BIOINF-2020
Class online
LabMMBA

Alineamientos de secuencias (Pairwise alignments)

BCH441 - Bioinformatics
Boris Steipe
http://steipe.biochemistry.utoronto.ca/abc/index.php/Bioinformatics_Main_Page

http://circos.ca/

Buenas prácticas en **Bioinformatics**

¿Cómo llevan a diario el seguimiento de sus experimentos de wet-lab?

¿Qué prácticas tienen regularmente en el laboratorio?

¿Cómo hacer esto en análisis bioinformáticos?

Necesitamos saber el tipo de sistema operativo que utilizan (Linux/Windows/macOS/otro)

Alineamiento de secuencias

¿Qué es una secuencia?

Una secuencia tiene: composición; largo; dirección; orden..

Un gen, un transcrito, una proteína, un fragmento de una de las anteriores, una **estructura**, una **función**

DNA

CAAGTTGGATYNNTAC

RNA

AUGCAUCRCGCNUG

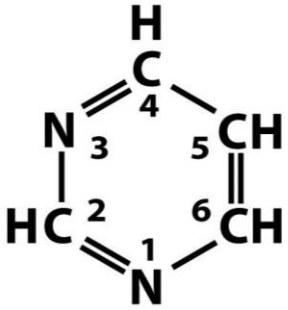
Protein

YGAXVLPMNGART

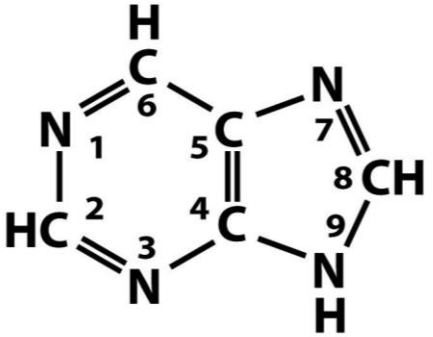
Alineamiento de secuencias

Nucleic acids

(5’ -> 3’)



Pyrimidine



Purine

A = Adenine
C = Cytosine
G = Guanine
T = Thymine
U = Uracyl

R = G A (puRine)
Y = T C (pYrimidine)
K = G T (Keto)
M = A C (aMino)
S = G C (Strong bonds)
W = A T (Weak bonds)

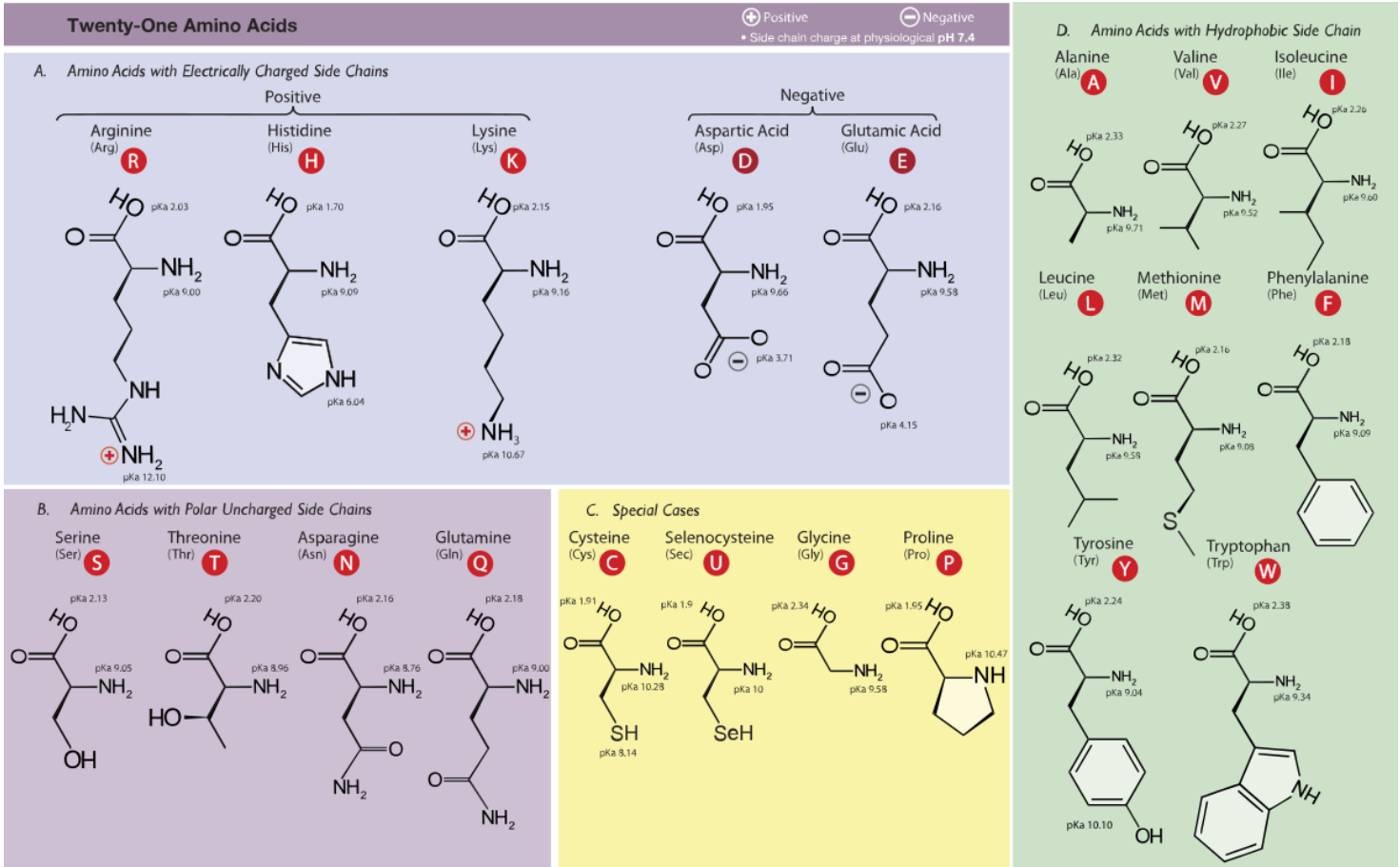
B = G T C (B is not A)
D = G A T (D is not C)
H = A C T (H is not G)
V = G C A (V is not T)

N = A G C T (aNy)

Alineamiento de secuencias

Proteins

(N-terminus -> C-terminus)



Alineamiento de secuencias

Proteins

B: D/N

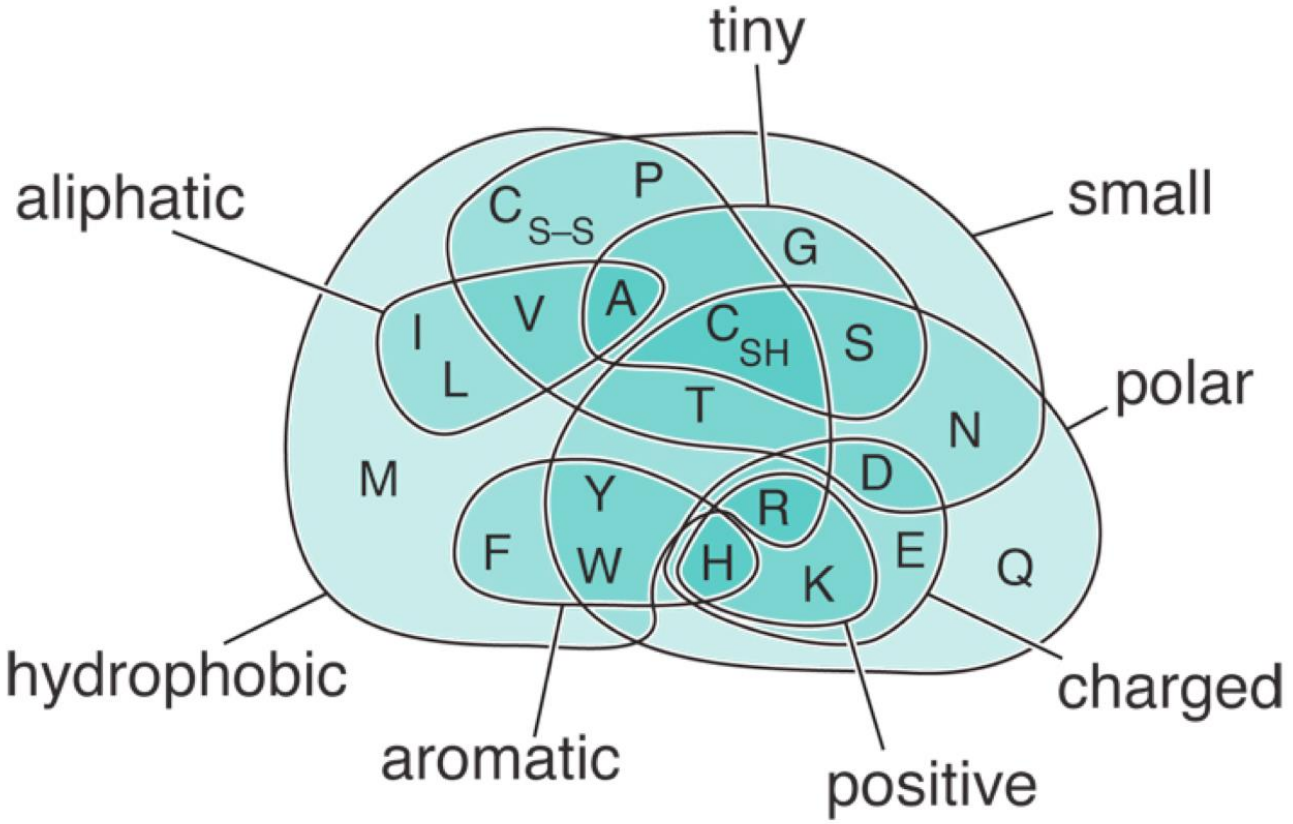
J: I/L

O: pyrrolysine

U: selenocysteine

X: unknown

Z: E/Q



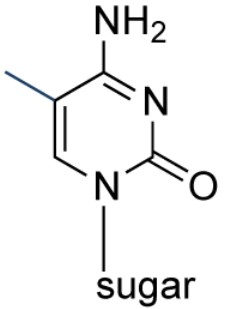
Propiedades fisicoquímicas importantes para identificar similitud

Alineamiento de secuencias

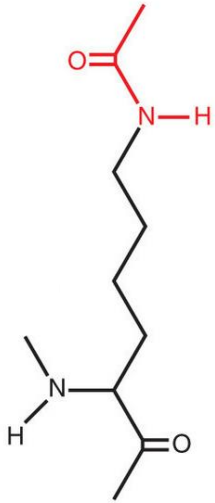
La representación de biomoléculas en secuencias es fundamental para la bioinformática, pero hay **limitaciones que considerar**!

1. Ambigüedad de secuencias (N, X)
2. En general, no representan modificaciones (metilaciones, acetilaciones, etc..)
3. La secuencia no almacena la información de anotación

```
CDS             /old_locus_tag="CTTA_0002"
                complement(1175..2188)
                /locus_tag="CTTA_RS00015"
                /old_locus_tag="CTTA_0002"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:WP_012203330.1"
                /note="Derived by automated computational analysis using
                gene prediction method: Protein Homology."
                /codon_start=1
                /transl_table=11
                /product="tripartite tricarboxylate transporter substrate
                binding protein"
                /protein_id="WP_149354226.1"
```



5-methylcytosine (mC)



Acetyl-lysine

Alineamiento de secuencias

Convenciones generales de secuencias:

1. código de nucleótidos IUPAC;
2. código de aminoácidos IUPAC;
3. 5’- a 3’-, o N- a C-;
4. Para genomas, la dirección de la hebra debería estar en el sentido de la replicación (bacterias), o de acuerdo a su posición en el cromosoma (desde el telómero de brazo corto es 5’->)
5. Unidades: bp o pb (basepair o pares de bases), Kb (10^3 bp), Mb (10^6 bp), Gb (10^9 bp).

Alineamiento de secuencias

¿Qué es un alineamiento de secuencias y para qué sirve?

1. Comparación entre dos (o más) secuencias
2. Aproximación a homología de secuencias (ortólogos/parologos)

¡Casi todos los análisis bioinfomáticos son derivados de alguna manera desde inferencias basadas en homología de secuencias!

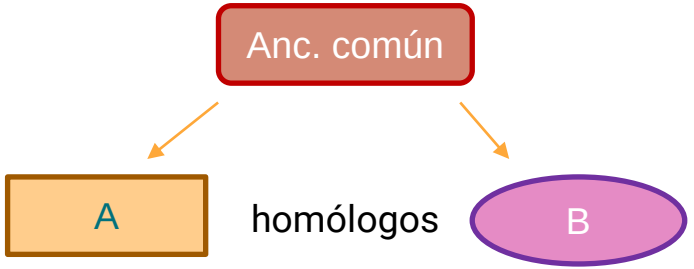
Alineamiento de secuencias

Homología vs Identidad vs Similitud

Dos genes son homólogos si divergieron de un ancestro común

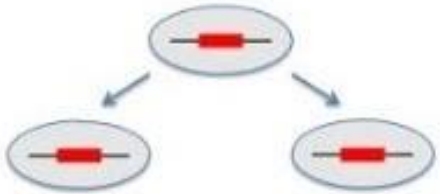
Homología es una cualidad

No existe algo altamente homólogo o 30% homólogo

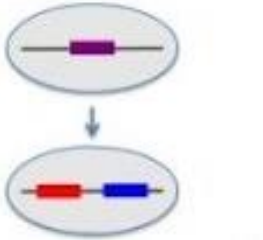


Dos genes son o no son homólogos, encontrando secuencias ortólogas o parálogas

Ortólogos



Parálogos



Alineamiento de secuencias

Homología vs Identidad vs Similitud

Dos genes son homólogos si divergieron de un ancestro común

Próteínas o genes homólogos, en general, tienen una estructura similar y usualmente poseen una función similar (hay que tener cuidado con esto)

Para identificar homología recurrimos a similitud de secuencias.

Parámetros de identidad (identity), similitud (similarity) y puntaje (scoring) mediante alineamiento de secuencias, además de otros parámetros!

(localización de residuos del sitio activo, estructura secundaria, patrones de residuos funcionales conservados, anotación de dominios proteicos)

Alineamiento de secuencias

Teniendo lo anterior en consideración y entrando de lleno en alineamientos:

1. ¿Qué tipo de alineamiento es necesario? Pairwise? Local? Global? Multiple?
2. ¿Qué sistema de puntaje será usado (scoring)?
3. ¿Cúal algoritmo debo usar?
4. ¿Qué método estadístico será usado para evaluar la significancia de un alineamiento?

Alineamiento de secuencias

Pairwise sequence alignments (PSA)

n = 2

```
  Q K E S G P S S S Y C
  |   | | |           |
V Q Q E S G L V R T T C
```

BLAST
EMBOSS (PSA)

Multiple sequence alignments (MSA)

n > 2

```
  Q K E S G P S S S Y C
  |   | | |           |
V Q Q E S G L V R T T C
  |   | |   | | |   | |
V Q K E S L L V R S T C
```

MAFFT
MUSCLE
ClustalW
Translator-X

Alineamiento de secuencias

Cuando comparamos secuencias estamos buscando evidencia de que hayan divergido de un ancestro común. Para esto, se deben alinear las secuencias considerando los posibles modelos evolutivos que hayan causado su divergencia. En este proceso podemos encontrar substituciones, inserciones o deleciones (indels).

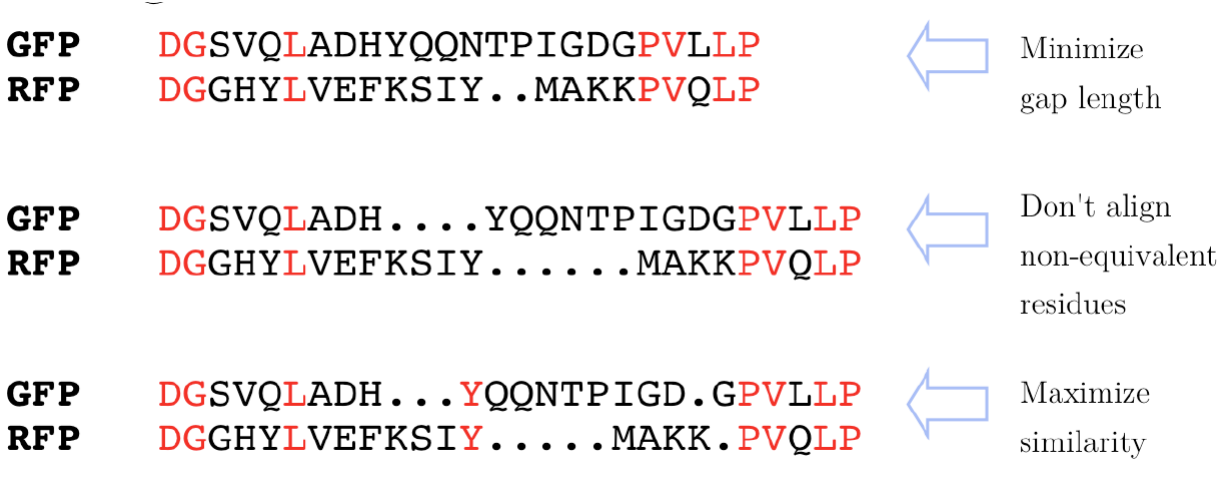
Un alineamiento es una forma de identificar regiones de similitud entre las secuencias.

```
GFP    DGSVQLADHYQQNTPIGDGPVLLP
RFP    DGGHYLVEFKSIY..MAKKPVQLP

GFP    DGSVQLADH....YQQNTPIGDGPVLLP
RFP    DGGHYLVEFKSIY......MAKKPVQLP

GFP    DGSVQLADH...YQQNTPIGD.GPVLLP
RFP    DGGHYLVEFKSIY.....MAKK.PVQLP
```

Pero, ¿cuál alineamiento es el correcto?

Alineamiento de secuencias

Para determinar cuales de los nucleótidos o residuos son causa de substitución o indels, se debe elegir una matriz de substitución que permite dar un valor numérico al alineamiento.



En alineamientos de secuencias los indels son referidos como "***gaps***".

Definiendo dos eventos:
"*gap opening*": aparición de un indel

"*gap extension*": extensión del indel en un carácter

Alineamiento de Secuencias

Matrices de substitución o puntuación

(Substitution or scoring matrices)

Estos definen la tasa de substitución de cada residuo en una secuencia a través del tiempo, entregando un valor numérico a cada mutación puntual definida.

BLOSUM62 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

PAM250 matrix

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

Alineamiento de Secuencias

Algoritmos de alineamiento

Alineamiento global (global alignment)

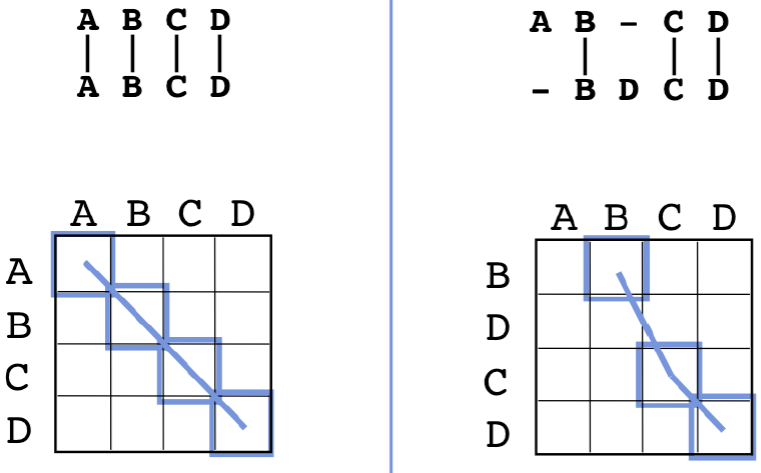
Premisa: El puntaje total del alineamiento depende de la suma de todos los puntajes en pares de caracteres, menos una penalización por cada indel y su extensión

El puntaje más alto posible es extendido en la mejor forma posible cada un residuo o nucleótido (carácteres)

Alineamiento de Secuencias

Needleman-Wunsch algorithm (1970):

Cualquier alineamiento puede ser representado como un camino en una matriz que conecta cada intersección de columnas y filas de dos caracteres alineados



Alineamiento de Secuencias

Needleman-Wunsch algorithm (1970):

El alineamiento óptimo esta dado por el camino que lleva a la suma más alta posible de todas las sumas de puntajes de a dos que contiene

First step: compile all pairwise scores into a matrix.

	A	B	C	D
B	0	**1**	0	0
D	0	0	0	**1**
C	0	0	**1**	0
D	0	0	0	**1**

Alineamiento de Secuencias

Needleman-Wunsch algorithm (1970):

El alineamiento óptimo esta dado por el camino que lleva a la suma más alta posible de todas las sumas de puntajes de a dos que contiene

Second step: The highest score in the last column and row is the is highest pairscore we put there from the scoring matrix. This is the Base Case, if we think about the recursion, because there is no previous score we had to consider.

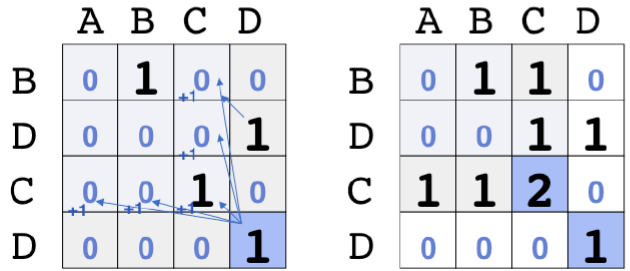
	A	B	C	D
B	0	**1**	0	0
D	0	0	0	**1**
C	0	0	**1**	0
D	0	0	0	**1**

Alineamiento de Secuencias

Needleman-Wunsch algorithm (1970):

El alineamiento óptimo esta dado por el camino que lleva a la suma más alta posible de todas las sumas de puntajes de a dos que contiene

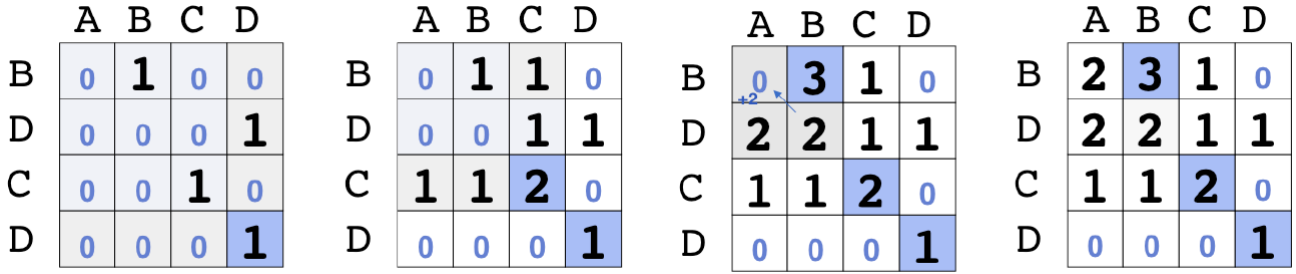
Third step: Extend the path. Assign to each cell of the next column and row the highest value we can get by adding to its current value a value from a previous cell **that could be part of an alignment path.**



Alineamiento de Secuencias

Needleman-Wunsch algorithm (1970):

El alineamiento óptimo esta dado por el camino que lleva a la suma más alta posible de todas las sumas de puntajes de a dos que contiene



A B - C D
| | |
- B D C D

(This example assumes: identity matrix - match=1, mismatch=0, no gap penalties)