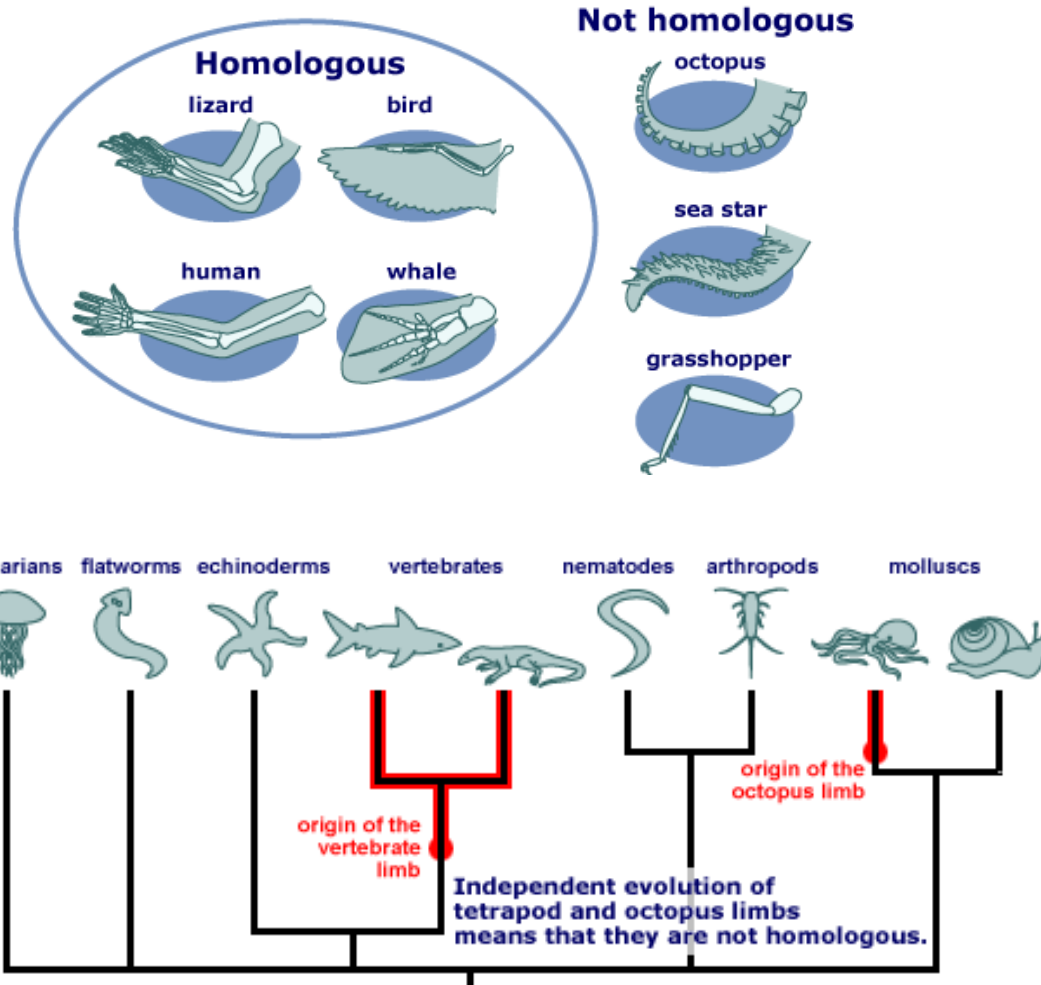


Alineamiento de Secuencias

Homología y Analogía

Homología: rasgos heredados a partir de un ancestro común.

Analogía: similitud debida a una evolución convergente.



Secuencias Homólogas

Dos secuencias son homólogas cuando comparten un ancestro común.

Similitud vs. homología

- La similitud es el parecido entre dos secuencias. Suele expresarse como el porcentaje de posiciones idénticas entre dos secuencias. Por ejemplo, dos secuencias pueden mostrar un 30% de identidad.
- En la homología no hay grados, las secuencias pueden ser homólogas o no.

Similitud NO IMPLICA homología

Tipos de secuencias homólogas

Ortólogas: Las dos secuencias derivan de un ancestro común a partir de especiación. Ejemplo: citocromo c de humanos y de ratón.

Parálogas: Las dos secuencias derivan de una secuencia ancestral a partir de duplicación génica. Ejemplo: alfa- y beta globinas humanas

Xenólogas: Una de las dos secuencias homólogas se ha adquirido por transferencia horizontal de genes.

Secuencias Homólogas

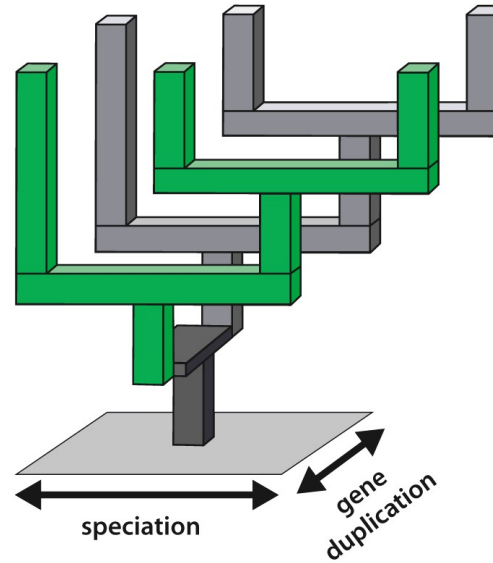
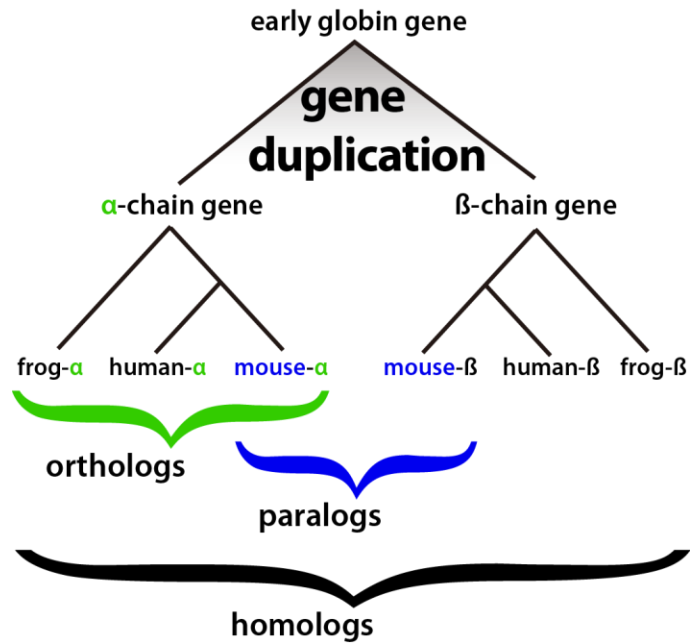
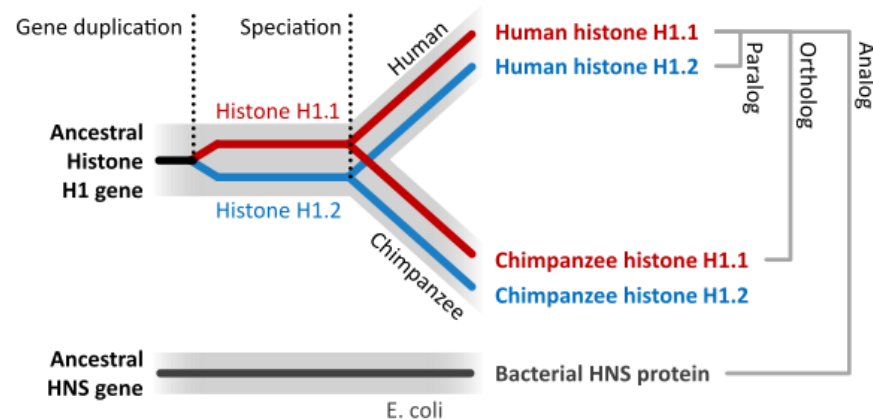
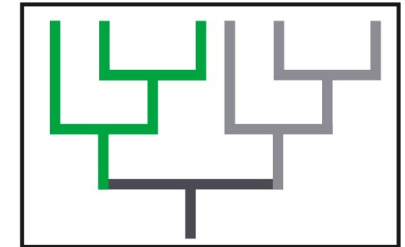
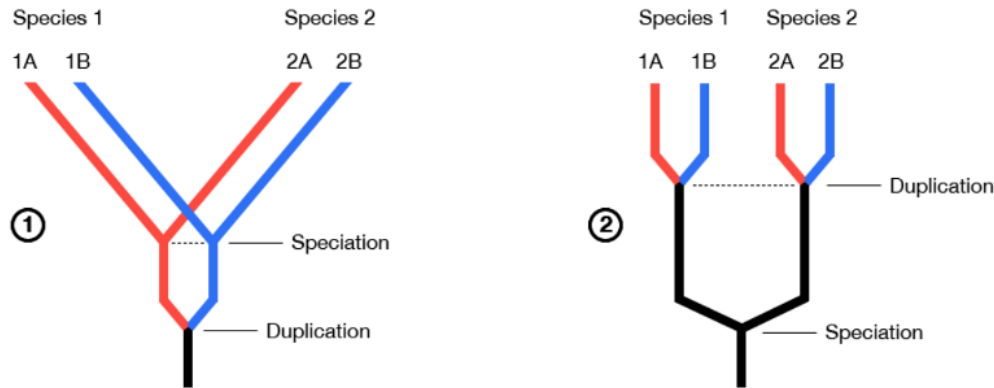


Figure 5.1 Phylogenomics: A Primer (© Garland Science 2013)



¿Qué secuencias son parálogas y ortólogas?

The diagrams below show gene duplication in two lineages.



A. In diagram 1, a paralog to 1A is _____. An ortholog to 1A is _____.

B. In diagram 2, a paralog to 1A is _____. An ortholog to 1A is _____.

Modelo de evolución de secuencias

Supongamos dos secuencias actuales (B y C), con un ancestro común (A), es decir, homólogas:



Tipos de mutaciones:

- Sustituciones
- Inserciones/Deleciones (Indels)

Ensembl Variation - Variant classification

Sequence variants

Type	Description	Example (Reference / Alternative)	
SNP	Single Nucleotide Polymorphism	Ref: . . . TTG A CGTA . . .	Alt: . . . TTG G CGTA . . .
Insertion	Insertion of one or several nucleotides	Ref: . . . TTGACGTA . . .	Alt: . . . TTGA TG CGTA . . .
Deletion	Deletion of one or several nucleotides	Ref: . . . TTG AC GTA . . .	Alt: . . . TTGGTA . . .
Indel	An insertion and a deletion, affecting 2 or more nucleotides	Ref: . . . TTG AC GTA . . .	Alt: . . . TTG GCT CGTA . . .
Substitution	A sequence alteration where the length of the change in the variant is the same as that of the reference.	Ref: . . . TTG AC GTA . . .	Alt: . . . TTG TA GTA . . .

Structural variants

Type	Description	Example (Reference / Alternative)	
CNV	Copy Number Variation: increases or decreases the copy number of a given region	Reference:	"Gain" of one copy:
			"Loss" of one copy:
Inversion	A continuous nucleotide sequence is inverted in the same position	Reference:	Alternative:
Translocation	A region of nucleotide sequence that has translocated to a new position	Reference:	Alternative:

Matriz de puntos: alineamiento de secuencias

a)

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

b)

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

c)

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

d)

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

Alineamiento 1:

AT--GCGTCGTT

ATCCGCGTC---

Alineamiento 2:

ATCGTCGT

ATCG-CGT

a) Las dos secuencias son idénticas en la parte alineada.

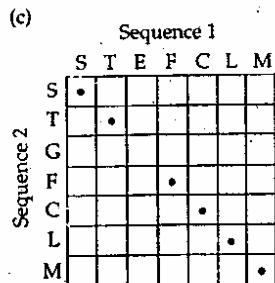
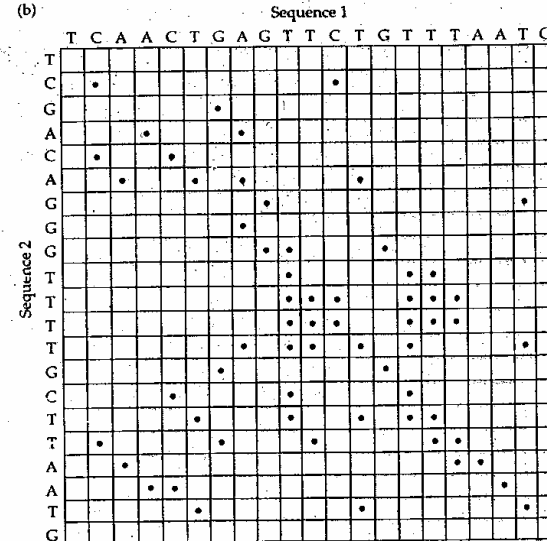
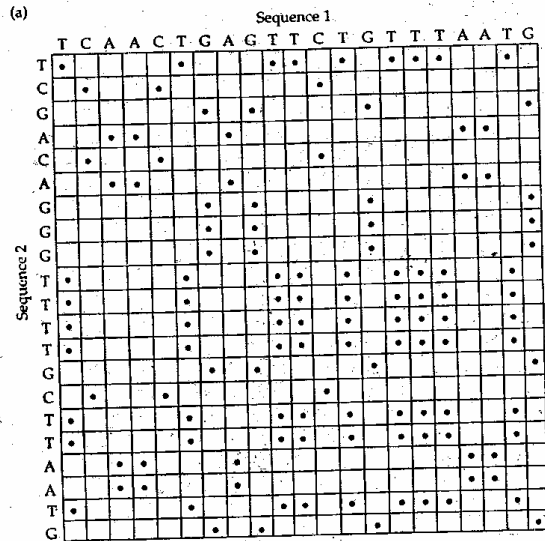
b) Las dos secuencias muestran un desemparejamiento debido a una sustitución; la posición (3,3) se queda en blanco.

c) Las secuencias difieren por un indel, dando lugar a un hueco. Nótese el quiebro en zig-zag de la diagonal.

d) Dos posibles alineamientos de las mismas secuencias. El alineamiento 1 supone cinco huecos, mientras que el 2 supone 1 hueco y 2 desemparejamientos.

Matriz de puntos: alineamiento de secuencias

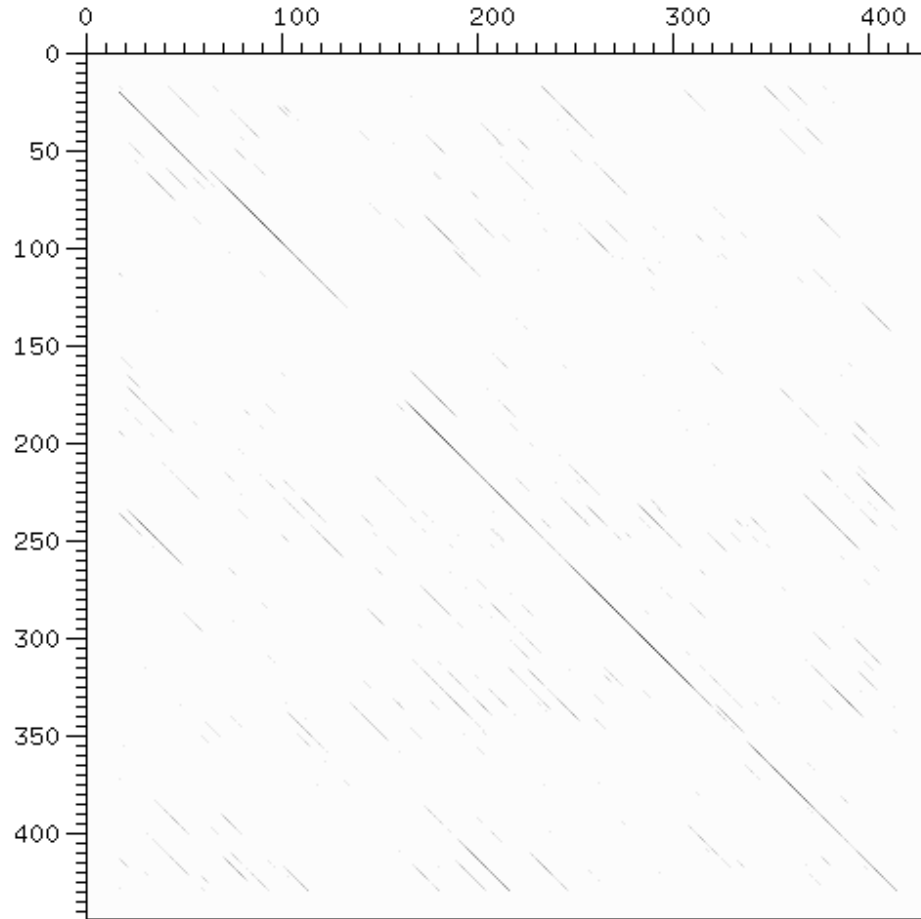
Visualizar correctamente los patrones compartidos entre dos secuencias depende de poder eliminar el ruido que pueden producir similitudes espúreas



- a) De esta manera, si pintamos un punto en cada nucleótido coincidente, la nube de puntos resultante oculta el patrón.
- b) Si utilizamos una ventana móvil de 3 nucleótidos y sólo pintamos un punto cuando 2 de los 3 son iguales, el patrón compartido se observa más claro.
- c) Esto mejor si en lugar del ADN alineamos la proteína resultante de ambas secuencias.

Utilidades de la matriz de puntos

Regiones de alta similitud entre parálogos: α - y β -globina humana



La matriz de puntos muestra en la diagonal las regiones con alta similitud de secuencia, coincidiendo con los exones de ambas secuencias.

Utilidades de la matriz de puntos

Homologías remotas: Human μ -crystallin vs. *Salmonella* glutamyl-tRNA reductase

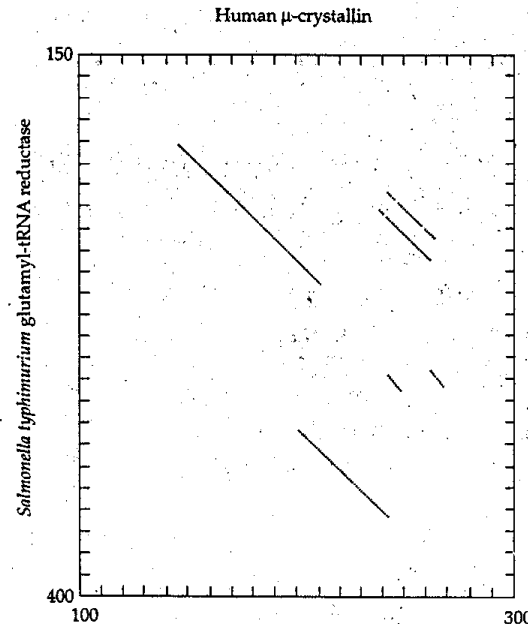
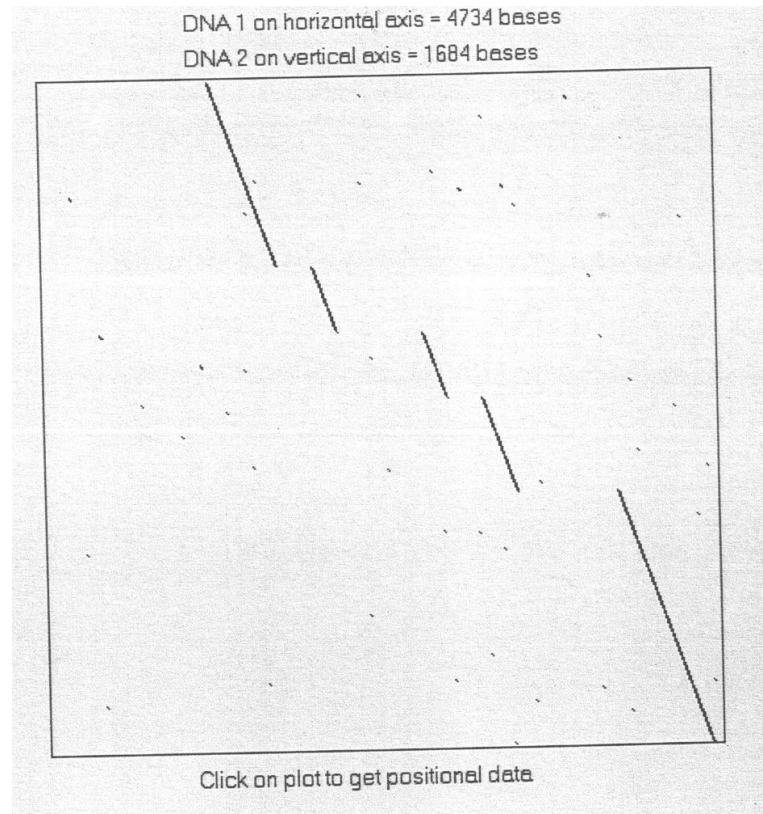


FIGURE 3.9 Dot matrix comparison of amino acid sequences between human μ -crystallin and glutamyl-tRNA reductase from *Salmonella typhimurium*. Only the C-terminal ends of the proteins were used. The window size was 60 amino acids and the stringency was 24 matches. Diagonals show regions of similarity. The vertical gap indicates that a coding region corresponding to approximately 75 amino acids has either been deleted from the human gene or inserted into the bacterial gene. The two diagonally oriented parallel lines in the upper right-hand corner of the plot most probably indicate that a small internal duplication has occurred in the bacterial gene. Modified from Segovia et al. (1997).

Las matrices de puntos pueden mostrarnos dominios homólogos en secuencias que supondríamos no-homólogas. En este caso se muestran dos regiones similares divididas por un indel en la diagonal y una probable duplicación interna en el gen bacteriano.

Utilidades de la matriz de puntos

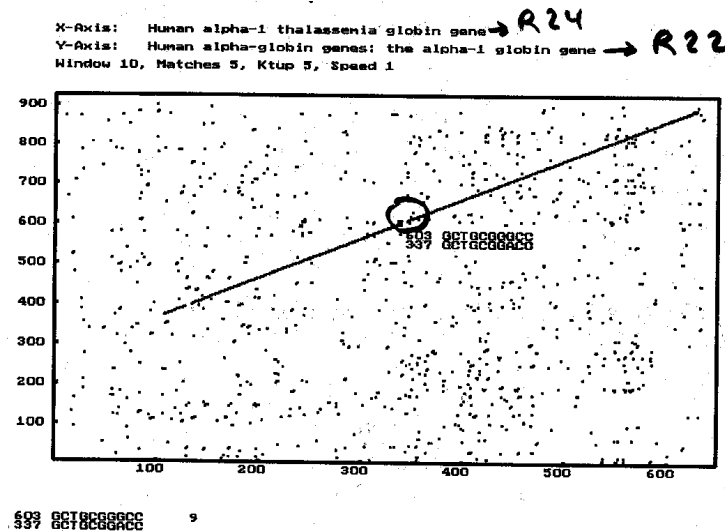
Estructura exones-intrones: Gen vs. ARNm maduro de la rodopsina de *Xenopus*



La estructura exones-intrones en el ARNm pueden ser visibles al comparar el ADN y el ARNm de un gen.

Utilidades de la matriz de puntos

Detección de mutaciones al comparar secuencias entre dos individuos



○
ALIGN
↓

```
Sequence 1 Human alpha-1 thalassemia globin gene (334-353)
Sequence 2 Human alpha-globin genes: the alpha-1 globin gene (600-619)

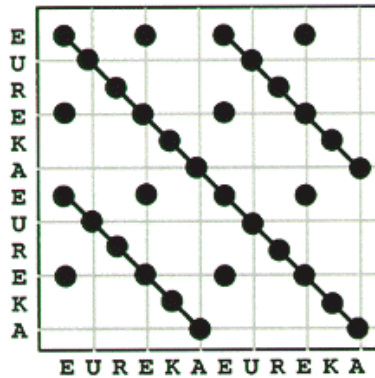
Sequence 1 GCGGCTGCGGACCTGGGCCC
          *****
Sequence 2 GCGGCTGCGGCGCTGGGCCC

Sequence 1 334-353
Sequence 2 600-619
Aligned    20
Matches    19
Mismatches 1
Homology   95%
```

Utilidades de la matriz de puntos

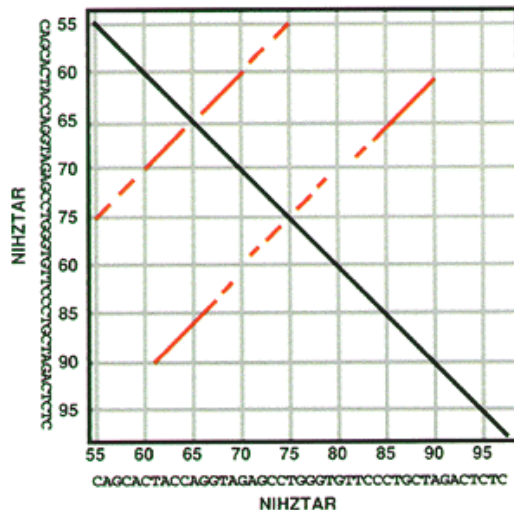
Detección de elementos repetidos directos (A) e invertidos (B)

A



(A) Elementos repetidos directos, representados por líneas de puntos paralelas a la diagonal.

B



(B) Elementos repetidos indirectos, representados por líneas de puntos perpendiculares a la diagonal.

Evaluación de alineamientos

Consideremos dos secuencias:

A: TCAGACGATTG (m=11)

B: TCGGAGCTG (n=9)

Emparejamientos - matches (x) = |
Desemparejamientos - mismatches (y) = *
Huecos - Gaps (z) = -

Se podrían realizar al menos tres alineamientos diferentes, según el parámetro que se desee minimizar:

(I) Reducir el número de desemparejamientos a cero:

```
TCAG-ACG-ATTG
|| | | | |
TC-GGA-GC-T-G
```

x=7 y=0 z=6

(II) Reducir el número de huecos al mínimo $|m-n| = 2$:

```
TCAGACGATTG
||*||****
TCGGAGCTG--
```

x=4 y=5 z=2 (o $z_2=1$)

(III) Alineamiento con un equilibrio entre desemparejamientos y huecos:

```
TCAG-ACGATTG
|| | | |*|*
TC-GGA-GCTG-
```

x=6 y=2 z=4

¿Cuál de estos alineamientos es más probable?

Evaluación de alineamientos: Método de la distancia (Waterman)

Comparemos los alineamientos I, II y III mediante dos sistemas de penalización para los huecos:

1) $D = y + w \cdot z$ (Penalización lineal por gap)

Con $w = 2$ tendríamos:

I: $D = 0 + (2 \times 6) = 12$

II: $D = 5 + (2 \times 2) = 9$

III: $D = 2 + (2 \times 4) = 10$

2) $D = y + \sum w_k \cdot z_k$ (Penalización flexible, huecos convexos)

Con $w_1 = 2$, $w_2 = 6$ tendríamos:

I: $D = 0 + (2 \times 6) = 12$

II: $D = 5 + (6 \times 1) = 11$

III: $D = 2 + (2 \times 4) = 10$

En este caso existen dos penalizaciones, w_1 para huecos de una posición y w_2 para huecos de más de una posición. De esa manera se reduce la penalización por extensión de un hueco.

I) $x=7$ $y=0$ $z=6$

TCAG-ACG-ATTG

| | | | | |

TC-GGA-GC-T-G

II) $x=4$ $y=5$ $z=2$ (o $z_2=1$)

TCAGACGATTG

| | * | | * * * *

TCGGAGCTG--

III) $x=6$ $y=2$ $z=4$

TCAG-ACGATTG

| | | | | * | *

TC-GGA-GCTG-

Nótese que con penalizaciones diferentes, los resultados podrían ser otros!

El alineamiento más probable depende de los parámetros y la penalización utilizada, es decir depende de lo que busquemos

Penalización por hueco

Se han propuesto diferentes aproximaciones a la penalización por apertura y extensión de hueco:

CHAPTER 3

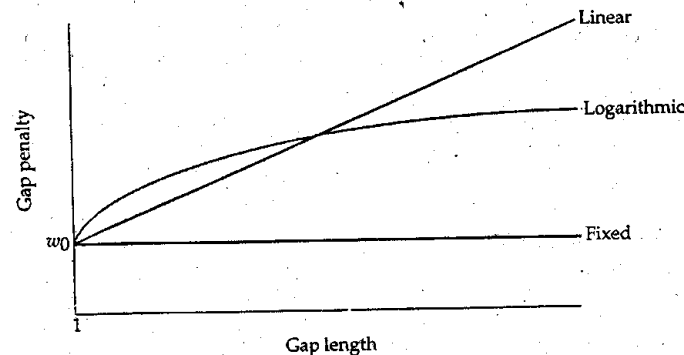
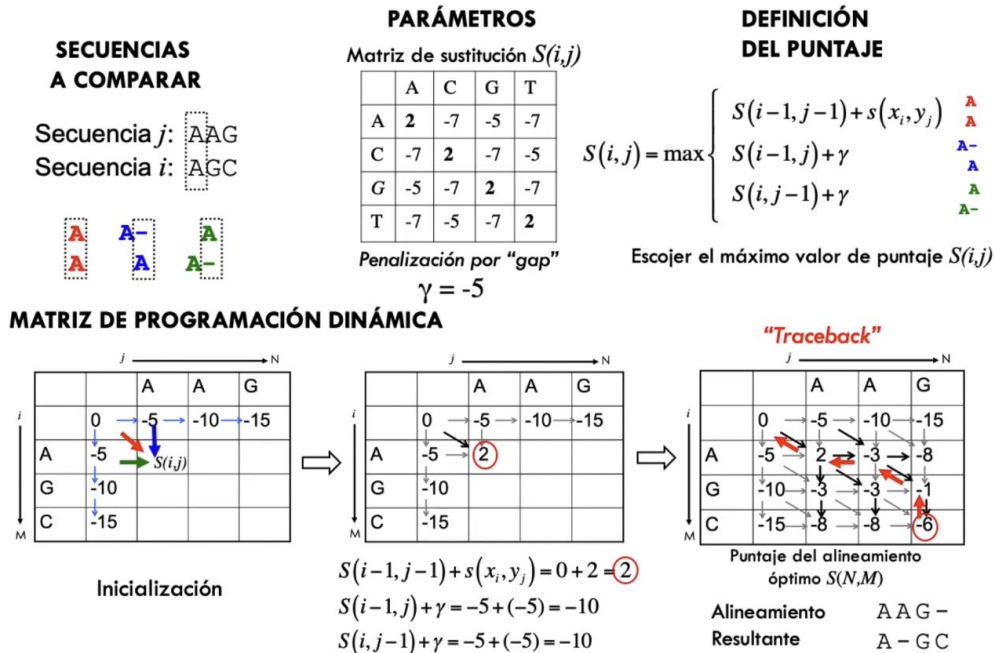


FIGURE 3.10 Three gap penalty systems. The gap-opening penalty in all three systems is w_0 . In the linear gap penalty system, the gap-extension cost is the product of the gap length minus 1 and the gap-extension penalty for increasing the gap by 1. In the logarithmic gap penalty system, the gap-extension penalty increases with the logarithm of the gap length. In the fixed gap penalty system, there are no gap-extension costs.

- Linear: la extensión y la apertura tienen un coste similar.
- Logarítmica: la relación entre la penalización por extensión y apertura se reduce cuanto mayor sea la extensión.
- Fija: No existe penalización por extensión.

Alineamiento global: Algoritmo de Needleman y Wunsch

- Es un ejemplo de algoritmo de programación dinámica
- Este algoritmo siempre termina y garantiza que la solución devuelta es la óptima
- Método cuantitativo: matrices de sustitución (información biológica previa) y penalización por apertura y extensión de hueco



1. Definimos matriz de sustitución y penalización por hueco
2. Definimos sistema de puntaje (score)
3. Creamos primera línea y columna con sumatorio de puntuación por huecos
4. Aplicamos sistema de score seleccionando siempre el valor mayor para la siguiente celda
5. Una vez completa, recorremos la matriz desde la celda inferior seleccionando la celda con mayor puntuación. Donde una flecha vertical supone un hueco en la secuencia horizontal, una diagonal un alineamiento y una flecha horizontal un hueco en la secuencia vertical.

EMBOSS Needle

Alineamiento global: Algoritmo de Needleman y Wunsch

EMBOSS Needle

```
#####
# Program: needle
# Rundate: Mon 15 Jul 2013 12:00:00
# Commandline: needle
# [-asequence] tsw:hba_human
# [-bsequence] tsw:hbb_human
# Align_format: srspair
# Report_file: hba_human.needle
#####
```

```
#-----
```

```
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBL0SUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:          9/149 ( 6.0%)
# Score: 292.5
#
#-----
```

```
HBA_HUMAN      1 MV-LSPADKTNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHF-D    48
                || |:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|
HBB_HUMAN      1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD  48

HBA_HUMAN     49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR  93
                || |:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|
HBB_HUMAN     49 LSTPDVAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH    98

HBA_HUMAN     94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                |||:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|
HBB_HUMAN     99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

```
#-----
#-----
```

EBLOSUM62

	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
A	0	1	0	4																	A
G	-3	0	-2	0	6																G
P	-3	-1	-1	-1	-2	7															P
D	-3	0	-1	-2	-1	-1	6														D
E	-4	0	-1	-1	-2	-1	2	5													E
Q	-3	0	-1	-1	-2	-1	0	2	5												Q
N	-3	1	0	-2	0	-2	1	0	0	6											N
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										H
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									R
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								K
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							M
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					L
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				V
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			W
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		Y
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

La puntuación tiene en cuenta la frecuencia relativa de cada aminoácido y sus probabilidades de sustitución

Alineamiento global: Algoritmo de Needleman y Wunsch

EMBOSS Needle

Length: The length of the alignment, including any gaps that have been introduced to construct the alignment.

Identity: This is a count of the number of positions over the length of the alignment where all of the residues or bases at that position are identical.

Similarity: This is a count of the number of positions over the length of the alignment where $\geq 51\%$ of the residues or bases at that position are similar.

Gaps: This is a count of the number of positions over the length of the alignment where there are one or more sequences with a gap.

Score: This is the score used by the program that calculated the alignment to determine which is the best possible alignment to report.

Markup Line: Is the line commonly placed between a pairwise alignment or at the bottom of alignments of 3 or more sequences that shows where sequences are mismatched, gapped, identical or similar.

In general the markup line uses a space for a mismatch or a gap, '.' for any small positive score, ':' for a similarity which scores more than 1.0, and '|' for an identity where both sequences have the same residue regardless of its score