

Tema 2: Alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST

Introducción a la Filoinformática:
IBMB-UNLP/CONICET, Argentina. 2-6 Julio 2018.

Pablo Vinuesa (vinuesa@cgc.unam.mx)

Programa de Ingeniería Genómica, CGG-UNAM, México
<http://www.cgc.unam.mx/~vinuesa/>

Todo el material del curso (presentaciones, tutoriales y datos de secuencias) lo encontrarás en:
<https://github.com/vinuesa/intro2phyloinfo>

Tema 2: alineamientos pareados y búsqueda de homólogos en bases de datos

- evolución de secuencias y clasificación de mutaciones
- indeles y gaps
- alineamientos globales (Needleman-Wunsch) vs. locales (Smith-Waterman);
- programación dinámica;
- dot plots;
- matrices de costo de sustitución, penalización de gaps y cuantificación de la similitud;
- evaluación estadística de la similitud entre pares de secuencias;
- escrutinio de bases de datos mediante BLAST; Búsquedas a nivel de DNA vs. AA;
- la familia BLAST e interpretación de resultados de búsqueda de secuencias homólogas
- prácticas: uso de NCBI BLAST en línea

Introducción a la filoinformática – pan-genómica y filogenómica. IBBM-UNLP/CONYCET, Argentina. Julio 2018

Protocolo básico para un análisis filogenético de secuencias moleculares

Tema 3:
alineamientos
pareados, búsquedas
de homólogos en
bases de datos

Colección de secuencias homólogas
• BLAST y FASTA

Alineamiento múltiple de secuencias
• Clustal, T-Coffee, muscle...

Análisis evolutivo del alineamiento y selección del modelo de sustitución más ajustado
• tests de saturación, modeltest, ...

Estima filogenética
• NJ, ME, MP, ML, Bayes ...

Pruebas de confiabilidad de la topología inferida
• proporciones de bootstrap
probabilidad posterior ...

Interpretación evolutiva y aplicación de las filogenias

Homología entre secuencias de DNA y proteína:
tipos de mutaciones en secs. codificadoras de proteínas

secuencia ancestral

pos. codón 123
codones ATG TGT TTT GAT GCA
AA M C F D A

especie A
ATG TAT TTT CAT GCA
M T F H A
no-sinónima

especie B
ATG --- TTC GAC GCA
M F D A
sinónimas y deleción en marco

especie C
ATG TGT TT- G ATG CAX
M C L M X
deleción fuera de marco

- Todas las mutaciones en 2ª posiciones resultan en sustituciones no sinónimas
- 96% de mutaciones en 1ª posiciones resultan en sustituciones no sinónimas
- Casi todas las sustituciones sinónimas ocurren en las 3ª posiciones
- las deleciones o inserciones en secs. codificadoras de aa suceden generalmente en múltiplos de tres nt; de no ser así se generan cambios de marco de lectura corriente abajo de la mutación, con frecuencia generando un pseudogen no funcional

Homología entre secuencias de DNA y proteína:
alineamiento y tipos de mutaciones

secuencia ancestral

pos. codón 123
codones ATG TGT TTT GAT GCA
AA M C F D A

alineamiento de sitios homólogos para tres secs.

especie A ATG TAT TTT CAT GCA
especie B ATG --- TTC GAC GCA
especie C ATG TGT TT- G ATG CAX

cambio de marco de lectura !!! posible pseudogen.

Transiciones (ti) purina - purina

α_{A-C}

β_{A-C}

β_{C-G}

β_{A-T}

β_{G-T}

Transiciones (ti) pirimidina - pirimidina

α_{C-G}

Transversiones (tv) pur. \leftrightarrow pyr.

β_{G-T}

- existen 4 tipos de ti y 8 de tv
- las tasas de sustitución de ti (α) son generalmente mucho más altas que las de tv (β)

© Pablo Vinuesa 2018,
vinuesa[at]cgc[dot]unam[dot]mx;
<http://www.cgc.unam.mx/~vinuesa/>

Estadísticos de Karlin-Altschul de similitud entre secuencias: frecuencias diana, lambda y entropía relativa

Los atributos más importantes de una matriz de sustitución son sus **frecuencias esperadas** o **diana** implícitas para cada par de aa en sus respectivos scores crudos. Estas frecuencias esperadas **representan el modelo evolutivo subyacente**. Los scores que han sido re-escalados y redondeados (scores representados en la matriz) son los **scores crudos** $s_{a,b}$. Para convertirlos a un **score normalizado** (log-odd score original) tenemos que multiplicarlos por λ , una constante específica para cada matriz. λ es aprox. igual al inverso del factor de escalamiento (c).

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$
$$p_{ab} = f_a f_b e^{\lambda s_{ab}} = \text{score normalizado}$$

por tanto, para despejar λ necesitamos $f_a f_b$ y encontrar el valor de λ para el que la suma de las frecuencias diana implícitas valga 1.

$$\sum_{a=1}^n \sum_{b=1}^n p_{ab} = \sum_{a=1}^n \sum_{b=1}^n f_a f_b e^{\lambda s_{ab}} = 1$$

Una vez calculada λ , se usa para calcular el **valor de expectación (E)** de cada **HSP (High Scoring Pair)** en el reporte de una búsqueda **BLAST**

Dado que las $f_a f_b$ de los residuos de algunas proteínas difieren mucho de las frecuencias de residuos empleadas para calcular las matrices PAM y BLOSUM, versiones recientes de BLASTP y PSI-BLAST incorporan una **"composition-based λ "** que es **"hit-específica"**

Programación dinámica y la generación de alineamientos pareados (globales y locales)

• Pares de secuencias pueden ser comparadas usando **alineamientos globales y locales**, dependiendo del objetivo de la comparación.

Un **alineamiento global** fuerza el alineamiento de ambas secuencias a lo largo de toda su longitud. Usamos aln. globales cuando estamos seguros de que la homología se extiende a lo largo de todas las secuencias a comparar. Este es el tipo de alineamientos que generan programas de alineamiento múltiple tales como clustal, T-Coffee o muscle.

(a)

P00001	1	MGDVERGKKIFIMKCSQCHTVEKGGKHKTPNLHGLFGRKTGQAPGYSYTAANKNKG	58
		D KG+ +F Q C T + K+ GP L G+ GRK G A G+Y+ N N G+	
P00090	1	Q-DAARGEAVF-----KQCMTCHRADKNMVGPA LGGVVGKAGTAAGFTYSPLNHNHSGEAGL	56
P00001	59	IWGEDTLMYELENPKKIYP-----GTMKMFVGIKKKEERADLIAYLKKAATNE	105
		+W ++ ++ YL +P Y+ T K M F + ++R D+ AYL AT +	
P00090	57	VWTQENI IAYLPDPNAYLKKFLTDKGQADKATGSKTMTF-KLANDQQQKDVAAYL--ATLK	114

Alineamiento global óptimo del citocromo C humano (105 residuos, SWISS-PROT acc. P00001) y citocromo C2 de Rhodopseudomonas palustris (114 residuos, SWISS-PROT acc. P00090).

La matriz de puntuación o ponderación ("scoring matrix") empleada fue **BLOSUM62**, con **costo de gaps afines** de -(11 + k). La puntuación del alineamiento global es de 131, usando el algoritmo de **Needleman-Wunsch**.

Alineamientos pareados y búsqueda de homólogos en bases de datos

Los **alineamientos pareados** son la base de los métodos de búsqueda de **secuencias homólogas** en bases de datos

- Si dos proteínas o genes se parecen mucho a lo largo de toda su longitud asumimos que se trata de proteínas o genes homólogos, es decir, descendientes de un mismo ancestro común (cenastro).
- Por ello una de las técnicas más utilizadas para detectar potenciales homólogos en bases de datos de secuencias se basa en la **cuantificación de la similitud entre pares de secuencias** y la determinación de la **significancia estadística** de dicho parecido. Estas magnitudes son las que reportan los estadísticos de **BLAST**.

```
>gi171548896|ref|PF_00669120.1| Translation elongation factor G:Small GTP-binding protein domain
[Nitrosomonas eutropha C71]
gi171486077|gb|EAO18626.1| Translation elongation factor G:Small GTP-binding protein domain
[Nitrosomonas eutropha C71]
Length=696

Score = 828 bits (2140), Expect = 0.0
Identities = 434/697 (62%), Positives = 541/697 (77%), Gaps = 9/697 (1%)

Query 1 MTRFSELEKTRNIGIMAHIDAGKTTTTERVLYTTGRIHIGETHEGASQMDWMAQEQERG 60
M++ LE+ RNIGIMAHIDAGKTTT+ER+L+VTG HK+GE H+GA+ MDMM QEQERG
Sbjct 1 MSKRNPLERYENIGIMAHIDAGKTTTSERLIFTTGVSHKIGEVHDDGAATMDMMEQEQERG 60

Query 61 XXXXXXXXXXXXN-----DHRINIIDTPGHVDFTVEVERSLRVLDGAVAVLDAQSGVE 113
ITITSAATT W +HRIN+IDTPGHVDFT+EVERSLRVLDGA V + GV+
Sbjct 61 ITITSAATTCFVGMAGNYPEHRINVIDTPGHVDFTIEVERSLRVLDGACTVFCVSGVGQ 120 (... truncado)
```

Programación dinámica y la generación de alineamientos pareados (globales y locales)

Un **alineamiento local** sólo busca los segmentos con la puntuación más alta. Se usa por ejemplo en el escrutinio de bases de datos de secuencias debido a que la homología entre pares de secuencias frecuentemente existe sólo a nivel de ciertos dominios, pero no a lo largo de toda la secuencia (**estructura modular de proteínas; genes discontinuos intrones-exones; barajado de exones ...**). **BLAST** y **FASTA** buscan alineamientos locales con alta puntuación (**HSPs** ó high-scoring pairs)

(b)

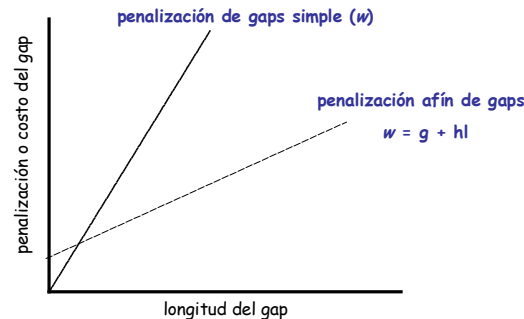
P13569	1221	EGGNAILLENISFSISPGQRVGLLGRGTSGSKSTLLSAFLRL-----NTEGEIQIDGVS	1273
		+ ++ +S ++ G+ + L+G +GSGKS +A L +L T GEI DG	
P33593	13	QAAQPLVHGVSLLQGRVLA LVGSGSGKSLTCAATLGILPAGVRQTAGEILADGKP	70
P13569	1274	WDSITL-----QQWRKAFGVIPQKVFI FSGTFRKNLDPYEQWSDQEIWKVADEV	1322
		L Q R AF + + + + + K AD+	
P33593	71	VSPCALRGIKIATIMQNPFSAFNPL-----HTMHTHARETCIALGKPADDA	116
P13569	1323	GLRSVIEQFP-GKLDVFLVDGGCVLSHGKQMLCLARSVLSKAKILLDEPSAHLDPV	1379
		L + IE VL +S G Q M +A +VL ++ ++ DER+ LD V	
P33593	117	TLTAATRAVGLENAARVLKLYPFMSGGMLQRMIMAMVLCESPFIIADEPTDLDVV	174

Alineamiento local óptimo del regulador de conductancia transmembranal de fibrosis cística de humano (1480 residuos, SWISS-PROT acc. P13569) y la proteína transportadora de Ni dependiente de ATP de E. coli (253 residuos, SWISS-PROT acc. P33593).

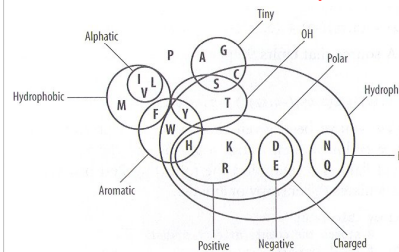
La matriz de puntuación o ponderación ("scoring matrix") empleada fue **BLOSUM62**, con **costo de gaps afines** de -(11 + k). La puntuación del alineamiento local es de 89, usando el algoritmo de **Smith-Waterman**.

alineamientos pareados y factores de penalización afines para gaps

- Dado que **un sólo evento mutacional puede insertar o eliminar varios nucleótidos** de una secuencia, un indel largo no debe de ser penalizado mucho más que otro más corto ubicado en la misma región de un gen. De ahí el uso de **factores de penalización afines para gaps** (affine gap penalties or costs), que cobran una penalidad relativamente alta por abrir un gap y una penalidad más baja por cada posición sobre la que se extiende.
- La calidad de un alineamiento depende en gran medida de los **valores de apertura y extensión de gap** elegidos.



Similitud entre pares de secuencias de AA



- Las matrices empíricas de sustitución entre AAs no reflejan necesariamente las relaciones químicas entre ellos. Se trata de una **definición puramente estadística basada en el análisis de frecuencias empíricas de sustituciones observadas en alineamientos de secs.** con un grado de divergencia definido

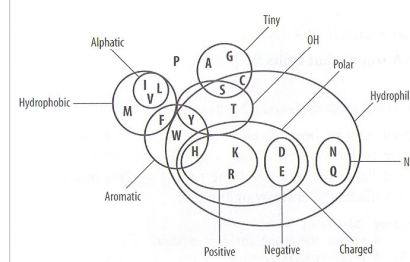
- Cada score de la matriz representa la tasa de sustitución esperada entre un par de AAs. Por tanto, los scores de los alineamientos pareados evaluados con estas matrices reflejan la distancia evolutiva existente entre las secuencias. Es importante notar que los scores son evolutivamente simétricos al no conocerse la dirección del cambio evolutivo.

Table 2 - The log odds matrix for BLOSUM62

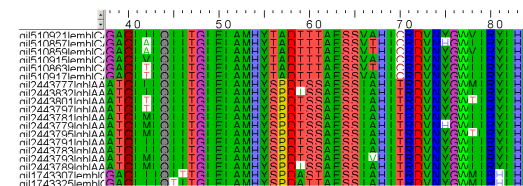
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	-1	-1	-1	-1	-1	-1	-1	1	1	1	-2	-3	-3	-1	-1	-1	0	-1	-1
C	9	3	-3	-4	-4	-4	-4	-4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
D		6	4	-2	-3	-3	-3	-3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
E			4	4	4	-2	-3	-3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
F				6	4	4	-2	-3	-1	-1	-1	-1	-1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

Matriz BLOSUM62

Similitud entre pares de secuencias de AA



- Este diagrama muestra a los **aminoácidos** agrupados atendiendo a sus características químicas y físicas.
- Desde una perspectiva evolutiva **esperamos encontrar más sustituciones entre aas. similares que entre los menos relacionados.**
- Estos patrones puede observarse en alineamientos múltiples como el mostrado abajo



Segmento de un
aln. múltiple de
citocromos C
de primates

Similitud entre pares de secuencias de AA

- Matrices de sustitución de AAs
log-odds scores

$$s(a,b) = (c) \log \frac{p_{ab}}{f_a f_b}$$

$s(a,b)$ = score del par a, b

Table 2 - The log odds matrix for BLOSUM 62

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

Matrix BLOSUM62

p_{ab} = verosimilitud de la hipótesis a testar; **frecuencia esperada o diana**, probabilidad con la que esperamos encontrar a y b apareados en un alineamiento múltiple

$f_{a|b}$ = verosimilitud de la hipótesis nula; **frecuencia de fondo**, probabilidad con la que esperamos encontrar a y b en cualquier proteína. Refleja su abundancia o frecuencia

c = Factor de escalamiento usado para multiplicar los lod scores (números reales) antes de ser redondeados a números enteros, tal y como se observa en la matriz. Los valores enteros redondeados resultantes se conocen como "raw scores".

Tema 2: Alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST

Estadísticos de Karlin-Altschul para alineamientos locales

Karlin, S., and Altschul, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A 87: 2264-268.

$$E = k m n e^{-\lambda S}$$

Esta ecuación indica que el número de alineamientos esperados por azar (E) durante una búsqueda de similitud en una base de datos de secuencias está en función de: el tamaño del espacio de búsqueda (m, n), el score normalizado (λS) del HSP y una constante de valor pequeño (k)

E Describe el ruido de fondo por azar presente en matches de dos secs.

m = número de símbolos en la secuencia problema

n = número de símbolos en la base de datos

k ≈ 0.1 constante de ajuste para considerar HSPs altamente correlacionados

Introducción a la filoinformática – pan-genómica y filogenómica. IBBM-UNLP/CONYCECT, Argentina. Julio 2018

BLAST: Basic Local Alignment Search Tool

BLAST consta de una familia de programas. Los 5 ppales son:

BLASTN (nt-nt), BLASTP (p-p), BLASTX (translated nt-p), TBLASTN (p-translated nt), usado en mapeo de prots contra DNA genómico TBLASTX (translated nt - translated nt) usado en la predicción de genes

y variantes de BLASTP como PSI- y PHI-BLAST

NCBI → BLAST

Updated: 5 Mar 2007 Data: test of new BLAST interface

About

- Getting started
- News
- FAQs

More info

- NSF 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI
- Contributors
- Mailing list
- Contact us

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

New BLAST beta Try it at: <http://ncbi.nlm.nih.gov/blast/beta>

Nucleotide <ul style="list-style-type: none">Quickly search for highly similar sequences (megablast)Quickly search for divergent sequences (discontiguous megablast)Nucleotide-nucleotide BLAST (blastn)Search for short, nearly exact matches (tblastx)Search trace archives with (megablast or discontiguous megablast)	Protein <ul style="list-style-type: none">Protein-protein BLAST (blastp)Position-specific iterated and pattern-motif initiated BLAST (PSI- and PHI-BLAST)Search for short, nearly exact matches (tblastx)Search the conserved domain database (rdblast)Protein homology by domain architecture (dbaf)
Translated <ul style="list-style-type: none">Translated query vs. protein database (tblastn)Protein query vs. translated database (blastx)Translated query vs. translated database (tblastx)	Genomes <ul style="list-style-type: none">Human, mouse, rat, chimp, cow, pig, dog, sheep, catChicken, puffer fish, zebrafishPhylopharyngeal, other insectsMicrobes, environmental samplesPlants, nematodesFungi, protozoa, other eukaryotes
Special <ul style="list-style-type: none">Search for gene expression data (GEO BLAST)Align two sequences (blast2seq)Screen for vector contamination (VecScreen)Immunoglobulin BLAST (igblast)SNP BLAST	Meta <ul style="list-style-type: none">Retrieve results

BLAST: Basic Local Alignment Search Tool

• Anatomía de un reporte de NCBI-BLAST estándar

BLASTP 2.2.13 [Nov-27-2005]

1

References:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1990), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 18:3298-3302.
RID: 1141782136-12667-92041342765.BLASTP4

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples 3,420,754 sequences; 1,167,289,757 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQ](#) [taxonomy reports](#)

Query: human_myoglobin
Length=154

Distribution of 500 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Color key for alignment scores
Query 0 30 60 90 120 150
40-60 60-80 80-200 >200

1.- Encabezado. Indica el programa de BLAST y su versión, con la fecha

Request ID

Indica la BD sobre la que se hizo la búsqueda, junto con el no. de secs contenida en ella y el no. de caracteres

Indica cual fue la query y su longitud

2.- Resumen gráfico de distribución de hits con respecto a la query.

escala de color que indica el score de los HSPs

Las barras indican la distribución de los HSPs (coordenadas) con respecto a la secuencia problema (query), indicando en una escala de color el score de los alns. medidos en bits

BLAST: Basic Local Alignment Search Tool

• Anatomía de un reporte de NCBI-BLAST estándar

3. Resúmenes de 1 línea. Indican el nombre de la sec. junto con el score más alto y E value más bajo encontrado para un HSP o grupo de HSPs

Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value	
gi 4885477 ref NP_005359.1 myoglobin [Homo sapiens] >gi 4495...	316	6e-86	Gene Info
gi 62511907 gb AA84516.1 myoglobin transcript variant 1 [Homo	315	1e-85	
gi 386872 gb AA59595.1 myoglobin	315	1e-85	
gi 229361 ref U71165.8 myoglobin	313	4e-85	
gi 127683 ref P02145 MYG_PANTR Myoglobin	312	9e-85	Structures
gi 151317414 ref P62735 MYG_HYLY Myoglobin >gi 151317413 ref P62734	311	1e-84	
gi 1276561 ref P02147 MYG_GORER Myoglobin	311	2e-84	
gi 229360 ref U71165.8 myoglobin	311	2e-84	
gi 155728442 emb CA90965.1 hypothetical protein [Pongo pygmaeus	310	5e-84	
gi 230638 pdb 2MM1 Myoglobin Mutant With Lys 45 Replaced By...	309	6e-84	
gi 127689 ref P02148 MYG_PONPY Myoglobin >gi 229570 ref U761377A	308	2e-83	
gi 62901707 ref P68086 MYG_BRYPA Myoglobin >gi 62901706 ref P68...	300	4e-81	

© Pablo Vinuesa 2018,
vinuesa[at]ccg[dot]unam[dot]mx;
<http://www.ccg.unam.mx/~vinuesa/>

BLAST: Basic Local Alignment Search Tool

Anatomía de un reporte de NCBI-BLAST estándar

4. **Alineamientos.** Representan la parte más voluminosa del reporte. Además de la información estadística, indica las coordenadas de inicio y fin de las secuencias query y subject. Si la búsqueda involucra secuencias de DNA, también se indica direccionalidad de las hebras Q/S (plus/plus; plus/minus).

```
>gi|475235461ref|NF_999401.1|g|127688|spl|P02189|MYG_PTG|6| Myoglobin [Sus scrofa]
gi|127688|spl|P02189|MYG_PTG|6| Myoglobin
gi|164547|gb|AAA31073.1|g| Myoglobin
length=154
normalized score
Score = 296 bits (758), Expect = 5e-80
Identities = 144/154 (93%), Positives = 148/154 (96%), Gaps = 0/154 (0%)

Query 1 MGLSDGEWQLVLNVWGKVEADIPGHGQEVLRLEFKGHPETLEKFDKFKHLKSEDEMKASE 60
      1 MGLSDGEWQLVLNVWGKVEAD+ GHGQEVLRLEFKGHPETLEKFDKFKHLKSEDEMKASE 60
Sbjct 1 MGLSDGEWQLVLNVWGKVEADVAGHGQEVLRLEFKGHPETLEKFDKFKHLKSEDEMKASE 60

Query 61 DLKKHGATVLTALGGILKKKGHHEAIEKPLAQSHATKKHIVPKYLEFISECIIQVLSKH 120
      61 DLKKHG TVLTALGGILKKKGHHEA+ PLAQSHATKKHIVPKYLEFISECIIQVLSKH 120
Sbjct 61 DLKKHGNVLTALGGILKKKGHHEAELTPLAQSHATKKHIVPKYLEFISECIIQVLSKH 120

Query 121 PGDFGADAQAGAMNKALELFRKDMASNYKELGFQG 154
      121 PGDFGADAQAGAM+KALELFR DMA+ YKELGFQG
Sbjct 121 PGDFGADAQAGAMSKALELFRNDMAAKYKELGFQG 154
```

BLAST: Basic Local Alignment Search Tool

RESUMEN de gapped-BLAST

- BLAST es un programa para búsqueda de secuencias similares a una sec. problema en bases de datos. BLAST puede ser usado en línea o localmente.
- Existen **diversos programas BLAST** para comparar todas las combinaciones posibles de secs. problema (aa y nt) con nt o aa DBs. (**BLASTN**, **BLASTP**, **BLASTX**, **TBLASTN**, **TBLASTX**) además de variantes de éstos que buscan similitudes en diversas DBs
- BLAST es una **versión heurística del algoritmo de Smith-Waterman** que encuentra matches locales cortos (**palabras**) que intenta extender en forma de alineamientos pareados
- El nuevo algoritmo **gapped-BLASTP** requiere al menos de dos palabras o hits no solapados con un score de al menos **T**, ubicados a una distancia máxima **A** el uno del otro, para invocar una extensión del segundo hit. Si el **HSP** generado tiene un score normalizado con un valor de al menos **Su** (**normalized ungapped score**) bits, se dispara una extensión con gap
- BLAST reporta además información relativa a la significancia estadística de los HSPs encontrados. El estadístico fundamental es el **valor de expectancia E (E-value)**, que indica el número de falsos positivos que cabe encontrar, dada la longitud de la secuencia problema, el tamaño de la base de datos explotada, y el score normalizado del HSP, tal y como indica la **ecuación de Karlin-Altschul**

$E = k m n e^{-\lambda S}$

- Si bien no existe una teoría estadística para evaluar explícitamente la significancia de alns. con gaps (no se puede estimar λ) éstas pueden obtenerse a partir de simulaciones *in silico*

BLAST: Basic Local Alignment Search Tool

Anatomía de un reporte de NCBI-BLAST estándar

5. **Pie de página.** Reporta los parámetros de búsqueda y varios estadísticos. Los más importantes son: **DB**, **T**, **E** y la **matriz de sustitución** o esquema de puntuación (match/mismatch) y **gap penalties** empleados

```
Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding
environmental samples
Posted date: Mar 6, 2006 5:22 AM
Number of letters in database: 327,455,400
Number of sequences in database: 872,833
Lambda K H
0.316 0.135 0.398
Gapped Lambda K H
0.267 0.0410 0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 872833
Number of Hits to DB: 3803460
Number of extensions: 145241
Number of successful extensions: 500
Number of sequences better than 10: 117
Number of HSP's better than 10 without gapping: 0
Number of HSP's gapped: 444
Number of HSP's successfully gapped: 121
Length of query: 154
Length of database: 327455400
Length adjustment: 111
Effective length of query: 43
Effective length of database: 327455400
Effective search space: 14080582200
Effective search space used: 9914550291
T: 11
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
S2: 66 (30.0 bits)
```

$E = k m n e^{-\lambda S}$

matriz de sustitución

gap penalties

E value umbral usado = 10; HSPs con gap

E value umbral usado = 10; HSPs no gap

neighborhood word threshold score

two-hit distance

extension attenuation parameter

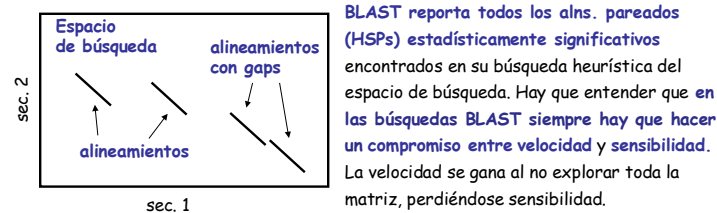
aln. threshold (ungapped)

aln. threshold (gapped)

BLAST: Basic Local Alignment Search Tool

• El algoritmo BLAST

El **espacio de búsqueda entre 2 secs.** puede ser visualizado como una gráfica con una sec. en cada eje. Sobre esta gráfica podemos visualizar **alineamientos** como una secuencia de pares de letras con o sin gaps. $Score = \text{sumatoria de scores individuales } p_{ab} - \text{costo gaps}$. **BLAST no explora todo el espacio de búsqueda entre dos secuencias (es un heurístico).**

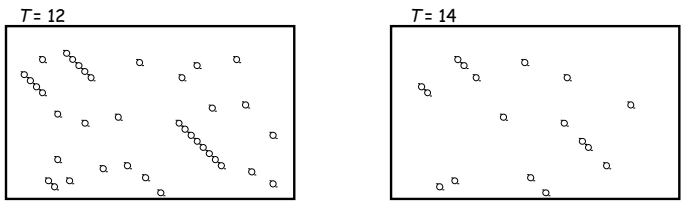


El algoritmo heurístico de BLAST sigue tres niveles de reglas para refinar secuencialmente HSPs (High Scoring Pairs) potenciales: **ensemillado**, **extensión** y **evaluación**. Estos pasos conforman una estrategia de refinamiento secuencial que le permite a BLAST muestrear todo el espacio de búsqueda sin perder tiempo en regiones de escasa similitud

BLAST: Basic Local Alignment Search Tool

• Ensemillado

El valor adecuado de T depende de los valores en la tabla de sustitución empleada, como del balance deseado entre velocidad y sensibilidad. A valores más altos de T , menos palabras son encontradas, reduciendo el espacio de búsqueda. Ello hace las búsquedas más rápidas, a costa de incrementar el riesgo de perder algún alineamiento significativo.

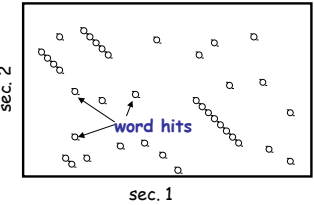


El **tamaño de palabra W** es otro parámetro que controla el número de word hits. $W=1$ producirá más hits que $W=5$. Cuanto más chico sea W más sensible y lenta la búsqueda. La interrelación entre W , T y la matriz de sustitución empleada es crítica, y su selección juiciosa es la mejor manera de controlar el balance entre velocidad y sensibilidad de BLAST

BLAST: Basic Local Alignment Search Tool

• Ensemillado

BLAST asume que los alineamientos significativos contienen "**palabras**" en común (serie de letras). BLAST primero determina la localización de todas las palabras comunes ("**word hits**"). **Sólo las regiones que contienen word hits serán usados como semillas de alineamientos.** Así se reduce mucho el espacio a explorar. BLAST usa el concepto de **vecindad** para definir un **word hit**. Esta contiene a la palabra misma y todas las demás cuyo score sea al menos tan grande como T cuando se compara con la matriz de ponderación. T corresponde a un umbral (Threshold) mínimo de score que han de tener las palabras encontradas.



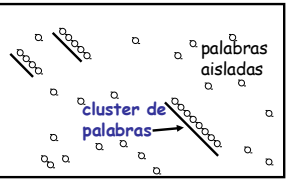
MPRDG	{	MPR	secuencia y
		PRD	palabras de
		RDG	3 letras

Palabra	Score (Blosum62)
RDG	17
KGD	14
QGD	13
RGE	13
EGD	12
...	

BLAST: Basic Local Alignment Search Tool

• Ensemillado

Las palabras tienden a agruparse en clusters en algunas regiones del espacio. BLAST usa el **two-hit algorithm** para seleccionar regiones con al menos dos palabras agrupadas dentro de una distancia definida sobre la diagonal. De esta manera **se eliminan palabras sin significancia, que carecen de vecinos**. Cuanto más grande la distancia impuesta al algoritmo (A), más palabras aisladas serán ignoradas, reduciéndose consecuentemente el espacio de búsqueda, incrementándose la velocidad a costa de perder sensibilidad.



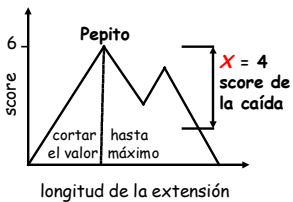
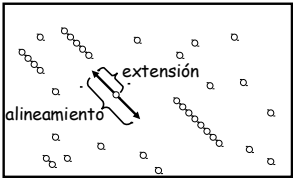
• **Detalles de implementación:** BLASTN vs. BLASTP
1. En **NCBI-BLASTN** las semillas son siempre palabras idénticas. T no es usado. Para hacer BLASTN más rápido se incrementa W , por hacerlo más sensible se disminuye W . El valor min. de $W=7$. El algoritmo de two-hit tampoco es usado por BLASTN ya que hits de palabras largas idénticas son raros.

2. En **BLASTP** (y otros programas basados en aa) se usan valores de W de 2 ó 3. Para hacer las búsquedas más rápidas $W=3$ y $T=999$, que elimina todas las palabras vecinas. La distancia (A) entre vecinos del algoritmo two-hit es por defecto = 40 aas. Las palabras que ocurren con una frecuencia significativamente mayor que la esperada por azar (FFF) corresponden frecuentemente a **regiones de baja complejidad** (rbc) que generalmente **son enmascaradas**. El uso de "**soft masking**" evita el ensemillado en rbc

BLAST: Basic Local Alignment Search Tool

• Extensión

Una vez que el espacio de búsqueda ha sido ensemillado, pueden generarse alineamientos pareados a partir de semillas individuales. La extensión acontece en ambas direcciones.



En el algoritmo de Smith-Waterman los puntos terminales de un aln. local son determinados después de haber evaluado todo el espacio de búsqueda. BLAST, al ser un algoritmo heurístico, tiene un mecanismo para no tener que explorar todo el espacio de búsqueda y sólo extiende una semilla hasta un determinado punto. Para ello se requiere de una variable X , que representa cuánto se permite caer al score del alineamiento después de haber pasado por un máximo. El algoritmo lleva la cuenta de los scores del alineamiento y de caída en base a la matriz de sustitución y de penalización de gaps

Ej. del control de extensión usando +1/-1 para match y mismatch respect., $X = 4$, (no gaps)

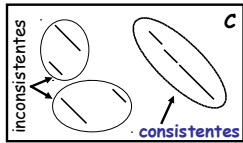
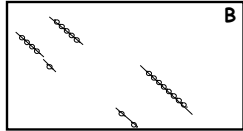
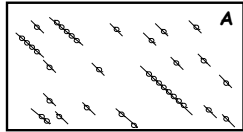
Pepito Pérez se fue a pescar al lago
Pepito López no vio a Arturo en casa

123456 54345 43 210 1 0 ... <- score aln.
000000 12321 23 456 5 6 ... <- score de caída

BLAST: Basic Local Alignment Search Tool

• Evaluación

Una vez extendidas las semillas, los alns. resultantes son evaluados para determinar si son estadísticamente significativos. Los que lo son se denominan HSPs (high scoring pairs)



Determinar la significancia de múltiples HSPs no es tan sencillo como sumar los scores de todos los alns. involucrados, ya que muchos corresponden a extensiones de palabras fortuitas, por lo que no todos los grupos de HSPs tienen sentido. Se define así un umbral de alineamiento (aln. threshold λT), basado en los scores de los alns. y que no considera por tanto el tamaño de la base de datos (BD). Cuanto más alto, menos alns. son considerados (Figs. A y B).

Idealmente la relación entre los HSPs debería de ser lo más parecida posible a alns. sin gaps globales, es decir, seguir las diagonales por la mayor distancia posible y no solaparse.

Grupos de HSPs que se comportan de esta manera se denominan grupos consistentes de HSPs (Fig. C). Para identificarlos, el algoritmo determina las coordenadas de todos los HSPs para cuantificar el solape. Este cálculo es cuadrático. Una vez organizados en grupos consistentes, se calcula un "final threshold" para cada grupo que considera todo el espacio de búsqueda (tamaño de la BD). BLAST reporta todos los que están por encima del E value de corte