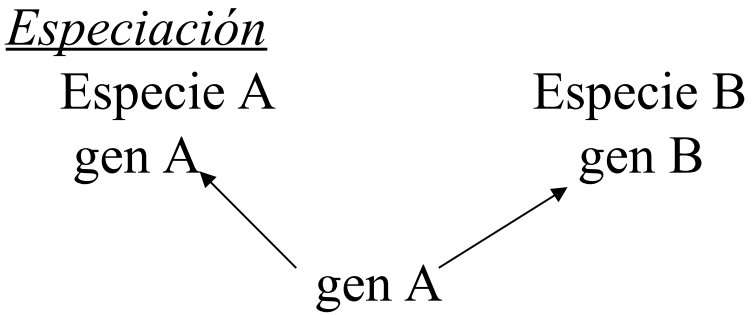
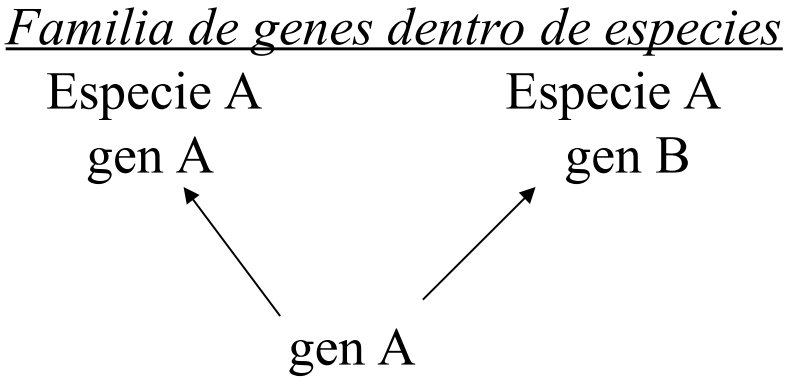
Genes y secuencias Homólogas



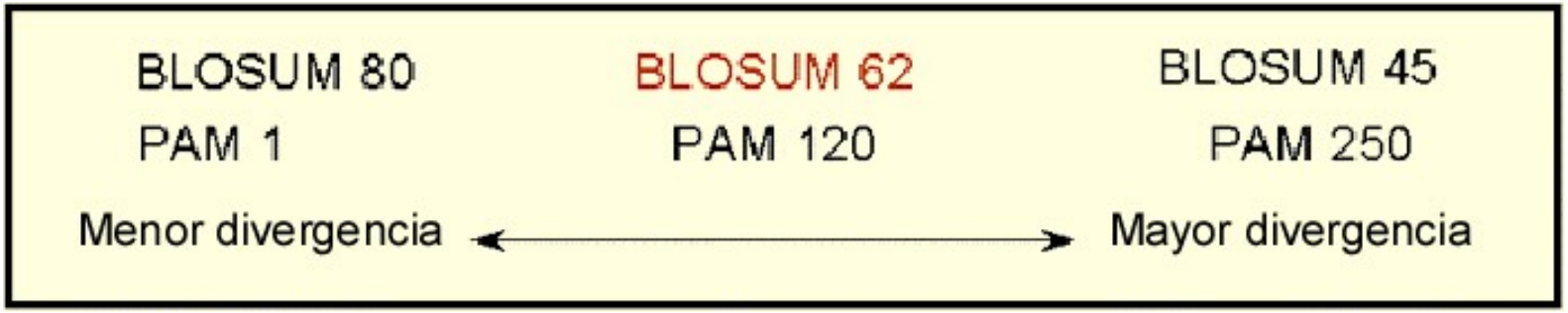
ortologos



paralogos

La homología de secuencias se refiere a la situación en la que las secuencias de dos o más proteínas o ácidos nucleicos son similares entre sí debido a que presentan un mismo origen evolutivo.

Que Matriz usar para alineamiento de proteinas?



✓ Las matrices PAM están diseñadas fundamentalmente para estudiar homología a nivel global entre secuencias. Tienen muy en cuenta el factor de distancia evolutiva.

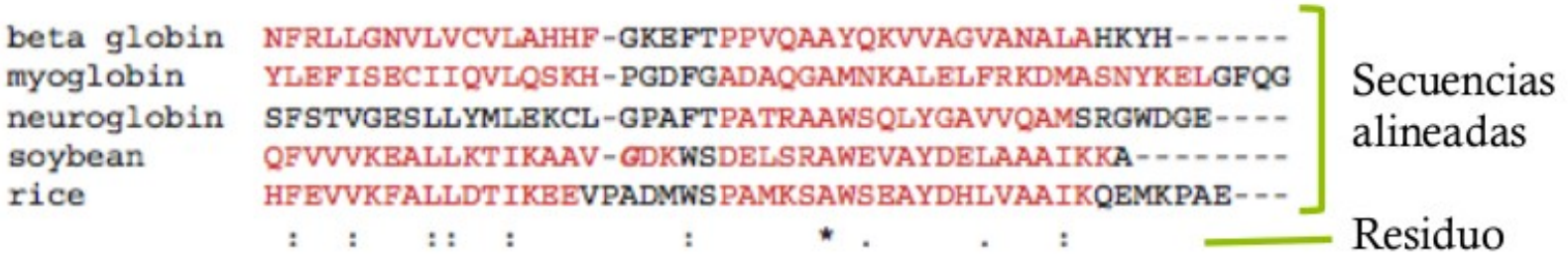
✓ Las matrices BLOSUM, en cambio, son preferibles cuando se estudian secuencias conservadas (sec. consenso). El factor del tiempo evolutivo es menos importante.

Alineamiento multiple de secuencias

- Para hacer un alineamiento, generalmente necesitamos seleccionar:
  1. Las secuencias homólogas a alinear
  2. El software que utilice una función de puntuación óptima
  3. Los parámetros adecuados (fundamentalmente huecos)
- No hay un alineamiento perfecto
  - Las secuencias evolucionan más rápido que las estructuras o funcionalidades (la secuencia puede variar y la estructura o función seguir invariante)

Alineamiento multiple de secuencias

- Un alineamiento múltiple es una colección de tres o más secuencias de aminoácidos o nucleótidos parcial o completamente alineados
  - Residuo: secciones homólogas de las secuencias, en un sentido
    - Evolutivo: presumiblemente provenientes de un ancestro común
    - Estructural: suelen ocupar lugares relevantes en la estructura 3D



[NOTA: A veces se llama residuo a cada columna del alineamiento]

Alineamiento multiple de secuencias para que?

- Dar información acerca de la función, estructura y evolución de una secuencia
  - Al conocer cómo se alinea respecto a un grupo de secuencias
  - Válido para análisis de genes, proteínas o poblaciones
- Encontrar miembros distantes de una familia de proteínas
  - Es muy frecuente que estén distantes, y el alineamiento de pares no suele ser lo suficientemente preciso para encontrarlos
- Clasificación y generación de BBDD de proteínas una vez secuenciado el genoma completo de un organismo
- Primer paso (y el más importante) en la generación de árboles filogenéticos

Alineamiento multiple de secuencias

- Existen cinco aproximaciones algorítmicas distintas al MSA
  1. Métodos exactos
  2. Alineamiento progresivo
  3. Aproximaciones iterativas
  4. Métodos basados en la consistencia
  5. Métodos basados en la estructura
- Las aproximaciones no son excluyentes
  - Las tres últimas, por ejemplo, utilizan alineamiento progresivo

Alineamiento multiple de secuencias Algoritmos Exactos

- Se basan en programación dinámica
  - Similar a un NW para alineamiento global de pares
- Aseguran un alineamiento óptimo, pero son lentos
  - No son factibles ni en espacio ni en tiempo si tenemos más de unas pocas secuencias
  - Para N secuencias de longitud media L, el coste en tiempo es O ($2^N L^N$)
- Se prefieren los métodos inexactos, mucho más rápidos
  - ClustalW: $O(N^4+L^2)$
  - MUSCLE: $O(N^4+ NL^2)$

Alineamiento multiple de secuencias

En la práctica problemas de alineamiento multiple se resuelven usando heuristicas

Alineamiento multiple Progresivo:

- Escoger 2 secuencias y alinearlas.
- Escoger una tercera secuencia y alinearla con respecto a las dos anteriores.
-Repetir hasta que todas las secuencias queden alineadas
-Existen diferentes opciones para escoger las secuencias y calcular los score de los alineamientos.

Alineamiento multiple de secuencias

Alineamiento multiple progresivo basado en Profiles.

CLUSTAL OMEGA
- Construir una matriz de distancias de todos los pares de secuencias usando la programacion dinámica.
-Alinear progresivamente los pares en orden decreciente de similitud

CLUSTALO – utiliza varias heuristicas para mejorar su presicion.

https://www.ebi.ac.uk/Tools/msa/clustalo/

Alineamiento multiple de secuencias

- Clustal implementa el algoritmo de Feng y Doolittle, que consta de 3 fases
  1. Alineamiento global 2 a 2 mediante el algoritmo de NW
     - Las puntuaciones de similitud se traducen a una matriz de distancias
  2. Se crea un árbol guía a partir de la matriz de distancias
  3. Se crea el alineamiento múltiple paso a paso
     1. Haciendo alineamientos de pares pero según las distancias

Alineamiento multiple de secuencias CLUSTAL Fase 1- Alineamiento Pares

- Ejemplo: cinco globinas muy conocidas, bastante distantes
  - NP_000509, NP_005359, NP_067080, 1FSL, 1D8U
  - Para 5 secuencias tendremos 10 alineamientos
  - Para *n* secuencias tendremos *n!/[2·(n-2)!]* alineamientos

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	beta_globin	147	2	myoglobin	154	25
1	beta_globin	147	3	neuroglobin	151	15
1	beta_globin	147	4	soybean	144	13
1	beta_globin	147	5	rice	166	21
2	myoglobin	154	3	neuroglobin	151	16
2	myoglobin	154	4	soybean	144	8
2	myoglobin	154	5	rice	166	12
3	neuroglobin	151	4	soybean	144	17
3	neuroglobin	151	5	rice	166	18
4	soybean	144	5	rice	166	43

Las puntuaciones se traducirán a distancias para que puedan usarse para generar el árbol

1 ← Mejor alineamiento

Alineamiento multiple de secuencias
CLUSTAL Fase 1- Similitud a Distancia

- Conversión de similitud a distancia (Feng y Doolittle)
  - Sea $S_{real(ij)}$ la similitud entre las secuencias *i* y *j*
    - Sea $S_{rand(ij)}$ la media de las similitudes calculadas para las 2 secuencias aleatorizadas muchas veces (p. ej. 1000)
    - Sea $S_{iden(ij)}$ la media de las similitudes identidad:

$$S_{iden(ij)} = \frac{S_{real(ii)} + S_{real(jj)}}{2}$$

    - Sea $S_{eff(ij)} = \frac{S_{real(ij)} + S_{rand(ij)}}{S_{iden(ij)} + S_{rand(ij)}} \times 100$
  - La distancia entre las secuencias *i* y *j* es $D_{ij} = -\ln S_{eff(ij)}$

http://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Feng-Doolittle

Alineamiento multiple de secuencias
CLUSTAL Fase 2 - Arbol Guia

- La longitud de las ramas depende de las distancias
- Se unen las ramas de las secuencias con distancias más cortas

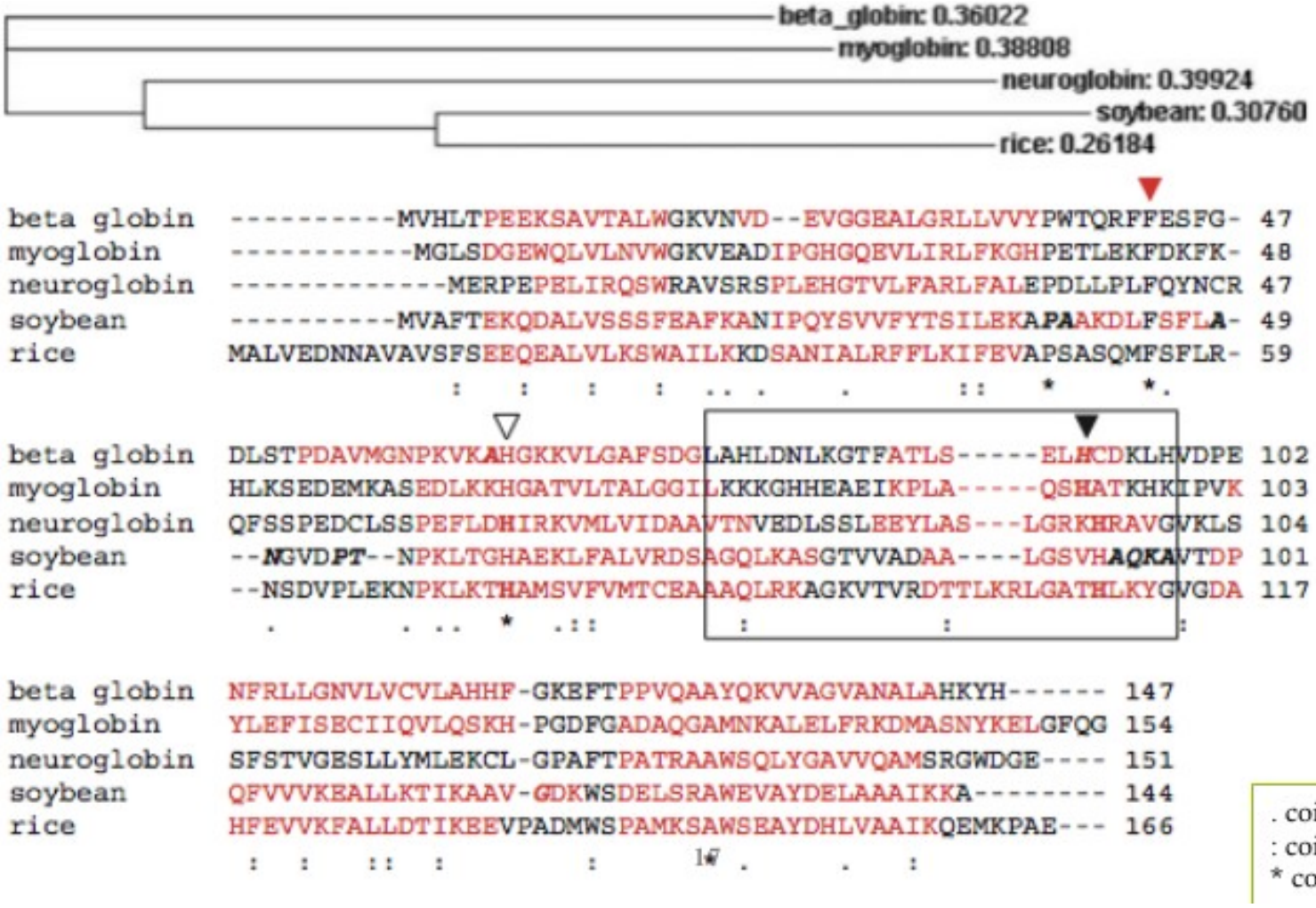
```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```

Formato Newick (.nwk)

Alineamiento multiple de secuencias
CLUSTAL Fase 3 – Alineamiento multiple

- Se seleccionan las dos secuencias más cercanas según el árbol guía
- Se realiza un alineamiento de pares entre ellas
- Se seleccionan las dos secuencias más cercanas siguientes
  - Si ninguna coincide con las anteriores, se realiza su alineamiento de pares
  - Si alguna coincide, se añade al alineamiento de pares, dando lugar a un alineamiento de 3+ secuencias, o *perfil*
- El alineamiento continúa hasta llegar a la raíz del árbol

Alineamiento multiple de secuencias
CLUSTAL Fase 3 – Alineamiento multiple



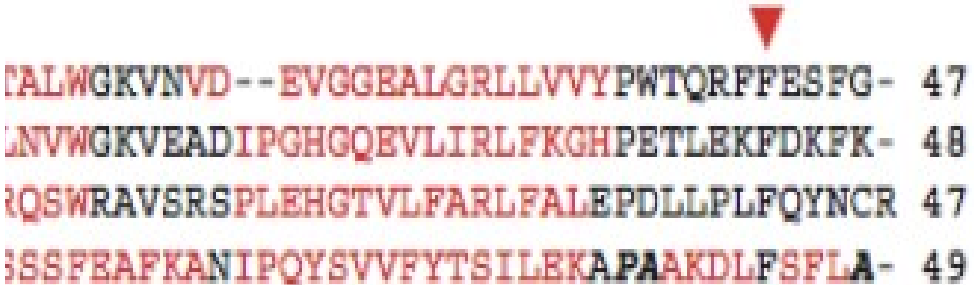
Alineamiento multiple de secuencias
Ejemplo - Interpretación

Hay una fenilanina muy conservada (flecha roja)

Hay una histidina muy conservada (flecha hueca)

Hay otra histidina, que a pesar de saberse que está muy conservada no se ha alineado bien (flecha negra)

Alineamiento multiple de secuencias
Ejemplo - Interpretación



Reading Multiple Sequence Alignment Output

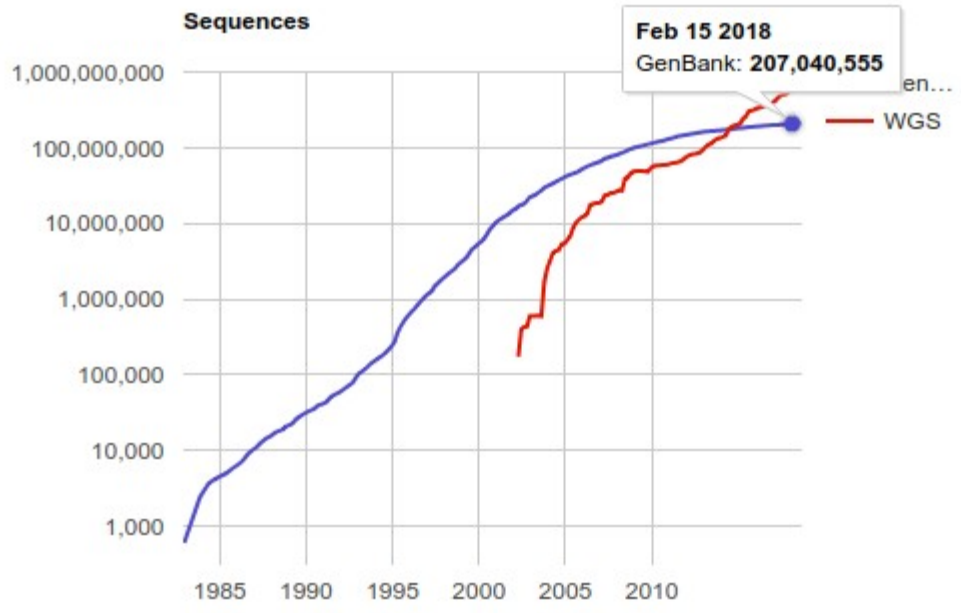
Symbol	Definition	Meaning
*	asterisk	positions that have a single and fully conserved residue
:	colon	conservation between groups of strongly similar properties with a score greater than .5 on the PAM 250 matrix
.	period	conservation between groups of weakly similar properties with a score less than or equal to .5 on the PAM 250 matrix

Alineamiento Rápido de secuencias

- El problema Biológico
- Estrategias de Búsqueda
- FASTA
- BLAST

El Problema biológico

- Los algoritmos de alineamiento Global y Local son muy lentos en la práctica.
- Considrear el escenario donde se requiere alinear una nueva secuencia – con una funcion desconocida, con las secuencias de una Base de Datos



A feb. del 2018 – 207 Millones de Secuencias en las BD de GENBANK

Problema con la gran cantidad de secuencias

- Crecimiento exponencial en el número y la longitud total de secuencias

- Una posible solución: Comparar solamente con "Organismos Modelo".

- Con una gran cantidad de secuencias existen probabilidades de que ocurran coincidencias aleatorias, por lo que se necesita un análisis estadistico.

Estrategias de Busqueda

Cómo aumentar la velocidad de búsqueda?

- Encontrar formas de limitar el número de compraciones.

Comparar secuencias a nivel de “palabras” para encontrar palabras comunes

- En este caso “palabras” significan sub - secuencias de longitud k.

Analizando el contenido de las palabras

Por ejemplo dada la secuencia I: TGATGATGAAGACATCAG
Entonces para k =8 el conjunto de k- tuples (o k-words) es:

TGATGATG
GATGATGA
ATGATGAA
TGATGAAG
...
GACATCAG

Existen n-k +1 k-tuples en una cadena de longitud n.

Analizando el contenido de las palabras

Se necesita considerar que si por lo menos una palabra de I no se encuentra en otra secuencia J, sabremos que I es diferente de J.
Se necesita considerar la significancia estadistica: Podrian haber coincidencias aleatorias.

Sean n= | I | y m = | J |

Analizando el contenido de las palabras

Los K-tuples de I se pueden listar en una Tabla (lista) de ocurrencias de palabras Lw(I)

Considerar k =2 y I= GCATCGGC:
GC, CA, AT, TC, CG, GG, GC
AT: 3
CA: 2
CG: 5
GC: 1, 7 ← Indices donde comienza el k-word GC en I
GG: 6
TC: 4

Construir la tabla Lw(I) toma O(n) tiempo.

K-words comunes

El número de k-words comunes en I y J puede ser computado usando las listas de ocurrencias Lw(I) y Lw(J)

Para cada palabra w en I existen |Lw(J)| ocurrencias en J

Por lo tanto, I y J tendrán $\sum_w |Lw(I)\ Lw(J)|$ palabras comunes

El tiempo de computo de esto sería $O(n+m+4^k)$

$O(n+m)$ para construir las listas

$O(4^k)$ tiempo para calcular la suma

K-words comunes

I = GCATCGGC
J = CCATCGCCATCG

L_w (I)	L_w (J)	Palabras comunes
AT: 3	AT: 3, 9	2
CA: 2	CA: 2, 8	2
	CC: 1, 7	0
CG: 5	CG: 5, 11	2
GC: 1, 7	GC: 6	2
GG: 6		0
TC: 4	TC: 4, 10	2
		10 en total

Propiedades de la Lista de palabras comunes

Coincidencias exactas se pueden encontrar utilizando búsquedas binarias $O(\log 4^k)$

Para valores grandes de k el tamaño de la tabla es demasiado grande para calcular el conteo de palabras comunes, de la manera mostrada

La tecnica de los ktuples puede combinarse con el algoritmo de alineamiento local para conseguir un método de alineamiento más rapido.

Alineamiento Rápido de secuencias

- El problema Biológico
- Estrategias de Búsqueda
- FASTA
- BLAST

FASTA

- Fasta Es un algoritmo de alineamiento de secuencias de varios pasos (Wilbur and Lipman, 1983)
- El formato de secuencias usado por el software FASTA es ampliamente usado por otros sistemas de análisis de secuencias
- La idea principal:
  - Escoger regiones de 2 secuencias que tienen algún grado se similitud
  - Efectuar el alineamiento local usando programacion dinámica en estas regiones.

FASTA - PASOS

·El algoritmo tiene cinco pasos:

1. Identificar las k-words comunes entre I y J
2. Puntuar las diagonales con coincidencias de k-words, identificar las 10 mejores diagonales.
3. Puntuar de nuevo las regiones iniciales con una matriz de scoring de sustituciones.
4. Juntar las regiones iniciales usando gaps, y penalizar por los gaps.
5. Realizar los alineamientos finales utilizando programación dinamica.

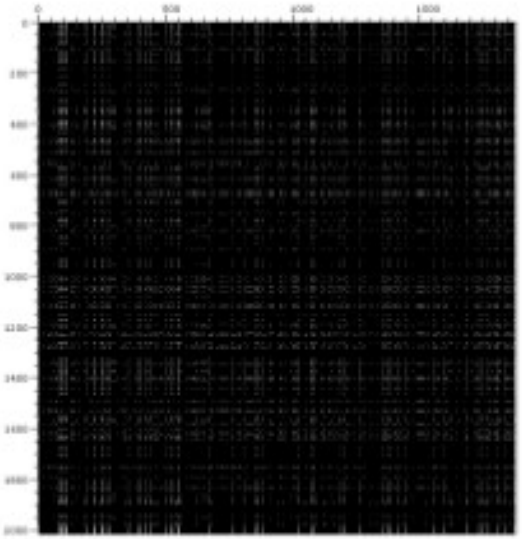
Comparaciones con Dot - Matrix

- La Coincidencia de k-words puede representarse con una matriz de puntos o dot matrix

- Un punto en la posicion (i,j) existe si la palabra que empieza en la posicion i en la secuencia I coincide exactamente con la palabra en la posicion j de la secuencia J

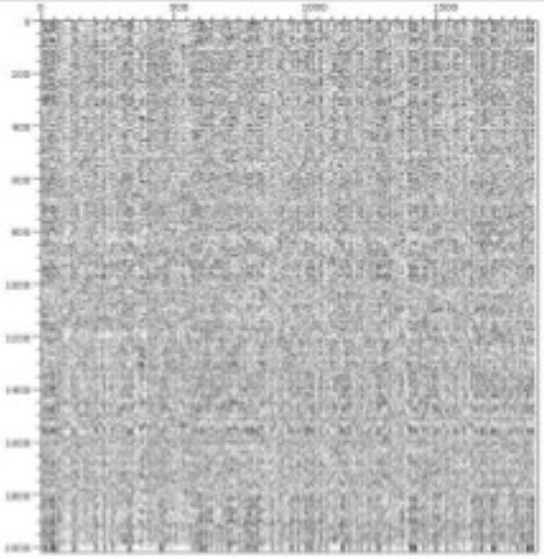
- Se pueden construir varias matrices para diferentes valores de k

Alineamiento Rápido de secuencias

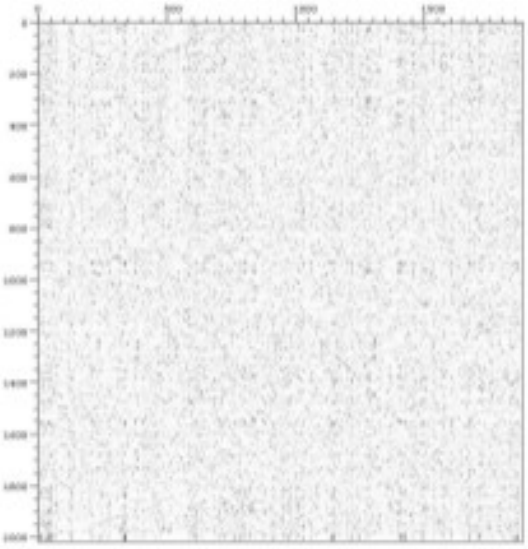


k=1

k=4

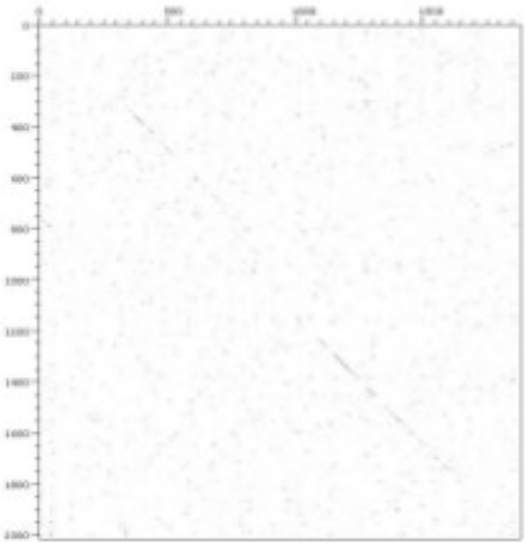


Dot matrix (k=1,4,8,16)
for two **DNA** sequences
X85973.1 (1875 bp)
Y11931.1 (2013 bp)



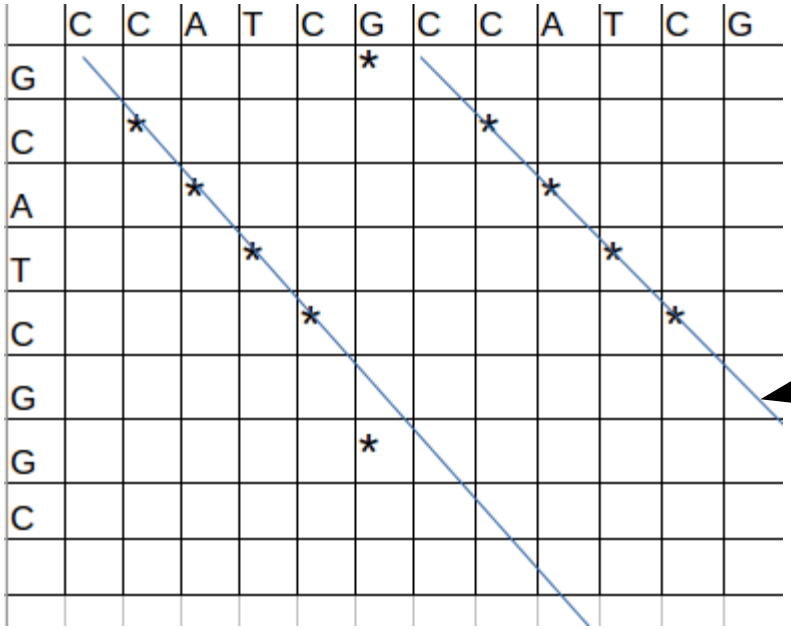
k=8

k=16



Calculando la suma de las Diagonales

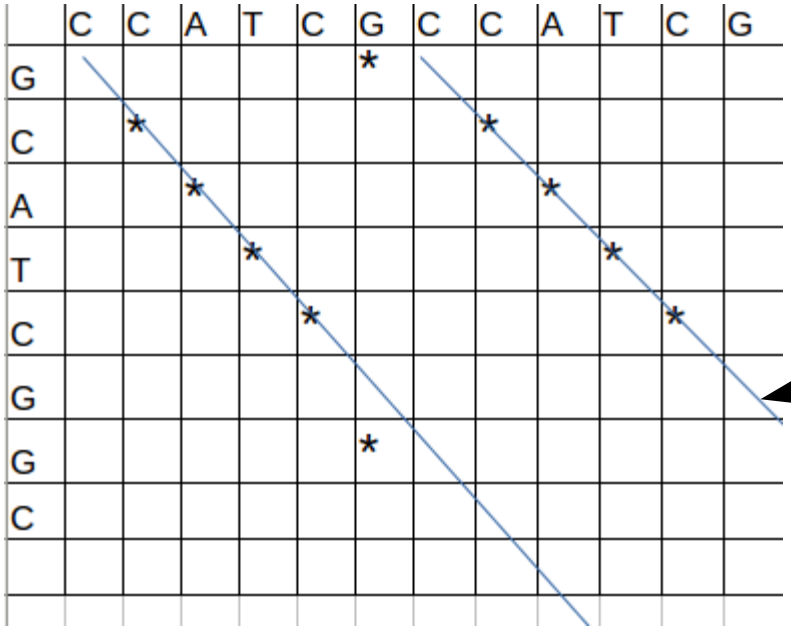
- Se requiere encontrar diagonales con altos puntajes en la matriz de puntos.
- Las diagonales se indexan por la diferencia: d = i -j



K=2
Diagonal:
d=i-j = -6

Calculando la suma de las Diagonales

- Se requiere encontrar diagonales con altos puntajes en la matriz de puntos.
- Las diagonales se indexan por la diferencia: d = i -j



K=2
Diagonal:
d=i-j = -6

Calculando la suma de las Diagonales

Ejemplo: Sean I =GCATCGGC y J = CCATCGCCATCG; k=2.

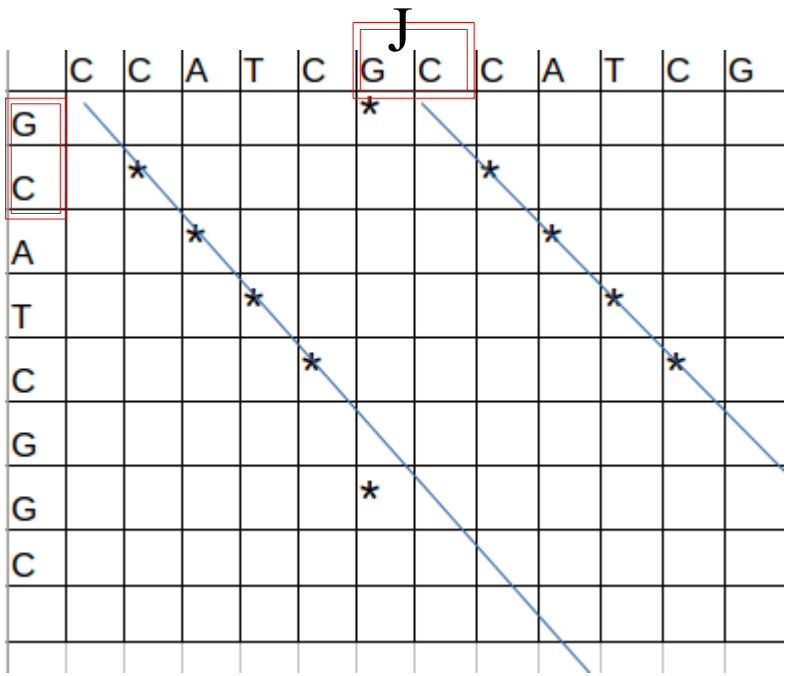
- Construir la lista de kwords $L_w(J)$
- Las sumas de las diagonales se calculan en una tabla indexada con la diferencia e inicializada en cero.

d.	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
S_d.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Calculando la suma de las Diagonales

Recorrer las kwords para I, buscar coincidencias en $L_w(J)$

Y actualizar los valores de las sumas de las diagonales.



Para la primera 2-word en I , GC: $L_{GC}(J)= 6$.

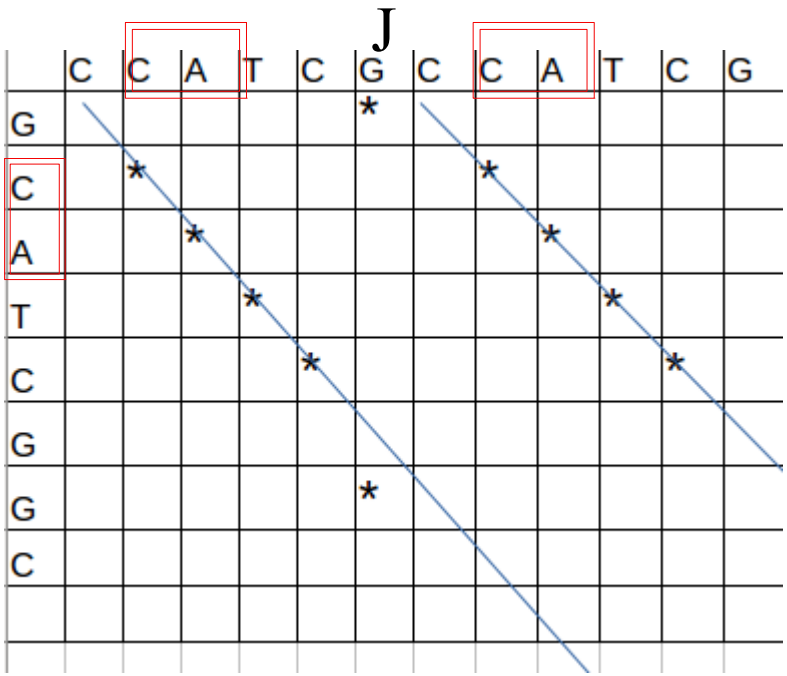
Luego se actualiza la suma de la diagonal d=i-j = 1-6=-5

$S_{-5} = S_{-5} +1 = 0+1 = 1$

Calculando la suma de las Diagonales

Recorrer las kwords para I, buscar coincidencias en $L_w(J)$

Y actualizar los valores de las sumas de las diagonales.



La siguiente 2-word en I es CA: $L_{CA}(J)= \{2,8\}$.

Se actualizan 2 sumas de las diagonales $d=i-j = 2-2=0$

$S_0 = S_0 +1 = 0+1 = 1$

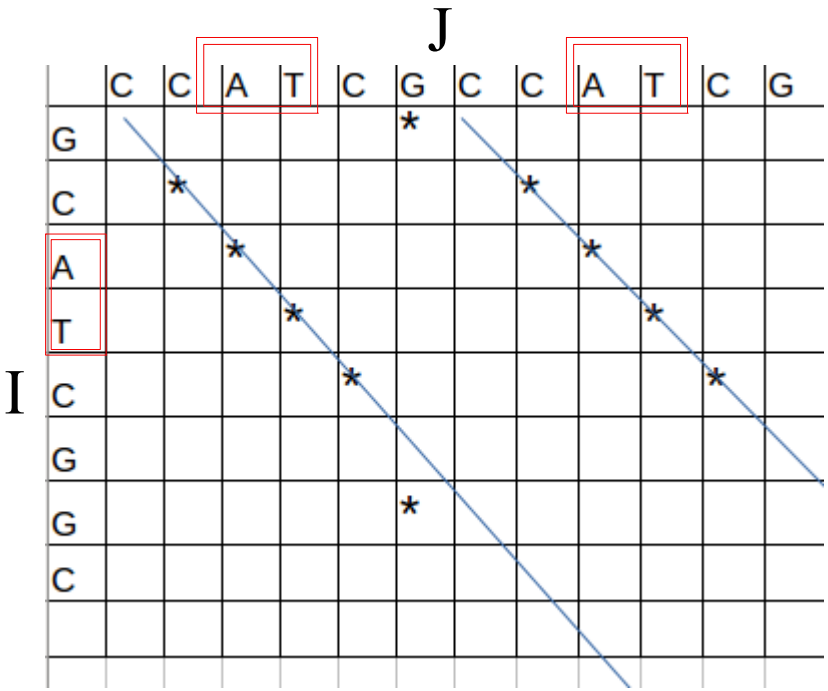
$d=i-j = 2-8=-6$

$S_{-6} = S_{-6} +1 = 0+1 = 1$

Calculando la suma de las Diagonales

Recorrer las kwords para I, buscar coincidencias en $L_w(J)$

Y actualizar los valores de las sumas de las diagonales.



La siguiente 2-word en I es AT: $L_{AT}(J)= \{3,9\}$.

Se actualizan 2 sumas de las diagonales $d=i-j = 3-3=0$

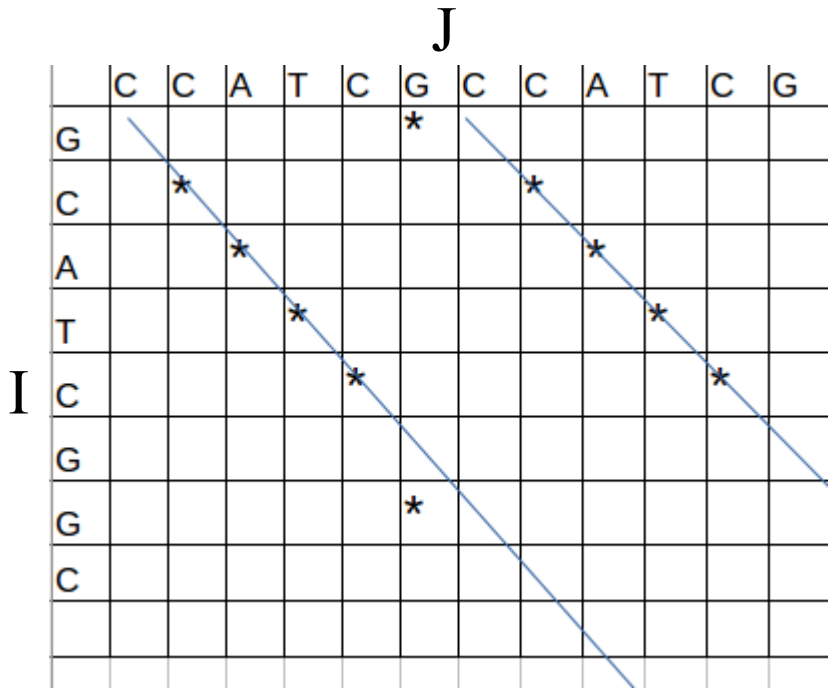
$S_0 = S_0 +1 = 1+1 = 2$

$d=i-j = 3-9=-6$

$S_{-6} = S_{-6} +1 = 1+1 = 2$

Calculando la suma de las Diagonales

d.	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
S_d.	0	0	0	0	4	1	0	0	0	0	4	1	0	0	0	0	0



Despues de recorrer todas las 2-words de I el resultado es el mostrado

Algoritmo para la suma de las Diagonales

$S_d := 0$ for all $1 - m <= d <= n-1$

Calcular $L_w(J)$ para todas las w -words

```
for i := 1 to n – k – 1 do
      w := I_i I_{i+1} ...I_{i+k-1}
      for j ∈ L_w(J) do
            d := i – j
            S_d := S_d + 1
      end
end
```

FASTA - PASOS

·El algoritmo tiene cinco pasos:

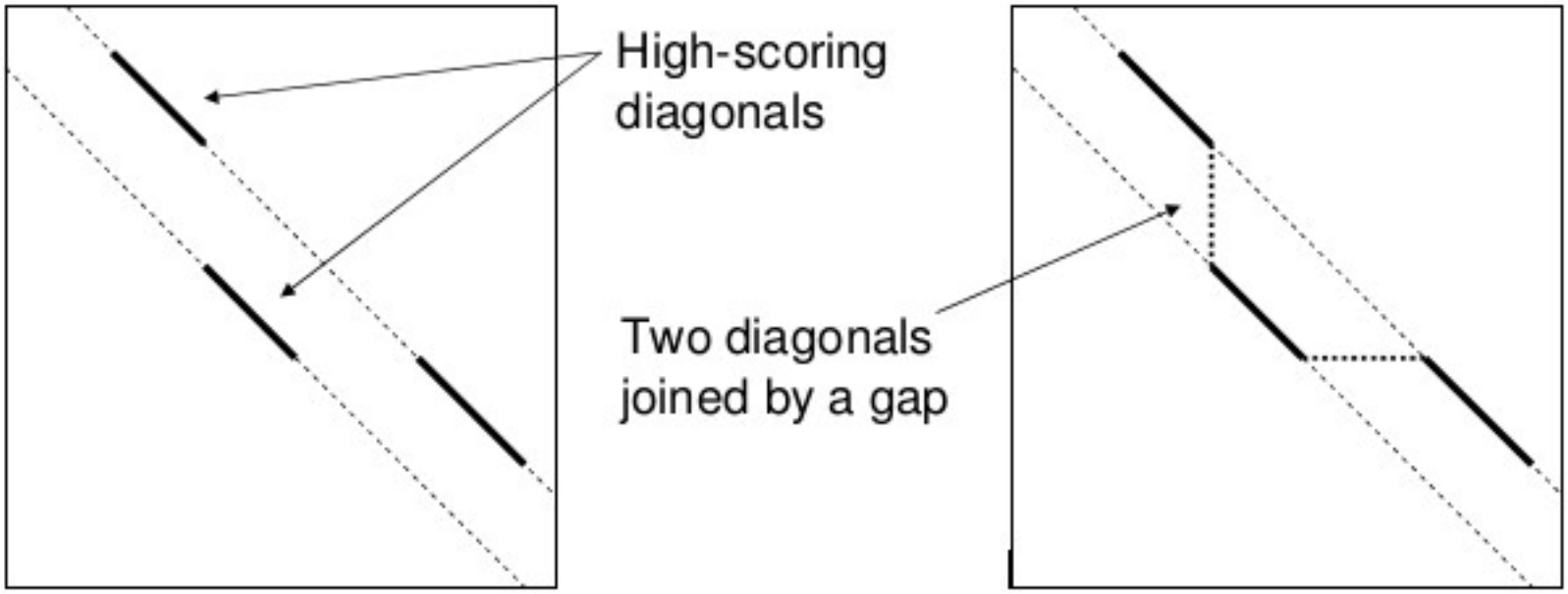
1. Identificar las k-words comunes entre I y J
2. Puntuar las diagonales con coincidencias de k-words, identificar las mejores diagonales.
3. Puntuar de nuevo las regiones iniciales con una matriz de scoring de sustituciones.
4. Juntar las regiones iniciales usando gaps, y penalizar por los gaps.
5. Realizar los alineamientos finales utilizando programación dinamica.

Recalculo de Score para las diagonales escogidas

- A Las diagonales escogidas en el paso anterior se les recalcula el scoring de acuerdo con una matriz de scoring
- Esto se hace para encontrar sub regiones con coincidencias mas cortas que el valor de k
- Los extremos de la diagonal que no coinciden – se descartan.

Juntando las diagonales

- Dos Diagonales pueden ser unidas con un gap si el alineamiento resultante tiene un score más alto
- Se utilizan gaps de separacion apertura y extension
- Se encuentra la mejor combinacion de diagonales.



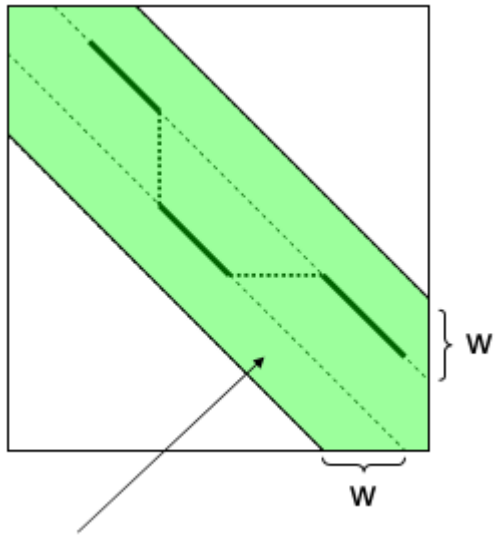
FASTA - PASOS

·El algoritmo tiene cinco pasos:

1. Identificar las k-words comunes entre I y J
2. Puntuar las diagonales con coincidencias de k-words, identificar las mejores diagonales.
3. Puntuar de nuevo las regiones iniciales con una matriz de scoring de sustituciones.
4. Juntar las regiones iniciales usando gaps, y penalizar por los gaps.
5. Realizar los alineamientos finales utilizando programación dinamica.

Alineamiento local en la region con el máximo score

- El último paso de FASTA es realizar el alineamiento local usando programación dinámica en la region con el máximo score.
- La region a ser alieneada cubre desde las diagonales con offset -w y +w hasta las diagonales con los score maximos
- Para secuencias grandes, esta region es pequeña comparada con toda la matriz nxm



Zona donde se llena la matriz para la programación dinámica

Propiedades de FASTA

- Es rápido comparado con el algoritmo de alineamiento local, usando solamente programación dinámica
  - Solamente una sección angosta de toda la matriz es alineada
- Incrementando el parametro k disminuye el numero de coincidencias, Se incrementa la especificidad, disminuye la sensitividad
- FASTA puede ser muy especifico para hallar regiones largas de baja similitud
  - El metodo especifico no produce muchos resultados incorrectos
  - El método sensitivo produce muchos de los resultados correctos

Propiedades de FASTA

Demo de FASTA en EBI
http://www.ebi.ac.uk/fasta/

Alineamiento Rápido de secuencias

- El problema Biológico
- Estrategias de Búsqueda
- FASTA
- BLAST

BLAST: Basic Local Alignment Search Tool

Algoritmo diseñado en 1990, por Altschul y otros.
Este algoritmo y algunas de sus variantes son los más usados en la actualidad
El BLAST tiene 3 partes:

1. Buscar alineamientos locales (seed hits) entre la secuencia que nos interesa y una secuencia de la Base de datos.
2. Extender los alineamientos locales encontrados en alineamientos locales con grandes scores.
3. Calcular los expected values E-Values y ordenar en un ranking los alineamientos locales

Los alineamientos locales con grandes scores se denominan High Scoring Segment Pairs (HSP)
En 1997 se introdujo Gapping BLAST, que permite introducir gaps en las alineaciones.

BLAST: Buscar Alineamientos locales (seed hits)

Primero se generan un conjunto de secuencias de proximidad dados los valores de **k**, la matriz de scoring y un limite **T**.
Las secuencias de proximidad de una k-word **w** incluyen todas las cadenas de longitud **k**, que cuando están alineadas con **w** tienen un score de alineamiento de por lo menos **T**

Por Ejemplo sea I = GCATCGGC, J = CCATCGCCATCG y k= 5, match score =1, mismatch score = 0 y T = 4

BLAST: Buscar seed hits

Por Ejemplo sea I = GCATCGGC, J = CCATCGCCATCG y k= 5, match score =1, mismatch score = 0 y T = 4

Esto (T=4) permite un mismatch en cada 5-word.
La proximidad de la primera k-word de I, GCATC es GCATC y las siguientes 15 secuencias:

$$\{A,T\}CCATC,\ G\{A,T\}GATC,\ GC\{C,T\}GTC,\ GCA\{A,G\}C\ C,\ GCAT\{A,G,T\}$$

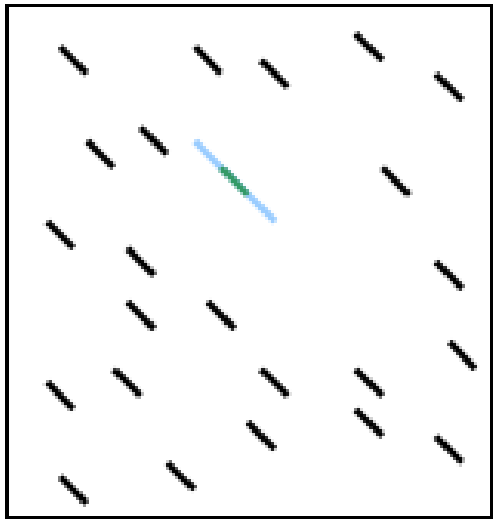
BLAST: Buscar seed hits

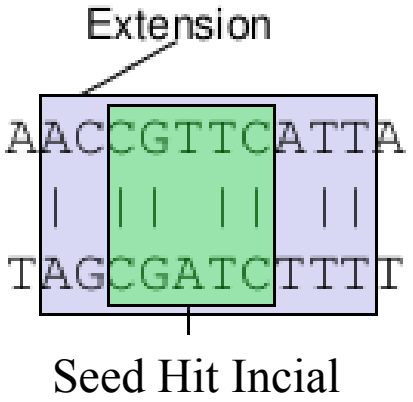
I = GCATCGGC, tiene 4 k-words (k=5), y por lo tanto 4 x 16 = 64 5-word seed hits que hay que buscar en la secuencia J

Estas Seed hits se pueden buscar en un tiempo proporcional a la suma de las longitudes de las word seed + la longitud de J + el numero de coincidencias.

Extender las seed hits – BLAST Original

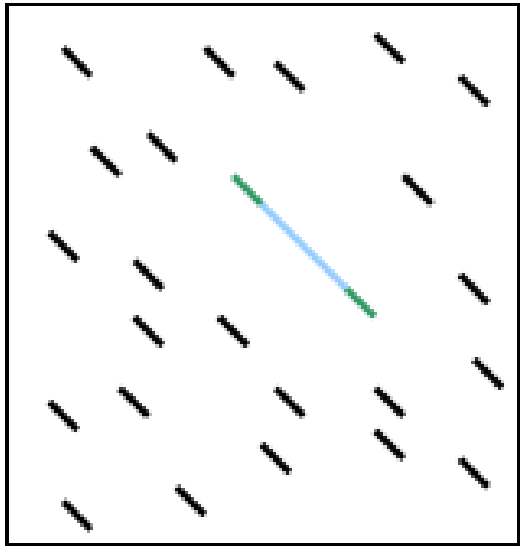
- Las Seed Hits iniciales son extendidas
- Las extensiones no agregan gaps al alineamiento.
- La secuencia es extendida en HSP (High Scoring segment Pairs) hasta que el score del alineamiento baja debajo del valor del score máximo obtenido menos el valor de un parámetro umbral.
- Todos los HSP estadisticamentte significantes se reportan.





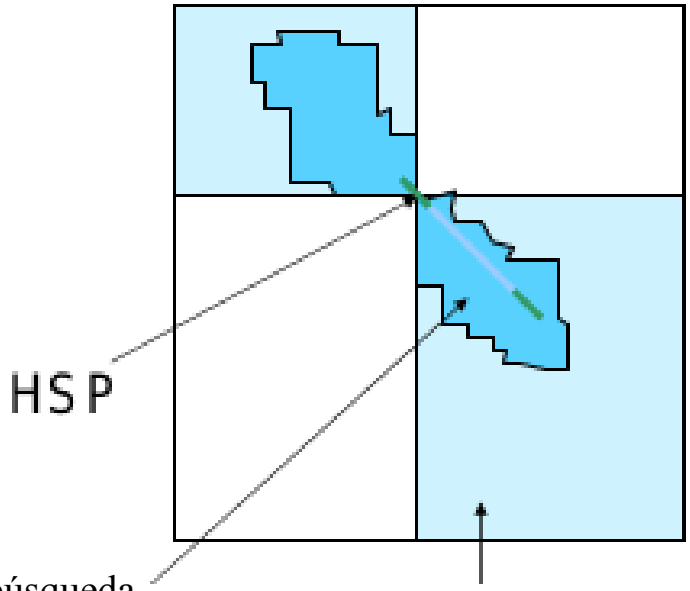
Extender las seed hits – Gapped BLAST

- En una version posterior 2 Seed Hits deberán encontrarse en una diagonal.
  - Las Seed Hits no deben sobreponerse.
  - Si estan mas cerca que un valor dado A (Un Nuevo parámetro- A), entonces se unen en una High Scoring Segment Pair - HSP
- El valor de umbral T se baja para alcanzar una sensitividad comparable
- Si la HSP resultante alcanza un score de por lo menos S_g – se registra una extencion con gaps (Gapped Extension)



Extension con Gaps de HSP

- Se realiza un alienamiento local empezandeo desde la HSP
- La matriz de programacion dinamica se llena en las direcciones hacia adelante y hacia atras – Ver figura.
- No se toman en cuenta celdas donde el valor X_g será menor que el mejor score de alineamiento encontrado hasta el momento.



E-value

El E-value – más conocido como EXPECT value – es una función del score, el tamaño de la base de datos y de la longitud de la secuencia 'query'.

E-value: Número de alineamientos con un score >=S que se espera encontrar si la base de datos es una colección de letras al azar. Cuanto menor es el E-value, mas significante es el alineamiento

Valores de E-value menores que 1e-6 ($1* 10^{-6}$) son generalmente muy buenos para proteínas, mientras que E<1e-2 ($1* 10^{-2}$) puede considerarse significativo.

Solo serán mostradas las secuencias cuyo E-value supere un determinado umbral

Variantes de BLAST

BLASTP: proteína – proteína

BLASTN: nucleótido – nucleótido

BLASTX: nucleótido (Trad.) – proteína

TBLASTN: proteína - nucleótido (Trad.)

TBLASTX: nucleótido (Trad.) - nucleótido (Trad.)

- http://www.ncbi.nlm.nih.gov/BLAST/

Otros Recursos web para el alineamiento multiple de secuencias

Clustal Omega:	http://www.ebi.ac.uk/Tools/msa/clustalo/
T-Coffee:	http://www.ebi.ac.uk/Tools/msa/tcoffee/
Kalign :	http://www.ebi.ac.uk/Tools/msa/kalign/
MAFFT:	http://www.ebi.ac.uk/Tools/msa/mafft/
MUSCLE:	http://www.ebi.ac.uk/Tools/msa/muscle/