

Tp 3

Alineamientos de a pares



Universidad
Nacional
de San Martín

CONICET



Objetivos y ejercicios del TP

Objetivos:

1. Entender cómo funciona el algoritmo de alineamiento de pares de secuencias de Needleman-Wunsch.
2. Aprender a interpretar un Dot-Plot
3. Comprender los conceptos de identidad, similitud y homología de secuencias

Ejercicios:

1. Realizar el alineamiento entre dos secuencias cortas de forma manual
2. Analizar el *dot-plot* generado a partir de un alineamiento de secuencias
3. Determinar qué especies están más relacionadas entre caballo, ballena enana y canguro rojo

Introducción

¿Qué es alinear de a pares?

Es la asignación uno-a-uno de correspondencias entre los elementos que componen dichas secuencias sin alterar su orden.

- **Match (M)**: Cuando los elementos enfrentados son equivalentes.
- **Mismatch (m)**: Cuando los elementos correspondientes son diferentes.
- **Gap (g)**: Cuando un elemento de una secuencia no tiene par en la otra y se enfrenta a un espacio, caracterizado por un guión (-).
 - Gap open: Cuando se abre un gap.
 - Gap extend: Cuando se agregan gaps a continuación de otro gap.

¿Para qué alineamos?

Para obtener una medida de similitud entre secuencias $\Rightarrow$ necesitamos un sistema de puntaje o *scoring*

Introducción

Algoritmos de alineamiento

- **Alineamientos globales:** Se aparean todos los elementos de una secuencia con todos los elementos de la otra. Este tipo de alineamientos se utiliza principalmente para comparar dos secuencias que son similares en longitud. También se denominan alineamientos de Needleman-Wunsch por sus creadores
- **Alineamientos locales:** (o de Smith-Waterman), parean únicamente parte de las secuencias y son útiles para identificar, por ejemplo, dominios en común.
- **Alineamientos mixtos:** Combinan los dos anteriores.



Dynamic programming

Es un algoritmo que utiliza matrices para alinear dos secuencias. Los pasos para completar esta matriz son:

1. Comenzar en el extremo superior izquierdo, con un puntaje inicial de 0.
2. Calcular el costo que tiene aparejado desplazarse de una celda a la otra para cada paso, dado el sistema de puntajes pre-establecido.
3. Elegir la opción más favorable, es decir aquella que maximice el puntaje global del alineamiento.
4. Guardar el puntaje con el que se llegó a una celda dada y el movimiento que originó dicho camino o path, indicado típicamente con una flecha.

Una vez que la matriz está completa en su totalidad se puede recorrer hacia atrás o realizar un *traceback*, desde el extremo inferior derecho al superior izquierdo, para reconstruir el alineamiento.

Dynamic programming

Match: $M=1$

Mismatch: $m=-1$

Gap: $g=-2$

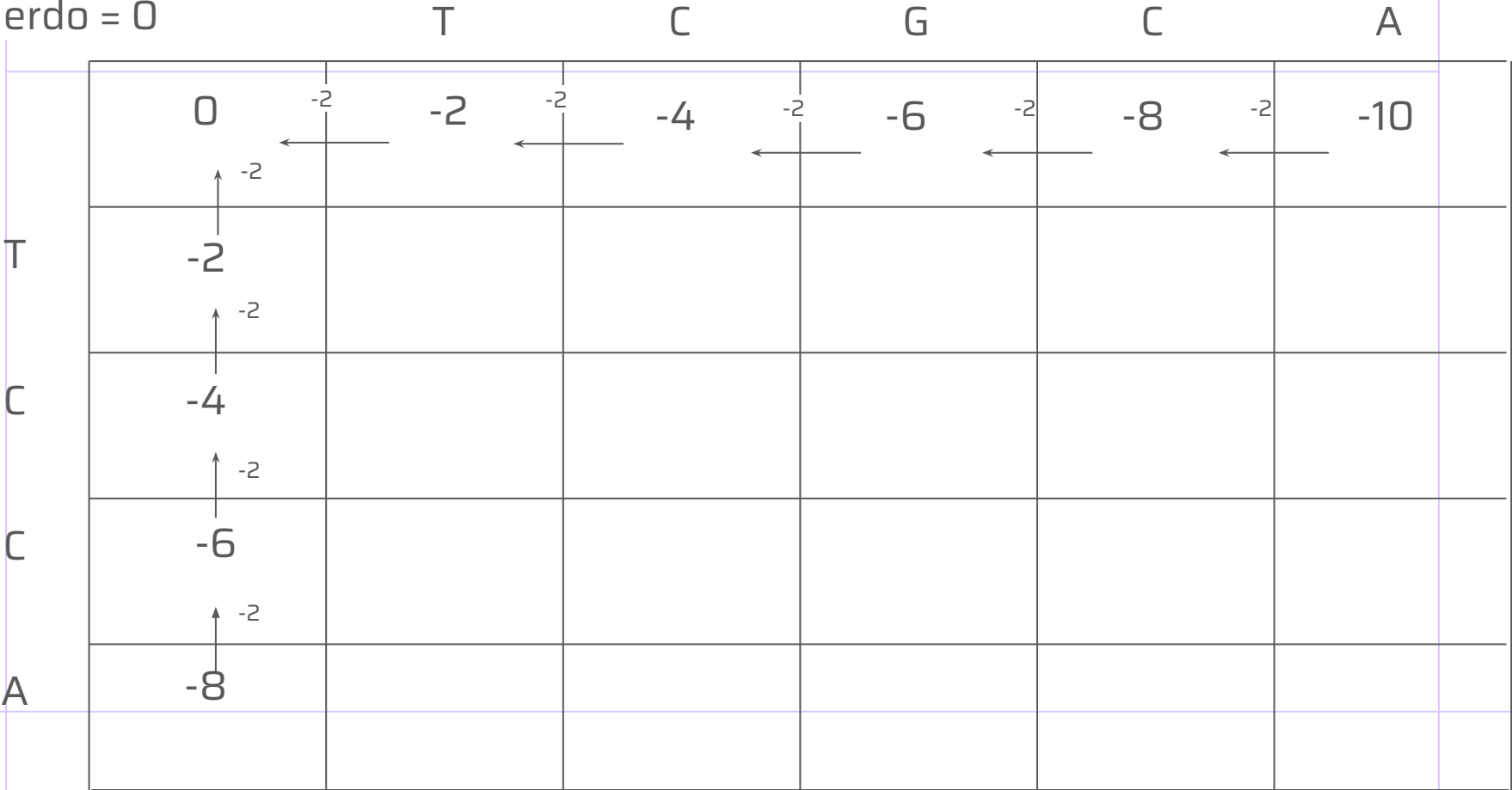
Primer paso:

Extremo superior

izquierdo = 0

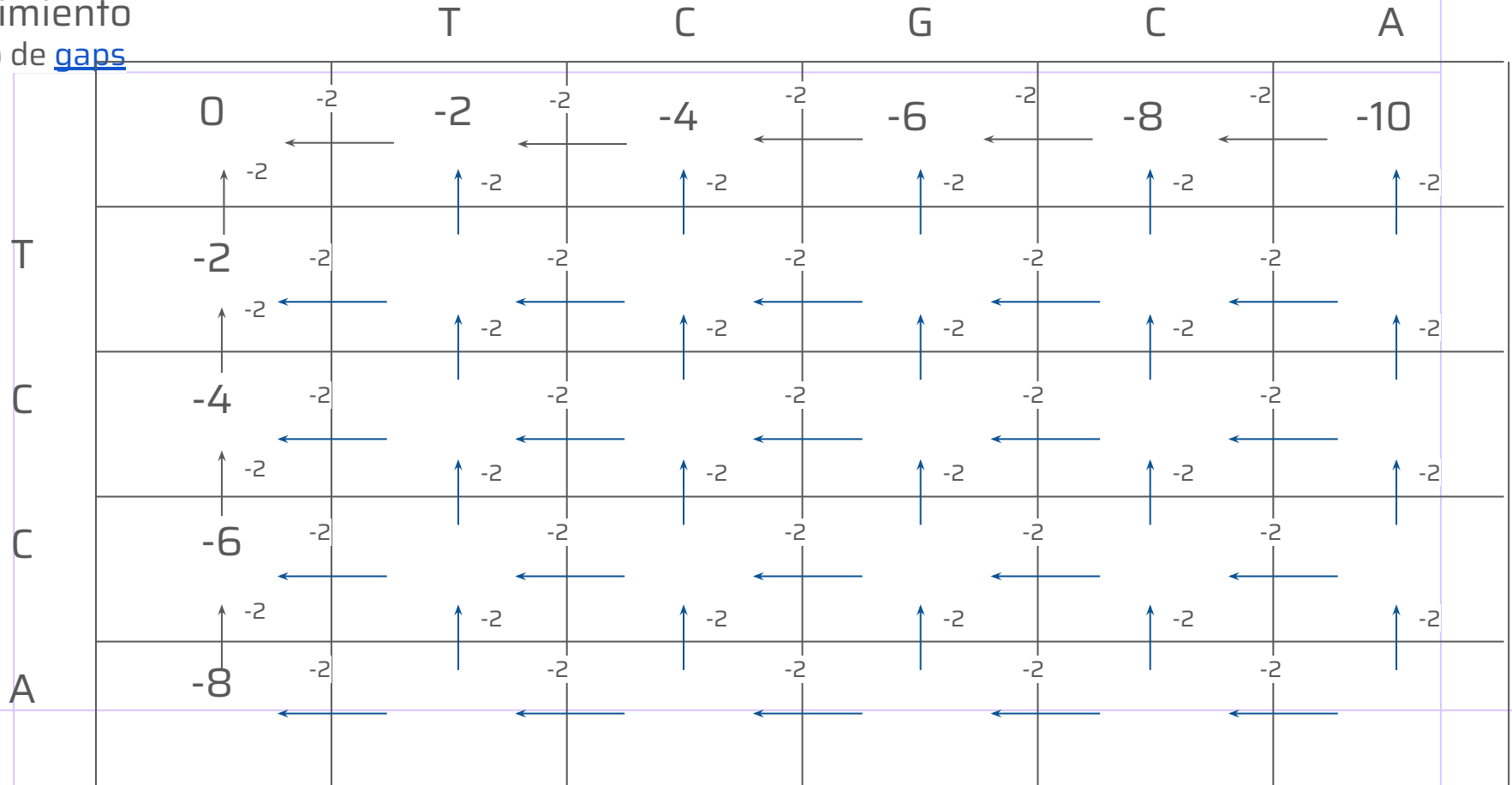
Dynamic programming

Match: M=1 Mismatch: m=-1
Gap: g=-2



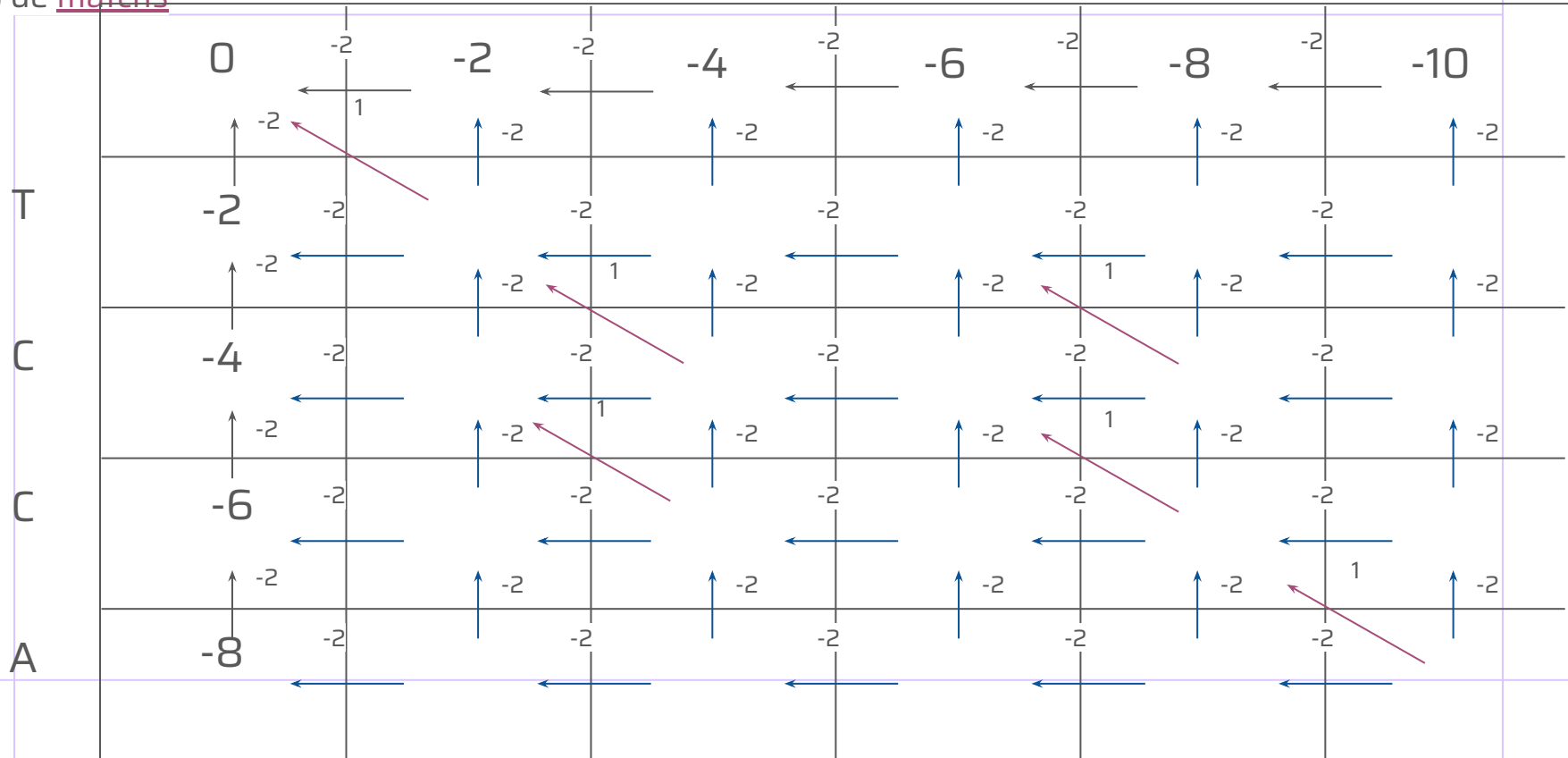
Gap: $g=-2$

Costo de gaps



Match: $M=1$ Mismatch: $m=-1$
Gap: $g=-2$

Costo de matches



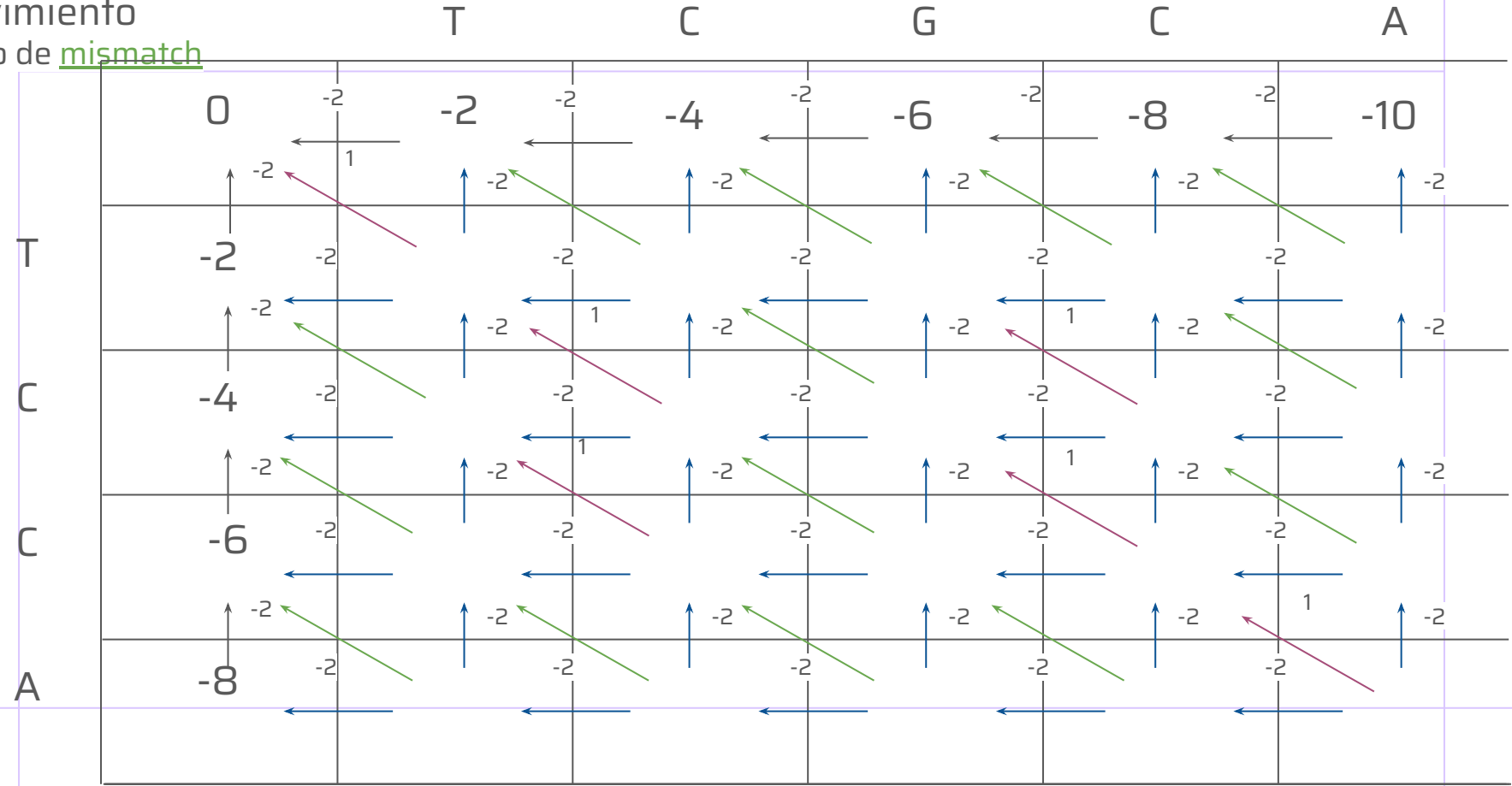
Segundo paso:

Dynamic programming

Match: $M=1$ Mismatch: $m=-1$
Gap: $g=-2$

Costos de cada
movimiento

Costo de mismatch

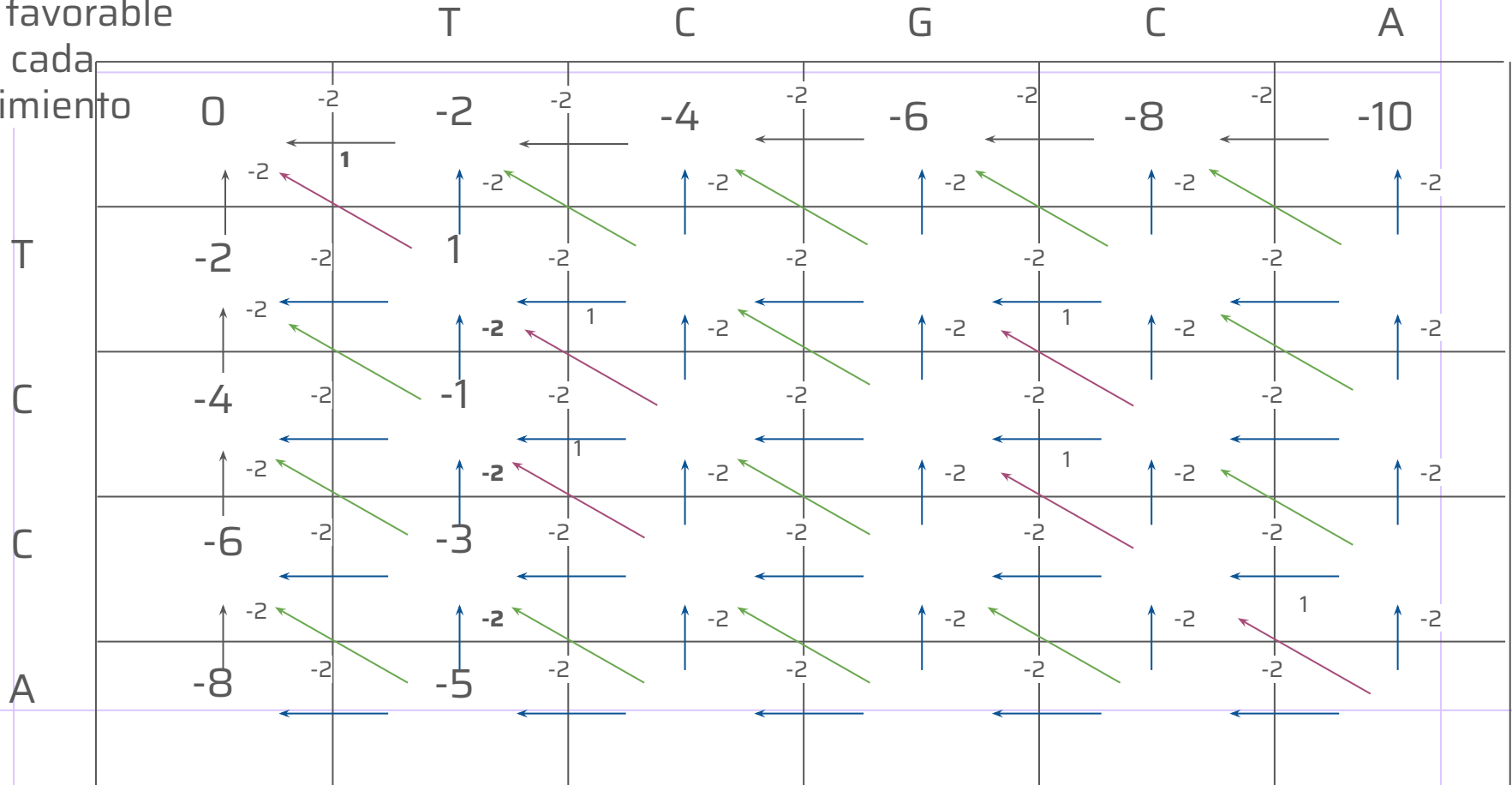


Tercer paso:

Elegir la opción
más favorable
para cada
movimiento

Dynamic programming

Match: $M=1$ Mismatch: $m=-1$
Gap: $g=-2$

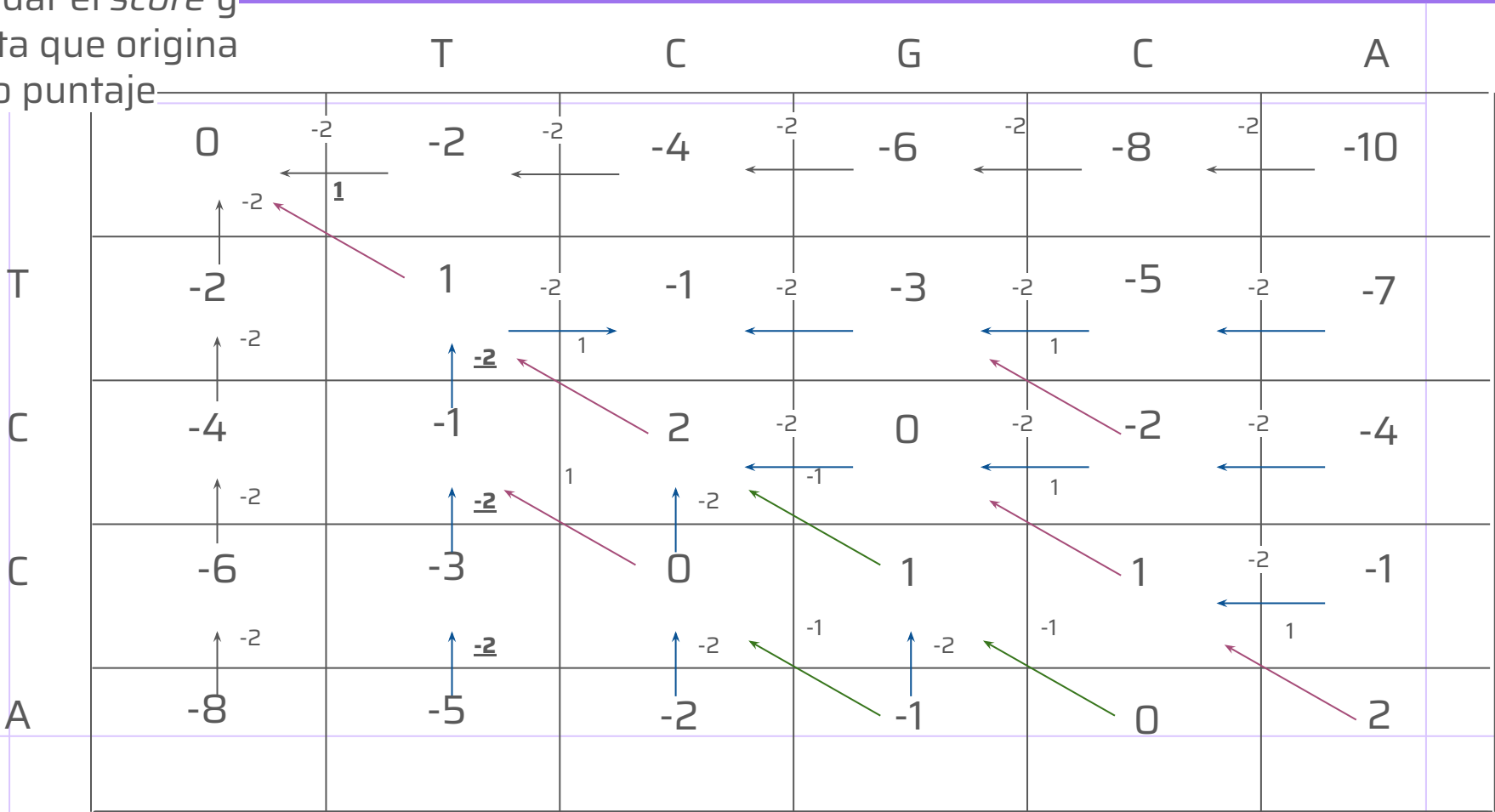


Cuarto paso:

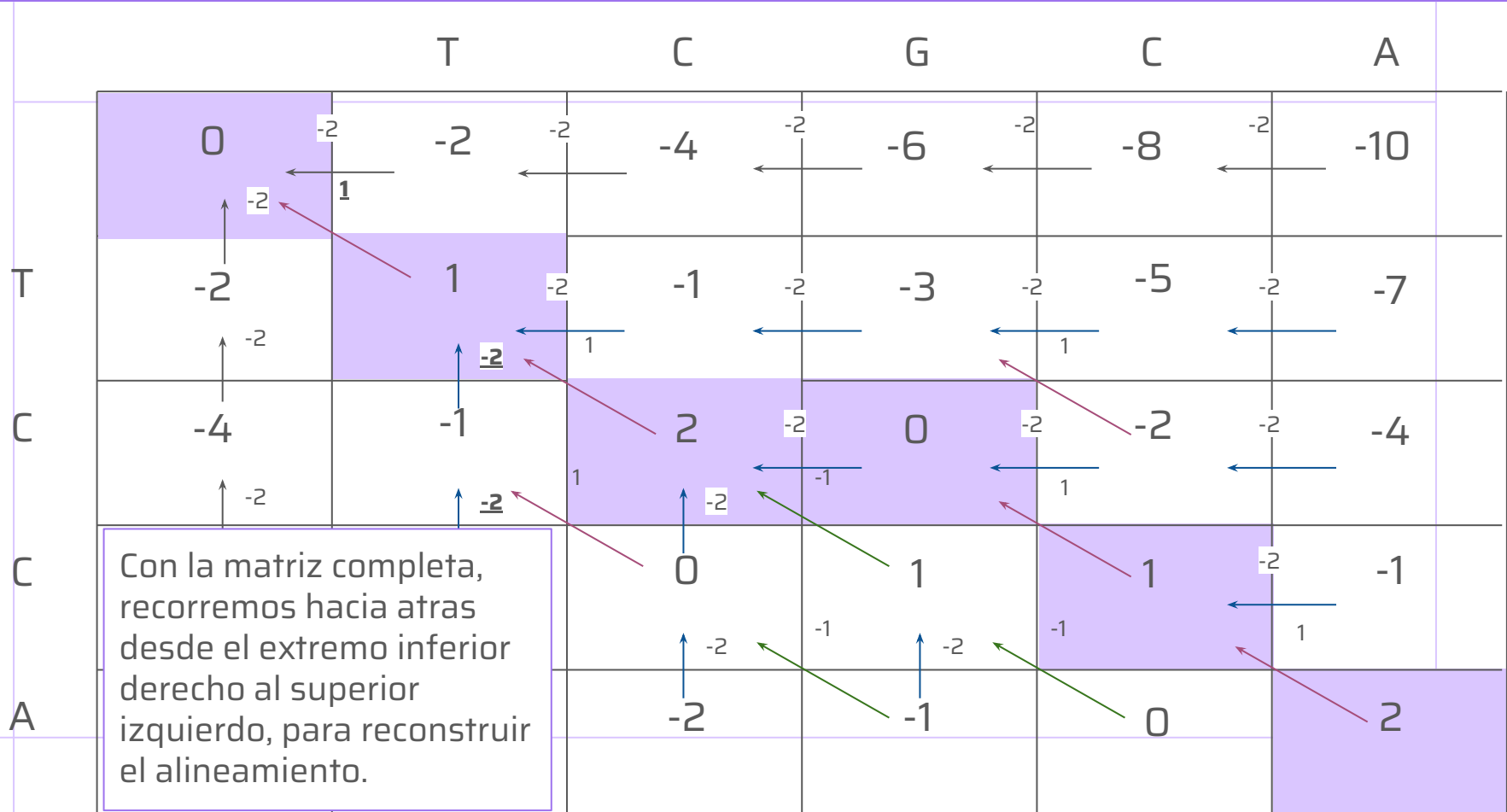
Guardar el *score* y la ruta que origina dicho puntaje

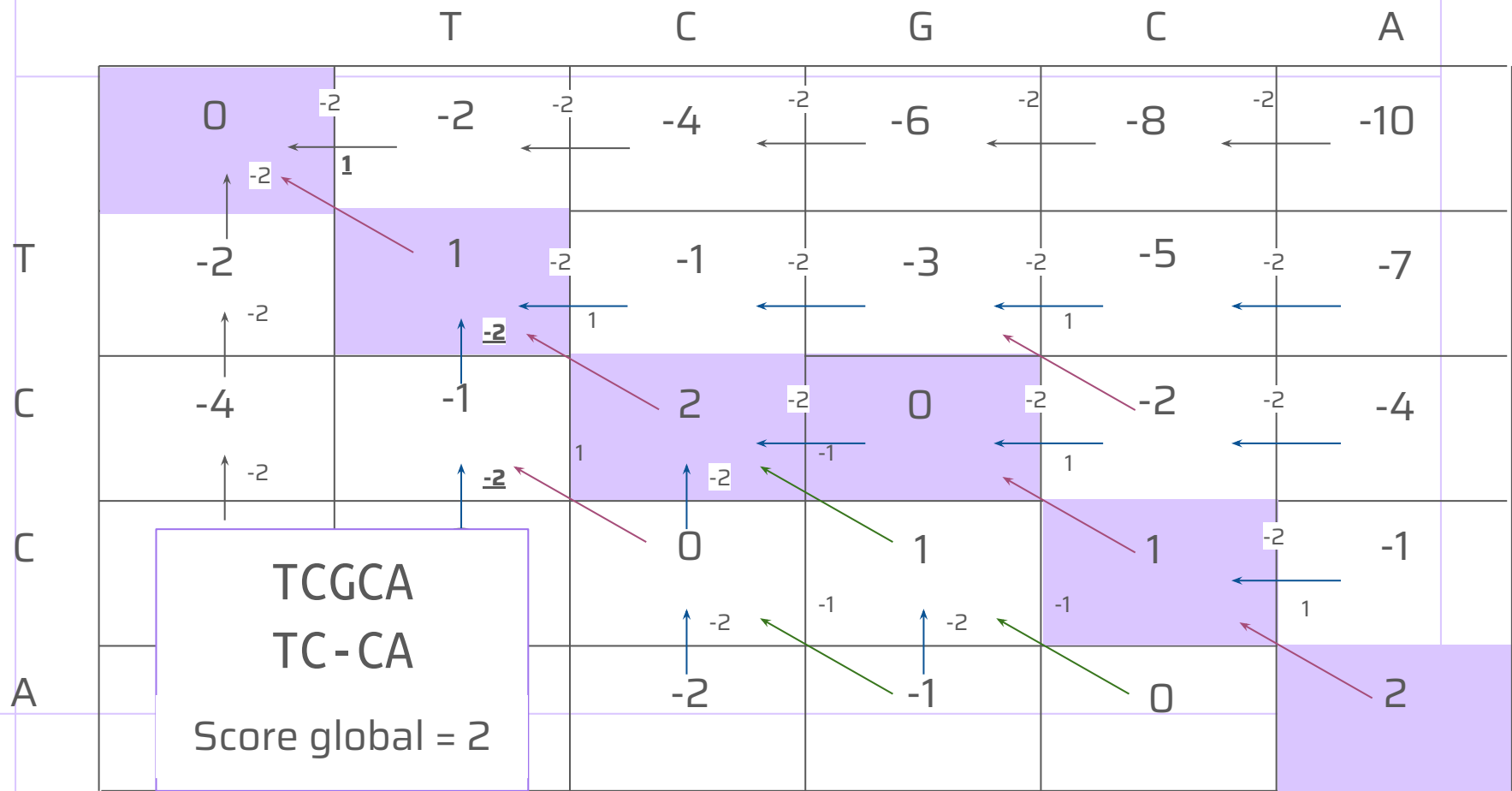
Dynamic programming

Match: M=1 Mismatch: m=-1
Gap: g=-2

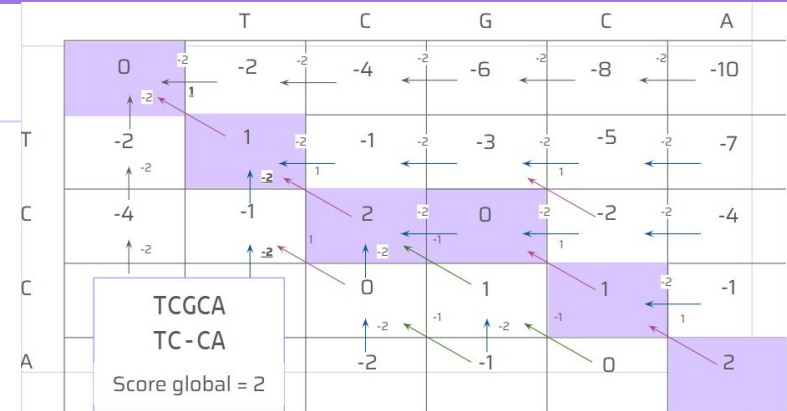


Dynamic programming



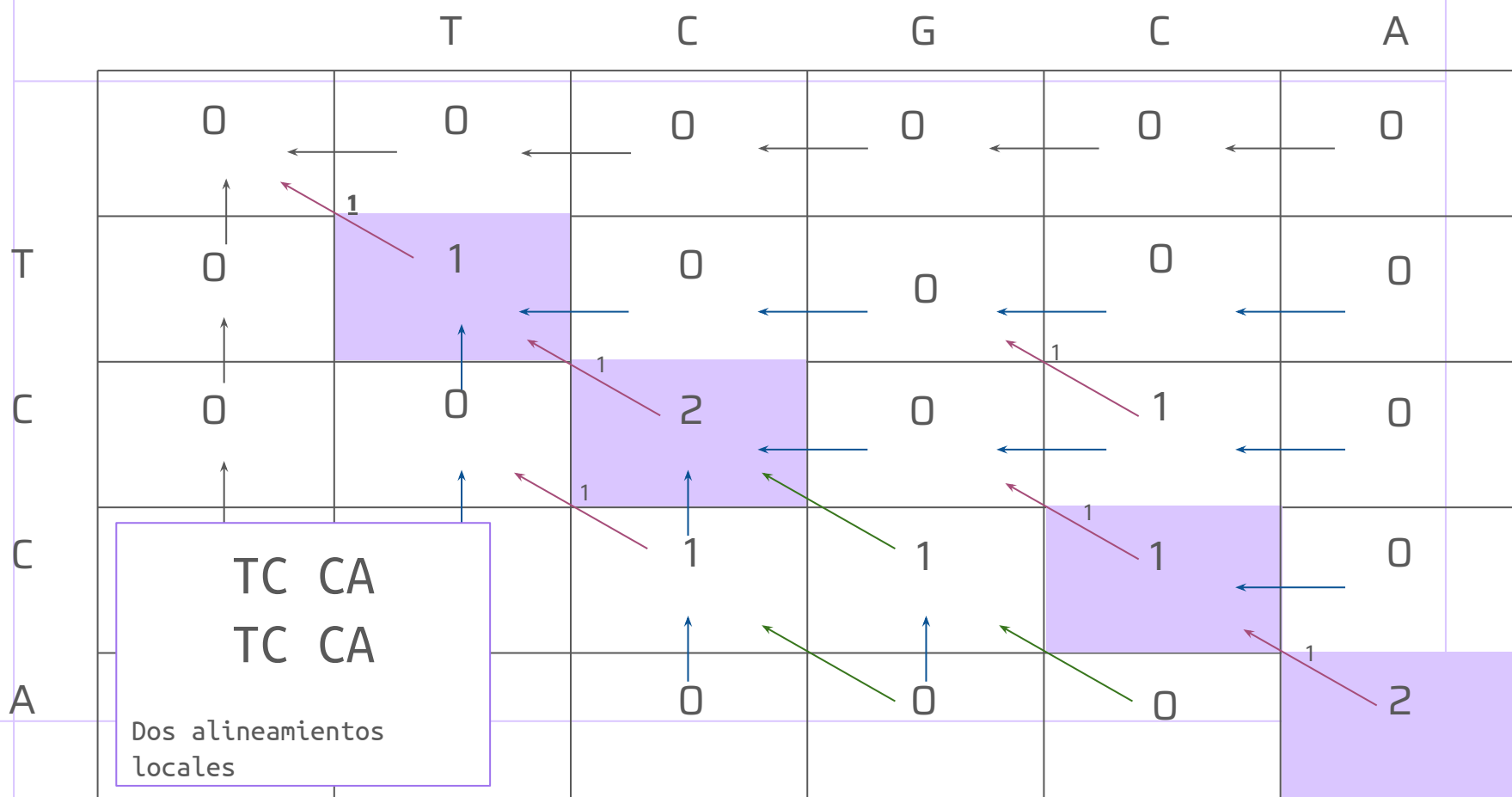


¿Y si fuese alineamiento local?

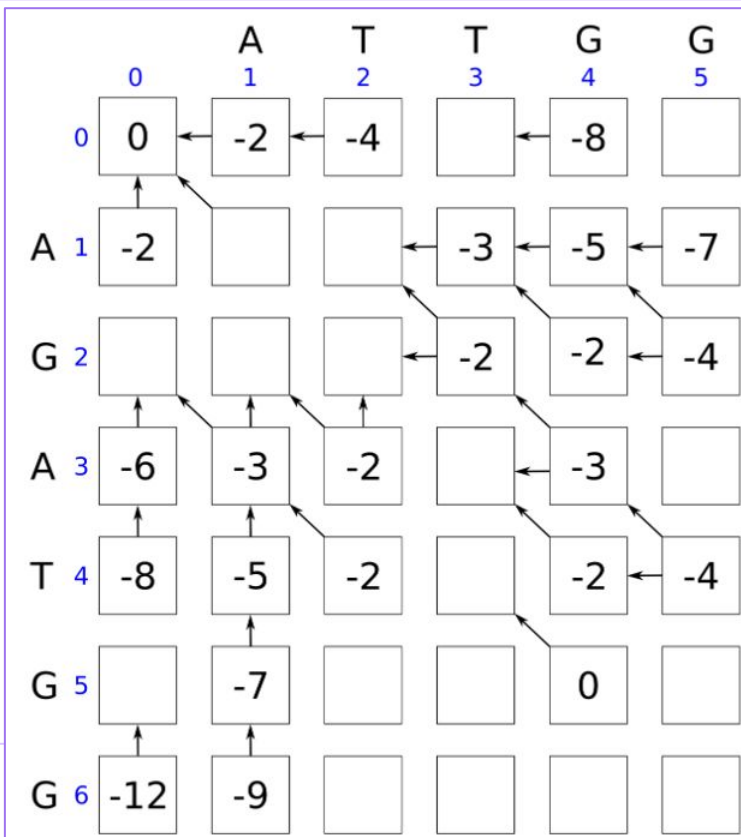


	Smith–Waterman algorithm	Needleman–Wunsch algorithm
Initialization	First row and first column are set to 0	First row and first column are subject to gap penalty
Scoring	Negative score is set to 0	Score can be negative
Traceback	Begin with the highest score, end when 0 is encountered	Begin with the cell at the lower right of the matrix, end at top left cell

¿Y si fuese alineamiento local?



Ejercicio 1: Dynamic programming

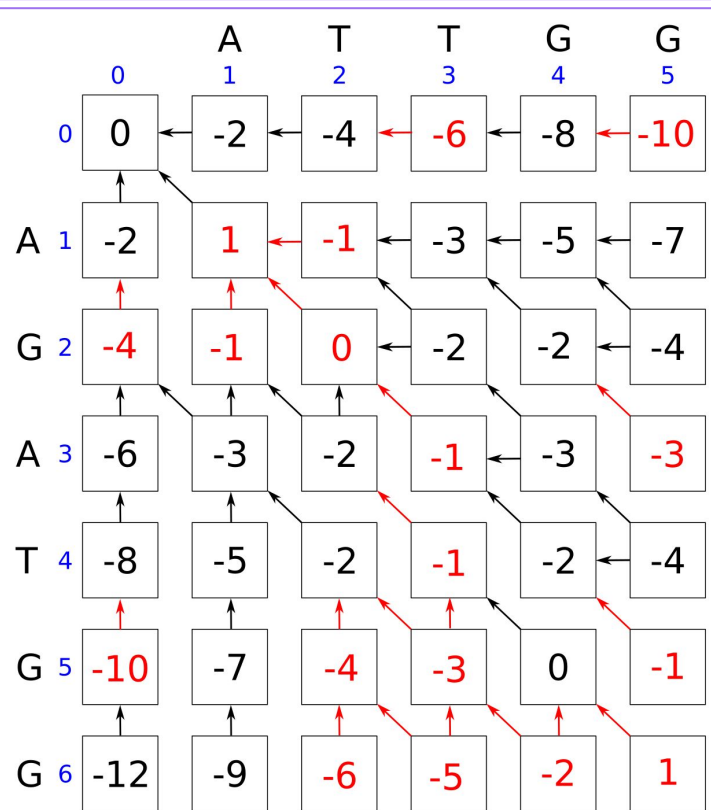


Alineamiento de las secuencias ATTGG con AGATGG, usando el esquema de puntajes: $M=1$, $m=-1$, $g=-2$

1. Extremo superior izquierdo = 0
2. Calcular el costo de desplazamiento para cada paso
3. Elegir la opción que maximice el puntaje global del alineamiento.
4. Guardar el puntaje con el que se llegó a una celda y el movimiento que originó dicho camino

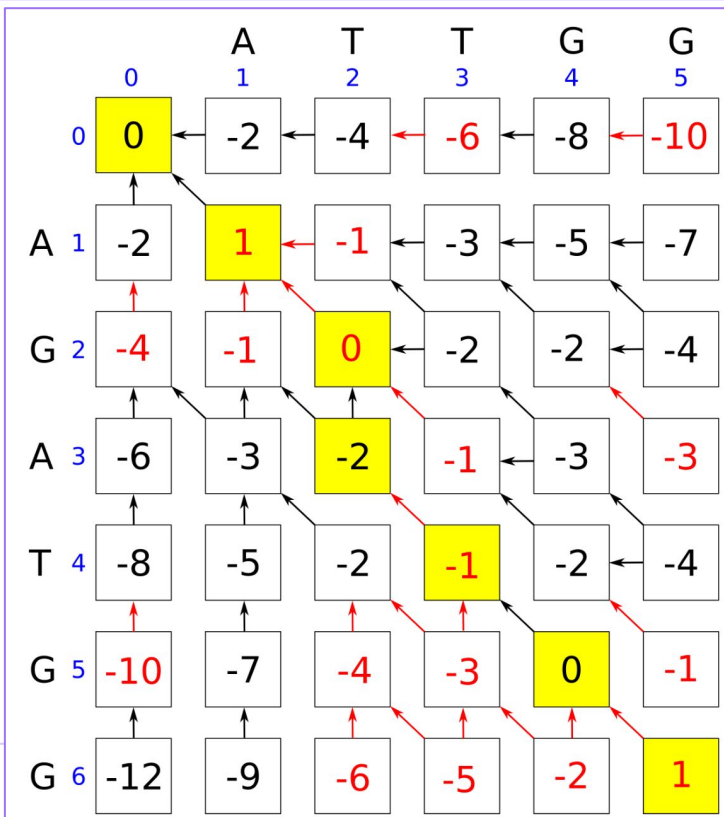
10' para pensar

Ejercicio 1: Dynamic programming



Alineamiento de las secuencias ATTGG con AGATGG, usando el esquema de puntajes: $M=1$, $m=-1$, $g=-2$

Ejercicio 1: Dynamic programming



Alineamiento de las secuencias ATTGG con AGATGG, usando el esquema de puntajes: $M=1$, $m=-1$, $g=-2$

Camino 1

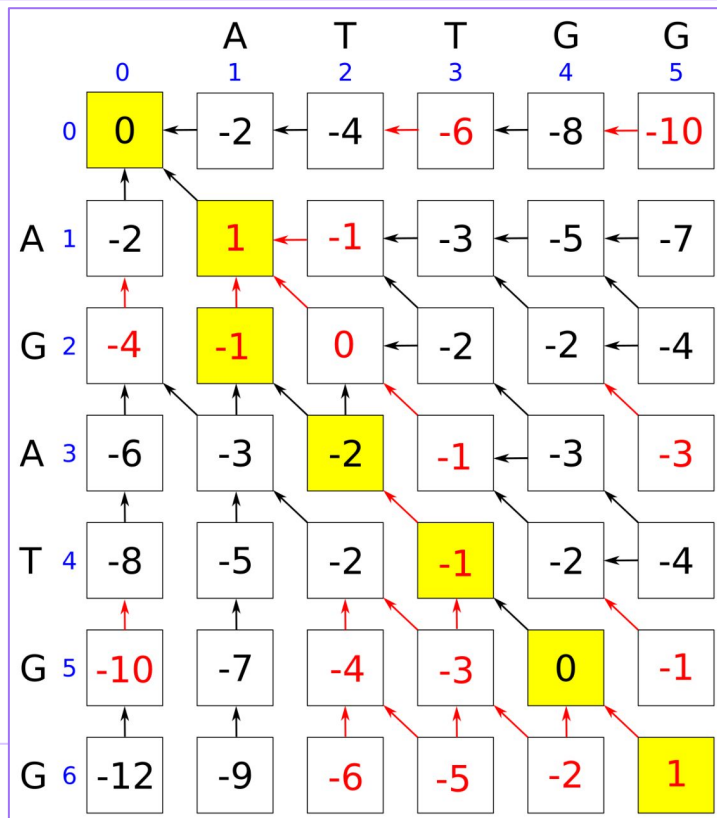
A T - T G G

| | | | |

A G A T G G

$$1 - 1 - 2 + 1 + 1 + 1 = 1$$

Ejercicio 1: Dynamic programming



Alineamiento de las secuencias ATTGG con AGATGG, usando el esquema de puntajes: $M=1$, $m=-1$, $g=-2$

Camino 2

A - T T G G

| | | | |

A G A T G G

$$1 - 2 - 1 + 1 + 1 + 1 = 1$$

Ejercicio 1: Dynamic programming

Alineamiento de dos secuencias usando la página. Qué son los colores verde y rosa. Verde = la celda seleccionada. Rosa = las celdas que están relacionadas con el score de la celda verde. Si hay mas de una, es porque las dos vías llevan al mismo score.

D		A ₁	T ₂	T ₃	G ₄	G ₅	G ₆
	0	-5	-10	-15	-20	-25	-30
A ₁	-5	1	-4	-9	-14	-19	-24
G ₂	-10	-4	0	-5	-8	-13	-18
A ₃	-15	-9	-5	-1	-6	-9	-14
T ₄	-20	-14	-8	-4	-2	-7	-10
G ₅	-25	-19	-13	-9	-3	-1	-6
G ₆	-30	-24	-18	-14	-8	-2	0
Score: 0							

Match 1 Mismatch -1 Gap -5

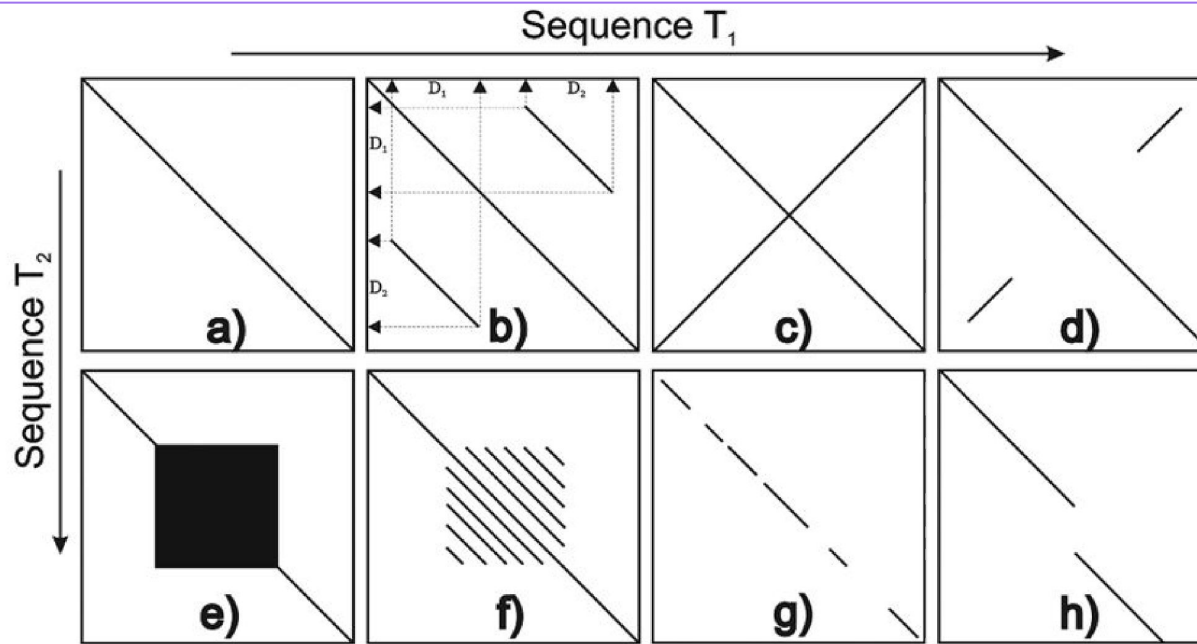
D		A ₁	T ₂	T ₃	G ₄	G ₅	G ₆
	0	-1	-2	-3	-4	-5	-6
A ₁	-1	1	0	-1	-2	-3	-4
G ₂	-2	0	-1	-2	0	-1	-2
A ₃	-3	-1	-2	-3	-1	-2	-3
T ₄	-4	-2	0	-1	-2	-3	-4
G ₅	-5	-3	-1	-2	0	-1	-2
G ₆	-6	-4	-2	-3	-1	1	0
Score: 0							

Match 1 Mismatch -5 Gap -1

Si se penaliza mucho el gap, hay un solo alineamiento con el mejor score (que es 0). Si se penaliza mucho el m pero no el gap, hay muchísimos alineamientos con muchos gaps adentro (para score 0 se permite un gap x match)

Dot plots

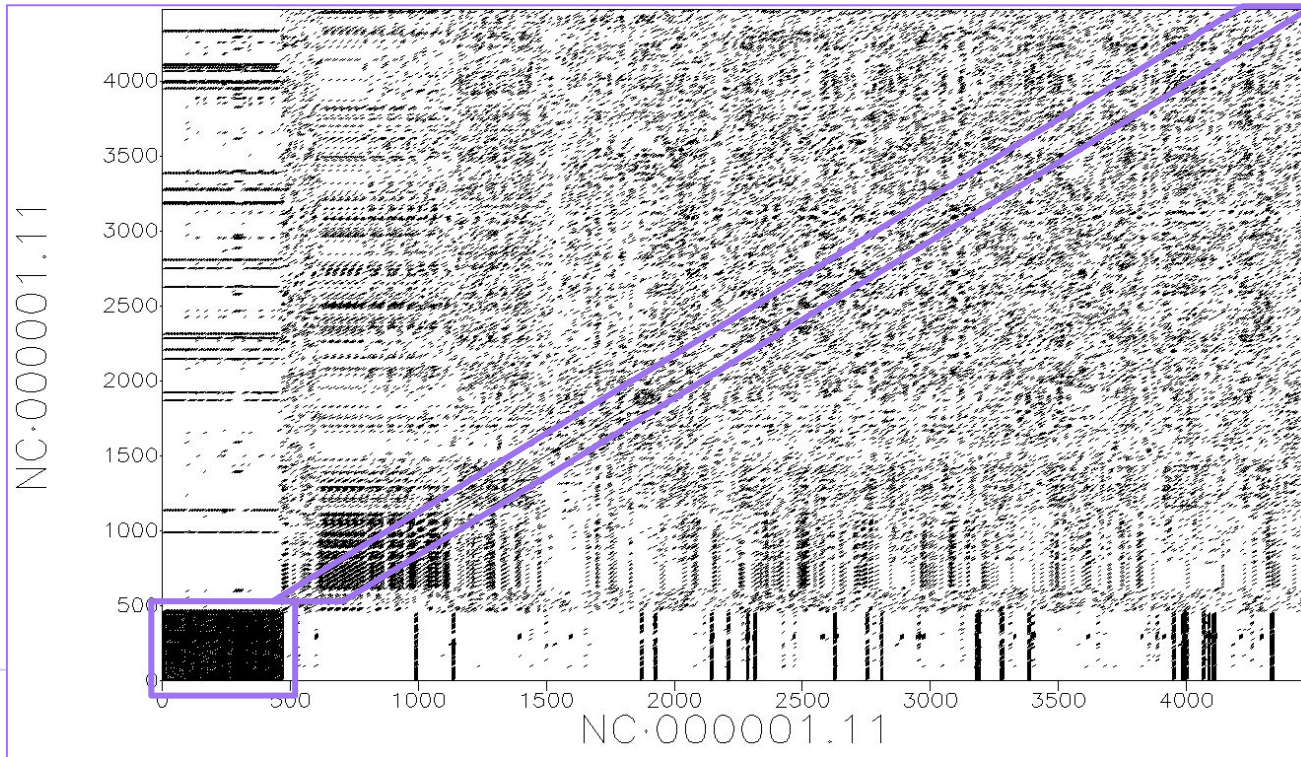
Los dot-plots son representaciones gráficas que dan un pantallazo sobre la similitud entre dos secuencias. En ellos se pueden identificar patrones que aporten información sobre la relación entre ambas secuencias.



- a) Match perfecto.
- b) Repeticiones.
- c) Palíndromo.
- d) Repeticiones invertidas.
- e) Zonas de baja complejidad (microsatelites).
- f) Zonas altamente repetitivas (minisatelites).
- g) Secuencias con alta conservación.
- h) Inserción o delección.

Ejercicio 2: Dot plots

Comparar la secuencia HS-ch11-fragment.fasta contra sí misma y observar los patrones del *dot plot*

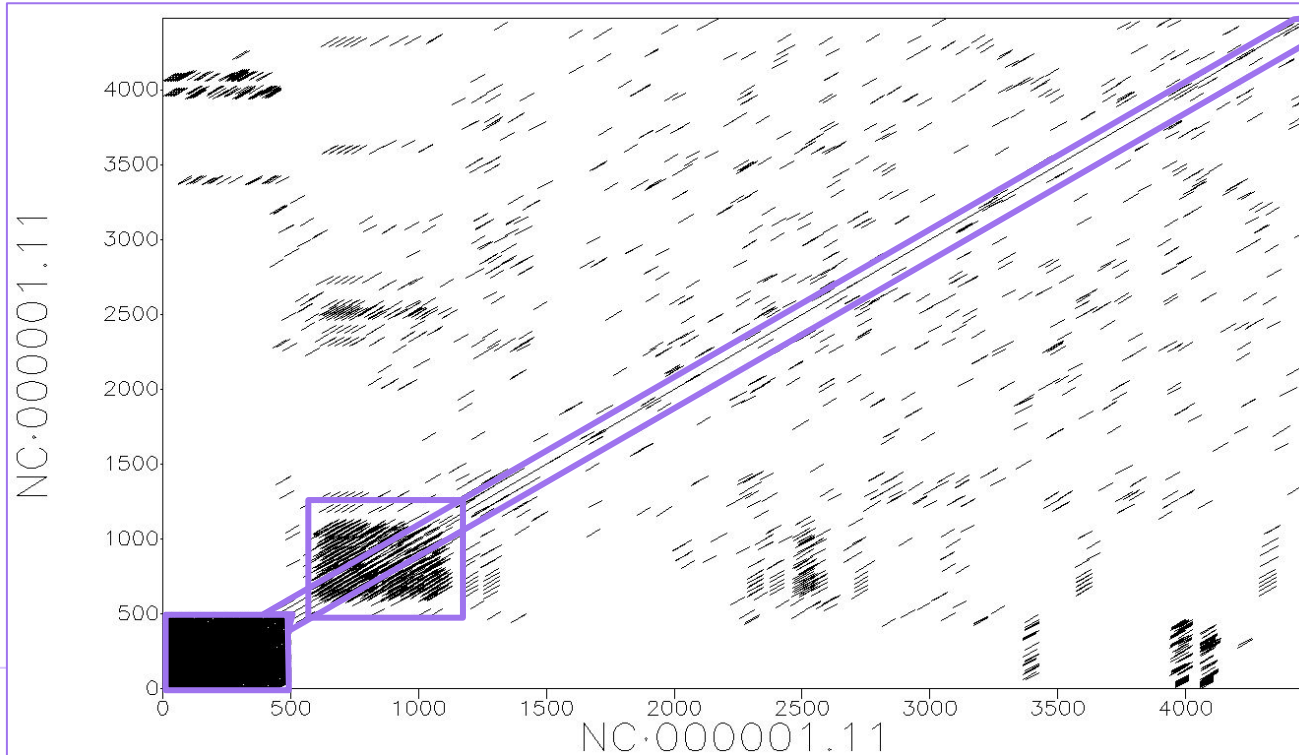


Se ven dos patrones y **mucho ruido**

⇒ Ajustemos los parámetros (por defecto *windowsize* 10 y *threshold* 23)

Ejercicio 2: Dot plots

Comparar la secuencia HS-ch11-fragment.fasta contra sí misma y observar los patrones del *dot plot*



⇒ Nuevos
parámetros:
window size 50 y
threshold 20

Microsatélites

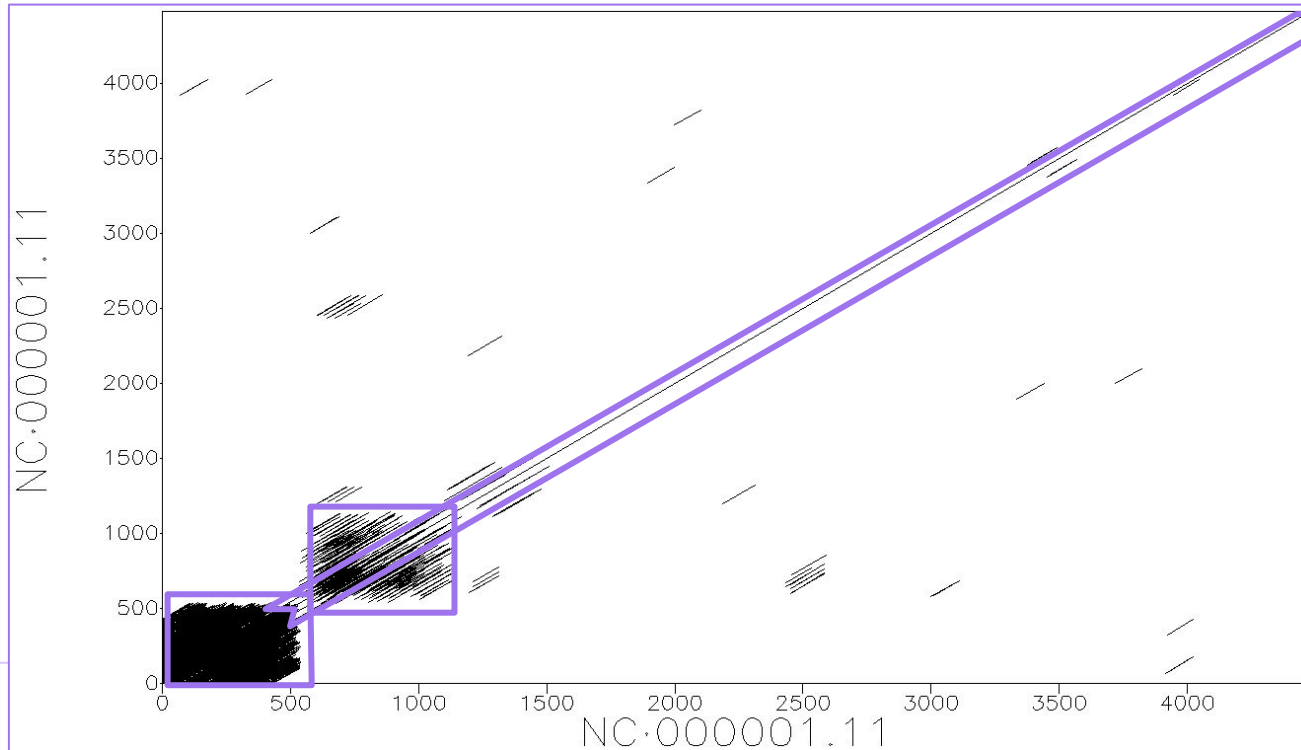
Minisatélites

Match perfecto

Repeticiones

Ejercicio 2: Dot plots

Comparar la secuencia HS-ch11-fragment.fasta contra sí misma y observar los patrones del *dot plot*



⇒ Nuevos
parámetros:
window size 100 y
threshold 20

Microsatélites

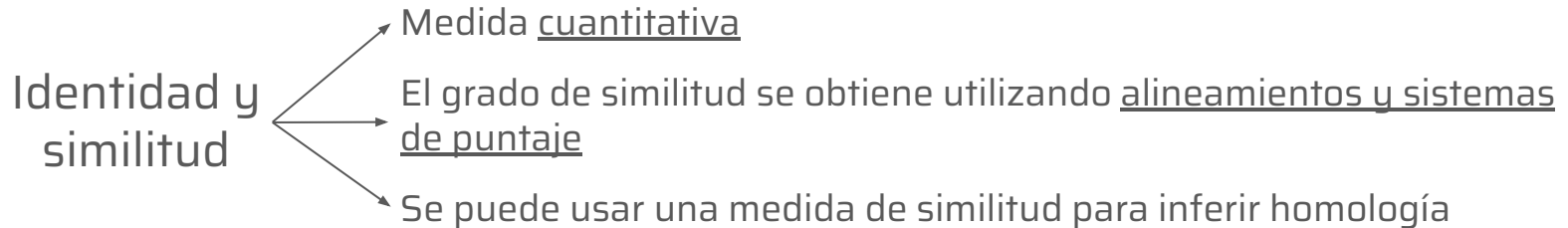
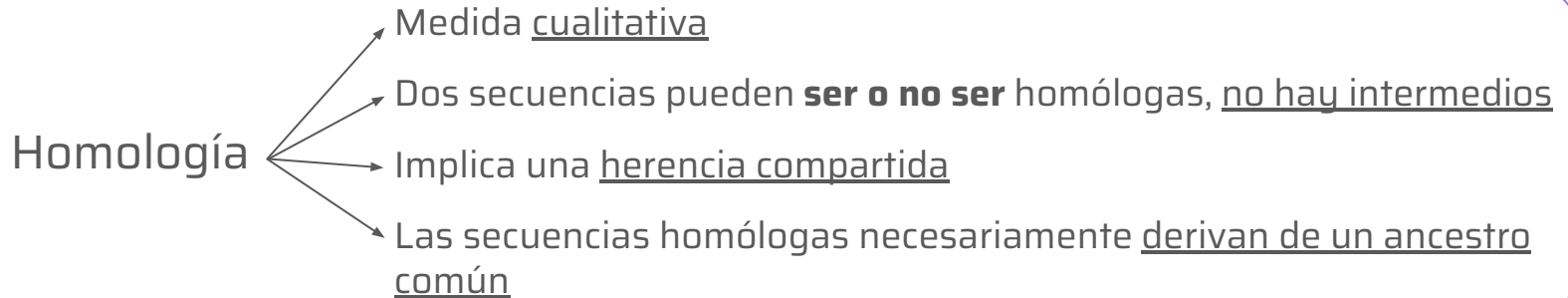
Minisatélites

Match perfecto

Repeticiones
muchas menos

Identidad, similitud y homología

No son sinónimos! Son **conceptos relacionados** con distinto significado



Ejercicio 3: Similitud y homología

Determinar qué especies están más relacionadas utilizando la ribonucleasa pancreática de caballo (*Equus caballus*), ballena enana (*Balaenoptera acutorostrata*) y canguro rojo (*Macropus rufus*).



```
# Aligned_sequences: 2
# 1: RNAS1_BALAC
# 2: RNAS1_MACRU
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 1.0
#
```



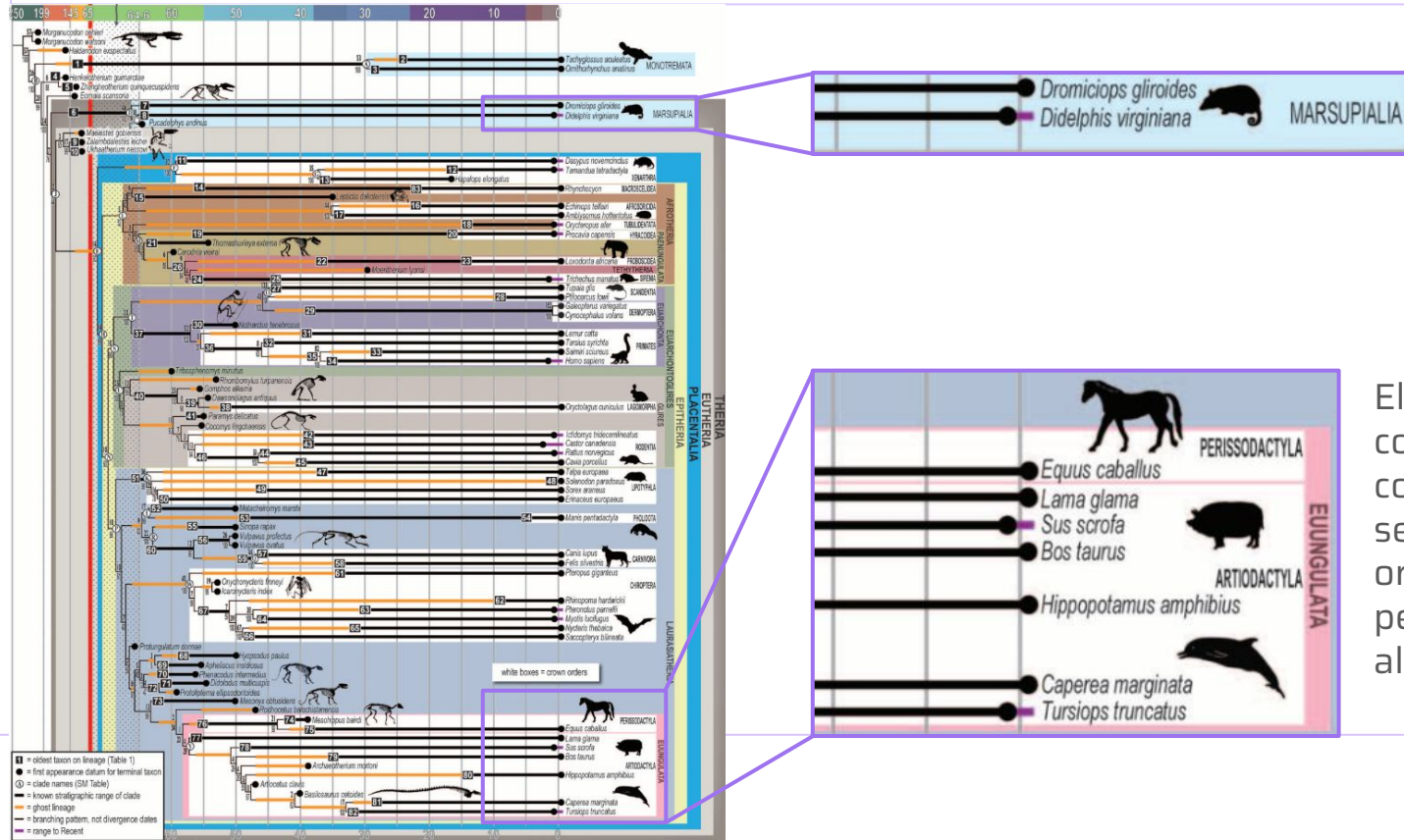
```
# Aligned_sequences: 2
# 1: RNAS1_HORSE
# 2: RNAS1_MACRU
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 1.0
#
```



```
# Aligned_sequences: 2
# 1: RNAS1_HORSE
# 2: RNAS1_BALAC
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 1.0
#
```

¿Qué secuencias son más similares? ¿Tiene sentido el resultado obtenido?

Ejercicio 3: Similitud y homología



El alineamiento con mayor *score* corresponde a secuencias de organismos pertenecientes al mismo clado

Parcialito

<https://forms.gle/fZBgvmkB3uX6eQaz5>

user: Ciencia de Datos
pass: u2Djfu20



Informe

Este TP tiene informe!

Se entrega hasta el 28/08 hasta las 23:59

¿Por qué pedimos informes?

Para que puedan aplicar los conocimientos y herramientas adquiridas durante la cursada a un problema puntual. También sirve para evaluar la capacidad técnica para producir resultados, así como también el criterio a la hora de evaluar el significado biológico de los mismos y la comprensión de las potencialidades y limitaciones de los algoritmos o herramientas empleadas.

Cada informe de cada trabajo práctico puede (o no) estar relacionado con alguno anterior. Por lo tanto: Guarden los archivos y sean prolijos.

IMPORTANTE: Grupos de TPs = Grupos de informes!