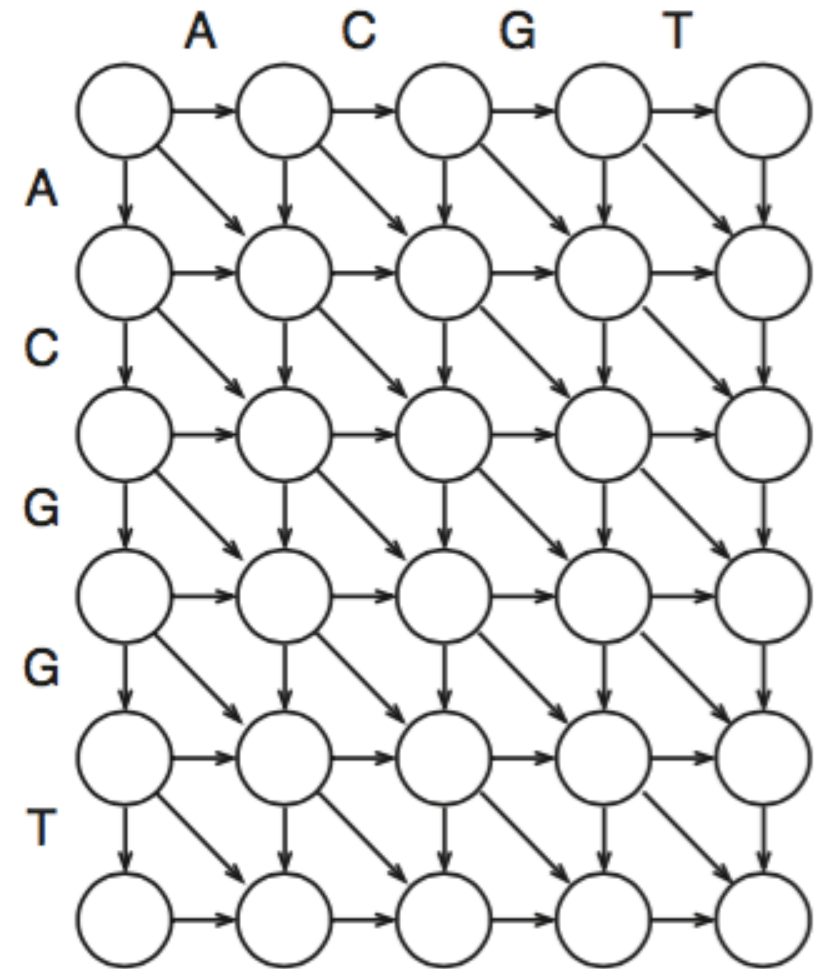


Introducción a la Bioinformática

Alineamiento de secuencias

Fernán Agüero
fernan@iib.unsam.edu.ar
Instituto de Investigaciones Biotecnológicas
Universidad Nacional de San Martín



Análisis comparativo

El alineamiento de secuencias es similar a otros tipos de análisis comparativo.

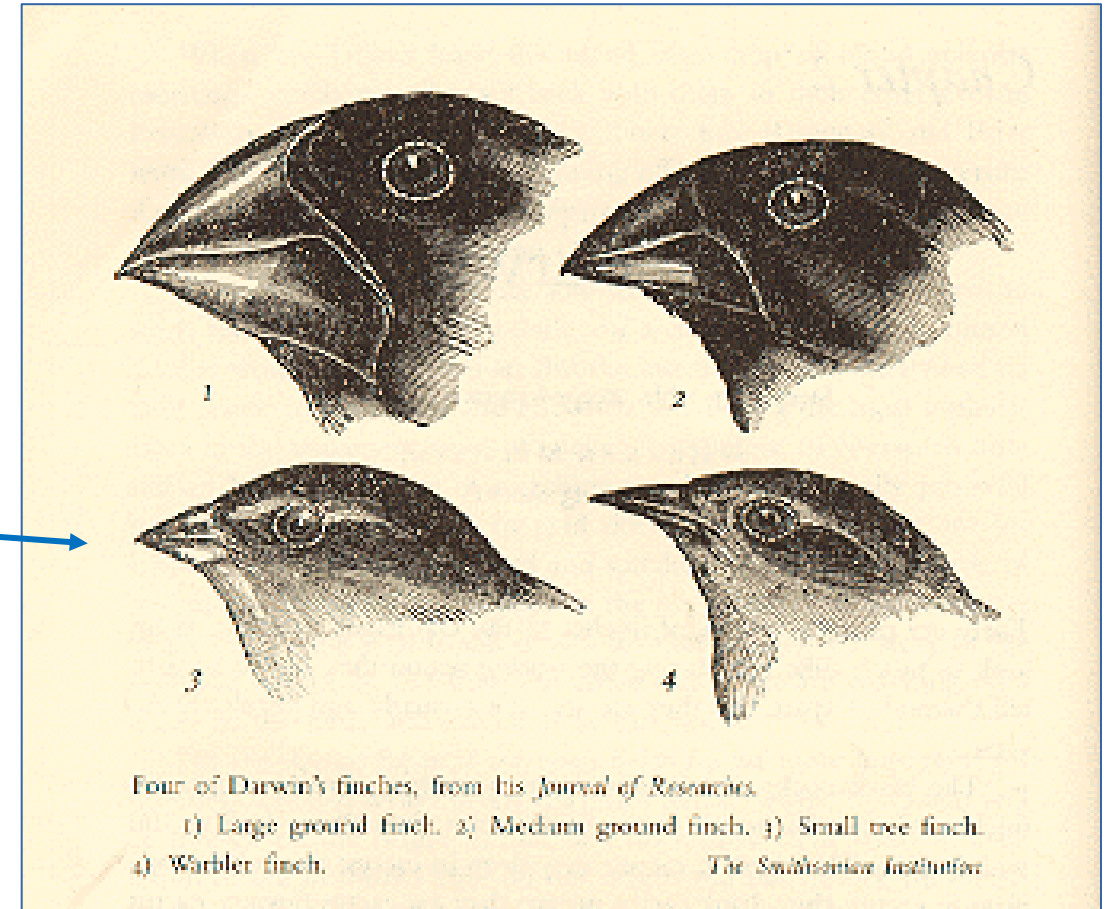
En ambos es necesario **cuantificar** similitudes y diferencias (scoring) entre un grupo relacionado de **entidades**

Picos de aves (Darwin)

Secuencias (Bioinformática)

GCTACTAGTTCGCTTAGC

GCTACTAGCTCTAGCGCGTATAGC



Finches of the Galápagos Islands observed by Charles Darwin on the voyage of HMS *Beagle*

Cómo comparar secuencias

Para poder comparar secuencias, tenemos que sistematizar la manera en que lo hacemos.

1 Alinear. Cómo?

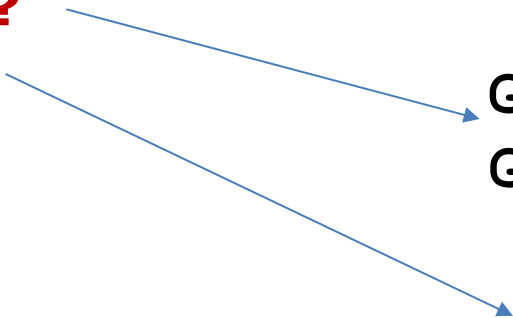
Por donde empezamos?

Comparamos las dos secuencias **letra a letra?**

Empezando **por la primera?**

Empezando **por la última?**

Tiene sentido?



GCTACTAGTTCGCTTAGC
GCTACTAGCTCTAGCGCGTATAGC

GCTACTAGTTCGCTTAGC
GCTACTAGCTCTAGCGCGTATAGC

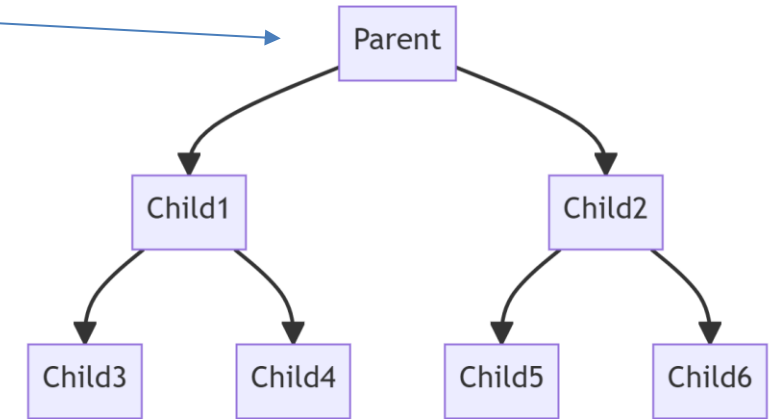
2 Cuantificar

Contar cuántas letras coinciden = **match**

Contar cuántas letras no coinciden = **mismatch**

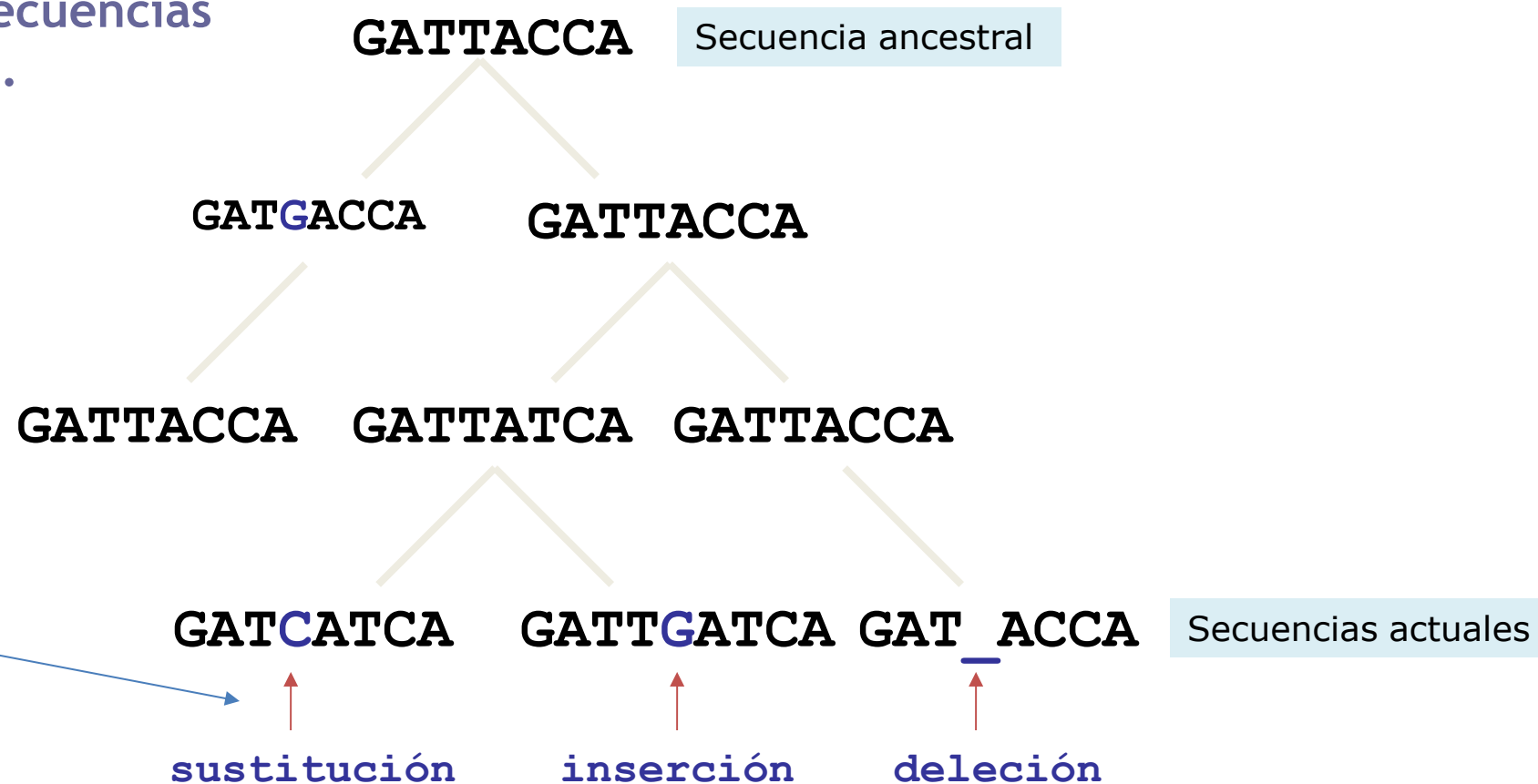
Homología vs similitud

- **Homología** entre dos entes biológicos implica una **herencia compartida**
- **Homología** es un término **cualitativo**
 - Se es homólogo o no se es
 - No es correcto decir que se es muy homólogo
 - O que algo es más homólogo que otra cosa
- **Similitud** implica una apreciación **cuantitativa** de algún carácter que se puede medir
 - Podemos usar una medida de similitud para **inferir** homología
 - Dos secuencias similares (tienen alto parecido) podrían (tal vez) tener un origen ancestral común



Análisis comparativo

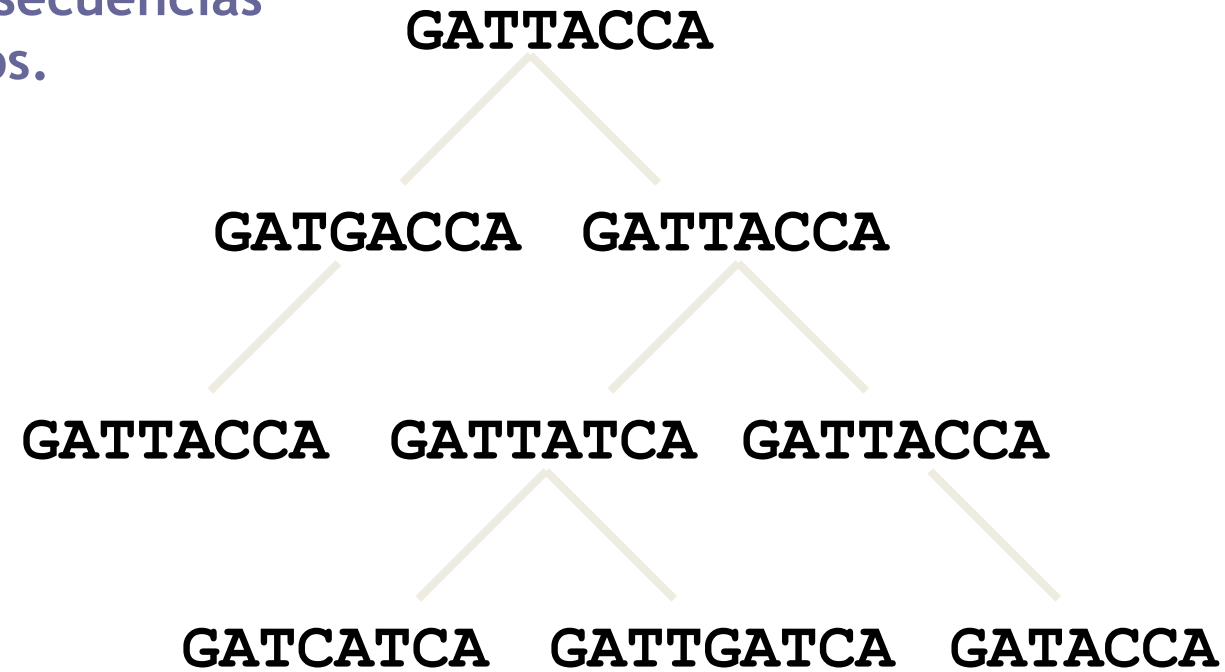
Los algoritmos que alinean secuencias **modelan** procesos evolutivos.



Deriva de un ancestro común a través de cambios incrementales debido a errores en la replicación del DNA, mutaciones, daño o crossing-over desigual.

Análisis comparativo

Los algoritmos que alinean secuencias **modelan** procesos evolutivos.

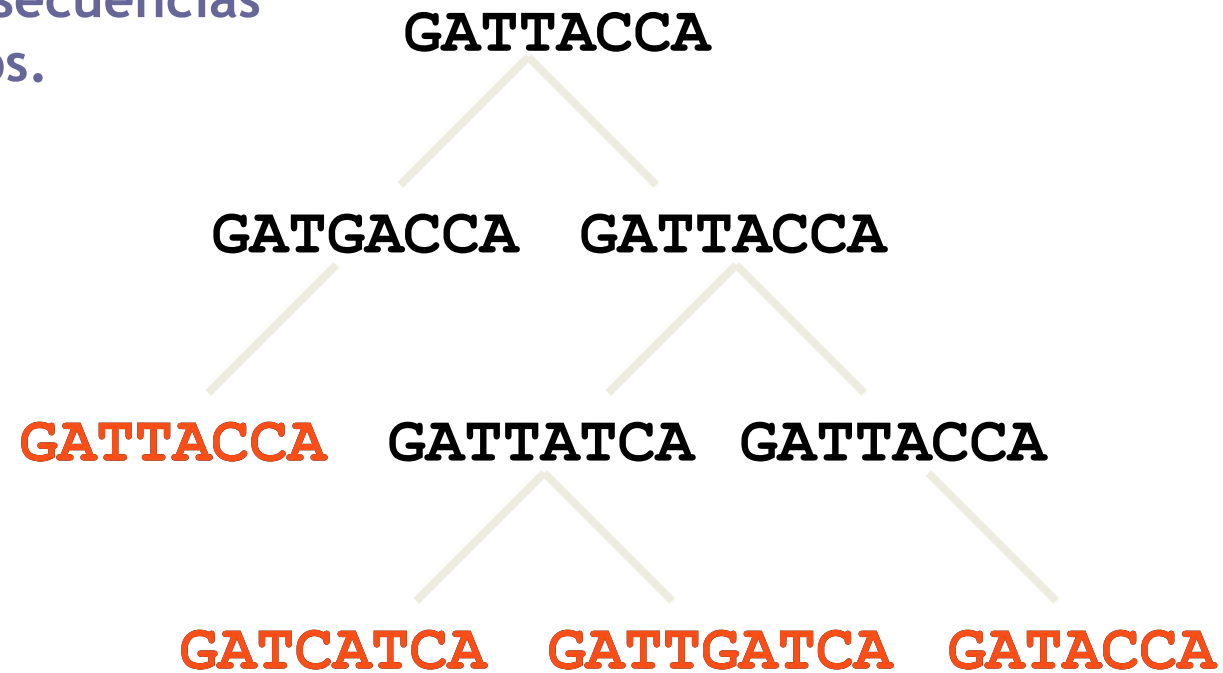


Evolución = deriva a partir de un ancestro común a través de cambios incrementales.

**Secuencias
actuales**

Análisis comparativo

Los algoritmos que alinean secuencias **modelan** procesos evolutivos.



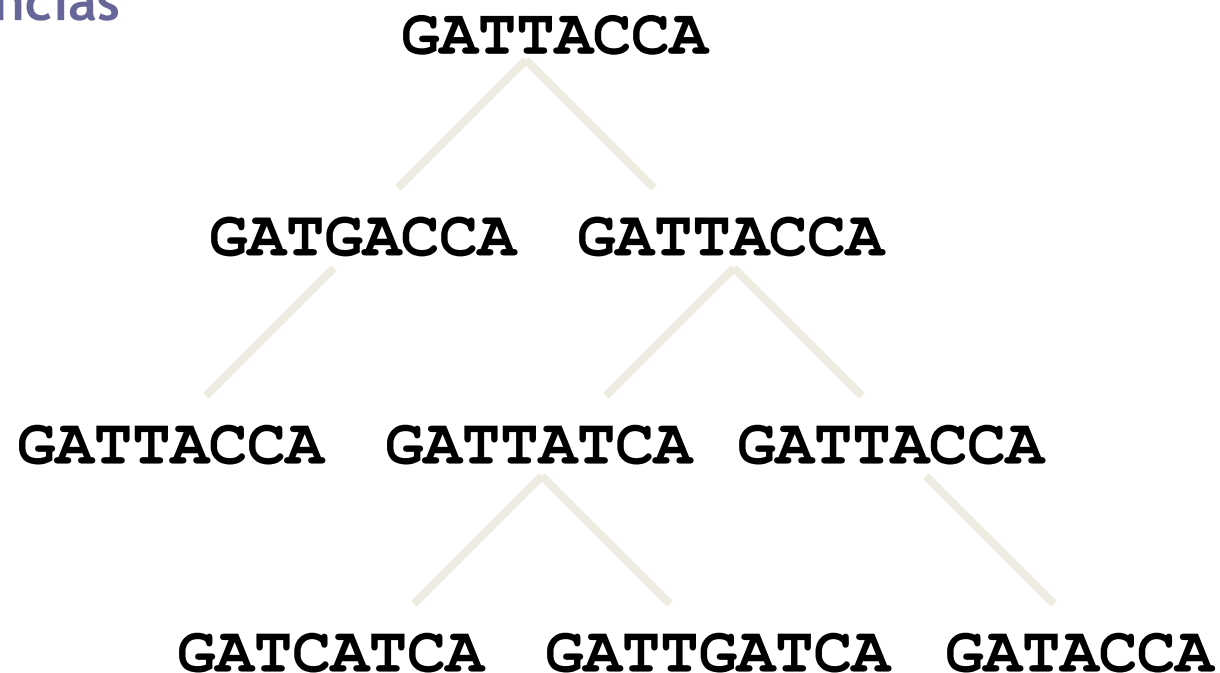
Evolución = deriva a partir de un ancestro común a través de cambios incrementales.

Sólo las secuencias actuales son conocidas, las secuencias ancestrales se postulan.

Secuencias
actuales

Los algoritmos que alinean secuencias **modelan** procesos evolutivos.

Evoución = deriva a partir de un ancestro común a través de cambios incrementales. Los cambios (mutaciones) que no matan a un individuo pueden pasar a la descendencia y fijarse en la población.



Homología implica una herencia común (un ancestro y una historia evolutiva compartida). Esto se puede inferir a partir de observaciones de **similitud** de secuencias.

Alineamientos de secuencias

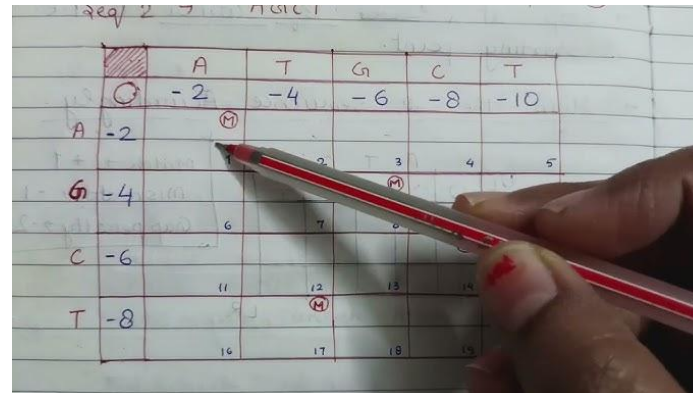
- Qué es un alineamiento?

- Es **un arreglo** de las secuencias de ADN, ARN o Proteínas para poder cuantificar similitudes
- Es **el procedimiento** de comparación de dos (o más) secuencias buscando lograr que una serie de caracteres individuales se encuentren en el mismo orden en ambas secuencias

SUJETO / OBJETO
VERBO / ACCIÓN

- Cómo alineamos dos secuencias?

- Usando un método (algoritmo)
 - programando una computadora
 - a mano, lapiz y papel (como en los viejos tiempos)



```
pairwiseAlignment(seq1, seq2)
1  m = seq1.length + 1
2  n = seq2.length + 1
3  for i = 0 to m
4    M[i, 0] = i.GAP
5  for j = 0 to n
6    M[0, j] = j.GAP
7  for i = 1 to m
8    for j = 1 to n
9      diag = M[i-1, j-1] + similarity(seq1[i], seq2[j])
10     up = M[i-1, j] + GAP
11     left = M[i, j-1] + GAP
12     M[i, j] = max(diag, up, left)
13  traceback(M[i, j])
14  distance = 1 - (numMATCH / totalAlignmentWithoutGap)
15  Return distance

clustalW-pairwiseAlignment(N, seq)
1  for i = 1 to N
2  for j = 1 to N
3  If (i < j)
4  distMatrix[i, j] = pairwiseAlignment(seq[i], seq[j])
```

Definición de alineamiento: tipos

- Alineamiento:** Cada base se usa a lo sumo una vez
- Alineamiento global:** Todas las bases se alinean con otra base o con un gap ("-")
- Alineamientos locales:** No hay necesidad de alinear todas las bases

Align **GATESLIKESCHEESE** and **GRATEDCHEESE**

G-ATESLIKESCHEESE	or	G-ATES	& CHEESE
GRATED-----CHEESE		GRATED	& CHEESE

Alineamientos buenos y malos?

Cuál es el 'mejor' alineamiento?

GCTACTAG-T-T--CGC-T-TAGC
GCTACTAGCTCTAGCGCGTATAGC

0 mismatches, 5 gaps

GCTACTAGTT-----CGCTTAGC
GCTACTAGCTCTAGCGCGTATAGC

3 mismatches, 1 gap

- Pero: necesitamos una medida **objetiva**
- **sistemas de puntaje (scoring)**
 - reglas para asignar puntos
 - el más simple: match, mismatch, gap

Cómo decidir cuál es el mejor?

- **Respuesta: el más significativo desde el punto de vista biológico**
- **Pero: necesitamos una medida **objetiva****
- **sistemas de puntaje (scoring)**
 - **reglas para asignar puntos**
 - **el más simple: match, mismatch, gap**

Cómo decidir cuál es el mejor?

- **Respuesta: el más significativo desde el punto de vista biológico**
- **Pero: necesitamos una medida **objetiva****
- **sistemas de puntaje (scoring)**
 - **reglas para asignar puntos**
 - **el más simple: match, mismatch, gap**

Cómo decidir cuál es el mejor?

- **Respuesta: el más significativo desde el punto de vista biológico**
- **Pero: necesitamos una medida **objetiva****
- **sistemas de puntaje (scoring)**
 - **reglas para asignar puntos**
 - **el más simple: match, mismatch, gap**

Un primer sistema de puntajes

Ejemplo de sistema de score

match = +1

mismatch = 0

gap = -1

G-ATESLIKESCHEESE
GRATED-----CHEESE

Un primer sistema de puntajes

Ejemplo de sistema de score

match = +1

mismatch = 0

gap = -1

G-ATESLIKESCHEESE
GRATED-----CHEESE

Score

$$(10 * 1) + (1 * 0) + (6 * (-1)) = +4$$

Cambiamos nuestro sistema de puntajes

Usando otro de sistema de puntajes?

match = +2

mismatch = 0

gap = -1

G-ATESLIKESCHEESE
GRATED-----CHEESE

Usando otro sistema de score

Score

$$(10 * 2) + (1 * 0) + (5 * (-1)) = +14$$

No se pueden comparar scores

- **no tiene sentido comparar scores de distintos alineamientos**
- **a menos que se especifique el sistema de scoring utilizado**

No se pueden comparar scores

- **Primera conclusión importante:**
 - **no tiene sentido comparar scores de distintos alineamientos**
 - **a menos que se especifique el sistema de scoring utilizado**

No se pueden comparar scores

- **Primera conclusión importante:**
 - **no tiene sentido comparar scores de distintos alineamientos**
 - **a menos que se especifique el sistema de scoring utilizado**

No se pueden comparar scores

- **Primera conclusión importante:**
 - **no tiene sentido comparar scores de distintos alineamientos**
 - **a menos que se especifique el sistema de scoring utilizado**

Gap penalties

gap opening penalty = -5

gap extension penalty = -1

1- Abrir un gap es costoso

GCTACTAG-T-T--CGC-T-TAGC
GCTACTAGCTCTAGCGCGTATAGC

$$\text{Penalty} = 5 * (-5) + 6 * (-1) = -31$$

2 - Extender un gap es menos costoso

GCTACTAGTT-----CGCTTAGC
GCTACTAGCTCTAGCGCGTATAGC

$$\text{Penalty} = 1 * (-5) + 6 * (-1) = -11$$

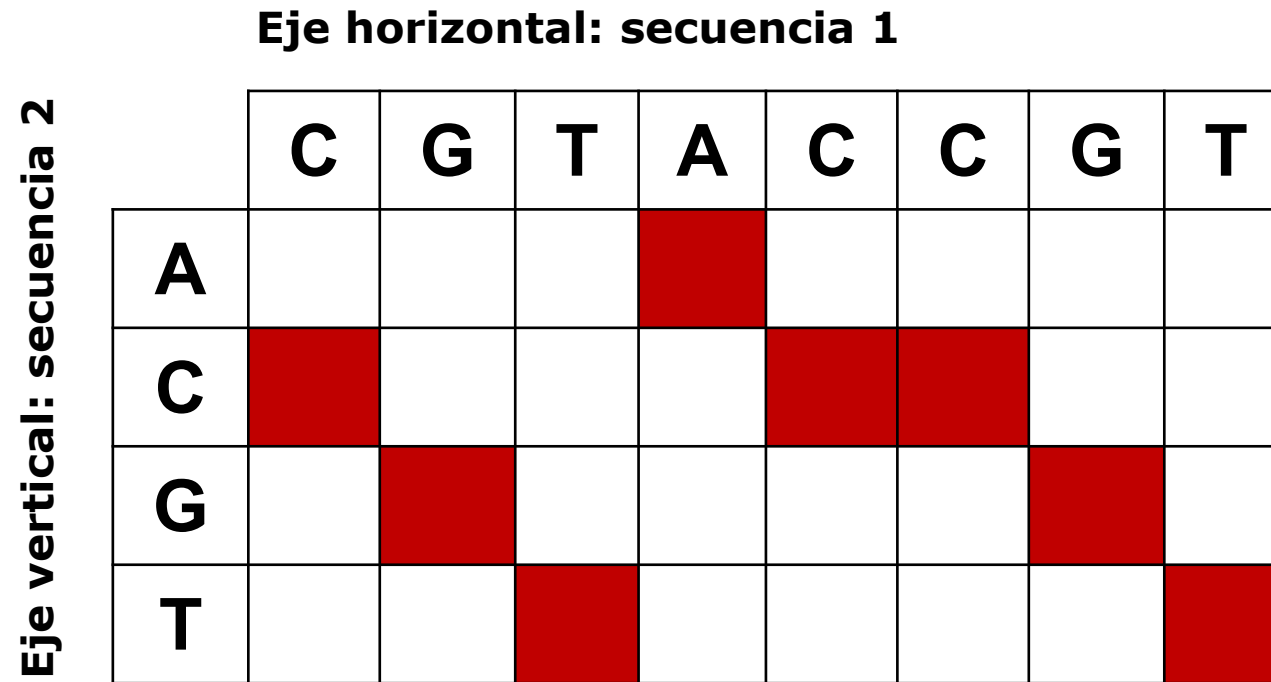
Dot plots: introducción

		Eje horizontal: secuencia 1							
		C	G	T	A	C	C	G	T
Eje vertical: secuencia 2	A								
	C								
	G								
	T								

Objetivo: marcar (con una X) o pintar las posiciones en la grilla donde hay coincidencia (match)

Fitch WM. Locating gaps in amino acid sequences to optimize the homology between two proteins. Biochem Genet. 1969 Apr;3(2):99-108. doi: 10.1007/BF00520346. PMID: 5364927.

Dot plots: introducción

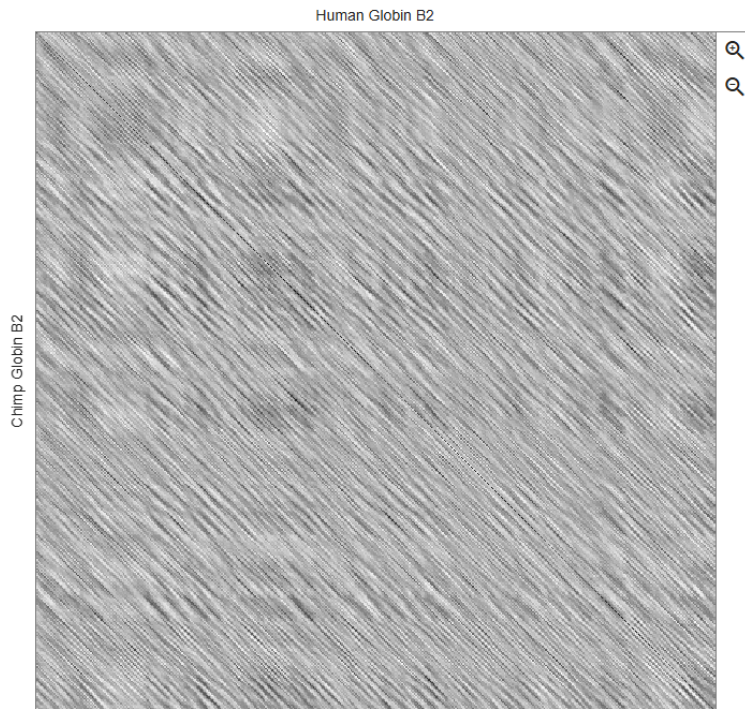


Objetivo: marcar (con una X) o pintar las posiciones en la grilla donde hay coincidencia (match)

Fitch WM. Locating gaps in amino acid sequences to optimize the homology between two proteins. Biochem Genet. 1969 Apr;3(2):99-108. doi: 10.1007/BF00520346. PMID: 5364927.

Dot Plots / Dot Matrix Plots

- Se colocan **"puntos" (dots)** en donde hay un match.
- Las **líneas diagonales** marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?



Human Globin B2 (SAI82133.1) vs Chimpanzee Globin B2 (SAI82151.1)

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

Dot Plots / Dot Matrix Plots

- Se colocan “puntos” (dots) en donde hay un match.
- Las líneas diagonales marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?
- Se aplican filtros para mejorar la comprensión del gráfico
 - **Ventana deslizante (sliding window)**
 - **Window size = 3**
 - **Stringency = 3**

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

Dot Plots / Dot Matrix Plots

- Se colocan “puntos” (dots) en donde hay un match.
- Las líneas diagonales marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?
- Se aplican filtros para mejorar la comprensión del gráfico
 - **Ventana deslizante (sliding window)**
 - **Window size = 3**
 - **Stringency = 3**

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

Dot Plots / Dot Matrix Plots

- Se colocan “puntos” (dots) en donde hay un match.
- Las líneas diagonales marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?
- Se aplican filtros para mejorar la comprensión del gráfico
 - **Ventana deslizante (sliding window)**
 - **Window size = 3**
 - **Stringency = 3**

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

Dot Plots / Dot Matrix Plots

- Se colocan “puntos” (dots) en donde hay un match.
- Las líneas diagonales marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?
- Se aplican filtros para mejorar la comprensión del gráfico
 - **Ventana deslizante (sliding window)**
 - **Window size = 3**
 - **Stringency = 3**

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

Dot Plots / Dot Matrix Plots

- Se colocan “puntos” (dots) en donde hay un match.
- Las líneas diagonales marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?
- Se aplican filtros para mejorar la comprensión del gráfico
 - **Ventana deslizante (sliding window)**
 - **Window size = 3**
 - **Stringency = 3**

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

Dot Plots / Dot Matrix Plots

- Se colocan “puntos” (dots) en donde hay un match.
- Las líneas diagonales marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?
- Se aplican filtros para mejorar la comprensión del gráfico
 - **Ventana deslizante (sliding window)**
 - **Window size = 3**
 - **Stringency = 3**

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

Dot Plots / Dot Matrix Plots

- Se colocan “puntos” (dots) en donde hay un match.
- Las líneas diagonales marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?
- Se aplican filtros para mejorar la comprensión del gráfico
 - **Ventana deslizante (sliding window)**
 - **Window size = 3**
 - **Stringency = 3**

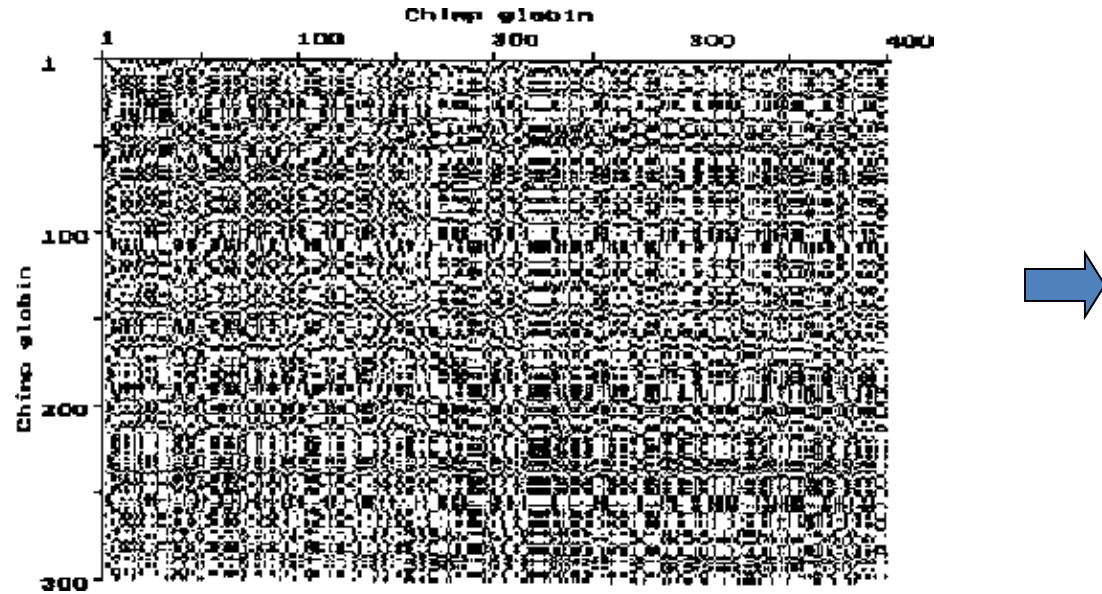
	G	A	T	C	A	A	C	T	G	A	C	G	T	A
G														
T														
T														
C														
A														
G														
C														
T														
G														
C														
G														
T														
A														
C														

Dot Plots / Dot Matrix Plots

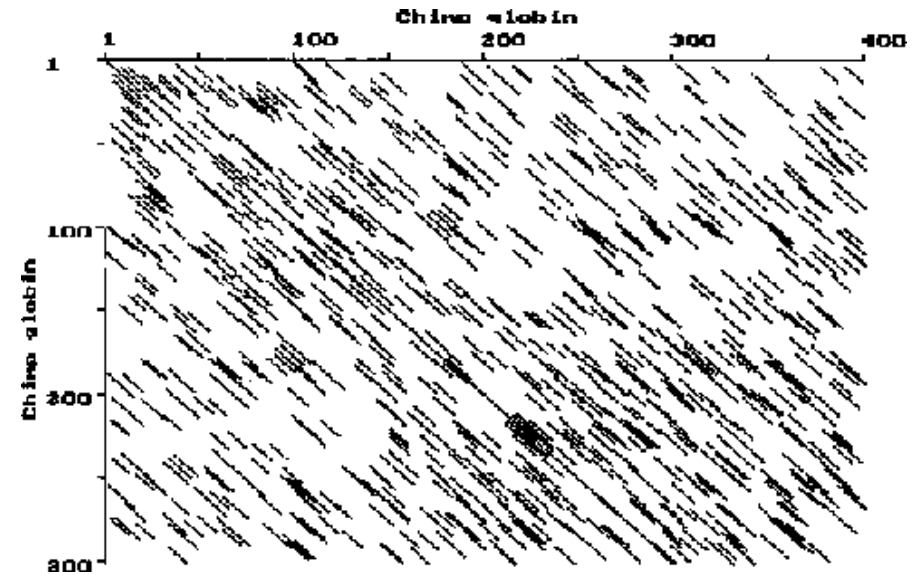
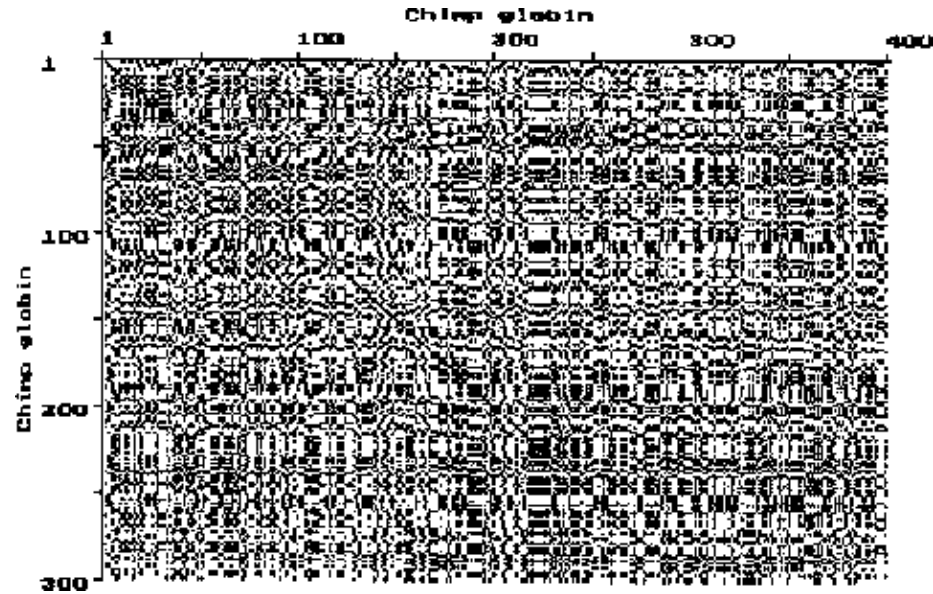
- Se colocan “puntos” (dots) en donde hay un match.
- Las líneas diagonales marcan regiones de identidad.
- Qué pasa cuando hay muchas coincidencias?
- Se aplican filtros para mejorar la comprensión del gráfico
 - **Ventana deslizante (sliding window)**
 - **Window size = 3**
 - **Stringency = 3**

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

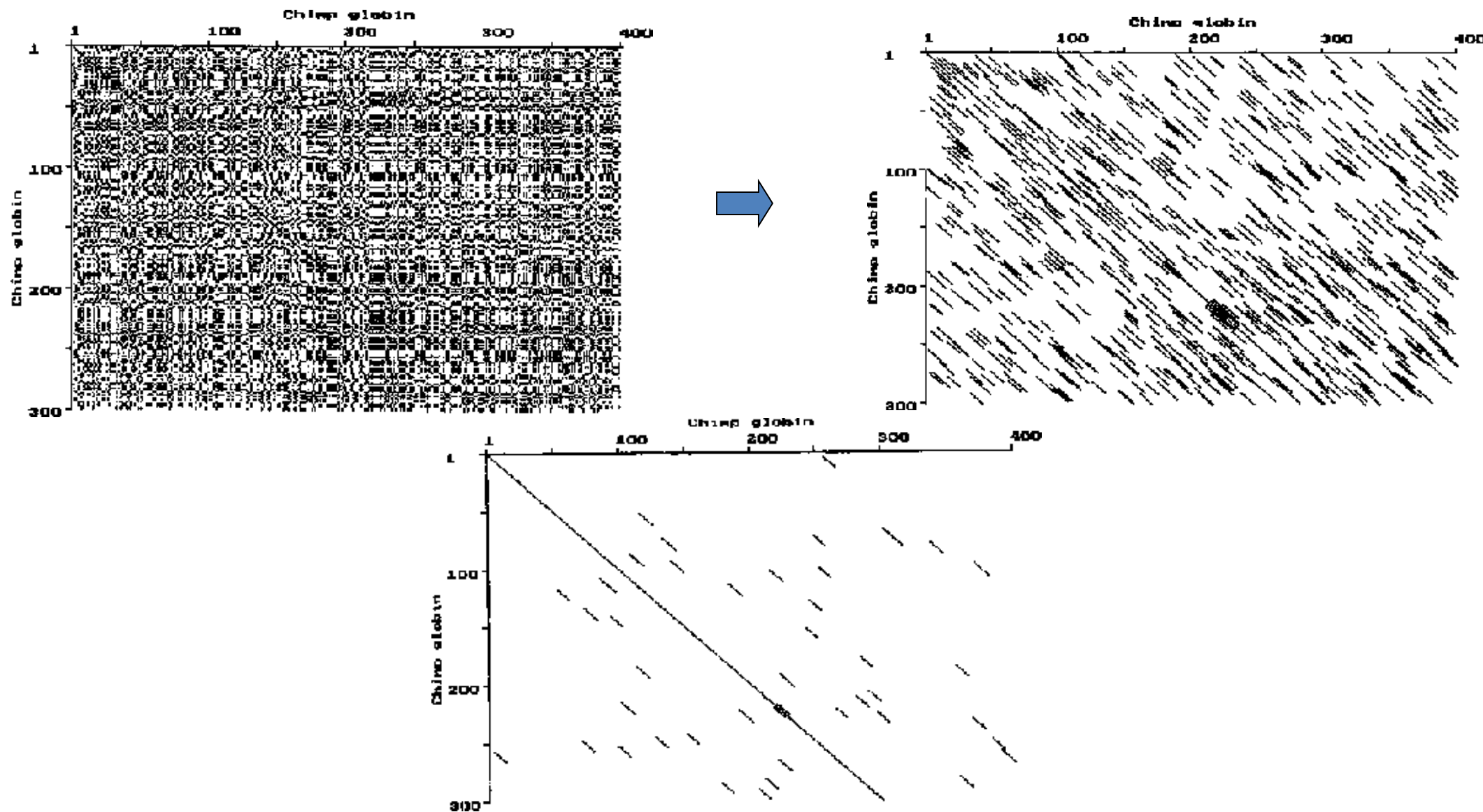
Limpiando Dot Matrix Plots con filtros



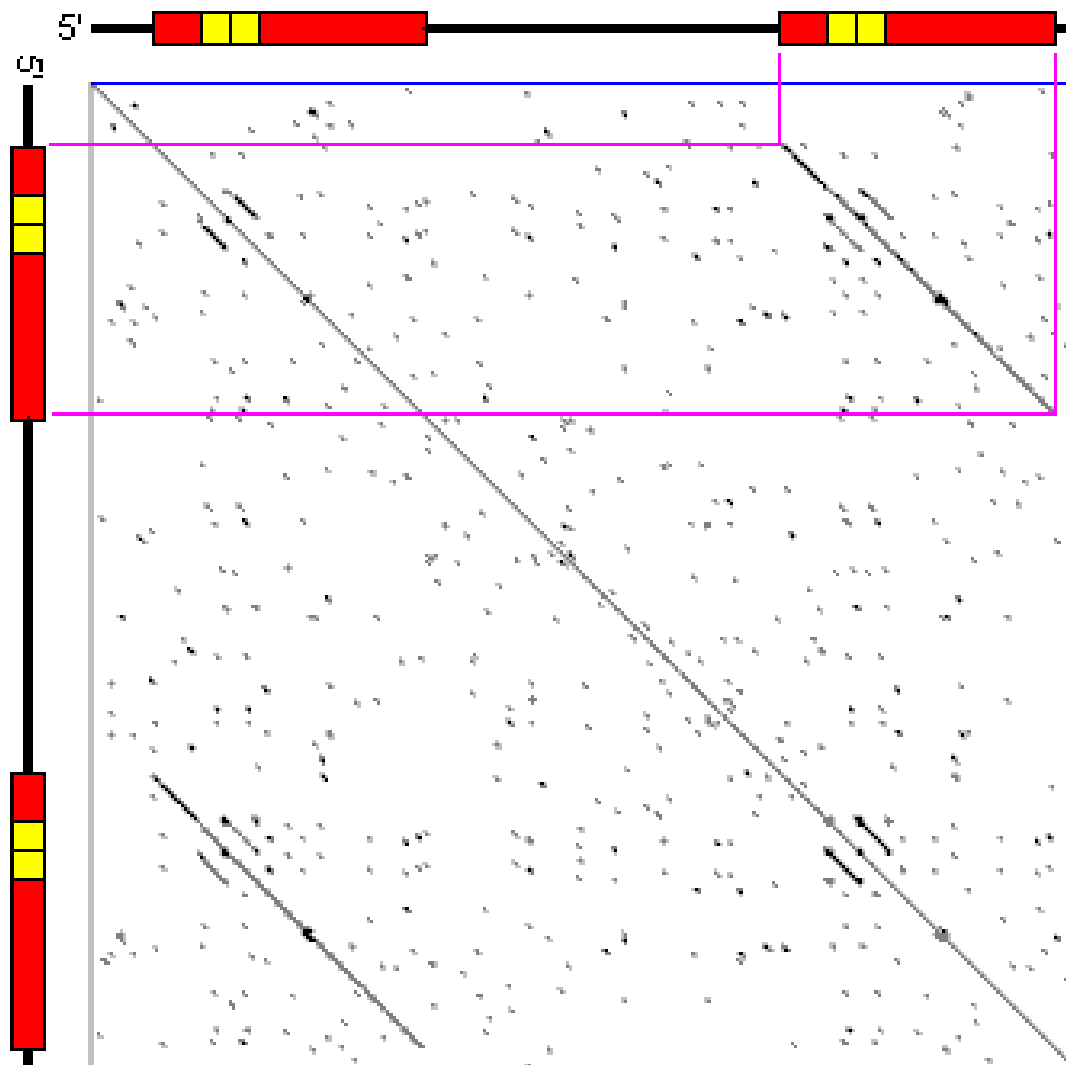
Limpiando Dot Matrix Plots con filtros



Limpiando Dot Matrix Plots con filtros



Dot Matrix Plot

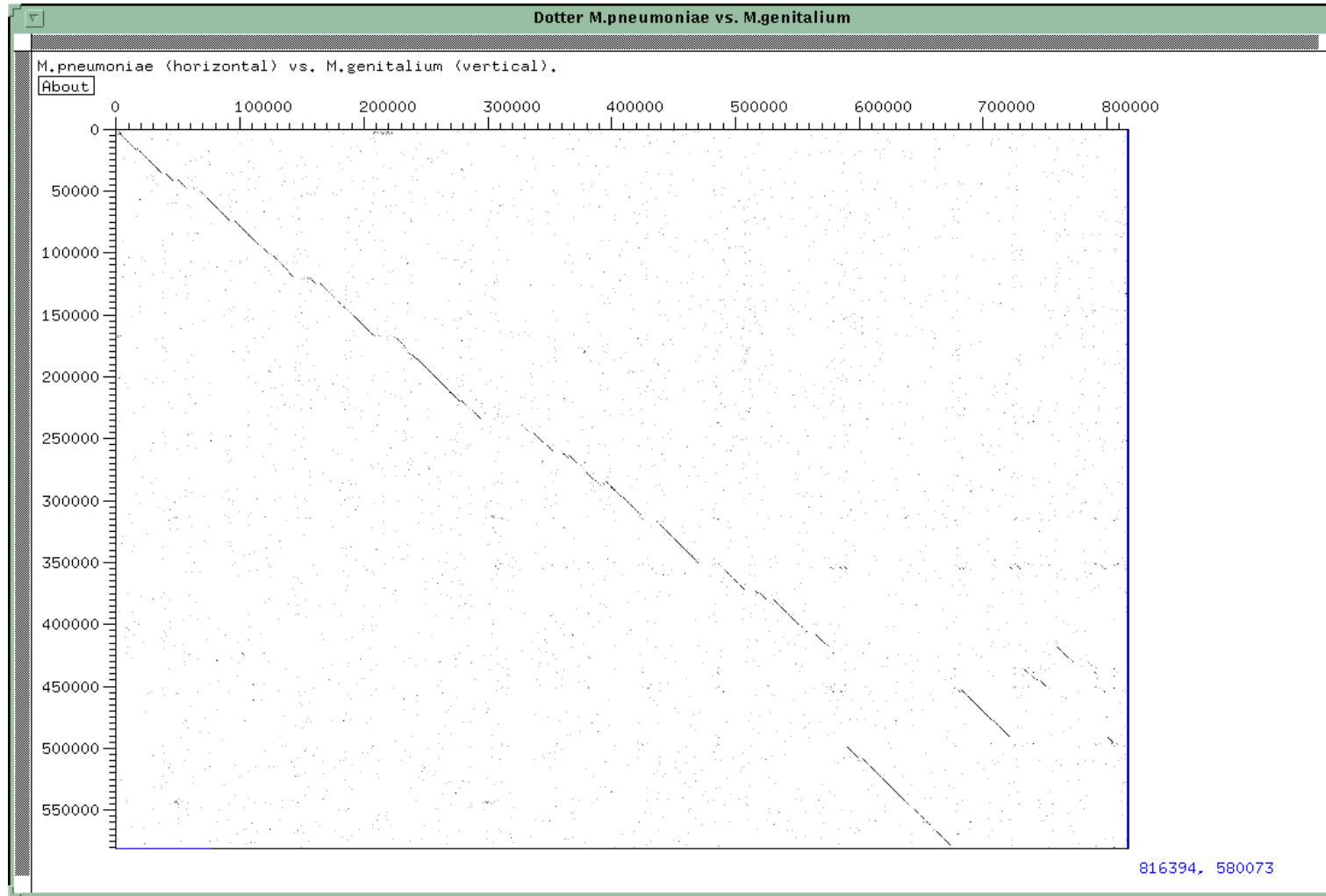


Clave

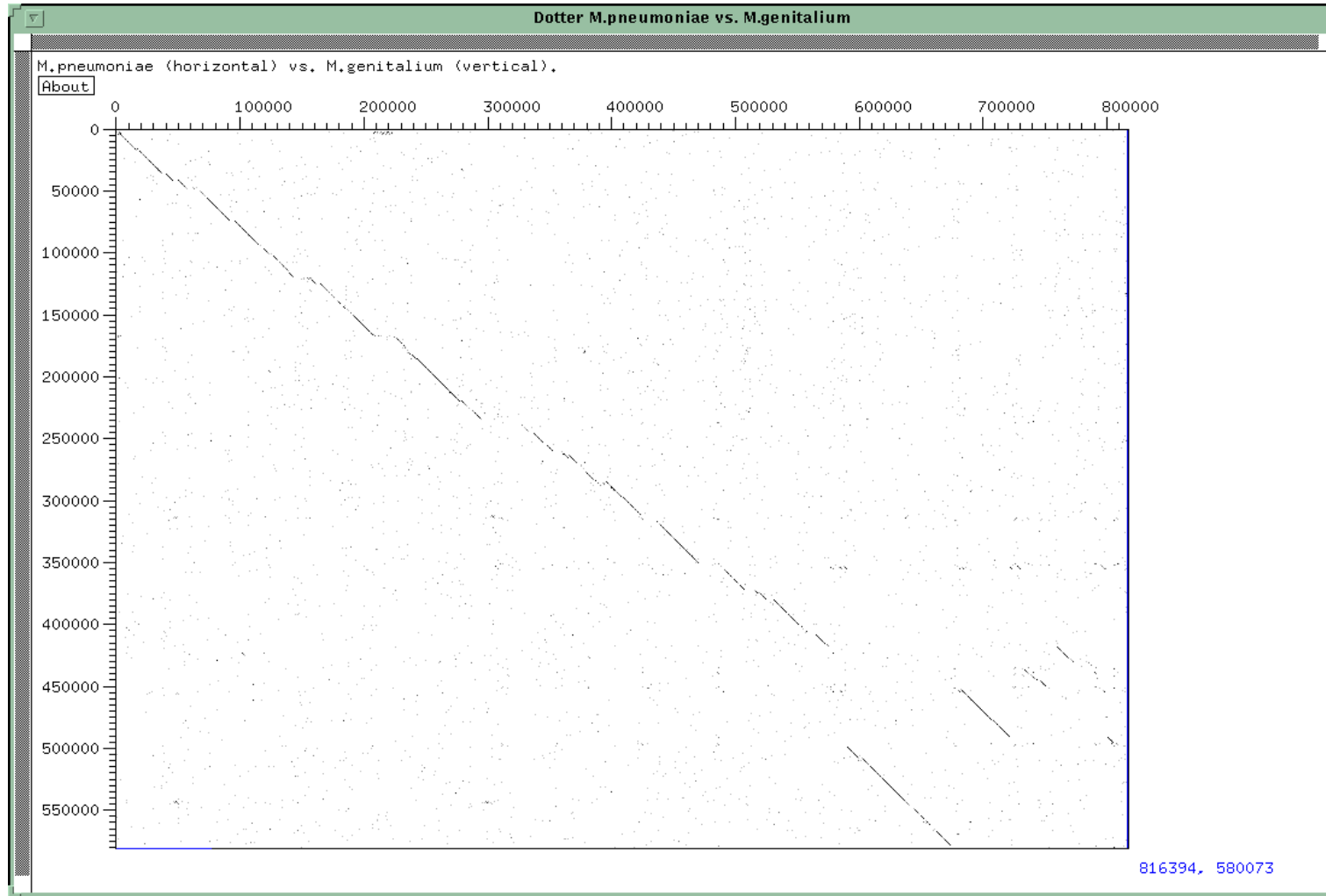
- Dominio A
- Motivo B

Qué nos muestra este Dot Plot?
Cuántas diagonales vemos?

Cómo leer Dot Matrix Plots



Cómo leer Dot Matrix Plots

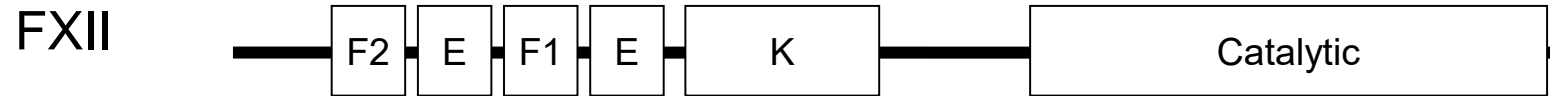


Qué nos muestran las interrupciones y corrimientos?

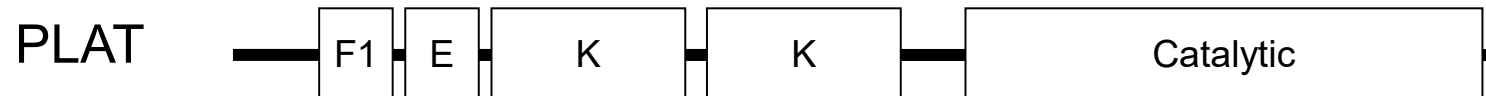
Dot plots para análisis exploratorio

Dominios mezclados confunden a los algoritmos de alineamiento.

Módulos (dominios) en el factor XII de coagulación (FXII) y en el activador de plasminógenos - tissue plasminogen activator (PLAT)



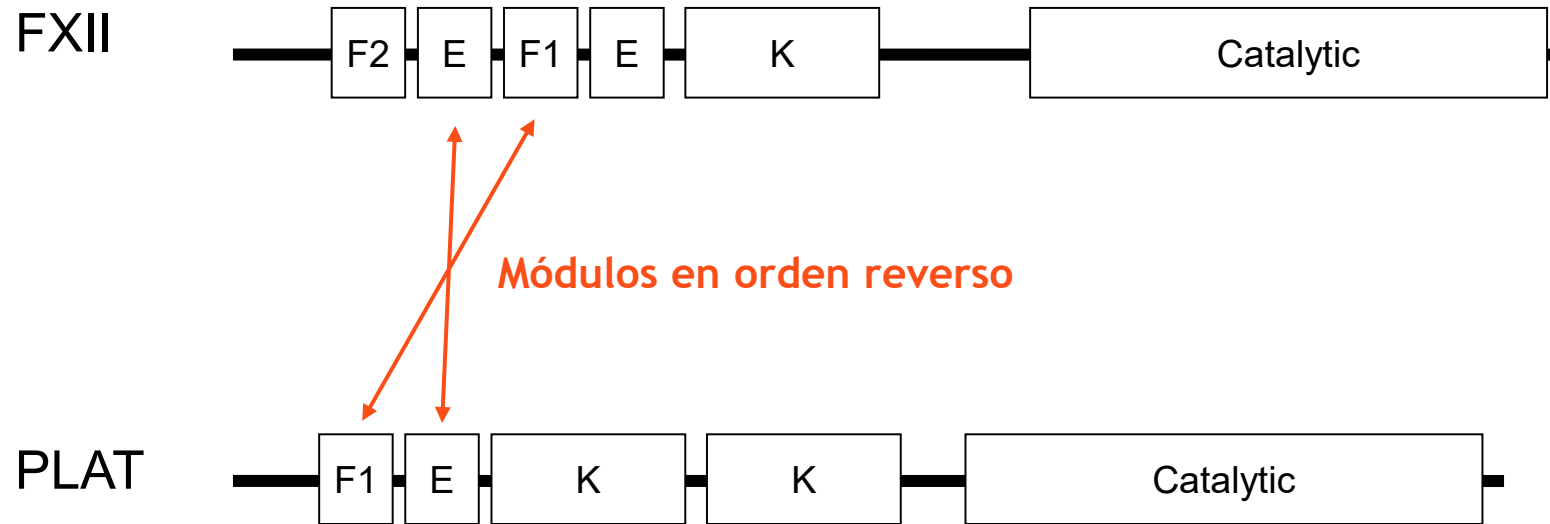
F1,F2	Fibronectin repeats
E	EGF similarity domain
K	Kringle domain
Catalytic	Serine protease activity



Dot plots para análisis exploratorio

Dominios mezclados confunden a los algoritmos de alineamiento.

Módulos (dominios) en el factor XII de coagulación (FXII) y en el activador de plasminógenos - tissue plasminogen activator (PLAT)

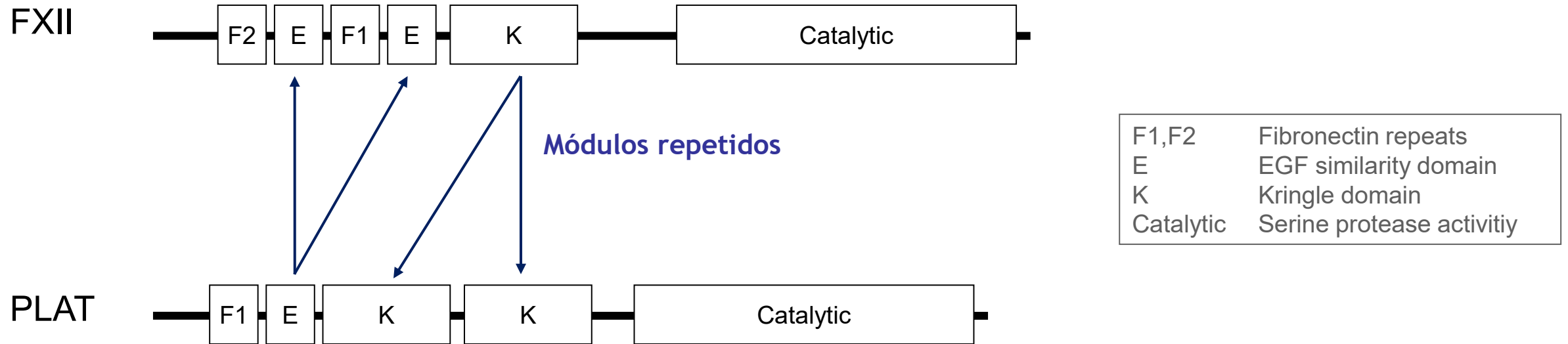


F1,F2	Fibronectin repeats
E	EGF similarity domain
K	Kringle domain
Catalytic	Serine protease activity

Dot plots para análisis exploratorio

Dominios mezclados confunden a los algoritmos de alineamiento.

Módulos (dominios) en el factor XII de coagulación (FXII) y en el activador de plasminógenos - tissue plasminogen activator (PLAT)

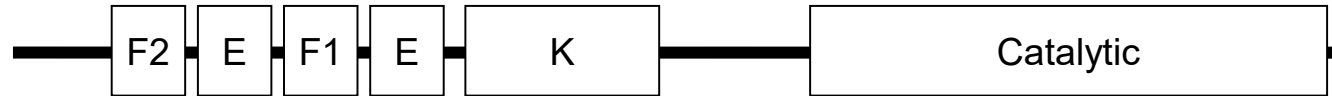


Dot plots para análisis exploratorio

Dominios mezclados confunden a los algoritmos de alineamiento.

Módulos (dominios) en el factor XII de coagulación (FXII) y en el activador de plasminógenos - tissue plasminogen activator (PLAT)

FXII



F1,F2	Fibronectin repeats
E	EGF similarity domain
K	Kringle domain
Catalytic	Serine protease activity

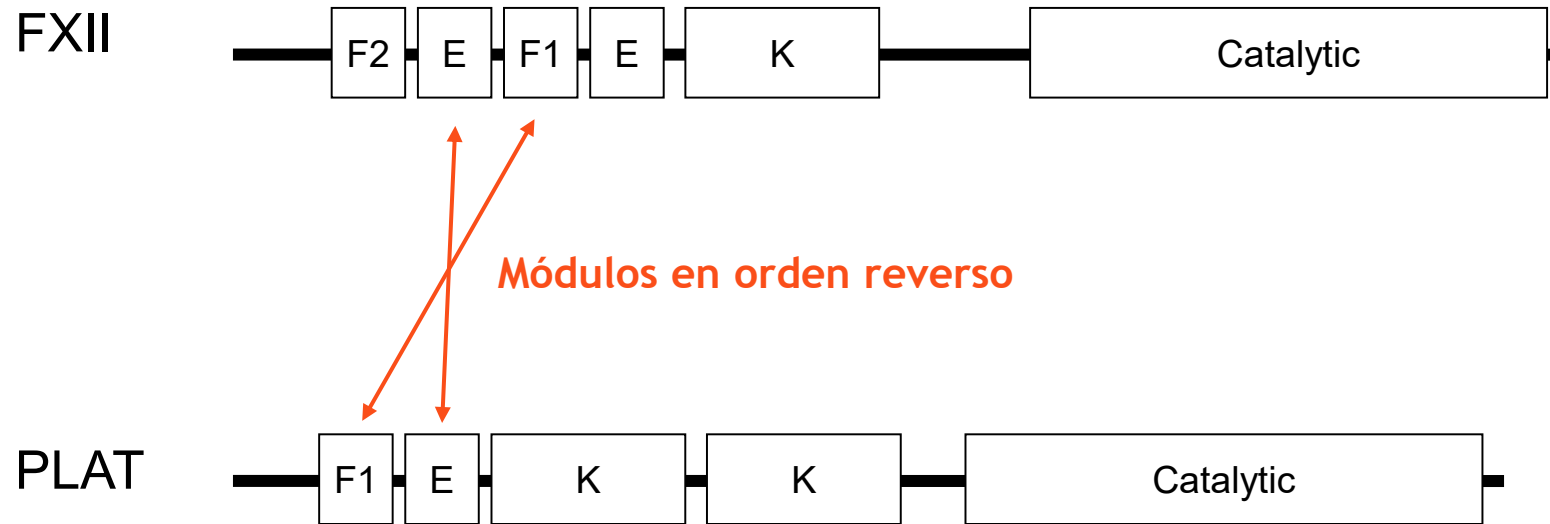
PLAT



Dot plots para análisis exploratorio

Dominios mezclados confunden a los algoritmos de alineamiento.

Módulos (dominios) en el factor XII de coagulación (FXII) y en el activador de plasminógenos - tissue plasminogen activator (PLAT)

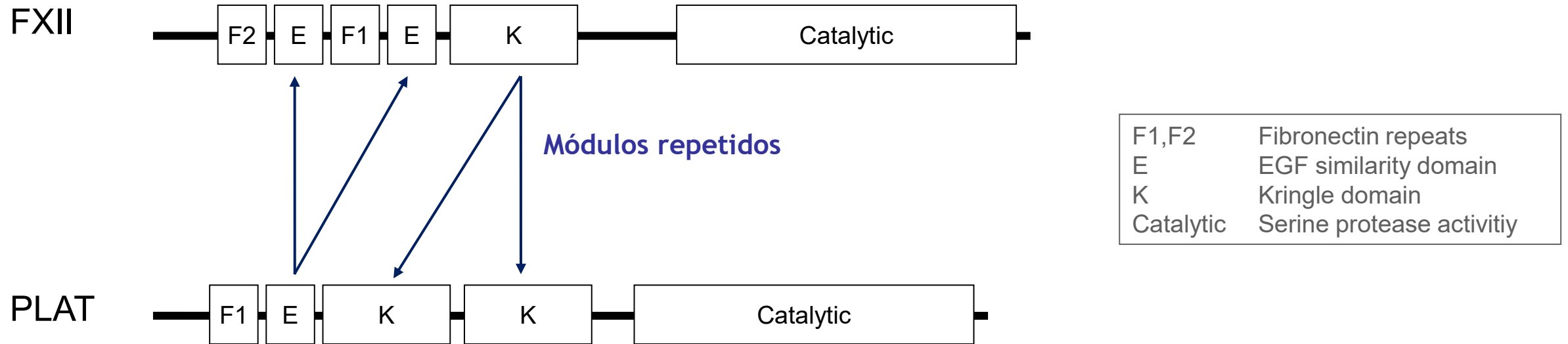


F1,F2	Fibronectin repeats
E	EGF similarity domain
K	Kringle domain
Catalytic	Serine protease activity

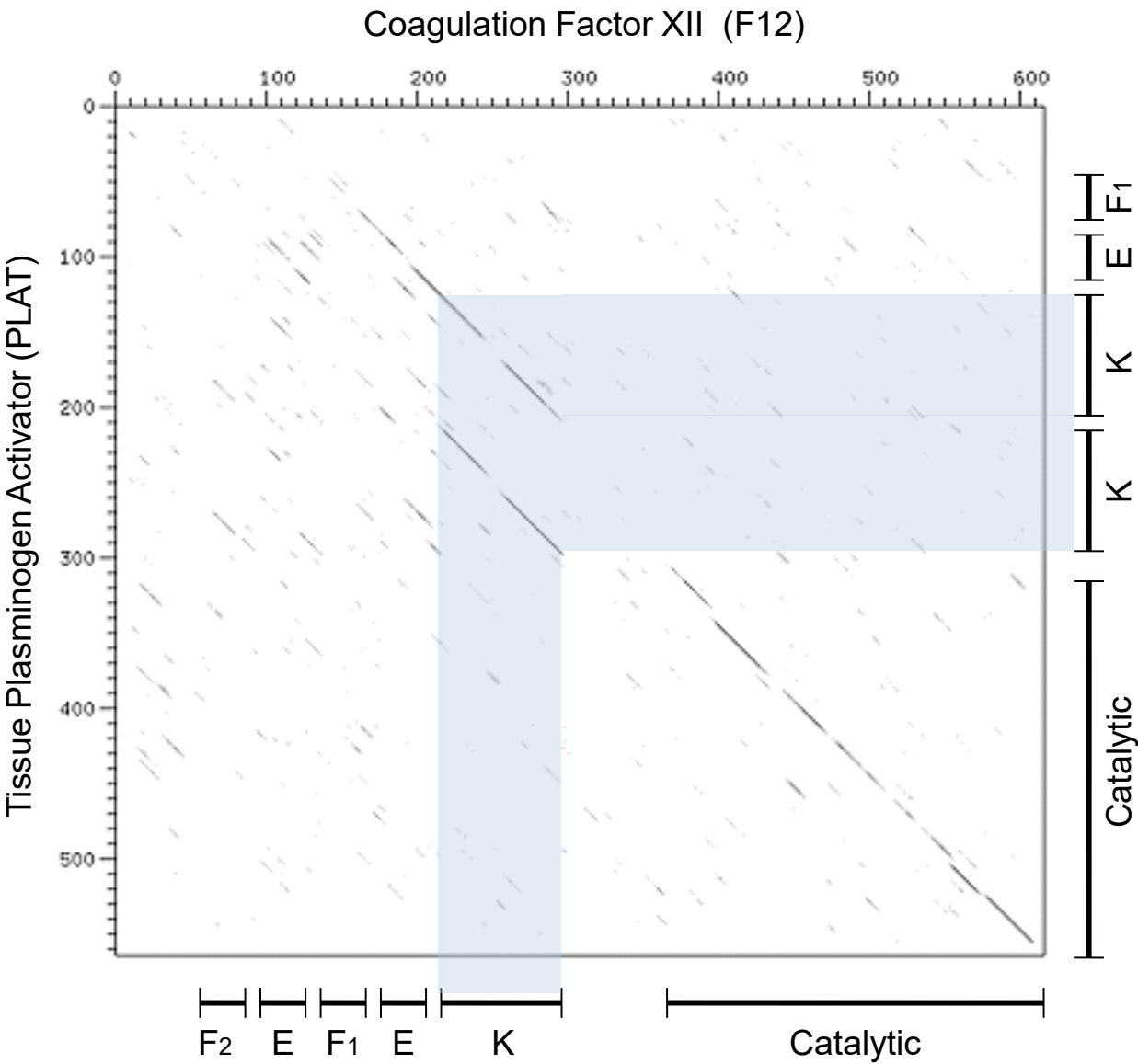
Dot plots para análisis exploratorio

Dominios mezclados confunden a los algoritmos de alineamiento.

Módulos (dominios) en el factor XII de coagulación (FXII) y en el activador de plasminógenos - tissue plasminogen activator (PLAT)



Dot plots para análisis exploratorio



Dominios repetidos muestran un patrón característico.

Dot plots: path graphs

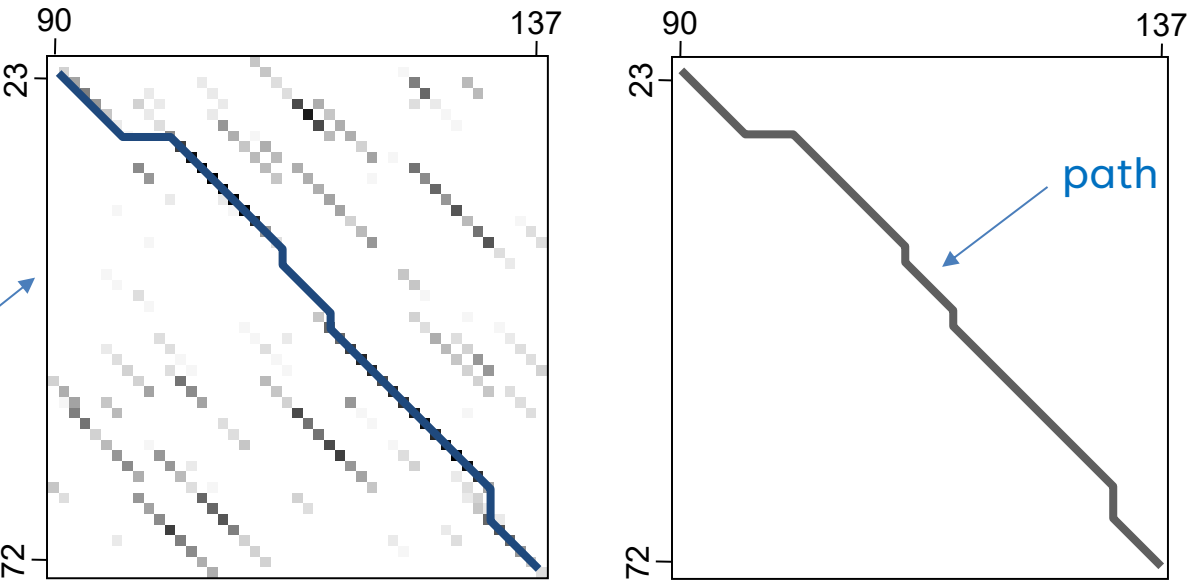
Los dot plots sugieren **caminos** (paths) a través del **espacio de alineamientos posibles**

Path graphs son representaciones explícitas de un alineamiento.

Cada path es un **alineamiento único**

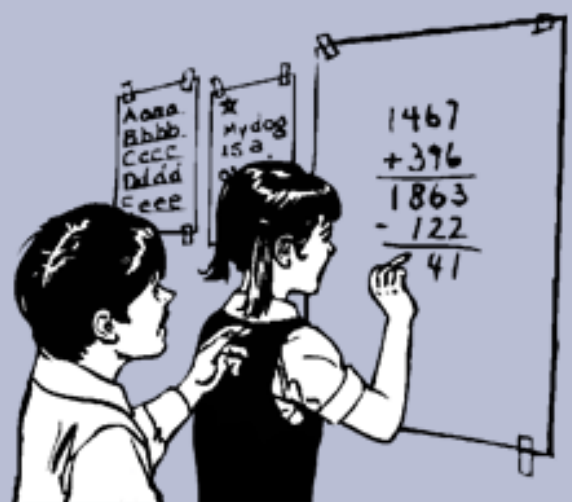
path graph

Dominios EGF conservados en la urokinase plasminogen activator (PLAU) y el tissue plasminogen activator (PLAT)



PLAU	90	EPKKVKDHCSKHSPCQKGGTCVNMP--SGPH-CLCPQHLTGNHCQKEK---CFE	137
PLAT	23	ELHQVPSNCD----CLNGGTCVSNKYFSNIHWCNCPKKFGGQHCEIDKSKTCYE	72

I'm definitely going to need coffee
for this shit.



your  cards
someecards.com

Path graphs: encontrar el mejor camino

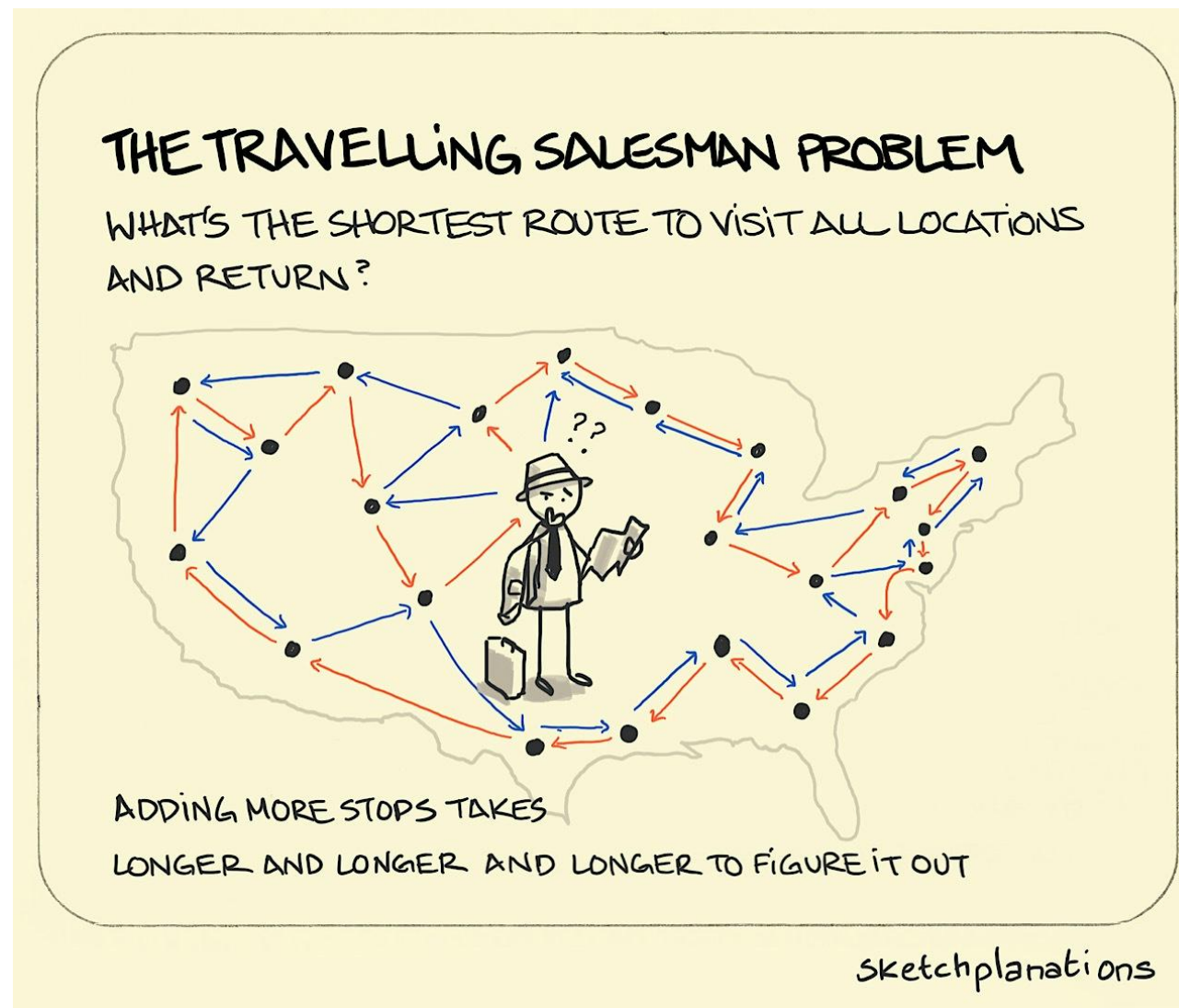
Los problemas que involucran encontrar **la mejor ruta** o camino posible entre varios (best-path problems) son comunes en computación científica.

Que necesitamos para encontrar el **mejor** camino?

Definir qué quiere decir mejor.

Una métrica cuantitativa:

- Kilómetros recorridos
- Combustible gastado
- Tiempo transcurrido

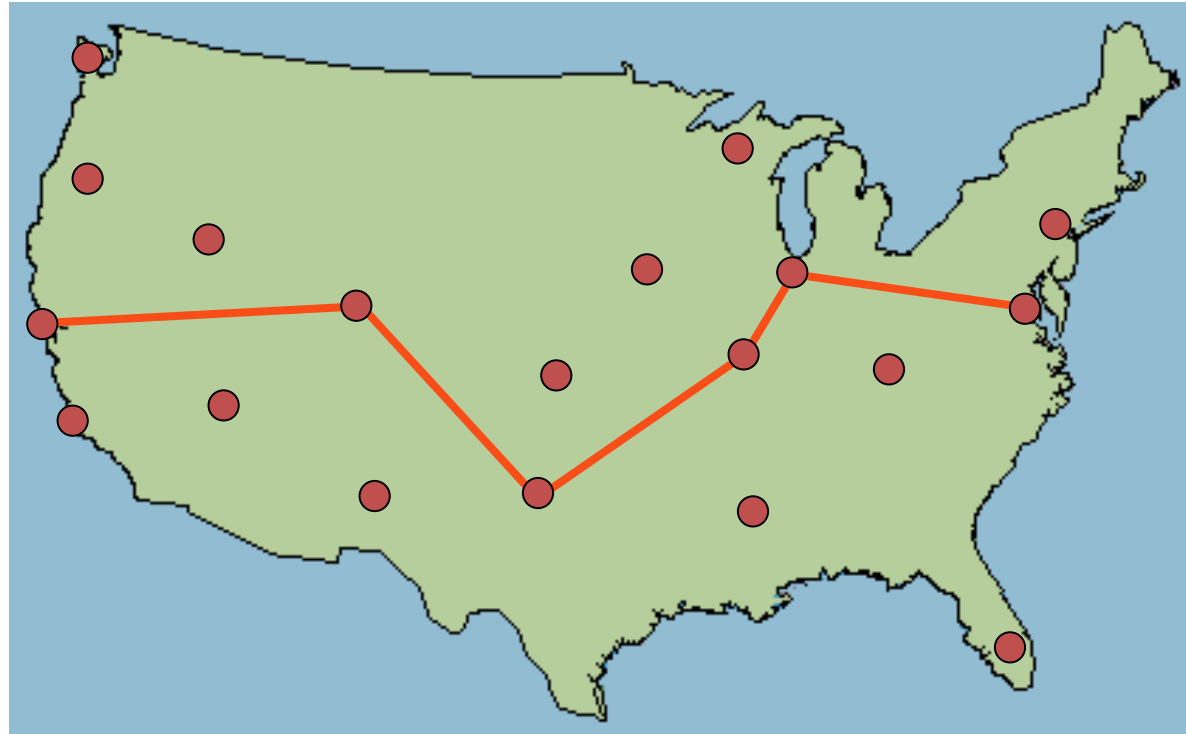


Path graphs: encontrar el mejor camino

Los problemas que involucran encontrar **la mejor ruta** o camino posible entre varios (best-path problems) son comunes en computación científica.

El algoritmo para encontrar el mejor camino entre dos extremos y pasando por varios puntos se llama **dynamic programming**

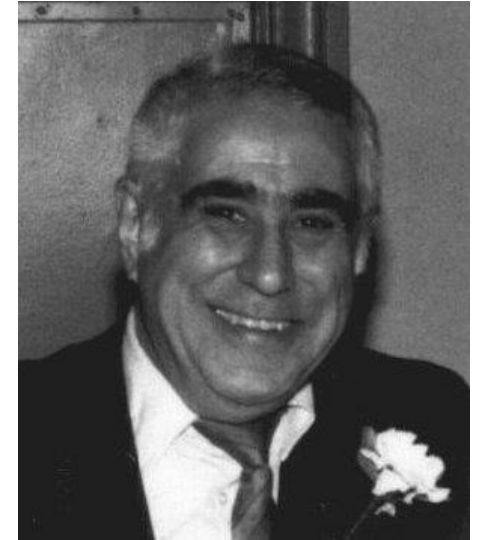
Rutear una llamada telefónica desde NY a San Francisco



1970

Saul Needleman & Christian Wunsch

Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970 Mar;48(3):443-53. doi: 10.1016/0022-2836(70)90057-4. PMID: 5420325.



Un ejemplo:

Construir un
alineamiento óptimo
entre estas dos
secuencias

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

Match: +1

Mismatch: -1

Gap: -1

Un ejemplo:

Construir un
alineamiento óptimo
entre estas dos
secuencias

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

Utilizando las
siguientes reglas de
scoring:

Match: +1

Mismatch: -1

Gap: -1

Dynamic programming: ejemplo

Ordenar las dos secuencias en una matriz bidimensional

Los vértices de cada celda se encuentran entre letras (bases).

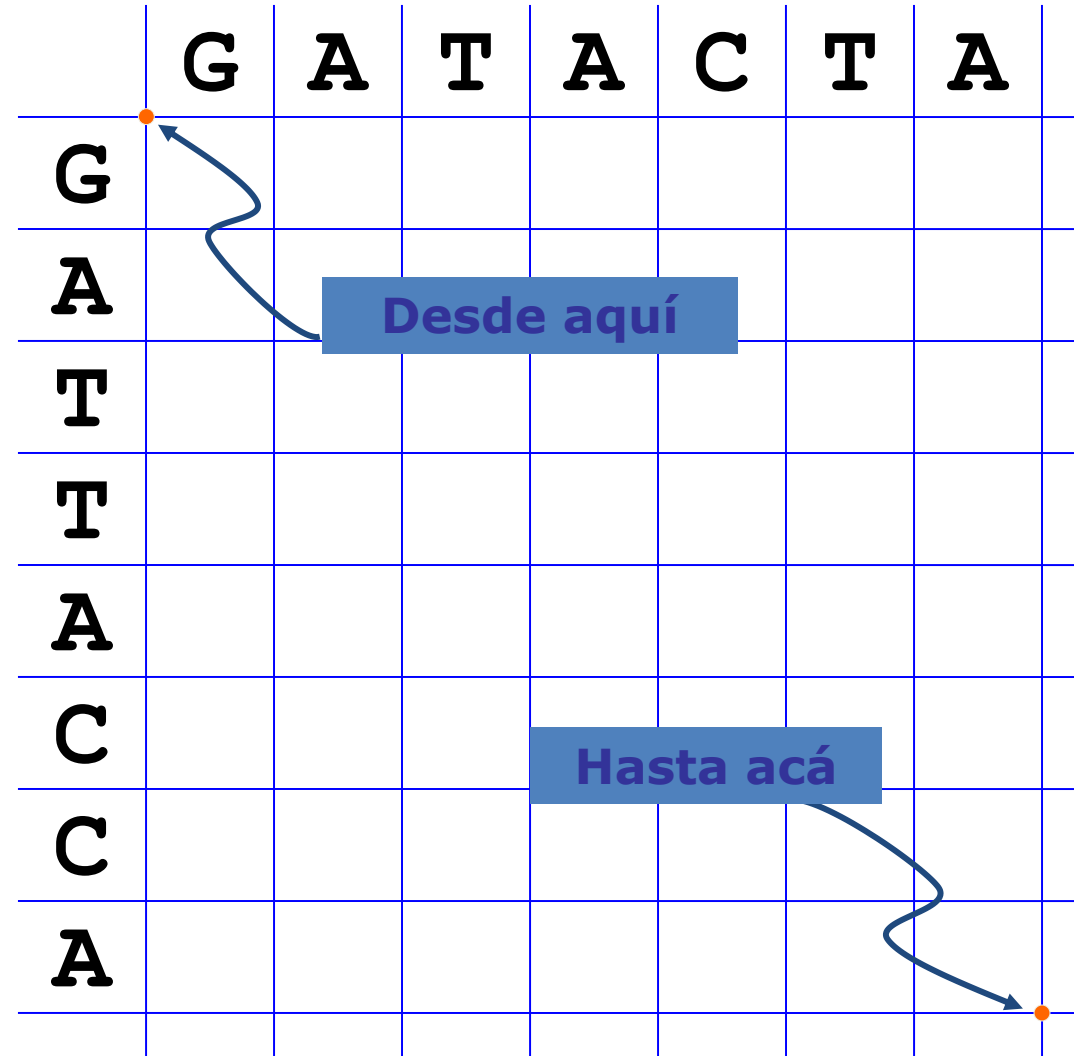
Needleman & Wunsch (1970)

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

Slides Dynamic Programming: Hugues Sicotte (NCBI)

Dynamic programming: ejemplo (cont.)

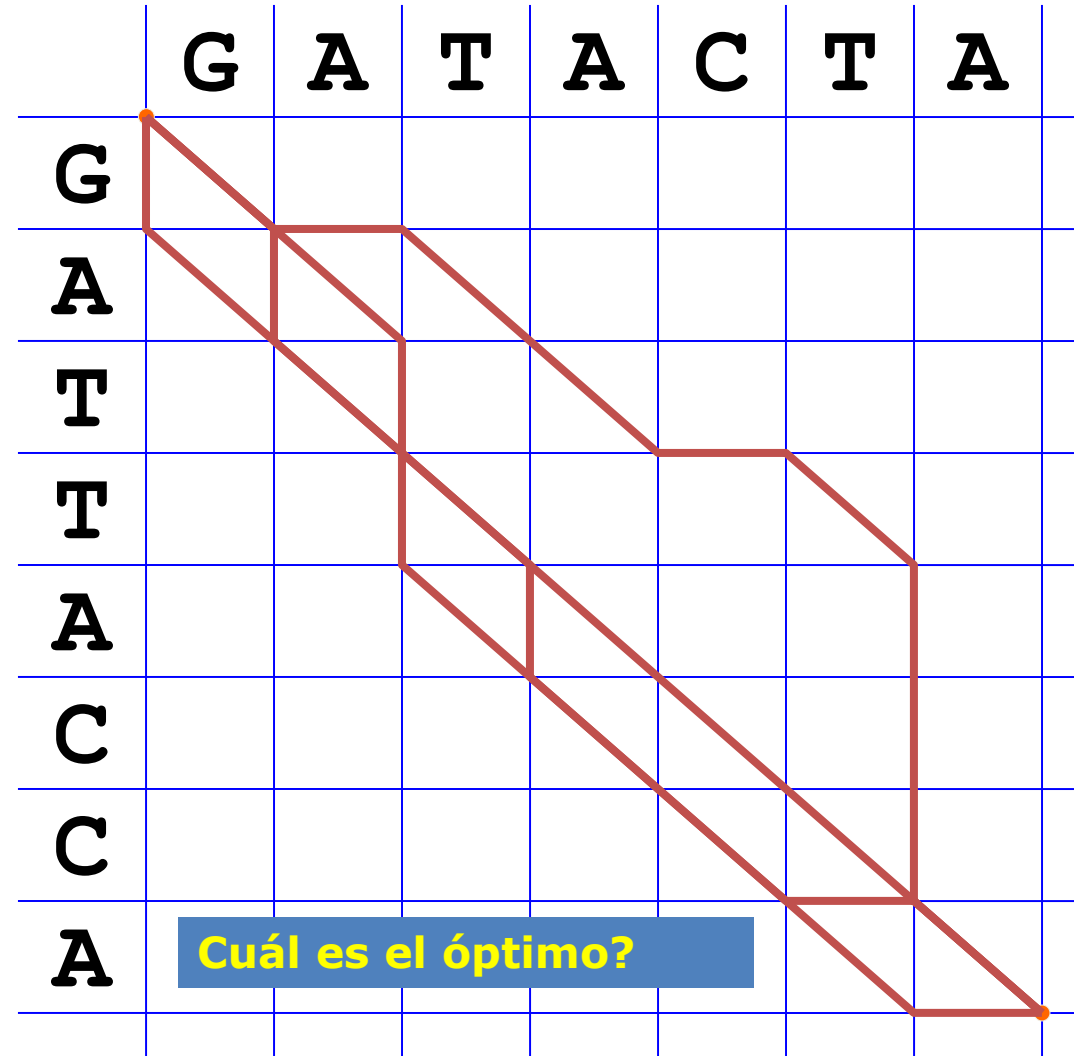
El objetivo es encontrar
la ruta (path) óptimo



Slides Dynamic Programming: Hugues Sicotte (NCBI)

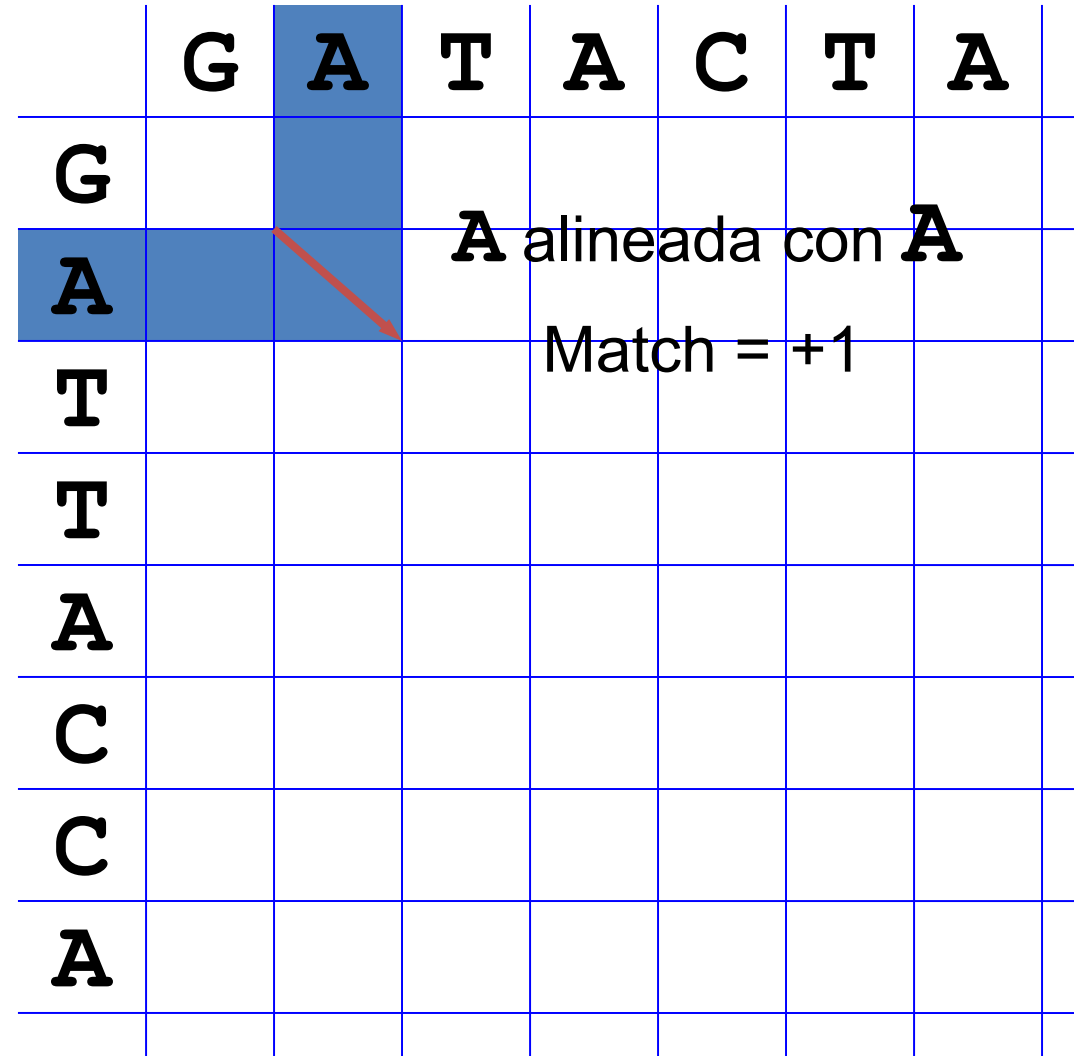
Dynamic programming: paths posibles

Cada path corresponde a
un alineamiento único



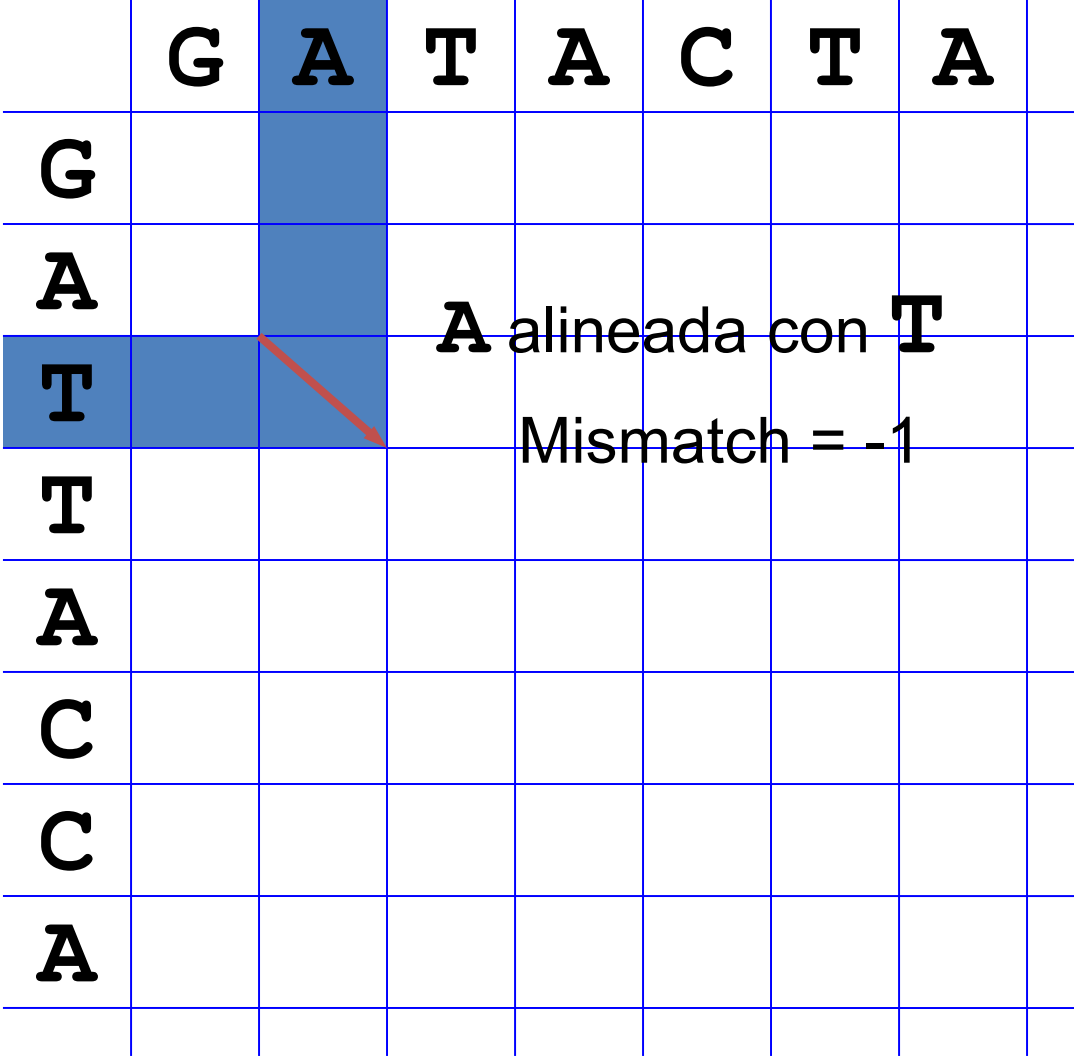
Dynamic programming: scores: match

El score para una ruta (path) es la suma incremental de los scores de sus pasos (diagonales o lados).



Dynamic programming: scores: mismatch

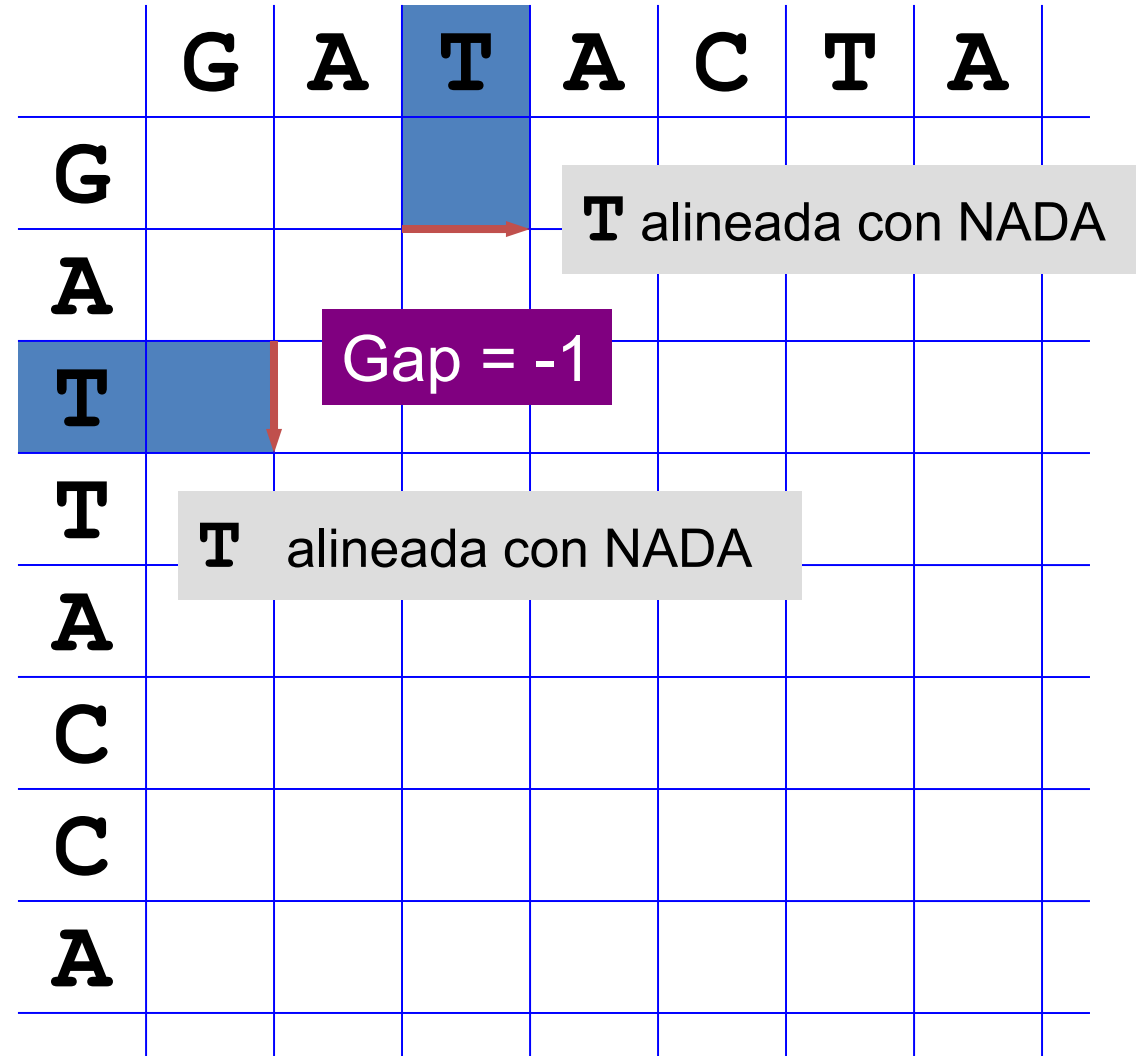
El score para una ruta (path) es la suma incremental de los scores de sus pasos (diagonales o lados).



Slides Dynamic Programming: Hugues Sicotte (NCBI)

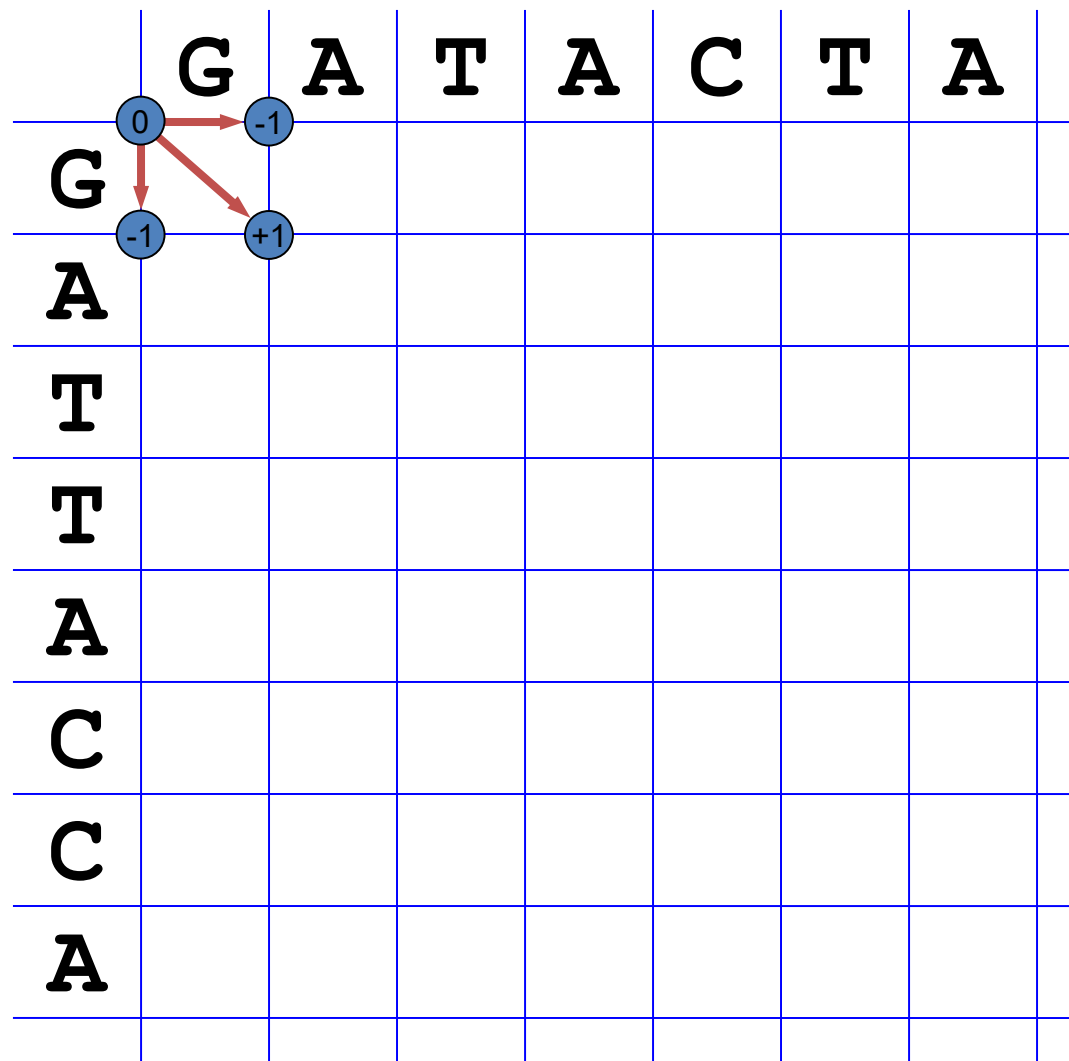
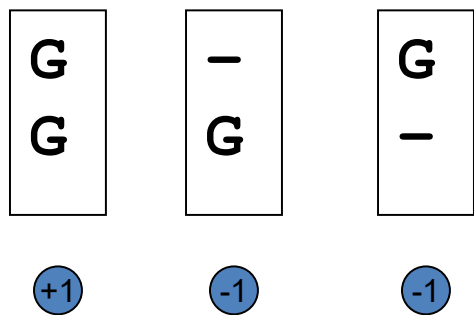
Dynamic programming: scores: gaps

El score para una ruta (path) es la suma incremental de los scores de sus pasos (diagonales o lados).



Dynamic programming: paso a paso (1)

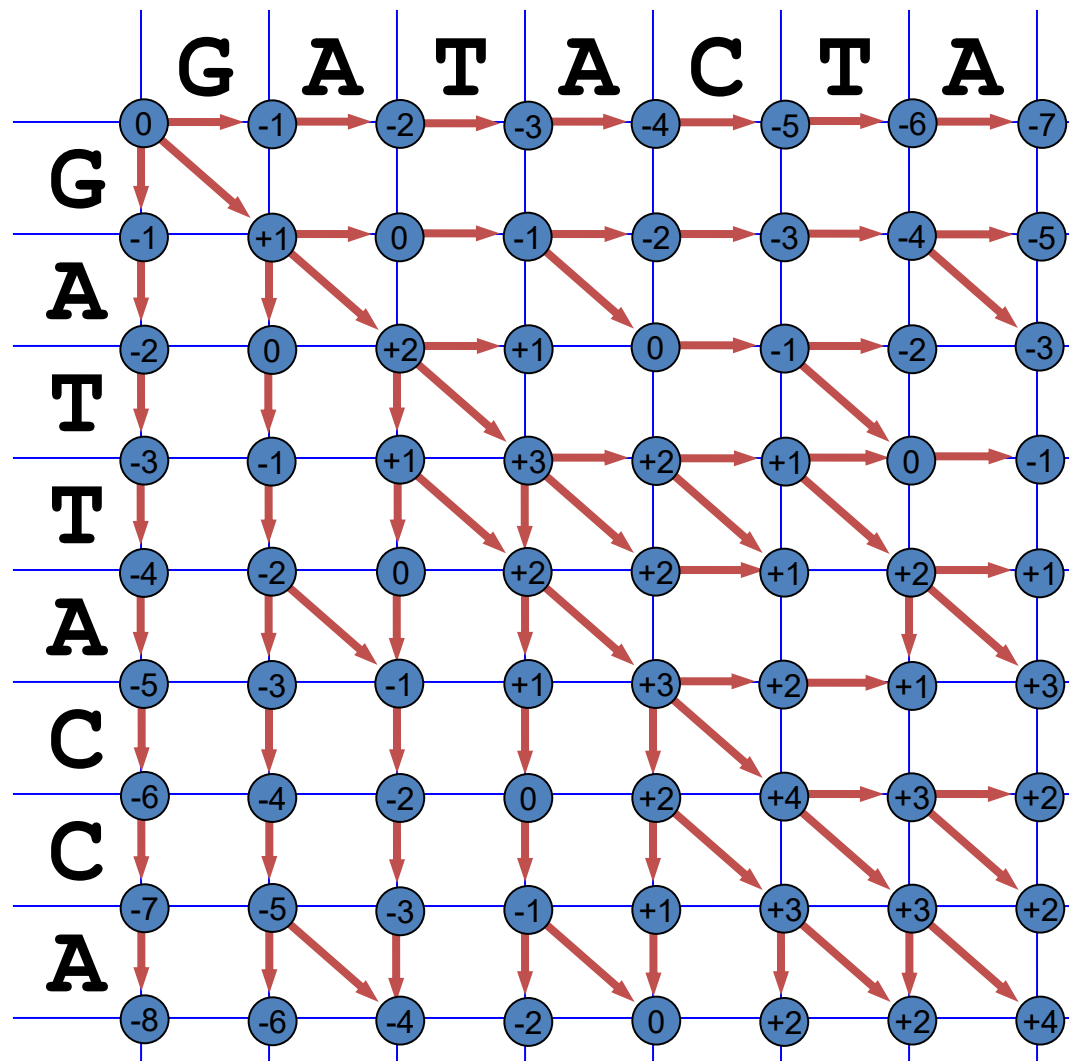
Extender el path paso por paso



Dynamic programming: paso a paso (7)

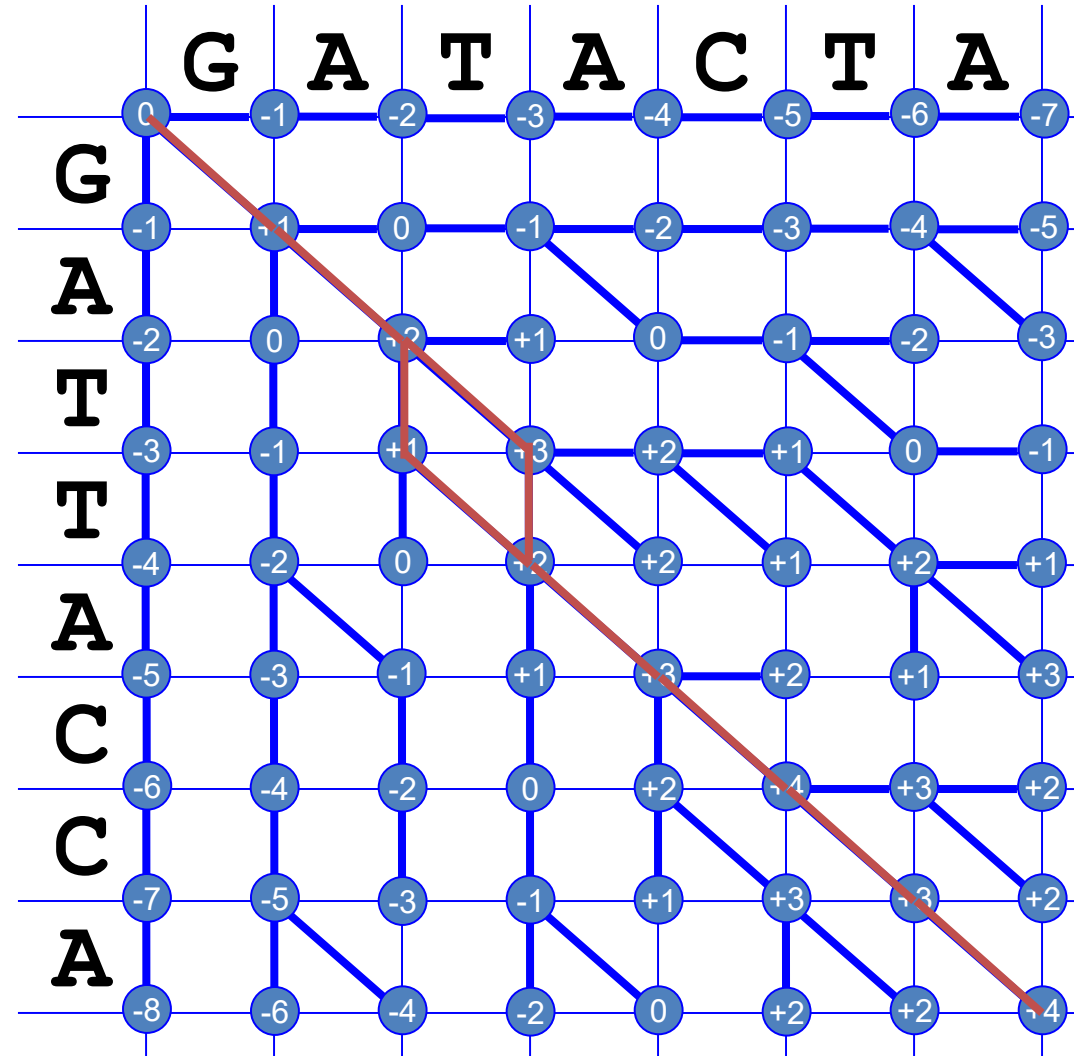
Incrementar el path paso a paso

Recordar el mejor subpath que lleva a cada punto en la matriz.



Dynamic programming: best path

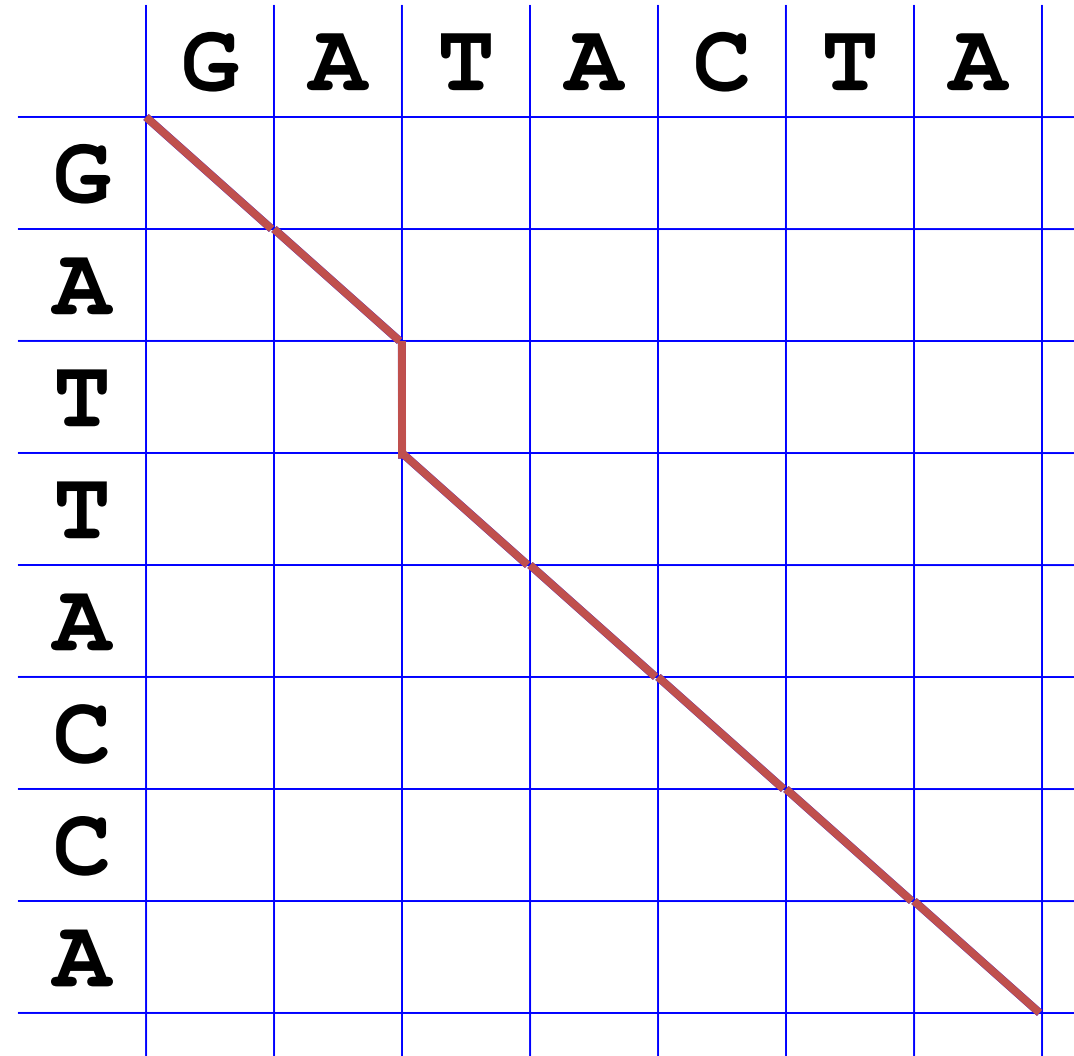
Recorrer el camino de atrás hacia adelante para obtener el mejor path y alineamiento.



Slides Dynamic Programming: Hugues Sicotte (NCBI)

Imprimir el alineamiento

GA-TACTA
GATTACCA



1981 Temple Smith & Michael Waterman

Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25;147(1):195-7. doi: 10.1016/0022-2836(81)90087-5. PMID: 7265238.



Temple Smith



Michael Waterman

- El método de NW fue modificado por **Smith & Waterman** para obtener **alineamientos locales**
- El método garantiza la obtención de un **alineamiento óptimo** (cuyo score no puede ser mejorado)
- La complejidad es proporcional al producto de las longitudes de las secuencias a alinear ($M \times N$)

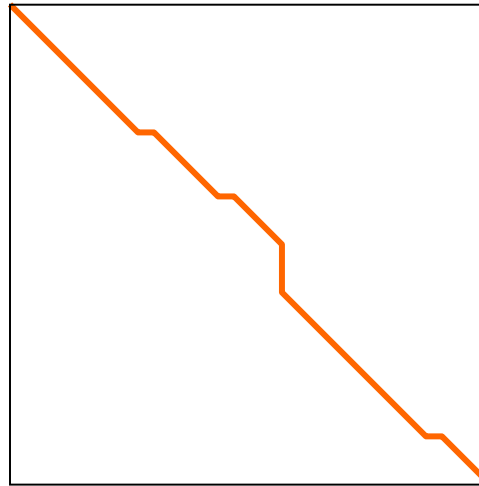
https://en.wikipedia.org/wiki/Smith-Waterman_algorithm

Similitud global y local

El algoritmo de programación dinámica puede ser implementado para alineamientos locales o globales.

Optimal global alignment

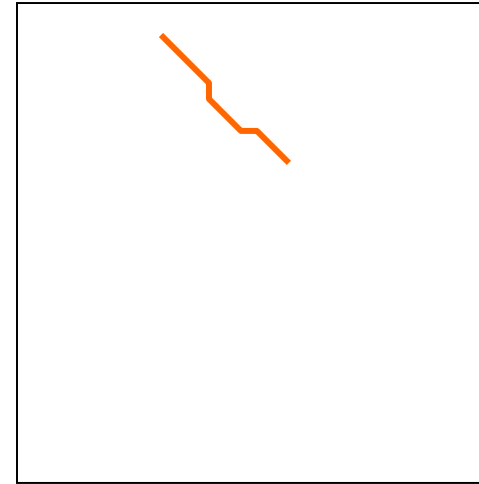
Needleman & Wunsch (1970)



Las secuencias se alinean esencialmente de un extremo a otro

Optimal local alignment

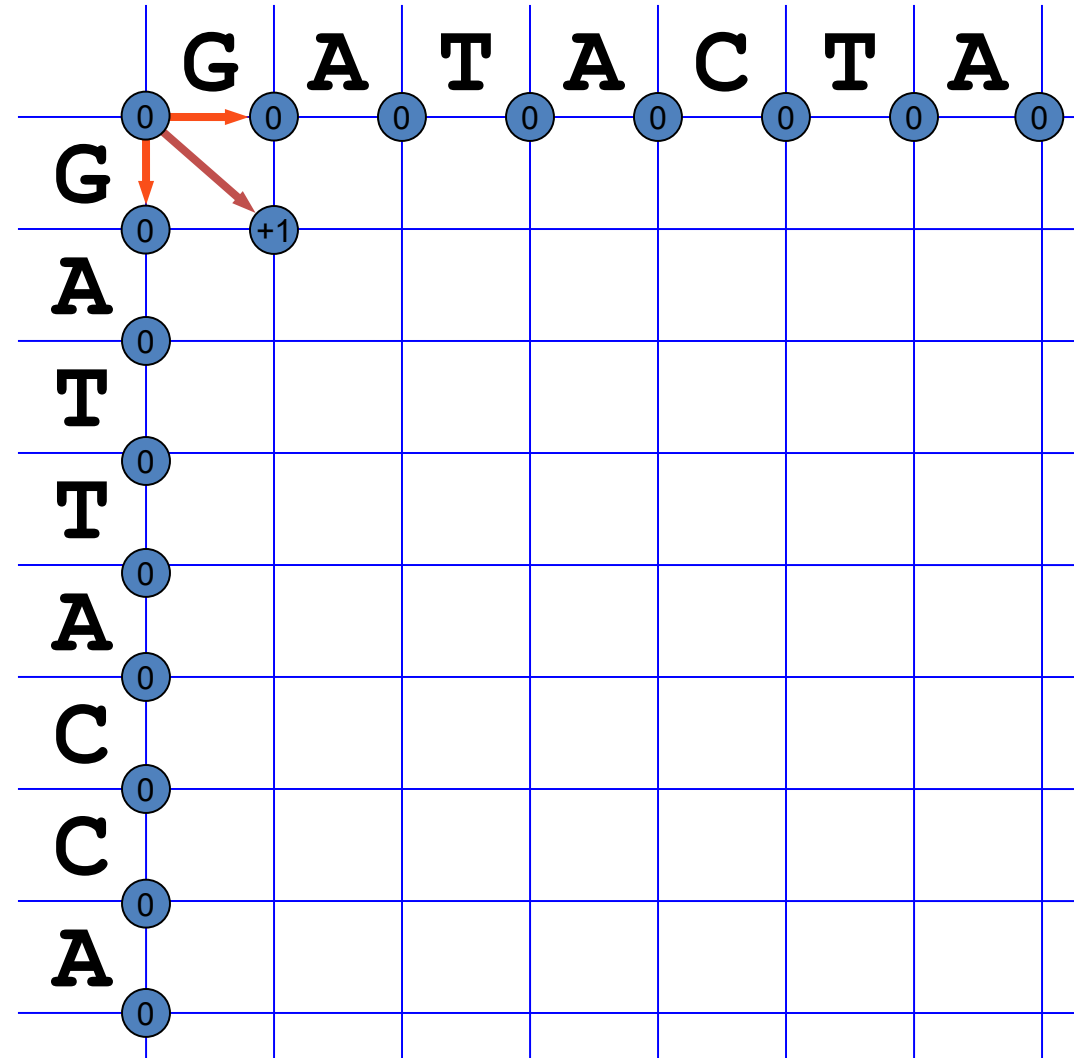
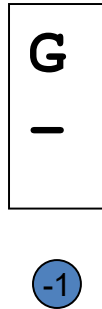
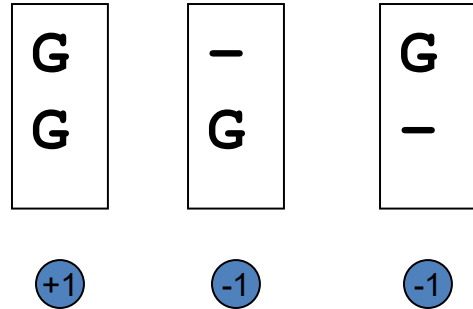
Smith & Waterman (1981)



Las secuencias se alinean en regiones pequeñas y aisladas

Smith-Waterman: paso a paso

Extender el path paso por paso

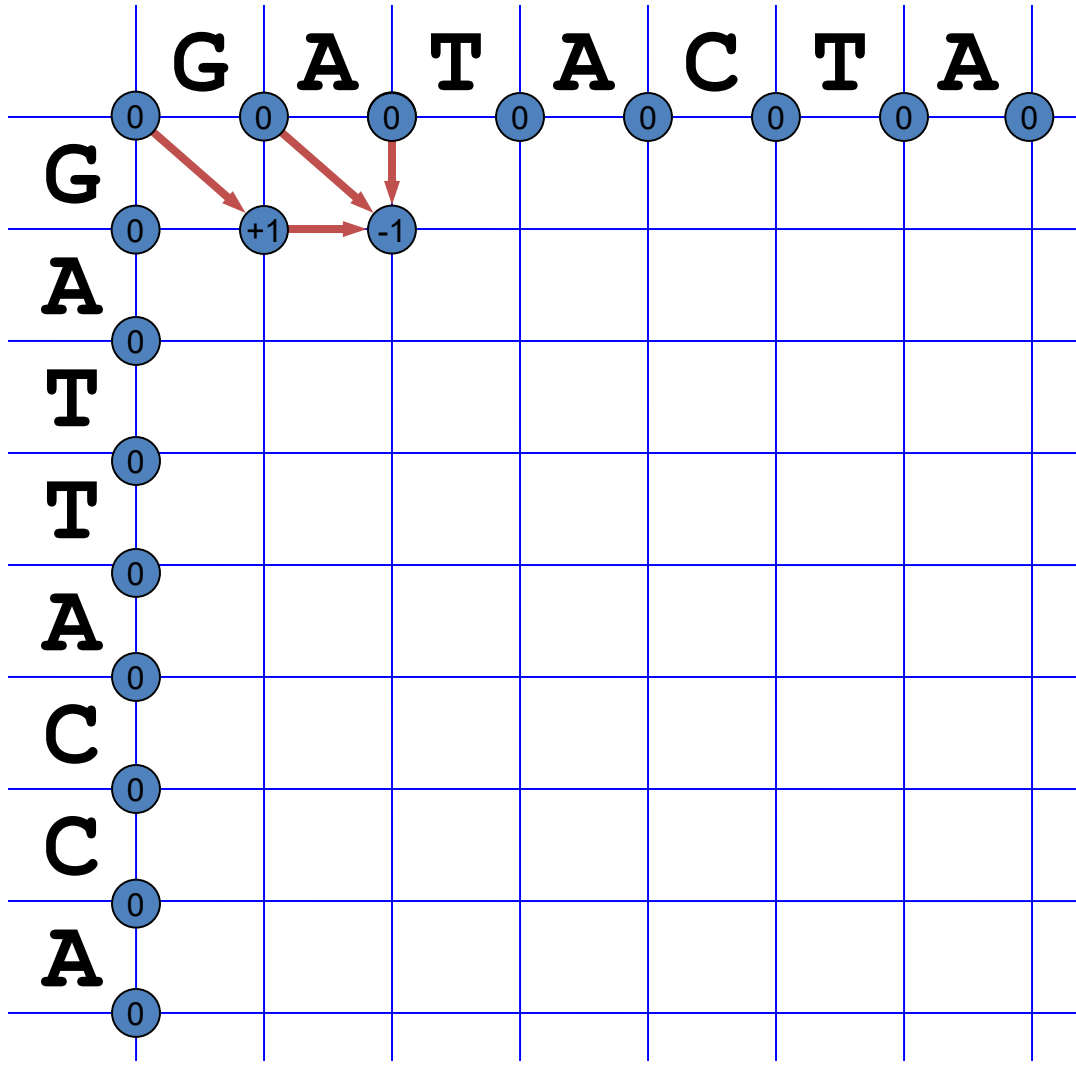


Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25;147(1):195-7. doi: 10.1016/0022-2836(81)90087-5. PMID: 7265238.

Smith-Waterman: paso a paso (3)

Incrementar el path paso a paso

Recordar el mejor subpath que lleva a cada punto en la matriz.



- Needleman-Wunsch web app desarrollada para enseñanza
 - <http://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Needleman-Wunsch>

Input:

Sequence *a*:

AATCG

Sequence *b*:

AACG

Optimization of:

Distance ☐

Similarity ☒

Scoring in *s*:

Match

Mismatch

Gap

Hint:

For similarity maximization,
match scores should be positive and all other scores lower.
For distance minimization the reverse applies.

Recursion:

$$D_{i,j} = \max \begin{cases} D_{i-1,j-1} + s(a_i, b_j) \\ D_{i-1,j} + s(a_i, -) \\ D_{i,j-1} + s(-, b_j) \end{cases} = \max \begin{cases} D_{i-1,j-1} + 1 & a_i = b_j \\ D_{i-1,j-1} + -1 & a_i \neq b_j \\ D_{i-1,j} + -2 & b_j = - \\ D_{i,j-1} + -2 & a_i = - \end{cases}$$

Output:

<i>D</i>		<i>A</i> ₁	<i>A</i> ₂	<i>C</i> ₃	<i>G</i> ₄
	0	-2	-4	-6	-8
<i>A</i> ₁	-2	1	-1	-3	-5
<i>A</i> ₂	-4	-1	2	0	-2
<i>T</i> ₃	-6	-3	0	1	-1
<i>C</i> ₄	-8	-5	-2	1	0
<i>G</i> ₅	-10	-7	-4	-1	2

Score: 2

Results

You can select a result to get the related traceback.

AATCG

** **

AA _CG

- **Smith-Waterman web app desarrollada para enseñanza**
 - <http://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Smith-Waterman>

Input:

Sequence *a*:

AATCG

Sequence *b*:

AACG

Scoring in *s*:

Match

1

Mismatch

-1

Gap

-2

Hint:

For similarity maximization,
match scores should be positive and all other scores lower.

Recursion:

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(a_i, b_j) \\ S_{i-1,j} + s(a_i, -) \\ S_{i,j-1} + s(-, b_j) \\ 0 \end{cases} = \max \begin{cases} S_{i-1,j-1} + 1 & a_i = b_j \\ S_{i-1,j-1} + -1 & a_i \neq b_j \\ S_{i-1,j} + -2 & b_j = - \\ S_{i,j-1} + -2 & a_i = - \\ 0 \end{cases}$$

Output:

<i>S</i>		<i>A</i> ₁	<i>A</i> ₂	<i>C</i> ₃	<i>G</i> ₄
	0	0	0	0	0
<i>A</i> ₁	0	1	1	0	0
<i>A</i> ₂	0	1	2	0	0
<i>T</i> ₃	0	0	0	1	0
<i>C</i> ₄	0	0	0	1	0
<i>G</i> ₅	0	0	0	0	2

Score: 2

Results

You can select a result to get
the related traceback.

AA

**

AA

CG

**

CG

- **Un algoritmo de alineamiento local, siempre produce alineamientos locales?**
- **Un algoritmo de alineamiento global siempre produce alineamientos globales?**

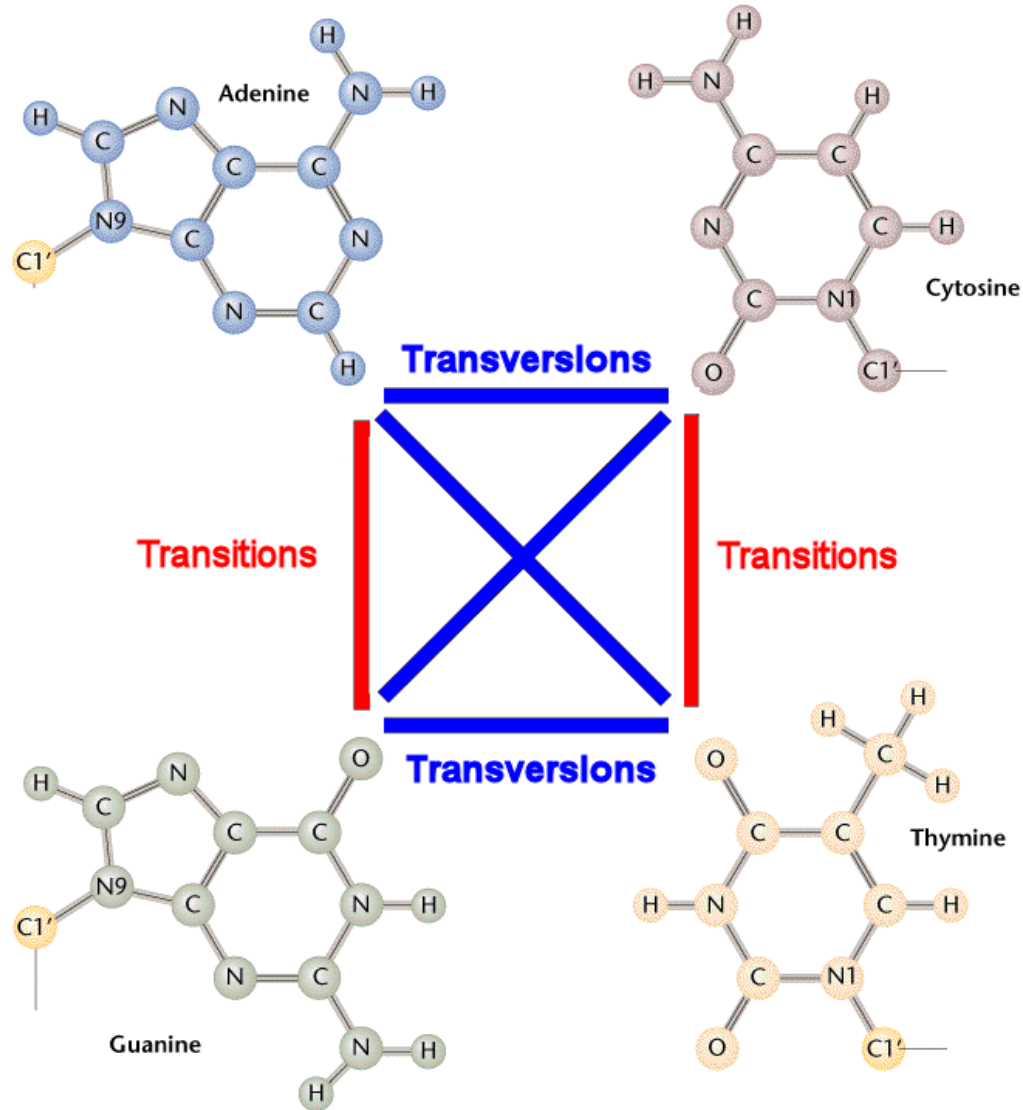
Alineamientos Globales vs Locales

NO

- **SW puede producir alineamientos que son esencialmente globales dependiendo del sistema de scoring (scores para match/mismatch/gaps)**
- **SW puede producir alineamientos que son esencialmente globales si las secuencias son parecidas en toda su longitud**
- **NW puede producir alineamientos locales dependiendo la penalidad asignada a los gaps en los extremos de un alineamiento global (o alterando significativamente el sistema de scoring)**

- **Un sistema de scoring simple, penaliza igual cualquier mismatch**
- **Biológicamente tiene sentido penalizar ciertos cambios y ser más permisivo con otros**
 - **En proteínas: residuos hidrofóbicos reemplazados entre sí.**
 - **En DNA: transversiones vs transiciones**
- **Una matriz no es otra cosa que un sistema de scoring que permite asignar puntajes individuales a cada una de las letras del alfabeto en uso.**

Transiciones vs Transversiones



IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

- **Un ejemplo de matriz de scoring podría ser el clásico ejemplo de penalizar más los cambios que alteran las propiedades químicas de un residuo (aa)**
 - **hidrofóbicos: Ile, Val, Leu, Ala**
 - **Polares (+): Lys, Arg**
 - **Polares (-): Glu, Asp**
 - **Aromáticos: Phe, Tyr, Trp**
 - **etc.**

Ile x Val = -1

Ile x Asp = -5

Phe x Tyr = -1

Phe x Gly = -8

- **Matrices PAM (Dayhoff, 1978)**

- **PAM = Point Accepted Mutations**
- **Proveen estimaciones de plausibilidad de cambio de un aminoácido en otro en proteínas homólogas**
- **Derivadas a partir de un grupo de secuencias > 85% similares**
- **Los cambios de aminoácidos observados son llamados “accepted mutations”**
- **1 PAM es el tiempo en el cual 1/100 aminoácidos se espera que hayan mutado**
- **Las matrices PAM se extrapolan matrices a períodos evolutivos más largos multiplicando por ellas mismas**



Margaret Dayhoff

- **Matrices PAM (Dayhoff, 1978)**
 - **PAM = Point Accepted Mutations**

	ORIGINAL AMINO ACID																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	2	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

PAM 1



Margaret Dayhoff

- **Matrices PAM (Dayhoff, 1978)**
 - **PAM = Point Accepted Mutations**

		ORIGINAL AMINO ACID																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
REPLACEMENT AMINO ACID	A Ala	13	6	9	9	5	8	9	12	5	8	6	7	7	4	11	11	11	2	4	9
	R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
	N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
	D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
	C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
	Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	2	1	4	3	3	1	2	3
	E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
	G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
	H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
	I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
	L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
	K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
	M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
	F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
	P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
	S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
	T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
	W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
	Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
	V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

PAM 250



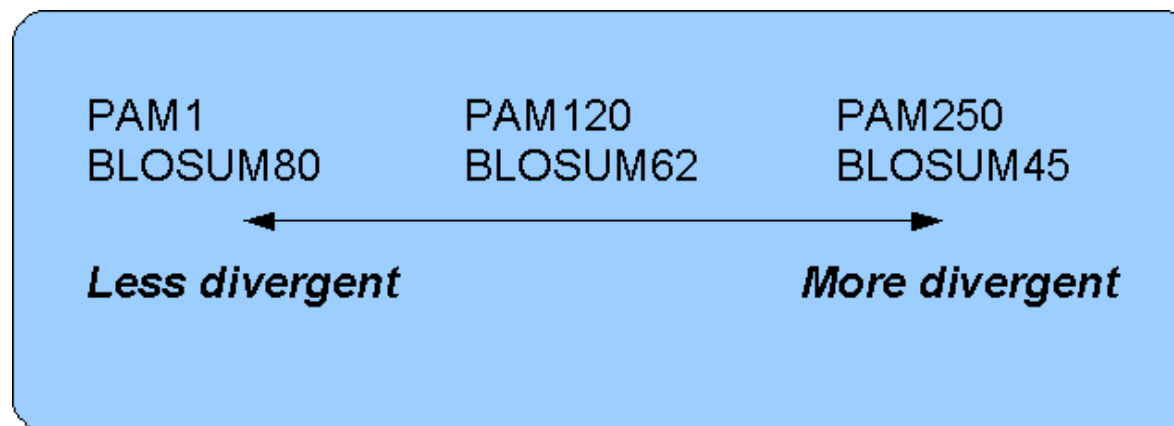
Margaret Dayhoff

- **BLOSUM (Henikoff)**

- **Blocks Amino Acid Substitution Matrices**
- **Sustituciones de amino ácidos observadas en un conjunto grande de 'blocks'**
- **Representan más de 500 familias de proteínas**
- **Se agrupan los blocks de acuerdo a su identidad y se generan matrices**
- **blocks 80% idénticos -> BLOSUM80**
- **Blocks 60% idénticos -> BLOSUM60**
- **etc**



Steven Henikoff



Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915-9. doi: 10.1073/pnas.89.22.10915. PMID: 1438297; PMCID: PMC50453.

https://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html

Sistemas de scoring: BLOSUM62

Algunas sustituciones son más comunes que otras

Los scores provienen de la observación de los tipos y frecuencias de sustitución en distintas familias proteicas

BLOSUM62

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

Sistemas de scoring: BLOSUM62: identidades

Las identidades tienen scores positivos, pero algunas son más valoradas que otras.

BLOSUM62

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

Sistemas de scoring: BLOSUM62: sustituciones

Algunas sustituciones tienen scores positivos, pero la mayoría son negativos.

BLOSUM62

[illegible]

- **PAM**
- **BLOSUM**
- **Otras**
 - **Comparación simple de propiedades químicas de amino ácidos**
 - **Análisis complejos de sustituciones en estructura secundaria de proteínas, a partir de alineamientos estructurales**
 - **Gonnet (1994). Sustitución de dipéptidos**
 - **Jones (1994) matriz específica de proteínas transmembrana**

Gonnet GH, Cohen MA, Benner SA. Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix. Biochem Biophys Res Commun. 1994 Mar 15;199(2):489-96. doi: 10.1006/bbrc.1994.1255. PMID: 8135790.

Jones DT, Taylor WR, Thornton JM. A mutation data matrix for transmembrane proteins. FEBS Lett. 1994 Feb 21;339(3):269-75. doi: 10.1016/0014-5793(94)80429-x. PMID: 8112466.

Referencias

Bioinformatics, 4th Ed. Baxevanis, Bader, Wishart, Wiley (2020)

Bioinformatics. Sequence and Genome analysis. David W Mount, CSHL Press (2001)

Introduction to Bioinformatics. Lesk, A. M. (2019). Oxford University Press.

Hugues Sicotte (NCBI). (slides Dynamic Programming)

Javascript-based implementations for various algorithms
<http://rna.informatik.uni-freiburg.de/Teaching/>

Will this day never end????

someecards
user card

