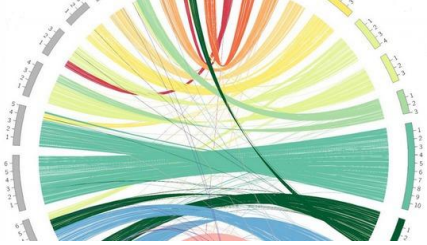
UNIVERSIDAD DE LA REPUBLICA

FACULTAD DE CIENCIAS
UDELAR | fcien.edu.uy



Curso Genomica Evolutiva

Edición 2020

Dot Plots

Viernes 26 Junio 2020

Fernando Alvarez

Se comparan dos secuencias: A (de longitud = *m*) y B (de longitud = *n*).

2.- Se escribe la secuencia A en la fila superior y la secuencia B (longitud = *n*) en la columna de la izquierda para construir una matriz con *m* columnas y *n* filas (*m* × *n*).

3.- Se compara cada letra de la secuencia A con cada letra de la secuencia B. Si coinciden los caracteres se marca esa posición con un punto. Si no, se deja en blanco.

Dot Plots

	A	A	T	C	G	T	T	C	G	A
T										
A										
T										
C										
G										
C										
G										
A										

Dot Plots

	A	A	T	C	G	T	T	C	G	A
T			•			•	•			
A	•	•								•
T			•			•	•			
C				•				•		
G					•				•	
C				•				•		
G					•				•	
A	•	•								•

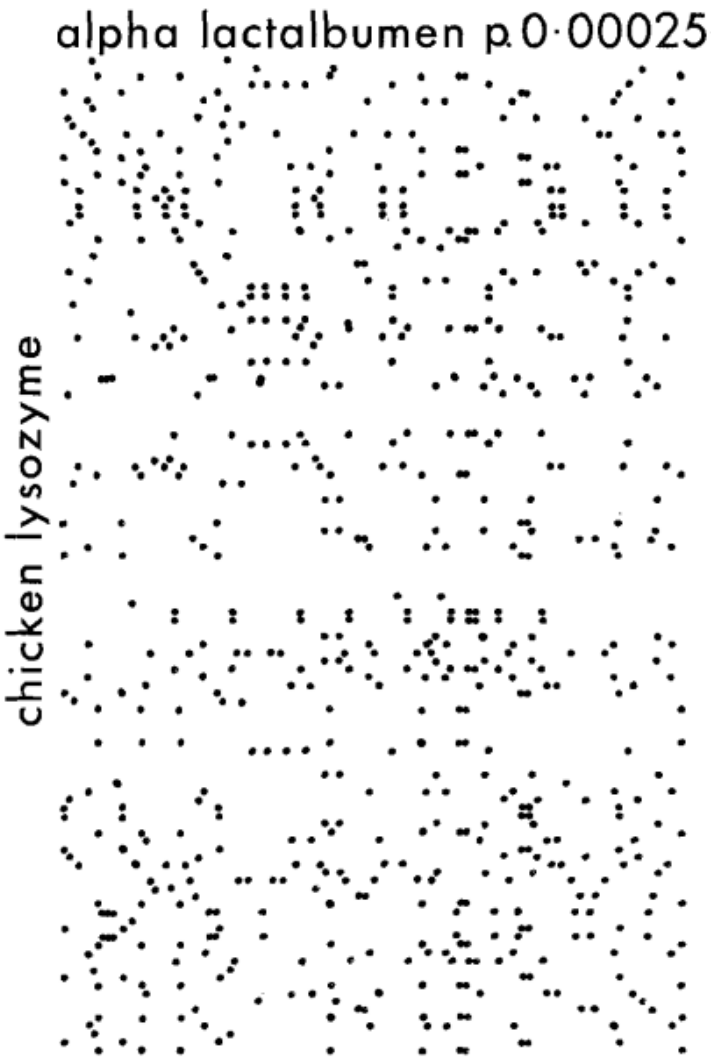


Fig.4. *The diagram obtained from a comparison of chicken lysozyme (left margin, N terminus at top) and bovine alpha lactalbumen (upper margin, N terminus at left end)*

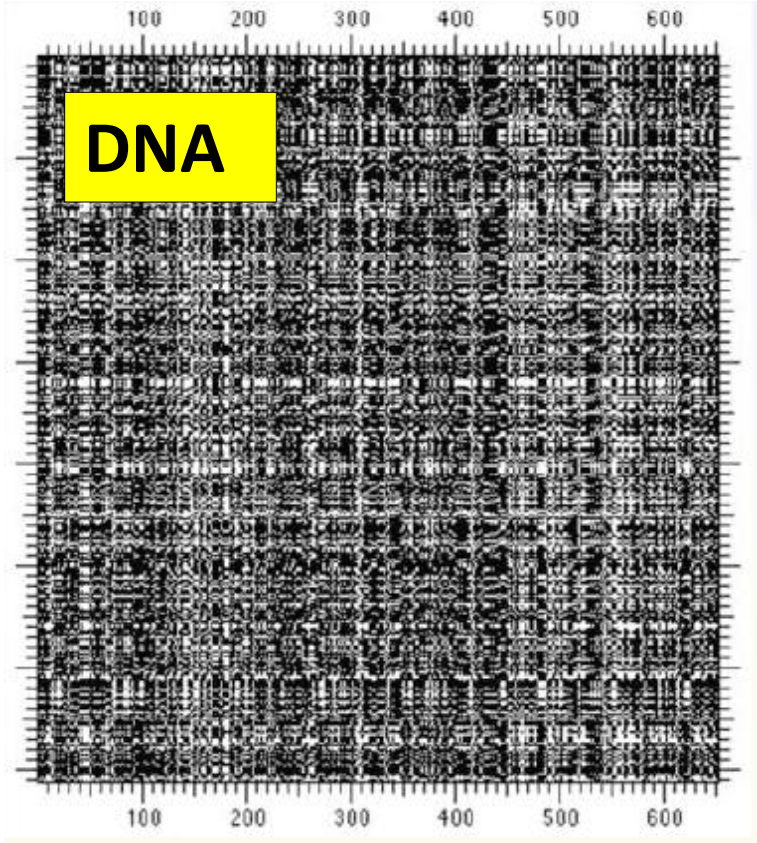
Es un método visual que detecta todas las coincidencias posibles entre dos secuencias. Es tarea del investigador determinar cuáles son relevantes.

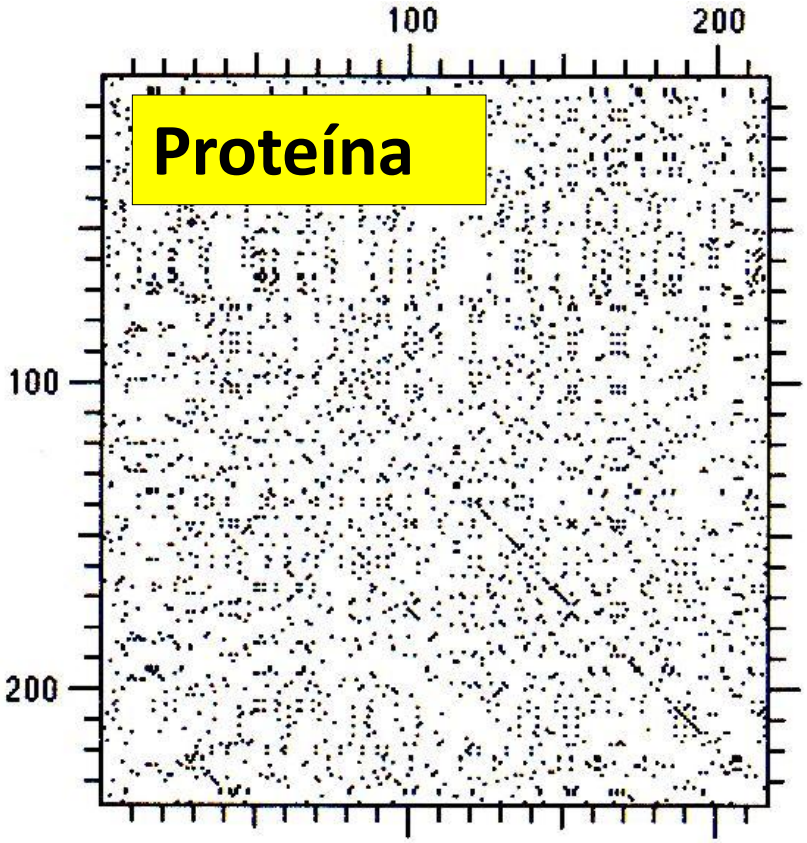
No proporciona un alineamiento de las secuencias pero nos da una idea de qué regiones deberían estar alineadas después de utilizar cualquiera de los otros métodos y nos puede ayudar a decidir cuál es el alineamiento óptimo.

Detecta relaciones entre las secuencias, o dentro de una misma secuencia que, de otra forma, serían muy difíciles de encontrar

Secuencia horizontal: gen/proteína *c2* del fago P22

Secuencia vertical: gen/proteína *cI* del fago λ



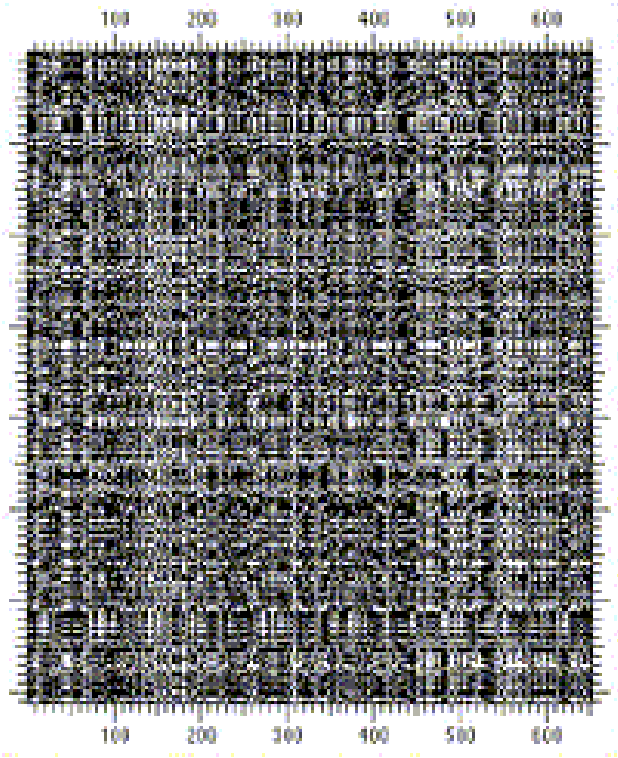


Como sólo hay 4 nucleótidos, aparecen muchas coincidencias por mero azar que generan ruido

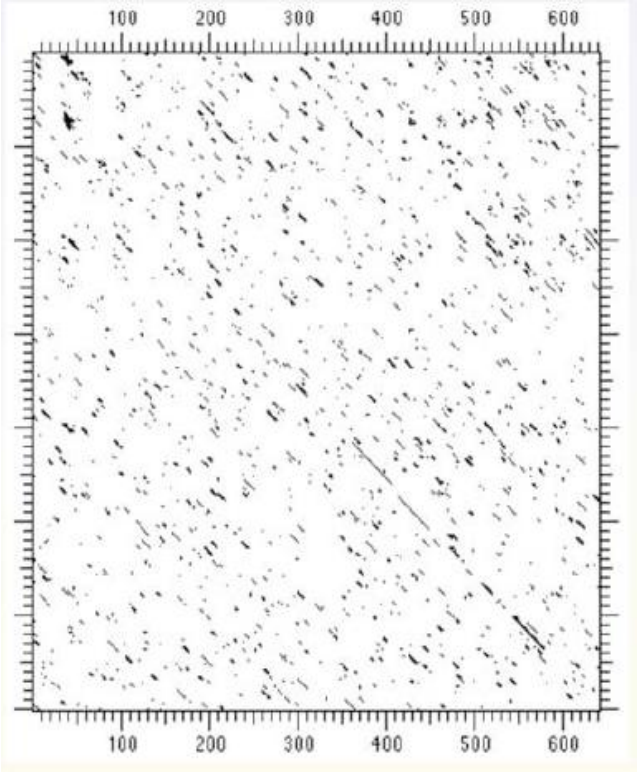
Como hay 20 aminoácidos, hay muchas menos coincidencias por azar y presenta mucho menos ruido

Genera mucho ruido, pero este se puede filtrar

- Secuencia horizontal: gen *c2* del fago P22
- Secuencia vertical: gen *cI* del fago λ



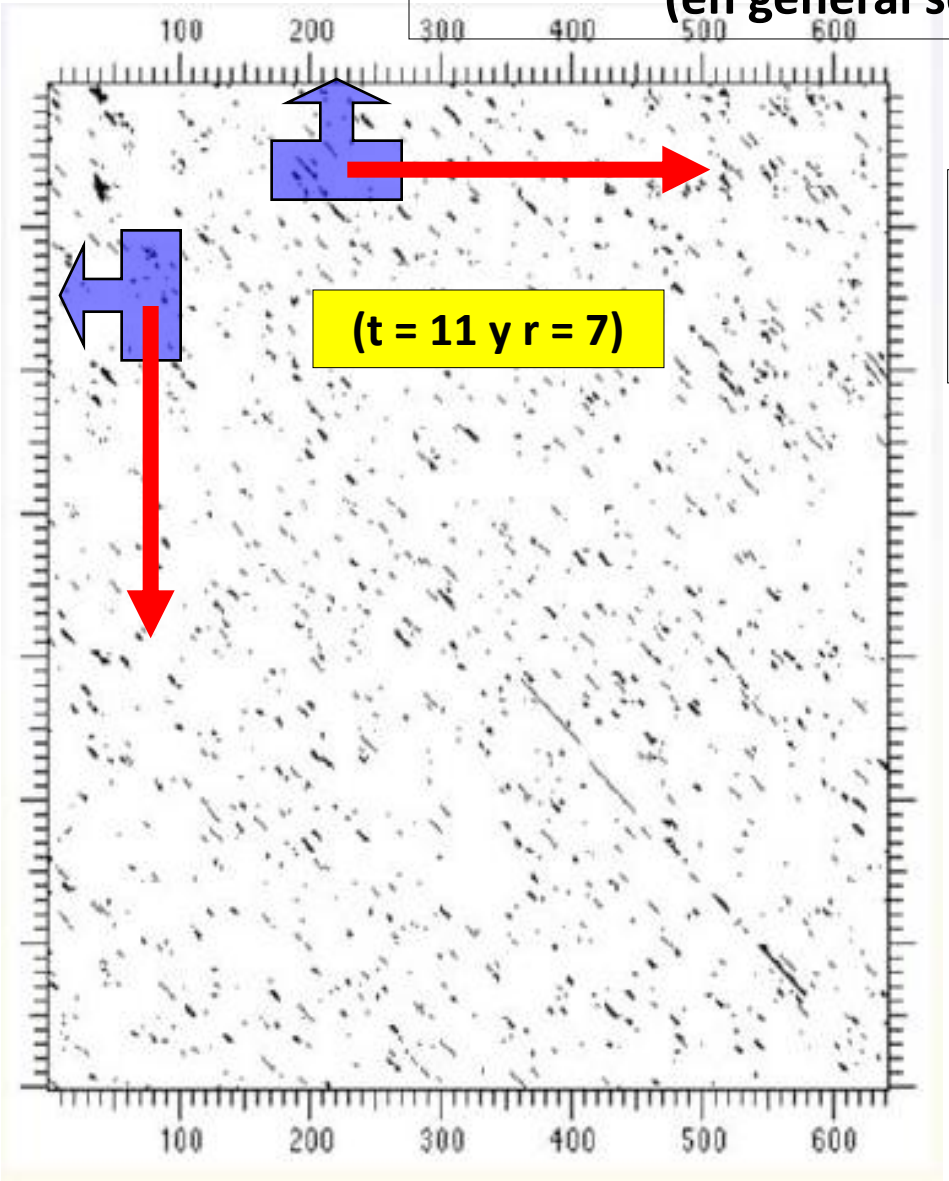
Sin filtrar



Aplicando un filtro

Filtradros, se comparan K-mers consecutivos

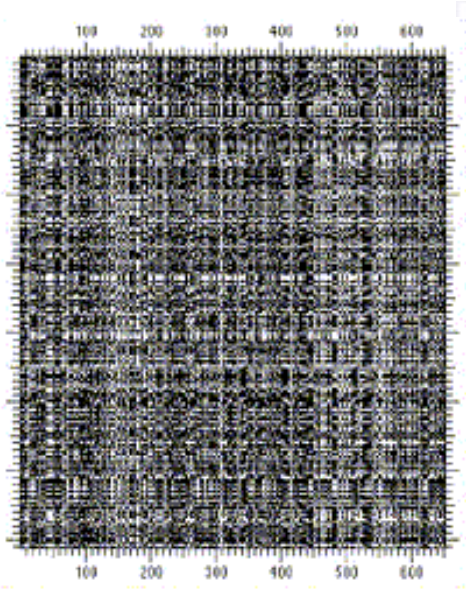
(en general solapantes o no)

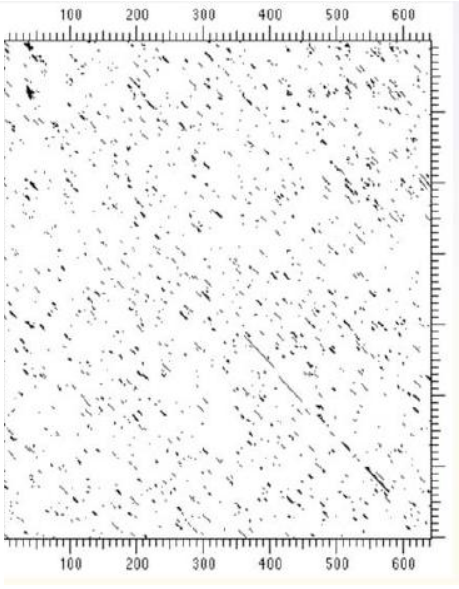


- TAMAÑO (t): es el número de letras del K-mer. Suele ser 15 en el caso del DNA y 2 ó 3 en el caso de proteínas.

- RIGOR (r): es el número mínimo de coincidencias que debe haber entre las dos K-mers para colocar un punto en la matriz

	A	A	T	C	G	T	T	C	G	A
T			•			•	•			
A	•	•								•
T			•			•	•			
C				•				•		
G					•				•	
C				•				•		
G					•				•	
A	•	•								•

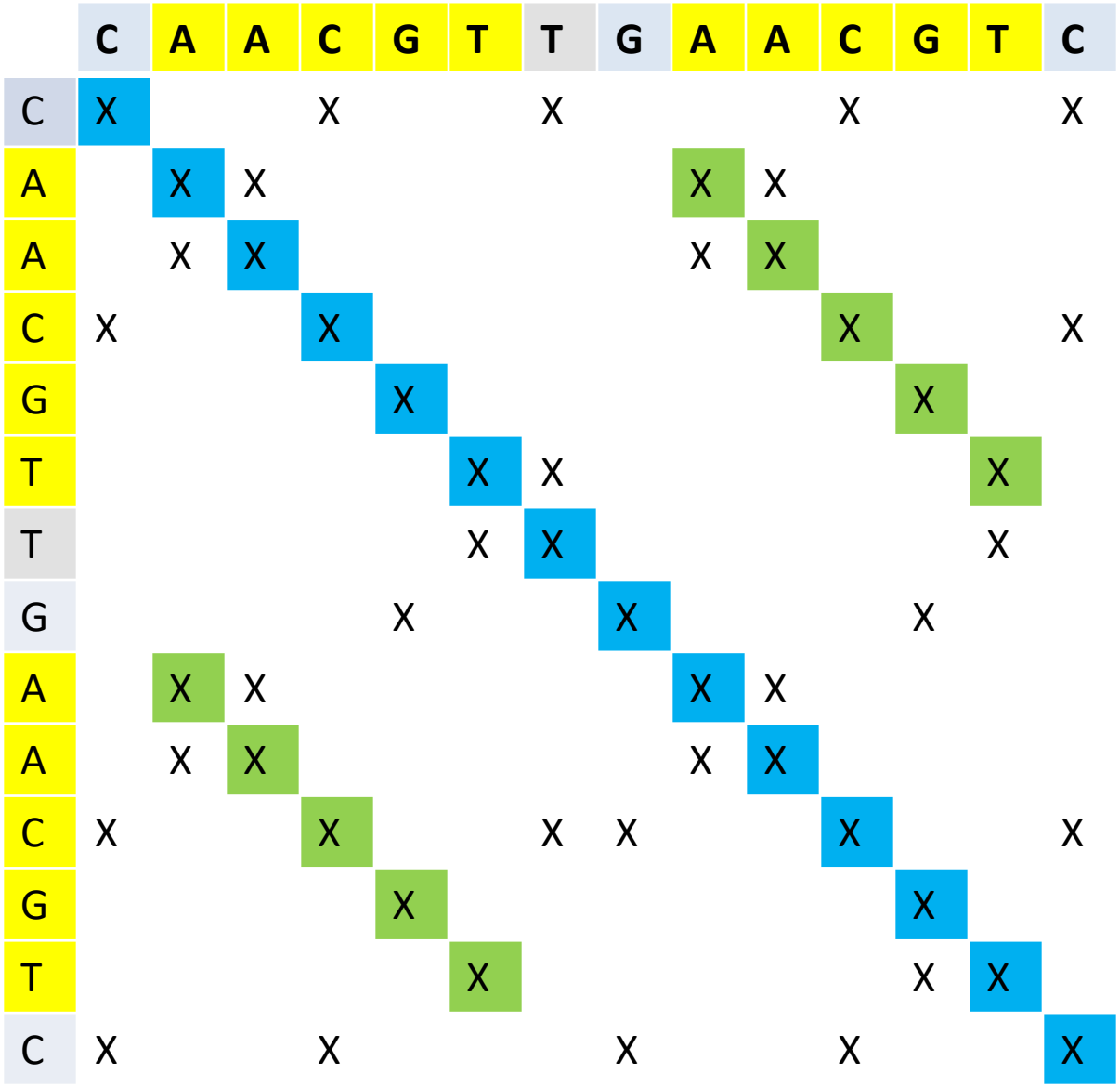


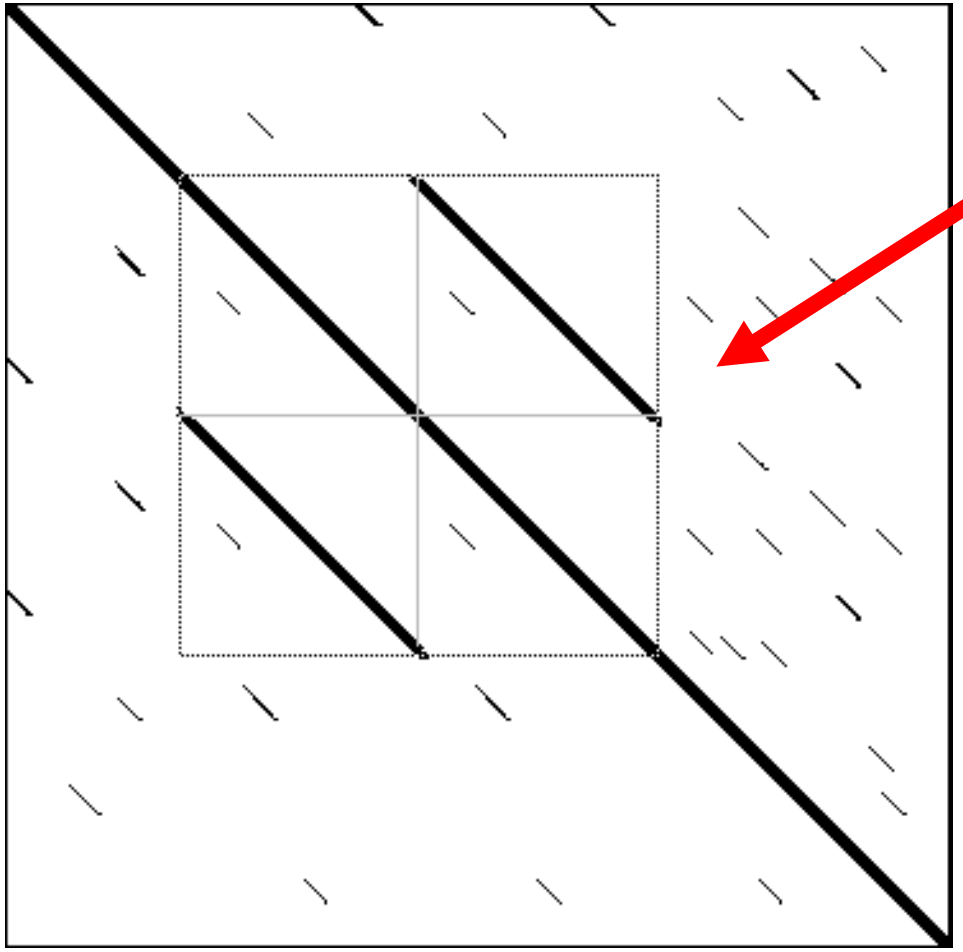


Filtrado usando k-mer=2

	AA	AT	TC	CG	GT	TT	TC	CG	GA
TA									
AT		X							
TC			X				X		
CG				X				X	
GC									
CG				X				X	
GA									X

Comparación de scuenceia consigo misma
Detección de repeticiones



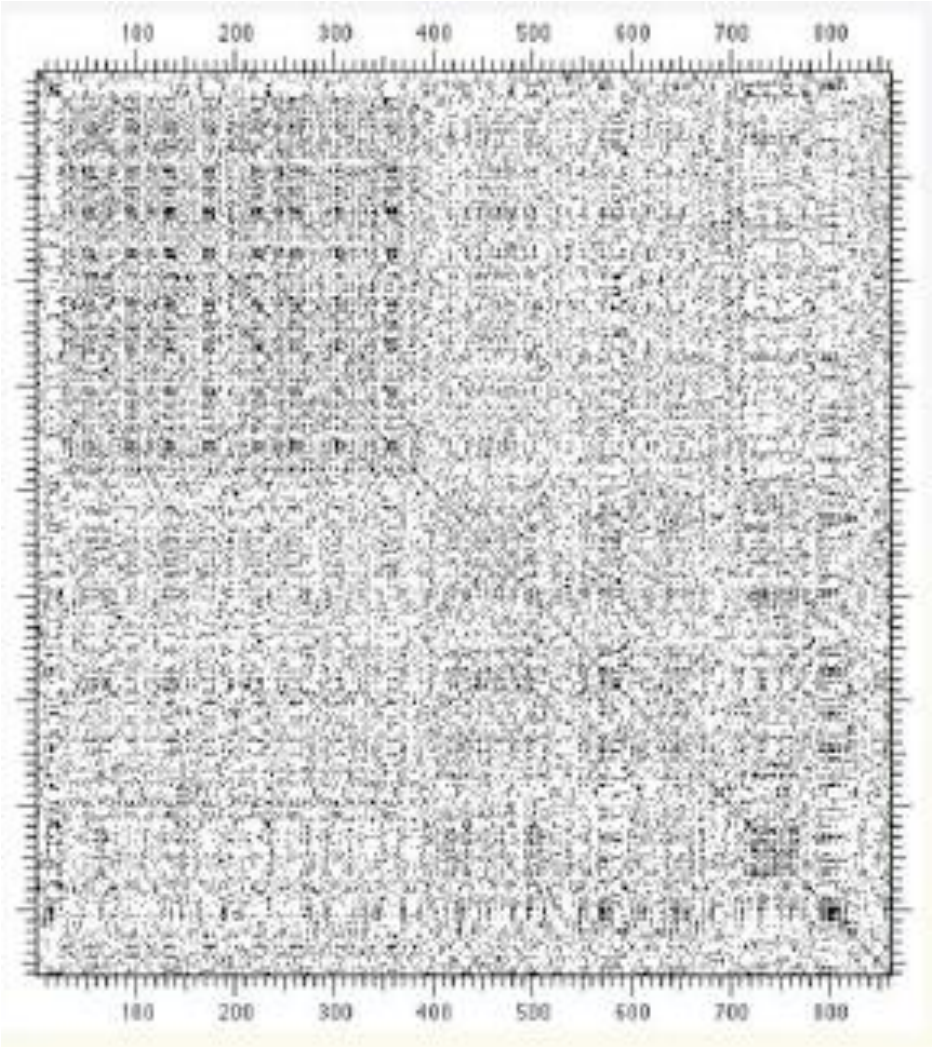


Repetición en tándem de un fragmento de la secuencia

…ABCDEFGEFGHIJKLMNO…

Comparación de una secuencia (ADN o proteína) consigo misma

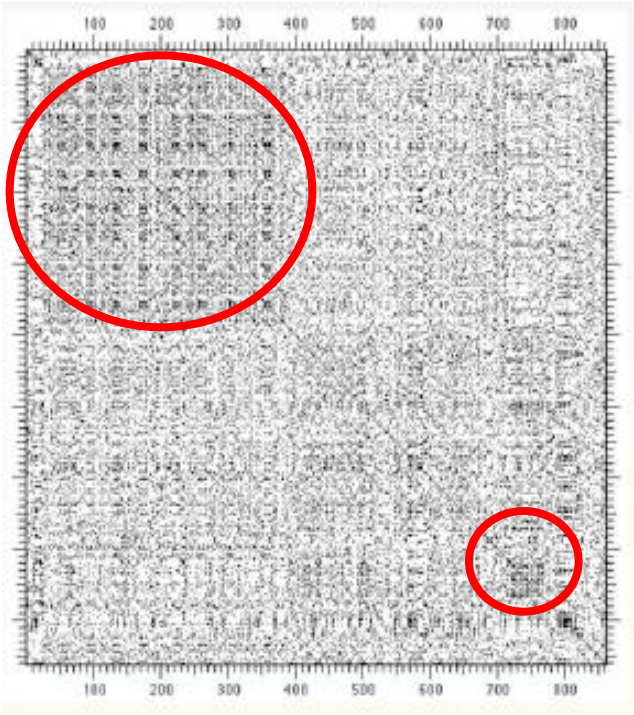
(Receptor LDL humano)



- Aparece una diagonal de lado a lado
- Hay simetría respecto a esa diagonal
- Las líneas paralelas a ambos lados de la diagonal corresponden a repeticiones de la secuencia.
- Las repeticiones invertidas o las secuencias palindrómicas aparecen como líneas perpendiculares a la diagonal principal
- Las áreas con alta densidad de puntos son repeticiones cortas de un mismo nucleótido o aminoácido (regiones de poca complejidad)
- Se ve mejor con un filtrado

Región de poca complejidad

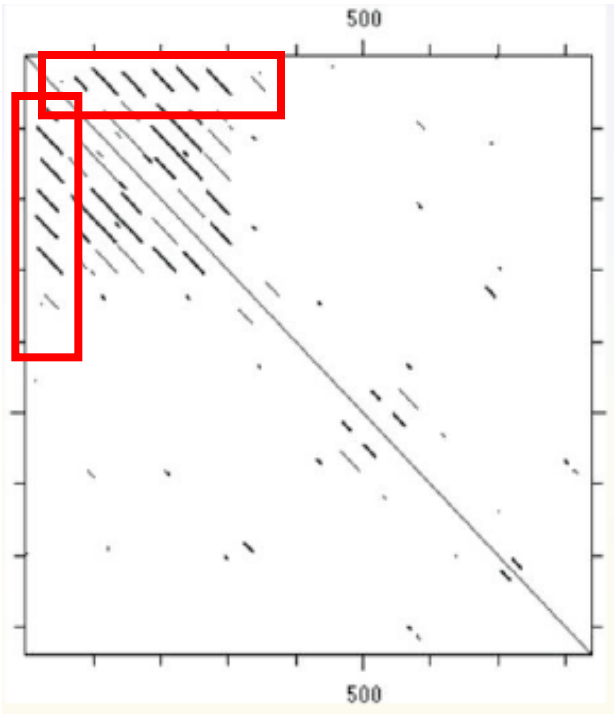
(t = 1 y r =1)



Receptor LDL humano (sin filtrar)

Regiones repetidas

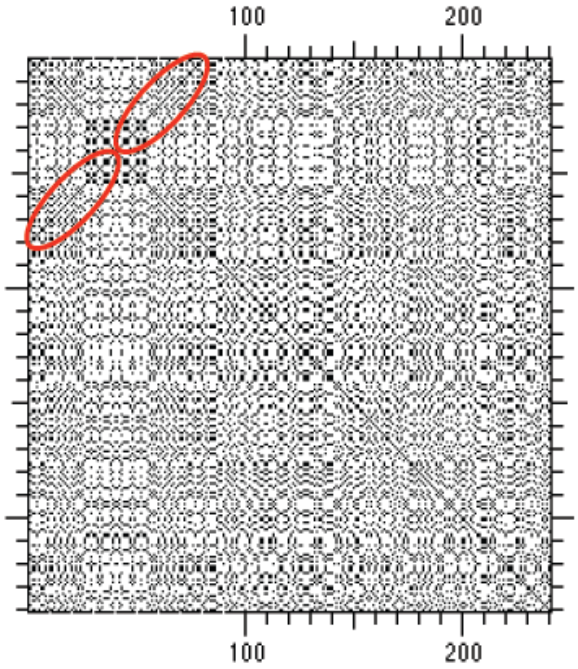
(t = 23 y r =17)



Receptor LDL humano (filtrado)

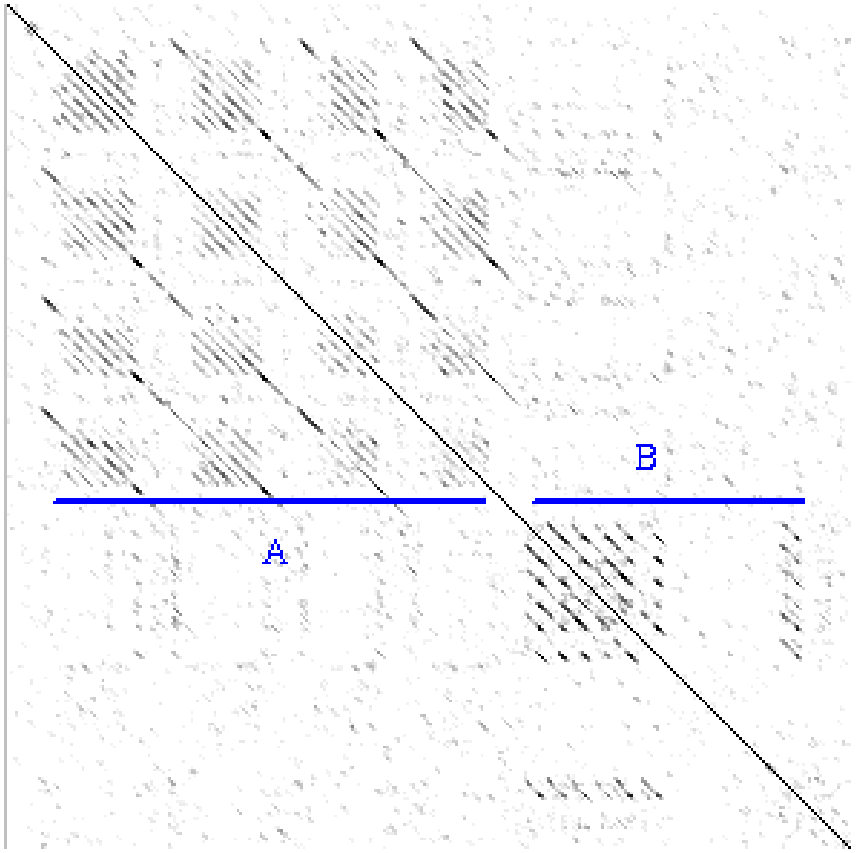
Repeticiones invertidas

(t = 1 y r =1)



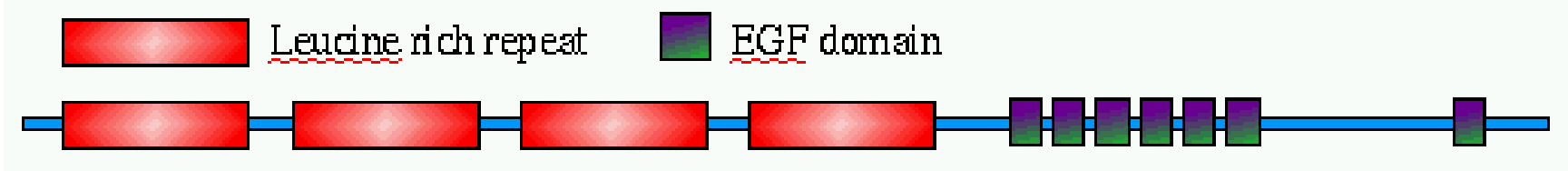
Factor de transcripción humano

Proteína SLIT de *Drosophila melanogaster*

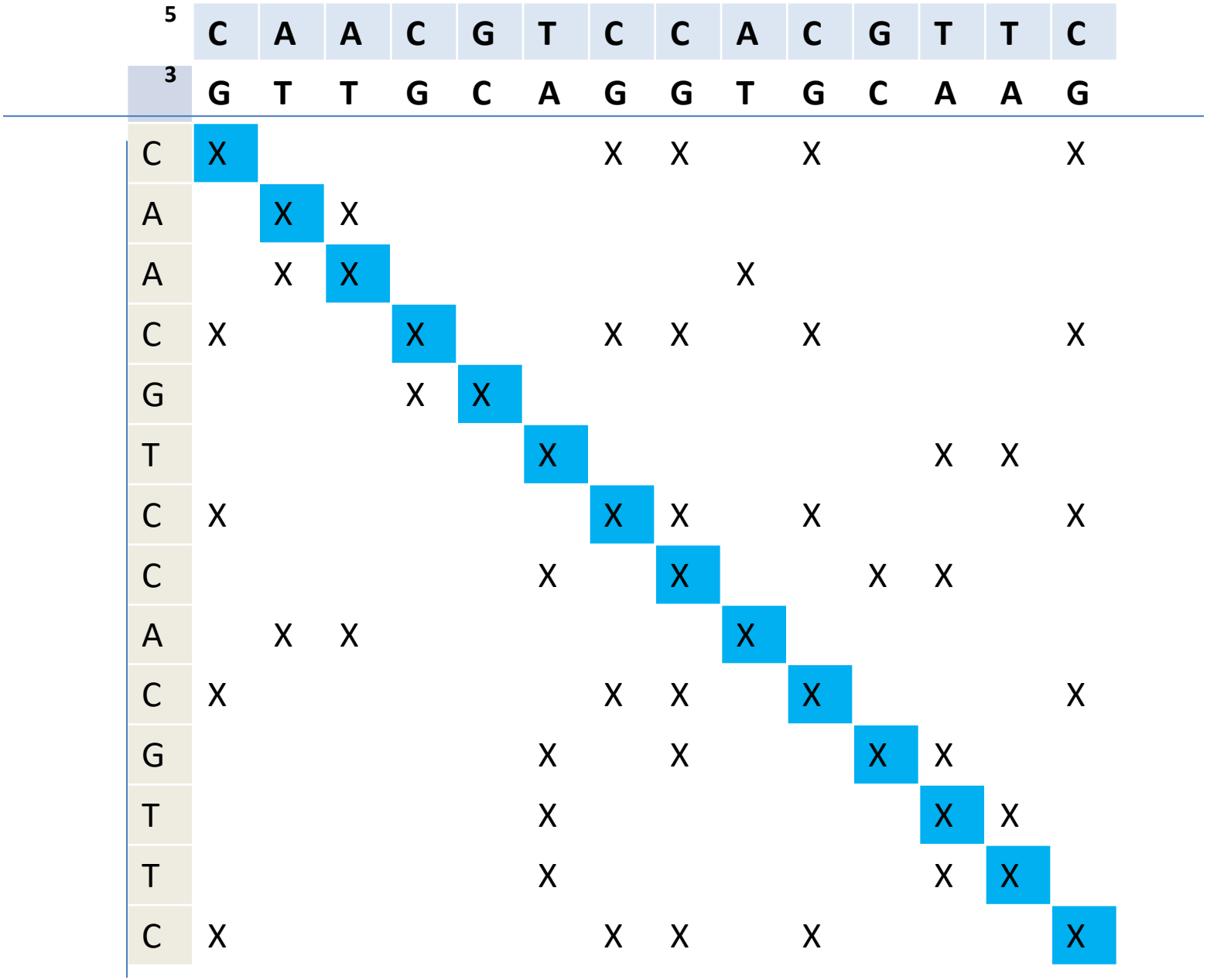


- En el extremo amino hay 4 regiones repetidas, ricas en leucina (A)

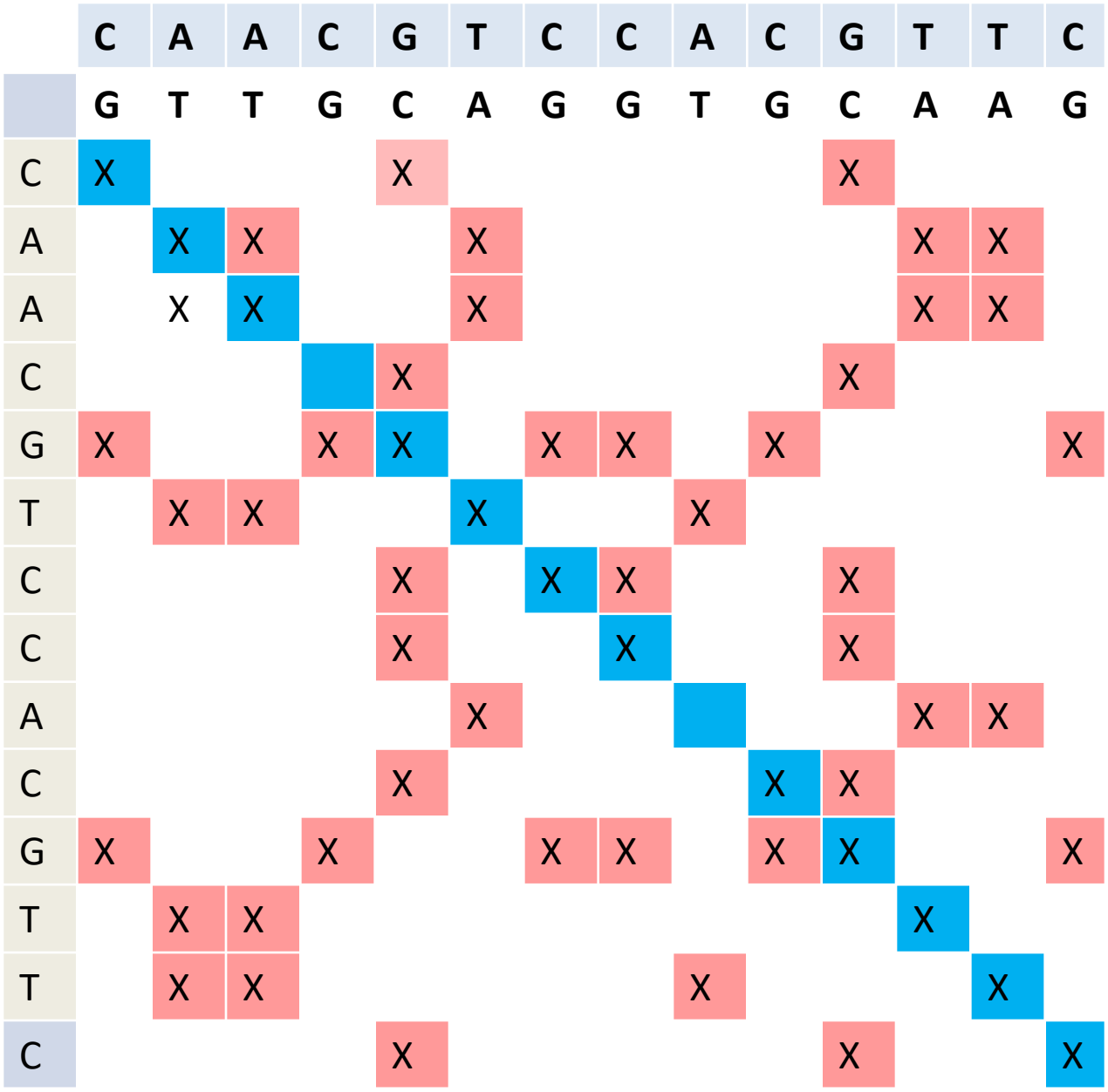
- Hay otro dominio que se repite unas 6 veces en un tramo pequeño y otra vez más cerca del extremo carboxilo (B). Es el dominio EGF.

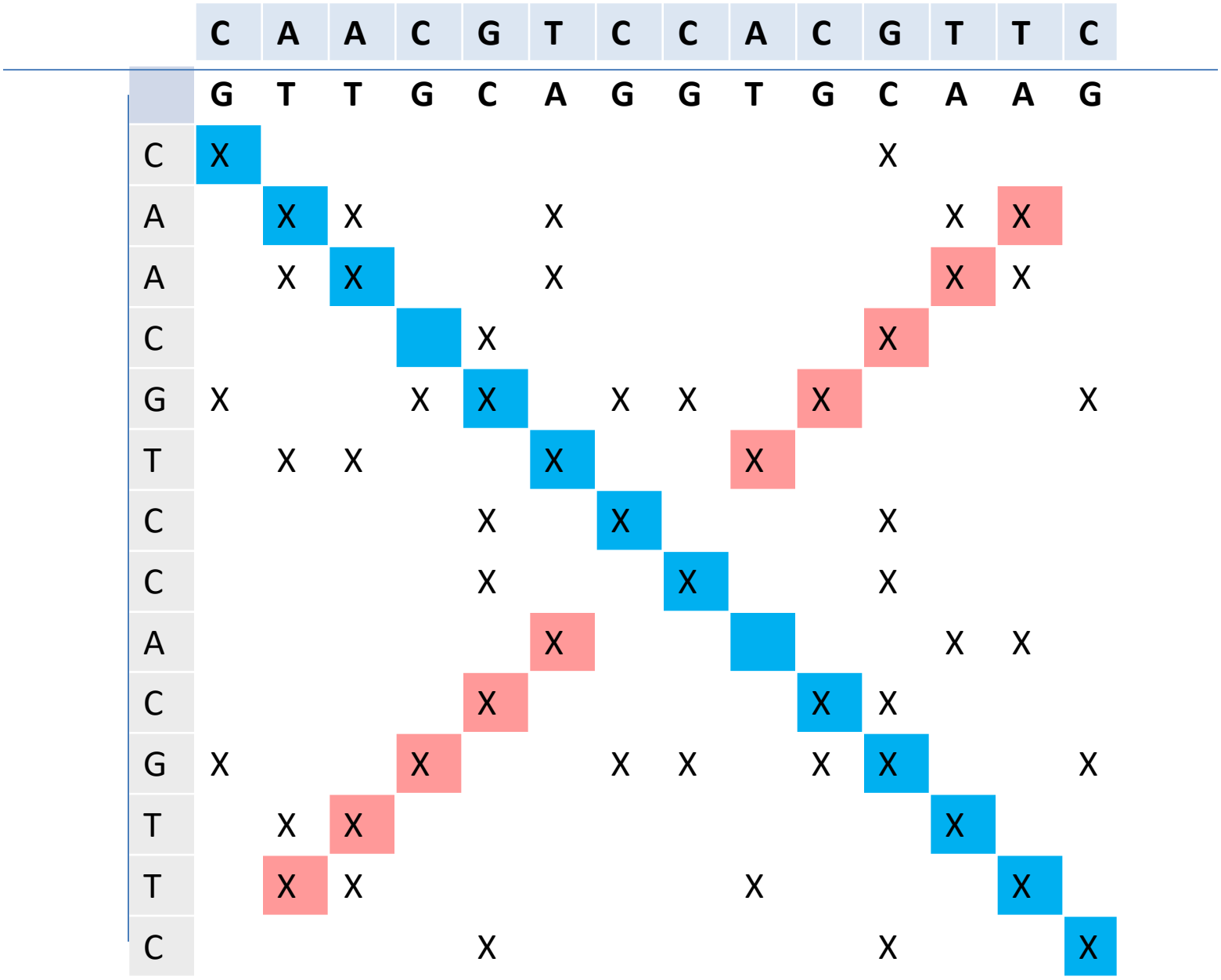


	C	A	A	C	G	T	C	C	A	C	G	T	T	C
C	X			X			X	X		X				X
A		X	X											
A		X	X						X					
C	X			X			X	X		X				X
G				X	X									
T						X						X	X	
C	X			X			X	X		X				X
C	X			X		X		X			X	X		
A		X	X						X					
C	X						X	X		X				X
G					X						X			
T						X						X	X	
T						X						X	X	
C	X			X			X	X		X				X

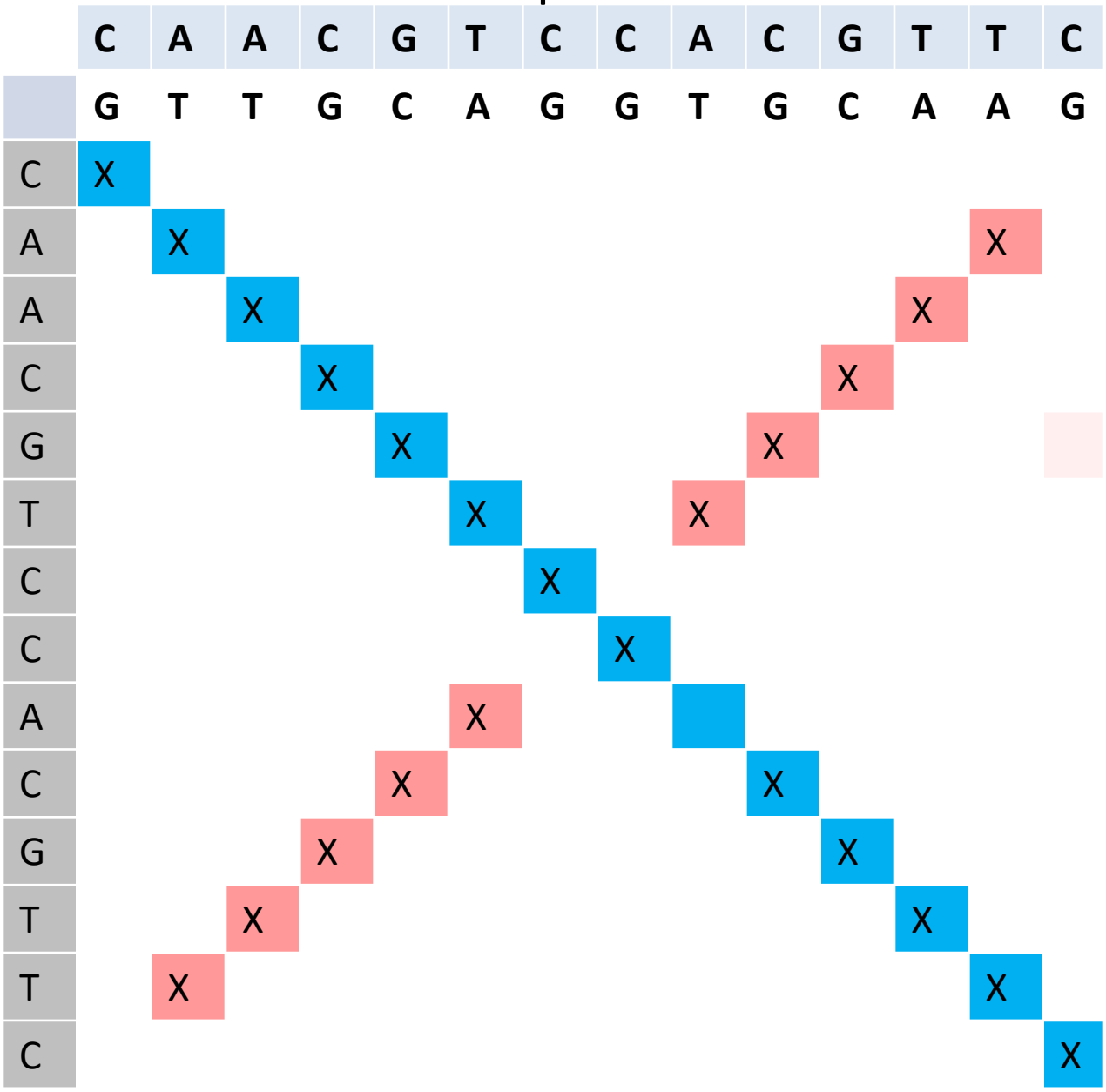

5 C A A C G T C C A C G T T C
3 G T T G C A G G T G C A A G
C X X X X
A X X
A X X X
C X X X X X X
G X X
T X X X
C X X X X X
C X X X X
A X X X
C X X X X X
G X X X X
T X X X
T X X X
C X X X X

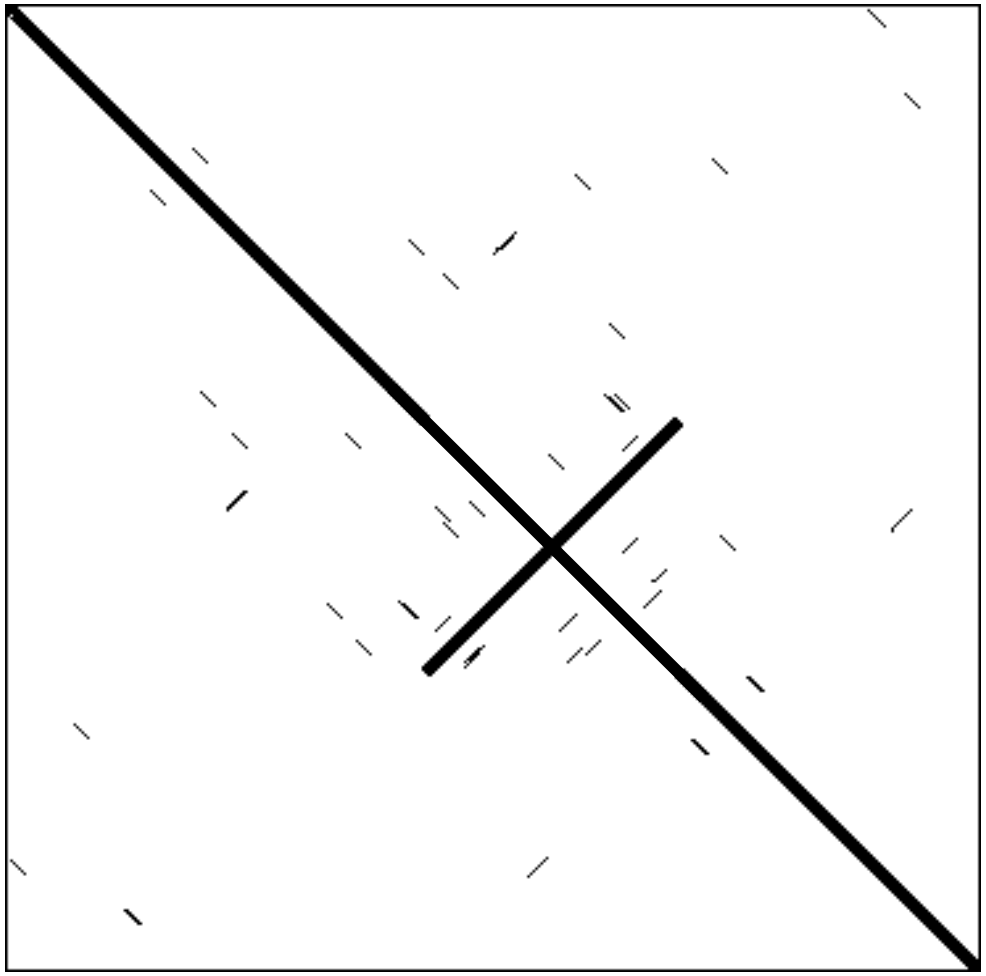
Comparación con hebra complementaria (en rojo)




C A A C G T C C A C G T T C
G T T G C A G G T G C A A G
C
A
A
C
G
T
C
C
A
C
G
T
T
C

Repetidos inversos



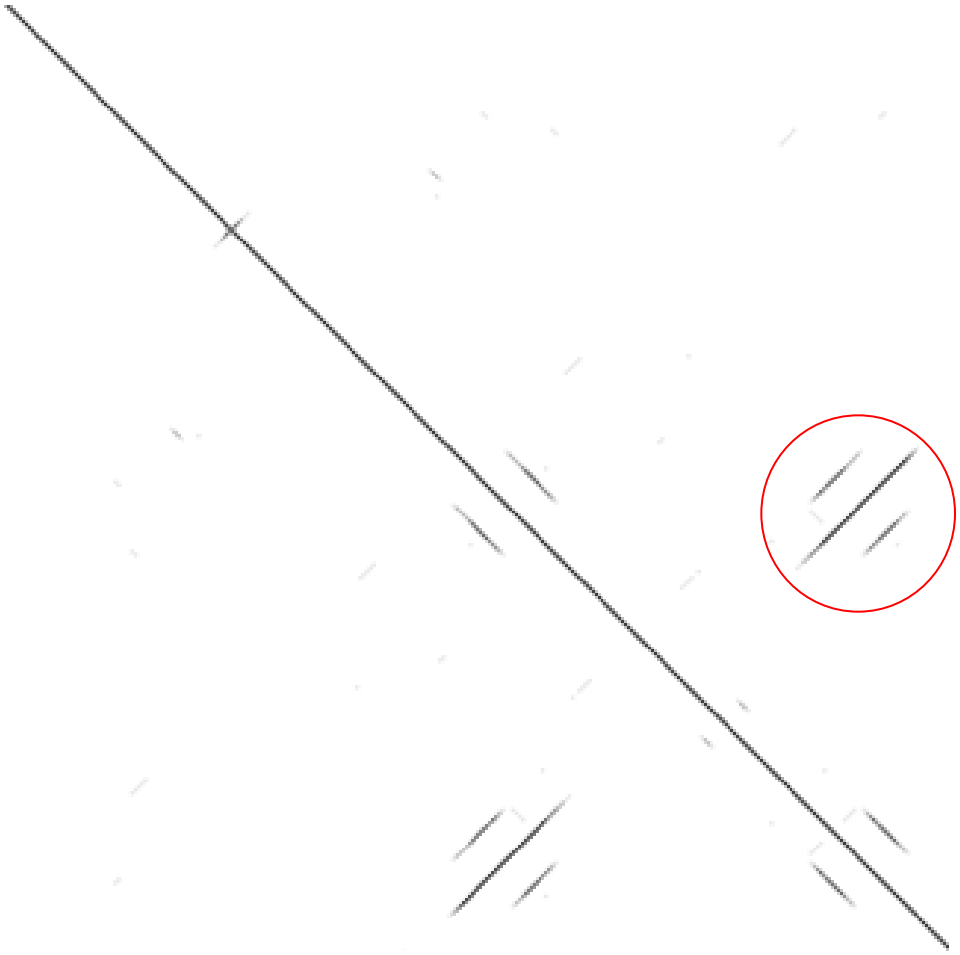


En las secuencias palindrómicas, la secuencia de una hebra se lee igual que la de su hebra complementaria:

5' GGCC 3'

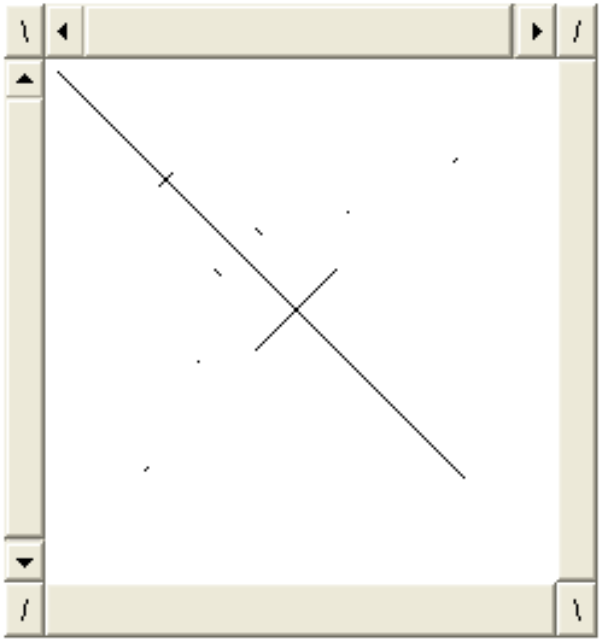
3' CCGG 5'

Las repeticiones invertidas se pueden encontrar en:



- Secuencias implicadas en la unión de los factores de transcripción
- Transposones
- Genes de retrovirus insertados en el genoma del huésped
- Genes duplicados
- Estructuras secundarias (*stem-loop*) del RNA (horquillas de terminación de la transcripción)

Horquilla de terminación en la secuencia del gen UTP-glucosa-1-fosfato uridililtransferasa de *Bacillus subtilis*



- En las regiones con apareamientos locales (estructuras *stem-loop*) la secuencia directa coincide con la de la hebra complementaria escrita en sentido inverso