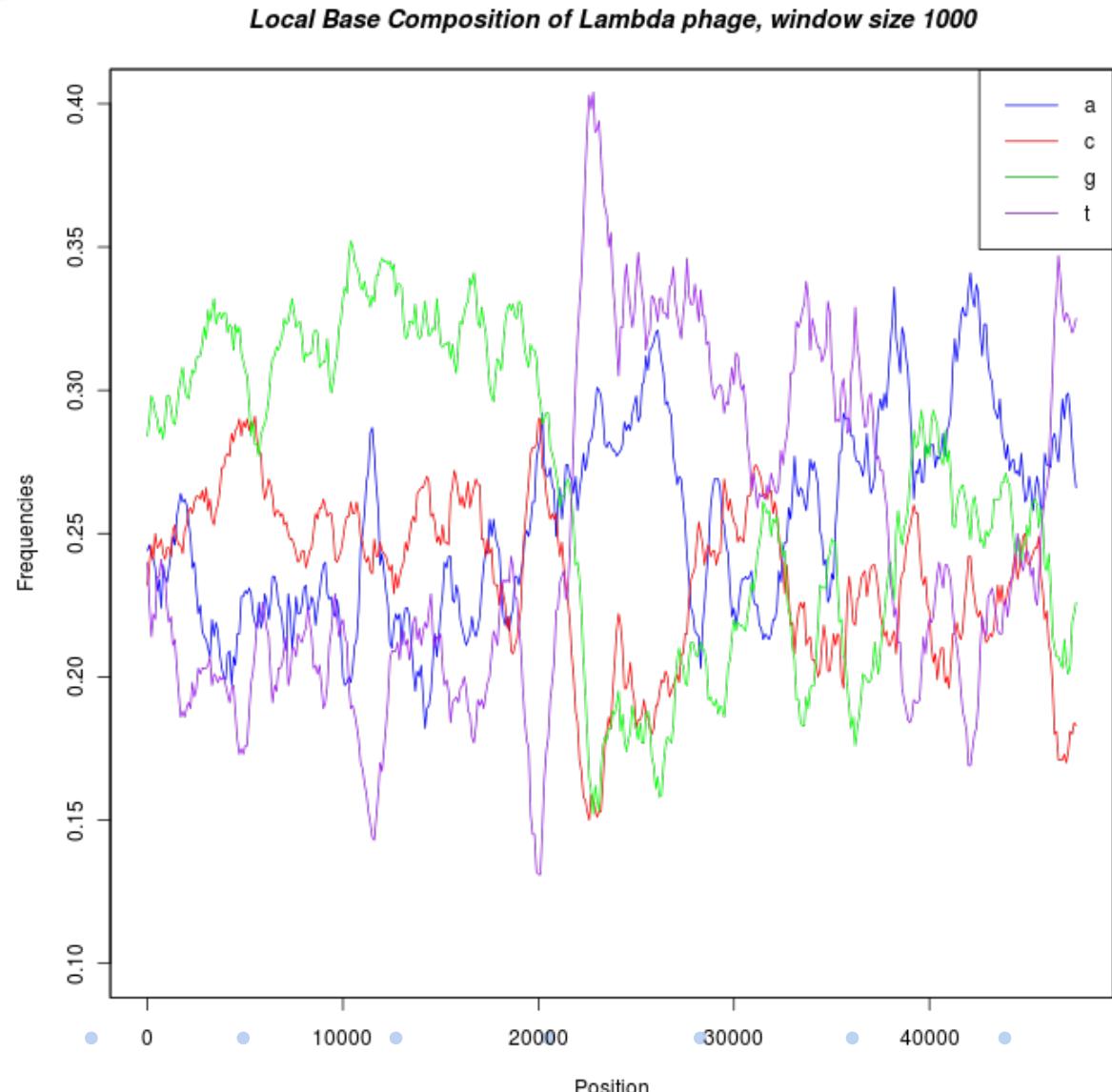
Ventanas de un tamaño excesivamente grande pierden información local y muestran una gráfica con poca variabilidad.



# Algoritmo: Composición de Bases Local

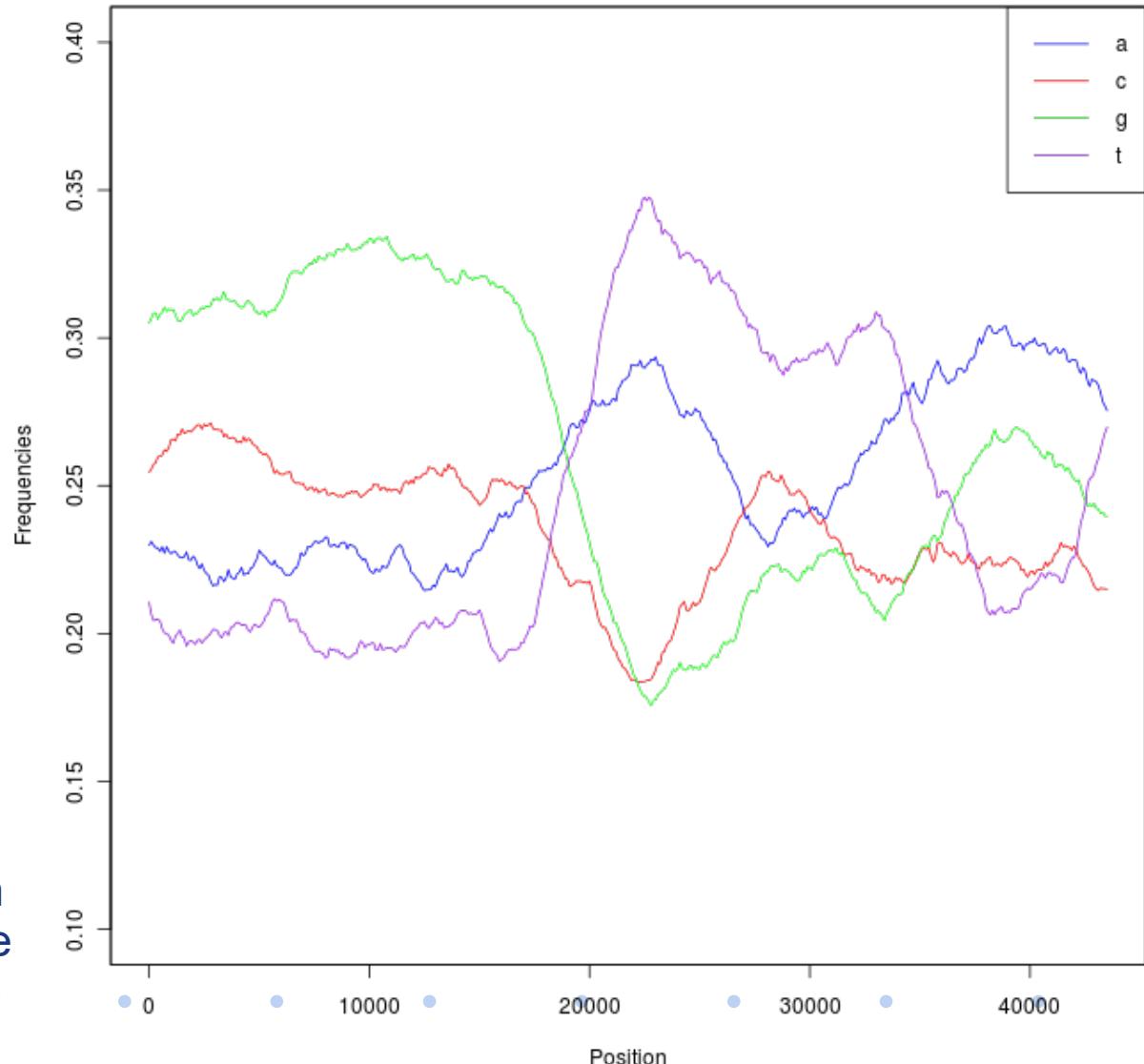
El tamaño de la ventana se determina usualmente de forma empírica.

Ventanas de un tamaño excesivamente pequeño recogen mucha variabilidad local y dificultan la comparación con la composición global de bases.



# Algoritmo: Composición de Bases Local

*Local Base Composition of Lambda phage, window size 5000*

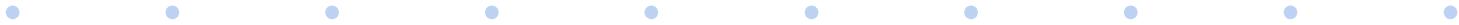


El tamaño de la ventana se determina usualmente de forma empírica.

Ventanas de un tamaño óptimo permiten identificar regiones del genoma con un comportamiento local que se diferencia significativamente del global.

# Contenido en GC Global

- No todos los nucleótidos tienen las mismas propiedades bioquímicas.
  - Entre *g* y *c* se establecen tres puentes de hidrógeno mientras que entre *a* y *t* dos.
  - El **contenido en GC** de un genoma, la proporción de los nucleótidos *g* y *c*, es una **característica muy específica de cada organismo en concreto**.
- 
- La **función GC** de la librería **seqinr** calcula el contenido en GC de una secuencia de DNA.



# Algoritmo: Contenido en GC Local



Entrada: Un vector que representa una secuencia de DNA (dna.sequence), la longitud de la ventana(window.length) y el desplazamiento (offset)

Paso 1: Inicialización de variables antes del bucle:

```
lowest ← 1  
highest ← window.length  
local.GC ← numeric(0)  
positions ← numeric(0)  
i ← 1
```

Paso 2: Mientras highest <= length(dna.sequence)

    Guardar valor local: local.GC[i] ← GC(dna.sequence[lowest:highest])

    Guardar posición actual: positions[i] ← lowest

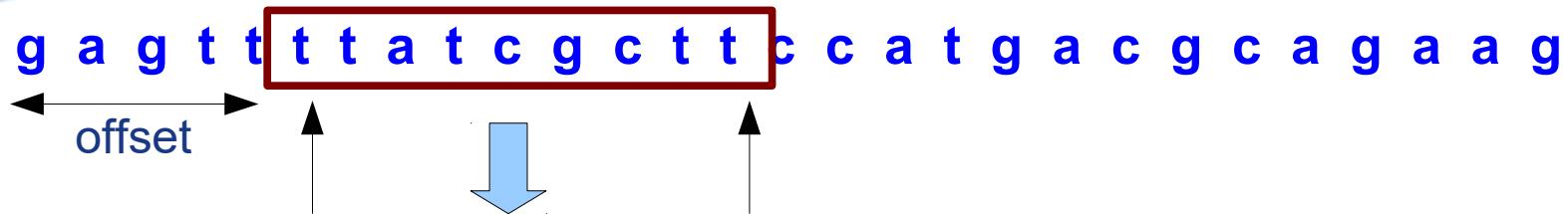
    Actualizar ventana: lowest ← lowest + offset

                        highest ← highest + offset

                        i ← i + 1

Paso 3: Devolver result una lista con local.GC y positions.

# Algoritmo: Contenido en GC Local


$$\text{lowest} \leftarrow \text{lowest} + \text{offset} \quad (\text{GC}_1, \text{GC}_2) \quad \text{highest} \leftarrow \text{highest} + \text{offset}$$

Entrada: Un vector que representa una secuencia de DNA (dna.sequence), la longitud de la ventana(window.length) y el desplazamiento (offset)

Paso 1: Inicialización de variables antes del bucle:

```
lowest ← 1  
highest ← window.length  
local.GC ← numeric(0)  
positions ← numeric(0)  
i ← 1
```

Paso 2: Mientras highest <= length(dna.sequence)

    Guardar valor local: local.GC[i] ← GC(dna.sequence[lowest:highest])

    Guardar posición actual: positions[i] ← lowest

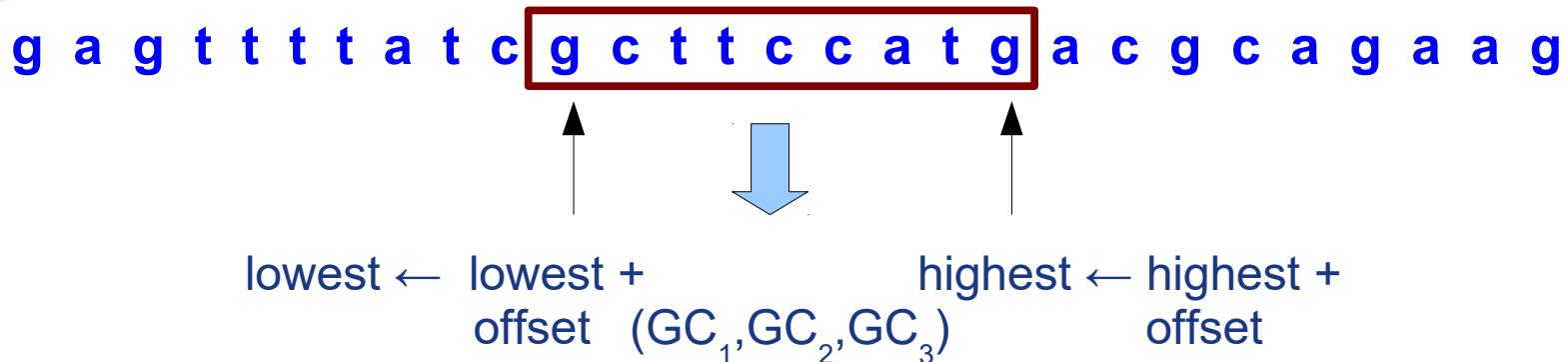
    Actualizar ventana: lowest ← lowest + offset

                        highest ← highest + offset

                        i ← i + 1

Paso 3: Devolver result una lista con local.GC y positions.

# Algoritmo: Contenido en GC Local



Entrada: Un vector que representa una secuencia de DNA (dna.sequence), la longitud de la ventana(window.length) y el desplazamiento (offset)

Paso 1: Inicialización de variables antes del bucle:

```
lowest ← 1  
highest ← window.length  
local.GC ← numeric(0)  
positions ← numeric(0)  
i ← 1
```

Paso 2: Mientras  $\text{highest} \leq \text{length}(\text{dna.sequence})$

    Guardar valor local:  $\text{local.GC}[i] \leftarrow \text{GC}(\text{dna.sequence}[\text{lowest}:\text{highest}])$

    Guardar posición actual:  $\text{positions}[i] \leftarrow \text{lowest}$

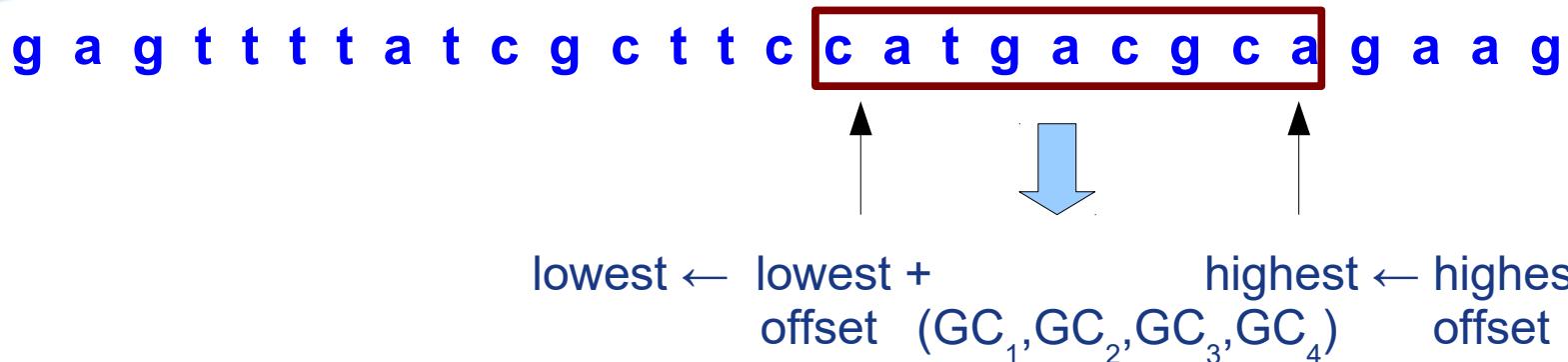
    Actualizar ventana:  $\text{lowest} \leftarrow \text{lowest} + \text{offset}$

$\text{highest} \leftarrow \text{highest} + \text{offset}$

$i \leftarrow i + 1$

Paso 3: Devolver result una lista con local.GC y positions.

# Algoritmo: Contenido en GC Local



Entrada: Un vector que representa una secuencia de DNA (dna.sequence), la longitud de la ventana(window.length) y el desplazamiento (offset)

Paso 1: Inicialización de variables antes del bucle:

```
lowest ← 1  
highest ← window.length  
local.GC ← numeric(0)  
positions ← numeric(0)  
i ← 1
```

Paso 2: Mientras highest <= length(dna.sequence)

    Guardar valor local: local.GC[i] ← GC(dna.sequence[lowest:highest])

    Guardar posición actual: positions[i] ← lowest

    Actualizar ventana: lowest ← lowest + offset

                        highest ← highest + offset

                        i ← i + 1

Paso 3: Devolver result una lista con local.GC y positions.

# Algoritmo: Contenido en GC Local



Entrada: Un vector que representa una secuencia de DNA (dna.sequence), la longitud de la ventana(window.length) y el desplazamiento (offset)

Paso 1: Inicialización de variables antes del bucle:

```
lowest ← 1  
highest ← window.length  
local.GC ← numeric(0)  
positions ← numeric(0)  
i ← 1
```

Paso 2: Mientras highest <= length(dna.sequence)

    Guardar valor local: local.GC[i] ← GC(dna.sequence[lowest:highest])

    Guardar posición actual: positions[i] ← lowest

    Actualizar ventana: lowest ← lowest + offset

                highest ← highest + offset

                i ← i + 1

Paso 3: Devolver result una lista con local.GC y positions.

# Algoritmo: Contenido en GC Local



Entrada: Un vector que representa una secuencia de DNA (dna.sequence), la longitud de la ventana(window.length) y el desplazamiento (offset)

Paso 1: Inicialización de variables antes del bucle:

```
lowest ← 1  
highest ← window.length  
local.GC ← numeric(0)  
positions ← numeric(0)  
i ← 1
```

Paso 2: Mientras highest <= length(dna.sequence)

    Guardar valor local: local.GC[i] ← GC(dna.sequence[lowest:highest])

    Guardar posición actual: positions[i] ← lowest

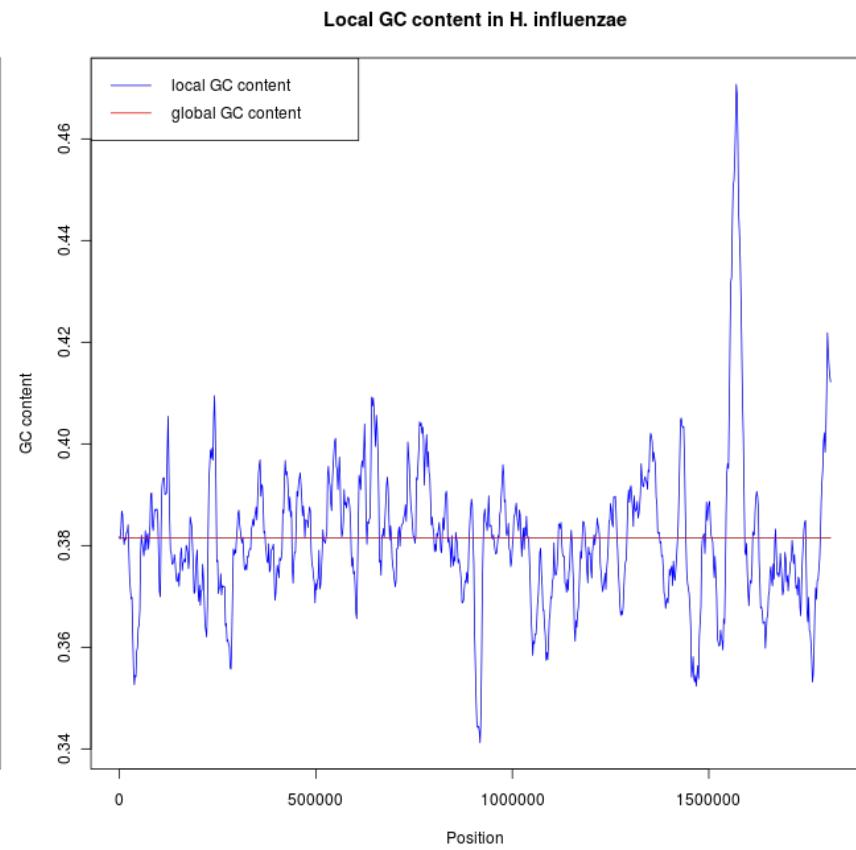
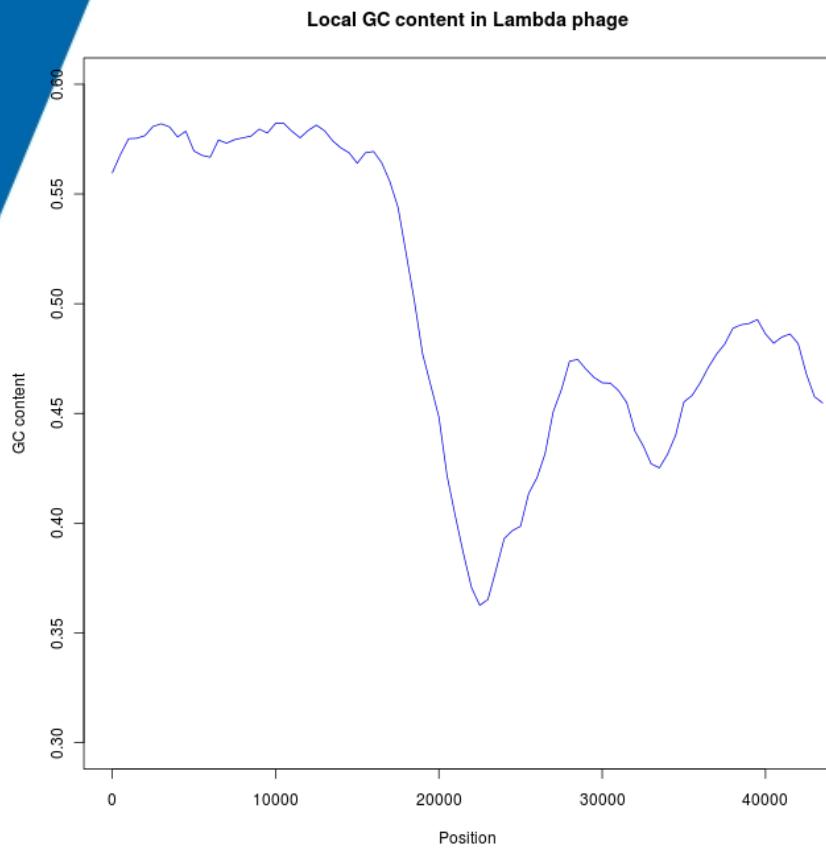
    Actualizar ventana: lowest ← lowest + offset

                highest ← highest + offset

                i ← i + 1

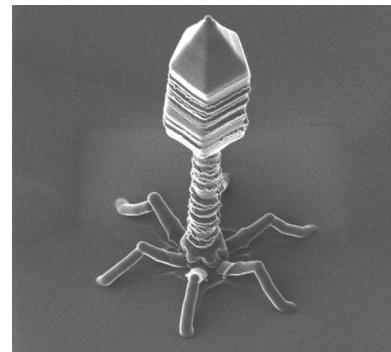
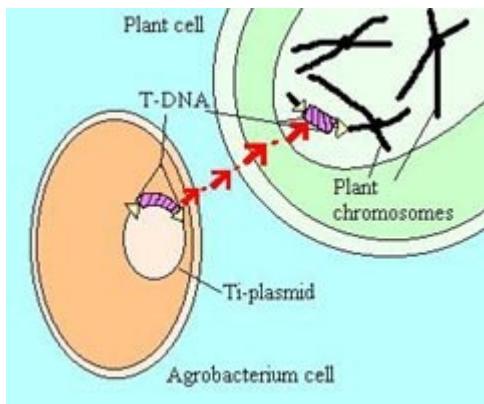
Paso 3: Devolver result una lista con local.GC y positions.

# Contenido en GC Local



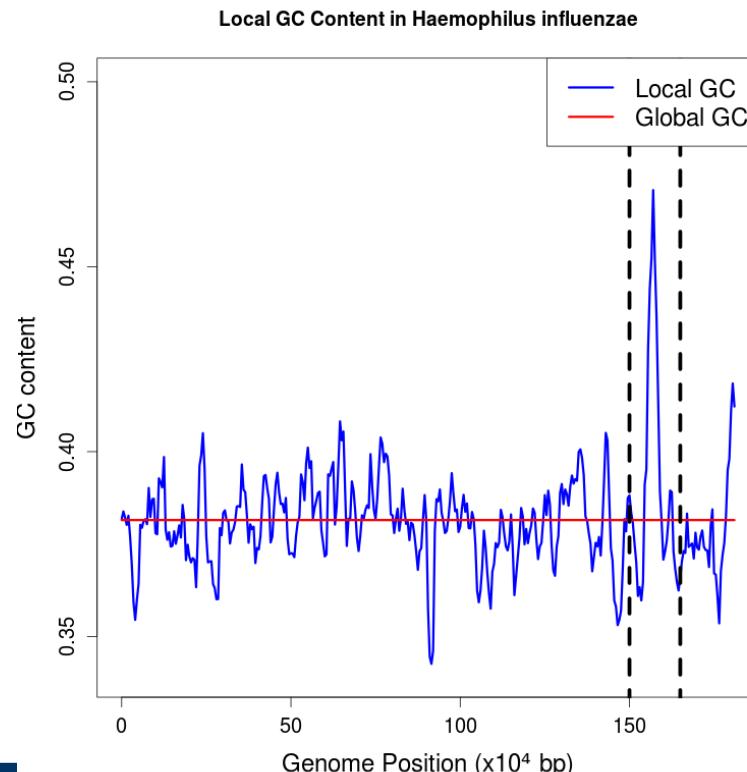
# Transferencia Horizontal de Genes

- La **transferencia horizontal de genes** hace referencia a la transmisión de genes entre organismos vivos (comúnmente de diferentes especies) sin que uno sea el progenitor del otro.



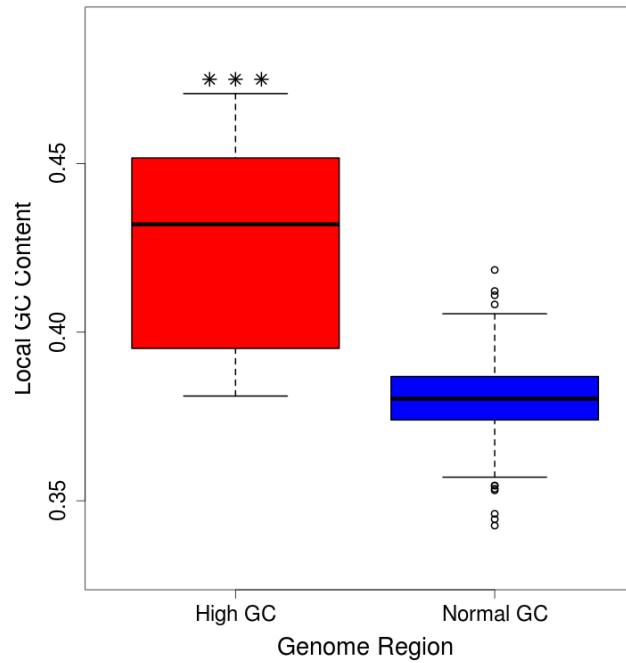
# Transferencia Horizontal de Genes

- Debido a que el contenido en GC es una de las características más específicas de cada especie la identificación de zonas en el genoma cuyo contenido en GC diverge significativamente del contenido GC global puede indicar la existencia de un evento de transferencia horizontal de genes.



# Transferencia Horizontal de Genes

- Debido a que el contenido en GC es una de las características más específicas de cada especie la identificación de zonas en el genoma cuyo contenido en GC diverge significativamente del contenido GC global puede indicar la existencia de un evento de transferencia horizontal de genes.



# Guión de la Unidad

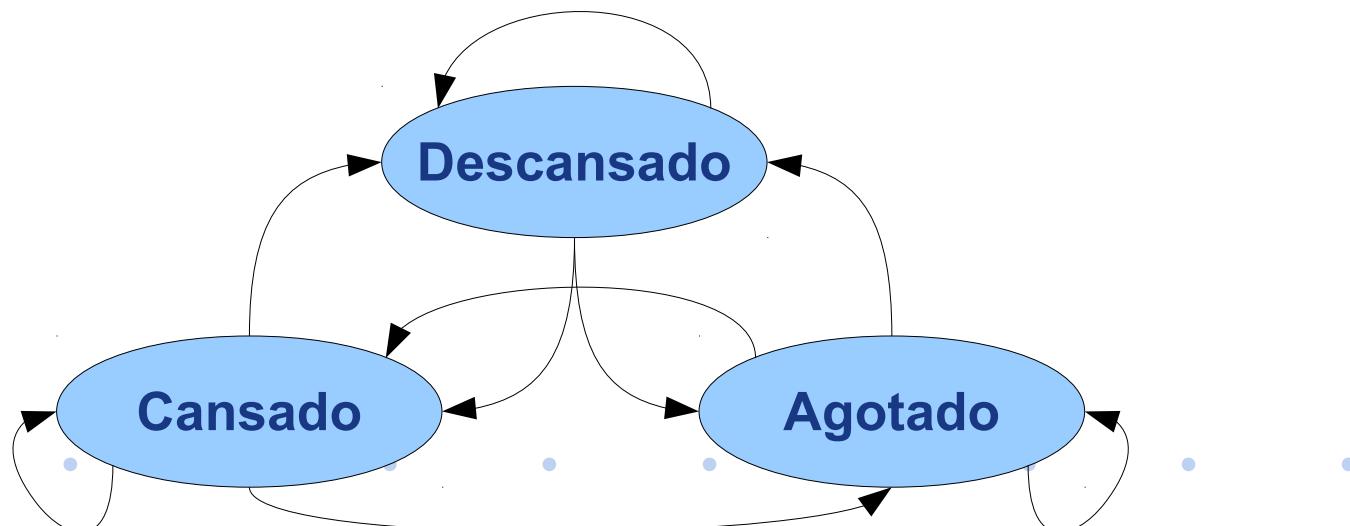
- Introducción histórica
- Definiciones Básicas
- Modelos de Secuencias de DNA
  - Modelos multinomiales:
    - Estimación de modelos multinomiales
    - Composición de bases
    - Contenido en GC
  - Modelos markovianos:
    - **Estimación de modelos markovianos**
    - **Frecuencia de k-meros**
- Sesgos en el Uso de Codones



# Modelos Markovianos

- Una **cadena de Markov** es un modelo de una sucesión o secuencia de eventos tal que la probabilidad de obtener un evento en una posición en particular sólo depende del evento observado en la posición anterior.

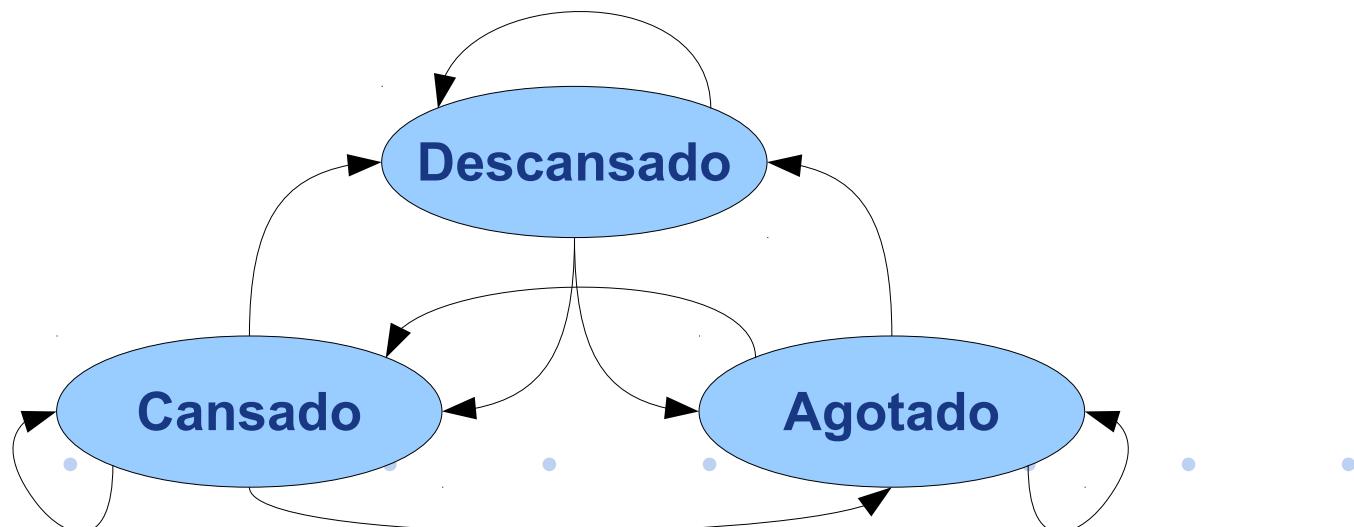
D → C → C → D → C → A → D → D → C → A → D



# Modelos Markovianos

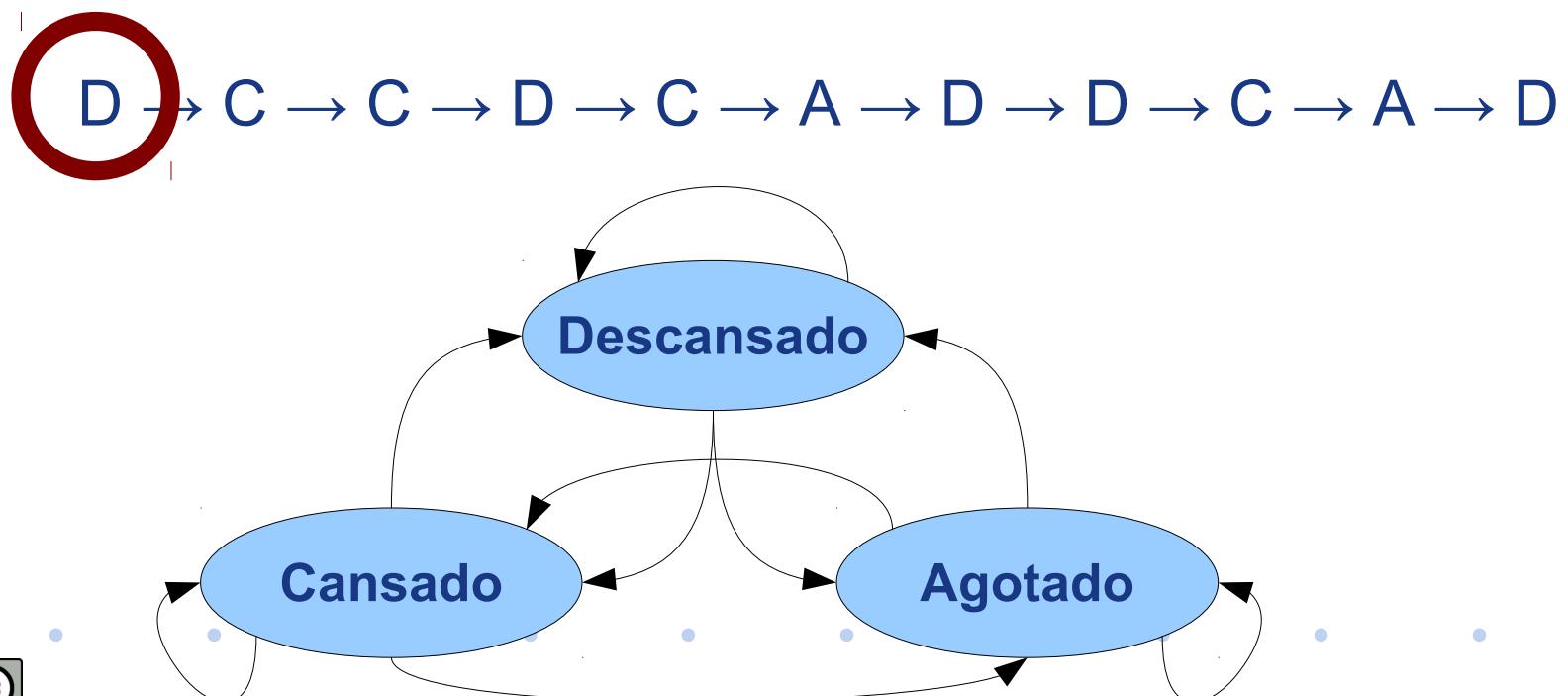
- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$

$D \rightarrow C \rightarrow C \rightarrow D \rightarrow C \rightarrow A \rightarrow D \rightarrow D \rightarrow C \rightarrow A \rightarrow D$



# Modelos Markovianos

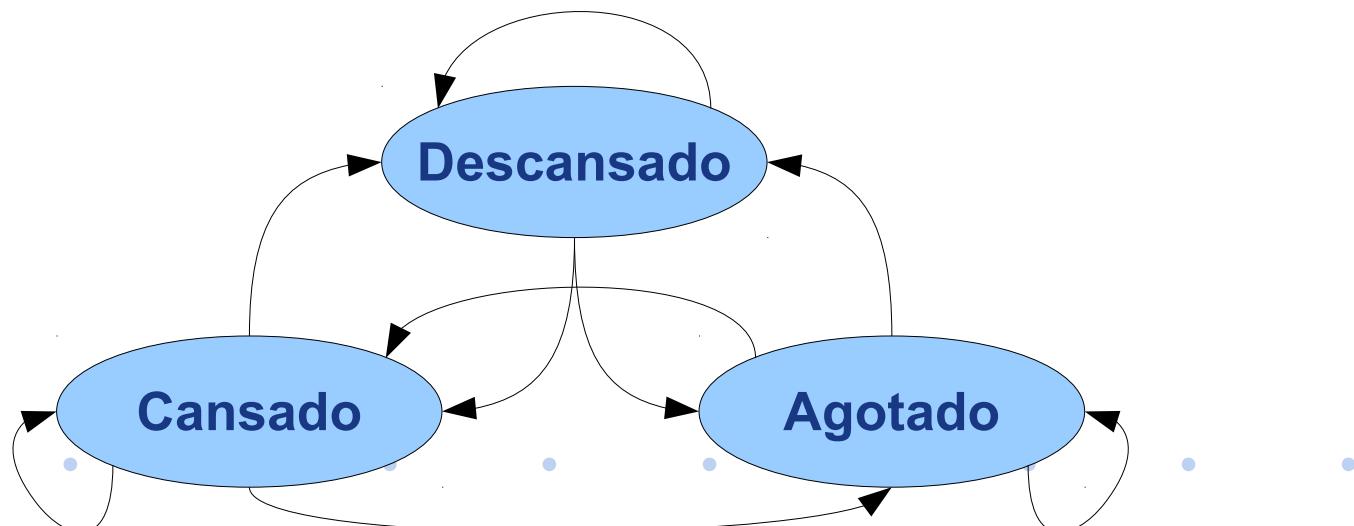
- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.



# Modelos Markovianos

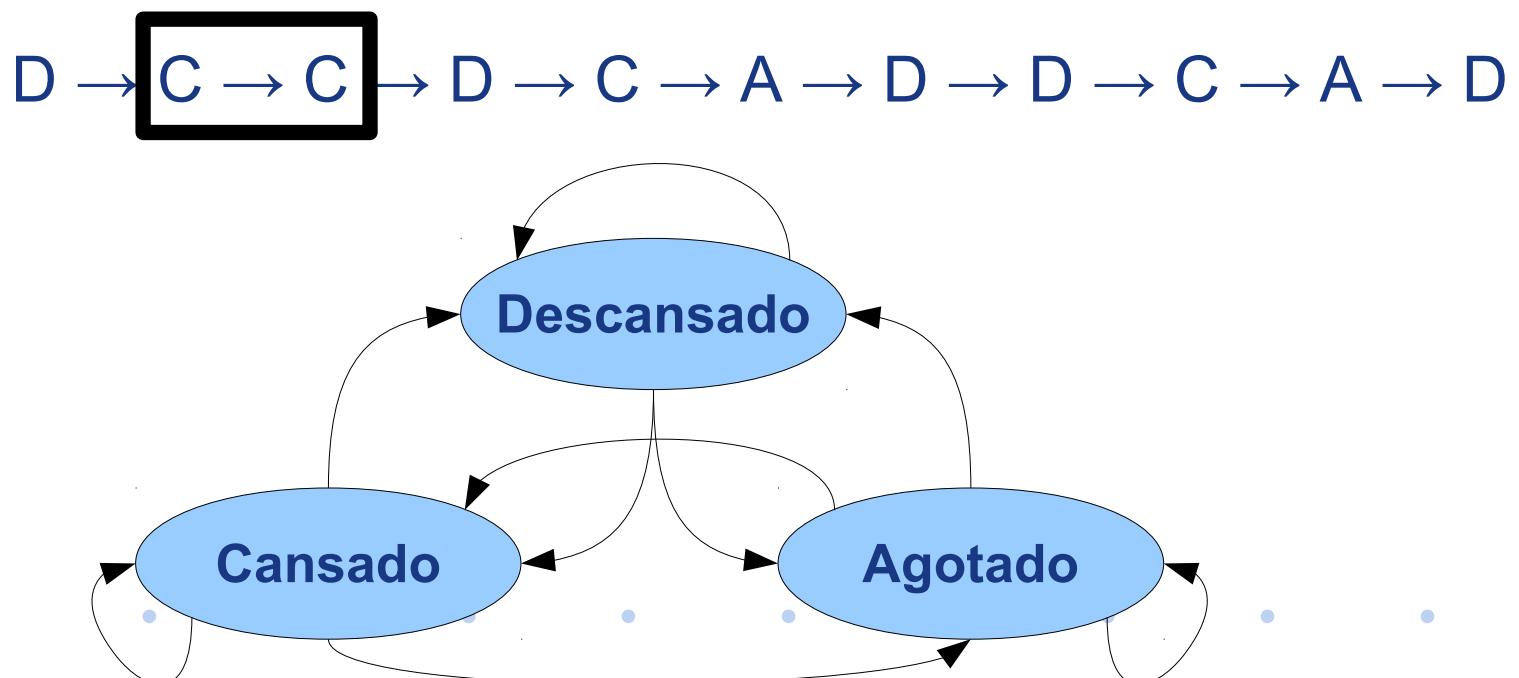
- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .

**D → C → C → D → C → A → D → D → C → A → D**



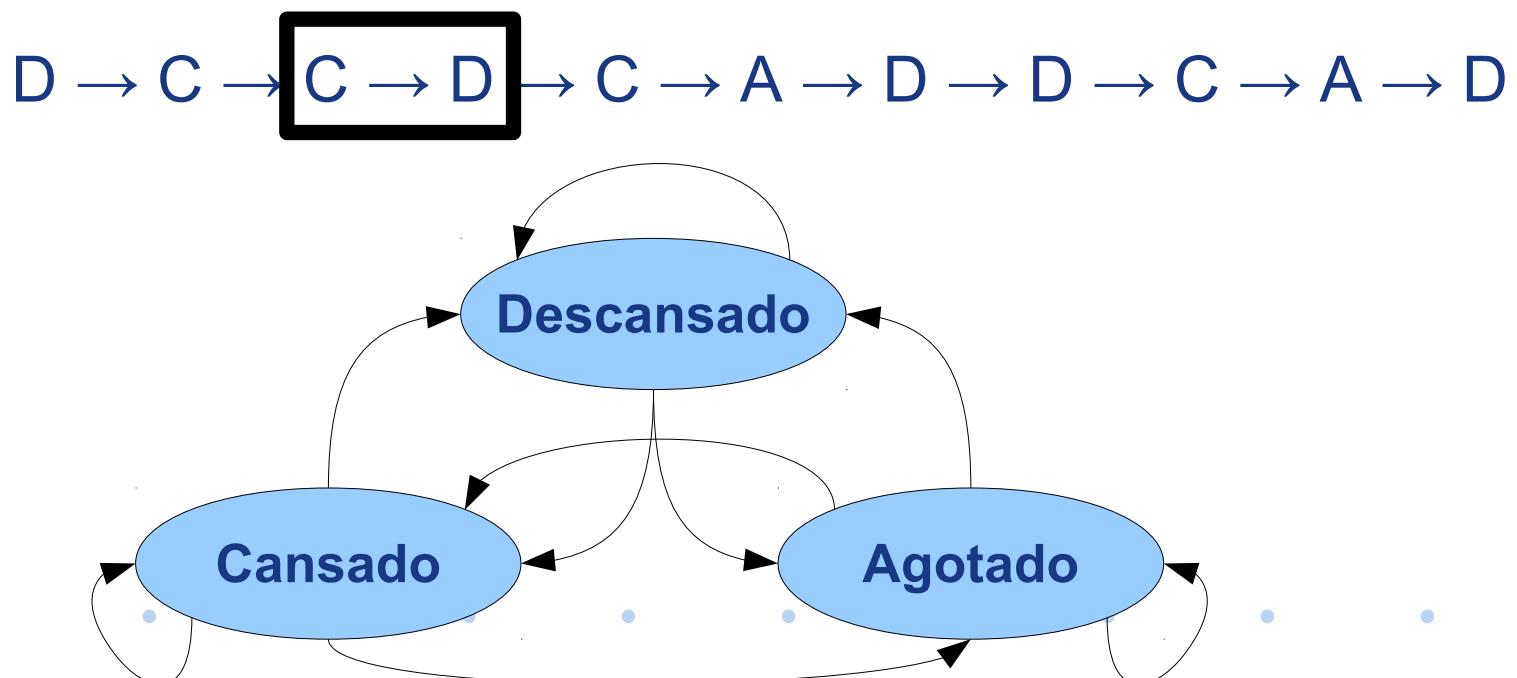
# Modelos Markovianos

- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .



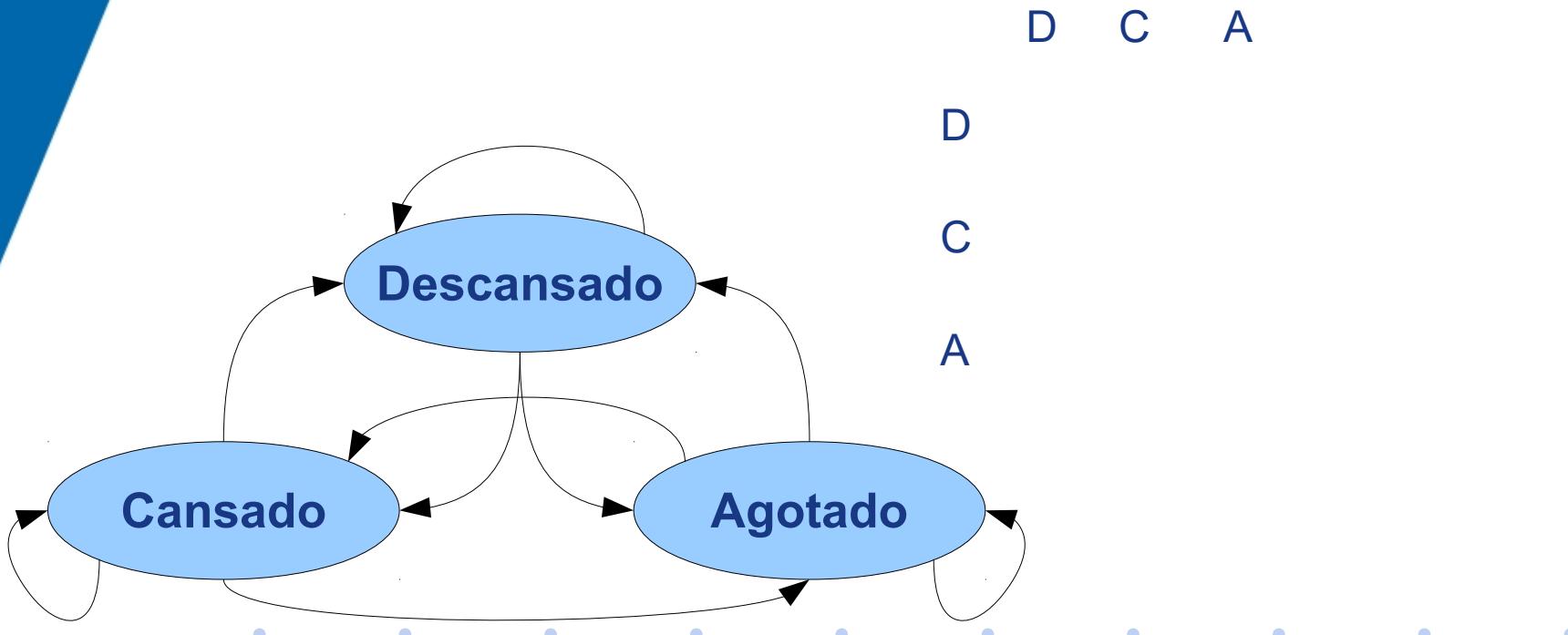
# Modelos Markovianos

- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .



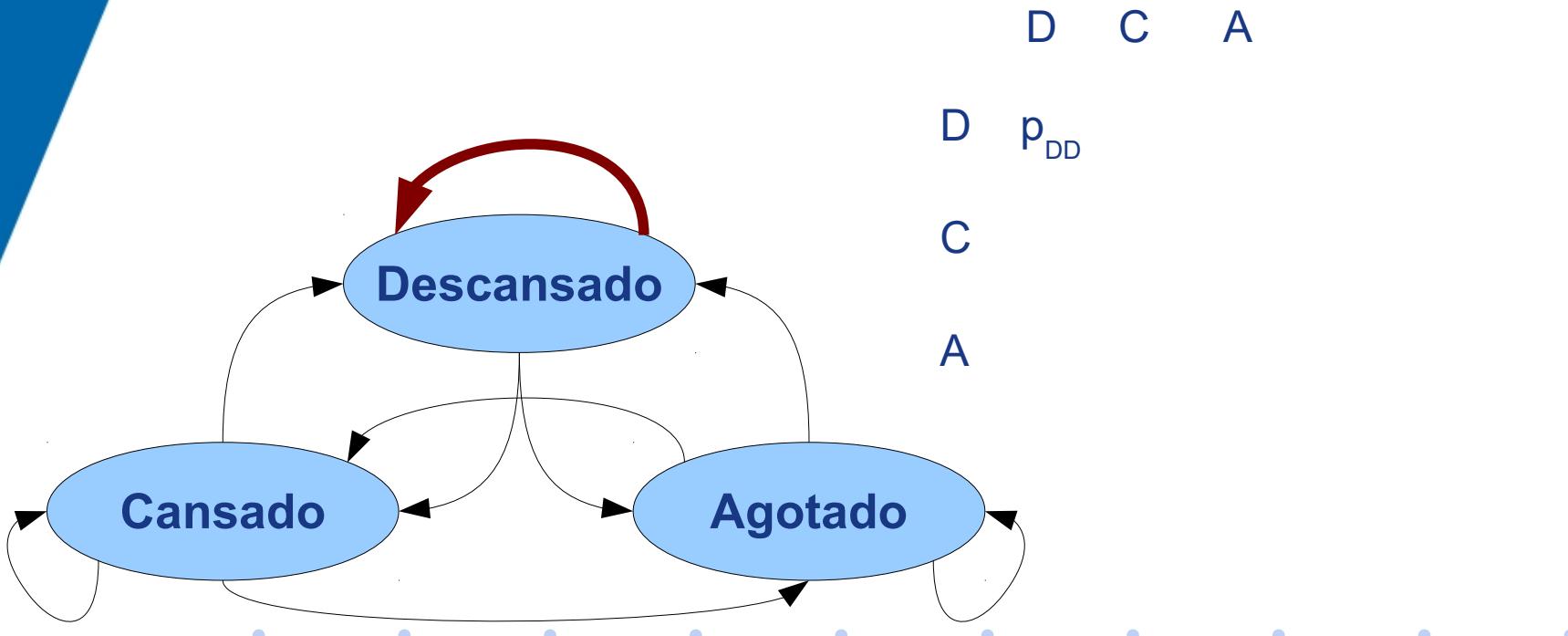
# Modelos Markovianos

- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .



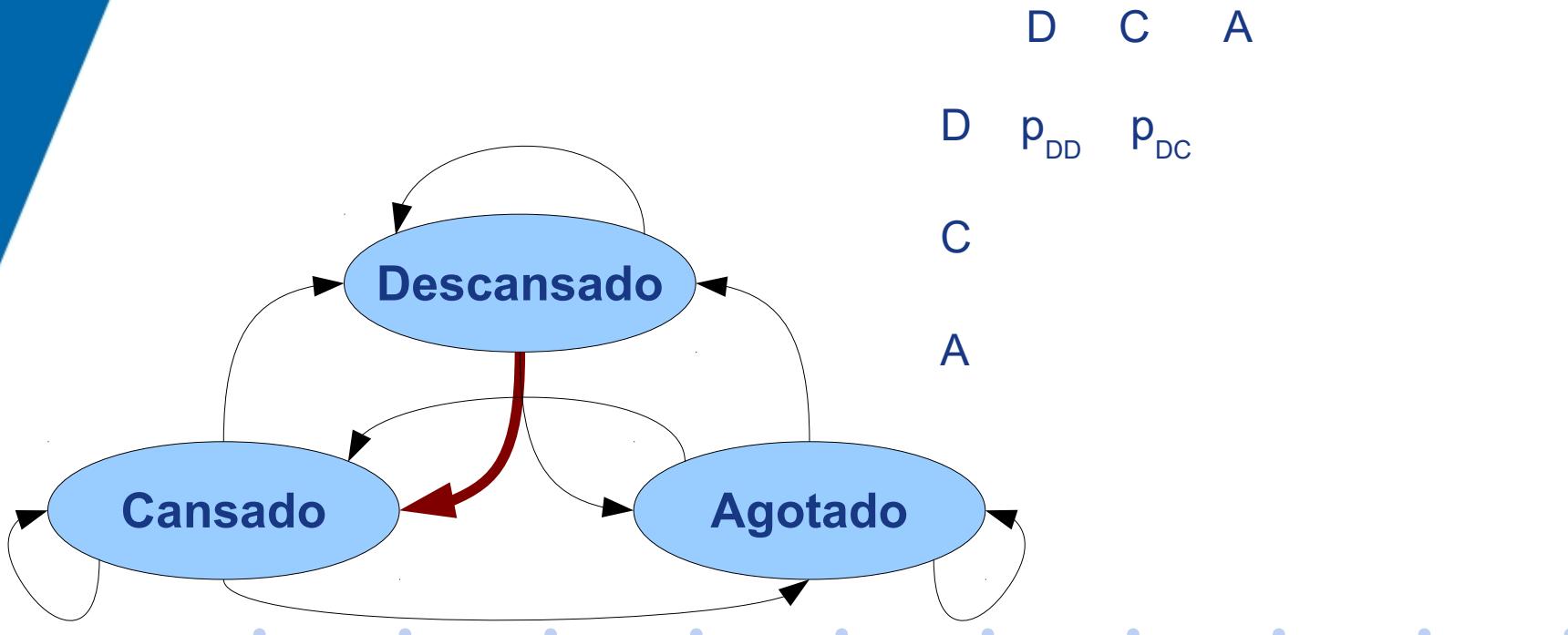
# Modelos Markovianos

- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .



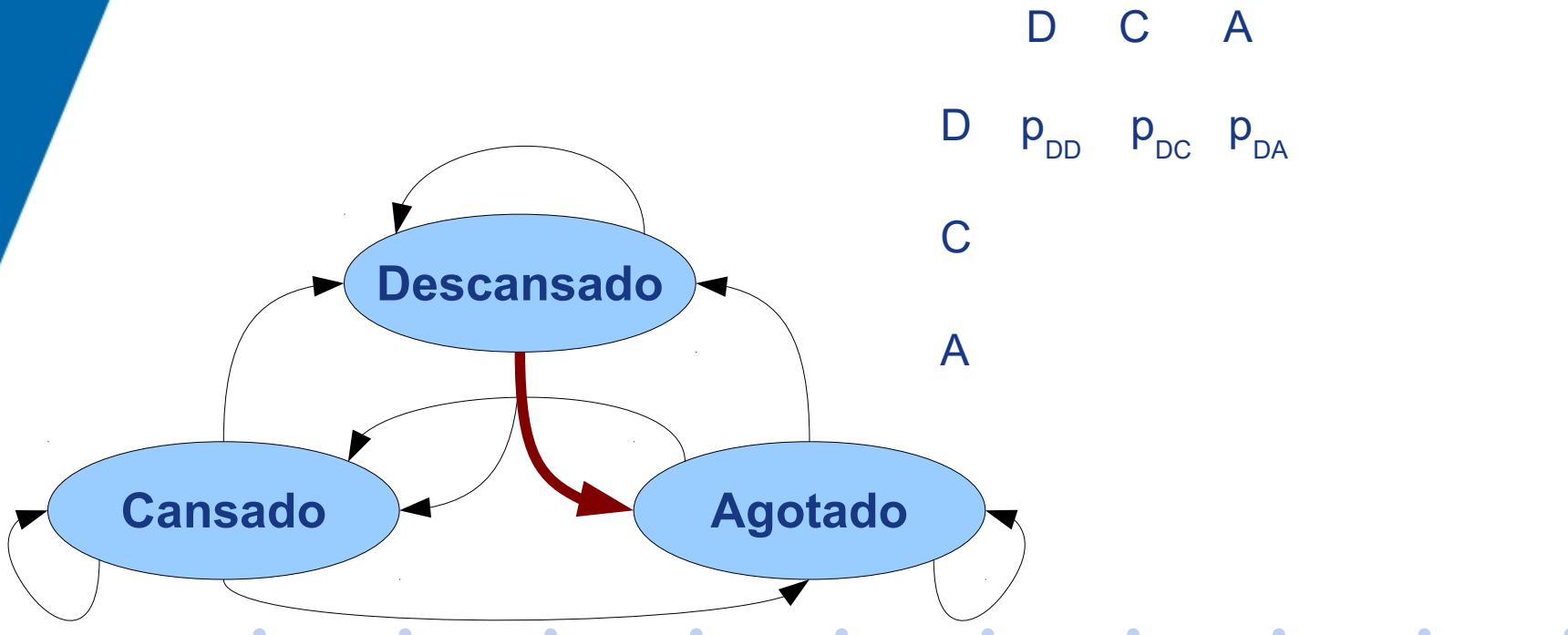
# Modelos Markovianos

- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .



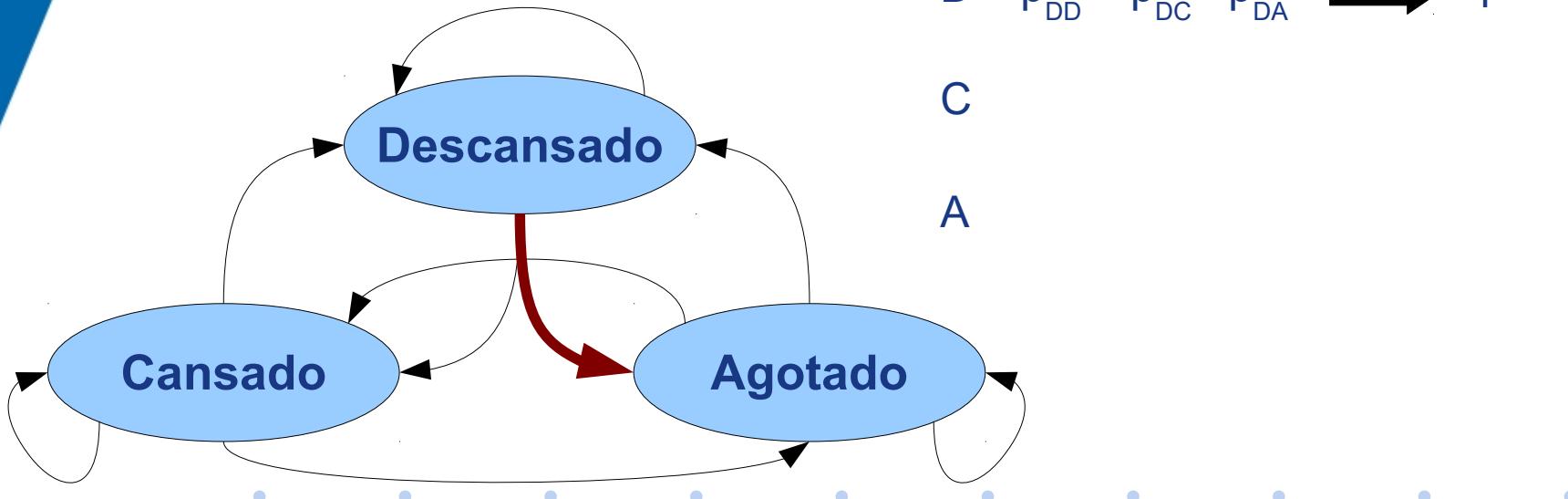
# Modelos Markovianos

- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .



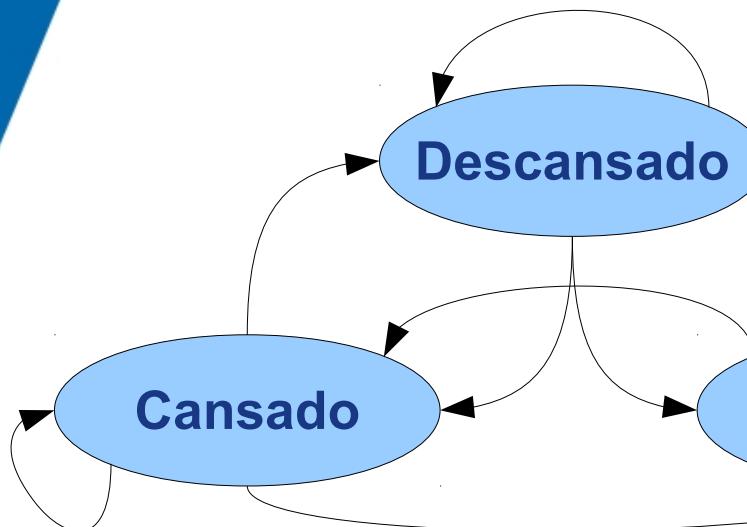
# Modelos Markovianos

- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .



# Modelos Markovianos

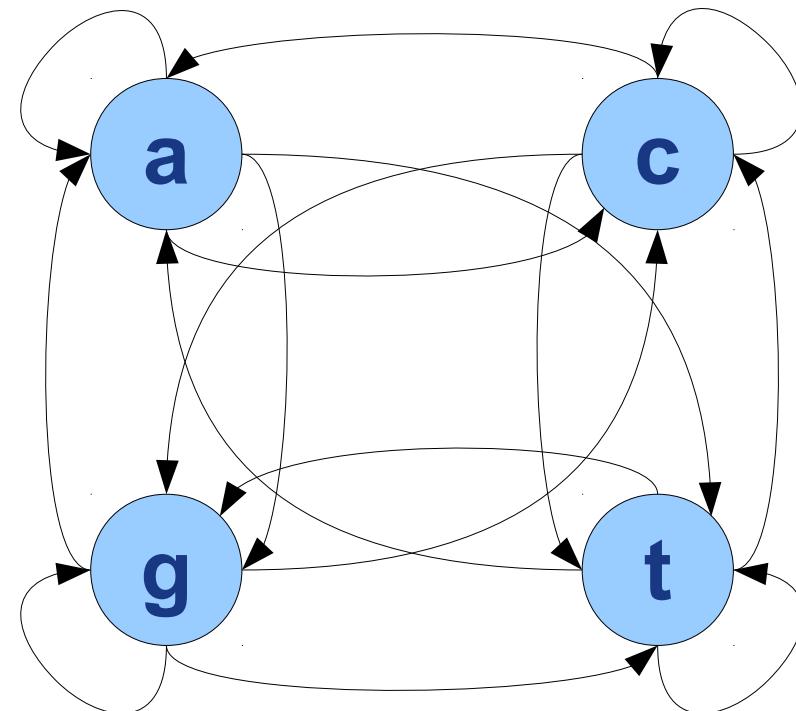
- Una **cadena de Markov** queda determinada por:
  - Vector de estados,  $S = \{ D, C, A \}$
  - Modelo multinomial con probabilidades iniciales.
  - Matriz de transición,  $T$ .



	D	C	A		
D	$p_{DD}$	$p_{DC}$	$p_{DA}$	1	
C	$p_{CD}$	$p_{CC}$	$p_{CA}$	1	
A	$p_{AD}$	$p_{AC}$	$p_{AA}$	1	

# Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

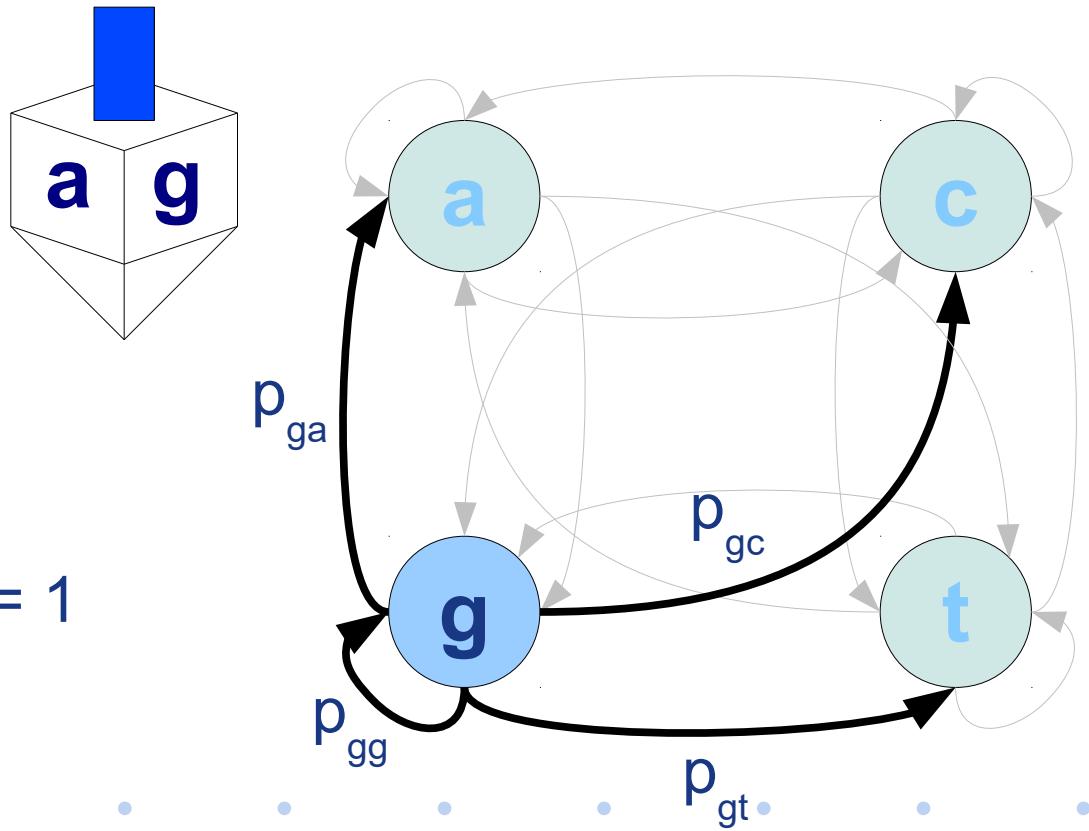


# Modelos Markovianos

g a g t t t a t c **g** c t t c c a t g a c g c a g a a g

$$\begin{aligned} p_{ga} &= P[a | g] \\ p_{gc} &= P[c | g] \\ p_{gg} &= P[g | g] \\ p_{gt} &= P[t | g] \end{aligned}$$

$$p_{ga} + p_{gc} + p_{gg} + p_{gt} = 1$$

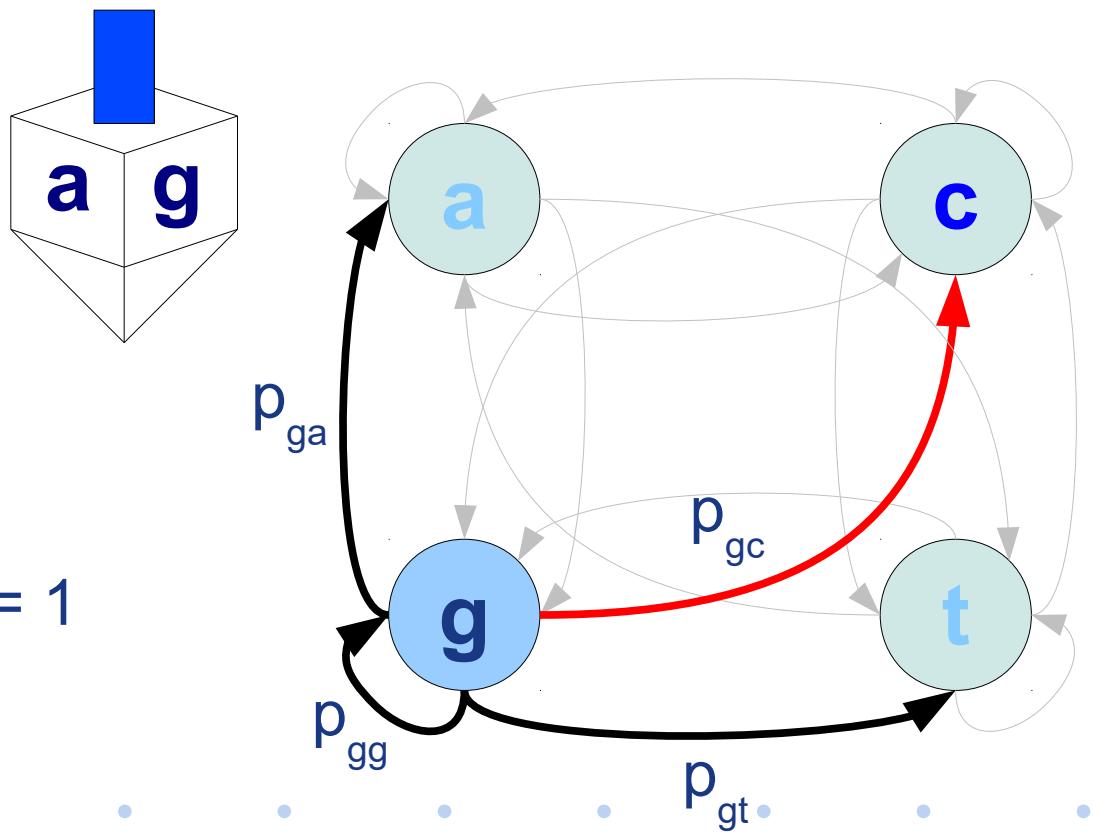


# Modelos Markovianos

g a g t t t a t c **g c** t t c c a t g a c g c a g a a g

$$\begin{aligned} p_{ga} &= P[a | g] \\ p_{gc} &= P[c | g] \\ p_{gg} &= P[g | g] \\ p_{gt} &= P[t | g] \end{aligned}$$

$$p_{ga} + p_{gc} + p_{gg} + p_{gt} = 1$$

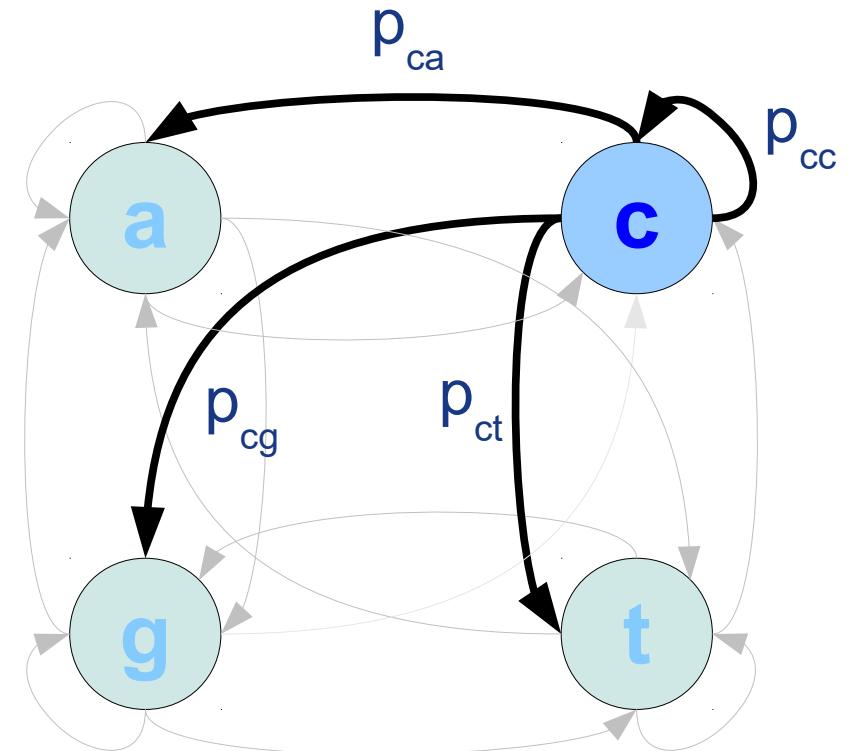
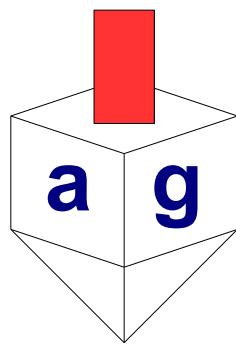


# Modelos Markovianos

g a g t t t a t c g **c t** t c c a t g a c g c a g a a g

$$\begin{aligned} p_{ca} &= P[a | c] \\ p_{cc} &= P[c | c] \\ p_{cg} &= P[g | c] \\ p_{ct} &= P[t | c] \end{aligned}$$

$$p_{ca} + p_{cc} + p_{cg} + p_{ct} = 1$$

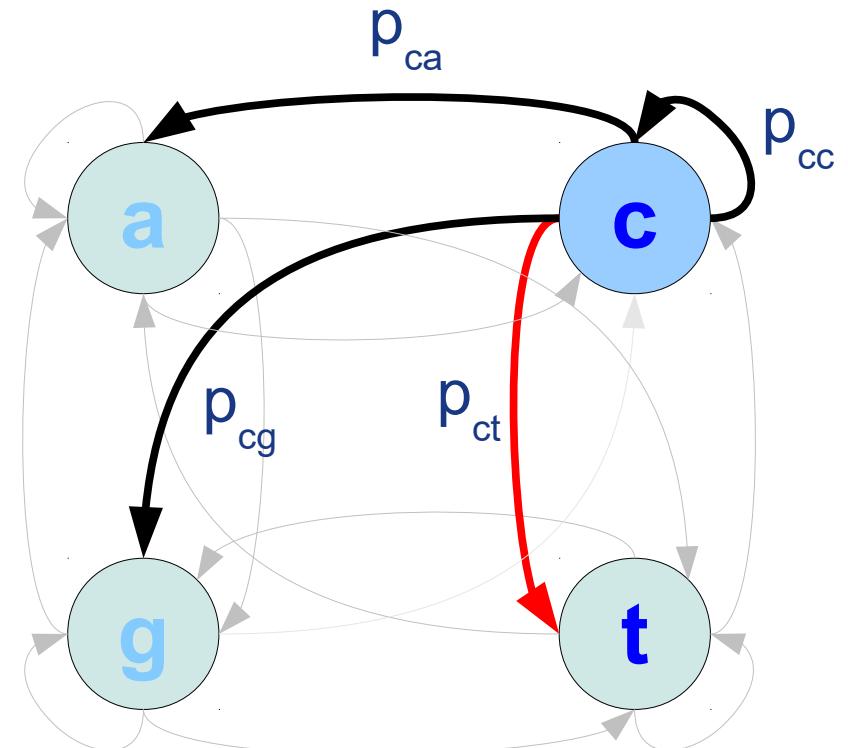
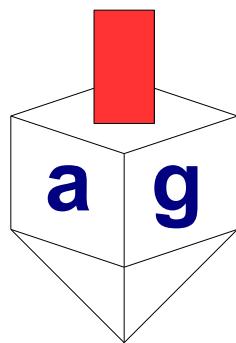


# Modelos Markovianos

g a g t t t a t c g **c t** t c c a t g a c g c a g a a g

$$\begin{aligned} p_{ca} &= P[a | c] \\ p_{cc} &= P[c | c] \\ p_{cg} &= P[g | c] \\ p_{ct} &= P[t | c] \end{aligned}$$

$$p_{ca} + p_{cc} + p_{cg} + p_{ct} = 1$$

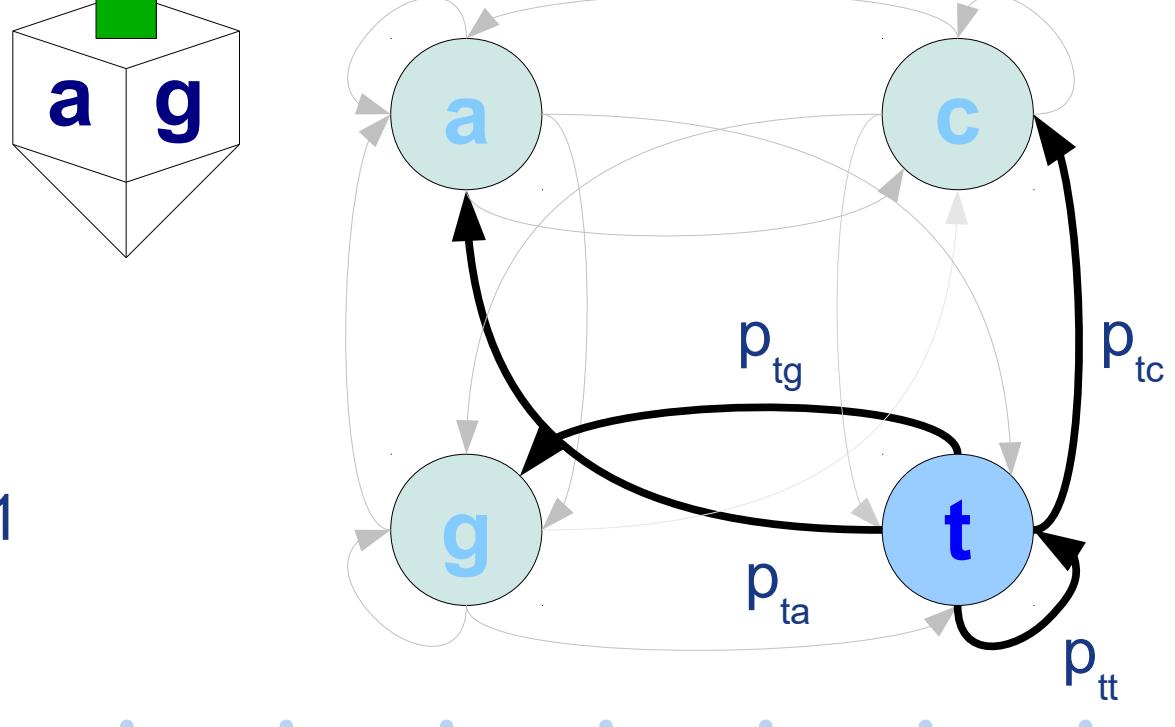
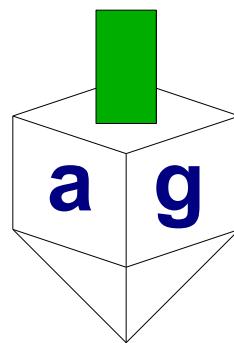


# Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

$$\begin{aligned} p_{ta} &= P[a | t] \\ p_{tc} &= P[c | t] \\ p_{tg} &= P[g | t] \\ p_{tt} &= P[t | t] \end{aligned}$$

$$p_{ta} + p_{tc} + p_{tg} + p_{tt} = 1$$

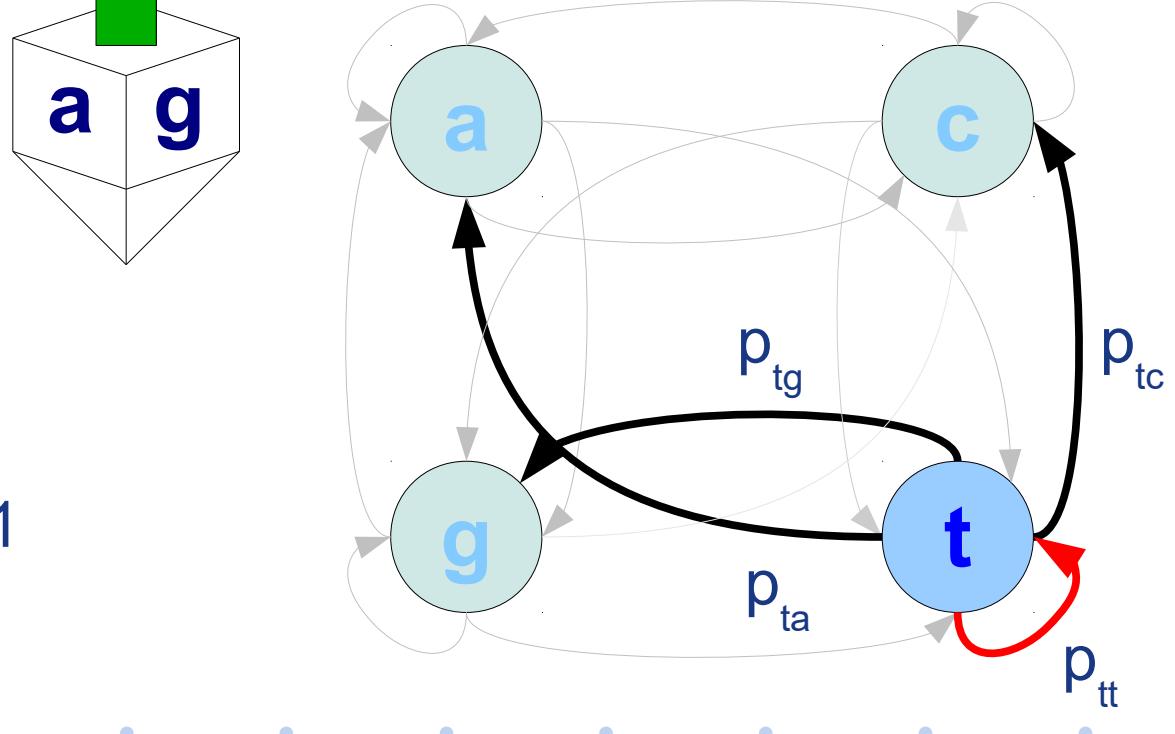
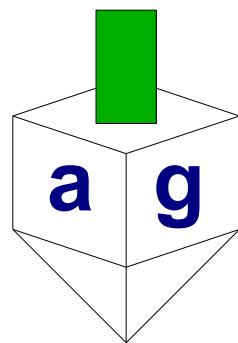


# Modelos Markovianos

g a g t t t a t c g c **t t** c c a t g a c g c a g a a g

$$\begin{aligned} p_{ta} &= P[a | t] \\ p_{tc} &= P[c | t] \\ p_{tg} &= P[g | t] \\ p_{tt} &= P[t | t] \end{aligned}$$

$$p_{ta} + p_{tc} + p_{tg} + p_{tt} = 1$$

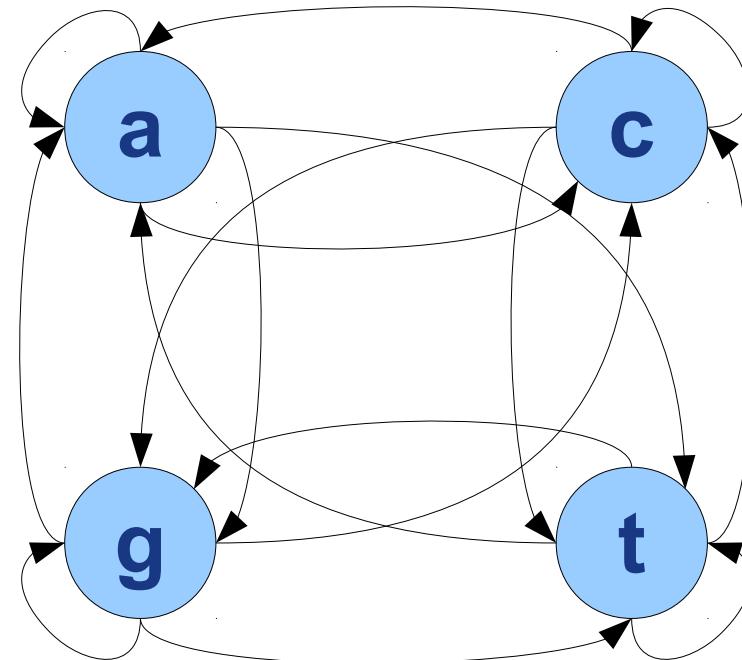


# Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

g

$$\begin{aligned}P[s_1 = a] &= p_a \\P[s_1 = c] &= p_c \\P[s_1 = g] &= p_g \\P[s_1 = t] &= p_t\end{aligned}$$



• • • • • • • • • •

# Modelos Markovianos

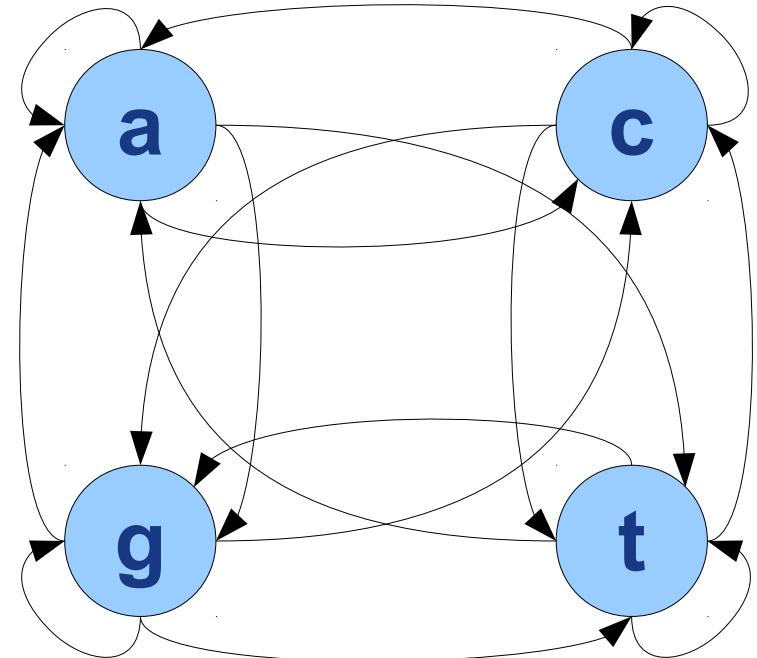
g a g t t t t a t c **g** c t t c c a t g a c g c a g a a g

- El modelo markoviana de una secuencia de DNA queda unívocamente definido por:
- Un vector de probabilidades iniciales:
- Una matriz de transición:

$$\mathbf{p}_0 = (p_a, p_c, p_g, p_t)$$

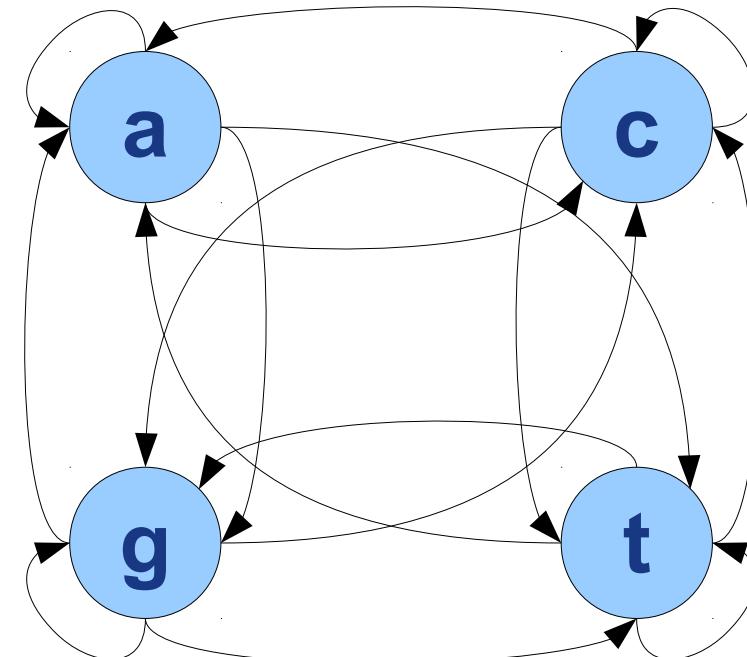
a c g t

$$\begin{array}{ccccc} & a & c & g & t \\ a & p_{aa} & p_{ac} & p_{ag} & p_{at} \\ c & p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ g & p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ t & p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{array}$$



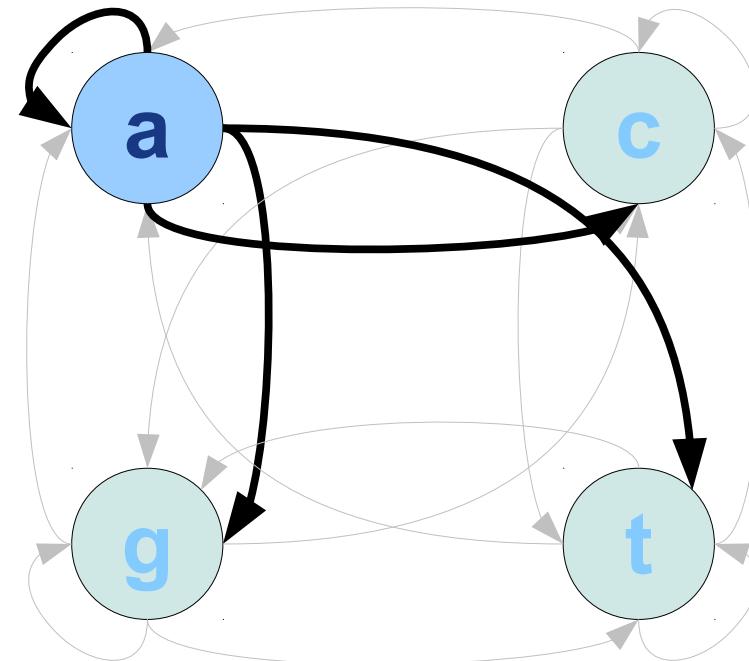
# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



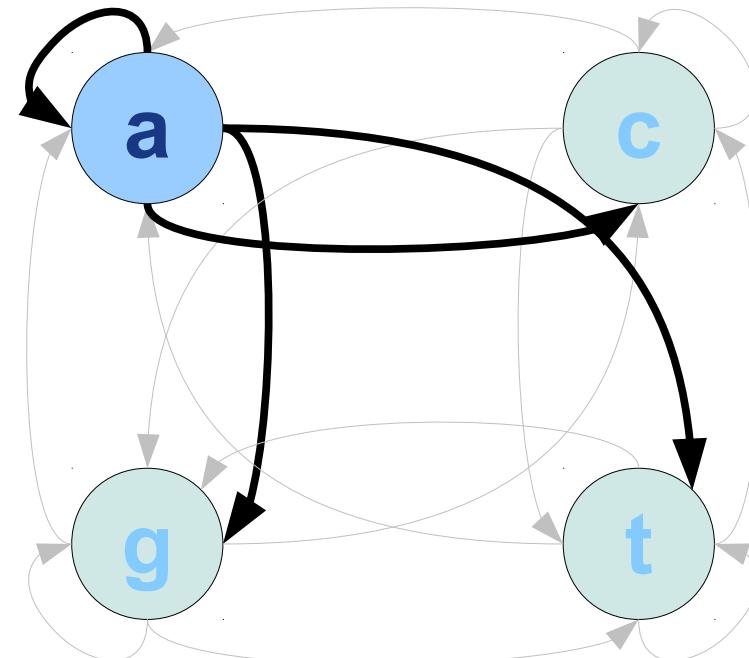
# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



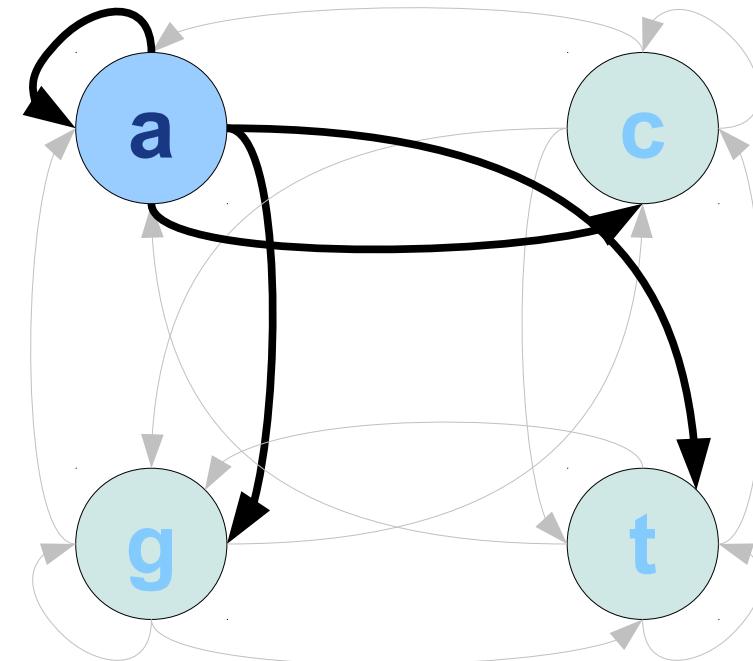
# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



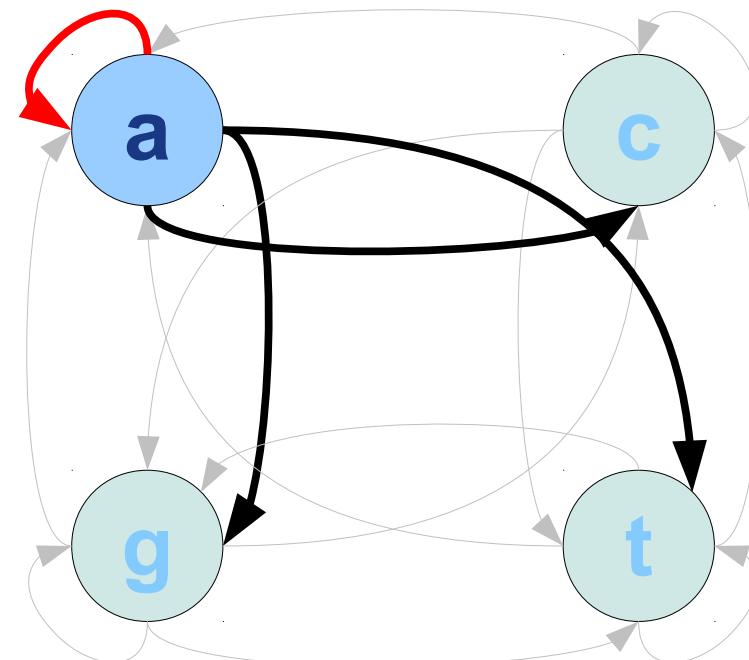
# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



# Estimación de Modelos Markovianos

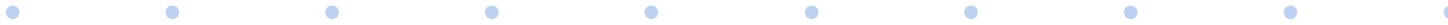
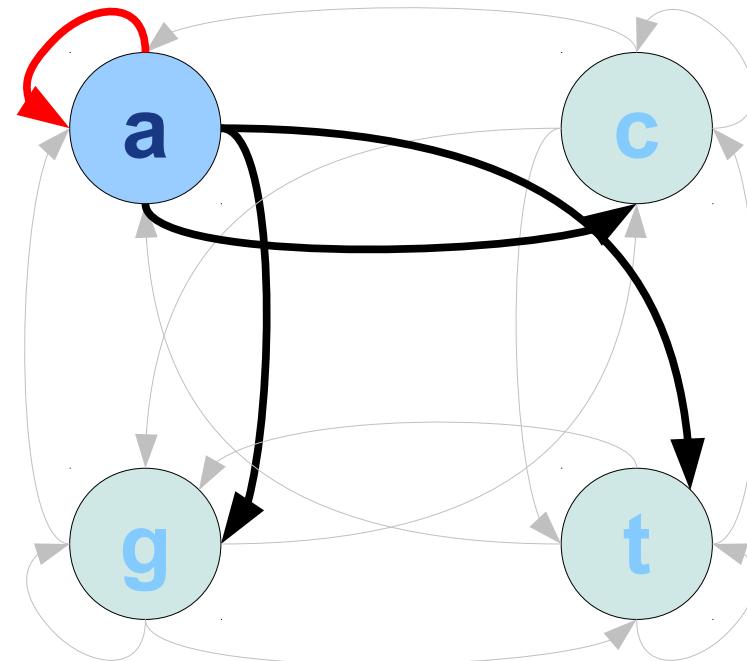
g a g t t t a t c g c t t c c a t g a c g c a g a a g



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

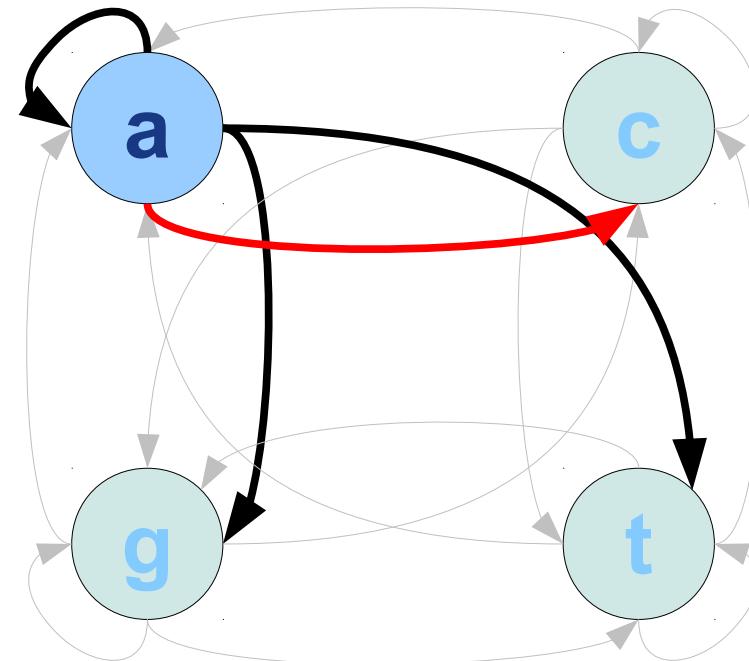
$\text{freq}(\text{"aa"}) = 1$



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

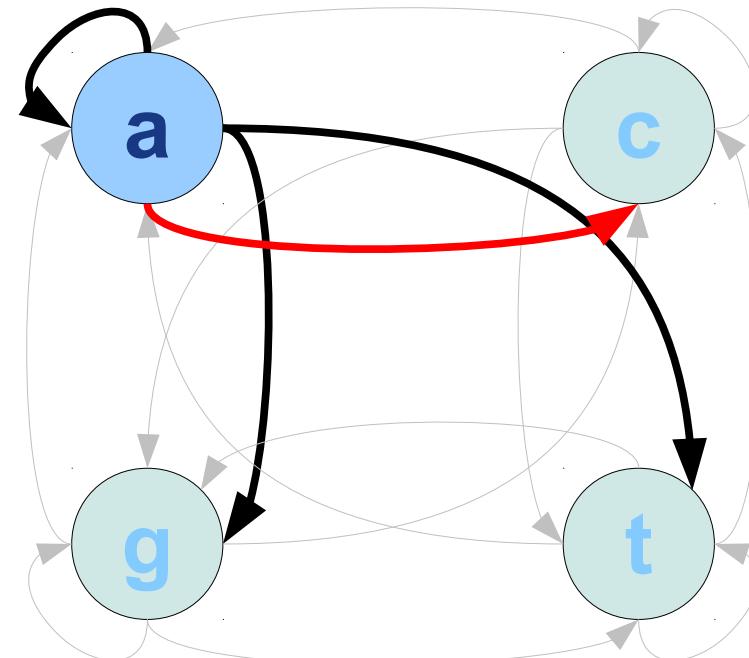
$\text{freq}(\text{"aa"}) = 1$



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

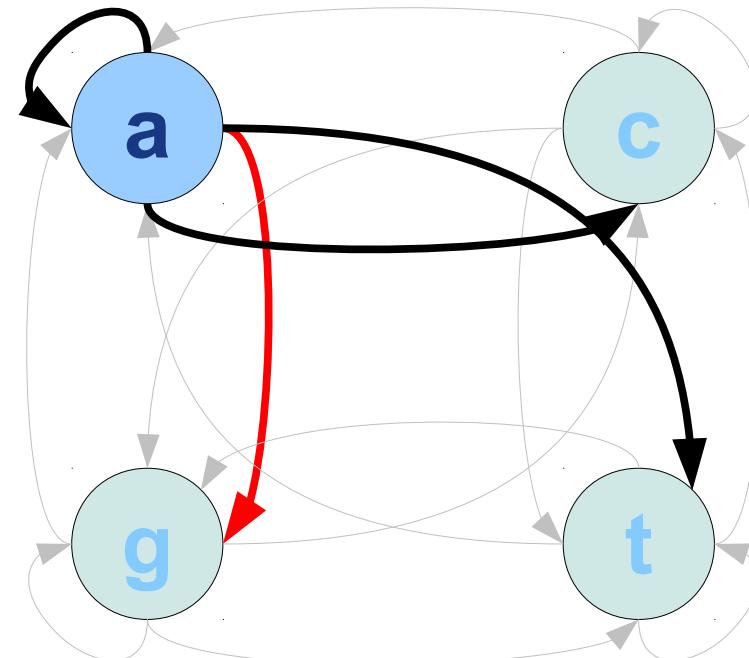
$$\begin{aligned}\text{freq("aa")} &= 1 \\ \text{freq("ac")} &= 1\end{aligned}$$



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

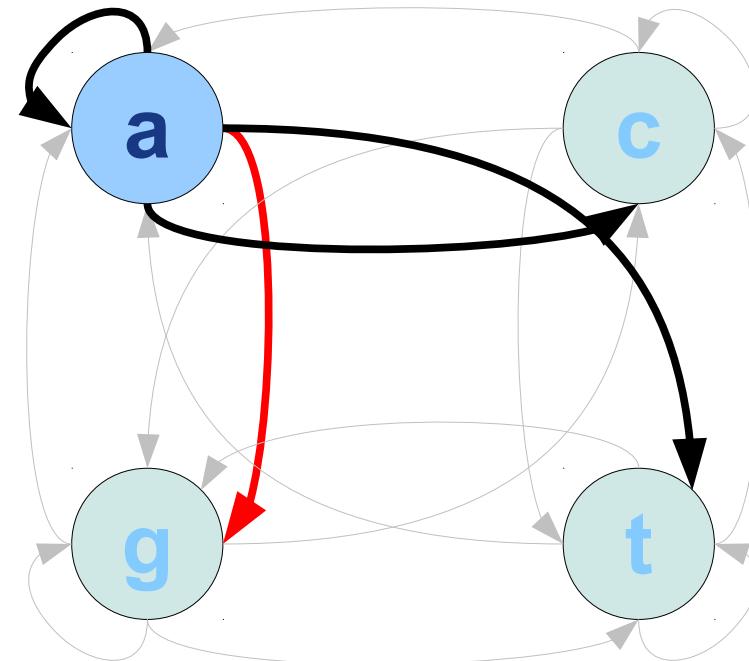
$\text{freq}(\text{"aa"}) = 1$   
 $\text{freq}(\text{"ac"}) = 1$



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

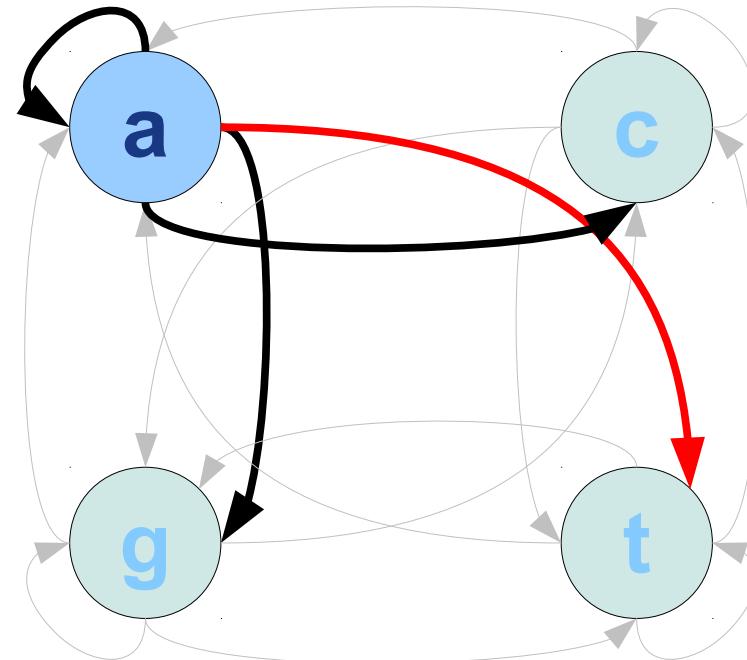
$\text{freq}(\text{"aa"}) = 1$   
 $\text{freq}(\text{"ac"}) = 1$   
 $\text{freq}(\text{"ag"}) = 3$



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

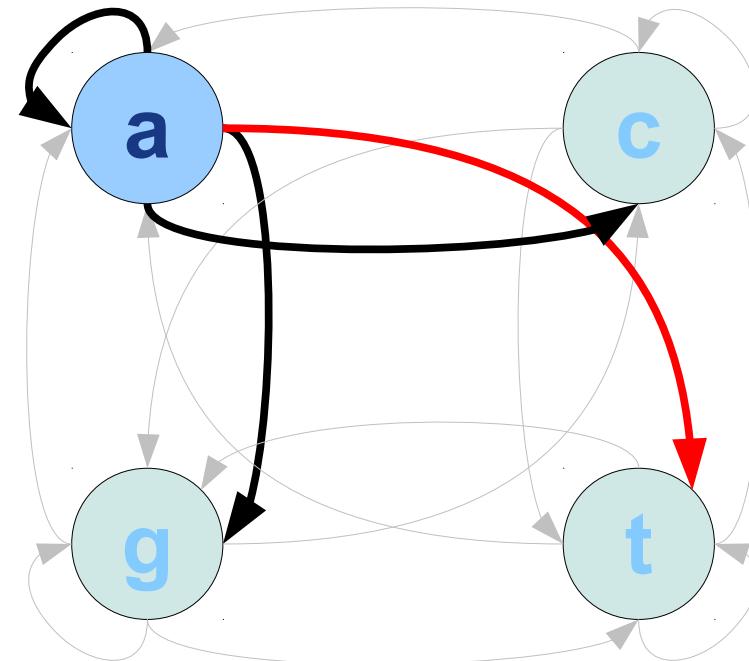
$\text{freq}(\text{"aa"}) = 1$   
 $\text{freq}(\text{"ac"}) = 1$   
 $\text{freq}(\text{"ag"}) = 3$



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

$\text{freq}(\text{"aa"}) = 1$   
 $\text{freq}(\text{"ac"}) = 1$   
 $\text{freq}(\text{"ag"}) = 3$   
 $\text{freq}(\text{"at"}) = 2$

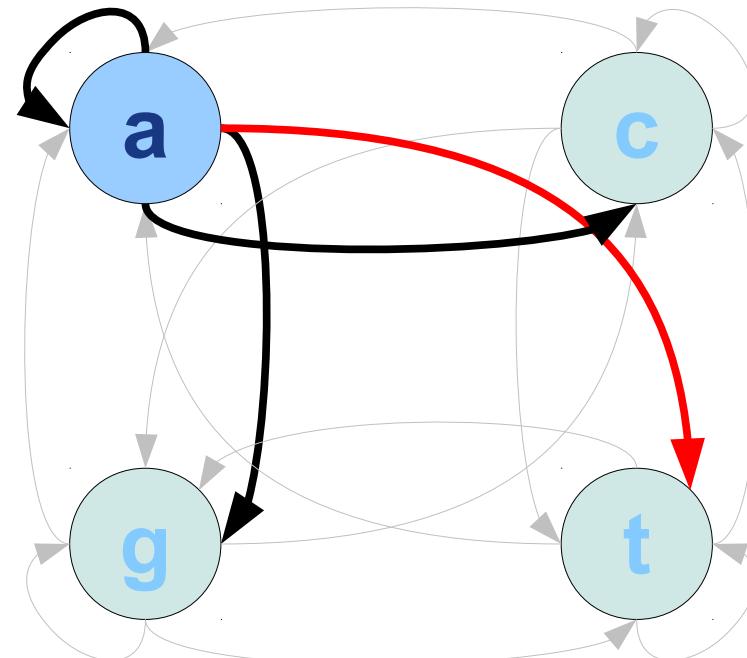


# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

$\text{freq}(\text{"aa"}) = 1$   
 $\text{freq}(\text{"ac"}) = 1$   
 $\text{freq}(\text{"ag"}) = 3$   
 $\text{freq}(\text{"at"}) = 2$

$\text{total}(\text{"ax"}) = 7$



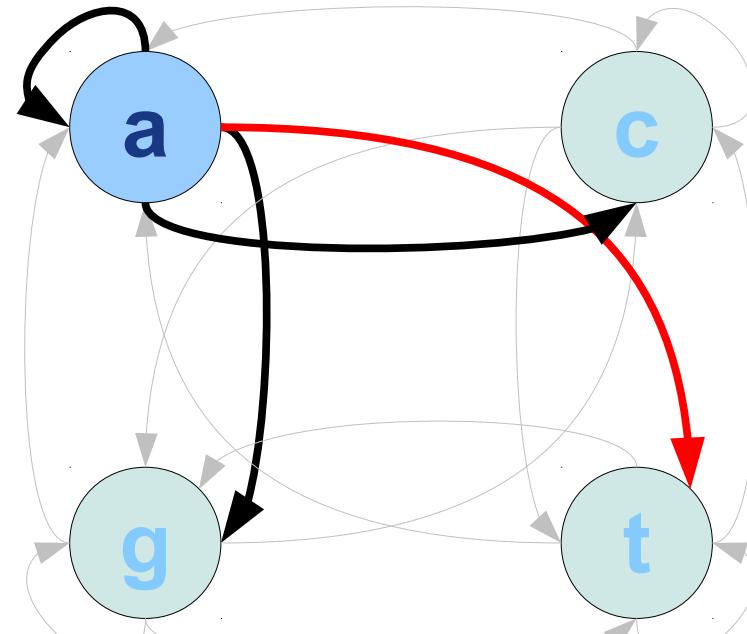
# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

$\text{freq}(\text{"aa"}) = 1$   
 $\text{freq}(\text{"ac"}) = 1$   
 $\text{freq}(\text{"ag"}) = 3$   
 $\text{freq}(\text{"at"}) = 2$

$\text{total}(\text{"ax"}) = 7$

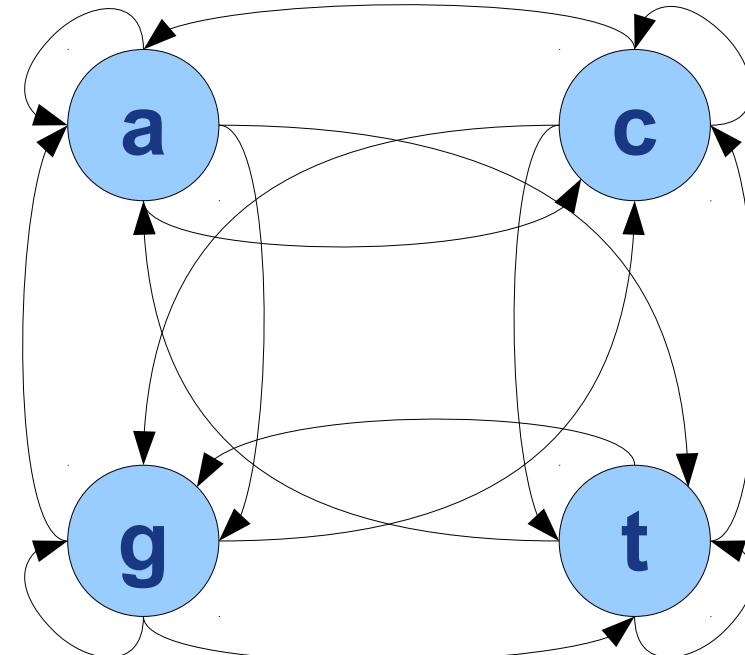
$p_{aa} = 1/7$     $p_{ac} = 1/7$   
 $p_{ag} = 3/7$     $p_{at} = 2/7$



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g

La función **count** de seqinr recibe como entrada una secuencia sequence y un número natural N y devuelve la frecuencia absoluta de todas las palabras de longitud N en sequence.



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$S = \{a, c, g, t\}, p_0 = (p_a, p_c, p_g, p_t)$$

$$\begin{array}{cccc} p_{aa} & p_{ac} & p_{ag} & p_{at} \\ p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{array}$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$S = \{a, c, g, t\}, p_0 = (p_a, p_c, p_g, p_t)$$

$$\begin{array}{cccc} p_{aa} & p_{ac} & p_{ag} & p_{at} \\ p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{array}$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

Paso 1: Calcular las frecuencias absolutas de dinucleotidos con count.



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$S = \{a, c, g, t\}, p_0 = (p_a, p_c, p_g, p_t)$$

$$\begin{matrix} p_{aa} & p_{ac} & p_{ag} & p_{at} \\ p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{matrix}$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

Paso 1: Calcular las frecuencias absolutas de dinucleotidos con count.

Paso 2: Crear una matriz 4x4 con el vector anterior



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$S=\{a,c,g,t\}, p_0 = (p_a, p_c, p_g, p_t)$$

$$\begin{matrix} p_{aa} & p_{ac} & p_{ag} & p_{at} \\ p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{matrix}$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

Paso 1: Calcular las frecuencias absolutas de dinucleotidos con count.

Paso 2: Crear una matriz 4x4 con el vector anterior

Paso 3: Dividir por la suma total de las frecuencias absolutas.



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$S=\{a,c,g,t\}, p_0 = (p_a, p_c, p_g, p_t)$$

$$\begin{matrix} p_{aa} & p_{ac} & p_{ag} & p_{at} \\ p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{matrix}$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

Paso 1: Calcular las frecuencias absolutas de dinucleotidos con count.

Paso 2: Crear una matriz 4x4 con el vector anterior

Paso 3: Dividir por la suma total de las frecuencias absolutas.

Paso 4: Nombrar cada fila y columna con a, c, g y t.



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$S=\{a,c,g,t\}, p_0 = (p_a, p_c, p_g, p_t)$$

$$\begin{matrix} p_{aa} & p_{ac} & p_{ag} & p_{at} \\ p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{matrix}$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

Paso 1: Calcular las frecuencias absolutas de dinucleotidos con count.

Paso 2: Crear una matriz 4x4 con el vector anterior

Paso 3: Dividir por la suma total de las frecuencias absolutas.

Paso 4: Nombrar cada fila y columna con a, c, g y t.

Paso 5: Calcular el modelo multinomial para las probabilidades iniciales.



# Estimación de Modelos Markovianos

g a g t t t a t c g c t t c c a t g a c g c a g a a g



$$S=\{a,c,g,t\}, p_0 = (p_a, p_c, p_g, p_t)$$

$$\begin{matrix} p_{aa} & p_{ac} & p_{ag} & p_{at} \\ p_{ca} & p_{cc} & p_{cg} & p_{ct} \\ p_{ga} & p_{gc} & p_{gg} & p_{gt} \\ p_{ta} & p_{tc} & p_{tg} & p_{tt} \end{matrix}$$

Entrada: Un vector de caracteres que representa una secuencia de DNA.

Paso 1: Calcular las frecuencias absolutas de dinucleotidos con count.

Paso 2: Crear una matriz 4x4 con el vector anterior

Paso 3: Dividir por la suma total de las frecuencias absolutas.

Paso 4: Nombrar cada fila y columna con a, c, g y t.

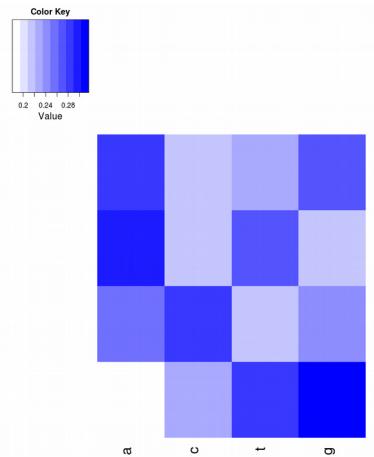
Paso 5: Calcular el modelo multinomial para las probabilidades iniciales.

Salida: Un vector formado por cuatro elementos que representan  $p_a$ ,  $p_c$ ,  $p_g$  y  $p_t$

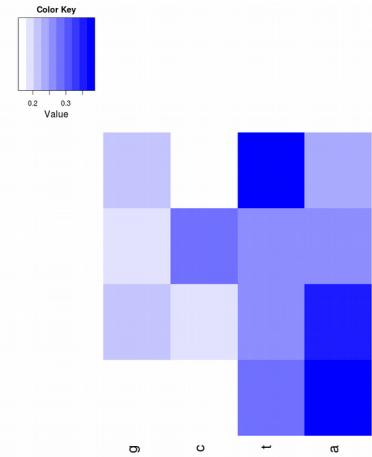
- Una matriz T que representa las transiciones

# Frecuencia de Dinucleótidos

- El modelo markoviano de la secuencia de un genoma suele referirse como frecuencia de dinucleótidos.
- La **frecuencia de dinucleótidos** es una característica específica del genoma de cada especie.



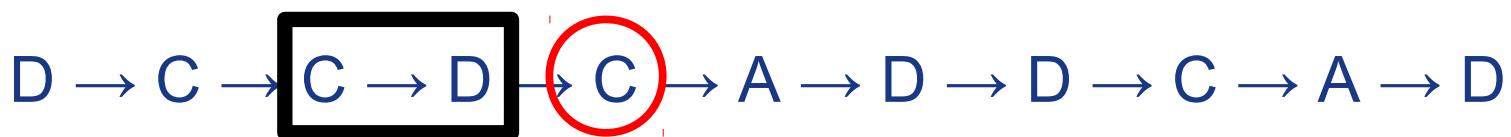
*Fago lambda*



*Haemophilus influenzae*

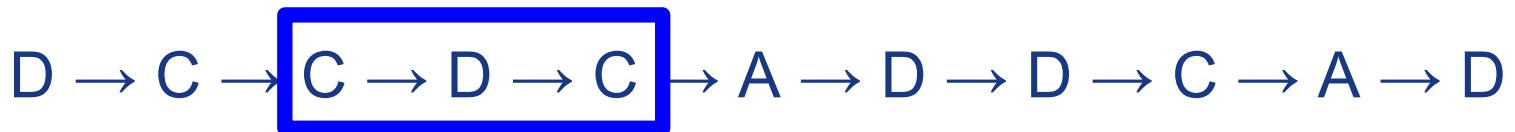
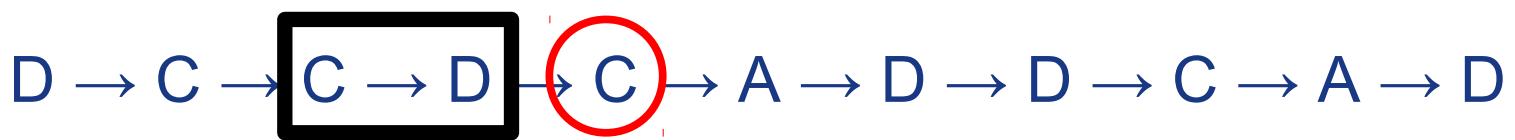
# Modelos Markovianos de Orden Superior

- Las **cadenas de Markov** donde la probabilidad de encontrarse en un estado depende de más de uno de los estados anteriores se denominan cadenas de Markov de orden superior.
- En las cadenas de Markov de orden  $k$  el estado actual depende de los  $k$  estados inmediatamente anteriores.



# Modelos Markovianos de Orden Superior

- Las **cadenas de Markov** de orden  $k$  se estiman a partir de la frecuencia de los  $(k+1)$ -meros.
- Se suelen referir como **frecuencias de k-meros** y son una característica específica de cada especie.



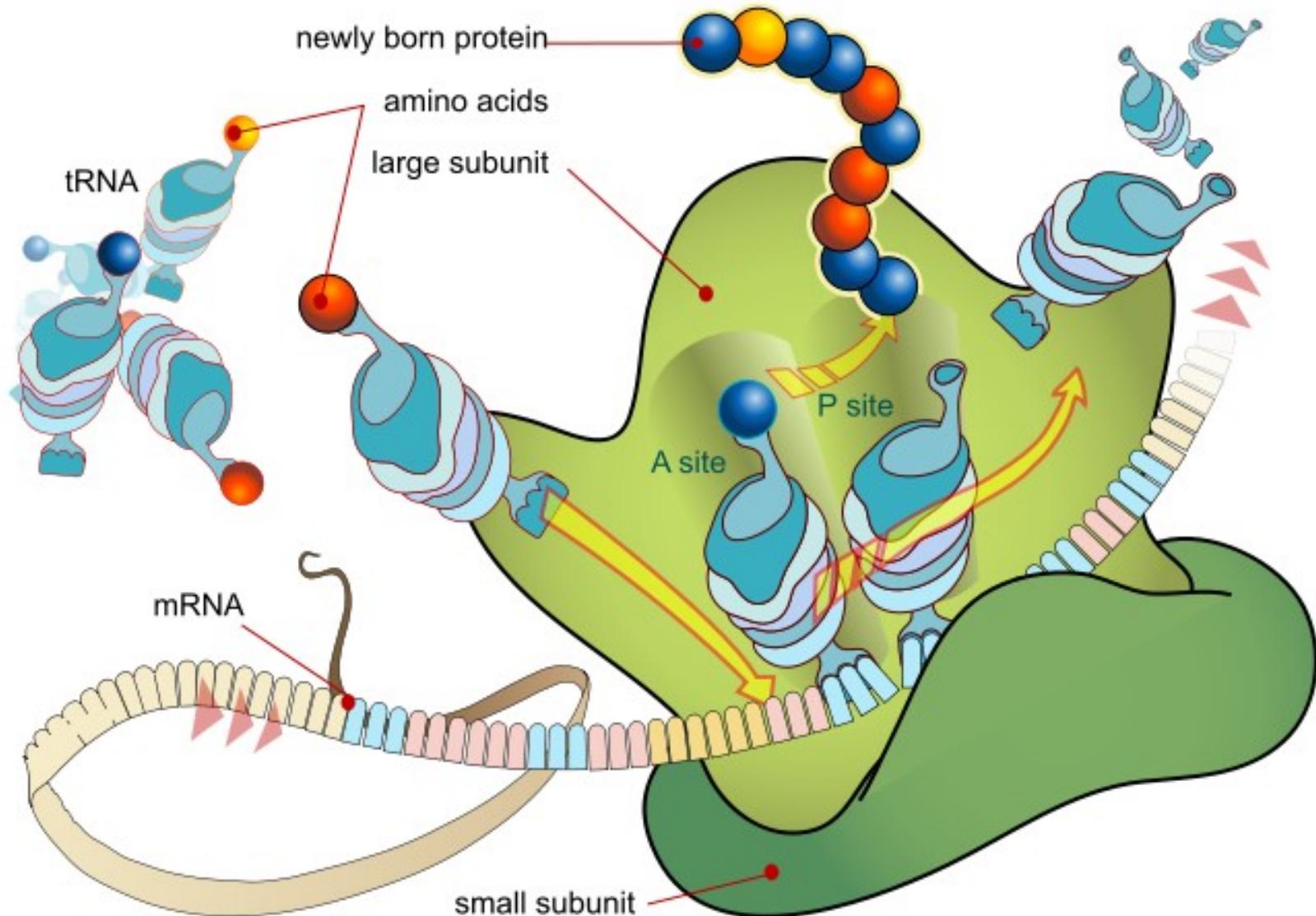
• • • • • • • • • •

# Guión de la Unidad

- Introducción histórica
- Definiciones Básicas
- Modelos de Secuencias de DNA
  - Modelos multinomiales:
    - Estimación de modelos multinomiales
    - Composición de bases
    - Contenido en GC
  - Modelos markovianos:
    - Estimación de modelos markovianos
    - Frecuencia de k-meros
- Sesgos en el Uso de Codones



# Traducción de Proteínas

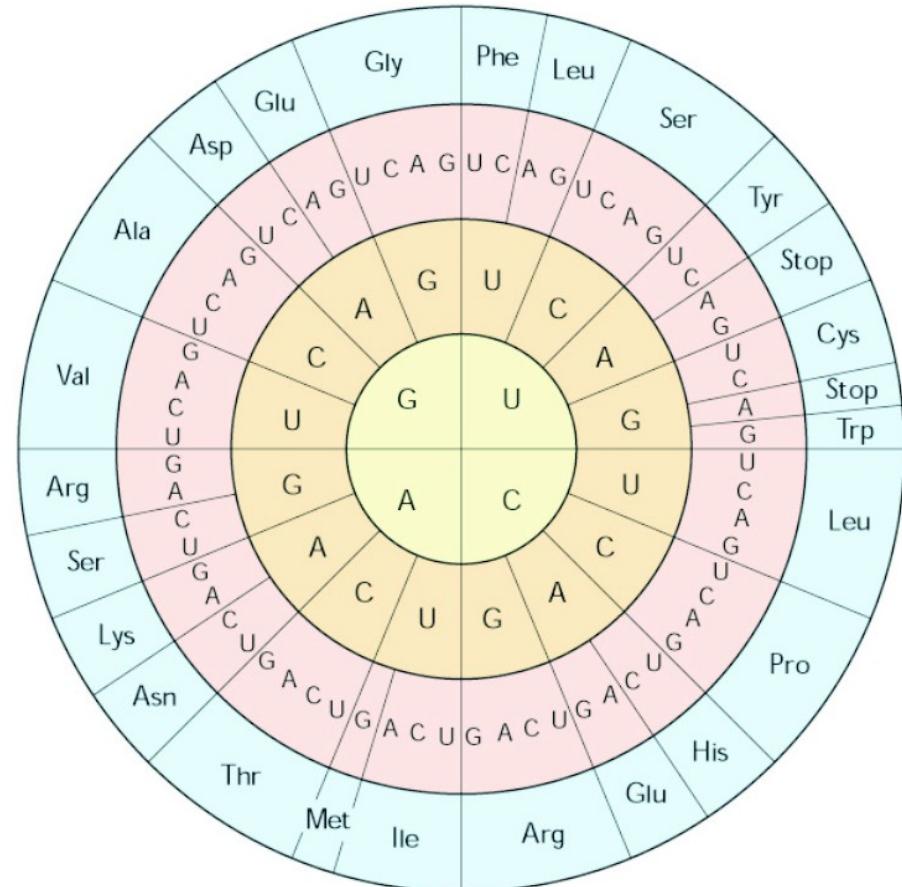


# Redundancia en el Código Genético

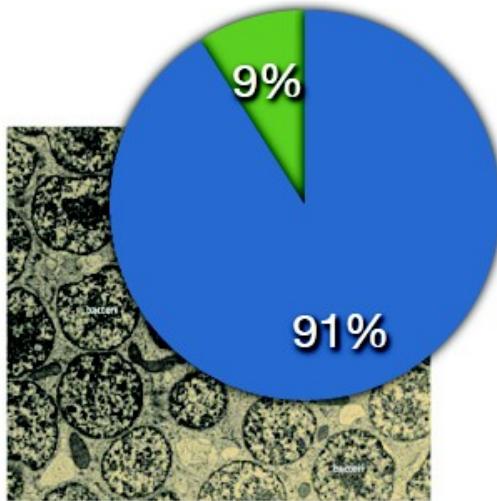
El código genético, la asociación de codones a aminoácidos, es redundante.

Diferentes codones codifican por el mismo aminoácido.

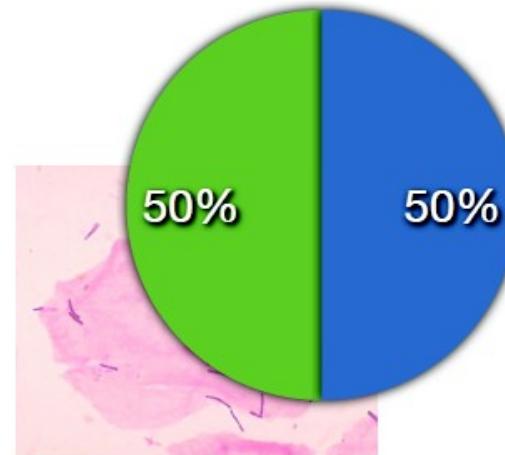
Por ejemplo, el aminoácido lisina puede ser codificado por los dos codones AAA y AAG.



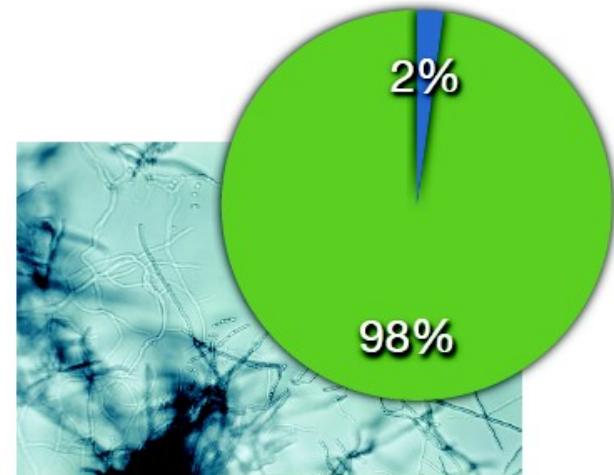
# Sesgos en el Uso de Codones



*Buchnera aphidicola*  
Endosimbionte de pulgones



*Lactobacillus acidophilus*  
Microbiota humana



*Streptomyces venezuelae*  
Producción antibiótico

# Sesgos en el Uso de Codones

- El uso de codones es una característica específica de un genoma.
- Se observan fuertes sesgos en el uso de codones en genomas bacterianos tales como *Escherichia coli* y eucariotas unicelulares tales como *Saccharomyces cerevisiae*. Los sesgos en el uso de codones de mamíferos parecen no ser tan marcados.



# Sesgos en el Uso de Codones

UUU	F	0.57	UCU	S	0.11	UAU	Y	0.53	UGU	C	0.42
UUC	F	0.43	UCC	S	0.11	UAC	Y	0.47	UGC	C	0.58
UUA	L	0.15	UCA	S	0.15	UAA	*	0.64	UGA	*	0.36
UUG	L	0.12	UCG	S	0.16	UAG	*	0.00	UGG	W	1.00
CUU	L	0.12	CCU	P	0.17	CAU	H	0.55	CGU	R	0.36
CUC	L	0.10	CCC	P	0.13	CAC	H	0.45	CGC	R	0.44
CUA	L	0.05	CCA	P	0.14	CAA	Q	0.30	CGA	R	0.07
CUG	L	0.46	CCG	P	0.55	CAG	Q	0.70	CGG	R	0.07
AUU	I	0.58	ACU	T	0.16	AAU	N	0.47	AGU	S	0.14
AUC	I	0.35	ACC	T	0.47	AAC	N	0.53	AGC	S	0.33
AUA	I	0.07	ACA	T	0.13	AAA	K	0.73	AGA	R	0.02
AUG	M	1.00	ACG	T	0.24	AAG	K	0.27	AGG	R	0.03
GUU	V	0.25	GCU	A	0.11	GAU	D	0.65	GGU	G	0.29
GUC	V	0.18	GCC	A	0.31	GAC	D	0.35	GGC	G	0.46
GUA	V	0.17	GCA	A	0.21	GAA	E	0.70	GGA	G	0.13
GUG	V	0.40	GCG	A	0.38	GAG	E	0.30	GGG	G	0.12

[Codon/a.a./fraction per codon per a.a.]

E. coli K12 data from the Codon Usage Database



# Sesgos en el Uso de Codones

UUU	F	0.46	UCU	S	0.19	UAU	Y	0.44	UGU	C	0.46
UUC	F	0.54	UCC	S	0.22	UAC	Y	0.56	UGC	C	0.54
UUA	L	0.08	UCA	S	0.15	UAA	*	0.30	UGA	*	0.47
UUG	L	0.13	UCG	S	0.05	UAG	*	0.24	UGG	W	1.00
CUU	L	0.13	CCU	P	0.29	CAU	H	0.42	CGU	R	0.08
CUC	L	0.20	CCC	P	0.32	CAC	H	0.58	CGC	R	0.18
CUA	L	0.07	CCA	P	0.28	CAA	Q	0.27	CGA	R	0.11
CUG	L	0.40	CCG	P	0.11	CAG	Q	0.73	CGG	R	0.20
AUU	I	0.36	ACU	T	0.25	AAU	N	0.47	AGU	S	0.15
AUC	I	0.47	ACC	T	0.36	AAC	N	0.53	AGC	S	0.24
AUA	I	0.17	ACA	T	0.28	AAA	K	0.43	AGA	R	0.21
AUG	M	1.00	ACG	T	0.11	AAG	K	0.57	AGG	R	0.21
GUU	V	0.18	GCU	A	0.27	GAU	D	0.46	GGU	G	0.16
GUC	V	0.24	GCC	A	0.40	GAC	D	0.54	GGC	G	0.34
GUA	V	0.12	GCA	A	0.23	GAA	E	0.42	GGA	G	0.25
GUG	V	0.46	GCG	A	0.11	GAG	E	0.58	GGG	G	0.25

[Codon/a.a./fraction per codon per a.a.]

Homo sapiens data from the Codon Usage Database



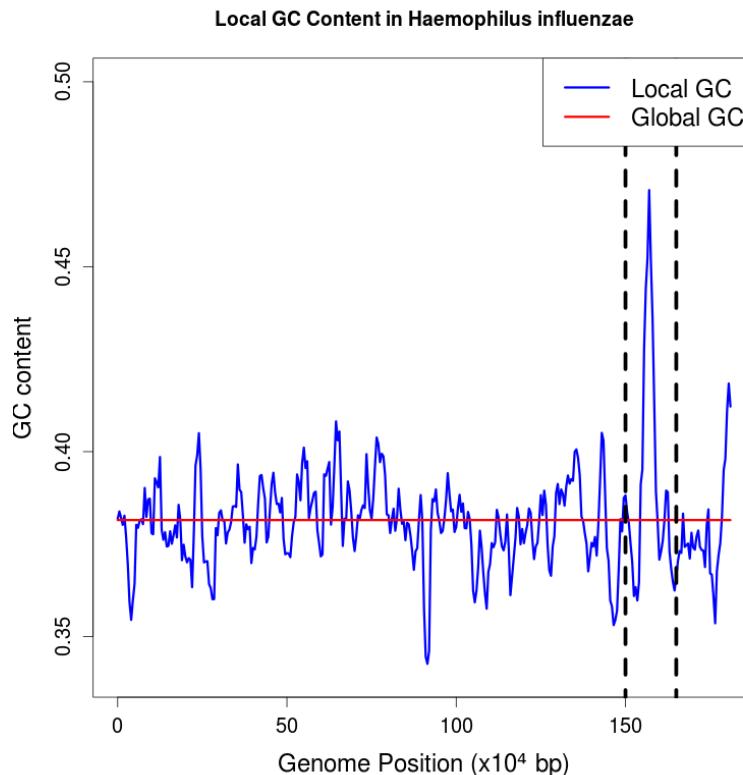
# Causas de los Sesgos en el Uso de Codones

- Se observa una correlación positiva entre el número de copias de genes codificantes por un tRNA específico y el uso del correspondiente codón.
- Otros investigadores defienden que los sesgos en el uso de codones surgen por las diferentes tasas de mutación.

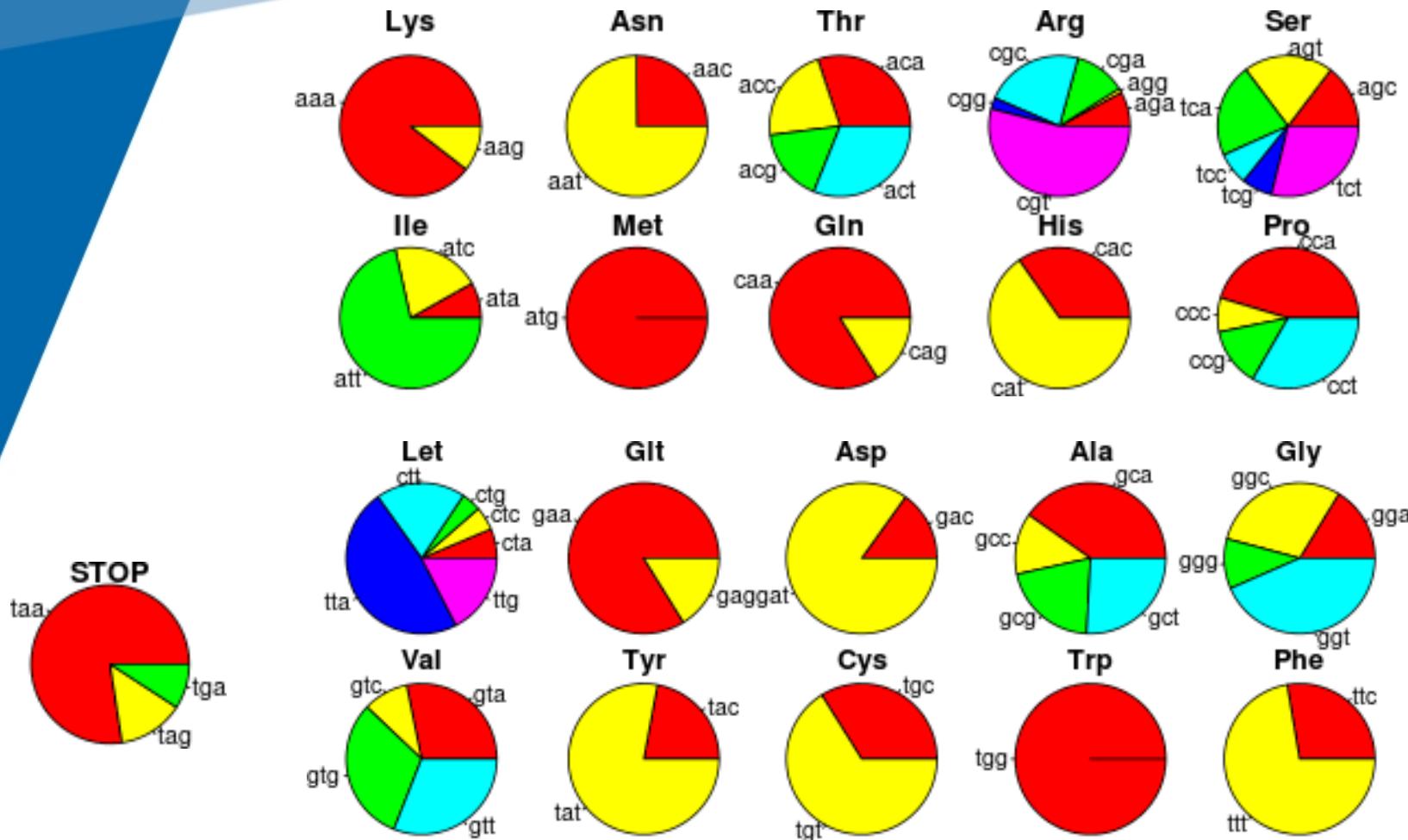


# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- Variaciones significativas en el uso de codones dentro del mismo genoma pueden indicar eventos de **transferencia horizontal de genes**.



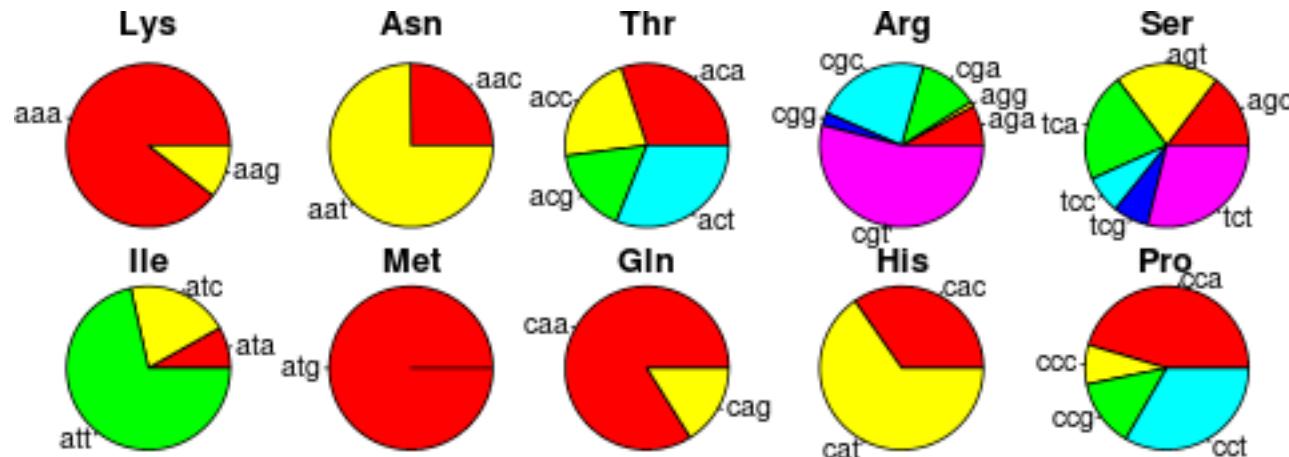
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes



- Uso de codones en *Haemophilus influenzae* •

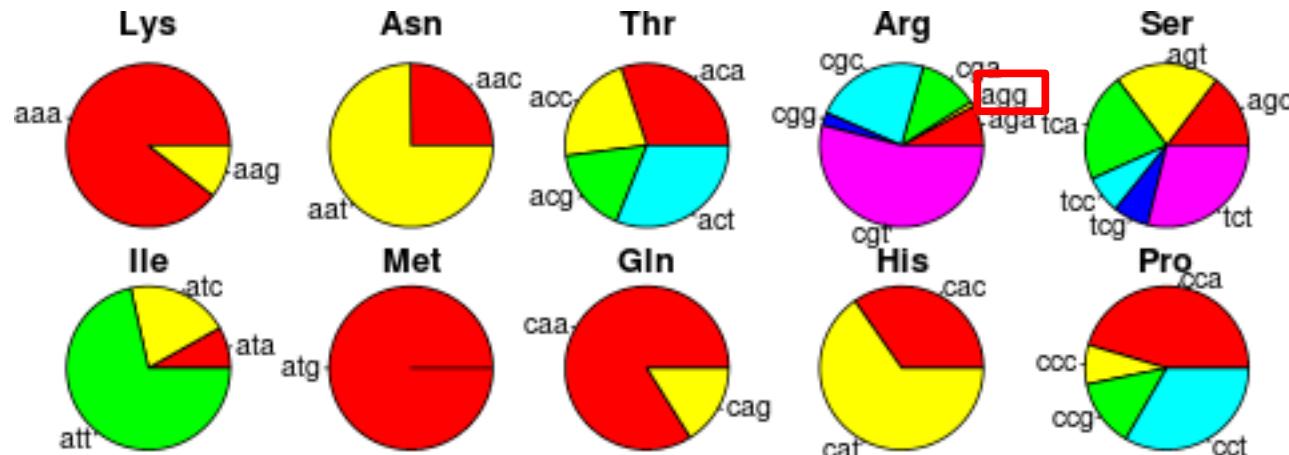
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- Se suelen considerar como **codones raros** aquellos cuyo uso es inferior al 5%.
- En el caso de *H. influenzae* sólo existe un codón raro, el **codón agg** que codifica por Arginina con un uso del 0.9 %.



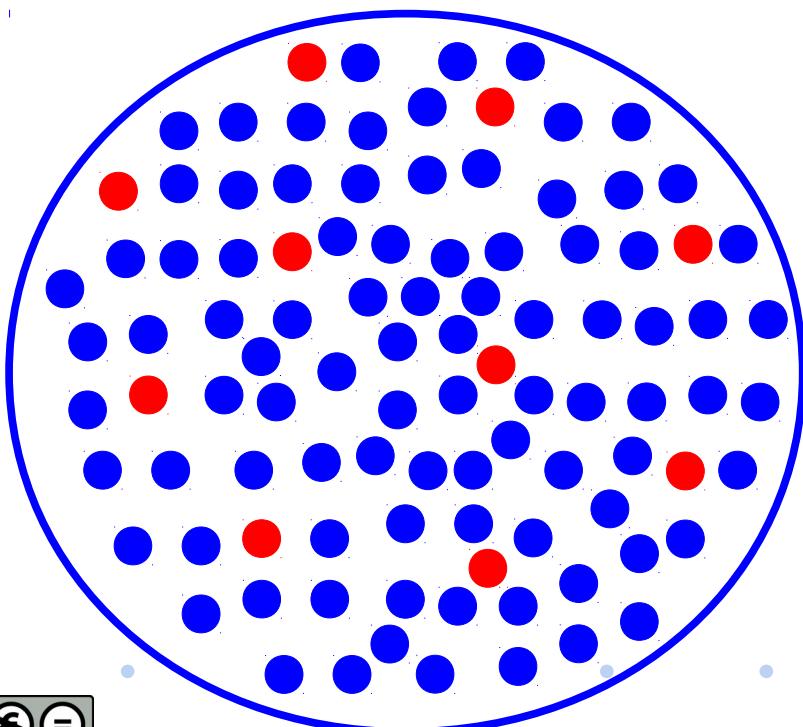
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- Se suelen considerar como **codones raros** aquellos cuyo uso es inferior al 5%.
- En el caso de *H. influenzae* sólo existe un codón raro, el **codón agg** que codifica por Arginina con un uso del 0.9 %.



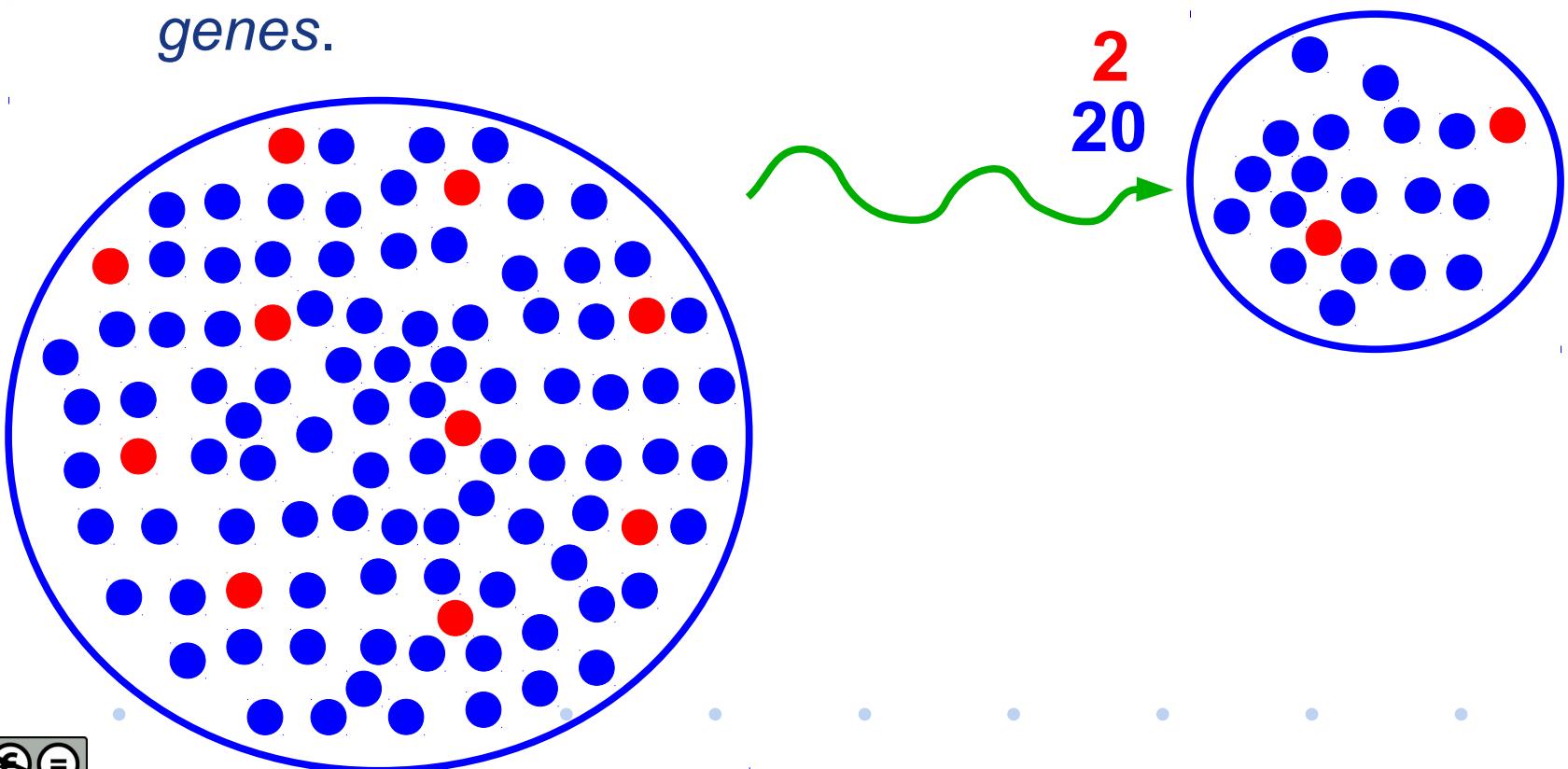
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- El **enriquecimiento significativo de codones raros** en regiones de un genoma constituye evidencias para la *transferencia horizontal de genes*.



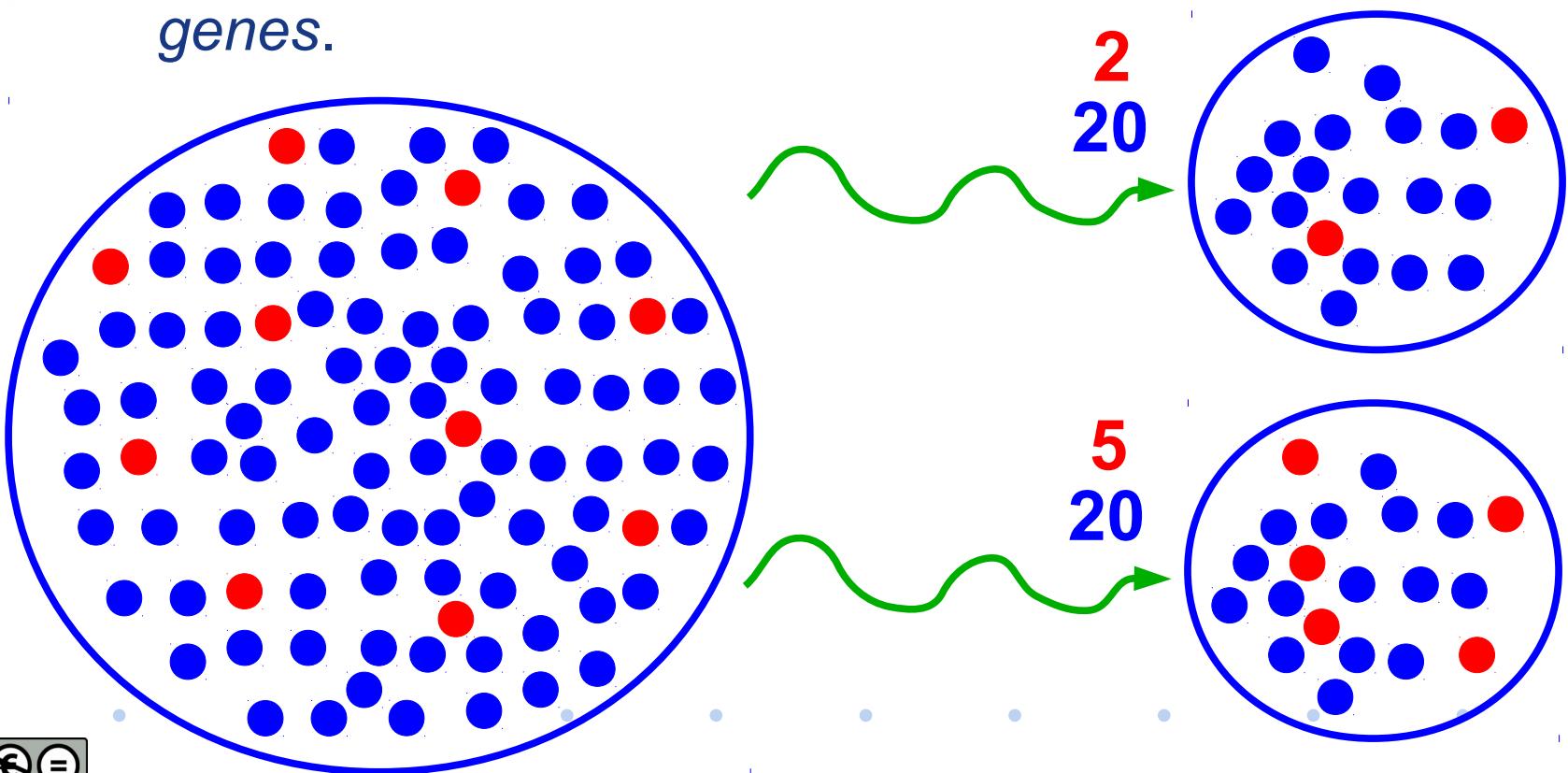
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- El **enriquecimiento significativo de codones raros** en regiones de un genoma constituye evidencias para la *transferencia horizontal de genes*.



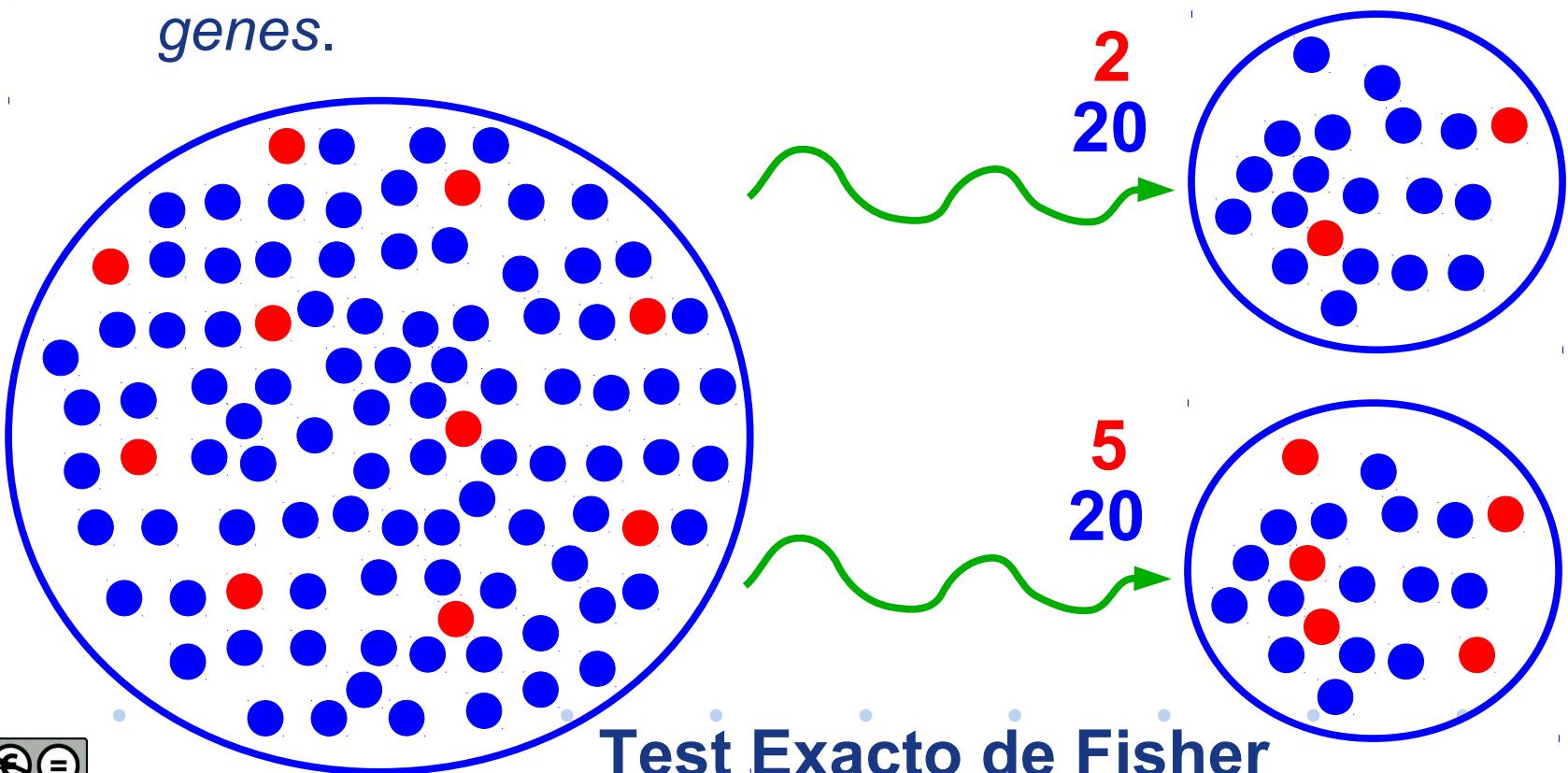
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- El **enriquecimiento significativo de codones raros** en regiones de un genoma constituye evidencias para la *transferencia horizontal de genes*.



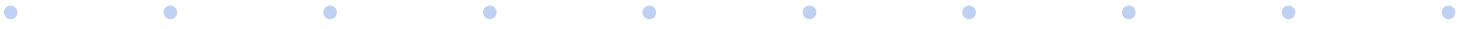
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- El **enriquecimiento significativo de codones raros** en regiones de un genoma constituye evidencias para la *transferencia horizontal de genes*.



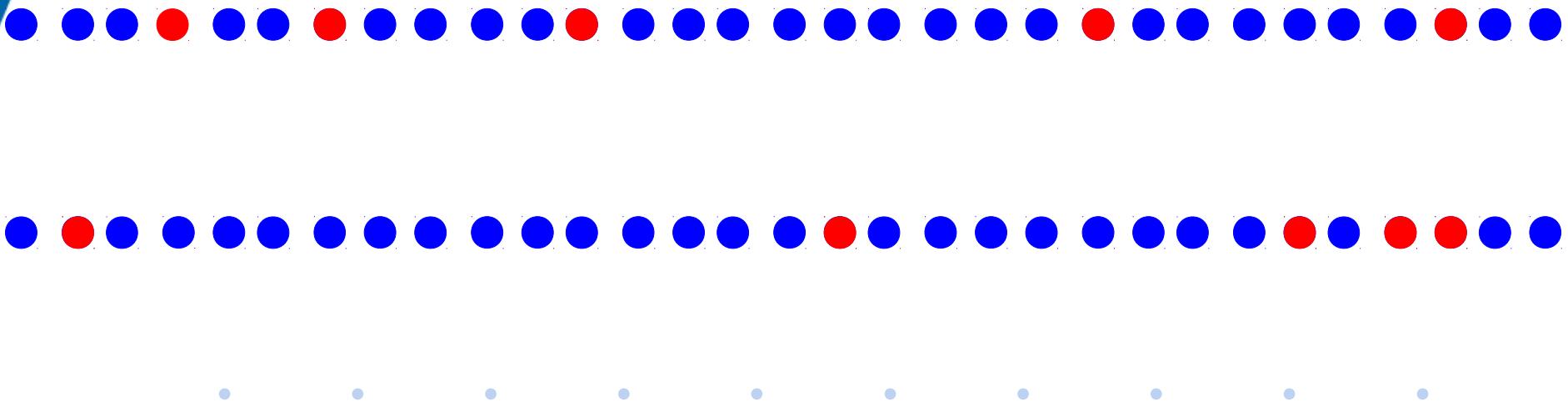
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- El **enriquecimiento significativo de codones raros** en regiones de un genoma constituye evidencias para la *transferencia horizontal de genes*.



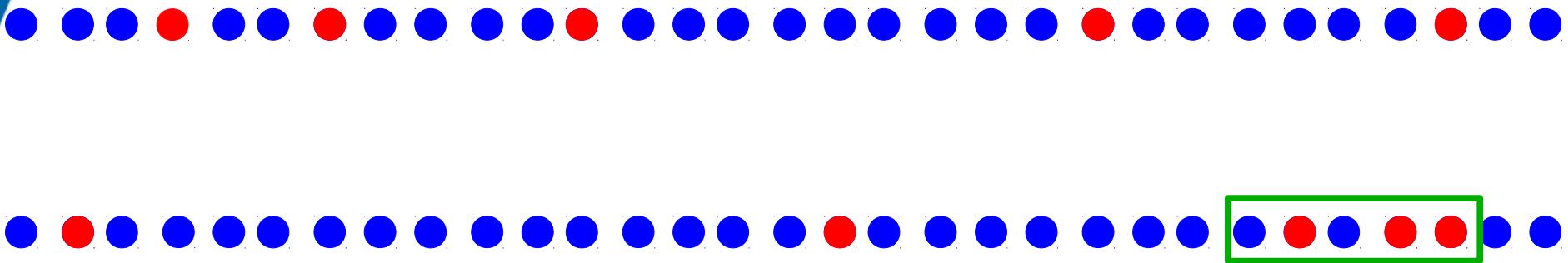
# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- El **enriquecimiento significativo de codones raros** en regiones de un genoma constituye evidencias para la *transferencia horizontal de genes*.



# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- El **enriquecimiento significativo de codones raros** en regiones de un genoma constituye evidencias para la *transferencia horizontal de genes*.



**Test Exacto de Fisher**

# Uso de Codones en la Identificación de Transferencia Horizontal de Genes

- El **enriquecimiento significativo de codones raros** en regiones de un genoma constituye evidencias para la *transferencia horizontal de genes*.

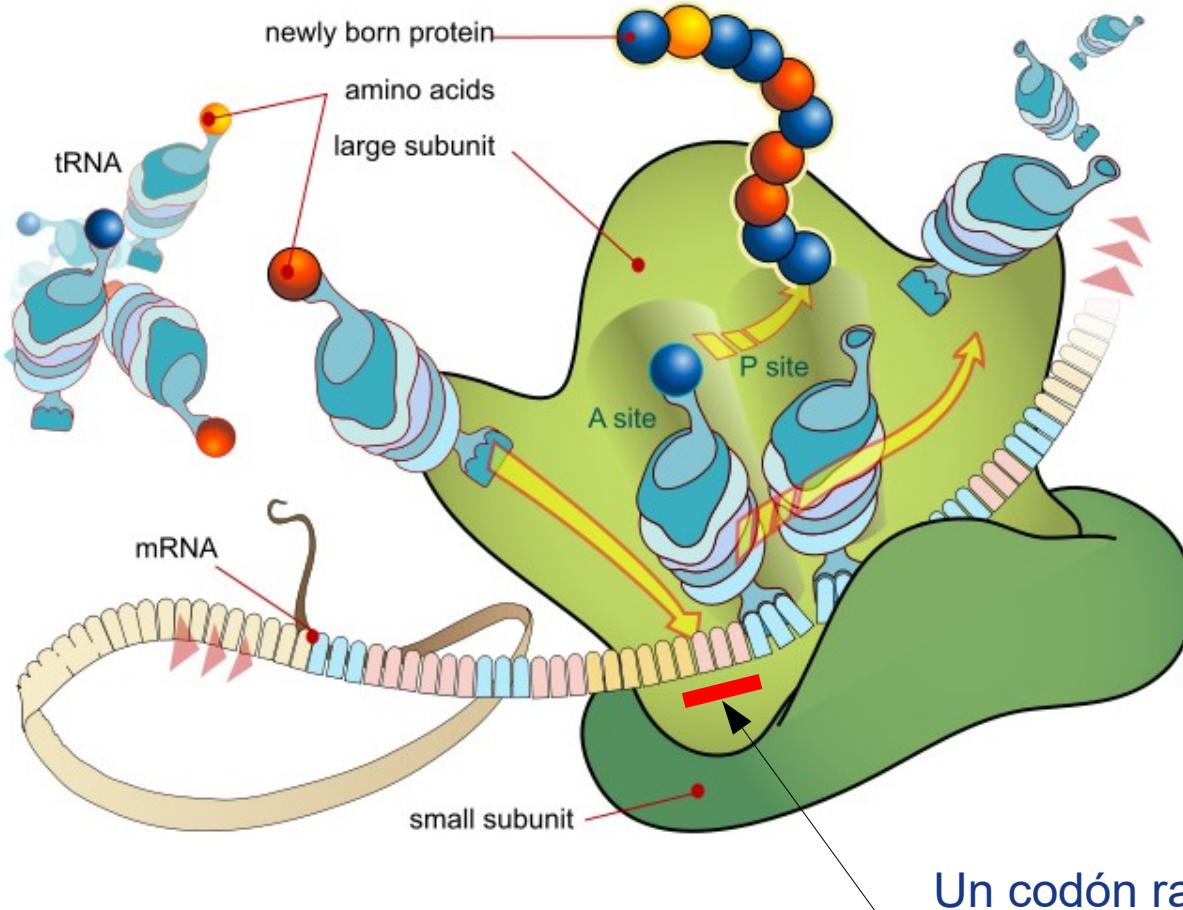
```
> fisher.test(matrix(c(22756,220,661,14),nrow=2),alternative="greater")
    Fisher's Exact Test for Count Data
data: matrix(c(22756, 220, 661, 14), nrow = 2)
p-value = 0.007597
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
1.29878      Inf
sample estimates:
odds ratio
2.190701
```



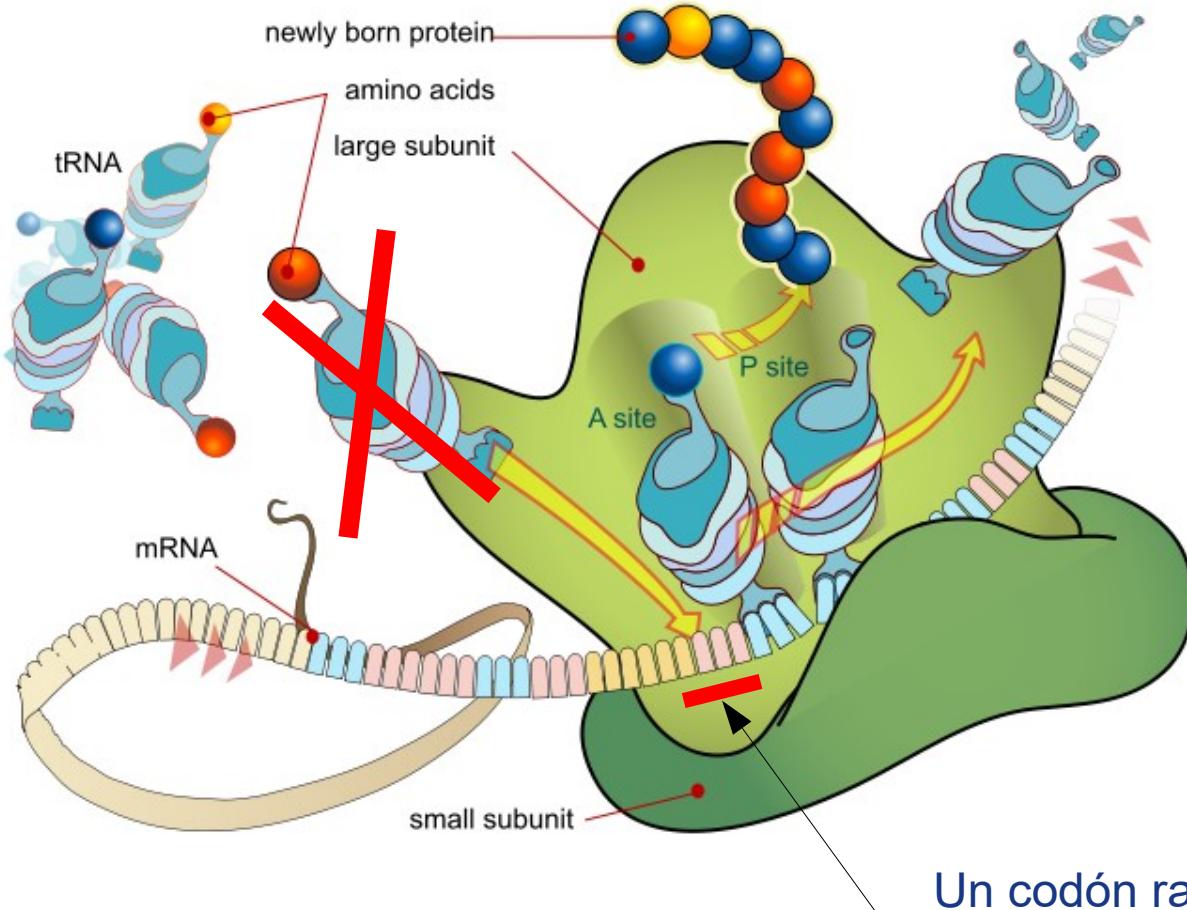
# Efectos del Sesgo en el Uso de Codones

- Los sesgos en el uso de codones en una secuencia codificante específica afecta la velocidad de traducción de la correspondiente proteína.
- La aparición de un **codon raro** en la secuencia codificante de una proteína puede producir un **estancamiento de los ribosomas en el correspondiente RNA**.

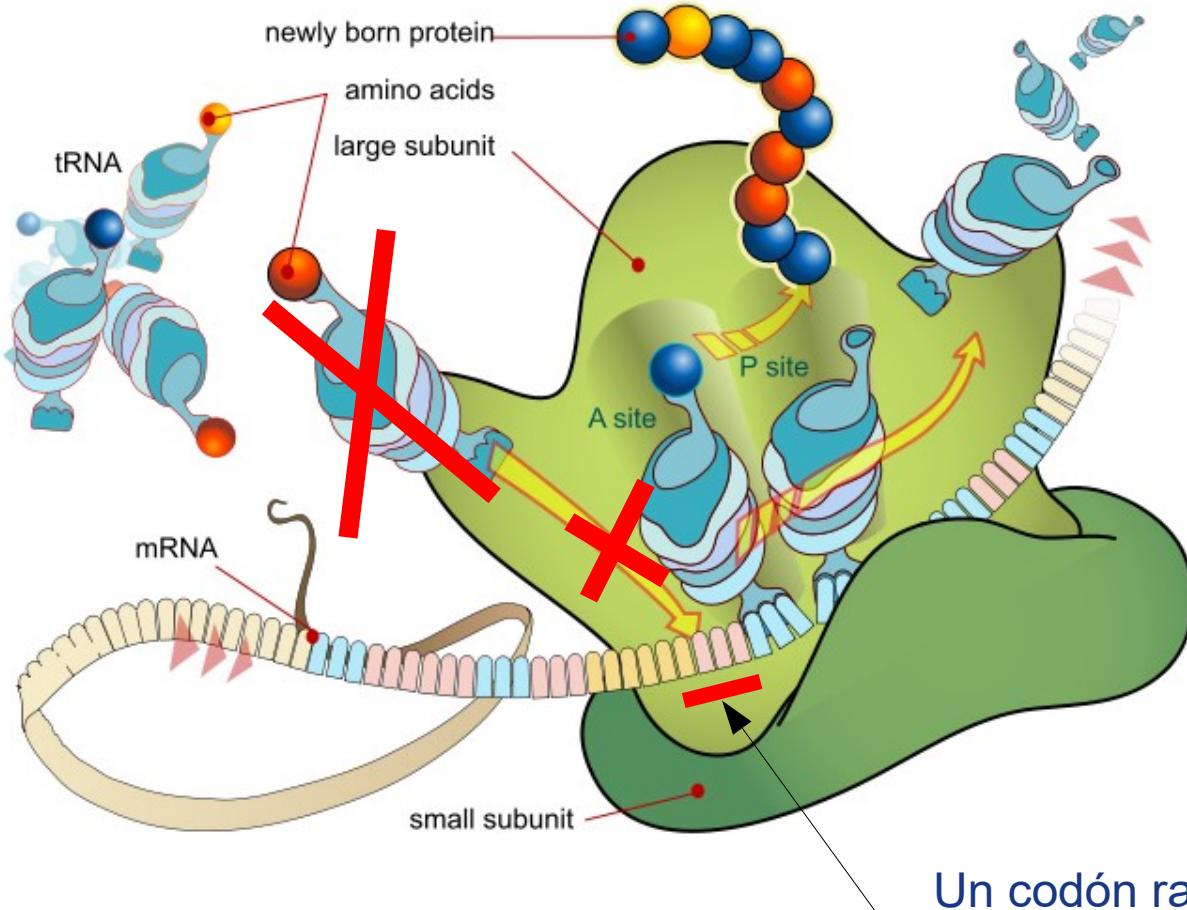




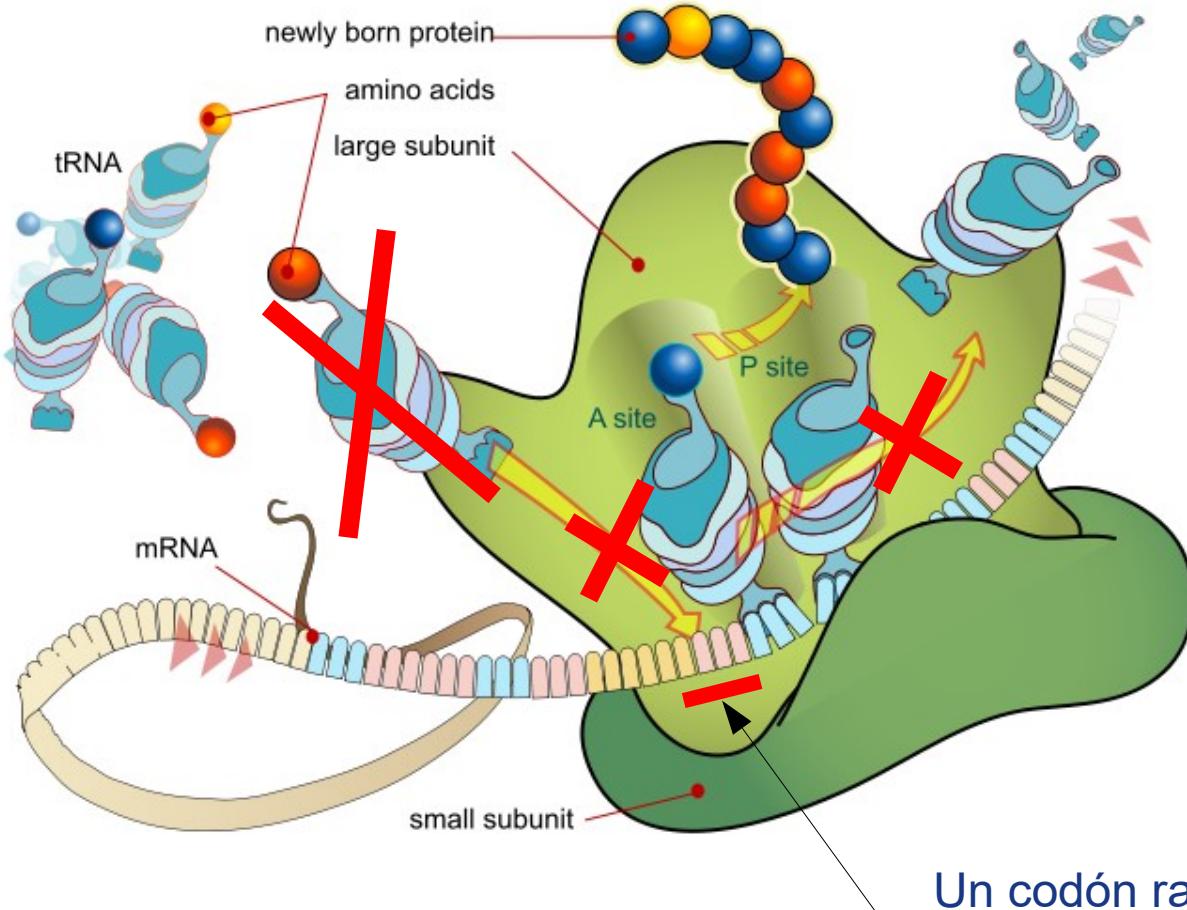
Un codón raro posee una  
muy baja copia del  
correspondiente tRNA



Un codón raro posee una muy baja copia del correspondiente tRNA



Un codón raro provoca un estancamiento en la progresión de los ribosomas durante la traducción



Un codón raro provoca un estancamiento en la progresión de los ribosomas durante la traducción

# Efectos del Sesgo en el Uso de Codones

- Regiones del genoma con CDS ricas en codones raros pueden indicar **eventos de transferencia horizontal de genes**.
- Secuencias codificantes ricas en codones raros producen una **excasa síntesis de proteínas**.
- Esta es una situación común en la **expresión heteróloga** de una proteína en un organismo con un uso de codones marcadamente diferente al organismo origen.



# Sesgos en el Uso de Codones

UUU F 0.57	UCU S 0.11	UAU Y 0.53	UGU C 0.42
UUC F 0.43	UCC S 0.11	UAC Y 0.47	UGC C 0.58
UUA L 0.15	UCA S 0.15	UAA * 0.64	UGA * 0.36
UUG L 0.12	UCG S 0.16	UAG * 0.00	UGG W 1.00
CUU L 0.12	CCU P 0.17	CAU H 0.55	CGU R 0.36
CUC L 0.10	CCC P 0.13	CAC H 0.45	CGC R 0.44
CUA L 0.05	CCA P 0.14	CAA Q 0.30	CGA R 0.07
CUG L 0.46	CCG P 0.55	CAG Q 0.70	CGG R 0.07
AUU I 0.58	ACU T 0.16	AAU N 0.47	AGU S 0.14
AUC I 0.35	ACC T 0.47	AAC N 0.53	AGC S 0.33
AUA I 0.07	ACA T 0.13	AAA K 0.73	AGA R 0.02
AUG M 1.00	ACG T 0.24	AAG K 0.27	AGG R 0.03
GUU V 0.25	GCU A 0.11	GAU D 0.65	GGU G 0.29
GUC V 0.18	GCC A 0.31	GAC D 0.35	GGC G 0.46
GUA V 0.17	GCA A 0.21	GAA E 0.70	GGA G 0.13
GUG V 0.40	GCG A 0.38	GAG E 0.30	GGG G 0.12

[Codon/a.a./fraction per codon per a.a.]

E. coli K12 data from the Codon Usage Database



# Sesgos en el Uso de Codones

UUU	F	0.46	UCU	S	0.19	UAU	Y	0.44	UGU	C	0.46
UUC	F	0.54	UCC	S	0.22	UAC	Y	0.56	UGC	C	0.54
UUA	L	0.08	UCA	S	0.15	UAA	*	0.30	UGA	*	0.47
UUG	L	0.13	UCG	S	0.05	UAG	*	0.24	UGG	W	1.00
CUU	L	0.13	CCU	P	0.29	CAU	H	0.42	CGU	R	0.08
CUC	L	0.20	CCC	P	0.32	CAC	H	0.58	CGC	R	0.18
CUA	L	0.07	CCA	P	0.28	CAA	Q	0.27	CGA	R	0.11
CUG	L	0.40	CCG	P	0.11	CAG	Q	0.73	CGG	R	0.20
AUU	I	0.36	ACU	T	0.25	AAU	N	0.47	AGU	S	0.15
AUC	I	0.47	ACC	T	0.36	AAC	N	0.53	AGC	S	0.24
AUA	I	0.17	ACA	T	0.28	AAA	K	0.43	AGA	R	0.21
AUG	M	1.00	ACG	T	0.11	AAG	K	0.57	AGG	R	0.21
GUU	V	0.18	GCU	A	0.27	GAU	D	0.46	GGU	G	0.16
GUC	V	0.24	GCC	A	0.40	GAC	D	0.54	GGC	G	0.34
GUA	V	0.12	GCA	A	0.23	GAA	E	0.42	GGA	G	0.25
GUG	V	0.46	GCG	A	0.11	GAG	E	0.58	GGG	G	0.25

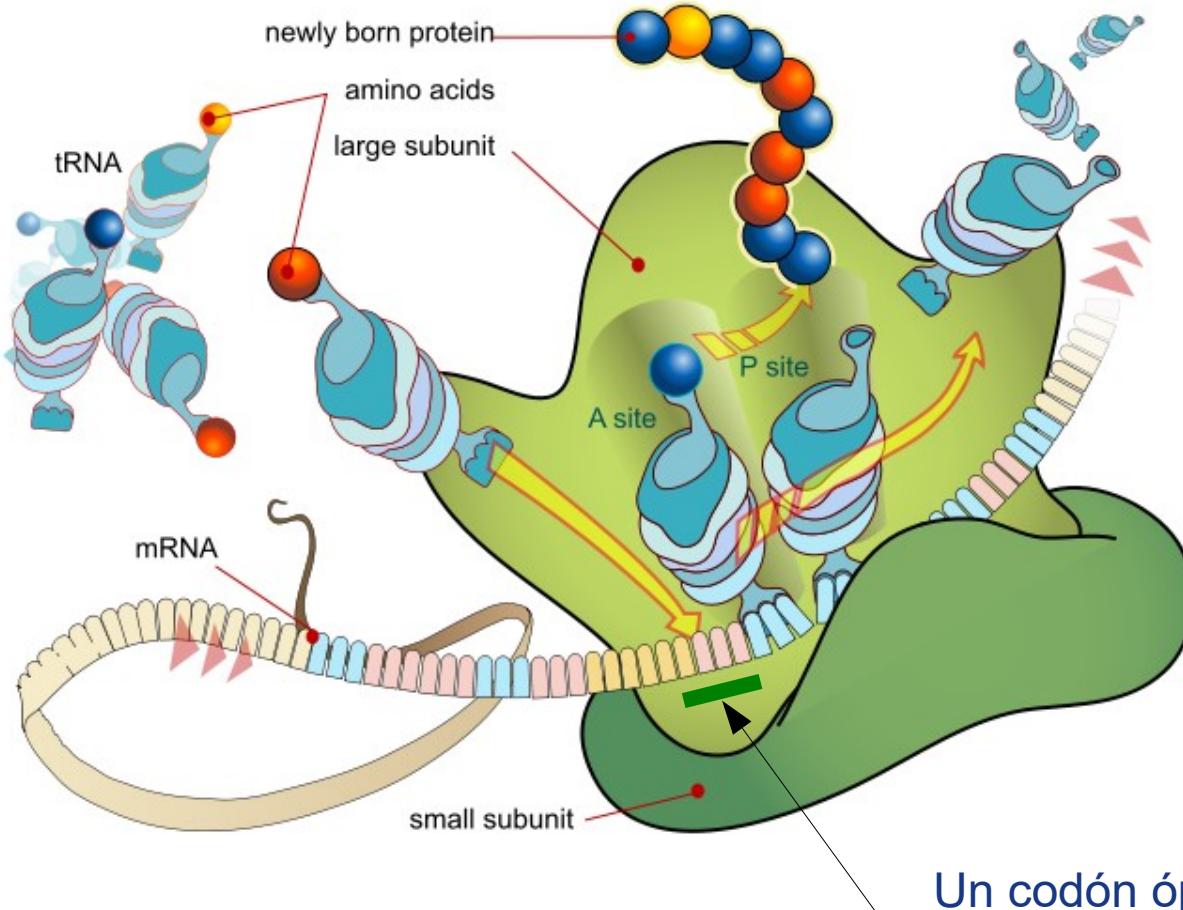
[Codon/a.a./fraction per codon per a.a.]

Homo sapiens data from the Codon Usage Database

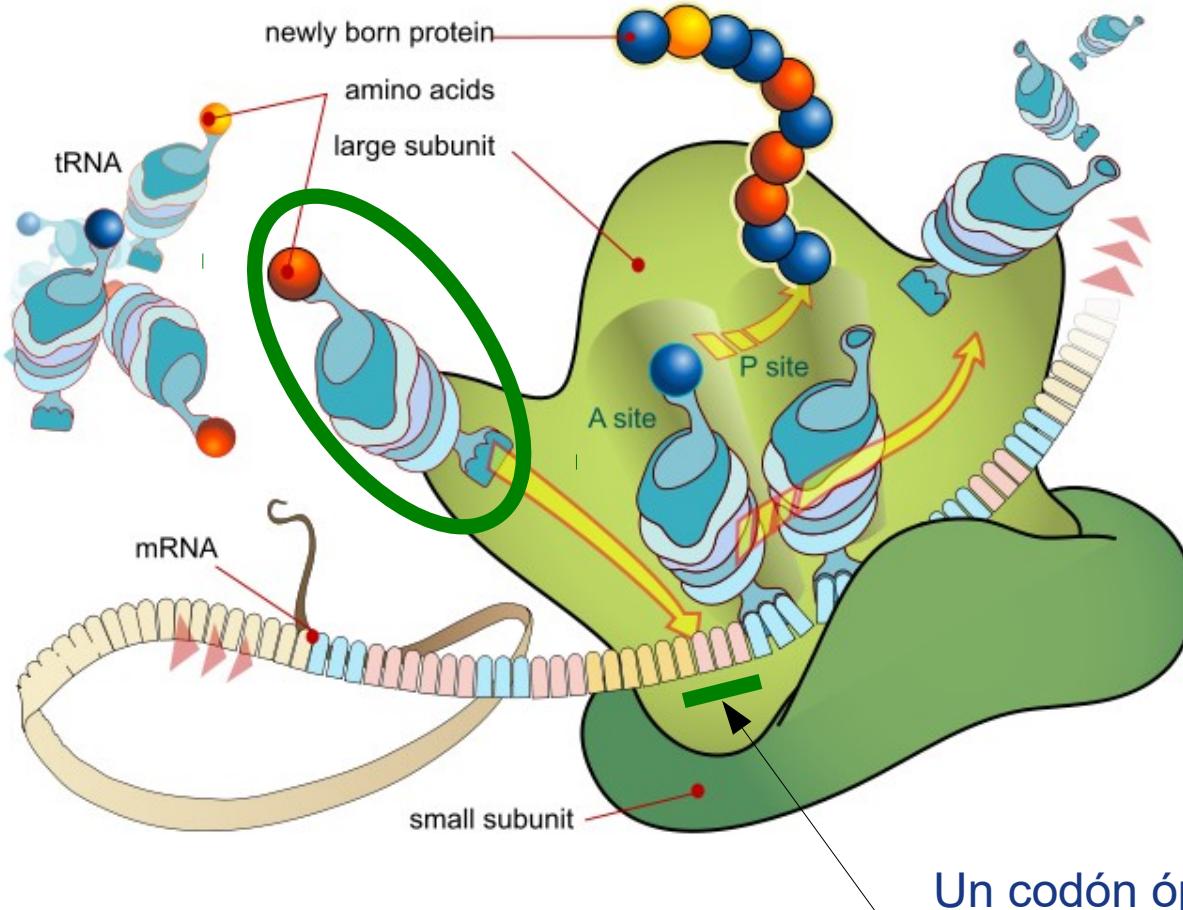
# Efectos del Sesgo en el Uso de Codones

- Los sesgos en el uso de codones en una secuencia codificante específica afecta la velocidad de traducción de la correspondiente proteína.
- La aparición de un **codon óptimo** en la secuencia codificante de una proteína puede **acelerar la traducción de la correspondiente proteína.**

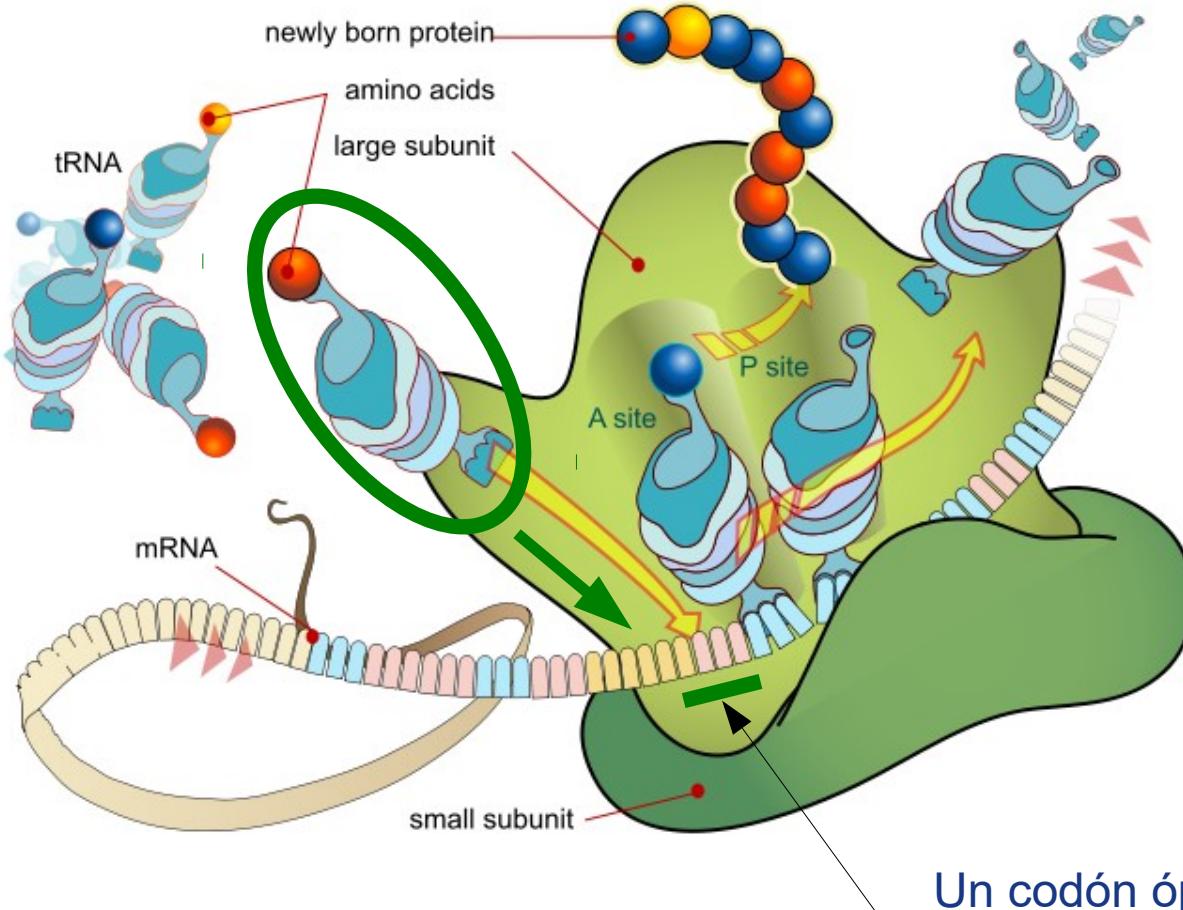




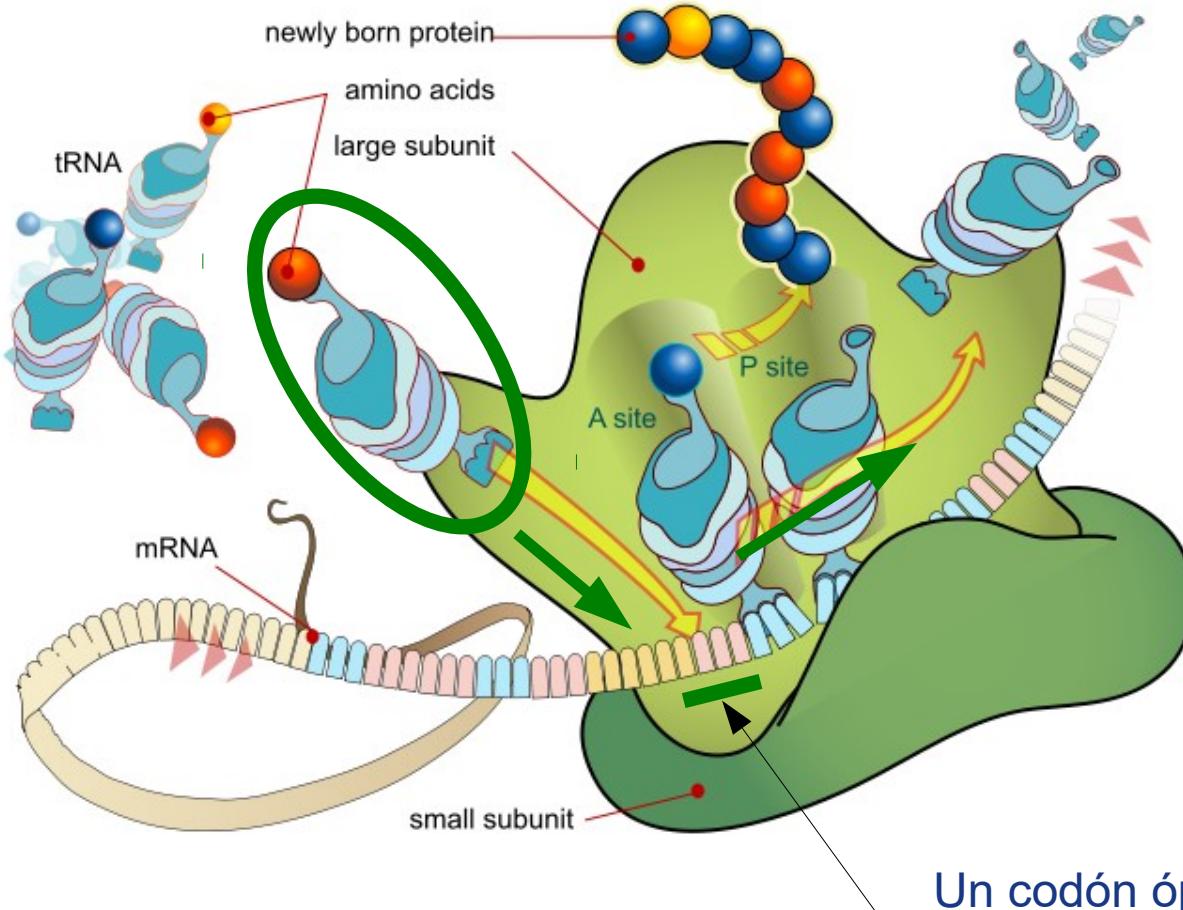
Un codón óptimo suele poseer una alta copia del correspondiente tRNA



Un codón óptimo suele poseer una alta copia del correspondiente tRNA



Un codón óptimo suele poseer una alta copia del correspondiente tRNA



Un codón óptimo suele poseer una alta copia del correspondiente tRNA

# Uso de Codones en la Optimización de la Expresión Heteróloga de Proteínas

- Uno de los campos de la biotecnología busca utilizar micro-organismos tales como *Escherichia coli* para la síntesis de enzimas de otros organismos que son de utilidad para el ser humano (i.e. la producción de insulina). Esto se conoce como **expresión heteróloga**.
- Si el uso de codones es marcadamente diferente entre el micro-organismo usado como destino y el organismo de donde se toma la secuencia de la enzima, la **síntesis suele ser subóptima**.
- La **optimización del uso de codones** consiste en generar una secuencia de DNA que codifique por la misma proteína pero que utilice codones óptimos para el genoma del micro-organismo destino.

# Estimación del Uso de Codones

- **Paso 1:** Utilizando la correspondiente base de datos obtener todas las secuencias de DNA codificantes por proteínas (CDS).

atgccgacagcaactggaaatggagctatttagctcaaagtgtactgcgaaaccgtcgaaggaaatt  
actactttccccctgattccatcaataaagaatattcaaagacagtaaacacccatacaactgctcttg  
gaaaggagaagcgagttattataccattgaagttgtatggacagcaaaaataaagatgctgctggtactac  
ccccataccaaagaaaaagccaaaaatataagaaggatattgctttggcgaggggtcaaagttgaac  
cttaa

...

atgatttcagtttgtataaaaactttcacgtcgtttagcatctccccaaacccatcctcgccatt  
tttgggctcctggtgcggttgtcattgattcccccttagtctcaacagtacaatcagaatggga  
tcaaccctaaaaactggttggtgtcaacgctgatgaaaatttcaattagctaatacctatcgatcaaa  
aatttgcctactctgattctttaatcagggtcaaattatccatcggtttaggatttcagggacgag  
atgatattttaaaaagttacattttgcaattaaatcagttagctcggtcagcctaa



# Estimación del Uso de Codones

- **Paso 2:** Dividir cada CDS según la pauta de lectura en codones.

atg ccg aca gca act tgg aat gga gct att tta gct caa agt gat cac tgc gaa acc gtc  
gaa gga aat tac tac ttt ccc cct gat tcc atc aat aaa gaa tat ttc aaa gac agt aac  
acc cat aca act tgc tct tgg aaa gga gaa gcg agt tat tat acc att gaa gtt gat gga  
cag caa aat aaa gat gct gct tgg tac tac ccc cat acc aaa gaa aaa gcc aaa aat  
ata gaa gga tat att gct ttt tgg cga ggg gtc aaa gtt gaa cct taa

...

atg att tca gtt agt gat aaa act ttt tct cac gtc gtt tta gca tct ccc caa ccc atc ctc gtc  
cat ttt tgg gct cct tgg tgc ggt ttg tgt cat ttg att ccc cct tta gtc tca aca gta caa tca  
gaa tgg gat caa ccc tta aaa ctg gtt ggt gtc aac gct gat gaa aat ttt caa tta gct aat  
acc tat cgt atc aaa aat ttg cct act ctg att ctt ttt aat cag ggt caa att atc cat cgt ttt  
gag gat ttt cag gga cga gat gat att tta aaa agt tta cat tct ttg caa tta aat cag tta  
gct cgt tca gcc taa . . . . . . . . . .

# Estimación del Uso de Codones

- **Paso 3:** Generar un vector de longitud 64 lleno de ceros cuyas componentes se nombren con los correspondientes codones. Este vector será un acumulador para las frecuencias absolutas de codones en las CDS (codon.freq.abs).

aaa	aac	aag	aat	aca	acc	acg	act	aga	agc	agg	agt	ata	atc	atg	att	caa	cac	cag
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cat	cca	ccc	ccg	cct	cga	cgc	cg <sup>g</sup>	cgt	cta	ctc	ctg	ctt	gaa	gac	gag	gat	gca	gcc
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gcg	gct	gga	ggc	ggg	ggt	gta	gtc	gtg	gtt	taa	tac	tag	tat	tca	tcc	tcg	tct	tgc
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tgg	tgt	tta	ttc	ttg	ttt													
0	0	0	0	0	0													

# Estimación del Uso de Codones

- **Paso 4.a:** Para cada CDS dividida en codones calcular la frecuencia absoluta de cada codon.

atg ccg aca gca act tgg aat gga gct att tta gct caa agt gat cac tgc gaa acc gtc  
gaa gga aat tac tac ttt ccc cct gat tcc atc aat aaa gaa tat ttc aaa gac agt aac  
acc cat aca act tgc tct tgg aaa gga gaa gcg agt tat tat acc att gaa gtt gat gga  
cag caa aat aaa gat gct gct tgg tac tac ccc cat acc aaa gaa aaa gcc aaa aat  
ata gaa gga tat att gct ttt tgg cga ggg gtc aaa gtt gaa cct taa



# Estimación del Uso de Codones

- **Paso 4.a:** Para cada CDS dividida en codones calcular la frecuencia absoluta de cada codon.

atg ccg aca gca act tgg aat gga gct att tta gct caa agt gat cac tgc gaa acc gtc  
gaa gga aat tac tac ttt ccc cct gat tcc atc aat **aaa** gaa tat ttc **aaa** gac agt aac  
acc cat aca act tgc tct tgg **aaa** gga gaa gcg agt tat tat acc att gaa gtt gat gga  
cag caa aat **aaa** gat gct gct tgg tac tac ccc cat acc **aaa** gaa **aaa** gcc **aaa** aat  
ata gaa gga tat att gct ttt tgg cga ggg gtc **aaa** gtt gaa cct taa

**aaa** aac aat aca acc act agt ata atc atg att caa cac cag cat ccc ccg cct cga gaa gac gat  
8 1 5 2 4 2 3 1 1 1 3 2 1 1 2 2 1 2 1 8 1 4  
gca gcc gcg gct gga ggg gtc gtt taa tac tat tcc tct tgc tgg tta ttc ttt  
1 1 1 5 5 1 2 2 1 4 4 1 1 2 4 1 1 2

# Estimación del Uso de Codones

- **Paso 4.a:** Para cada CDS dividida en codones calcular la frecuencia absoluta de cada codon.

atg ccg aca gca act tgg aat gga gct att tta gct caa agt gat cac tgc gaa acc gtc  
gaa gga aat tac tac ttt ccc cct gat tcc atc aat aaa gaa tat ttc aaa gac agt **aac**  
acc cat aca act tgc tct tgg aaa gga gaa gcg agt tat tat acc att gaa gtt gat gga  
cag caa aat aaa gat gct gct tgg tac tac ccc cat acc aaa gaa aaa gcc aaa aat  
ata gaa gga tat att gct ttt tgg cga ggg gtc aaa gtt gaa cct taa

aaa **aac** aat aca acc act agt ata atc atg att caa cac cag cat ccc ccg cct cga gaa gac gat  
1   **1**   5   2   4   2   3   1   1   1   3   2   1   1   2   2   1   2   1   8   1   4  
gca gcc gcg gct gga ggg gtc gtt taa tac tat tcc tct tgc tgg tta ttc ttt  
1   1   1   5   5   1   2   2   1   4   4   1   1   2   4   1   1   2



# Estimación del Uso de Codones

- **Paso 4.a:** Para cada CDS dividida en codones calcular la frecuencia absoluta de cada codon.

atg ccg aca gca act tgg **aat** gga gct att tta gct caa agt gat cac tgc gaa acc gtc  
gaa gga **aat** tac tac ttt ccc cct gat tcc atc **aat** aaa gaa tat ttc aaa gac agt aac  
acc cat aca act tgc tct tgg aaa gga gaa gcg agt tat tat acc att gaa gtt gat gga  
cag caa **aat** aaa gat gct gct tgg tac tac ccc cat acc aaa gaa aaa gcc aaa **aat**  
ata gaa gga tat att gct ttt tgg cga ggg gtc aaa gtt gaa cct taa

aaa aac **aat** aca acc act agt ata atc atg att caa cac cag cat ccc ccc cct cga gaa gac gat  
1 1 5 2 4 2 3 1 1 1 3 2 1 1 2 2 1 2 1 8 1 4  
gca gcc gcg gct gga ggg gtc gtt taa tac tat tcc tct tgc tgg tta ttc ttt  
1 1 1 5 1 2 2 1 4 4 1 1 2 4 1 1 2

# Estimación del Uso de Codones

- **Paso 4.a:** Para cada CDS dividida en codones calcular la frecuencia absoluta de cada codon.

atg ccg aca gca act tgg aat gga gct att tta gct caa agt gat cac tgc gaa acc gtc  
gaa gga aat tac tac ttt ccc cct gat tcc atc aat aaa gaa tat ttc aaa gac agt aac  
acc cat aca act tgc tct tgg aaa gga gaa gcg agt tat tat acc att gaa gtt gat gga  
cag caa aat aaa gat gct gct tgg tac tac ccc cat acc aaa gaa aaa gcc aaa aat  
ata gaa gga tat att gct ttt tgg cga ggg gtc aaa gtt gaa cct taa

aaa	aac	aat	aca	acc	act	agt	ata	atc	atg	att	caa	cac	cag	cat	ccc	ccg	cct	cga	gaa	gac	gat
1	1	5	2	4	2	3	1	1	1	3	2	1	1	2	2	1	2	1	8	1	4
gca	gcc	gcg	gct	gga	ggg	gtc	gtt	taa	tac	tat	tcc	tct	tgc	tgg	tta	ttc	ttt				
1	1	1	5	5	1	2	2	1	4	4	1	1	2	4	1	1	2				

# Estimación del Uso de Codones

- **Paso 4.a:** Para cada CDS dividida en codones calcular la frecuencia absoluta de cada codon.

atg ccg aca gca act tgg aat gga gct att tta gct caa agt gat cac tgc gaa acc gtc  
gaa gga aat tac tac ttt ccc cct gat tcc atc aat aaa gaa tat ttc aaa gac agt aac  
acc cat aca act tgc tct tgg aaa gga gaa gcg agt tat tat acc att gaa gtt gat gga  
cag caa aat aaa gat gct gct tgg tac tac ccc cat acc aaa gaa aaa gcc aaa aat  
ata gaa gga tat att gct ttt tgg cga ggg gtc aaa gtt gaa cct taa



aaa	aac	aat	aca	acc	act	agt	ata	atc	atg	att	caa	cac	cag	cat	ccc	ccg	cct	cga	gaa	gac	gat	
1	1	5	2	4	2	3	1	1	1	3	2	1	1	1	2	2	1	2	1	8	1	4
gca	gcc	gcg	gct	gga	ggg	gtc	gtt	taa	tac	tat	tcc	tct	tgc	tgg	tta	ttc	ttt					
1	1	1	5	5	1	2	2	1	4	4	1	1	2	4	1	1	2					

# Estimación del Uso de Codones

- **Paso 4.a:** Para cada CDS dividida en codones calcular la frecuencia absoluta de cada codon (cds.codon.freq.abs).

atg ccg aca gca act tgg aat gga gct att tta gct caa agt gat cac tgc gaa acc gtc  
gaa gga aat tac tac ttt ccc cct gat tcc atc aat aaa gaa tat ttc aaa gac agt aac  
acc cat aca act tgc tct tgg aaa gga gaa gcg agt tat tat acc att gaa gtt gat gga  
cag caa aat aaa gat gct gct tgg tac tac ccc cat acc aaa gaa aaa gcc aaa aat  
ata gaa gga tat att gct ttt tgg cga ggg gtc aaa gtt gaa cct taa



**table**

aaa	aac	aat	aca	acc	act	agt	ata	atc	atg	att	caa	cac	cag	cat	ccc	ccg	cct	cga	gaa	gac	gat	
1	1	5	2	4	2	3	1	1	1	3	2	1	1	1	2	2	1	2	1	8	1	4
gca	gcc	gcg	gct	gga	ggg	gtc	gtt	taa	tac	tat	tcc	tct	tgc	tgg	tta	ttc	ttt					
1	1	1	5	5	1	2	2	1	4	4	1	1	2	4	1	1	2					

# Estimación del Uso de Codones

- **Paso 4.b:** Para cada CDS actualizar el acumulador de frecuencias absolutas de codones (codon.freq.abs).

aaa	aac	aat	aca	acc	act	agt	ata	atc	atg	att	caa	cac	cag	cat	ccc	ccg	cct	cga	gaa	gac	gat	
1	1	5	2	4	2	3	1	1	1	3	2	1	1	1	2	2	1	2	1	8	1	4
gca	gcc	gcg	gct	gga	ggg	gtc	gtt	taa	tac	tat	tcc	tct	tgc	tgg	tta	ttc	ttt					
1	1	1	5	5	1	2	2	1	4	4	1	1	2	4	1	1	2					

# Estimación del Uso de Codones

- **Paso 4.b:** Para cada CDS actualizar el acumulador de frecuencias absolutas de codones (codon.freq.abs).

aaa aac aat aca acc act agt ata atc atg att caa cac cag cat ccc ccg cct cga gaa gac gat  
1 1 5 2 4 2 3 1 1 1 3 2 1 1 2 2 1 2 1 8 1 4

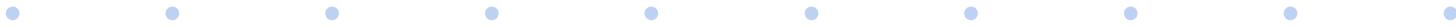
gca gcc gcg gct gga ggg gtc gtt taa tac tat tcc tct tgc tgg tta ttc ttt  
1 1 1 5 5 1 2 2 1 4 4 1 1 2 4 1 1 2



**names**

aaa aac aat aca acc act agt ata atc atg att caa cac cag cat ccc ccg cct cga gaa gac gat  
gca gcc gcg gct gga ggg gtc gtt taa tac tat tcc tct tgc tgg tta ttc ttt

**codons.in.cds**



# Estimación del Uso de Codones

- **Paso 4.b:** Para cada CDS actualizar el acumulador de frecuencias absolutas de codones (codon.freq.abs).

aaa aac aat aca acc act agt ata atc atg att caa cac cag cat ccc ccg cct cga gaa gac gat  
1 1 5 2 4 2 3 1 1 3 2 1 1 2 2 1 2 1 8 1 4

gca gcc gcg gct gga ggg gtc gtt taa tac tat tcc tct tgc tgg tta ttc ttt  
1 1 1 5 5 1 2 2 1 4 4 1 1 2 4 1 1 2



**names**

aaa aac aat aca acc act agt ata atc atg att caa cac cag cat ccc ccg cct cga gaa gac gat  
gca gcc gcg gct gga ggg gtc gtt taa tac tat tcc tct tgc tgg tta ttc ttt

**codons.in.cds**

`codon.freq.abs[codons.in.cds] ← codon.freq.abs[codons.in.cds] + cds.codon.freq.abs`



# Estimación del Uso de Codones

- **Paso 5:** Para cada aminoácido extraer las frecuencias absolutas de los codones sinónimos y almacenar este vector en la componente de una lista.

Lys	Asn	Thr	Arg	Ser	Ile	Met	Gln	His	Pro	Leu
Glt	Asp	Ala	Gly	Val	Tyr	Cys	Trp	Phe	STOP	

aaa	aac	aag	aat	aca	acc	acg	act	aga	agc	agg	agt	ata	atc	atg	att	
55546	16970	13637	43638	14454	28045	11120	21619	10314	8575	3441	23202	9761	22506	24521	65668	
caa	cac	cag	cat	cca	ccc	ccg	cct	cga	cgc	cgg	cgt	cta	ctc	ctg	ctt	
58860	7169	14559	17634	9611	22571	7023	23567	11261	12711	8195	16350	16829	15953	11607	16650	
gaa	gac	gag	gat	gca	gcc	gcg	gct	gga	ggc	ggg	ggt	gta	gtc	gtg	gtt	
69746	14797	16343	49882	19586	25415	14377	31849	35195	12192	19641	18698	13574	18186	14751	34257	
taa	tac	tag	tat	tca	tcc	tcg	tct	tga	tgc	tgg	tgt	tta	ttc	ttg	ttt	
2699	9347	972	33627	12377	11311	6396	21940		649	2836	19146	10647	69209	9145	19568	43797



# Estimación del Uso de Codones

- **Paso 5.a:** Para cada aminoácido extraer las frecuencias absolutas de los codones sinónimos y almacenar este vector en la componente de una lista (syn.codon.rel.freq).

Lys	Asn	Thr	Arg	Ser	Ile	Met	Gln	His	Pro	Leu					
Glt	Asp	Ala	Gly	Val	Tyr	Cys	Trp	Phe		STOP					
aaa	aac	aag	aat	aca	acc	acg	act	aga	agc	agg	agt	ata	atc	atg	att
55546	16970	3637	3638	14454	28045	11120	21619	10314	8575	3441	23202	9761	22506	24521	65668
caa	cac	cag	cat	cca	ccc	ccg	cct	cga	cgc	cg	cgt	cta	ctc	ctg	ctt
58860	7169	14559	17634	9611	22571	7023	23567	11261	12711	8195	16350	16829	15953	11607	16650
gaa	gac	gag	gat	gca	gcc	gcg	gct	gga	ggc	ggg	ggt	gta	gtc	gtg	gtt
69746	14797	16343	49882	19586	25415	14377	31849	35195	12192	19641	18698	13574	18186	14751	34257
taa	tac	tag	tat	tca	tcc	tcg	tct	tga	tgc	tgg	tgt	tta	ttc	ttg	ttt
2699	9347	972	33627	12377	11311	6396	21940	649	2836	19146	10647	69209	9145	19568	43797

# Estimación del Uso de Codones

- **Paso 5.a:** Para cada aminoácido extraer las frecuencias absolutas de los codones sinónimos y almacenar este vector en la componente de una lista (syn.codon.rel.freq).

Lys	Asn	Thr	Arg	Ser	Ile	Met	Gln	His	Pro	Leu					
Glt	Asp	Ala	Gly	Val	Tyr	Cys	Trp	Phe		STOP					
aaa	aac	aag	aat	aca	acc	acg	act	aga	agc	agg	agt	ata	atc	atg	att
55546	16970	3637	3638	14454	28045	11120	21619	10314	8575	3441	23202	9761	22506	24521	65668
caa	cac	cag	cat	cca	ccc	ccg	cct	cga	cgc	cg	cgt	cta	ctc	ctg	ctt
58860	7169	14559	17634	9611	22571	7023	23567	11261	12711	8195	16350	16829	15953	11607	16650
gaa	gac	gag	gat	gca	gcc	gcg	gct	gga	ggc	ggg	ggt	gta	gtc	gtg	gtt
69746	14797	16343	49882	19586	25415	14377	31849	35195	12192	19641	18698	13574	18186	14751	34257
taa	tac	tag	tat	tca	tcc	tcg	tct	tga	tgc	tgg	tgt	tta	ttc	ttg	ttt
2699	9347	972	33627	12377	11311	6396	21940	649	2836	19146	10647	69209	9145	19568	43797

Lys  
aaa aag  
55546 13637



# Estimación del Uso de Codones

- **Paso 5.a:** Para cada aminoácido extraer las frecuencias absolutas de los codones sinónimos y almacenar este vector en la componente de una lista (syn.codon.rel.freq).

Lys	Asn	Thr	Arg	Ser	Ile	Met	Gln	His	Pro	Leu
Glt	Asp	Ala	Gly	Val	Tyr	Cys	Trp	Phe		STOP

aaa	aac	aag	aat	aca	acc	acg	act	aga	agc	agg	agt	ata	atc	atg	att
55546	16970	13637	43638	14454	28045	11120	21619	10314	8575	3441	23202	9761	22506	24521	65668
caa	cac	cag	cac	cca	ccc	ccg	cct	cga	cgc	cgg	cgt	cta	ctc	ctg	ctt
58860	7169	14559	17634	9611	22571	7023	23567	11261	12711	8195	16350	16829	15953	11607	16650
gaa	gac	gag	gat	gca	gcc	gcg	gct	gga	ggc	ggg	ggt	gta	gtc	gtg	gtt
69746	14797	16343	49882	19586	25415	14377	31849	35195	12192	19641	18698	13574	18186	14751	34257
taa	tac	tag	tat	tca	tcc	tcg	tct	tga	tgc	tgg	tgt	tta	ttc	ttg	ttt
2699	9347	972	33627	12377	11311	6396	21940	649	2836	19146	10647	69209	9145	19568	43797



# Estimación del Uso de Codones

- **Paso 5.a:** Para cada aminoácido extraer las frecuencias absolutas de los codones sinónimos y almacenar este vector en la componente de una lista (syn.codon.rel.freq).

Lys	Asn	Thr	Arg	Ser	Ile	Met	Gln	His	Pro	Leu
Glt	Asp	Ala	Gly	Val	Tyr	Cys	Trp	Phe	STOP	

aaa	aac	aag	aat	aca	acc	acg	act	aga	agc	agg	agt	ata	atc	atg	att
55546	16970	13637	43638	14454	28045	11120	21619	10314	8575	3441	23202	9761	22506	24521	65668
caa	cac	cag	cat	cca	ccc	ccg	ccc	cga	cgc	cg	cgt	cta	ctc	ctg	ctt
58860	7169	14559	17634	9611	22571	7023	23567	11261	12711	8195	16350	16829	15953	11607	16650
gaa	gac	gag	gat	gca	gcc	gcf	gct	gga	ggc	ggg	ggt	gta	gtc	gtg	gtt
69746	14797	16343	49882	19586	25415	14377	31849	35195	12192	19641	18698	13574	18186	14751	34257
taa	tac	tag	tat	tca	tcc	tcg	tct	tga	tgc	tgg	tgt	tta	ttc	ttg	ttt
2699	9347	972	33627	12377	11311	6396	21940	649	2836	19146	10647	69209	9145	19568	43797

Lys	Asn	Thr	...	...	...										
aaa	aag	aac	aat	aca	acc	acg	act	...	...	...	...	...	...	...	...
55546	13637	16970	43638	14454	28045	11120	21619	...	...	...	...	...	...	...	...

# Estimación del Uso de Codones

- **Paso 5.b:** Normalizar cada componente (vector) de la lista syn.codon.rel.freq dividiendo por la suma de sus elementos.

Lys	Asn	Thr	...
aaa 55546	aag 13637	aac 16970	aat 43638

# Estimación del Uso de Codones

- **Paso 5.b:** Normalizar cada componente (vector) de la lista syn.codon.rel.freq dividiendo por la suma de sus elementos. Devolver syn.codon.rel.freq como el uso de codones.

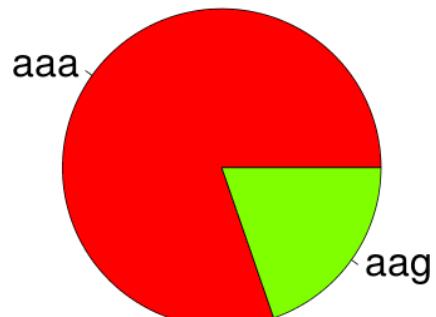
Lys  
aaa aag  
0.80 0.20

Asn  
aac aat  
0.28 0.72

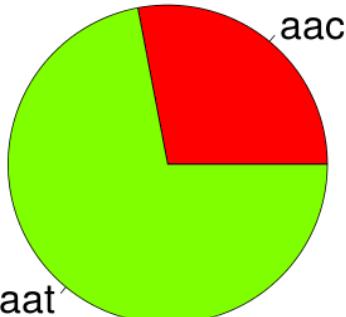
Thr  
aca acc acg act  
0.19 0.37 0.15 0.29

...

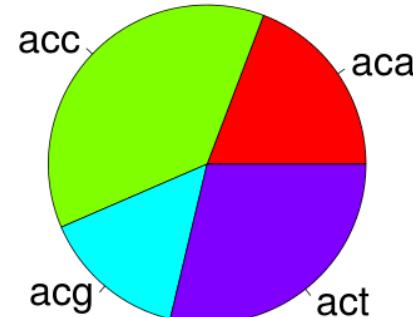
**Lys**



**Asn**



**Thr**



# Optimización del Uso de Codones

- **Entrada:**

- La CDS de una proteína a expresar heterólogamente (prot.cds).
- El uso de codones del organismo a utilizar para la síntesis como una lista de vectores (codon.usage). Cada elemento de esta lista está nombrado con un aminoácido y el correspondiente vector consiste en el uso de los codones que codifican el aminoácido.

```
>PEDF
ATGCAGGCCCTGGTGCTACTCCTCTGCATTGGAGCCCTCCTCGGGCACAGCAGCTGCCAGAACCTGCCA
GCCCCCCGGAGGAGGGCTCCCCAGACCCCCGACAGCACAGGGCGCTGGTGGAGGAGGAGGATCCTTCCTT
...
GCCCAGGGCTGCAGCCTGCCAACCTCACCTTCCCCTGGACTATCACCTAACCGCCTTCATCTTCGT
ACTGAGGGACACAGACACAGGGGCCCTCTTCATTGGCAAGATTCTGGACCCCAGGGGCCCTAA
```

Ala	Cys	Gln	Let	Met
gca 0.40	gcc 0.13	gcg 0.21	gct 0.26	tgc 0.34
				tgt 0.66



# Optimización del Uso de Codones

- **Paso 1:**

- Dividir en codones prot.cds.
- Traducir a aminoácidos y definir prot.

ATG CAG GCC CTG GTG CTA CTC CTC TGC ATT GGA GCC CTC CTC GGG CAC AGC AGC TGC CAG AAC CCT

...

GAG GGA CAC AGA CAC AGGG GCC CTT CTC TTC ATT GGC AAG ATT CTG GAC CCC AGG GGC CCC TAA

Met Gln Ala Let Val Let Let Let Cys Ile Gly

Ala				Cys		Gln		Let				Met		
gca	gcc	gcg	gct	tgc	tgt	caa	cag	cat	ctc	ctg	ctt	tta	ttg	atg
0.40	0.13	0.21	0.26	0.34	0.66	0.84	0.16	0.06	0.05	0.04	0.19	0.48	0.17	1



# Optimización del Uso de Codones

- **Paso 2:**
  - Para cada aminoácido usar codon.usage para determinar el codón óptimo.
  - Almacenar el codón óptimo en prot.cds.opt.

ATG CAG GCC CTG GTG CTA CTC CTC TGC ATT GGA GCC CTC CTC GGG CAC AGC AGC TGC CAG AAC CCT

...

GAG GGA CAC AGA CAC AGGG GCC CTT CTC TTC ATT GGC AAG ATT CTG GAC CCC AGG GGC CCC TAA

Met Gln Ala Let Val Let Let Let Cys Ile Gly

Ala				Cys		Gln		Let				Met		
gca	gcc	gcg	gct	tgc	tgt	caa	cag	cat	ctc	ctg	ctt	tta	ttg	atg
0.40	0.13	0.21	0.26	0.34	0.66	0.84	0.16	0.06	0.05	0.04	0.19	0.48	0.17	1

# Optimización del Uso de Codones

- **Paso 3:**
  - Convertir los codones de prot.cds.opt a secuencia dna.seq.opt.
  - Devolver dna.seq.opt

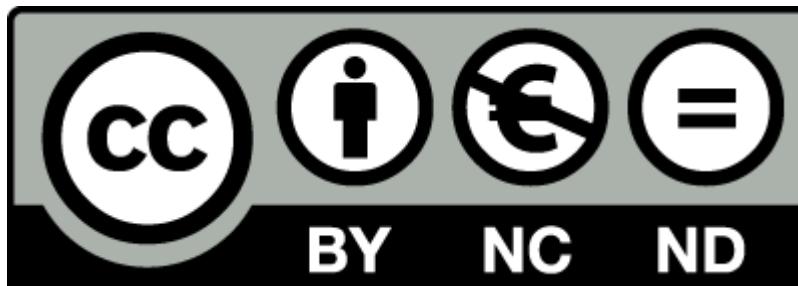
ATG CAG GCC CTG GTG CTA CTC CTC TGC ATT GGA GCC CTC CTC GGG CAC AGC AGC TGC CAG AAC CCT

...  
GAG GGA CAC AGA CAC AGGG GCC CTT CTC TTC ATT GGC AAG ATT CTG GAC CCC AGG GGC CCC TAA

Met Gln Ala Let Val Let Let Let Cys Ile Gly

Ala				Cys		Gln		Let				Met	
gca	gcc	gcg	gct	tgc	tgt	caa	cag	cat	ctc	ctg	ctt	tta	ttg
0.40	0.13	0.21	0.26	0.34	0.66	0.84	0.16	0.06	0.05	0.04	0.19	0.48	0.17





This work is licensed under the Creative Commons Attribution-NonCommercial NoDerivs 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>.

Parte de estas transparencias están basadas en los apuntes para la asignatura Informática Aplicada a la Bioquímica desarrollados por Francisco J. Romero-Campero e Ignacio Pérez Hurtado de Mendoza.

