Tema 5. Alineamiento de secuencias

Actualizado en: 24/01/2023

Objetivos

- Motivación y contexto
- Alineamientos de pares de secuencias
  - Alineamientos globales
  - Alineamientos locales
  - Matrices de similitud
- Alineamientos múltiples
  - Clustal
  - Muscle

Motivación y contexto

Glosario

Homología: Relación entre secuencias que descienden de un ancestro común. No es cuantificable.

Similitud: Grado (cuantificable) en que se parecen dos secuencias.

Identidad: El porcentaje de resíduos idénticos entre dos secuencias alineadas es una entre muchas posibles medidas de similitud.

Alinear: Colocar secuencias de DNA, RNA o proteínas para buscar las regiones de similitud que pueda haber entre ellas como consecuencia de sus relaciones funcionales, estructurales o evolutivas.

Ejemplo

GTCGTAGAATA

CACGTAG--TA

Aplicaciones

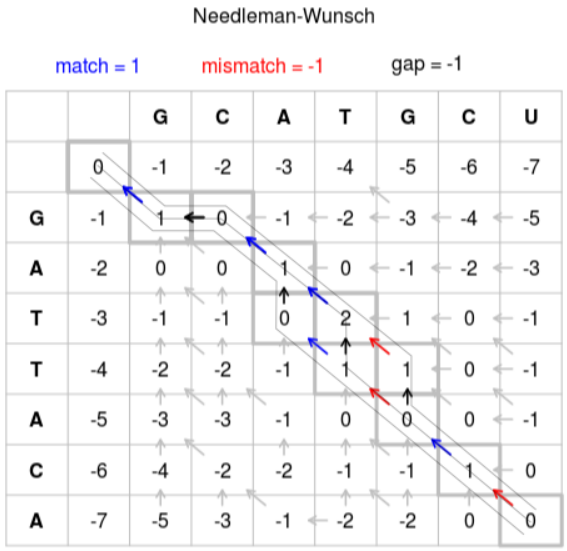
- Identificar homología para inferencia filogenética o funcional.
- Localizar el origen genómico de una secuencia corta (mapear).
- Identificar intrones en la secuencia de un gen.
- Reconocer sintenia y reordenación cromosómica.
- Identificar dominios proteicos conservados.
- Encontrar genes en un genoma secuenciado.
- etc.

Paradigma de análisis bioinformático

- Alinear es un problema de **optimización**.
- El alineamiento *óptimo* depende de la finalidad.
- Es habitual utilizar procedimientos **heurísticos**.
- Es necesario evaluar la calidad del resultado.
- Los errores del alineamiento se arrastran en los pasos siguientes.

Alineamiento de pares de secuencias

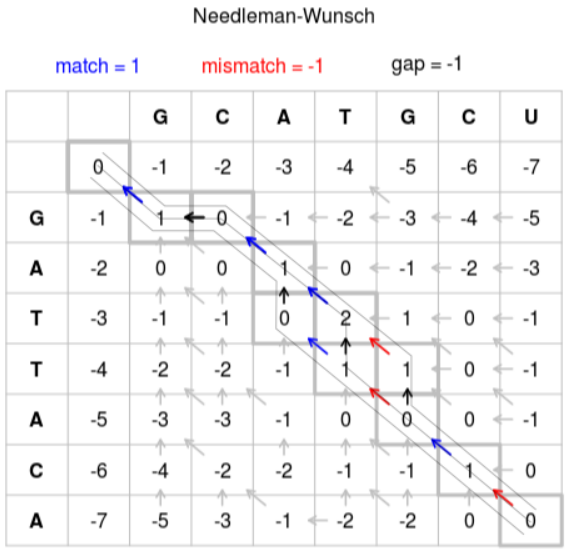
Alineamientos globales



		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Algoritmo de Needleman-Wunsch (programación dinámica). Garantiza el alineamiento *óptimo*. La puntución total es la suma de las puntuaciones de cada posición.

Alineamientos globales



		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Tres alineamientos de puntuación óptima de 0:

```
GCATG-CU      GCA-TGCU      GCAT-GCU
G-ATTACA      G-ATTACA      G-ATTACA
```

Esquemas de puntuación

- Simple:
  - Coincidencia: +1
  - *Mismatch*: -1
  - *indel*: -1
- Para favorecer alineamientos con pocos gaps más largos en lugar de muchos y cortos, se penaliza más la **aparición** que la **extensión** del *gap*.
- Matrices de similitud: puntúan cada cambio posible y definen lo que entendemos por un *buen* alineamiento.

Matrices de similitud

	A	G	C	T
A	1	-1	-1	-1
G	-1	1	-1	-1
C	-1	-1	1	-1
T	-1	-1	-1	4

Este esquema de puntuación valora más las coincidencias de timinas en el alineamiento.

Matriz PAM

Margaret Dayhoff introdujo las matrices PAM en 1978 para el alineamiento de proteínas.

Una matriz PAM_n corresponde al tiempo suficiente para que *n* mutaciones hayan aparecido entre 100 aminoácidos.

PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

BLOSUM (*block substitution matrix*)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

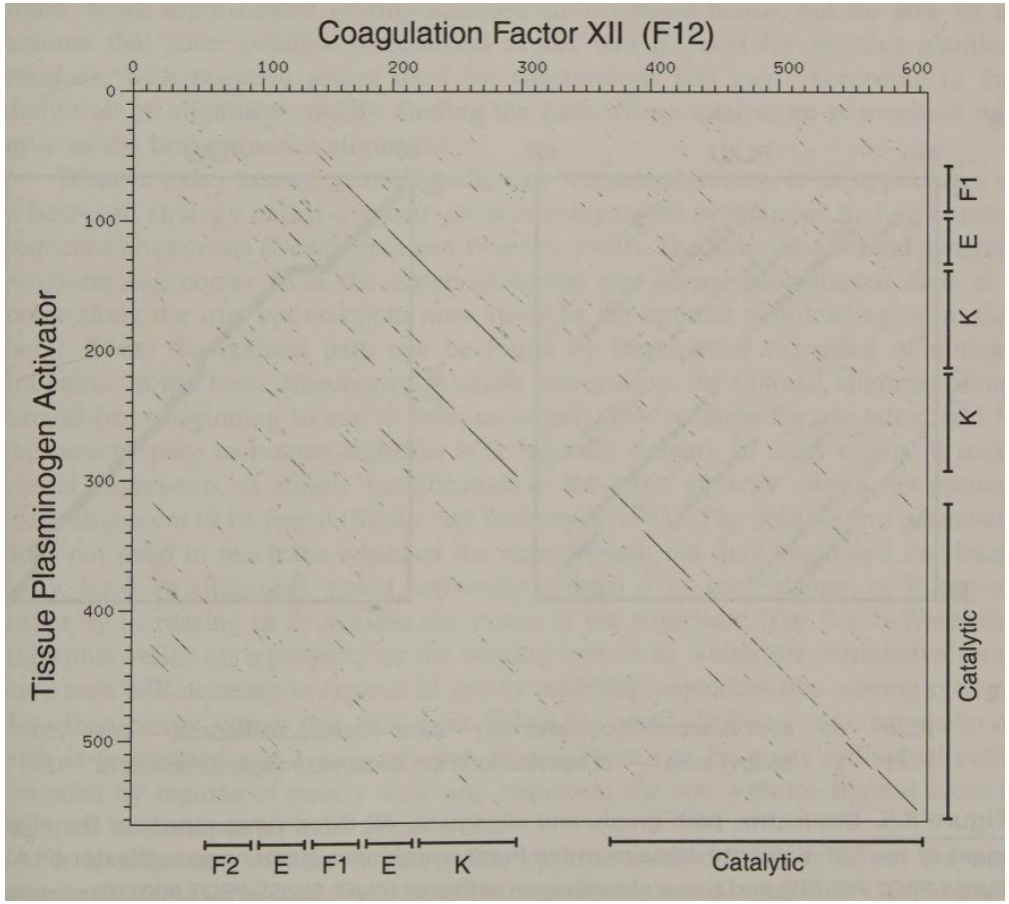
BLOSUM62

BLOSUM vs PAM

Correspondencia aproximada entre matrices PAM y BLOSUM

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM62
PAM200	BLOSUM50
PAM250	BLOSUM45

Alineamientos locales



Algoritmo de Smith-Waterman

Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$

$W_1 = 2$

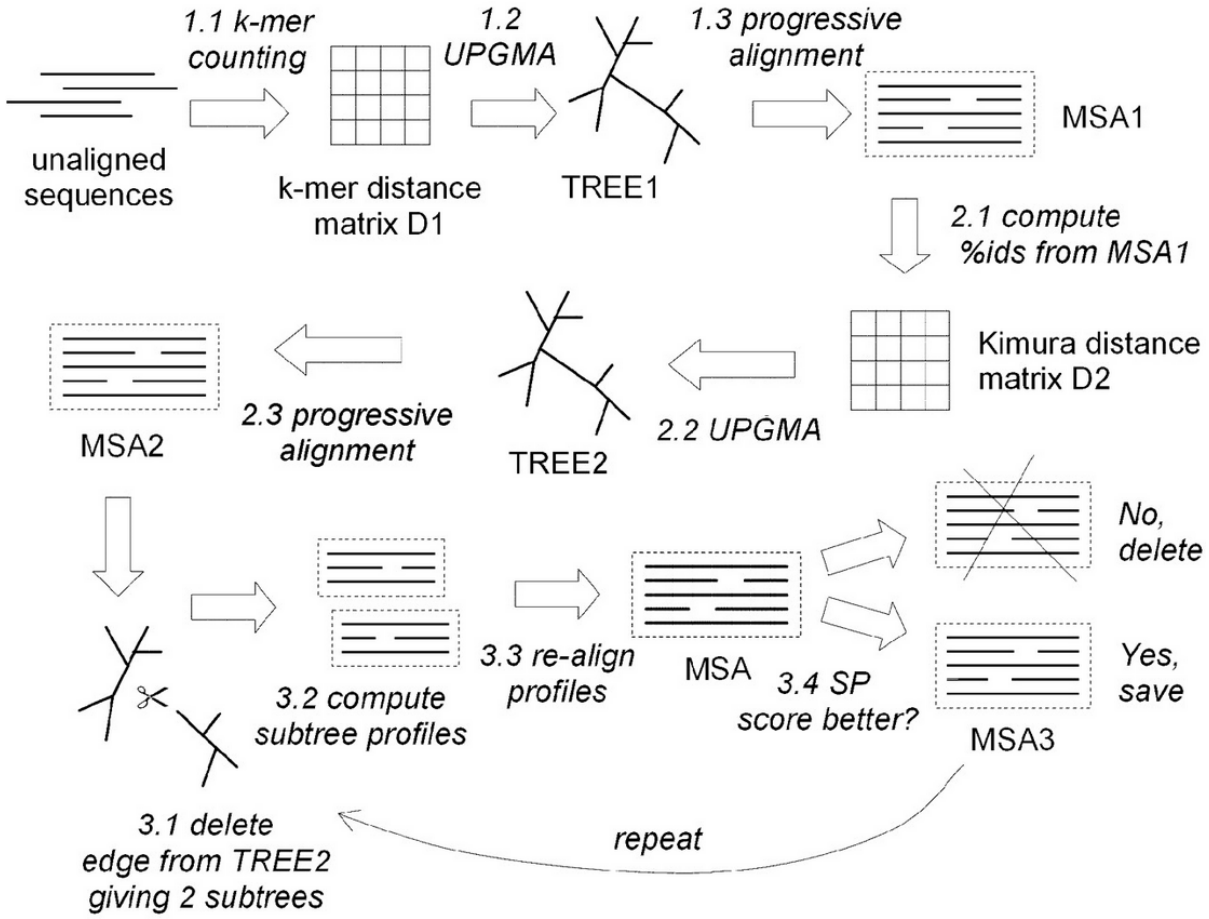
Alineamientos múltiples

CLUSTAL

- Diferentes versiones: CLUSTAL (1988), CLUSTAL V (1992), CLUSTAL W (1994), CLUSTAL X (1997), CLUSTAL 2 (2007), CLUSTAL Ω (2011).
- **Alineamiento progresivo**: se empieza alineando las parejas de secuencias más semejantes.
- El **árbol guía** que determina el orden de incorporación de secuencias en el alineamiento se obtiene de una comparación inicial de todos los pares de secuencias.

http://www.clustal.org/

MUSCLE



https://www.drive5.com/muscle/manual/index.html

Otros programas

Nombre	Descripción	Licencia
DECIPHER	Alineamiento global, progresivo e iterativo en R	GPL
MAFFT	Alineamiento local y global, progresivo e iterativo	BSD
T-Coffee	Alineamiento local y global, progresivo. Puede usar estructura 3D. Evalúa el alineamiento	GPL2