Búsqueda en bases de datos

Similitud, homología.

Métodos heurísticos.

Recordemos ...

- El alineamiento por parejas es el proceso de alinear dos secuencias hasta conseguir el máximo de identidad entre ellas,
  y en el caso de AA también el máximo de conservación, para establecer
  el *grado de similaridad* y la *posibilidad de homología* entre ambas secuencias.

Homología vs similaridad

- Similaridad
  - El grado de relación entre dos secuencias. Se basa en la combinación de identidad y conservación
  - Identidad
    - El grado en que 2 secuencias de AA o nucleótidos son invariantes.
  - Conservación
    - Cambios en una posición especifica de una secuencia de aminoacidos que mantienen las propiedades quimicas de la secuencia original
- Homología:
  - Similaridad atribuible a la descendencia de un ancestro común

Infiriendo homología

- Si, al alinear dos secuencias, se obtiene una puntuación “elevada” se puede inferir que ambas secuencias son posibles homólogos.
- Esto sugiere que para encontrar secuencias homólogas a una dada, en una base de datos, tan sólo hace falta alinear ésta con todas las de la BD.
- Las secuencias con mayor puntuación serán posibles homólogas de la secuencia problema.

Predicción de la función de un gen o una proteína

- La evolución es un proceso conservativo
  - Cambian los residuos en una secuencia
  - Pero se conservan las propiedades bioquímicas y los procesos fisiológicos
- Si somos capaces de encontrar *secuencias homólogas* a la secuencia problema podemos concluir que ésta “debe de tener” propiedades similares a las de la secuencia conocida.
- La búsqueda (el hallazgo, de hecho) de secuencias homólogas puede ser una vía para predecir la función de una proteína o un gen.

Un método para encontrar homólogos

- Dada una secuencia problema y una base de datos de secuencias en donde se desea encontrar homólogos de dicha secuencia problema …
  - Se escoge un sistema de puntuación (matriz, costes, …) y se alinea la secuencia con cada una de las que contiene la base de datos.
  - Se ordenan las secuencias de mayor a menor puntuación.
  - Las secuencias de la BD con mayor puntuación y un mínimo grado de similaridad/identidad son
    - *Similares* a la secuencia problema
    - Candidatas a ser *homólogas*.

Algunos problemas…

- El método sugerido presenta algunas dificultades obvias
  - ¿Cuánto es “un mínimo de similaridad”?
  - ¿Cómo se obtiene un método que seleccione el máximo de homólogos reales y descarte las similaridades no debidas a homología?
  - Que algoritmo es el más adecuado para realizar los alineamientos?

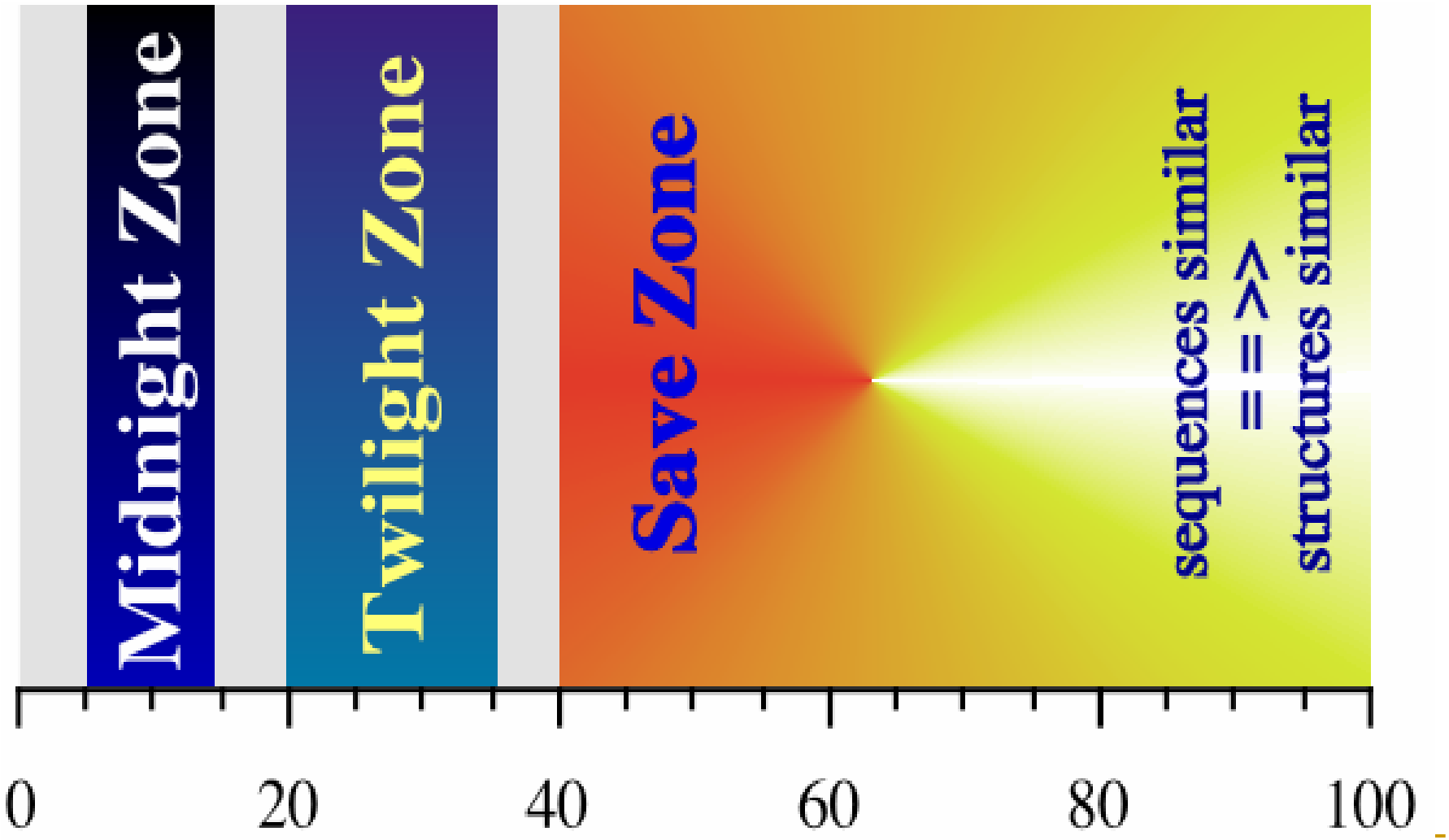
Grado de similaridad y homología

Rule-of-thumb sequence comparison

100% identical	fairly identical
one point mutation	may knock out function
ten mutations (spread)	similar structure
ten mutations (adjacent)	may be: Δ function, Δ localisation, Δ local structure
ten identical residues	may be similar: aspects of function local structure
similar motifs	typically relevant for functional aspects
similar domains	possibly relevant for function and structure

Fuente: Bukhard Rost (2006)

Zones



Fuente: Bukhard Rost (2006)

Sensibilidad y especificidad

Exitos y fracasos en la búsqueda

- Si conociéramos TODAS las coincidencias entre una secuencia problema y una base de datos …
- Podríamos distinguir si, una vez declarada una secuencia cómo homóloga a la secuencia problema esta homolgía es cierta o falsa.
- Es decir podríamos distinguir entre.
  - Verdaderos Positivos (TP): homólogos auténticos.
  - Falsos positivos (FP): similaridad debida al azar.
  - Verdaderos Negativos (TN): secuencias no relacionadas.
  - Falsos Negativos (FN): homólogos no detectados.

Verdaderos/Falsos Positivos/Negativos

Realidad / Detección	Secuencias homólogas	Secuencias no homólogas
Secuencias consideradas relacionadas	Verdaderos Positivos *(Homólogos auténticos)*	Falsos Positivos *(Similaridad debida al azar)*
Secuencias consideradas como no relacionadas	Falso Negativo *(Homólogos no detectados)*	Verdaderos Negativos *(Secuencias no relacionadas)*

Sensibilidad frente a Especificidad

- Sensibilidad= TP /(TP+FN)
  - *Capacidad del algoritmo para detectar auténticos positivos*
  - *% de coincidencias bien identificadas*
- Especificidad = TN / (TN+FP)
  - *Capacidad del algoritmo para declarar negativas las secuencias sin relación.*
  - *% de negativos correctos*

El compromiso entre sensibilidad y especificidad

- Si en una búsqueda colocamos el umbral alto →
  - Cuesta localizar los positivos→ Pocos FP
  - Pero tendremos más falsos negativos

  Es decir un umbral alto suele conllevar una baja sensibilidad y una alta especificidad
- AL reves si colocamos un umbral bajo
  - Tendremos muchos positivos → Tambien más FP
  - Pero habran menos falsos negativos

  Es decir un umbral bajo conlleva una alta sensibilidad y una baja especificidad
- *Idealmente:mirar de lograr un equilibrio,*
- *O en todo caso decidir que error nos interesa más controlar en cada situación*

Scoring table: PAM 300 Gap open 12; Gap extend 2

Result No.	Score	Query Match	Length	DB	ID	Description	Pred. No.
1	1390	98.5	335	1	PTP1_YEAST	Protein-tyrosine phosp	4.73e-207
2	343	24.3	711	1	PYP2_SCHPO	Protein-tyrosine phosp	1.97e-33
3	339	24.0	550	1	PYP1_SCHPO	Protein-tyrosine phosp	8.11e-33
4	324	23.0	1442	1	PTPG_MOUSE	Protein-tyrosine phosp	1.61e-30
5	321	22.7	2314	1	PTPZ_HUMAN	Receptor-type protein-	4.60e-30
6	321	22.7	2316	1	PTPZ_RAT	Receptor-type protein-	4.60e-30
7	316	22.4	1445	1	PTPG_HUMAN	Protein-tyrosine phosp	2.65e-29
8	316	22.4	1422	1	PTPG_CHICK	Protein-tyrosine phosp	2.65e-29
9	303	21.5	1457	1	PTPK_MOUSE	Receptor-type protein-	2.47e-27
10	299	21.2	434	1	PTN1_CHICK	Protein-tyrosine phosp	9.91e-27
90	174	12.3	750	1	PTP2_YEAST	Protein-tyrosine phosp	7.99e-09
91	138	9.8	928	1	PTP3_YEAST	Protein-tyrosine phosp	3.15e-04
95	110	7.8	478	1	YDIU_SHIFL	Hypothetical UPF0061 p	5.47e-01
96	109	7.7	478	1	YDIU_ECO57	Hypothetical UPF0061 p	7.02e-01
108	101	7.2	296	1	RN15_YEAST	mRNA 3'-end processing	4.91e+00
121	98	6.9	341	1	YH10_YEAST	Hypothetical 37.9 kDa	9.92e+00

< 0.05

< 1.00

```
RESULT   90
ID    PTP2_YEAST     STANDARD;      PRT;   750 AA.
DE    Protein-tyrosine phosphatase 2 (EC 3.1.3.48) (PTPase 2).

  DB  1;  Score     174;  Match 26.9%;  QryMatch 12.3%;  Pred. No. 7.99e-09;
  Matches    76;  Conservative   69;  Mismatches  89;  Indels  49;  Gaps  11;

           ******.*.*.      *.*.. *****.*   * *.**..      *... **. . .
Db    465 NDYINANYLKLT----QINPDFKYIATQAPLPSTMDDFWKVI----TLNKVKVIISLNSD 516
Qy     77 ndyinasyvkvnvpgqsiepgy-yiatqgptrktwdqfwqmcyhncpldni-vivmvtpl 134

            * *  *   **    .  ..*... . **.  . *.      ..*    . ... . .
Db    517 DELNLRKWDIYWNNLSYSNHTIKLQNTWENICNINGCVLRVFQVKKTAPQNDNISQDCDL 576
Qy    135 veynrekcyqywprgg-vddtvriaskwespggandmtqfpsdlkiefvnvhkvkdyytv 193

             .   **.    *.    * .. .  * *                    . .*.  .  .*
Db    577 PHNGDLTSITMAVSEPFIVYQLQYKNWLDSCGVDMNDIIKLHKVKNSLLFNPQSFITSLE 636
Qy    194 tdi-kltptdplvgpvktvhhfyfdlwkd------------------mnkpeevvpime 233

             .*          .. *   . ..*..****** **** *..** *.        ..  .
Db    637 KDVCKPDLIDDNNSELHLDTANSSPLLVHCSAGCGRTGVFVTLDFLL------SILSPTT 690
Qy    234 --lc---------ahshslnsrgnpiivhcsagvgrtgtfialdhlmhdtldfkniters 282

            **..    . *.***  ** .** **. ***   *..  *.*
Db    691 NHSNKIDVWNMTQDLIFIIVNELRKQRISMVQNLTQYIACYEA 733
Qy    283 rhsdrate-eytrdlieqivlqlrsqrmkmvqtkdqflfiyha 324
```

Low sensitivity, many false negatives

High selectivity, few false positives

Scoring table: PAM 50 Gap open 40; Gap extend 7

Result No.	Score	Query % Match	Length	DB	ID	Description	Pred. No.
1	3660	100.0	335	1	PTP1_YEAST	Protein-tyrosine phosp	0.00e+00
2	254	6.9	1445	1	PTPG_HUMAN	Protein-tyrosine phosp	6.39e-48
3	251	6.9	1422	1	PTPG_CHICK	Protein-tyrosine phosp	5.98e-47
4	248	6.8	1442	1	PTPG_MOUSE	Protein-tyrosine phosp	5.55e-46
5	226	6.2	711	1	PYP2_SCHPO	Protein-tyrosine phosp	5.73e-39
6	217	5.9	2316	1	PTPZ_RAT	Receptor-type protein-	3.81e-36
7	208	5.7	595	1	PTN6_MOUSE	Protein-tyrosine phosp	2.36e-33
8	208	5.7	613	1	PTN6_RAT	Protein-tyrosine phosp	2.36e-33
9	208	5.7	1216	1	PTPO_HUMAN	Receptor-type protein-	2.36e-33
10	207	5.7	1452	1	PTPM_MOUSE	Receptor-type protein-	4.79e-33
31	181	4.9	550	1	PYP1_SCHPO	Protein-tyrosine phosp	3.43e-25
69	151	4.1	750	1	PTP2_YEAST	Protein-tyrosine phosp	1.35e-16
90	120	3.3	928	1	PTP3_YEAST	Protein-tyrosine phosp	1.83e-08
91	117	3.2	171	1	VH01_RACVI	Dual specificity prote	9.89e-08
92	117	3.2	171	1	DUSP_VACCV	Dual specificity prote	9.89e-08
93	117	3.2	171	1	DUSP_VACCC	Dual specificity prote	9.89e-08
105	108	3.0	551	1	CC14_YEAST	Probable protein-tyros	1.33e-05
108	97	2.7	489	1	MSG5_YEAST	Protein-tyrosine phosp	3.62e-03
110	93	2.5	468	1	YOPH_YERPS	Protein-tyrosine phosp	2.47e-02
111	93	2.5	468	1	YOPH_YEREN	Protein-tyrosine phosp	2.47e-02
140	84	2.3	312	1	DCTD_YEAST	Deoxycytidylate deamin	1.41e+00

< 1.00

RESULT 69
ID PTP2_YEAST STANDARD; PRT; 750 AA.
DE Protein-tyrosine phosphatase 2 (EC 3.1.3.48) (PTPase 2).

DB 1; Score 151; Match 62.5%; QryMatch 4.1%; Pred. No. 1.35e-16;
Matches 15; Conservative 4; Mismatches 5; Indels 0; Gaps 0;

```
          .*  ****** **** *..** *.
Db    660 SPLLVHCSAGCGRTGVFVTLDFLL 683
Qy    246 npiivhcsagvgrtgtfialdhlm 269
```

High sensitivity, few false negatives

Low selectivity, many false positives

Algoritmos de búsqueda

Busqueda basada en PD

- Una forma razonable de comparar una secuencia con las de una base de datos es mediante alineamientos locales.
- Algoritmo: Smith-Waterman
  - Encuentra el mejor alineamiento en cada caso
  - Sólo impone una restricción: Puntuación > 0
  - Proporciona la mejor sensibilidad

Inconvenientes de la búsqueda basada en programación dinámica

- La busqueda basada en PD proporciona una gran sensibilidad pero
  - Es poco específica→ Pocos falsos negativos: Fàcil perder las “homologías remotas”
  - Es necesariamente lenta.
- Alternativa: Métodos heurísticos
  - Aproximaciones a SW con restricciones que:
    - Aumentan la especificidad (aunque baja la sensibilidad)
    - Són mucho más rápidas

Métodos heurísticos

FastA y BLAST

Principios básicos de los métodos heurísticos (1)

- La PD proporciona el mejor alineamiento entre dos secuencias, dado un sistema de puntuación.
- Los métodos heurísticos recorren el espacio de búsqueda con métodos rápidos y aproximados para
  - encontrar las secuencias en la BD que es más probable que se parezcan a la problema y
  - localizar la región de similaridad entre secuencias.
- El proceso de alineamiento está restringido
  - A las secuencias seleccionadas
  - A una porción dentro de las secuencias

Principios básicos de los métodos heurísticos (2)

- Los métodos utilizados (BLAST/FastA) se denominan *heurísticos*: se basan en un método empírico que usa reglas de decisión para encontrar soluciones.
- Casi siempre encuentran secuencias relacionadas en una BD aún cuando no pueden garantizar una solución óptima.
- Son mucho más rápidos (50-100 veces) que los métodos basados en PD

Búsqueda de una secuencia en una Base de datos con FastA o con BLAST

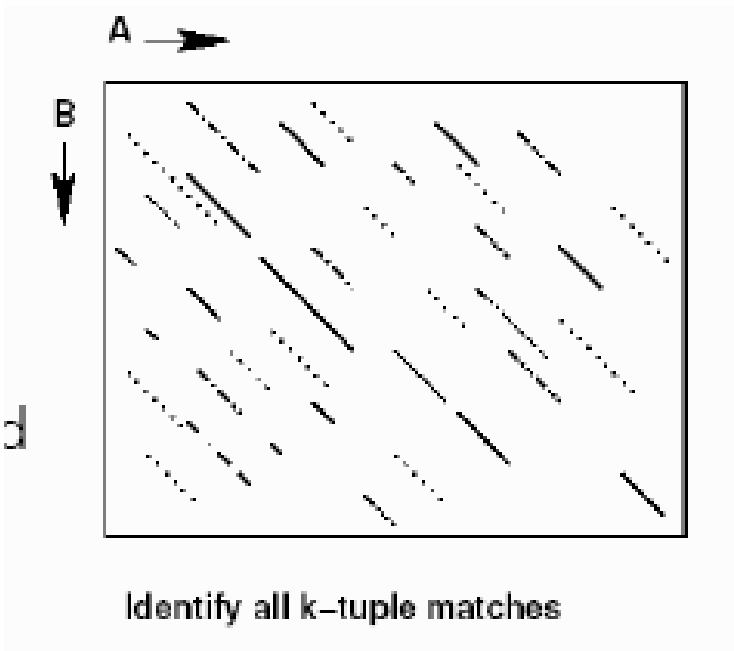
- Se realiza con FastA o con BLAST
- Uso de ambos programas es parecido.
- En ambos casos, para utilizarlos de manera eficiente, interesa conocer…
  - Qué hace el programa (el algoritmo).
  - Que parámetros de entrada necesita.
  - Que resultados proporciona y como debemos interpretarlos.

FastA

Introducción a FastA

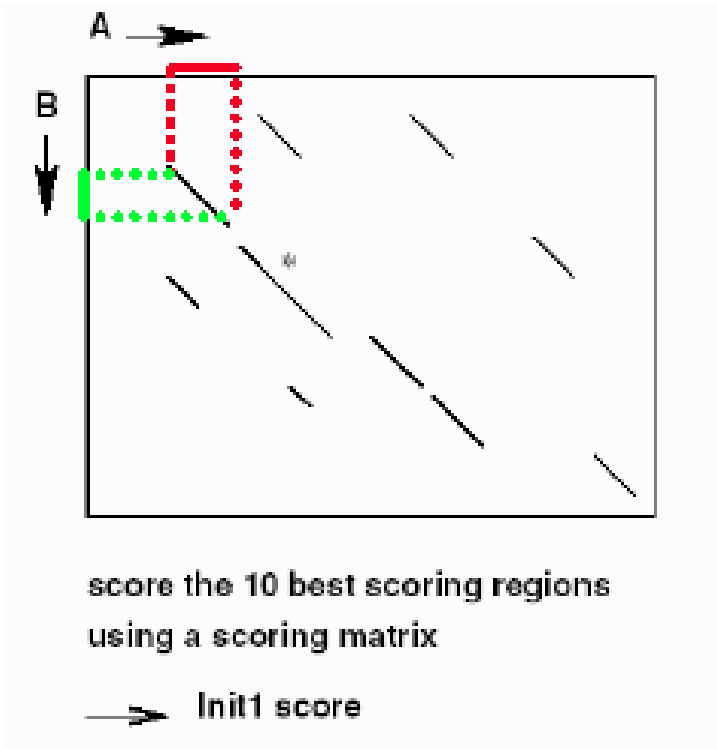
- Desarrollado a partir de 1985 por Pearson y Lipman.
- Se trata de una familia de programas
  - fasta3, fastxy3, tfastxy,…
- Desde sus primeras versiones admite uso de *gaps* y proporciona valores de significación.
- Se inicia buscando coincidencias idénticas entre la secuencia y cada una de las de la BD: se asocia con *busqueda por identidad*

El algoritmo de FastA (1)



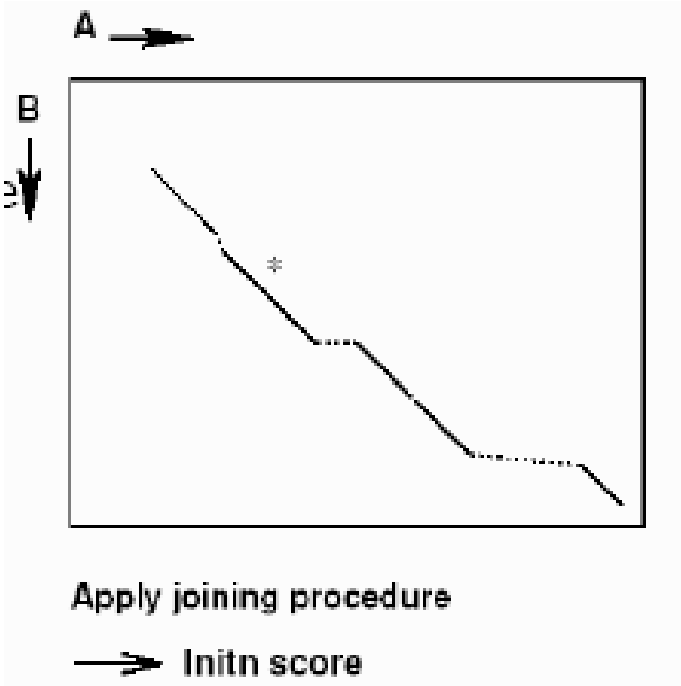
- Empieza localizando regiones de la secuencia problema y la de la BD con una alta densidad de coincidencias exactas de un tamaño mínimo (k-tuplas o palabras).
- Típicamente
  - K=2 para AA
  - K=6 para AN
- Recuerda la idea del “dot-plot”: busca rachas de coincidencias que formarían diagonales en un dot-plot.

El algoritmo de FastA (2)



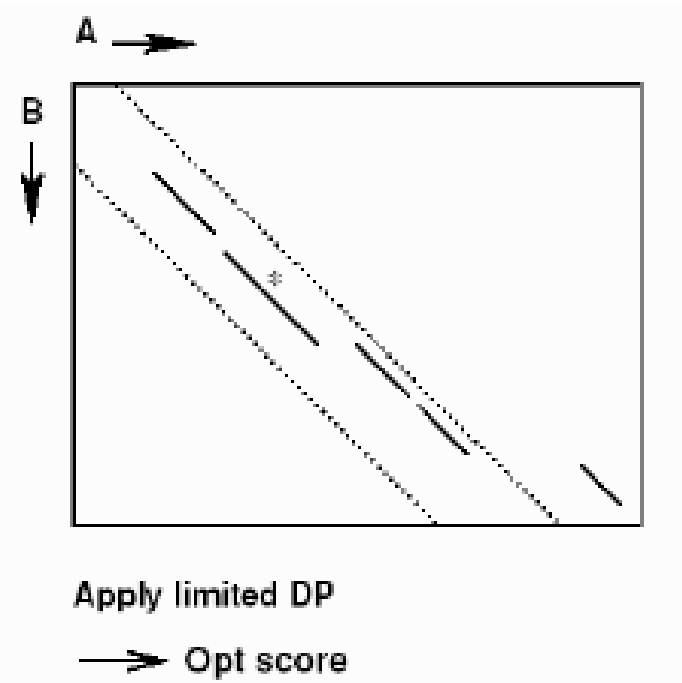
- Junta todas las k-tuplas que estan en la misma diagonal, no muy alejadas, creando “regiones”.
- Las 10 regiones que puntúen más alto se vuelven a evaluar conuna matriz de sustitución.
- La puntuación del valor más alto se guarda cómo valor de inicio: **init1 score**.

El algoritmo de FastA (3)



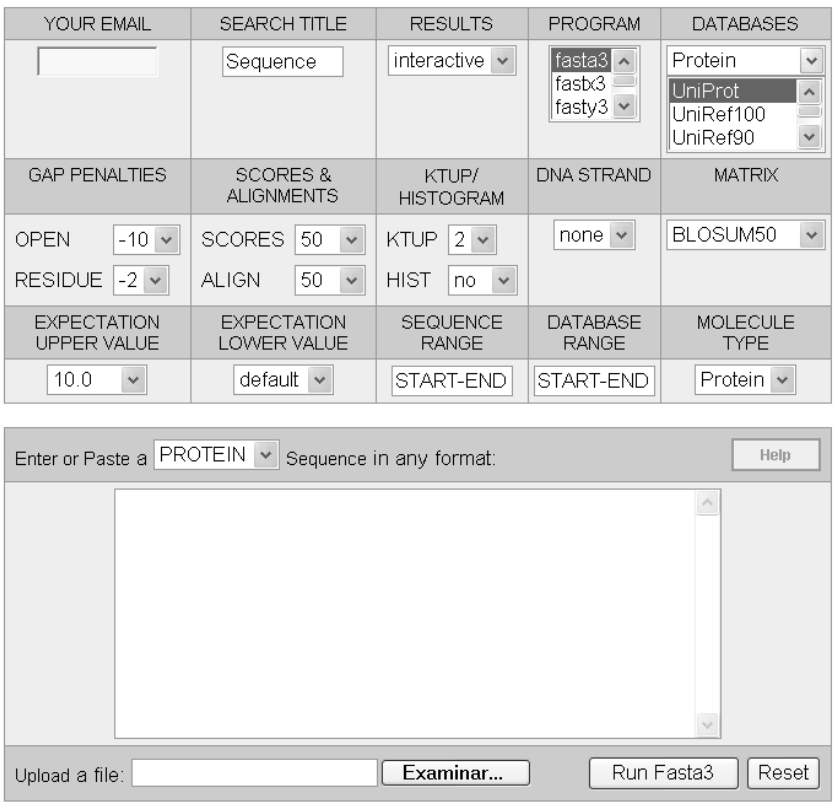
- Se determina si las regiones iniciales de las diversas diagonales pueden unirse para formar un alineamiento aproximado con *gaps.*
- Sólo se unen fragmentos que no se superpongan.
- La puntuación de las regiones agrupadas es la suma de los puntos menos una penalización de union por cada gap.

El algoritmo de FastA (4)



- Tras obtener la región que mejor puntúa se realiza un alineamiento local con una variante de Smith y Waterman, restringido a una banda alrededor de la región.
- La puntuación de este alineamiento es el **opt score**

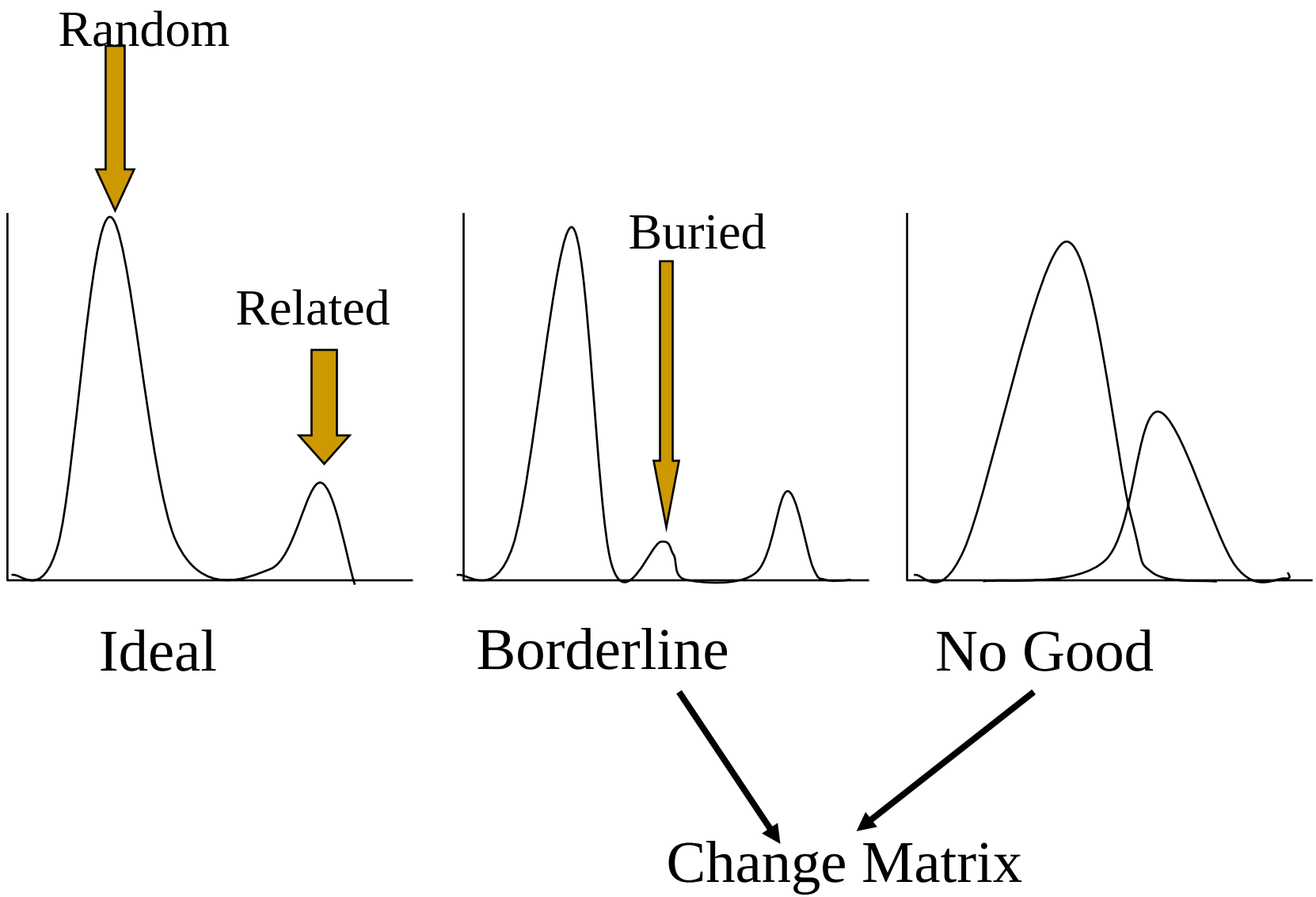
Parámetros de entrada



- Los habituales para un alineamiento:
  - Tipo de secuencia /B. Datos
  - Matriz de puntuación,
  - Coste de apertura /extensión de gaps
- Y además
  - Tamaño de palabra (identidad mínima requerida)
  - E() valores para limitar la búsqueda
  - Histograma de puntuaciones

El resultado de la búsqueda con FastA

- El resultado de la búsqueda és una lista de “hits” con algunas informaciones para c.u.
  - Nombre descripción, longitud
  - Puntuaciones obtenidad en cada etapa
  - Puntuación normalizada por la longitud (z-score).
  - Porcentaje de identidad y longitud de “overlap”
  - Valor esperado E(): Cuantos hits esperaríamos encontrar por azar con puntuación mayor…
  - Histograma de las puntuaciones y distribucion aleatoria de puntuaciones.


Random
Related
Buried
Ideal
Borderline
No Good
Change Matrix

Ejemplo

- En este enlace puedes encontrar el tutorial y el ejemplo de uso de FastA en 2can

BLAST

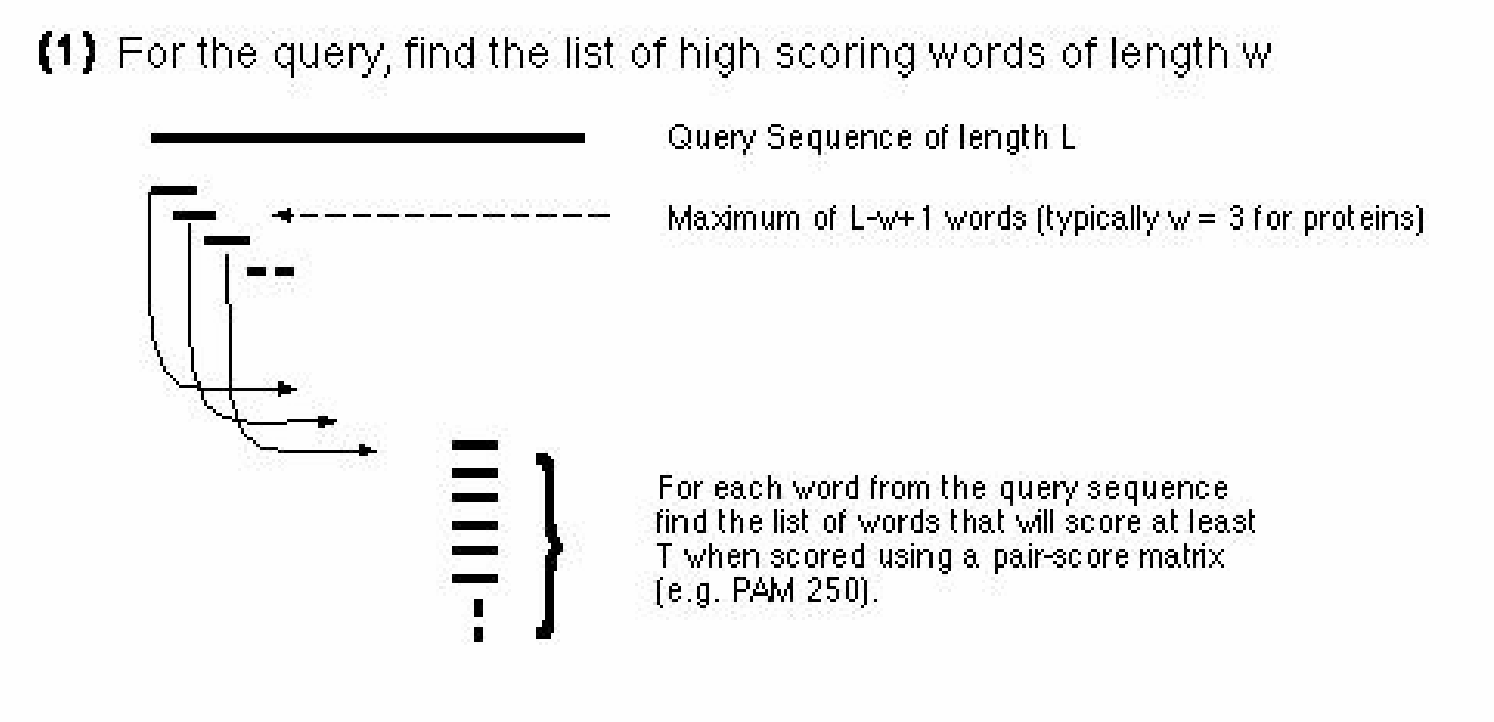
Introducción a BLAST

- Desarrollado a partir de 1990 por Alschultz …
- Modelo matemático solido pero hasta versiones recientes no incorpora los gaps
- Se inicia buscando similaridades altas entre la secuencia y cada una de las de la BD: se asocia con *busqueda por similaridad*.
- Se ha diversificado y existen muchas variantes y extensiones por lo que en muchos casos ha desplazado a FastA

El algoritmo de BLAST

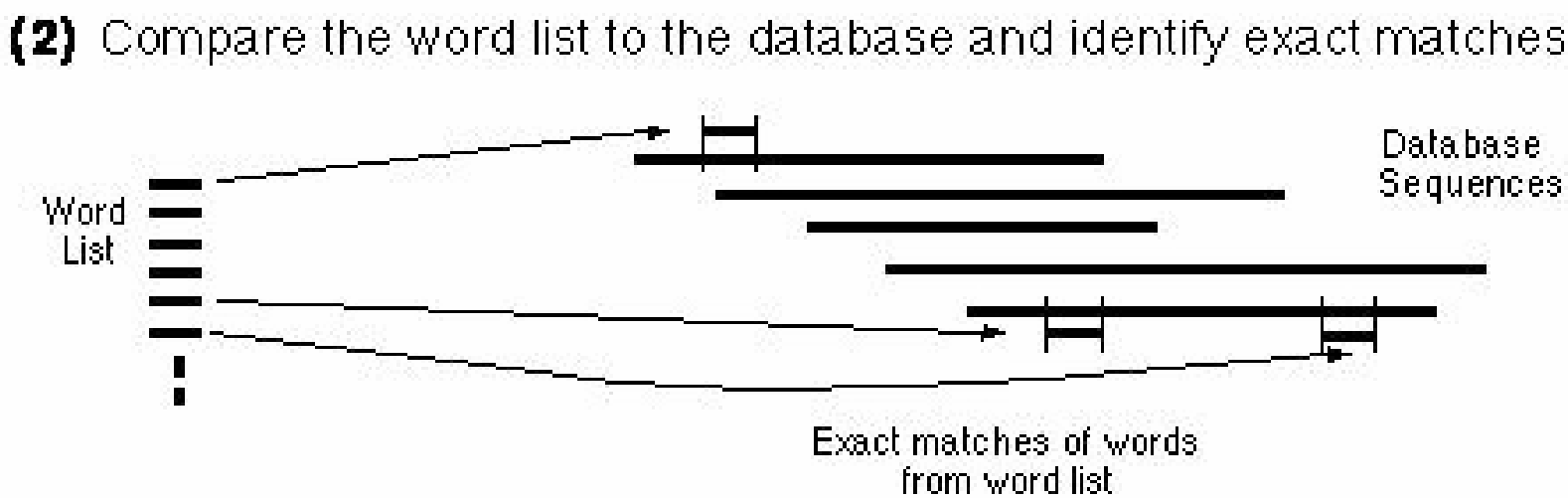
- Procede en tres fases
  - Empieza compilando una lista preliminar de alineamientos cuyo valor supere un umbral minimo HSP: *High-scoring Segment Pairs,*
  - El algoritmo escanea la BD en busca de fragmentos coincidentes con los HSP formando cortos alineamientos con las secuencias de la BD.
  - Los alineamientos se extienden hasta
    - sobrepasar un umbral ➔ coincidencia
    - No superar el umbral ➔ se descarta

BLAST (1)



- Compilar todas las palabras de tamaño *n* que generan una puntuación superior al umbral (HSP)

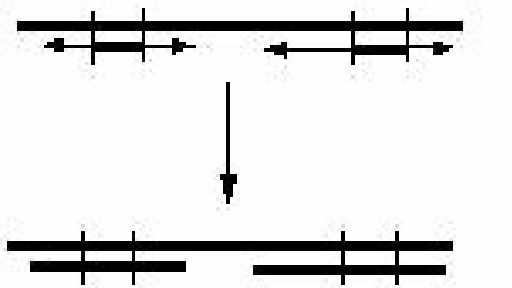
BLAST (2)



- Comparar estas palabras con las secuencias en la BD pera identificar las identidades exactas (“hits”)

BLAST (3)

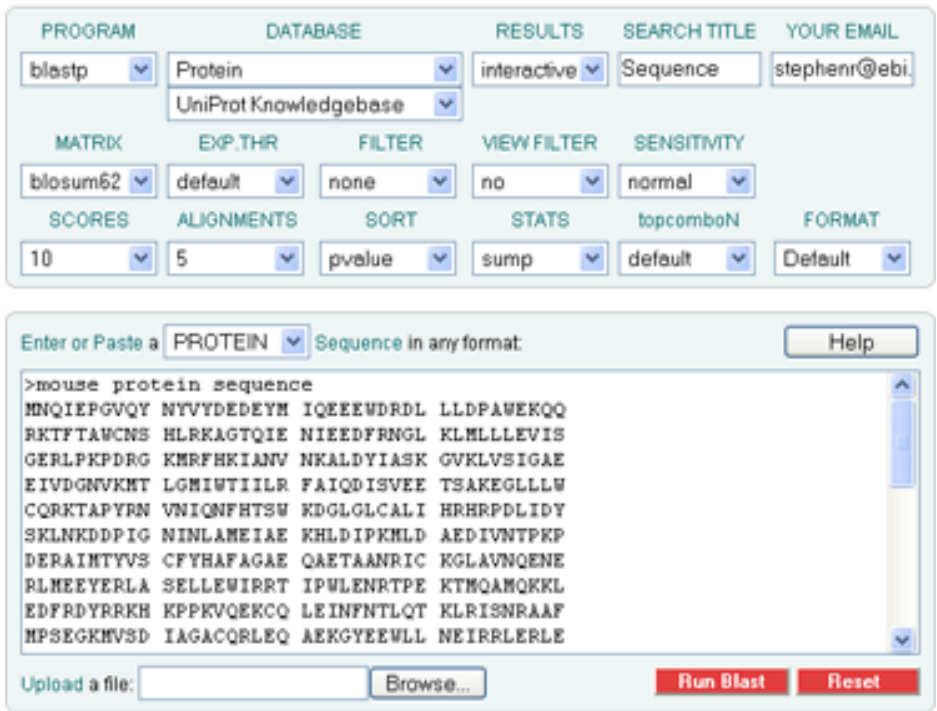
(3) For each word match, extend the alignment in both directions to find alignments that score greater than a threshold of value S



Maximal Segment Pairs (MSPs)

- Extender las palabras que han superado el umbral en ambas direcciones intentando mejorar la puntuacion
- La extensión concluye si la puntuacion desciende por debajo de otro umbral, si llega a cero, o si se acaba la secuencia.

Parámetros de entrada



- Los habituales para un alineamiento:
  - Tipo de secuencia /B. Datos
  - Matriz de puntuación,
  - Coste de apertura /extensión de gaps
- Y además
  - Tamaño de palabra (identidad mínima requerida).
  - E() valores para limitar la búsqueda
  - Sensibilidad deseada (funcion del tamaño de palabra, *w*, y el umbral *T*)

Blast output (1)

- Parameters used in the search
- The list of hits
- Database accession codes, name, description, general information about the hit
- Score in bits, the alignment score expressed in units of information. Usually 30 bits are required for significance
- Expectation value E(), how many hits we expect to find by chance with this score, when comparing this query to the database.

Blast output (2)

- The information for each hit
- A header including hit name, description, length
- Composite expectation value
- Score and expectation value
  - how many identical residues
  - how many residues contributing positively to the score
- The local alignment itself

Sequences producing significant alignments:	Score (bits)	E Value
gi|131557|sp|P25044|PTP1_YEAST Protein-tyrosine phosphatase...	705	0.0
gi|417567|sp|P32586|PYP2_SCHPO Protein-tyrosine phosphatase...	148	2e-35
gi|417568|sp|P32587|PYP3_SCHPO Protein-tyrosine phosphatase...	142	1e-33
gi|33112421|sp|Q13332|PTNS_HUMAN Receptor-type protein-tyro...	137	3e-32
gi|1709906|sp|P23468|PTPD_HUMAN Protein-tyrosine phosphatas...	135	2e-31
gi|462551|sp|Q05909|PTPG_MOUSE Protein-tyrosine phosphatase...	134	3e-31
gi|3183128|sp|Q62656|PTPZ_RAT Receptor-type protein-tyrosin...	133	7e-31
gi|20455509|sp|P35992|PTP1_DROME Protein-tyrosine phosphata...	132	8e-31
gi|462550|sp|P23470|PTPG_HUMAN Protein-tyrosine phosphatase...	132	9e-31
gi|125978|sp|P10586|PTPF_HUMAN LAR protein precursor (Leuko...	132	1e-30
gi|266860|sp|P29461|PTP2_YEAST Protein-tyrosine phosphatase...	85	3e-16
gi|731478|sp|P40048|PTP3_YEAST Protein-tyrosine phosphatase...	49	1e-05
gi|2499759|sp|P80994|VH01_RACVI Dual specificity protein ph...	36	0.13
gi|138374|sp|P07239|DUSP_VACCV Dual specificity protein pho...	35	0.18
gi|138373|sp|P20495|DUSP_VACCC Dual specificity protein pho...	35	0.19
gi|418237|sp|P33064|DUSP_VARV Dual specificity protein phos...	35	0.33
gi|1168807|sp|Q00684|CC14_YEAST Probable protein-tyrosine p...	33	1.2

Score ↓ E-Value ↑

Significación de los resultados

E-values, p-values y bit-scores

- Dado que los programas de búsqueda heurística tan sólo encuentran coincidencias aproximadas conviene poder cuantificar cuan aproximadas son
- Esto se hace mediante distintos estadísticos
  - E-value
  - P-value
  - Bit-scores

E-values

- Dada una secuencia que ha obtenido una puntuacion E-value es el *número esperado de puntuaciones iguales o superiores a las de dicha secuencia atribuibles al azar.*
- Un E-value de 10 para una coincidencia significa, que, en una base de datos de secuencias aleatorias del mismo tamaño en la que se ha realizado la búsqueda, se podría esperar encontrar hasta 10 coincidencias con la misma puntuación o similar.
- El E- value es la medida de corte más utilizada en las búsquedas en bases de datos. Sólo se informa de las coincidencias que superan un nivel mínimo
- El E-value oscila entre 0 y cualquier valor

P-values

- Refleja la probabilidad de obtener por azar una puntuación superior o igual a la observada
- Se relaciona con el E-value en que: $P=1-e^{-E}$
- Un P-valor de 0.03 significa que hay una probabilidad (>=) 3% de encontrar una puntuación superior a la observada simplemente por azar
- Si E<0,01 Los P-valores y los E-valores son similares
- Los p-valores oscilan entre 0 y 1

Bit scores

- El valor de la puntuaciones obtenidas por un emparejamiento carecen de sentido si no se tiene en cuenta el tamaño de la base de datos y el sistema de puntuación
- Los *Bit-scores* normalizan las puntuaciones para independizarlas de ambos factores de forma que podamos compararlas

Los parámetros importan

- En las transparencias siguientes se muestra un ejemplo extraído del libro de J. Pevsner *Bioinformatics and Functional Genomics* que muestra cómo cambian los resultados de BLAST al realizar una búsqueda con distintos valores de los parámetros.
- La secuencia buscada es la NP_006735 “retinol binding protein” (RBP)

Changing E, T & matrix for blastp nr RBP

Expect	10 (T=11)	1 (T=11)	10,000 (T=11)	10 (T=5)	10 (T=11)	10 (T=16)	10 (BL45)	10 (PAM70)
#hits to db	129m							
#sequences	1,043,455							
#extensions	5.2m							
#successful extensions	8,367							
#sequences better than E	142							
#HSPs>E (no gapping)	53							
#HSPs gapped	145							
X1, X2, X3	16 (7.4 bits) 38 (14.6 bits) 64 (24.7 bits)							

Fuente: ***Bioinformatics and Functional Genomics*****. J. Pevsner**

Changing E, T & matrix for blastp nr RBP

Expect	10 (T=11)	1 (T=11)	10,000 (T=11)	10 (T=5)	10 (T=11)	10 (T=16)	10 (BL45)	10 (PAM70)
#hits to db	129m	129m	129m					
#sequences	1,043,455	1.0m	1.0m					
#extensions	5.2m	5.2m	5.2m					
#successful extensions	8,367	8,367	8,367					
#sequences better than E	142	86	6,439					
#HSPs>E (no gapping)	53	46	6,099					
#HSPs gapped	145	88	6,609					
X1, X2, X3	16 (7.4 bits) 38 (14.6 bits) 64 (24.7 bits)	16 38 64	16 38 64					

Fuente: *Bioinformatics and Functional Genomics*. J. Pevsner

Changing E, T & matrix for blastp nr RBP

Expect	10 (T=11)	1 (T=11)	10,000 (T=11)	10 (T=5)	10 (T=11)	10 (T=16)	10 (BL45)	10 (PAM70)
#hits to db				112m	112m	112m		
#sequences				907,000	907,000	907,000		
#extensions				508m	4.5m	73,788		
#successful extensions				11,484	7,288	1,147		
#sequences better than E				125	124	88		
#HSPs>E (no gapping)				48	48	48		
#HSPs gapped				127	126	90		
X1, X2, X3								

Fuente: ***Bioinformatics and Functional Genomics*****. J. Pevsner**

Changing E, T & matrix for blastp nr RBP

Expect	10 (T=11)	1 (T=11)	10,000 (T=11)	10 (T=5)	10 (T=11)	10 (T=16)	10 (BL45)	10 (PAM70)
#hits to db	129m						386m	129m
#sequences	1,043,455						1.0m	1.0m
#extensions	5.2m						30.2m	19.5m
#successful extensions	8,367						9,088	13,873
#sequences better than E	142						110	82
#HSPs>E (no gapping)	53						60	66
#HSPs gapped	145						113	99
X1, X2, X3	16 (7.4 bits) 38 (14.6 bits) 64 (24.7 bits)						22 51 85	15 35 59

Fuente: ***Bioinformatics and Functional Genomics*****. J. Pevsner**

Changing E, T & matrix for blastp nr RBP

Expect	10 (T=11)	1 (T=11)	10,000 (T=11)	10 (T=5)	10 (T=11)	10 (T=16)	10 (BL45)	10 (PAM70)
#hits to db	129m	129m	129m	112m	112m	112m	386m	129m
#sequences	1,043,455	1.0m	1.0m	907,000	907,000	907,000	1.0m	1.0m
#extensions	5.2m	5.2m	5.2m	508m	4.5m	73,788	30.2m	19.5m
#successful extensions	8,367	8,367	8,367	11,484	7,288	1,147	9,088	13,873
#sequences better than E	142	86	6,439	125	124	88	110	82
#HSPs>E (no gapping)	53	46	6,099	48	48	48	60	66
#HSPs gapped	145	88	6,609	127	126	90	113	99
X1, X2, X3	16 (7.4 bits) 38 (14.6 bits) 64 (24.7 bits)	16 38 64	16 38 64				22 51 85	15 35 59

Fuente: ***Bioinformatics and Functional Genomics*****. J. Pevsner**