Alineamiento de pares de secuencias

Rodrigo Santamaría

Alineamiento de pares de secuencias

Introducción

Definiciones

Ejemplo

Algoritmos

Matrices de puntuación

Objetivo

- Determinar si una secuencia de nucleótidos o aminoácidos (un gen o una proteína) está relacionada con otra
- Esto nos permite determinar
  - Si evolucionaron desde un ancestro común
  - Si tienen funciones comunes
  - En el caso de proteínas, si tienen formas similares

Homología

- Dos secuencias son **homólogas** si comparten un ancestro común
  - No hay grados, o se es homólogo o no (~~"soy un 30% tu padre"~~)
  - Dos proteínas homólogas suelen tener una estructura 3D similar
  - Dos proteínas o genes homólogos suelen tener secuencias parecidas

Ortología y Paralogía

- **Ortólogo**: secuencias homólogas en diferentes diferentes especies que vienen de un ancestro común
  - P. ej.: el gen de la mioglobina en ratas y humanos tiene un ancestro común hace 80 millones de años (MYA)
- **Parálogo**: secuencias homólogas pero debidas a un mecanismo distinto a la evolución
  - Típicamente, duplicación genética
- Simplificación (no totalmente equivalente)
  - Ortólogo: homólogo entre diferentes especies
  - Parálogo: homólogo dentro de una misma especie

Similitud e Identidad

- **Similitud**: grado de coincidencia entre dos secuencias
  - "El grado de similitud entre dos secuencias es de un 45%"
- **Identidad**: coincidencia total entre dos secuencias
  - Aunque muchas veces se usa como sinónimo de similitud
- **Analogía**: grado muy alto de similitud entre dos secuencias
- La homología no siempre garantiza analogía
  - La β-globina y la neuroglobina sólo comparten un 22% de sus secuencias a pesar de ser homólogas

Alineamiento de pares

- Colocación de dos secuencias para que se maximice su similitud

```
ROJO
ROSSO   +2
ROUGE   +2
RED     +1
```

```
ROJ—O
ROSSO
**   *  +3 (75%)
```

```
RO—JO
ROUGE
**     +2 (50%)
```

```
ROJO
RED—
*       +1 (25%)
```

- Las secuencias son largas y la capacidad de combinación alta
  - Necesidad de métodos algorítmicos para realizar el alineamiento

Huecos (gaps)

- Separaciones añadidas artificialmente a una secuencia para maximizar su alineamiento con otra(s).
  - Representan posibles mutaciones sufridas durante la evolución que han causado divergencia entre las secuencias:
    - Inserciones
    - Deleciones
- Un hueco puede ocurrir en medio o en los extremos de la secuencia

Alineamiento global y local

- **Alineamiento global**: las secuencias se alinean a lo largo de toda su longitud, intentando alinear secuencias completas
  - Se introducen huecos para igualar las longitudes de secuencia
  - Útil para secuencias muy parecidas y de longitud similar
- **Alineamiento local**: sólo se alinean la partes más parecidas de la secuencia
  - Favorece encontrar patrones similares dentro de la secuencia
  - Un alineamiento local es una combinación de muchos alineamientos globales de secuencias cortas.

Alineamiento de pares. Ejemplo

- Un ejemplo biológico: α-globina (NCBI NP_00508) y β-globina (NP_000509) en humano
  - Proteínas homólogas con estructura 3D similar.
- Algoritmo BLAST de pares del NCBI para proteínas
  - http://blast.ncbi.nlm.nih.gov/Blast.cgi

1. "Enter query sequence":
   1. introducir NP_000508 o su secuencia FASTA
   2. Seleccionar "align two or more sequences
2. "Enter subject sequence": introducir NP_000509 o su FASTA
3. Click "BLAST"

NCBI/ BLAST/ blastp suite

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

BLASTP programs search protein subjects using a protein query. more...

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>gi|4504345|ref|NP_000508.1| hemoglobin subunit alpha [Homo sapiens]
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVAD
ALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLAS
VSTVLTSKYR
```

Query subrange

From

To

Or, upload file Seleccionar archivo No se h...rchivo

Job Title gi|4504345|ref|NP_000508.1| hemoglobin subunit...

Enter a descriptive title for your BLAST search

☑ **Align two or more sequences**

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence Clear

NP_000509

Subject subrange

From

To

Or, upload file Seleccionar archivo No se h...rchivo

Program Selection

Algorithm

◉ blastp (protein-protein BLAST)

Choose a BLAST algorithm

BLAST Search protein sequence using Blastp (protein-protein BLAST)

☑ **Show results in a new window**

```
Length=147

GENE ID: 3043 HBB | hemoglobin, beta [Homo sapiens] (Over 100 PubMed links)

Score =  114 bits (286),  Expect = 5e-31, Method: Compositional matrix adjust.
Identities = 63/145 (43%), Positives = 88/145 (61%), Gaps = 8/145 (6%)

Query  3    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQV  56
            L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F DLS      G+ +V
Sbjct  4    LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61

Query  57   KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  116
            K HGKKV  A ++ +AH+D++    + LS+LH  KL VDP NF+LL + L+  LA H
Sbjct  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121

Query  117  EFTPAVHASLDKFLASVSTVLTSKY  141
            EFTP V A+  K +A V+  L  KY
Sbjct  122  EFTPPVQAAYQKVVAGVANALAHKY  146
```

- **Query y Sbjct**: secuencias. Ambas se alinean para maximizar su similitud
- **Coincidencias**: línea entre Query y Sbjct
  - Letra: coincidencia idéntica
  - +: coincidencia conservada (aminoácidos básicos, ácidos, hidrófobos…)
  - Espacio en blanco: no hay coincidencia
- **Score**: valor del algoritmo de alineamiento
- **Identities/Positives**: porcentaje de coincidencias idénticas/conservadas
- **Gaps**: porcentaje de huecos incluidos en el alineamiento

Alineamiento de pares de secuencias

Introducción

Algoritmos

- Diagramas de puntos
- Alineamiento global
- Alineamiento local

Matrices de puntuación

Algoritmos de alineamiento

- Procedimientos de comparación entre dos secuencias mediante alineamientos óptimos
- Una secuencia por sí sola no es muy informativa
  - Necesitamos compararla con otras para descubrir relaciones
    - Funcionales: dos secuencias tienen la misma función o similar
    - Estructurales: dos secuencias de aminoácidos (proteínas) tienen la misma estructura 3D
    - Evolutivas: dos secuencias proceden de un ancestro común

Tipos de algoritmos de alineamiento

- Diagramas de puntos
- Algoritmos dinámicos
  - Alineamiento global
    - Needleman and Wunsch (1970)
  - Alineamiento local
    - Smith and Waterman (1981)

Diagramas de puntos (dot plots)

- Es el algoritmo más sencillo
  - No requiere computación (data de 1970)
    - Aunque se puede programar
  - No requiere hipótesis biológicas
  - Verificación visual
- Se coloca una secuencia en el eje X y otra en el eje Y
  - En cada lugar donde las secuencias coincidan se dibuja un punto
  - Las secuencias similares aparecerán en forma de líneas diagonales

Dot plots: ejemplo

TCCGACTTGAGATTACAGAAGTCG

Cada vez que haya una coincidencia en un nucleótido, colocamos un punto.

Por ejemplo, en el caso del primer nucleótido de la secuencia horizontal (T)

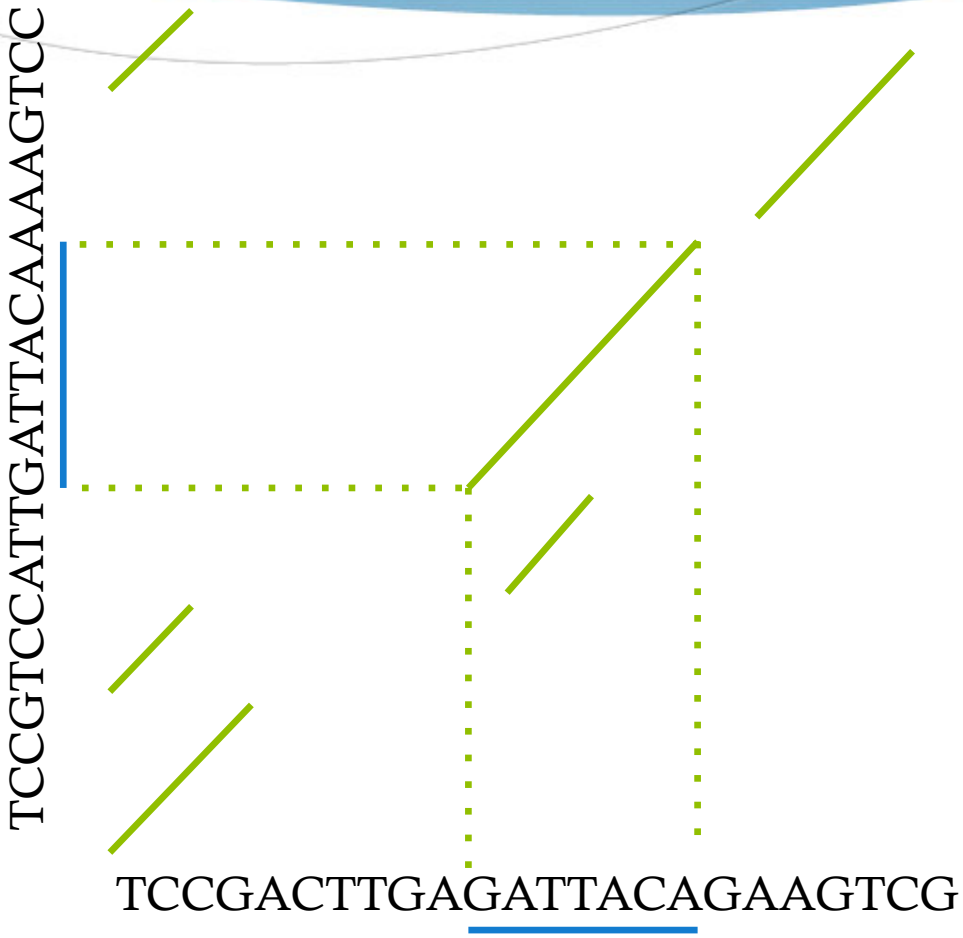
Dot plots: ejemplo

TCCGACTTGAGATTACAGAAGTCG

Continuamos con secuencias de 2, 3, 4, etc.

Se suele establecer un umbral mínimo, o tamaño de ventana (p. ej. 3)

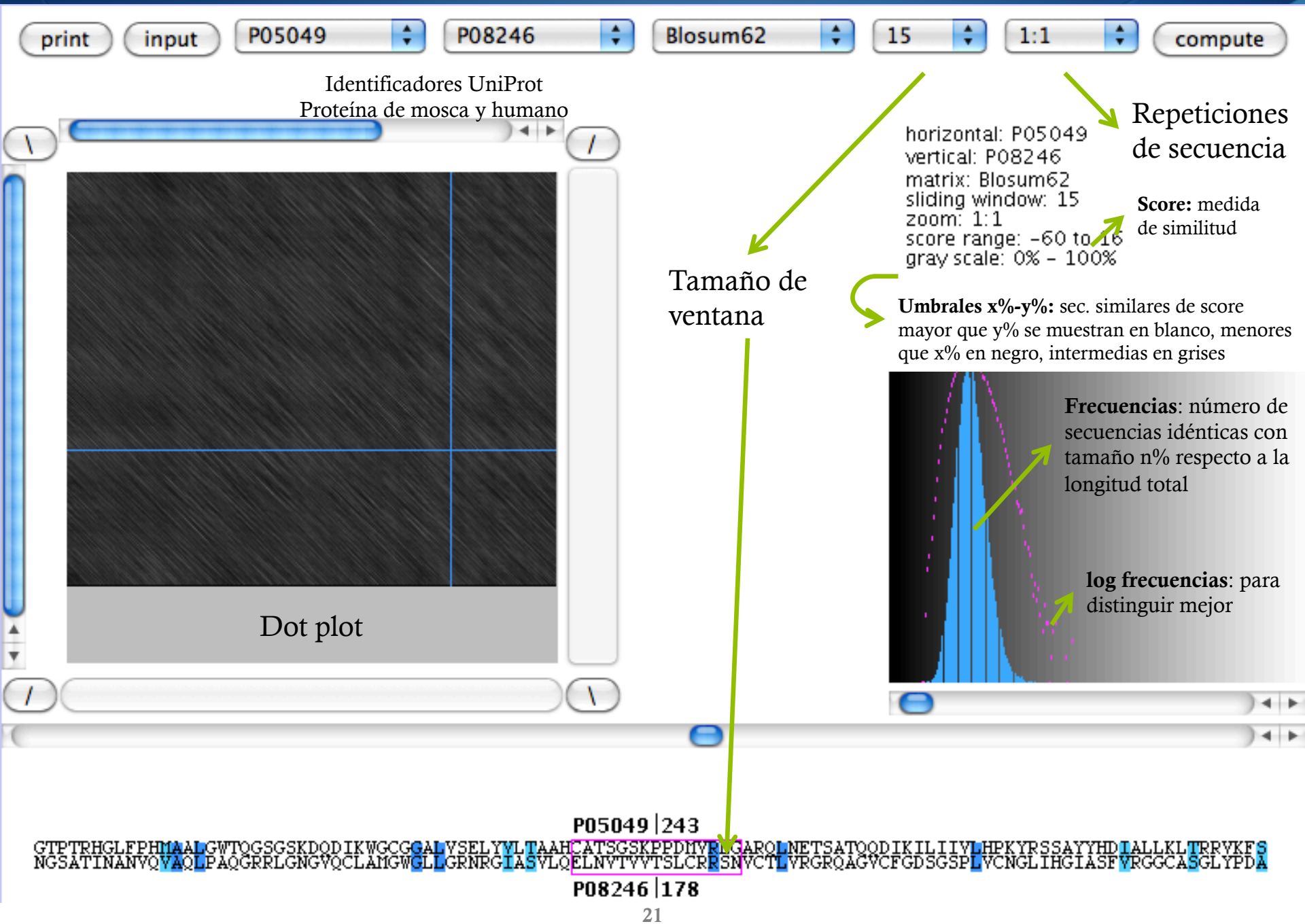
Dot plots: ejemplo

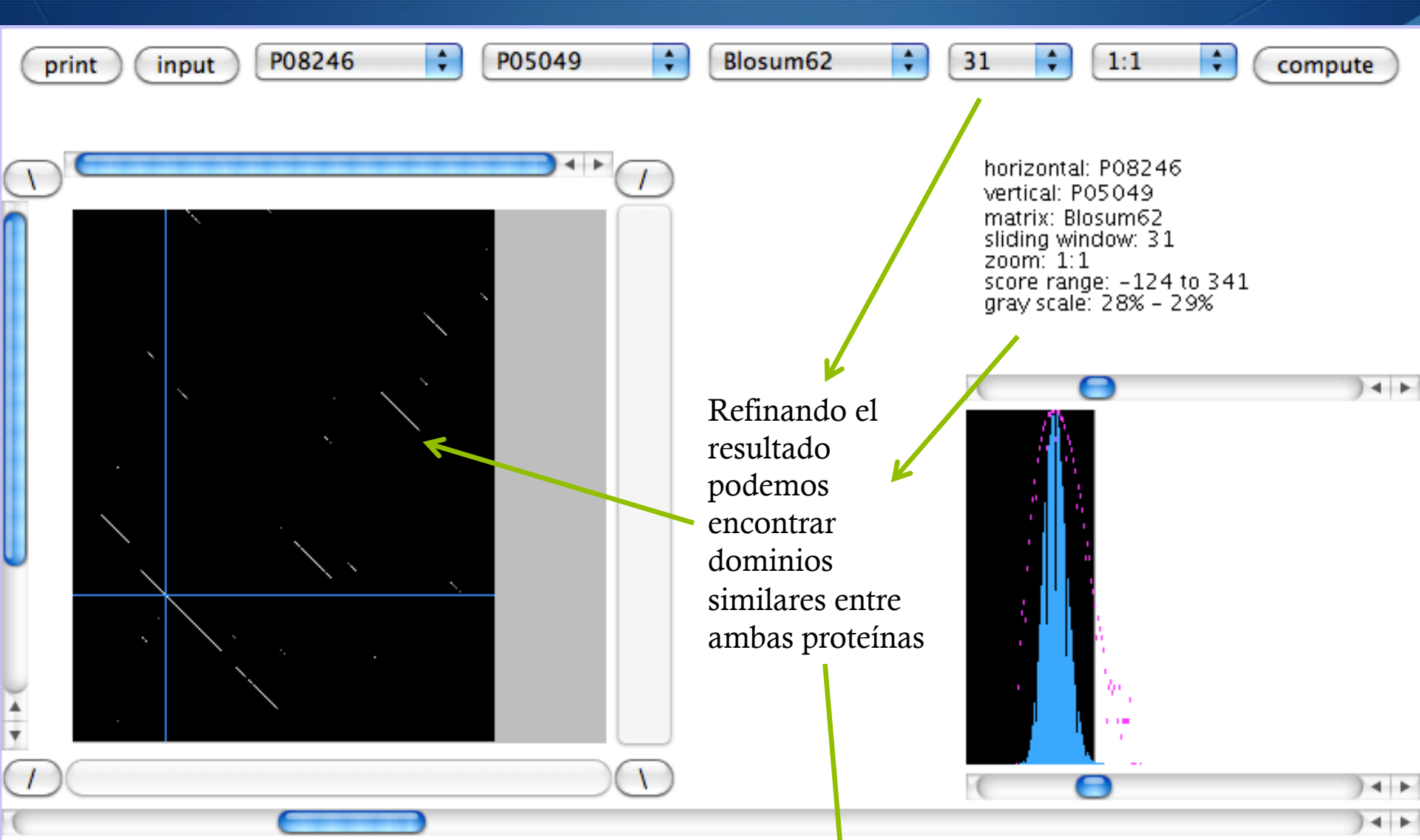


Los puntos suelen sustituirse con líneas

Dot plots: Dotlet

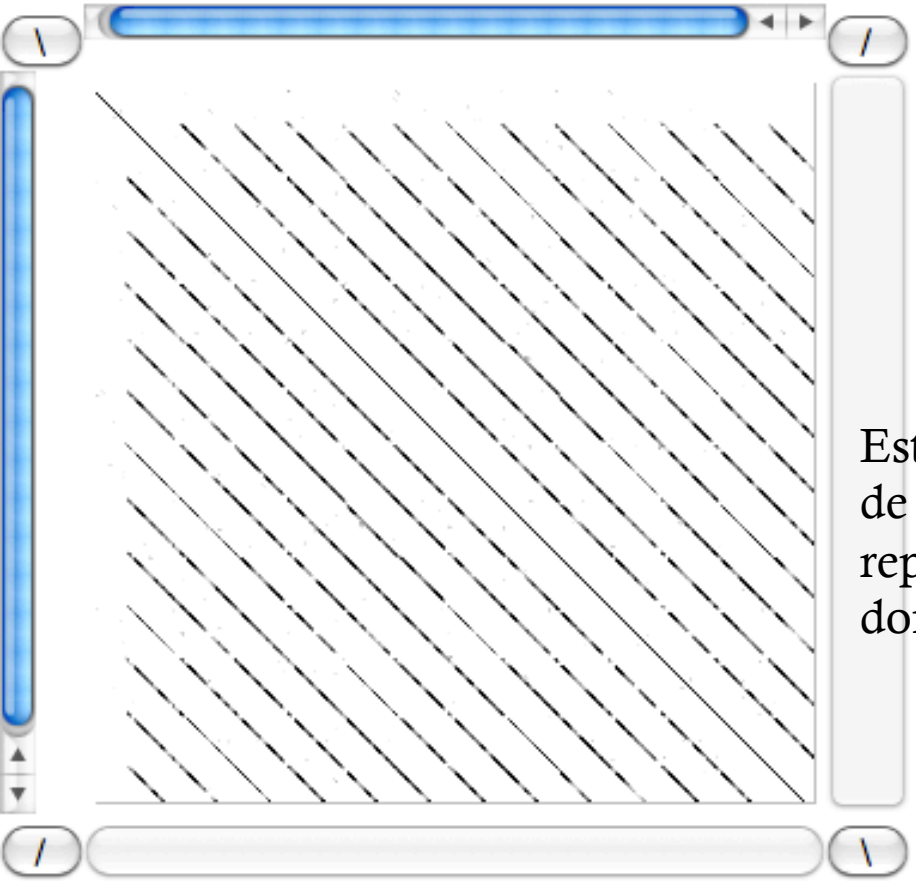
- Applet Java para cálculo y visualización de dot plots
  - Applet: pequeño programa que se carga en un navegador y que se ejecuta en la máquina cliente
  - Java: lenguaje de programación
- Fácil de usar
  - Regla general: “buscar una señal clara”
  - Veremos varios ejemplos
  - URL: http://myhits.isb-sib.ch/cgi-bin/dotlet


print
input
P05049
P08246
Blosum62
15
1:1
compute
Identificadores UniProt
Proteína de mosca y humano
horizontal: P05049
vertical: P08246
matrix: Blosum62
sliding window: 15
zoom: 1:1
score range: -60 to 16
gray scale: 0% - 100%
Repeticiones de secuencia
Score: medida de similitud
Tamaño de ventana
Umbrales x%-y%: sec. similares de score mayor que y% se muestran en blanco, menores que x% en negro, intermedias en grises
Frecuencias: número de secuencias idénticas con tamaño n% respecto a la longitud total
log frecuencias: para distinguir mejor
Dot plot
P05049 243
GTPTRHGLFPHMAALGWTQGSGSKDQDIKWGCGGALVSELYVLTAAHCATSGSKPPDMVRVGAPQLNETSATQQDIKILIIVLHPKYRSSAYYHDIALLKLTRRVKFS
NGSATINANVQVAQLPAQGRRLGNGVQCLAMGWGLLGRNRGIASVLQELNVTVVTSLCRRSNVCTLVRGRQAGVCFGDSGSPLVCNGLIHGIASFVRGGCASGLYPDA
P08246 178

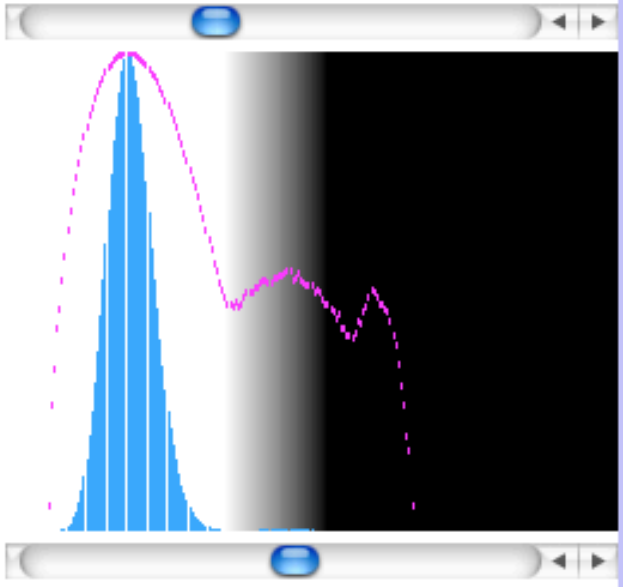

print
input
P08246
P05049
Blosum62
31
1:1
compute
horizontal: P08246
vertical: P05049
matrix: Blosum62
sliding window: 31
zoom: 1:1
score range: -124 to 341
gray scale: 28% - 29%
Refinando el resultado podemos encontrar dominios similares entre ambas proteínas
P08246 | 68
CVLPALLLGGTALASEIVGGRRARPHAWPFMVSLQLRGGHFCGATLIAPNFVMSAAHCVANVNVRAVRVVLGAHNLSRREPTRQVFAVQRIFENGYDPVNLLNDIVIL
CVPSVPLIVGGTPTRHGLFPHMAALGWTQGSGSKDQDIKWGCGGALVSELYVLTAAHCATSGSKPPDMVRLGARQLNETSATQQDIKILIIVLHPKYRSSAYYHDIAL
P05049 | 233

print input globin gla globin gla Blosum62 15 1:7 compute

Globina del caracol *Biomphalaria glabrata*
(código NCBI de proteínaCAJ44466)

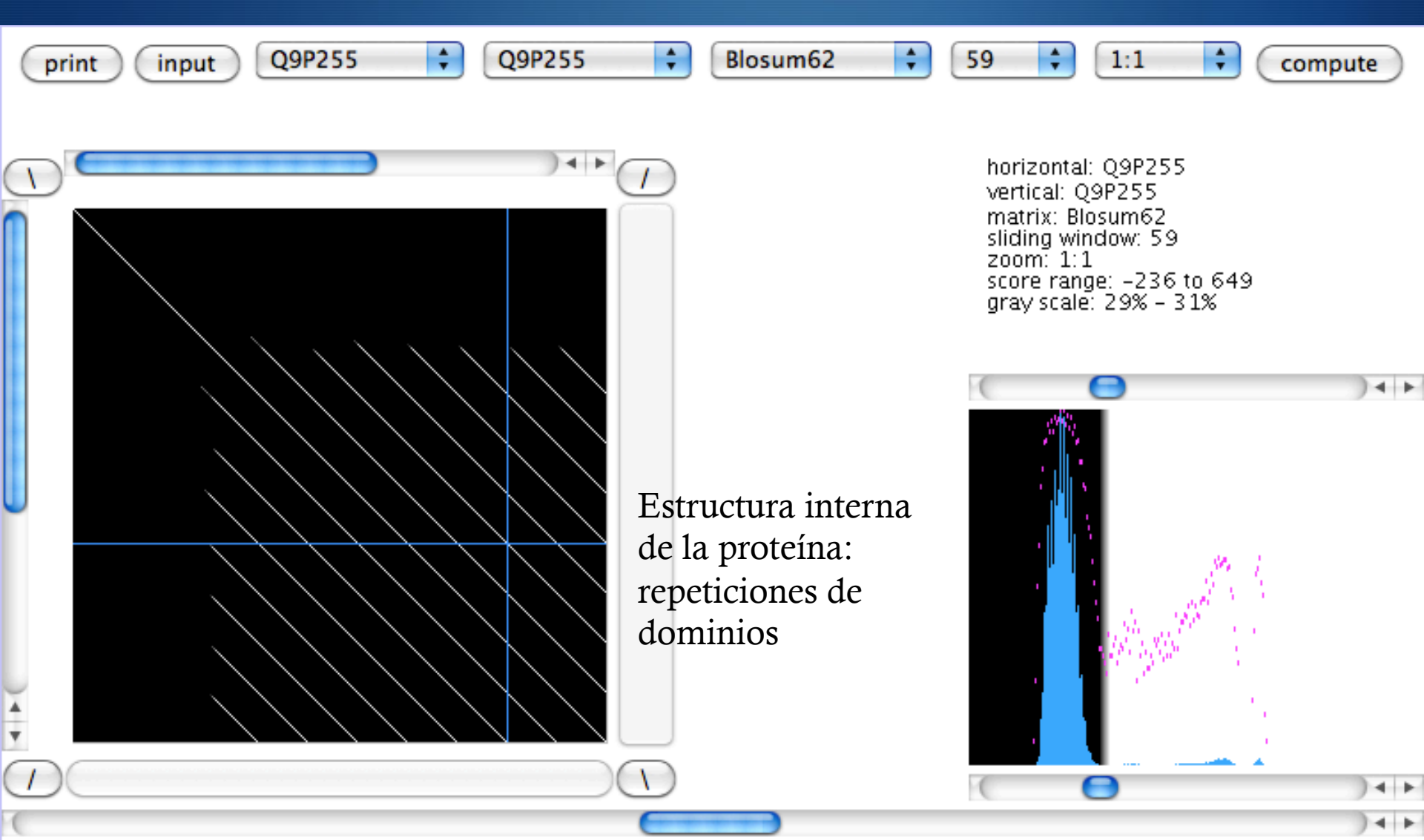


horizontal: globin gla
vertical: globin gla
matrix: Blosum62
sliding window: 15
zoom: 1:7
score range: -60 to 165
gray scale: 53% - 36%

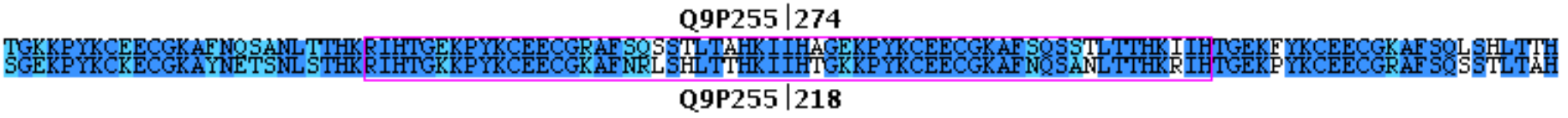


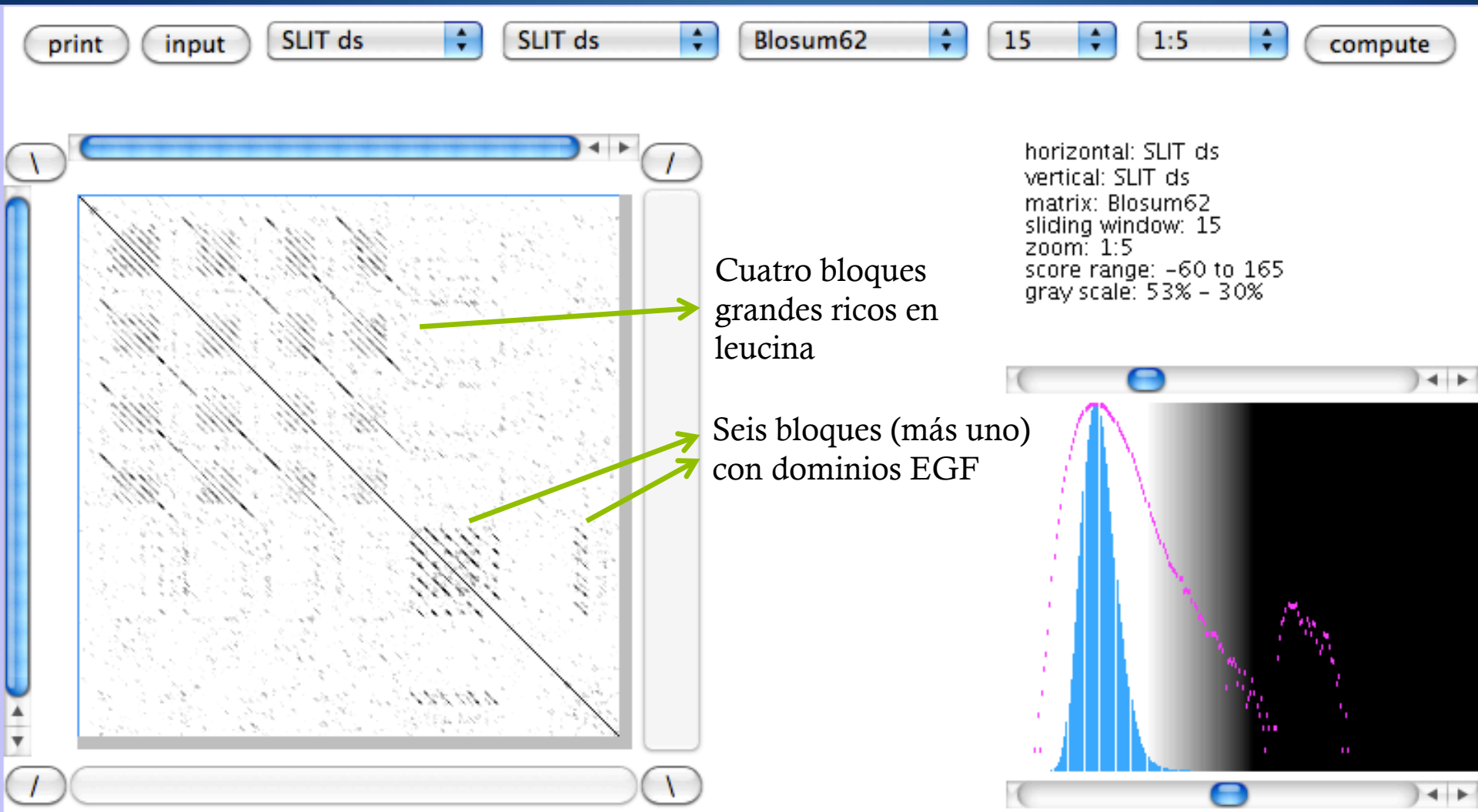
Estructura interna de la proteína: repeticiones de dominios

Hay un dominio que se repite 13 veces

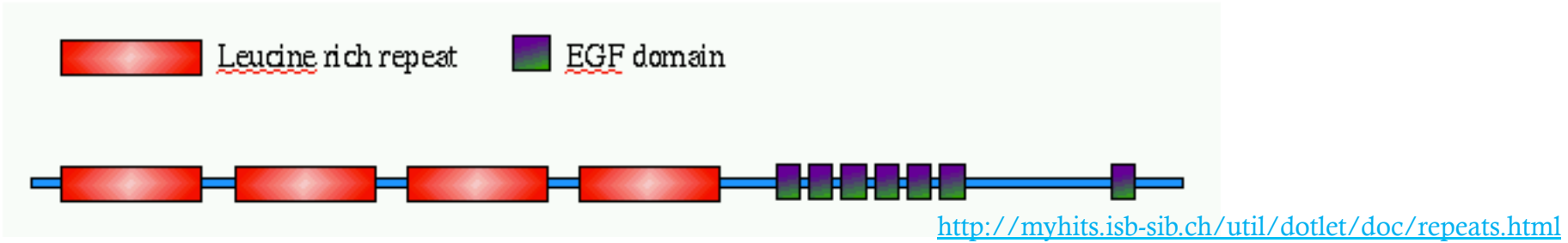


Estructura interna de la proteína: repeticiones de dominios





Repetición de dos dominios en la proteína SLIT de *Drosophila melanogaster* (P24014):



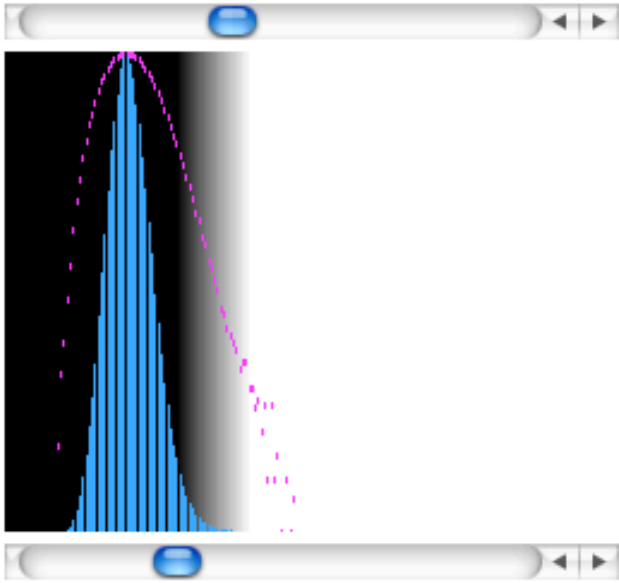
http://myhits.isb-sib.ch/util/dotlet/doc/repeats.html

Dotplot: consideraciones

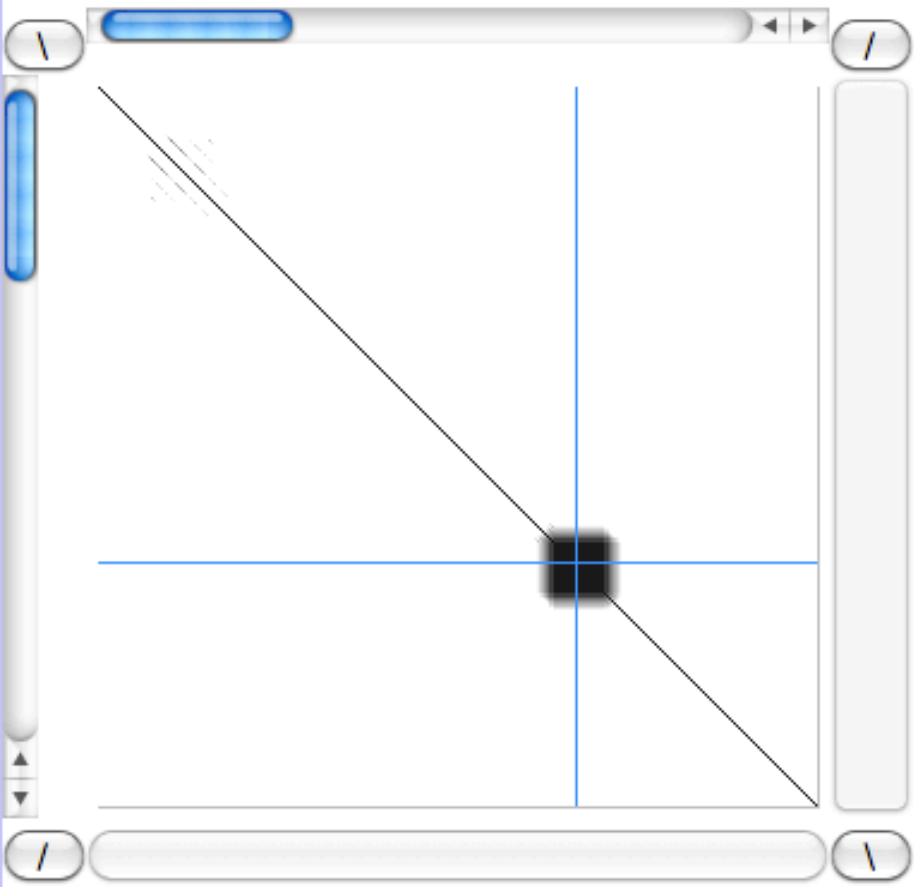
- Comparando una secuencia consigo misma
  - El dot plot es simétrico
  - El número de repeticiones de un patrón viene dado por el número de diagonales por arriba o por abajo
  - La longitud del patrón repetido lo da la longitud de la línea
- Ahora veamos algunos ejemplos más
  - Regiones de baja complejidad
  - Exones e intrones
  - Conservación entre especies

print | input | globin gla | cytog hs | Blosum62 | 13 | 1:7 | compute

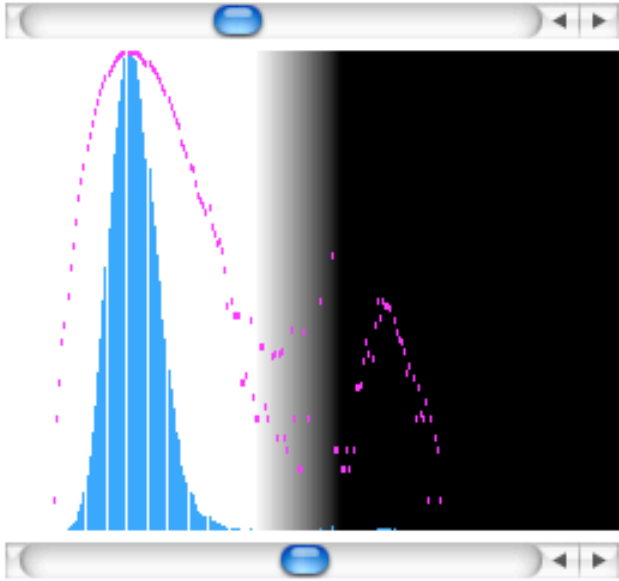
horizontal: globin gla
vertical: cytog hs
matrix: Blosum62
sliding window: 13
zoom: 1:7
score range: -52 to 143
gray scale: 28% - 40%



Lieb et al. *Red blood with blue-blood ancestry: Intriguing structure of a snail hemoglobin*. PNAS, 2006



horizontal: SERA PLAFG
vertical: SERA PLAFG
matrix: Blosum62
sliding window: 15
zoom: 1:1
score range: -60 to 165
gray scale: 55% - 41%



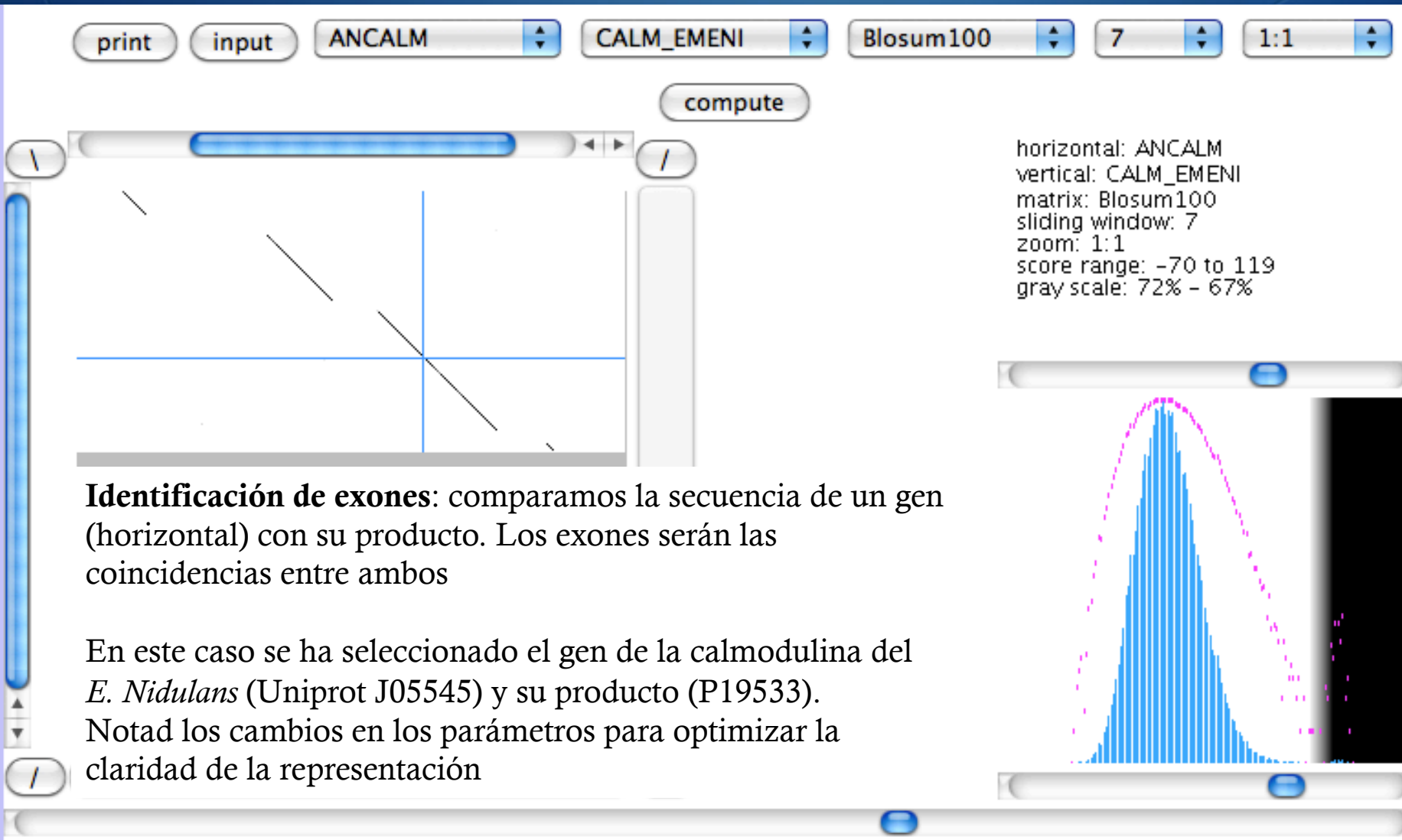
SERA PLAFG|207

NSGSLEKKKYVKLPSNGTTGEQGSSTGTVRGDTEPISDSSSSSSSSSSSSSSSSSSSSSSSSSSSSESLPANGPDSPTVKPPRNLQNICETGKNFKLVVYI

SNSGSLEKKKYVKLPSNGTTGEQGSSTGTVRGDTEPISDSSSSSSSSSSSSSSSSSSSSSSSSSSSSESLPANGPDSPTVKPPRNLQNICETGKNFKLVVY

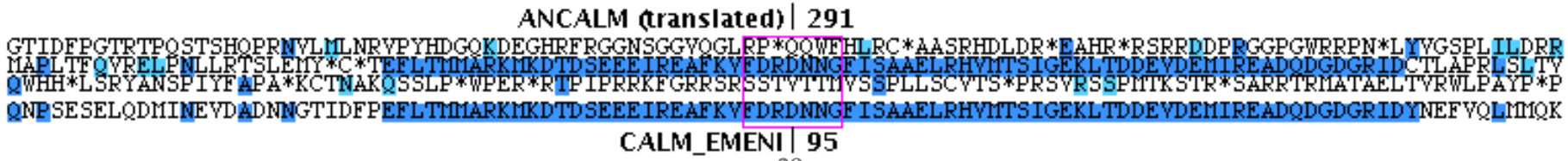
SERA PLAFG|206

Las **regiones de baja complejidad** (periodicidad de uno o pocos aminoácidos/nucleótidos) aparecen como rectángulos, como por ejemplo en esta proteína del P. Falciparum (UniProt P69192)



Identificación de exones: comparamos la secuencia de un gen (horizontal) con su producto. Los exones serán las coincidencias entre ambos

En este caso se ha seleccionado el gen de la calmodulina del *E. Nidulans* (Uniprot J05545) y su producto (P19533). Notad los cambios en los parámetros para optimizar la claridad de la representación



Dotlet

- Permite detectar fácilmente
  - Dominios internos en una secuencia
  - Dominios conservados entre dos secuencias
  - Zonas de baja complejidad
  - Exones e intrones
  - Etc.
- Limitaciones
  - Complejidad computacional → $O(n^2)$
  - ¿Y si hay huecos?

Necesidad de otros algoritmos

- Queremos el mejor alineamiento de **todos los posibles**
  - El que nos da una "puntuación" mejor
- El número de posibles alineamientos permitiendo huecos para dos secuencias de longitud n es $\binom{2n}{n}$
  - Para n=30 → 10^{17}
- **Programación dinámica**: el problema se divide en subproblemas, de manera que la solución a los subproblemas simplifica la solución del problema global

Programación dinámica

- Alineamiento global: Needleman-Wunsch (NW)
  - 1970, PMID: 5420325
- Alineamiento local: Smith-Waterman (SW)
  - 1981, PMID: 7265238

Identification of Common Molecular Subsequences

The identification of maximally homologous subsequences among sets of long sequences is an important problem in molecular sequence analysis. The problem is straightforward only if one restricts consideration to contiguous subsequences (segments) containing no internal deletions or insertions. The more general problem has its solution in an extension of sequence metrics (Sellers 1974; Waterman *et al.*, 1976) developed to measure the minimum number of "events" required to convert one sequence into another.

Puntuaciones

- Estos algoritmos funcionan en base a un sistema de puntuaciones de cuán parecidas son dos secuencias
- Por ejemplo
  - +1 por cada elemento igual (match)
  - -1 por cada elemento desigual (mismatch)
  - -1 por cada hueco introducido (gap)
  - Esta es una puntuación muy simple que se usará sólo para ilustrar el funcionamiento de los algoritmos NW y SW
    - Las puntuaciones se refinarán con PAM y BLOSUM más adelante

Alineamiento global Needleman-Wunsch: inicio

- Creamos una matriz, con una secuencia en horizontal y la otra en vertical.
- La primera fila y columna contienen valores de distancia al origen (gap scores)
  - Asegura el alineamiento hacia atrás y hasta el origen

		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1										
E	↑ -2										
L	↑ -3										
I	↑ -4										
C	↑ -5										
A	↑ -6										
N	↑ -7										

NW: llenado o inducción

- Para cada celda se calculan tres valores, que son la suma de una celda adyacente más el match/ mismatch (MM) de la celda actual
  - MM + celda superior
  - MM + celda izquierda
  - MM + celda superior izquierda
- Para cada celda
  - Se le asigna el máximo de los tres valores
  - Se le asigna la dirección a la celda que propició ese valor
  - En caso de que sean valores iguales, se elige un criterio de desempate

P≠C → MM=-1

0-1=-1
-1-1=-2
-1-1=-2

		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6	↖ -7	↖ -8	↖ -9	↖ -10
E	↑ -2	↖ -2	↖ -2	↖ -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8
L	↑ -3	↖ -3	↖ -3	↑ -2	↖ 0	← -1	← -2	← -3	← -4	← -5	← -6
I	↑ -4	↖ -4	↖ -4	↑ -3	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6
C	↑ -5	↖ -3	← -4	↑ -4	↑ -2	↖ -2	↖ 0	← -1	← -2	← -3	← -4
A	↑ -6	↑ -4	↖ -4	↖ -5	↑ -3	↖ -1	↑ -1	↖ 1	← 0	← -1	← -2
N	↑ -7	↑ -5	↖ -5	↖ -5	↑ -4	↑ -2	↖ -2	↑ 0	↖ 2	← 1	← 0

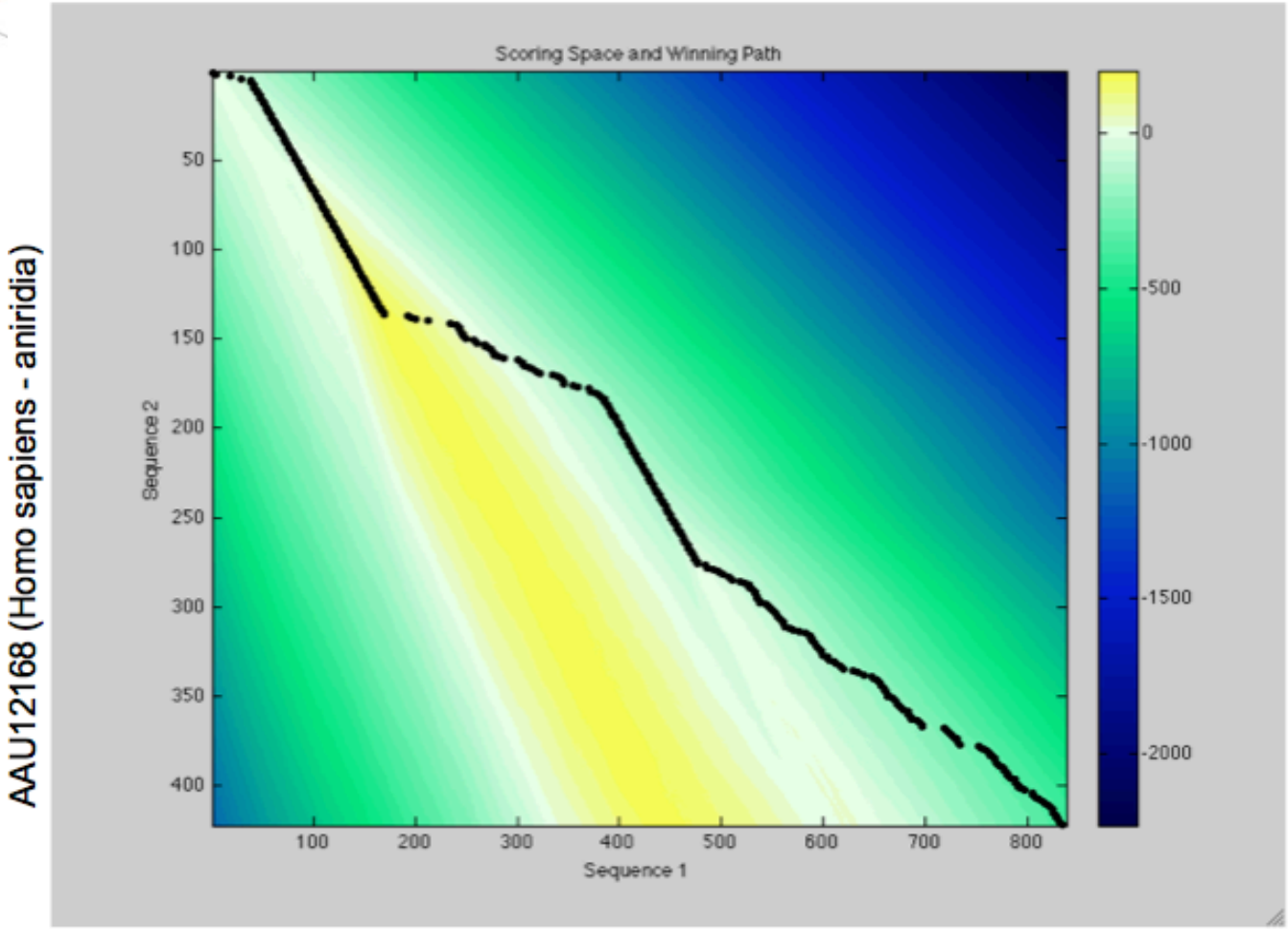
NW: "trace-back"

- Seguimos la ruta con mejor score
  - Comenzando en la esquina inferior derecha
  - Siguiendo las flechas
    - ← desplazamiento de la cadena vertical respecto a la horizontal
    - ↑ desplazamiento de la horizontal respecto a la vertical
    - ↖ no hay desplazamiento

		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6	↖ -7	↖ -8	↖ -9	↖ -10
E	↑ -2	↖ -2	↖ -2	↖ -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8
L	↑ -3	↖ -3	↖ -3	↑ -2	↖ 0	← -1	← -2	← -3	← -4	← -5	← -6
I	↑ -4	↖ -4	↖ -4	↑ -3	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6
C	↑ -5	↖ -3	← -4	↑ -4	↑ -2	↖ -2	↖ 0	← -1	← -2	← -3	← -4
A	↑ -6	↑ -4	↖ -4	↖ -5	↑ -3	↖ -1	↑ -1	↖ 1	← 0	← -1	← -2
N	↑ -7	↑ -5	↖ -5	↖ -5	↑ -4	↑ -2	↖ -2	↑ 0	↖ 2	← 1	← 0

COELACANTH
-PELICAN--

Needleman-Wunsh: ejemplo real



Alineamiento local y global

Global

- Útiles cuando las secuencias tienen un tamaño muy parecido
- Un alineamiento local hace múltiples alineamientos globales sobre subsecuencias

Local

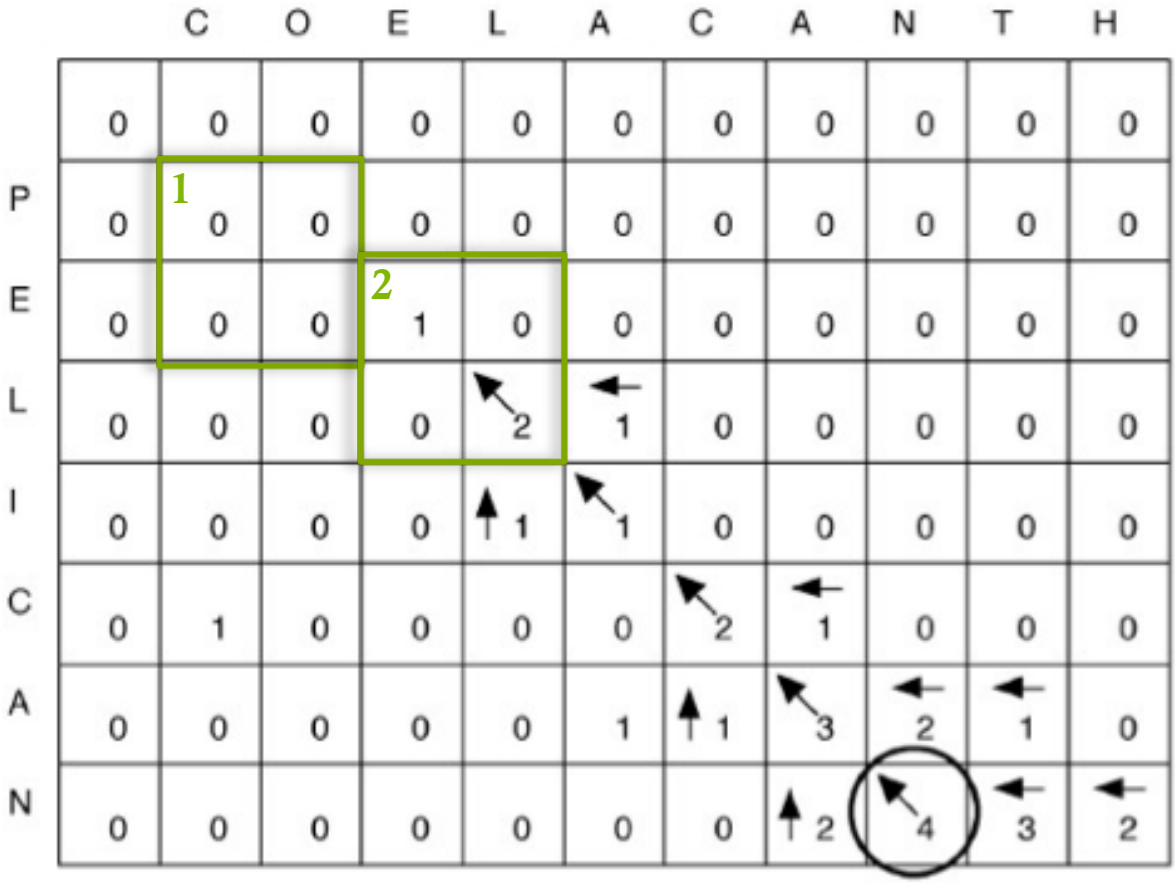
- Útil con secuencias de distinto tamaño pero que se espera que tengan regiones comunes
- Los algoritmos tienen un coste computacional más alto

```
Global      FTFTALILLAVAV
            F--TAL-LLA-AV
Local       FTFTALILL-AVAV
            --FTAL-LLAAV--
```

Alineamiento local Smith-Waterman

- Modificación del algoritmo NW
- La matriz se construye igual excepto que:
  - La primera fila y columna son 0s, en vez de -1,-2,-3,-4…
  - Mismatch es -1 y match 1
    - Si algún valor de celda queda negativo, lo ponemos a 0
  - El trace-back comienza en la puntuación más alta y termina cuando llegue a un 0

Smith-Waterman

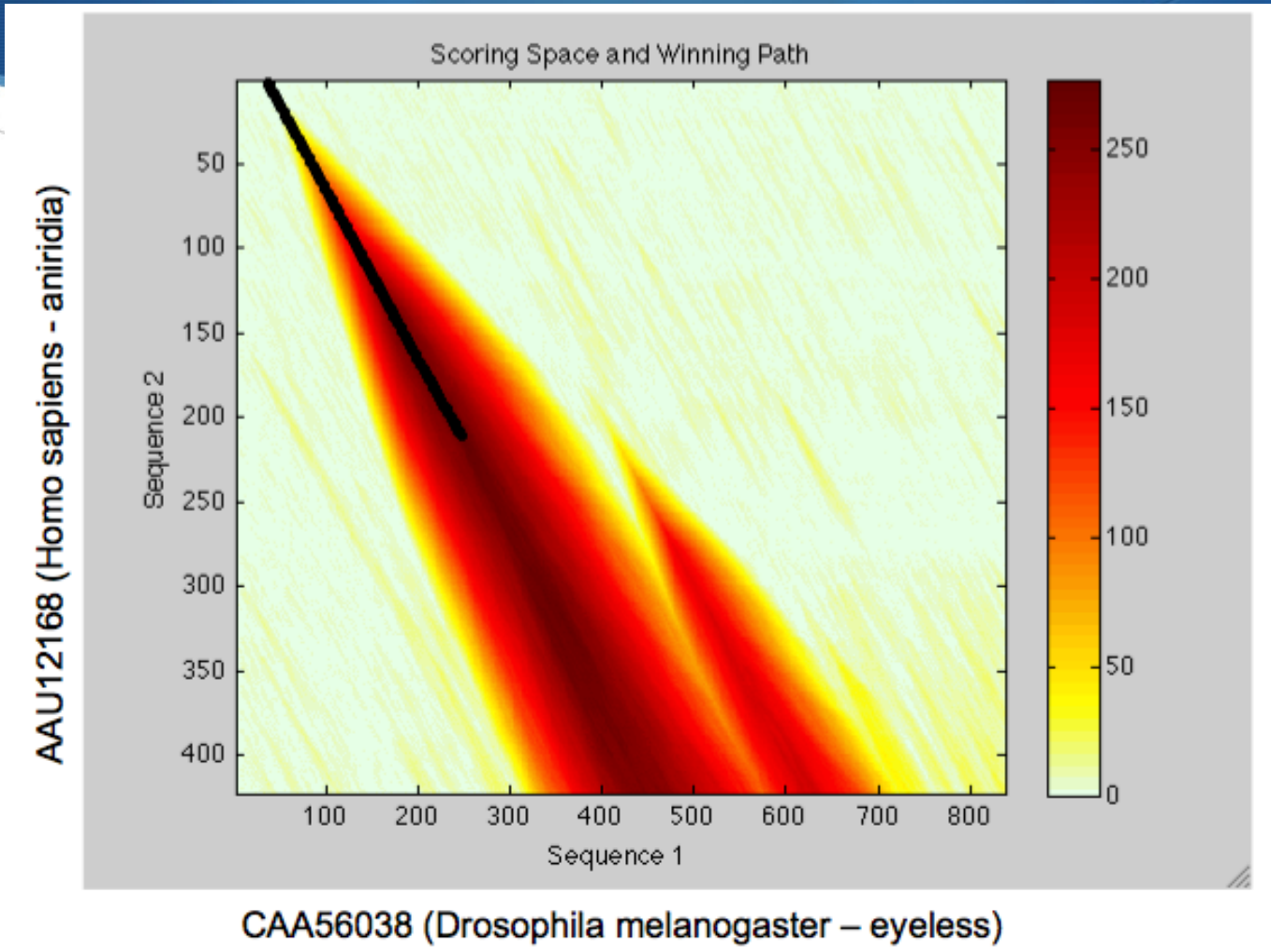


1 O vs E → -1 + max(0,0,0) → -1 → 0

2 L vs L → 1 + max(0,1,0) → 2

COELACANTH
-PELICAN--

Smith-Waterman: ejemplo real



Alineamiento de pares de secuencias

Introducción

Algoritmos

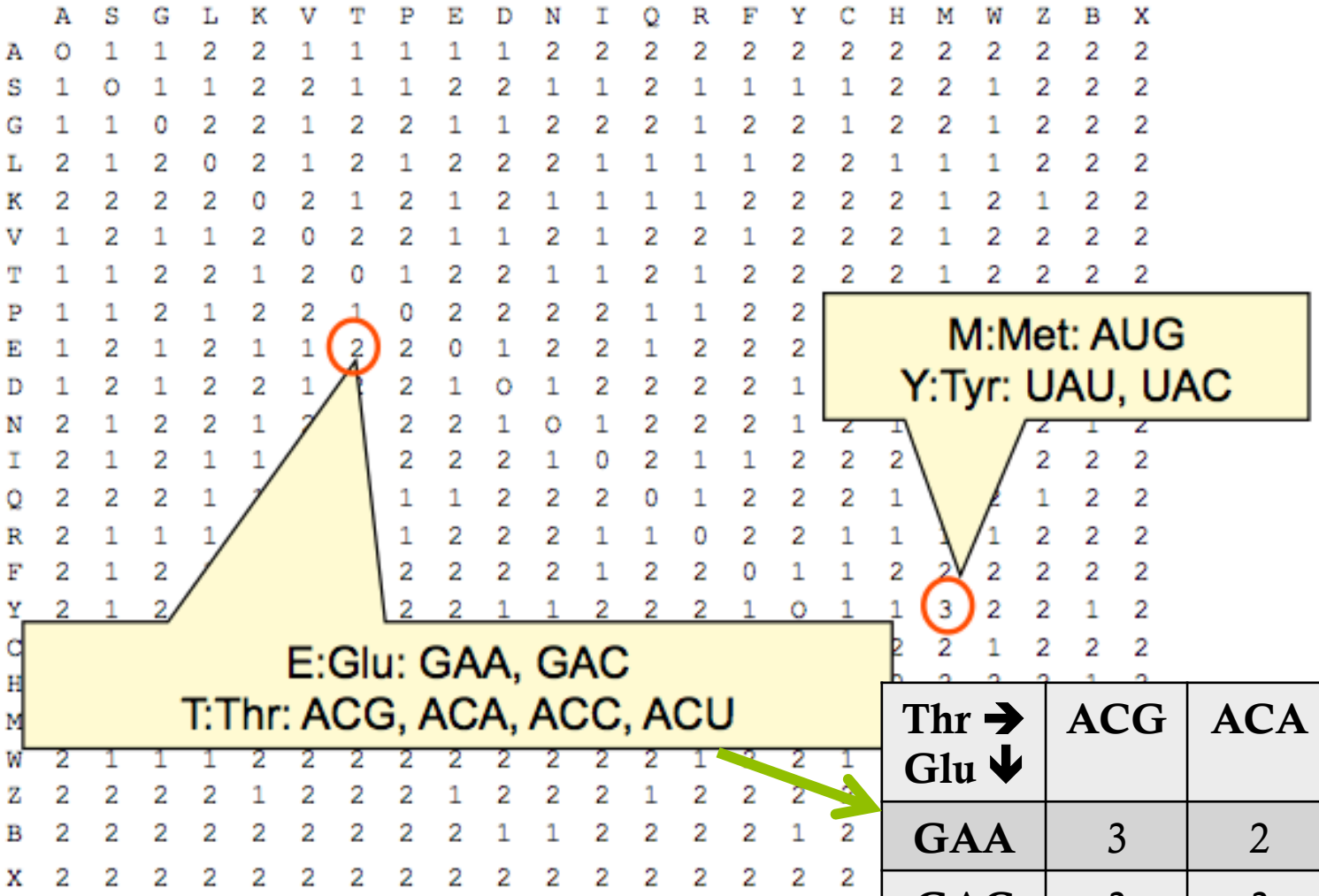
Matrices de puntuación

PAM

BLOSUM

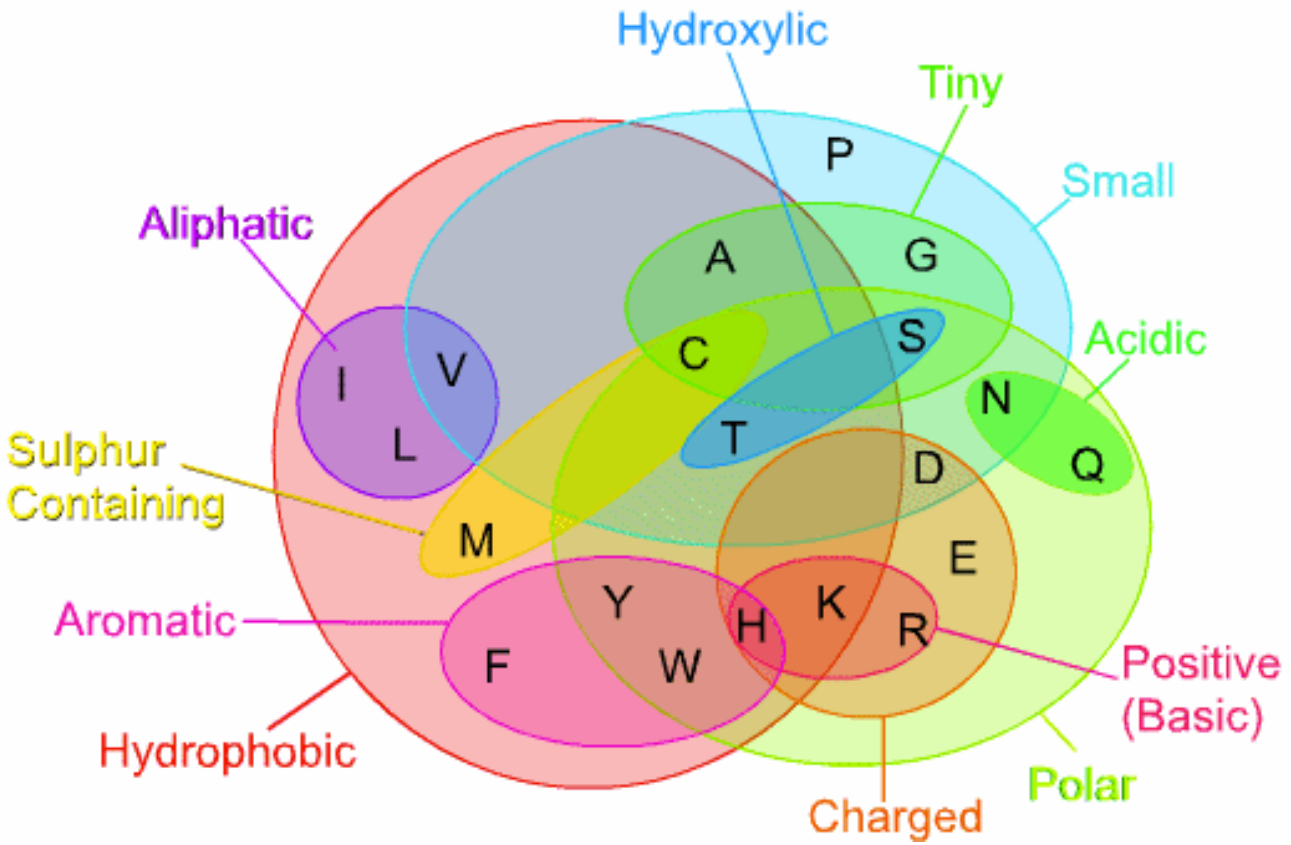
Matrices de puntuación

- Necesitamos métodos de puntuación más sofisticados, con sentido biológico:
  - Genético: coste de mutación de un aminoácido en otro
  - Químico: similitud en las características de los aminoácidos
  - Evolutivo: frecuencia evolutiva de cambio en aminoácidos
    - Lod scores
    - PAM
    - BLOSUM



Thr → Glu ↓	ACG	ACA	ACC	ACU
GAA	3	2	3	3
GAC	3	3	2	3

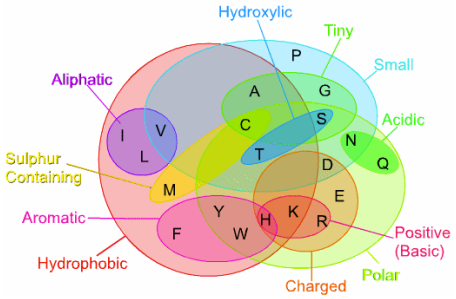
- **Matriz de coste de mutación**
  - Número mínimo de cambios de base requeridos para convertir un aminoácido en otro distinto (entre 1 y 3)



- **Matriz de similitud de características químicas**
  - El coste de mutación tiene en cuenta el código genético, pero no tiene en cuenta la presión selectiva en los cambios de un aminoácido a otro

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

Construcción de las matrices a partir de resultados experimentales

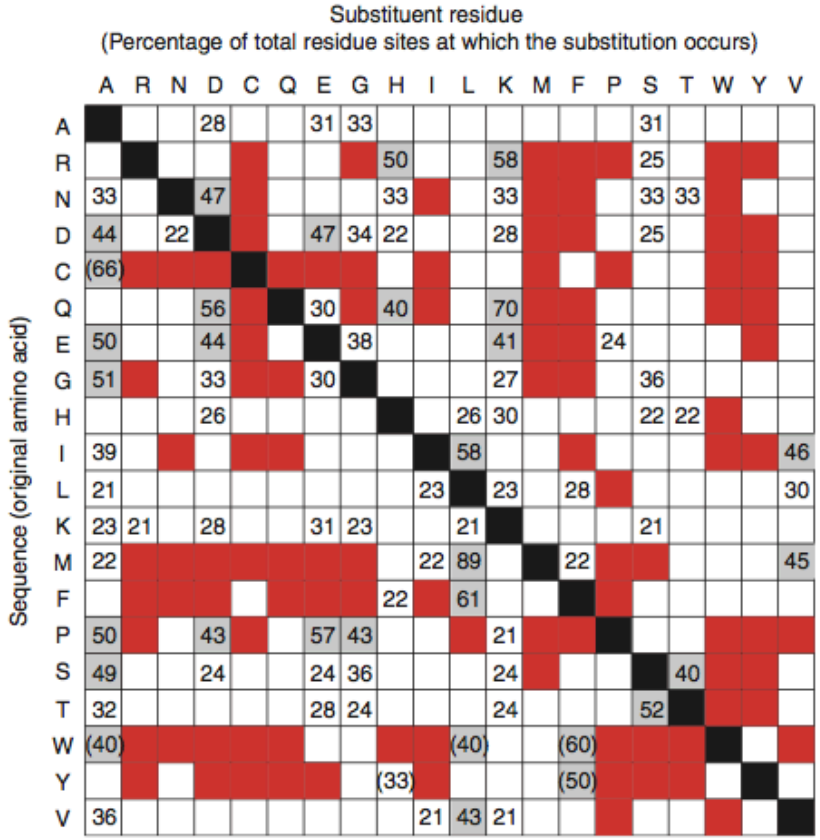


Matrices de puntuación

- Matriz con puntuaciones (lod scores) para todas las sustituciones de aminoácidos posibles
  - Las puntuaciones representan estimaciones de la probabilidad de sustitución respecto a una sustitución aleatoria
  - Se basa en las evidencias conocidas de sustituciones evolutivas
- Los lod scores son números reales, pero
  - Usualmente se representan mediante números enteros (*raw scores*) multiplicándolos antes por un factor de escala

Matrices de puntuación: historia

- 1965: Emile Zuckerkandl y **Linus Pauling** diseñan la primera matriz de puntuaciones para distintas secuencias de globina
  - Rojo: sustituciones que nunca ocurren
  - Blanco: sustituciones que ocurren en con una frecuencia menor al 20%
    - Con número si entre 21 y 39%
  - Gris: sustituciones que ocurren con una frecuencia del X% (>=40%)
    - Con paréntesis si se tienen pocas evidencias para esa sustitución



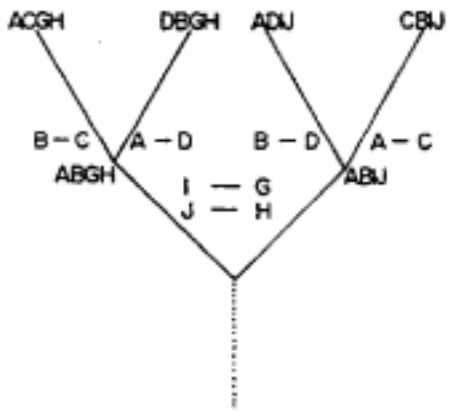
Matrices de puntuación: historia

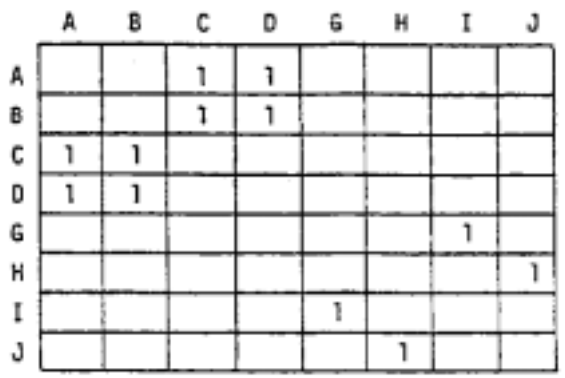
- 1978: **Dayhoff** et al. estudiaron 1572 sustituciones en 71 grupos de proteínas muy parecidas entre sí
  - Accepted Point Mutation (PAM): sustitución de un aminoácido por otro, que ha sido aceptada por la selección natural para una determinada proteína
- 1992: Henikof y Henikof mejoran la matriz de Dayhoff con una base de datos de 500 alineamientos (BLOCKS) entre proteínas poco parecidas entre sí

Margaret Dayhoff

Matriz de Dayhoff

- Se reconstruye un árbol filogenético para cada uno de los 71 grupos de proteínas (parecidas entre sí, >=85% de similitud)
  - Y se determinan cuántos reemplazos (mutaciones aceptadas) hay respecto al ancestro común





	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

A_{LEU}

A_{ij}

1572 mutaciones aceptadas (PAMs) estudiadas entre los distintos aminoácidos, multiplicadas por 10
Por ejemplo, 20.7 de las 1572 PAMs ocurren entre Leu y Met
Leu tiene un total de 142.8 mutaciones relacionadas

Matriz de Dayhoff

Para Ala se toma un valor arbitrario de 100

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

m_j

Mutabilidad relativa según Dayhoff et al. 1978

$$M_{ij} = \frac{\lambda m_j A_{ij}}{A_j}$$

- M_{ij} – probabilidad de que el aminoácido j cambie al aminoácido i en un intervalo evolutivo dado
- λ - constante de proporcionalidad (para Dayhoff, ~0.013)
- $M_{MET\text{-}LEU} = \lambda \cdot m_{LEU} \cdot A_{MET\text{-}LEU} / A_{LEU} = 0.013 \cdot 40 \cdot 207 / 1428 = 0.08\%$

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A	98.67	0.02	0.09	0.10	0.03	0.08	0.17	0.21	0.02	0.06	0.04	0.02	0.06	0.02	0.22	0.35	0.32	0.00	0.02	0.18
R	0.01	99.13	0.01	0.00	0.01	0.10	0.00	0.00	0.10	0.03	0.01	0.19	0.04	0.01	0.04	0.06	0.01	0.08	0.00	0.01
N	0.04	0.01	98.22	0.36	0.00	0.04	0.06	0.06	0.21	0.03	0.01	0.13	0.00	0.01	0.02	0.20	0.09	0.01	0.04	0.01
D	0.06	0.00	0.42	98.59	0.00	0.06	0.53	0.06	0.04	0.01	0.00	0.03	0.00	0.00	0.01	0.05	0.03	0.00	0.00	0.01
C	0.01	0.01	0.00	0.00	99.73	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.01	0.00	0.03	0.02
Q	0.03	0.09	0.04	0.05	0.00	98.76	0.27	0.01	0.23	0.01	0.03	0.06	0.04	0.00	0.06	0.02	0.02	0.00	0.00	0.01
E	0.10	0.00	0.07	0.56	0.00	0.35	98.65	0.04	0.02	0.03	0.01	0.04	0.01	0.00	0.03	0.04	0.02	0.00	0.01	0.02
G	0.21	0.01	0.12	0.11	0.01	0.03	0.07	99.35	0.01	0.00	0.01	0.02	0.01	0.01	0.03	0.21	0.03	0.00	0.00	0.05
H	0.01	0.08	0.18	0.03	0.01	0.20	0.01	0.00	99.12	0.00	0.01	0.01	0.00	0.02	0.03	0.01	0.01	0.01	0.04	0.01
I	0.02	0.02	0.03	0.01	0.02	0.01	0.02	0.00	0.00	98.72	0.09	0.02	0.21	0.07	0.00	0.01	0.07	0.00	0.01	0.33
L	0.03	0.01	0.03	0.00	0.00	0.06	0.01	0.01	0.04	0.22	99.47	0.02	0.45	0.13	0.03	0.01	0.03	0.04	0.02	0.15
K	0.02	0.37	0.25	0.06	0.00	0.12	0.07	0.02	0.02	0.04	0.01	99.26	0.20	0.00	0.03	0.08	0.11	0.00	0.01	0.01
M	0.01	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.05	0.08	0.04	98.74	0.01	0.00	0.01	0.02	0.00	0.00	0.04
F	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.08	0.06	0.00	0.04	99.46	0.00	0.02	0.01	0.03	0.28	0.00
P	0.13	0.05	0.02	0.01	0.01	0.08	0.03	0.02	[illegible]	[illegible]	0.02	0.02	0.01	0.01	99.26	0.12	0.04	0.00	0.00	0.02
S	0.28	0.11	0.34	0.07	0.11	0.04	0.06	0.16	[illegible]	[illegible]	0.01	0.07	0.04	0.03	0.17	98.40	0.38	0.05	0.02	0.02
T	0.22	0.02	0.13	0.04	0.01	0.03	0.02	0.02	0.01	0.11	0.02	0.08	0.06	0.01	0.05	0.32	98.71	0.00	0.02	0.09
W	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	99.76	0.01	0.00
Y	0.01	0.00	0.03	0.00	0.03	0.00	0.01	0.00	0.04	0.01	0.01	0.00	0.00	0.21	0.00	0.01	0.01	0.02	99.45	0.01
V	0.13	0.02	0.01	0.01	0.03	0.02	0.02	0.03	0.03	0.57	0.11	0.01	0.17	0.01	0.03	0.02	0.10	0.00	0.02	99.01

[Sin título]

Matriz con probabilidades de mutación PAM1

1 PAM se define como la unidad de divergencia evolutiva en la que han ocurrido un 1% de mutaciones entre dos secuencias de proteínas

PAMn

- Podemos multiplicar la matriz PAM1 por sí misma n veces para obtener las probabilidades de sustitución si hay un n% de probabilidades de que cada aminoácido de la cadena haya mutado.
  - Útil para estudiar sustituciones entre aminoácidos en proteínas muy distintas, con probabilidad de sustitución muy baja
  - PAM60, PAM80, PAM100, PAM250
- Multiplicación de matrices

$$M_1 = \begin{bmatrix} 3 & 4 \\ 0 & 2 \end{bmatrix} \quad M_2 = \begin{bmatrix} 5 & -2 \\ 2 & 1 \end{bmatrix}$$

$$M_{12} = \begin{bmatrix} (3)(5)+(4)(2) & (3)(-2)+(4)(1) \\ (0)(5)+(2)(2) & (0)(-2)+(2)(1) \end{bmatrix} = \begin{bmatrix} 23 & -2 \\ 4 & 2 \end{bmatrix}$$

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

q_{ij}

Matriz con probabilidades de mutación PAM250

Para conseguir esta matriz, se multiplica la matriz PAM1 por sí misma 250 veces

Matriz de puntuación: lod scores

- Las matrices PAM calculadas no se utilizan directamente en los algoritmos de alineamiento, requieren de una normalización previa para facilitar los cálculos
- **Lod (log odd) score**: puntuación del logaritmo de la frecuencia de sustitución de un aminoácido por otro
  - Las propiedades del logaritmo permiten sumar los lod scores, en vez de multiplicarlos, durante el alineamiento, lo cual es computacionalmente menos costoso
    - $log(m \cdot n)=log(m)+log(n)$
  - Los lod scores son simétricos, simplificando las matrices.
  - Los lod scores suelen simplificarse a enteros (raw scores), simplificando los cálculos

lod scores en Dayhoff

$$S_{ij} = 10 \times \log_{10} \frac{q_{ij}}{p_i}$$

- S_{ij} – lod score: logaritmo de la frecuencia con la que el aminoácido *i* se convierte en el aminoácido *j*
- p_i – frecuencia normalizada de aparición del aminoácido *i*
  - Es una normalización de la mutabilidad relativa
- q_{ij} – probabilidad de sustitución de *i* por *j* en PAMn

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

Lod scores: ejemplo

- El lod score de la sustitución de Alanina (A) en Arginina (R) en PAM250 es:
  - $S_{A,R}=10x\log_{10}(q_{AR}/p_A)=10x\log_{10}(0.06/0.087)=-1.61 \sim -2$
- Significado: la posibilidad de que haya una sustitución de A por R en dos cadenas (en un alineamiento regido por el estudio evolutivo de Dayhoff en PAM250) es $10^{-0.161} \sim$ un 70% de la posibilidad de que se dé esa correspondencia aleatoriamente
  - ¿Y en $S_{A,A}$, que es igual a 2? → $2=10x\log_{10}(x)$ → $x=10^{0.2}=1.58$

		Ala	Arg	Asn	Asp	Cys	Gln
		A	R	N	D	C	Q
Ala	A	13	6	9	9	5	8
Arg	R	3	17	4	3	2	5
Asn	N	4	4	6	7	2	5

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

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

lod score PAM250

- Es una matriz simétrica
  - Debido a la normalización y redondeo al entero más cercano
  - Simplifica los cálculos de los algoritmos de alineamiento
- Se dan lod scores simplificados al entero más cercano (raw scores)
  - Simplifican los cálculos de los algoritmos de alineamiento
  - Puede introducir pequeños errores (asumibles)

¿Qué PAM uso?

- El uso de una matriz PAM u otra depende de la similitud entre las secuencias
  - Para secuencias muy parecidas PAM10 puede dar buenos alineamientos
  - Para secuencias muy distintas PAM250 puede ser mejor
- A priori no siempre sabemos la similitud de las secuencias
  - Puede ser necesario repetir los alineamientos con varias matrices y quedarse con la que dé el mejor alineamiento

BLOSUM

- BLOcks amino acid SUbstitution Matrices
  - Se construyen a partir de la base de datos BLOCKS, que contiene "segmentos alineados sin huecos, correspondientes a las regiones de proteínas más conservadas"
    - BLOCKS contenía inicialmente 500 patrones, ahora ~2000
- Cada matriz BLOSUM se refiere a un % de similitud en las secuencias de partida
  - BLOSUM45, BLOSUM62, BLOSUM80

Bloque

- Ejemplo de bloque:

```
Block IPB012530

ID   BAGE; BLOCK
AC   IPB012530; distance from previous block=(0,8)
DE   B melanoma antigen
BL   AAE;   width=43; seqs=8; 99.5%=1725; strength=1441
BAGE1_HUMAN|Q13072   (   1) MAARAVFLALSAQLLQARLMKEESPVVSWRLEPEDGTALCFIF  29
BAGE2_HUMAN|Q86Y30   (   1) MAAGVVFLALSAQLLQARLMKEESPVVSWRLEPEDGTALDVHF  26
BAGE3_HUMAN|Q86Y29   (   1) MAAGVVFLALSAQLLQARLMKEESPVVSWRLEPEDGTALDVHF  26
BAGE4_HUMAN|Q86Y28   (   1) MAAGAVFLALSAQLLQARLMKEESPVVSWWLEPEDGTALXXXX  26
BAGE5_HUMAN|Q86Y27   (   1) MAAGAVFLALSAQLLQARLMKEESPVVSWRLEPEDGTALCFIF  27

Q08ER0|Q08ER0_HUMAN  (   1) MAAGVVFLALSAQLLQARLMKEESPVVSWRLEPEDGTALDVHF  26
Q29RY1|Q29RY1_HUMAN  (   1) MAAGVVFLALSAQLLQARLMKEESPVVSWRLEPEDGTALDGVS  30
Q1V5E6|Q1V5E6_VIBAL  (   9) LIAASLWLAASAQALEAKLHKDDLPVLSPEVQHETASKRVTSR 100
//
```

Cálculo BLOSUM

...A...
...A...
...A...
...A...
...A...
...A...
...A...
...K...

- En BLOSUM no conocemos el árbol filogenético: cualquier secuencia se considera un ancestro posible de las demás

Pares posibles

$$f_{AA} = 6+5+4+3+2+1 = 21 \qquad f_{AK} = 7$$

Probabilidades de los pares

$$q_{AA} = \frac{f_{AA}}{f_{AA}+f_{AK}} = \frac{21}{21+7} = 0.75 \qquad q_{AK} = \frac{f_{AK}}{f_{AA}+f_{AK}} = \frac{7}{21+7} = 0.25$$

Probabilidad de que A o K estén en un par

$$p_A = q_{AA} + \frac{q_{AS}}{2} = 0.75 + \frac{0.25}{2} = 0.875 \qquad p_K = \frac{q_K}{2} = 0.125$$

Probabilidad "esperada" de que ocurra AA o AK

$$e_{AA} = p_A \times p_A = 0.76525 \qquad e_{AK} = 2 \times p_A \times p_K = 0.21875$$

lod score

$$s_{ij} = 2 \times \log_2 \frac{q_{ij}}{e_{ij}}$$

$$s_{AA} = 2 \times \log_2 \frac{q_{AA}}{e_{AA}} \cong -0.06 \qquad s_{AK} = 2 \times \log_2 \frac{q_{AK}}{e_{AK}} \cong 0.39$$

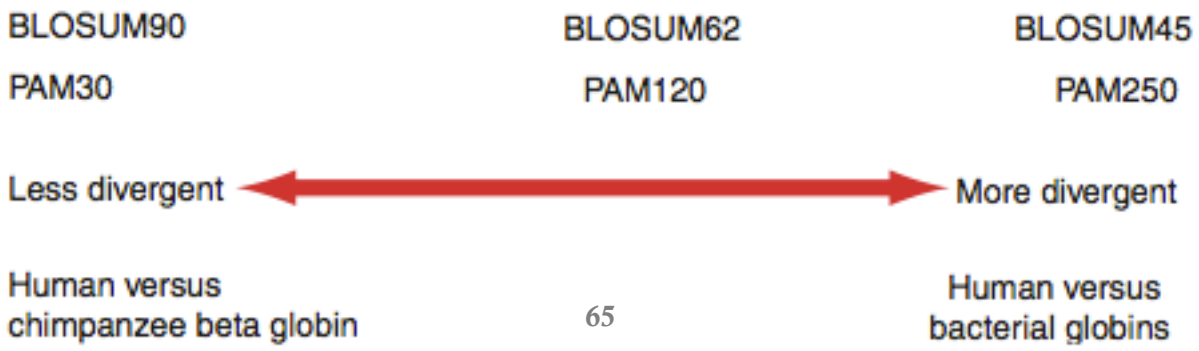
BLOSUM62

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	−1	1	1	−2	−1	−3	−2	5							
M	−1	−2	−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

Según Henikoff & Henikoff (1992) BLOSUM62 es mejor alternativa que PAM y otros BLOSUM para un alineamiento genérico.
Es utilizada como matriz por defecto por la mayoría de los programas de alineamiento y de búsquedas en bases de datos

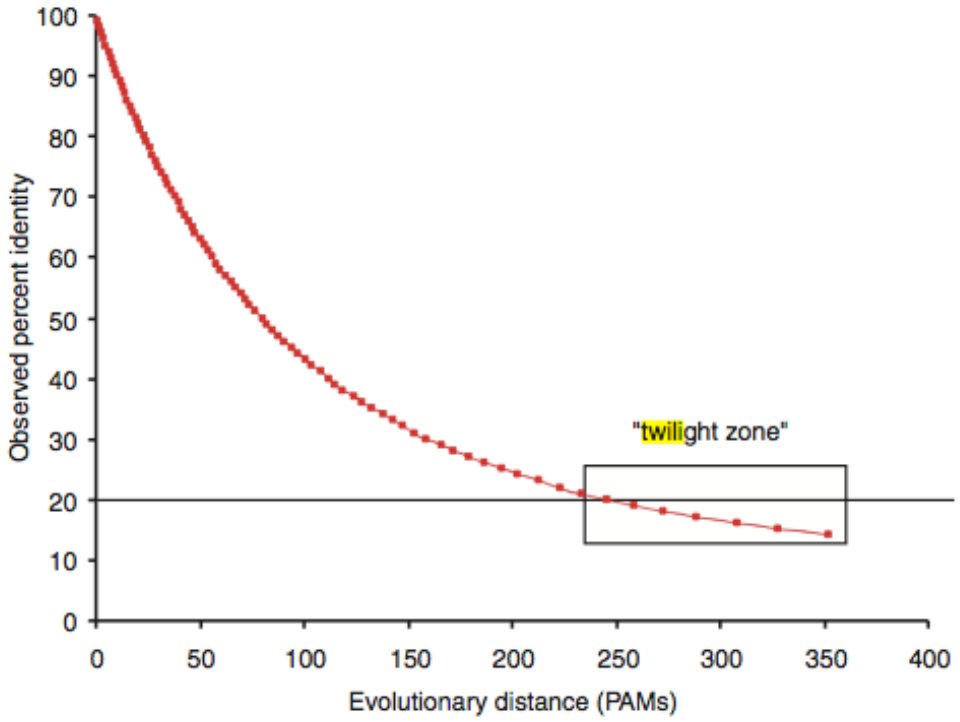
PAM vs BLOSUM

- PAM: basada en familias de proteínas muy relacionadas
  - Útil para alinear secuencias parecidas
- BLOSUM: basada en observaciones de alineamientos entre secuencias muy distintas
  - Útil para alinear secuencias distintas



La dimensión desconocida

- ¿Hasta qué punto dos proteínas pueden divergir de forma detectable?
  - PAM250 considera que hay un 250% de probabilidades de que cada aminoácido haya mutado → equivale a una similitud del ~20%
  - El área de similitud entorno al 20% se llama la "twilight zone" o dimensión desconocida
    - Las secuencias pueden estar relacionadas, pero no podemos detectarlo mediante el alineamiento



Alineamiento de proteínas vs alineamiento de ADN

- ¿Por qué no hablamos de matrices de puntuación para ADN?
  - Normalmente el alineamiento de proteínas nos da más información:
    - Muchos cambios de un nucleótido de un codón no varían el aminoácido resultante → más estable
    - Muchos aminoácidos comparten propiedades biofísicas
    - Muchas proteínas comparten estructura o regiones estructurales
    - El ADN sufre distintas modificaciones pos-translacionales que pueden influir en la proteína que codifica

Matrices de puntuación para nucleótidos

- No obstante, existen matrices para nucleótidos

	A	T	C	G
A	4	-5	-5	-5
T	-5	4	-5	-5
C	-5	-5	4	-5
G	-5	-5	-5	4

Matriz utilizada por BLAST

	A	T	C	G
A	2			
T	-7	2		
C	-7	-5	2	
G	-5	-7	-7	2

Matriz PAM1 para nucleótidos
La transición (mutación entre purinas A-G, o entre pirimidinas C-T) es más común que la transversión (de purina a pirimidina)

Resumen

- La bioinformática se basa en la comparación de datos. Para maximizar la calidad de la comparación de dos secuencias, necesitamos alinearlas. La comparación de secuencias nos ayuda a inferir funciones, relaciones, dominios, etc.
- Para realizar el alineamiento, debemos decidir cómo evaluar las diferencias entre las secuencias (matriz de puntuación) y qué estrategia usar para maximizar la similitud (algoritmo)
- El algoritmo puede ser global (NW) o local (SW). NW alinea la secuencia completa a la vez y es útil para secuencias similares en longitud. SW es una evolución de NW y es útil para secuencias más variables, encontrando en general mejores soluciones, a costa de una mayor complejidad computacional.
- Las matrices de puntuación más usadas son PAM y BLOSUM, ambas basadas en evidencias evolutivas. Las matrices tienen distintas versiones, que corresponden a distintas suposiciones sobre la "distancia evolutiva" entre las dos secuencias a alinear

Preguntas para debate

- Al calcular los lod scores ¿crees que los errores que introduce este cálculo adicional son asumibles? Prueba a calcular $S_{R,A}$ como se calculó $S_{A,R}$ en la diapositiva 58
- Si quieres comparar dos proteínas, hay una matriz "ideal" que utilizar? ¿Hay algún modo de conocer cuál es la mejor matriz de puntuación a utilizar?
- Muchas proteínas tienen varios dominios. Imagina una que tiene un dominio que evoluciona lentamente y otro que lo hace rápidamente. Para compararla con otra proteína, ¿usarías dos alineamientos distintos (por ej. con matrices PAM40 y PAM250) o uno solo, con una matriz intermedia?

Lecturas adicionales

- Pevsner, 2009: Ch 3 *Pairwise Sequence Alignment*
- Yi-Kuo Yu and Stephen F. Altschul, *The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions*, Bioinformatics. 2005 Apr 1;21(7):902-11,
  - PMID: 15509610
- Eddy SR., *Where did the BLOSUM62 alignment score matrix come from?* Nat Biotechnol. 2004 Aug;22(8):1035-6.
  - PMID: 15286655

Genome Valence es un proyecto de Ben Fry para visualizar el proceso de alineamiento de pares con BLAST
Se representan las dos secuencias a alinear, que se van rompiendo en pedazos y uniéndose si se alinean

http://benfry.com/genomevalence/