BLAST

Rodrigo Santamaría

BLAST

Introducción

Definición

Familia BLAST

Algoritmo

Salida

Estrategias

Otros programas

Introducción

- BLAST: Basic Local Alignment Search Tool
  - Altschul et al. 1990 (PMID 2231712)
  - Altschul et al. 1997 (PMID 9254694)
- Es el software más importante en bioinformática
  - Importancia de los estudios de similitud de secuencias
  - Rápido incluso con BBDD muy grandes
  - Fiable a nivel informático y estadístico
  - Flexible, con multitud de parámetros ajustables

Introducción

- BLAST permite seleccionar una secuencia (*query*) y realizar alineamientos de pares de secuencias con todas las secuencias de base de datos entera (*target*)
  - Realiza millones de alineamientos
  - Y devuelve los más relacionados con la query

Introducción

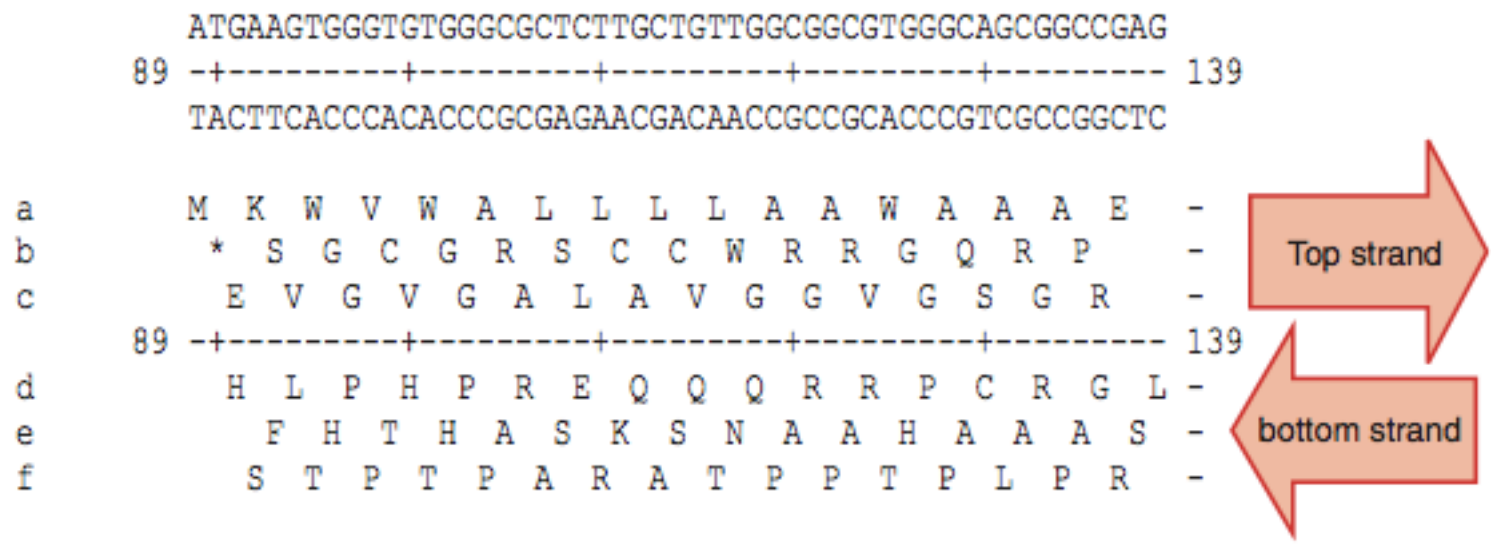
- Needleman-Wunsch (1970) hace alineamientos globales, cuando normalmente estamos interesados en locales
- Smith-Waterman (1981) hace alineamientos locales óptimos, pero no es útil en búsquedas de bases de datos porque es muy intensivo computacionalmente
- BLAST (1990) hace alineamientos locales subóptimos, pero suficientemente sensibles y muy rápidos.
  - Además es accesible online (http://www.ncbi.nlm.nih.gov/BLAST)

Introducción
Familia BLAST

Programa	Query	Target (DB)	Usos
blastn	Nucleótido	Nucleótido	Por ejemplo para comparar secuencias entre especies o detectar elementos repetitivos
blastp	Proteína	Proteína	Por ejemplo identificar regiones comunes entre proteínas, identificar proteínas comunes para estudios filogenéticos
blastx	Nucleótido traducido a proteína	Proteína	Determinar si una secuencia de ADN corresponde a una proteína conocida. Blastx convierte la secuencia de ADN a las 6 posibles proteínas y las compara con las proteínas de una DB
tblastn	Proteína	Nucleótido traducido a proteína	Por ejemplo comprobar si una determinada proteína aparece en ADN genómico de otras especies
tblastx	Nucleótido traducido a proteína	Nucleótido traducido a proteína	Comparar posibles proteínas de una cadena de ADN con posibles proteínas de una DB de ADN. Útil para encontrar coincidencias no dadas por métodos tradicionales o que no están aún en las bases de datos de proteínas. Computacionalmente alto

Introducción
Familia BLAST

- "Nucleótido traducido a proteína"
  - Una cadena de ADN da lugar a 6 cadenas de amino ácidos
    - 2 sentidos (*strands*) de ADN
    - 3 posibles marcos de lectura o *reading frames* (+0,+1,+2)



Program	Query	Number of database searches	Database
blastp	protein	1	protein

Use blastp to compare a protein query to a database of proteins.

Program	Query	Number of database searches	Database
blastn	DNA	1	DNA

Use blastn to compare both strands of a DNA query against a DNA database.

Program	Query	Number of database searches	Database
blastx	DNA	6	protein

Blastx translates a DNA sequence into six protein sequences using all six possible reading frames, and then compares each of these proteins to a protein database.

Program	Query	Number of database searches	Database
tblastn	protein	6	DNA

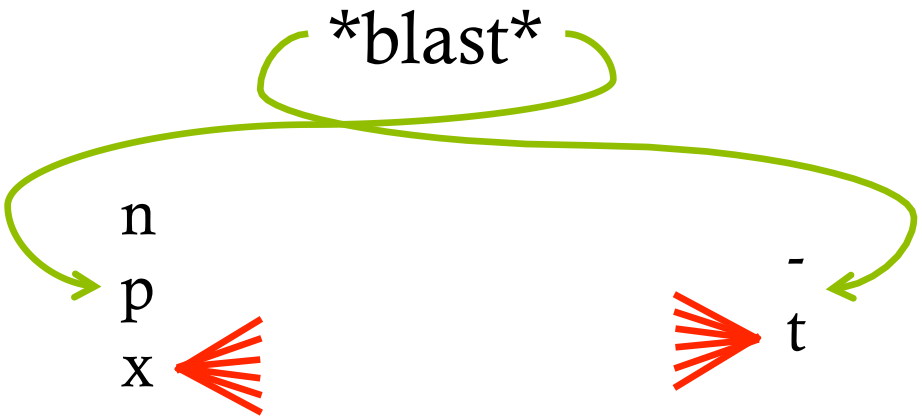
Tblastn is used to translate every DNA sequence in a database into six potential proteins, and then to compare your protein query against each of those translated proteins.

Program	Query	Number of database searches	Database
tblastx	DNA	36	DNA

Tblastx is the most computational intensive BLAST algorithm. It translates DNA from both a query and a database into six potential proteins, then performs 36 protein-protein database searches.

Familia BLAST

- Reglas nemotécnicas
  - n, p, x se refieren a la query
  - t se refiere al target (translated)
  - x/t indican que el query/target está traducido



Selección de Bases de Datos (NCBI BLAST)

- Proteínas (para blastp, blastx)
  - GenBank (RefSeq), SwissProt, PIR, PRF
  - BD no redundante (nr): Combina todas las anteriores, eliminando duplicados
- ADN (para blastn, tblastn, tblastx)
  - Genoma humano/ratón + transcritos
  - BD no redundante (nr): nucleótidos combinados (sin duplicados) de GenBank, EMBL, DDBJ y PDB
  - Otras BBDD de secuencias particulares

BLAST

Introducción

Algoritmo

- Fases
- Evaluación Estadística

Salida

Opciones

Protocolos

Otros programas

Algoritmo BLAST

1. **List**: se compila una lista preliminar de alineamientos posibles (*palabras*), según la secuencia de la query
2. **Scan**: se busca en la base de datos por secuencias que coinciden con las palabras, según un umbral T
3. **Extend**: se extienden los pares de palabras para encontrar aquéllos que superen un umbral S, reportándose como coincidencias.

Fase 1 - listado

- Se divide la secuencia en palabras, y para cada palabra w se lista el conjunto de palabras S_w (a veces llamadas *semillas*) con un nivel de coincidencia por encima de un umbral T
- Para proteínas
  - Las palabras tienen un tamaño por defecto de 3 (20^3=8000 palabras)
  - Para cada palabra, se identifican las palabras que se parecen a ella por encima de un umbral T, usando como puntuación una de las matrices vistas (PAM, BLOSUM)
- Para genes
  - Las palabras tienen un tamaño por defecto de 11 ($4^{11} \sim 10^6$)
  - Se identifican palabras que coincidan exactamente ($S_w = w$)
    - No hay umbral T

Phase 1: compile a list of words (w = 3) above threshold T

- Query sequence: human beta globin NP_000509 (includes ... VTALWGKVNVD...)
- words derived from query sequence (HBB):

VTA TAL ALW LWG WGK GKV KVN VNV NVD

- generate a list of words matching query (both above and below T). Consider LWG in the query and the scores (derived from a BLOSUM62 matrix) for various words:

examples of words above threshold 11

LWG	4+11+6=21
IWG	2+11+6=19
MWG	2+11+6=19
VWG	1+11+6=18
FWG	0+11+6=17
AWG	0+11+6=17
LWS	4+11+0=15
LWN	4+11+0=15
LWA	4+11+0=15
LYG	4+ 2+6=12

threshold

examples of words below threshold 11

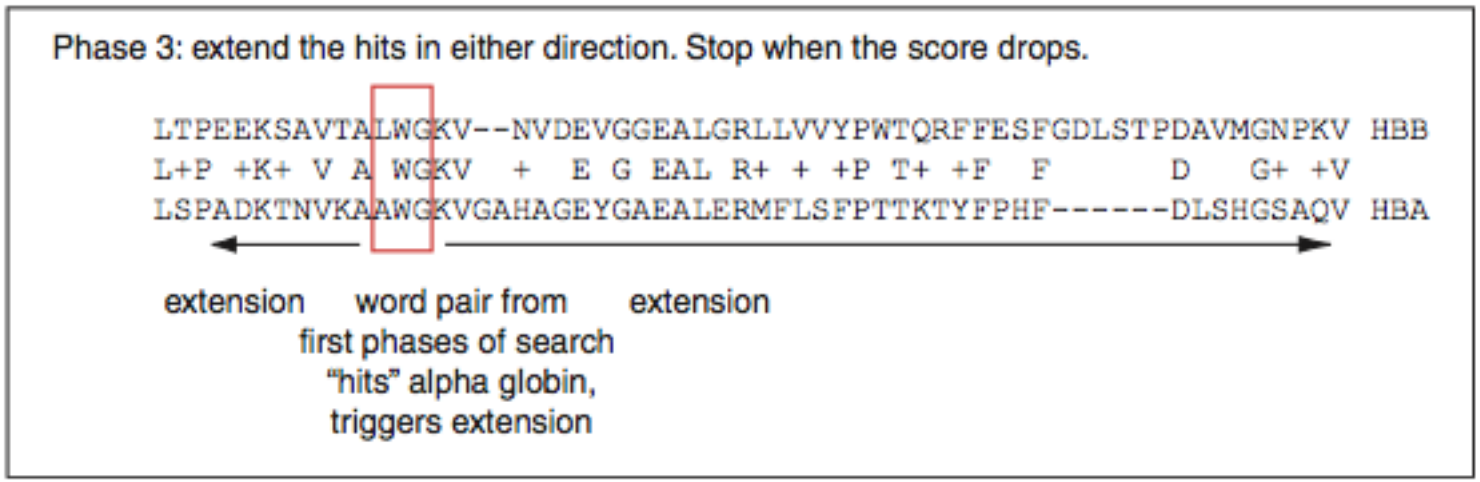
LFG	4+ 1+6=11
FWS	0+11+0=11
AWS	-1+11+0=10
CWS	-1+11+0=10
IWC	2+11-3=10

Fase 2 - búsqueda

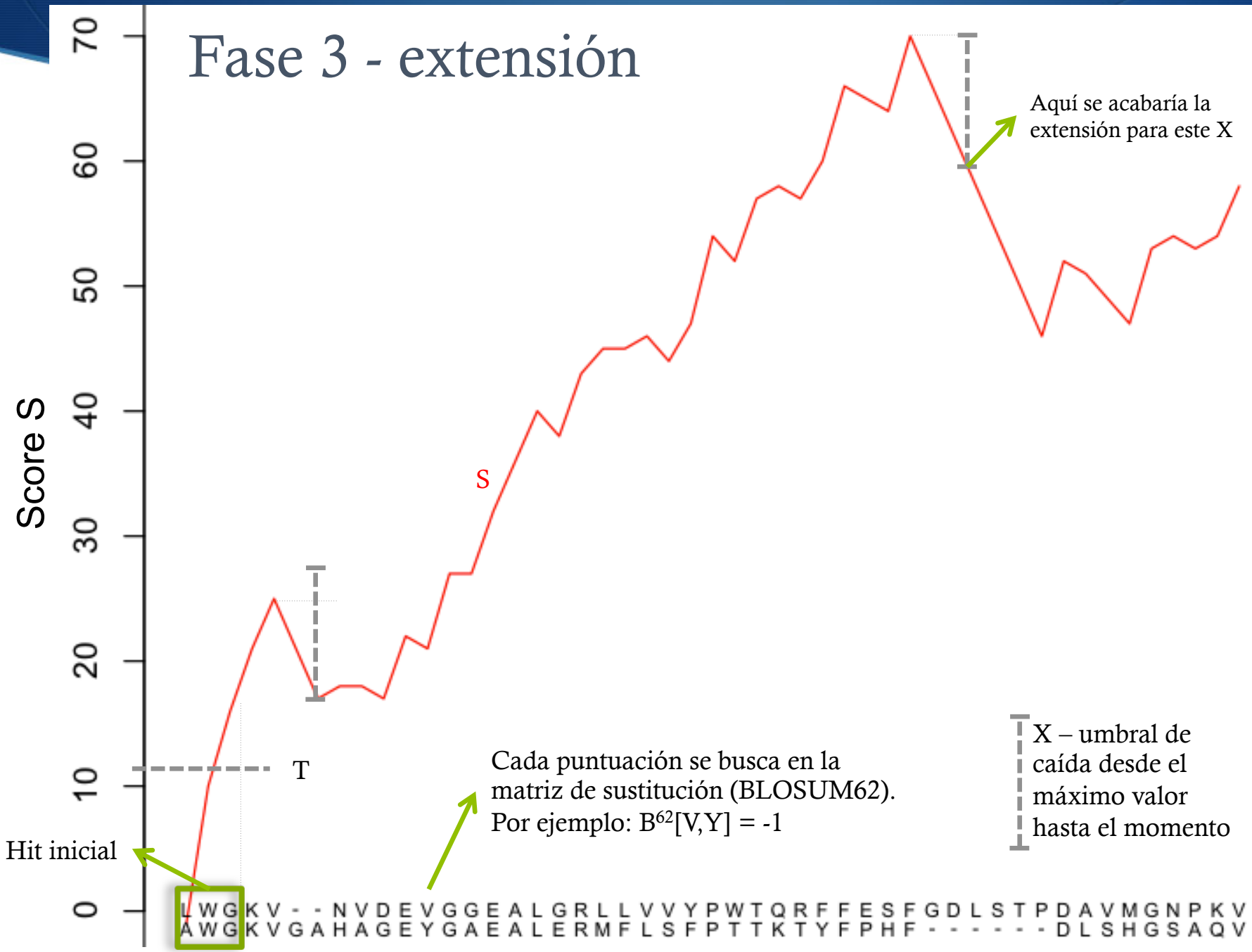
- Para cada palabra w se escanea la BD en busca de registros que coincidan con alguna de las palabras en S_w (*hits*)
  - En el caso de genes, que coincidan exactamente con w
- (Altschul et al., 1997) acelera el algoritmo buscando coincidencias con dos palabras w_1 y w_2 que se encuentren a una distancia $<A$
  - Conocido como el método de dos hits (*two-hit method*).
  - Encuentra 3 veces más coincidencias, pero extiende sólo 1/7 de ellas (es decir, acelera la fase 3)

Fase 3 - extensión

- Para cada coincidencia, se extiende la palabra en ambas direcciones, hasta que el valor de coincidencia baja por debajo de un umbral X
  - De nuevo, se usa para calcular el valor de coincidencia una de las matrices vistas (PAM, BLOSUM)



Fase 3 - extensión

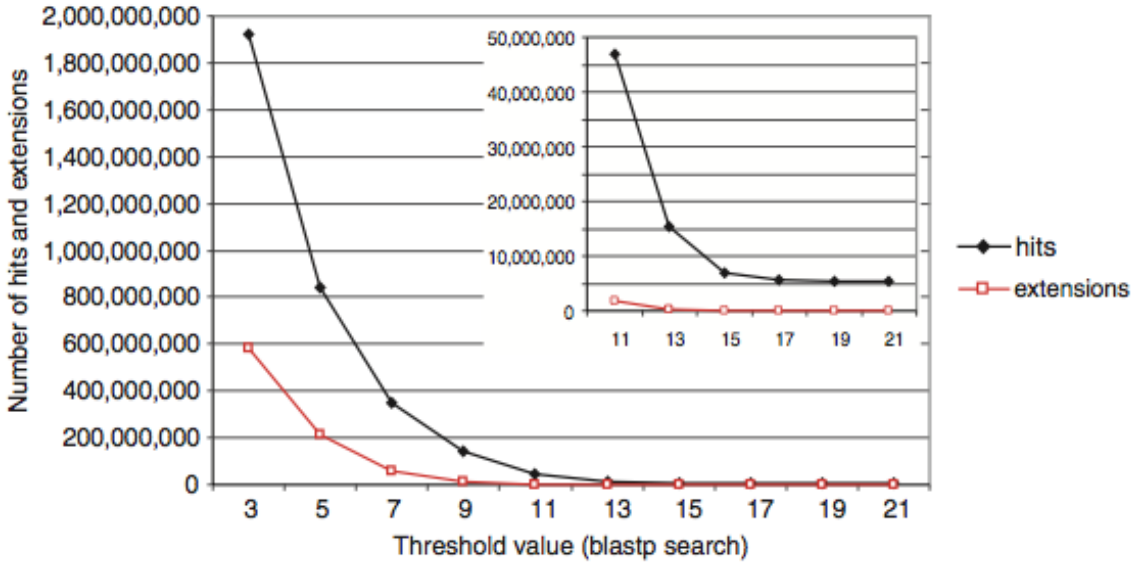


Fase 3 - extensión

- La versión actual BLAST de NCBI utiliza tres umbrales *X*, correspondientes a tres fases de extensión:
  - Primera fase: la primera extensión termina cuando se llega a una caída>*X1* o a un hueco (gap)
  - Segunda fase: la segunda extensión termina cuando hay una caída>*X2* (contando huecos)
  - Tercera fase: la extensión termina cuando hay una caída>*X3* (contando huecos)

Valores y umbrales

- *T*: determina las palabras que se consideran coincidencias inicialmente
- *X*: determina hasta dónde se considera coincidencia al extender las palabras
- Matrices de puntuación
- *S:* valor de puntuación, depende de las secuencias a comparar y de *X*, *T* y la matriz de puntuación



Parámetros

- Umbral T ↑ velocidad ↑ sensibilidad ↓
- Tamaño de palabra ↑ velocidad ↑ sensibilidad ↓
- Matriz de puntuación ~ características evolutivas
- La elección adecuada de estos tres parámetros es clave para modular la sensibilidad y velocidad de BLAST

Resumen

query

...VTALWGKVNVD...

División en palabras de tamaño determinado

VTA **LWG** KVN

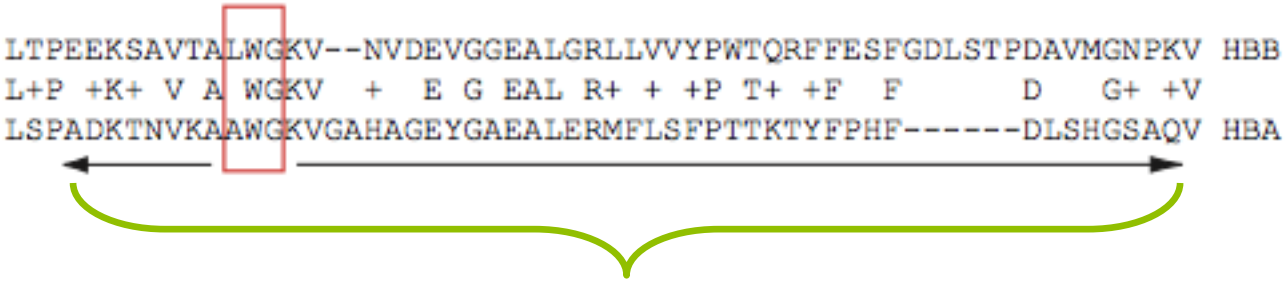
TAL WGK VNV

ALW GKV NVD

Listar palabras que coinciden por encima de T

LWG IWG MWG VWG FWG
AWG LWS LWN LWA LYG

Buscar coincidencias en la BD
Extender hasta que el nivel de coincidencia baje del umbral X

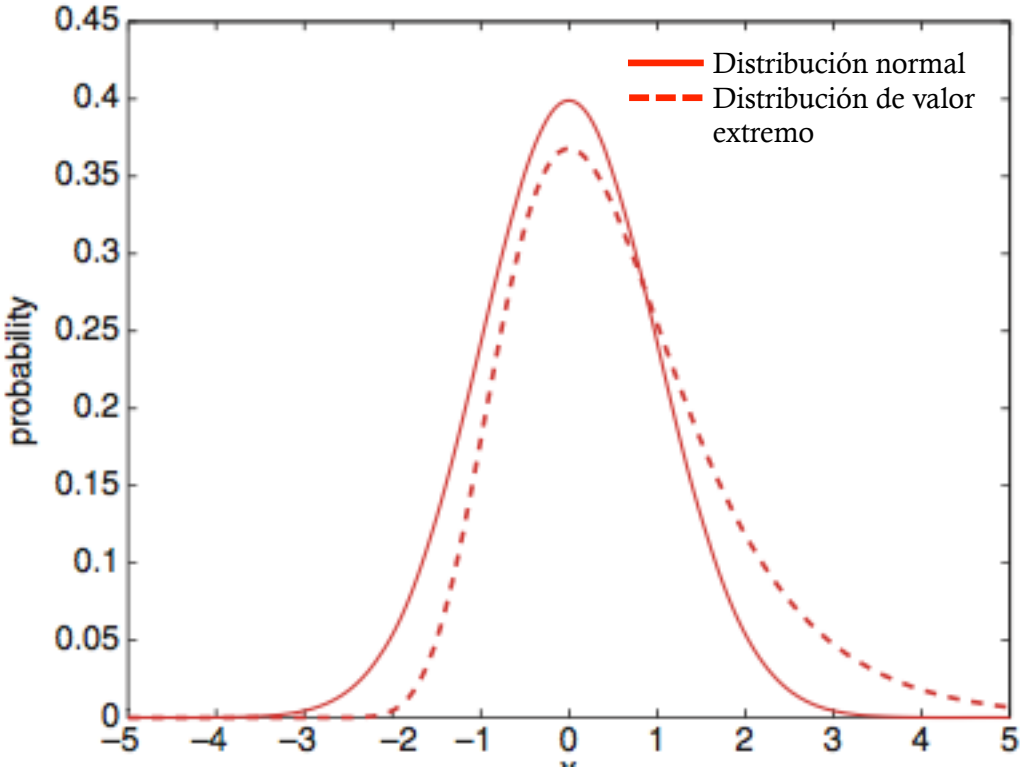


high-scoring segment pair (HSP)

Evaluación estadística

- Queremos calcular una medida cuantitativa de la probabilidad de que los alineamientos encontrados lo sean por azar
- Para ello, usamos una distribución de valores extremos en vez de una distribución normal
  - ¿Por qué?: La caída rápida de la distribución normal a la derecha hace que se sobreestime la significatividad del alineamiento → los alineamientos aleatorios no siguen una distribución normal

Distribución de valor extremo



- A partir de esta distribución, para dos secuencias con longitud *n* y *m*, el número esperado (*E*-valor) de hits con coincidencia >= *S* es
  - $E=Kmne^{-\lambda S}$
    - λ - factor de bajada
    - *K* - factor de escala
- *S* es un *raw score* (no tiene en cuenta la distribución de probabilidad)

Bit scores

- El bit score (*S'*) normaliza el raw score (*S*) en función de los métodos de puntuación y los tamaños de las secuencias:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- El E-valor de *S'* es: $E = mn \times 2^{-S'}$
  - *m* y *n* son los tamaños de la secuencias query y target
- *S'* permite comparar scores de búsquedas sobre distintas DDBB o realizadas con distintas matrices de puntuación

K y λ

- Dependen de
  - La matriz de puntuación
  - La penalización de gaps

Scoring matrix	Gap opening penalty[b]	Gap extension penalty[b]	K	λ
BLOSUM50	∞[a]	0-∞	0.232	0.11
BLOSUM50	15	8–15	0.09	0.222
BLOSUM50	11	8–11	0.05	0.197
BLOSUM50	11	1	—	—
BLOSUM62	∞[a]	0-∞	0.318	0.13
BLOSUM62	12	3–12	0.1	0.305
BLOSUM62	8	7–8	0.06	0.270
BLOSUM62	7	1	—	—
PAM250	∞[a]	0-∞	0.229	0.09
PAM250	15	5–15	0.06	0.215
PAM250	10	8-10	0.031	0.175
PAM250	11	1	—	—

E y p valores

- p-valor: probabilidad de tener un alineamiento por casualidad utilizando score *S* o mayor.
  - *E* y *p* son modos parecidos pero distintos de representar la significatividad de un alineamiento
  - La relación entre *p* y *E* es: $p = 1 - e^{-E}$
  - Para $E \leq 0.05$ → $p \sim E$

E	*p*
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950
0.0001	0.0001000

BLAST

Introducción
Algoritmo
Salida
- Cabecera
- Resúmenes
- Alineamiento
- Pie

Protocolos
Otros programas

Salida BLAST

- La estructura básica de una salida BLAST es
  - **Cabecera**: detalles sobre la consulta (query y target)
  - **Resúmenes de una línea**: alineamientos significativos
  - **Alineamientos**: detalles de cada alineamiento significativo
  - **Pie**: detalles sobre la consulta (resto de parámetros)
- La salida es similar para toda la familia BLAST
  - Veremos la salida de blastp y algunas peculiaridades del alineamiento en blastn y blastx

Cabecera

BLASTP 2.2.25+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference for compositional score matrix adjustment: Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schaffer, and Yi-Kuo Yu (2005) "Protein database searches using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109.

RID: YYHSBSST014

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
14,316,990 sequences; 4,903,270,308 total letters
Query= gi|4504349|ref|**NP_000509.1| hemoglobin subunit beta** [Homo sapiens]

Resúmenes de una línea

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Bit score

Accession	Description	Max score	Total score	Query coverage	E value	Links
AAX37051.1	hemoglobin beta [synthetic construct]	301	301	100%	2e-80	
AAX29557.1	hemoglobin beta [synthetic construct]	301	301	100%	2e-80	
NP_000509.1	hemoglobin subunit beta [Homo sapiens] >gi|55635219|ref	301	301	100%	3e-80	U G M
P02024.2	RecName: Full=Hemoglobin subunit beta; AltName: Full=B	300	300	100%	7e-80	
ACU56984.1	beta-globin [Homo sapiens]	299	299	100%	7e-80	
AAN84548.1	beta globin chain variant [Homo sapiens]	299	299	100%	7e-80	G
AAZ39780.1	beta globin [Homo sapiens] >gb|AAZ39781.1| beta globin	299	299	100%	7e-80	G
AAD19696.1	hemoglobin beta chain [Homo sapiens]	299	299	100%	8e-80	G M
1COH_B	Chain B, Structure Of Haemoglobin In The Deoxy Quaterna	298	298	99%	1e-79	S
AAF00489.1	hemoglobin beta subunit variant [Homo sapiens] >gb|AAA8	298	298	100%	1e-79	G M
2YRS_B	Chain B, Human Hemoglobin D Los Angeles: Crystal Struct	298	298	99%	2e-79	S
1DXU_B	Chain B, High-Resolution X-Ray Study Of Deoxy Recombina	297	297	99%	3e-79	S
1HDB_B	Chain B, Analysis Of The Crystal Structure, Molecular Mode	297	297	99%	3e-79	S
1DXV_B	Chain B, High-Resolution X-Ray Study Of Deoxy Recombina	297	297	98%	3e-79	S
AAL68978.1	mutant beta-globin [Homo sapiens]	297	297	100%	3e-79	G
2DXM_D	Chain D, Neutron Structure Analysis Of Deoxy Human Hem	297	297	98%	3e-79	S
1NQP_B	Chain B, Crystal Structure Of Human Hemoglobin E At 1.73	297	297	99%	3e-79	S
1K1K_B	Chain B, Structure Of Mutant Human Carbonmonoxyhemog	297	297	99%	3e-79	S
AAN11320.1	hemoglobin beta chain variant Hb S-Wake [Homo sapiens]	296	296	100%	5e-79	G M
XP_002822173.1	PREDICTED: hemoglobin subunit beta-like [Pongo abelii]	296	296	100%	5e-79	G M
1Y85_B	Chain B, T-To-T(High) Quaternary Transitions In Human He	296	296	98%	5e-79	S
1YE0_B	Chain B, T-To-T(High) Quaternary Transitions In Human He	296	296	99%	6e-79	S
1O1O_B	Chain B, Deoxy Hemoglobin (A,C:v1m,V62l; B,D:v1m,V67l	296	296	99%	6e-79	S

Alineamientos

```
>sp|P09968.2|GLB3_PETMA  RecName: Full=Globin-3; AltName: Full=Globin III
Length=150

 Score = 58.2 bits (139),  Expect = 6e-14, Method: Compositional matrix adjust.
 Identities = 31/118 (26%), Positives = 57/118 (48%), Gaps = 5/118 (4%)

Query  4    LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+  EK+ + + W  V  N +  G + L +     P  Q FF  F  L+T D +  +  V
Sbjct  12   LSAAEKTKIRSAWAPVYSNYETTGVDILVKFFTSTPAAQEFFPKFKGLTTADQLKKSADV  71

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVIA  116
            + H ++++ A +D +  +D+ +      L +L   H     VDP+ F++L  V+   +A
Sbjct  72   RWHAERIINAVNDAVVSMDDTEKMSMKLGDLSGKHAKSFQVDPQYFKVLAAVIADTVA  129
```

No coincidencia

coincidencia

No coincidencia exacta, pero score positivo

gap

Alineamientos - blastn

```
>ref|XM_001267620.1| Neosartorya fischeri NRRL 181 polyketide synthase, putative (NFIA_045430)
mRNA, complete cds
Length=5334

 Score = 37.4 bits (40),  Expect = 2.1
 Identities = 22/23 (95%), Gaps = 0/23 (0%)
 Strand=Plus/Minus

Query  137   CCAGGGTCGCCCCGGCAACCACG  159
             ||||| ||||||||||||||||
Sbjct  3040  CCAGGCTCGCCCCGGCAACCACG  3018
```

inverso complementario

```
>ref|XM_001092331.1| PREDICTED: Macaca mulatta similar to Trafficking protein particle
complex protein 2 (Sedlin) (MBP-1-interacting protein 2A)
(MIP-2A) (TRAPPC2), mRNA
Length=1454

 Score = 37.4 bits (40),  Expect = 2.1
 Identities = 20/20 (100%), Gaps = 0/20 (0%)
 Strand=Plus/Plus

Query  182  GACTGCACCAGAGCCATGGT  201
            ||||||||||||||||||||
Sbjct  52   GACTGCACCAGAGCCATGGT  71
```

Alineamientos - blastx

Marco de lectura

```
 Score = 69.7 bits (169),  Expect(2) = 3e-33
 Identities = 45/83 (54%), Positives = 45/83 (54%), Gaps = 38/83 (45%)
 Frame = +1

Query  166  GHLSPDIVAEQKKLEAADLVIFQ----------------------------------  234
            GHLSPDIVAEQKKLEAADLVIFQ
Sbjct  77   GHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPF  136

Query  235  -SKKAVLSITTGGSGSMYSLQGI  300
             SKKAVLSITTGGSGSMYSLQGI
Sbjct  137  RSKKAVLSITTGGSGSMYSLQGI  159
```

Pie

```
 Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
excluding environmental samples from WGS projects
    Posted date:  Jun 7, 2011  4:38 PM
  Number of letters in database: 608,303,012
  Number of sequences in database:  14,316,990

Lambda     K      H
   0.320    0.137    0.422
Gapped
Lambda     K      H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 14316990
Number of Hits to DB: 150690867
Number of extensions: 5974360
Number of successful extensions: 11879
Number of sequences better than 100: 127
Number of HSP's better than 100 without gapping: 0
Number of HSP's gapped: 11811
Number of HSP's successfully gapped: 127
Length of query: 147
Length of database: 4903270308
Length adjustment: 110
Effective length of query: 37
Effective length of database: 3328401408
Effective search space: 123150852096
Effective search space used: 123150852096
T: 11
A: 40
X1: 16 (7.4 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
S2: 67 (30.4 bits)
```

K y λ para el cálculo de E-valores

Matriz de puntuación

Penalización para gaps en hits iniciales y en extensiones

Umbral para hits iniciales

Distancia entre los 2 hits

Umbrales de extensión

Umbrales finales: se desechan alineamientos con S menor que estos umbrales

Más información sobre los resultados de BLAST en http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastn.html

BLAST

Introducción
Algoritmo
Salida
Estrategias
Consideraciones generales
Significatividad estadística
Modificación del nº de resultados
Protocolos
Otros programas

Consideraciones generales

- BLAST es una gran herramienta para explorar BBDD de secuencias
- Para obtener los mejores resultados posibles es esencial:
  - Definir la cuestión que se quiere responder
  - Evaluar cuál será la secuencia de entrada
  - Evaluar cuál será la BD de búsqueda
  - Evaluar cuál será el algoritmo a utilizar
  - Evaluar cuáles serán los parámetros del algoritmo
- Debemos considerar estos puntos *a priori*, antes de conocer ningún resultado
  - De lo contrario, caemos en una espiral de *ensayo y error* sin criterio.

Consideraciones generales

- **Buena aproximación**: tratar las búsquedas BLAST como un experimento científico más
  - Hipótesis (pregunta)
  - Diseño experimental (secuencia, BD, algoritmo, parámetros)
  - Resultados (salida)
  - Interpretación
- **Mala aproximación**: realizar búsquedas con una hipótesis o diseño pobre, y analizar los resultados en función de si obtengo lo que quería o no
  - Luego modifico el diseño, hasta que los resultados que obtenga confirmen lo que quería oír
  - Esta manipulación de los datos es posible en muchos casos, dada la flexibilidad de parámetros de BLAST.

Consideraciones generales

- Salida de BLAST
  - Un usuario novato se queda en los resúmenes de una línea
  - Un usuario avanzado examina los alineamientos y su estadística
  - Un profesional lee la sección final
    - Examina el espacio de búsqueda
      - Umbrales de listado W, T, A
      - Umbrales de extensión X, S
      - Matriz de puntuación

Consideraciones generales

- Debemos ser capaces de discernir, prever y corregir tres aspectos fundamentales respecto a la salida del algoritmo
  - El hecho de obtener demasiados alineamientos
  - El hecho de obtener muy pocos alineamientos
  - La significatividad de los alineamientos

Significatividad de los alineamientos

- Tras realizar una búsqueda con BLAST, obtenemos una lista de alineamientos ordenados por su E-valor
  - Un E-valor pequeño es probable que indique un alineamiento significativo, no solo estadísticamente sino biológicamente
    - Sin embargo, también puede ser un falso positivo
    - Además, por ejemplo, hay proteínas homólogas con una similitud baja, y por tanto el alineamiento no tendrá un E-valor muy bajo
- Consideremos un blastp sobre la proteína RBP4 (AAH20633.1), miembro de la familia de las lipocalinas

Significatividad de los alineamientos

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
NP_006735.2	retinol-binding protein 4 precursor [Homo sapiens] >sp|P02753.3|	420	420	100%	1e-117	U G M
AAH20633.1	Retinol binding protein 4, plasma [Homo sapiens]	419	419	100%	4e-117	G M
3FMZ_A	Chain A, Crystal Structure Of Retinol-Binding Protein 4 (Rbp4) In C	389	389	92%	4e-108	S
CAH72329.1	retinol binding protein 4, plasma [Homo sapiens]	388	388	92%	7e-108	G
EAW50068.1	retinol binding protein 4, plasma, isoform CRA_b [Homo sapiens]	386	386	92%	2e-107	G
CAA24959.1	unnamed protein product [Homo sapiens]	385	385	100%	6e-107	G M
1JYD_A	Chain A, Crystal Structure Of Recombinant Human Serum Retinol-	383	383	90%	2e-106	S
1BRP_A	Chain A, Crystal Structure Of The Trigonal Form Of Human Plasma	383	383	90%	2e-106	S
1JYJ_A	Chain A, Crystal Structure Of A Double Variant (W67IW91H) OF Re	378	378	90%	5e-105	S
1QAB_E	Chain E, The Structure Of Human Retinol Binding Protein With Its	372	372	89%	4e-103	S
2WQA_E	Chain E, Complex Of Ttr And Rbp4 And Oleic Acid >pdb|2WQA|F C	372	372	88%	6e-103	S
3BSZ_E	Chain E, Crystal Structure Of The Transthyretin-Retinol Binding Pr	372	372	87%	6e-103	S
2WR6_A	Chain A, Structure Of The Complex Of Rbp4 With Linoleic Acid	367	367	87%	1e-101	S
2WQ9_A	Chain A, Crystal Structure Of Rbp4 Bound To Oleic Acid	367	367	86%	1e-101	S
AAF69622.1	PRO2222 [Homo sapiens]	332	332	78%	7e-91	M
CAA26553.1	RBP [Homo sapiens]	208	208	49%	8e-54	G
CAB46489.1	unnamed protein product [Homo sapiens]	151	151	34%	1e-36	G
AAC02945.1	mutant retinol binding protein [Homo sapiens]	92.0	92.0	22%	1e-18	G
AAC02946.1	mutant retinol binding protein [Homo sapiens]	75.5	75.5	17%	1e-13	G
NP_001638.1	apolipoprotein D precursor [Homo sapiens] >sp|P05090.1|APOD_H	55.5	55.5	76%	1e-07	U G M
AAB32200.1	apolipoprotein D, apoD [human, plasma, Peptide, 246 aa]	54.7	54.7	68%	2e-07	
2HZQ_A	Chain A, Crystal Structure Of Human Apolipoprotein D (Apod) In C	44.3	44.3	68%	3e-04	S
AAB35919.1	apolipoprotein D [Homo sapiens]	42.4	42.4	36%	0.001	G M
EAW88163.1	progestagen-associated endometrial protein (placental protein 14,	40.4	40.4	87%	0.004	G M
EAW88165.1	progestagen-associated endometrial protein (placental protein 14,	40.0	40.0	62%	0.004	G
CAB43305.1	hypothetical protein [Homo sapiens]	40.0	40.0	62%	0.005	M
NP_001018059.1	glycodelin precursor [Homo sapiens] >ref|NP_002562.2| glycodeli	40.0	40.0	62%	0.005	U G M
1IW2_A	Chain A, X-Ray Structure Of Human Complement Protein C8gamm	38.5	38.5	60%	0.014	S
NP_000597.2	complement component C8 gamma chain precursor [Homo sapien	32.3	32.3	56%	0.97	U G M

Significatividad de los alineamientos

- Primer paso, ir mirando por orden de E-valor
  - Los primeros alineamientos son muy perfectos (E~10^{100}), con secuencias de nombres similares a RBP4
    - Redundancias que no ha podido resolver NCBI-BLAST, debido a que no son secuencias totalmente idénticas, etc.
    - Podemos solucionarlo utilizando RefSeq en vez de nr como DB
  - Hacia la mitad de la lista, con E-valor alrededor de 10^{-10} tenemos dos secuencias, RBP y apopiloprotein D
    - Estas dos secuencias son muy distintas a pesar de tener E valores similares
    - Importancia de inspeccionar los alineamientos

```
>gb|AAC02946.1|  mutant retinol binding protein [Homo sapiens]
Length=36

 GENE ID: 5950 RBP4 | retinol binding protein 4, plasma [Homo sapiens]
(Over 100 PubMed links)

 Score = 75.5 bits (184),  Expect = 1e-13, Method: Compositional matrix adjust.
 Identities = 34/36 (94%), Positives = 35/36 (97%), Gaps = 0/36 (0%)

Query  84   NWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGN  119
            NWDVCADMV TFTDTEDPAKFKMKYWGVASFLQKG+
Sbjct  1    NWDVCADMVDTFTDTEDPAKFKMKYWGVASFLQKGS  36
```

Una pequeña parte de RBP4 se alinea con el 94% de RBP

```
>ref|NP_001638.1|  apolipoprotein D precursor [Homo sapiens]
 sp|P05090.1|APOD_HUMAN  RecName: Full=Apolipoprotein D; Short=Apo-D; Short=ApoD; Flags:
Precursor
 gb|AAB59517.1|  apolipoprotein D precursor [Homo sapiens]
 ▶14 more sequence titles
 Length=189

 GENE ID: 347 APOD | apolipoprotein D [Homo sapiens] (Over 10 PubMed links)

 Score = 55.5 bits (132),  Expect = 1e-07, Method: Compositional matrix adjust.
 Identities = 46/163 (28%), Positives = 79/163 (48%), Gaps = 31/163 (19%)

Query  14   GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSA  73
              G+A    +  +  V+ENFD  ++ G WY + +K P         I A +S+ E G++
Sbjct  18   AEGQAFHLGKCPNPPVQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMENGKIKV  76

Query  74   T-----AKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKY-WGVASFLQKGNDDHWIVDT  127
                 A G V  +       +    T   +  +PAK ++K+ W + S         +WI+ T
Sbjct  77   LNQELRADGTVNQI-------EGEATPVNLTEPAKLEVKFSWFMPS------APYWILAT  123

Query  128  DYDTYAVQYSC----RLLNLDGTCADSYSFVFSRDPNGLPPEA  166
            DY+ YA+ YSC     +L ++D       ++++ +R+PN LPPE
Sbjct  124  DYENYALVYSCTCIIQLFHVD------FAWILARNPN-LPPET  159
```

Esta secuencia, con E-valor parecido, tiene una identidad mucho menor debido a su tamaño

Significatividad de los alineamientos

- Mucho más abajo, encontramos el componente complementario 8 gamma (NP_000597)
  - Tiene un E-valor muy malo (0.97) y puntuación baja (32.3)
  - La identidad es baja (25%) e incluye tres huecos
  - Parece razonable pensar que las dos proteínas no están relacionadas
    - ¡Pero lo cierto es que son homólogas!
    - BLAST es sólo una ayuda al descubrimiento de proteínas homólogas

Significatividad de los alineamientos

- Hay varias cuestiones que podemos considerar a la hora de decidir si dos proteínas son similares
  - **E-valor**: es una primera pista, pero cuidado con falsos positivos y con proteínas homólogas con baja identidad
  - **Tamaño**: dos secuencias homólogas no necesariamente tienen el mismo tamaño, o pueden compartir sólo alguna región
  - **Regiones o motivos**: por ejemplo RBP4 y 8-gamma comparten un motivo G*W típico de la familia de las lipocaínas
  - **Función biológica**: todas las lipocaínas son pequeñas, hidrófilas…
  - **Estructura 3D**: si las dos proteínas comparten alguna estructura bien conservada es otra evidencia para su homología
  - …

BLAST

Reducción del listado de resultados

- Es frecuente terminar con listas muy largas de alineamientos
- Algunas estrategias para reducirlas
  - Usar “refseq” como BD: elimina muchas entradas redundantes
  - Limitar los alineamientos a un solo organismo
  - Limitar a una porción de la secuencia (por ejemplo, un dominio o región característico de la proteína)
  - Ajustar los parámetros
    - Matriz de puntuación
    - Umbrales T, X, S, E

Reducción del listado de resultados

- Secuencias de baja complejidad
  - Secuencias con poca variación de nucleótidos/aminoácidos
    - PPCDPPPPPKDKKKKDDGPP
    - AAATAAAAAAAATAAAAAT
  - Suelen dar lugar a falsas coincidencias
  - NCBI-BLAST tiene una opción para filtrarlas

Incremento del listado de resultados

- También es común terminar con listas muy cortas de alineamientos (o no encontrar alineamientos en absoluto)
  - Muchos genes/proteínas no tienen coincidencias o no se conocen
- Algunas estrategias para intentar aumentar las coincidencias
  - Ajustar los umbrales, especialmente el umbral de E-valor
  - Probar matrices PAM más altas o BLOSUM más bajas
  - Buscar en BBDD adicionales
  - …

BLAST

Introducción

Algoritmo

Salida

Estrategias

Protocolos

- Mapeo y Exploración
- Descubrimiento de genes

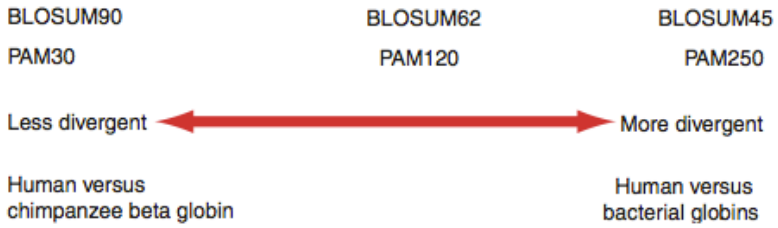
Otros programas

Protocolos

- La mayoría de las búsquedas BLAST son de dos tipos
  - **Mapeos**: encontrar la posición de una secuencia en otra
    - Se espera una coincidencia casi exacta de secuencias
    - El objetivo es encontrar la localización, no asociar las secuencias
    - Ejemplos: encontrar un gen en genoma o una región en proteína
  - **Exploraciones**: encontrar secuencias funcionalmente afines
    - Las estadísticas son de gran importancia
      - E, matriz de puntuación, % de identidad
    - Y también el conocimiento biológico
      - Filogenético, funcional, estructural, etc.
    - Ejemplo: encontrar proteínas homólogas (RBP4)

Parámetros y protocolos

- Match/mismatch de nucleótidos (blastn)
  - 99% de identidad (mapeo): +1/-3
  - 75% de identidad (exploración): +1/-1
- Matriz de puntuación en aminoácidos (blastp/n)
  - Mapeo: normalmente del mismo organismo o similar
    - BLOSUM62, BLOSUM80, PAM30
  - Exploración:
    - Entre organismos parecidos (BLOSUM62)
    - Entre organismos muy distintos (BLOSUM45)



Parámetros y protocolos

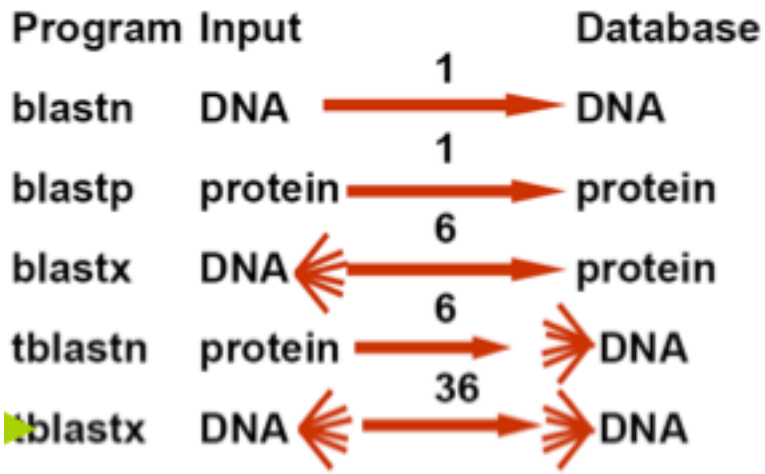
- Tamaño de palabra (en blastn)
  - Por defecto 11
  - Mapeo: puede ser más larga → más rápido
    - Nunca mayor que la secuencia a mapear
  - Exploración: según lo que se busque, se puede jugar con el tamaño
    - Mejor 9 que 11 en proteínas, evita codones degenerados.
- Penalización de gaps (blatsp/n)
  - Exploración: penalización alta
    - Un bloque funcional suele tener pocos huecos
  - Mapeo: penalización alta
    - O muy alta, hasta eliminar los saltos, en algunos casos

Parámetros y protocolos

- E-valor (blastp/n)
  - Exploración: depende de lo estricta que sea
    - Importancia de inspeccionar cada alineamiento por separado
  - Mapeo: muy bajo (en realidad poco importante)

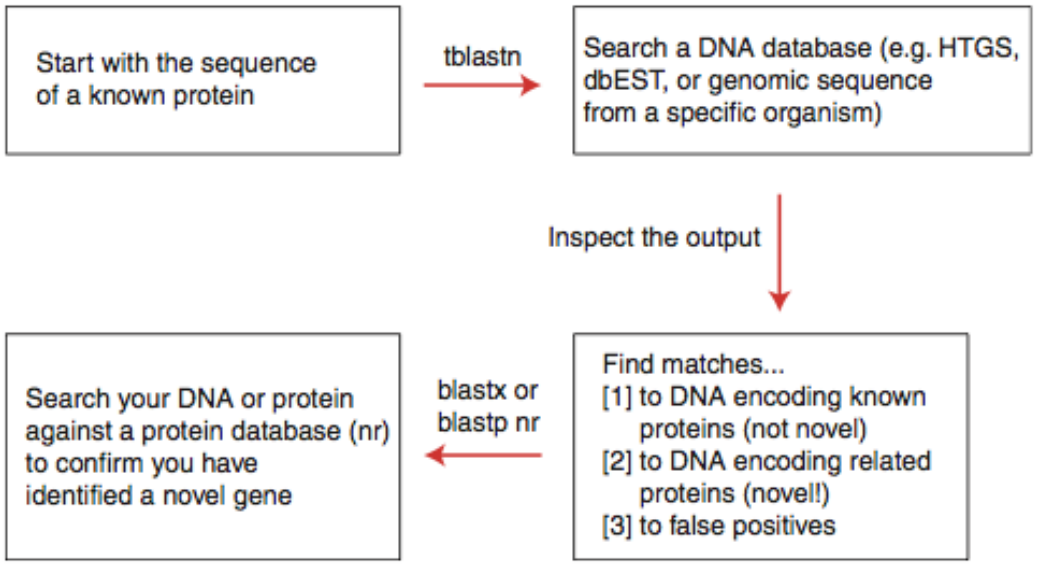
Tipos de mapeos

- Mapeos
  - Gen sobre genoma: blastn
  - Proteína sobre genoma: tblastn
- Exploraciones
  - Proteínas homólogas: blastp
  - Genes coincidentes: blastn
  - Genes que codifican proteínas: blastx
  - Genes posibles sobre secuencias genómicas: tblastx



Descubrimiento de genes

- Encontrar un nuevo gen en bioinformática equivale a descubrir una secuencia de ADN que no está anotada en una BD
  - No sustituye a las aproximaciones experimentales, si no que las complementa



Descubrimiento de genes Ejercicio

1. Elegir una proteína, incluyendo la especie y su identificador
   - p. ej. la beta-globina humana, con id NP_000509
2. Hacer un tblastn sobre una BD de ADN genómico
   - Evaluar E valores, puntuaciones y alineamientos, determinando:
     - Coincidencia perfecta: ya descubierta
     - Coincidencia cercana: posiblemente no descubierta aún
     - Falso positivo: un resultado que no es homólogo
   - Elegir una coincidencia como candidato a nueva proteína
3. Reunir información sobre la nueva proteína y su especie
   - Básicamente, la secuencia devuelva por el tblastn y de dónde viene
4. Demostrar que el gen y su proteína correspondiente son nuevos
   - Hacer un blastx de la secuencia contra la base de datos nr del NCBI

Descubrimiento de genes Ejercicio

- Consideraciones
  - Elección de la BD
    - Utilizaremos una BD de Expressed Sequence Tags (cadenas cortas -800 bases- de ADN expresadas en una región de un organismo en algún momento de su desarrollo)
    - Hay más probabilidades de encontrar un gen nuevo en un organismo que no haya sido anotado exhaustivamente
      - Mejor evitar humano, ratón o *S. cerevisiae*

Descubrimiento de genes Ejercicio

- Consideraciones
  - Para el ejercicio, consideraremos que el gen es "nuevo" si, realizando un blastx/blastp de su secuencia contra la BD no redundante (nr) del NCBI
    - Si hay una coincidencia del 100% de identidad con alguna proteína de la BD, de la misma especie que el gen "nuevo", el gen NO es nuevo (incluso si tiene por nombre *unknown*)
    - Si la mejor coincidencia tiene <100% de identidad, es posible que sea nueva
    - Si hay una coincidencia del 100%, pero en una especie distinta de la que empezaste, es un gen nuevo
    - Si no hay coincidencias con la proteína original de búsqueda, has encontrado un gen/proteína que no es homólogo con la búsqueda original. Probablemente haya habido un error, hay que volver a empezar o elegir otra proteína.

BLAST

Introducción

Algoritmo

Salida

Estrategias

Protocolos

Otros programas

PSI, RPS, PHI-BLAST

PatternHunter, BLASTZ, MegaBLAST

PSI-BLAST

- Position Specific Iterative BLAST
- Existen proteínas homólogas que no presentan similitud
  - Podemos usar distintas matrices BLOSUM o PAM en nuestra búsqueda BLAST para maximizar la sensibilidad de su alineamiento, pero a veces esto no es suficiente
- PSI-BLAST es más sensible que BLAST
  - Su objetivo es encontrar proteínas distantemente relacionadas

PSI-BLAST

- PSI-BLAST consta de 5 pasos
  1. Búsqueda blastp normal de una secuencia contra una BD
  2. Construcción de un alineamiento múltiple de las secuencias coincidentes, y creación de una matriz de búsqueda especializada (position-specific scoring matrix, PSSM) basada en dicho alineamiento
  3. Nueva búsqueda usando como matriz de puntuación la PSSM
  4. Se evalúa la significatividad estadística de las coincidencias
  5. Se repite el proceso, pero usando en 1 la PSSM para calcular la matriz de puntuación, hasta convergencia o nº máximo de iteraciones

PSI-BLAST
PSMM

- Filas: residuos de la secuencia de búsqueda
- Columnas: todos los aminoácidos
- La puntuación para un aminoácido puede variar dependiendo de su posición en la secuencia

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8	L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
14	A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15	A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2	-1	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
41	Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
...																					

PSI-BLAST

- Ventaja: las PSSM aportan mucha sensibilidad al método
  - Permiten asociar secuencias débiles en cuanto a su similitud pero fuertes en su relación biológica
- Desventaja: también aumentan el número de falsos positivos
  - Corrupción de la PSSM: cuando se asocia un falso positivo con la secuencia, éste se incorpora a la PSSM e incrementa la probabilidad de la inclusión de nuevos falsos positivos.

RPS-BLAST

- Reverse Position Specific BLAST
- Utiliza, en vez de una matriz de puntuación tradicional, una batería con muchas PSSMs

PHI-BLAST

- Pattern-Hit Initiated BLAST
  - A menudo la proteína de búsqueda contiene un patrón o "firma" en forma de conjunto de residuos que la definen como parte de una familia (una enzima, una secuencia que define un dominio funcional o estructural o una función conocida)
- PHI-BLAST permite encontrar proteínas
  - Relacionadas significativamente con la proteína de búsqueda (blastp)
  - Que contengan un determinado patrón
    - GXW[YF] → G seguido de cualquier aa (X), seguido de W, e Y ó F

DNA genómico y BLAST

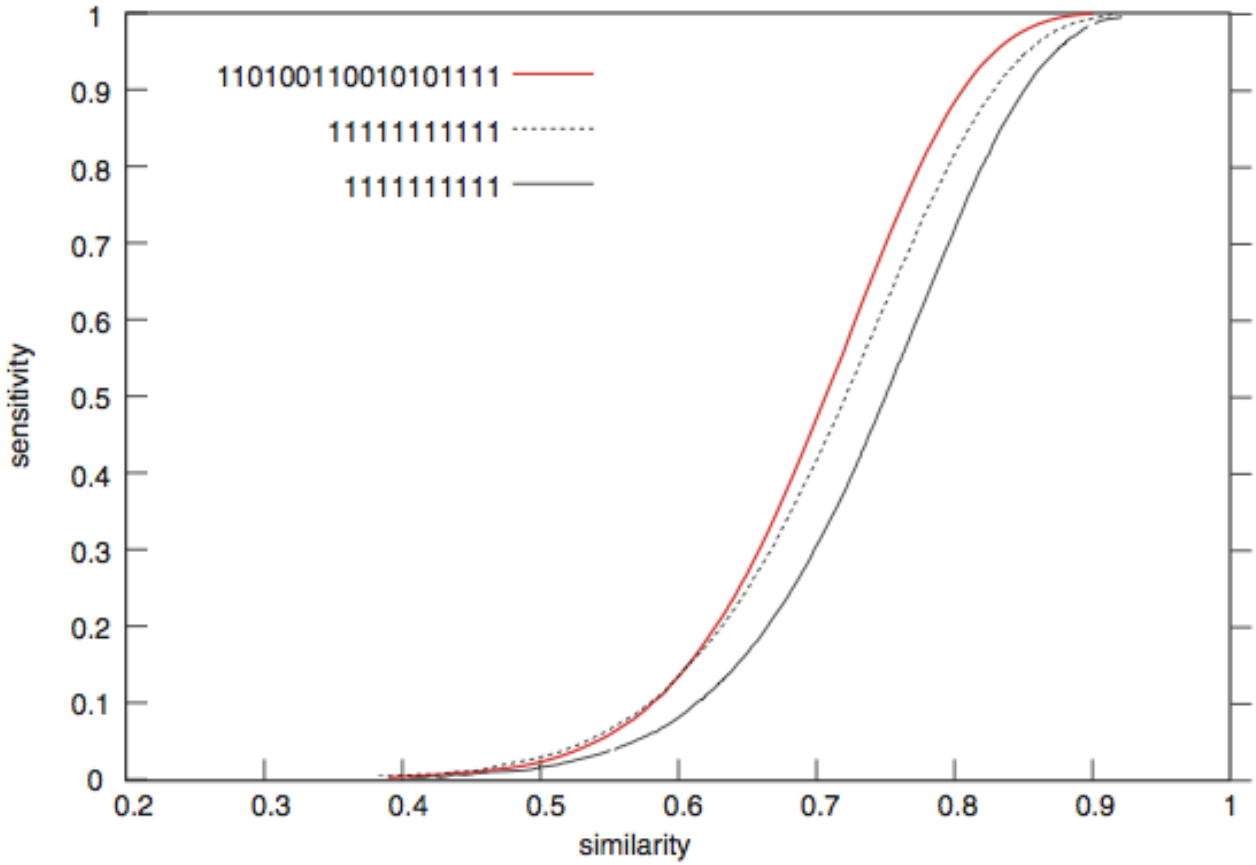
- Las BBDD de ADN genómico crecen rápidamente, y cada vez es más común buscar una proteína o secuencia de ADN contra un genoma.
- Este es un problema con características específicas
  - El ADN genómico tiene exones e intrones → queremos los exones
  - Queremos contemplar diferencias muy pequeñas tales como SNPs
  - También compararemos secuencias de organismos muy distintos, con deleciones, duplicaciones, inversiones y translocaciones.
- Veremos algunas herramientas para búsquedas de este tipo

PatternHunter

- blastn busca palabras de tamaño 11 que coincidan *exactamente*
  - Si 1 es una coincidencia, el patrón que se busca es 11111111111
- PatternHunter cambia el patrón a 110100110010101111
  - Permite mismatches (0s) entre medias, lo que incrementa el número de hits
    - P. ej. para dos secuencias de 64 nucleótidos con un 70% de similitud, blastn tiene un 30% de posibilidades de reportar una coincidencia, y PatternHunter un 47%
- BLASTZ y MegaBLAST adoptan este tipo de patrones

PatternHunter

- Incremento de la sensibilidad con el patrón "flexibe" de PatternHunter comparado con los patrones tradicionales de tamaño 10 y 11



BLASTZ

- Se desarrolló para alinear el genoma humano y el de ratón
  - Muy útil para comparar secuencias largas de genoma
- Básicamente es una modificación del BLAST con huecos
  - Busca coincidencias cortas casi exactas
  - Extiende sin permitir huecos
  - Hace una segunda extensión permitiendo huecos
- Mejoras
  - Eliminación de patrones ambiguos o repetitivos de ambas secuencias
  - Uso de coincidencias tipo PatternHunter (1110100110010101111)
  - Segunda búsqueda tras una coincidencia, en regiones adyacentes, usando un tamaño de palabra menor (7)

MegaBLAST

- Incrementa el tamaño de palabras de 11 a 28 (o hasta 64)
  - Esto acelera mucho su velocidad respecto a blastn
  - Aunque lo hace menos sensible
- Discontiguous-MegaBLAST adopta la estrategia de "palabra discontinua" de PatternHunter

Resumen

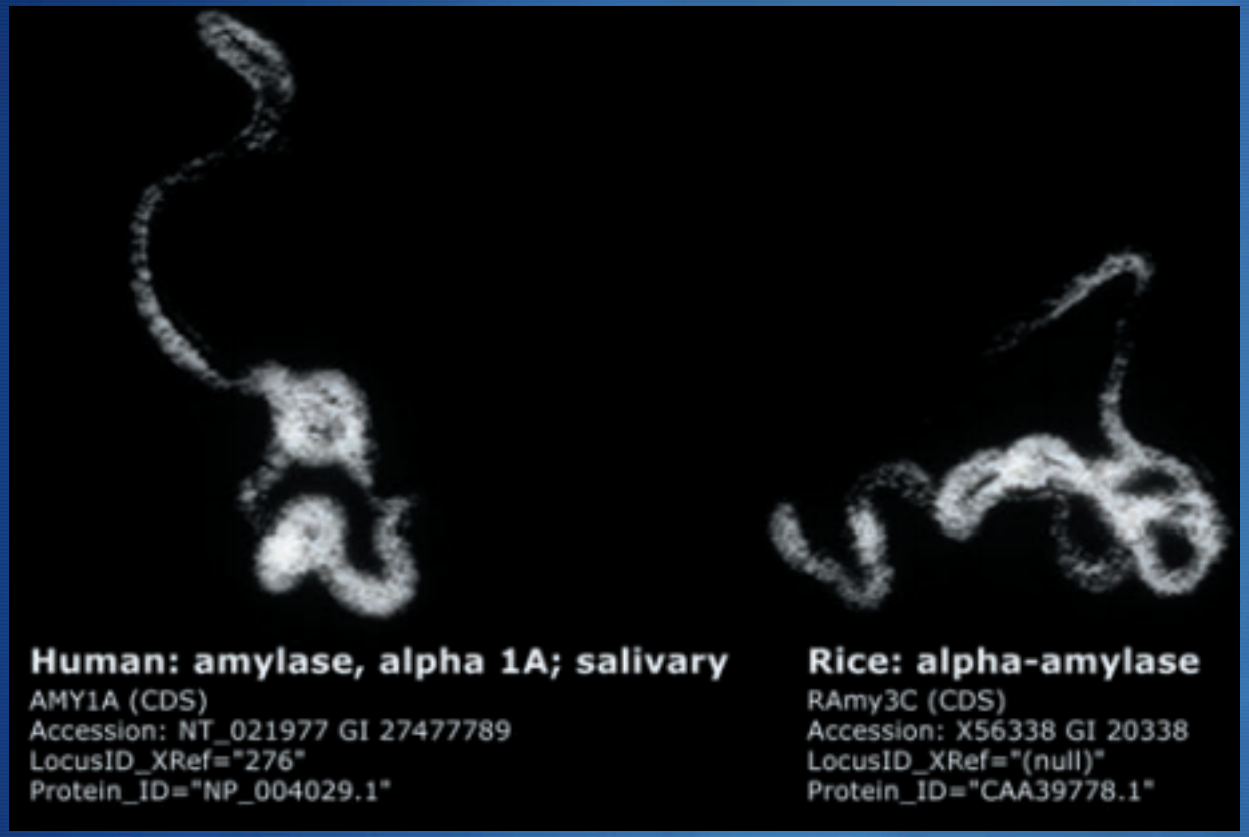
- BLAST es **una herramienta indispensable para encontrar relaciones** de una secuencia con las millones de secuencias existentes en las bases de datos públicas, a través de alineamientos de pares.
- El algoritmo de BLAST **divide la secuencia** de entrada, **busca "trozos" similares** en la BD **y** cuando los encuentra **expande el trozo** según una determinada métrica. Es un modo de alineamiento de pares muy **efectivo computacionalmente**.
- Existen **varios algoritmos** BLAST dependiendo de si comparamos nucleótidos, proteínas, etc.
- BLAST utiliza **una matriz BLOSUM o PAM**, pero versiones más avanzadas y cada vez más utilizadas, como PSI-BLAST, usan **matrices específicas** dependientes de la posición de los aminoácidos en la secuencia (PSSM).
- Es muy importante en una búsqueda BLAST la **elección de** la base de datos, el algoritmo y los **parámetros** más adecuados. Es de vital importancia también saber **interpretar los resultados** y discriminar los alineamientos significativos de los no significativos.

Preguntas para debate

- ¿Considerarías significativo un E-valor de 1, 0.05 o 10^{-5}? ¿Depende de la búsqueda particular que estés realizando?
- ¿Por qué un programa como BLAST debe tener un compromiso entre sensibilidad y especificidad? ¿Cómo hace blastp para ello? (consultar Altschul et al. 1990)
- ¿Por qué BLAST tiene tantas opciones? ¿NCBI-BLAST tiene pocas o muchas? ¿Convendría simplificarlo?

Lecturas adicionales

- Pevsner, 2009: Ch 4 *Basic Local Alignment Search Tool (BLAST)*
- Stephen F. Altschul et al., *Basic Local Alignment Search Tool.* J. Mol. Biol. 1990; 215:403-410
  - PMID: 2231712
- Stephen F. Altschul et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Research. 1997 Jul 16;25(17):3389-3402
  - PMCID: PMC146917



Ecce homology es una instalación artística que visualiza el proceso de búsqueda de secuencias similares entre el hombre y el arroz con BLAST

El espectador puede modificar y seleccionar los genes a alinear con sus movimientos

Proyecto
http://www.viewingspace.com/cccehomology.html
Paper:
http://la.remap.ucla.edu/~jburke/publications/West-et-al-2005_Ecce-Homology.pdf