

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

Introducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2022, 1-5 Agosto, 2022 CCG-UNAM, Cuernavaca, Mor. México  
<https://github.com/vinuesa/TIB-filoinfo>


**Introducción a la Filoinformática: Pan-genómica y Filogenómica microbiana– NNB & CCG-UNAM, 1-5 Agosto 2022**

Pablo Vinuesa ([vinuesa\[at\]ccg.unam.mx](mailto:vinuesa[at]ccg.unam.mx); [@pvinmex](https://twitter.com/pvinmex))  
Centro de Ciencias Genómicas, CCG-UNAM, Cuernavaca, México  
<http://www.ccg.unam.mx/~vinuesa/>

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

**• Tema 2: alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST**

- 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**



**Protocolo básico para un análisis filogenético de secuencias moleculares**

**Tema 2:**  
alineamientos pareados, búsquedas de homólogos en bases de datos

**Colección de secuencias homólogas**  
• BLAST, diamond

**Alineamiento múltiple de secuencias**  
• clustalo, mafft, 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 delección en marco

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

- Todas las mutaciones en 2<sup>as</sup> posiciones resultan en sustituciones **no** sinónimas
- 96% de mutaciones en 1<sup>as</sup> posiciones resultan en sustituciones no sinónimas
- Casi todas las sustituciones sinónimas ocurren en las 3<sup>as</sup> posiciones
- las delecciones 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  
 $\alpha_{A-C}$   
 $\beta_{A-C}$   
 $\alpha_{C-G}$   
Transiciones (ti) pirimidina - pirimidina  
 $\beta_{C-G}$   
 $\alpha_{G-T}$   
 $\beta_{G-T}$   
Transversiones (tv) pur. <-> pyr.

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

Licencia Creative Commons 4.0, no comercial  
Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)  
<https://creativecommons.org/licenses/by-nc/4.0/>

Alineamientos pareados y búsqueda de homólogos en bases de datos

Los alineamientos pareados son la base de lo 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**.

>gi|71548936|ref|PF\_00669120.1| Translation elongation factor G:Small GTP-binding protein domain [Nitrosomonas eutropha C71]  
gi|71486077|gb|EA018626.1| 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 MTREFSLEKTRNIGIMAHIDAGKTTTTERVLYTTPRIHKIGETHEGASQMDWMAQEQERG 60  
M++ LE+ RNIGIMAHIDAGKTTT+ER+L+YTG HK+GE H+GA+ MDWM QEQERG  
Sbjct 1 MSERNPLERYNIGIMAHIDAGKTTTTERILFTYTVSHKLGVEVHDGAATMDWMEQEQERG 60

Query 61 XXXXXXXXXXXXWN-----DHRINIIDTPGHVDFTVEVERSLRVLDGAVVLDAQSGVE 113  
ITITSAAIT W +HRIN+IDTPGHVDFT+EVERSLRVLDGA V + GV+  
Sbjct 61 ITITSAAITCFWKGAGNYFEHRINVIDTPGHVDFTIEVERSLRVLDGACTVFCVSGVGQV 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 EGGNAILNENISFSISPGQRVGLLGRVSGKSTLLSAFLRL-----NTEGEIQIDGVS 1273  
+ ++ +S ++ G+ + L+G +GSGKS +A L +L T GEI DG  
QAAQPLVHGVSLTLQGRVLAIVGGSGSKSLTCAATLGILPAGVRQTAGEILADGKP 70

P13569 1274 WDSITL-----QQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEV 1322  
L Q R AF + + + + + K AD+  
P33593 71 VSPCALRGIKIATIMQNPRAFNP-----HTMTHARETCLALGKPADDA 116

P13569 1323 GLRSVIEQFP-GKLDVFLVDGGCVLSHGKQMLCLARSVLSKAKILLDEPSAHLDPV 1379  
L + IE VL +S G Q M +A +VL ++ ++ DEP+ LD V  
P33593 117 TLTAATEAVGLENAARVLKLYPFEMSGGMLQRMIMAVLCESPFTIADEPTTDLDVV 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**.

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 MGDVERGKKIFIMKCSQCHTVEKGGKHKTFPNLHGLFGRKGTQAPGYSYTAANKNK---GI 58  
D KG+ +F QC T + K+ GP L G+ GRK G A G++Y+ N N G+  
P00090 1 Q-DAARKEAVF-----KQCMTCRADKNMVGPGALGGVGRKAGTAAGFTYSPLNHNHSGEAGL 56

P00001 59 IWGEDTLMEYLENPKKYIP-----GTKMIFVGIKKKEERADLIAYLKATNE 105  
+W ++ ++ YL +P Y+ TKM F + ++R D+ AYL AT +  
P00090 57 VWTQENIIAYLPDPNAYLKKFLTDKGQADKATGSKMTF-KLANDQQRKDVAAAYL--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 factores de penalización para gaps afines

• 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 para gaps afines** (*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.

penalización o costo del gap

penalización de gaps simple ( $w$ )

penalización de gaps afines

$w = g + hl$

longitud del gap

Licencia Creative Commons 4.0, no comercial Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) <https://creativecommons.org/licenses/by-nc/4.0/>

© Pablo Vinuesa 2022, @pvinmex vinuesa[at]ccg[dot]unam[dot]mx; <http://www.ccg.unam.mx/~vinuesa/>

### 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

- 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.

Matriz BLOSUM62

### Similitud entre pares de secuencias de AA

- **Matrices de sustitución de AAs log-odds scores (lod-scores)**

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

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

Matriz BLOSUM62

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

$f_a f_b$  = 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" o **puntuaciones crudas**.

### 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**  $S_{ab}$ . Para convertirlos a un **score normalizado** (log-odd score original) tenemos que multiplicarlos por  $\lambda$ , una constante específica para cada matriz.  $\lambda$  es ~ igual al inverso del factor de escalamiento (c ).

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

por tanto, para despejar  $\lambda$  necesitamos  $f_a f_b$  y encontrar el valor de  $\lambda$  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  $\lambda$ , se usa para calcular el **valor de expectation (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  $\lambda$ "** que es **"hit-específica"**

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 ( $\lambda S$ )** 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

NCBI-BLAST: Basic Local Alignment Search Tool

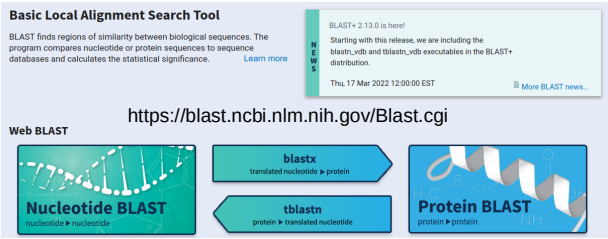
BLAST consta de una **familia de programas**, los cuales seleccionaremos en función de:

1. el tipo de secuencia problema (nt | p) [nucleótido | proteína]
2. el tipo de secuencia de la base de datos ainterrogar (nt | p)

Los 5 principales 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**, diseñados para aumentar sensibilidad



BLAST: Basic Local Alignment Search Tool

• Anatomía de un reporte de NCBI-BLAST estándar

BLASTP 2.2.13 [Nov-27-2005]

**Reference:**

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, 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.

RID: 1141782136-12667-92041342765.BLA8TQ4

**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

1

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

Request ID

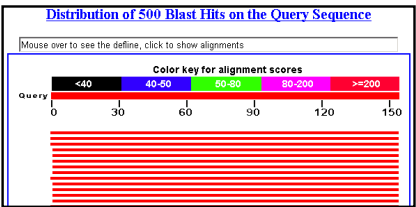
Indica la **BD** sobre la que se hizo la búsqueda, junto con el no. de secs contenida en ella y el no. de caracteres

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	
<a href="#">gi 4885477 ref NP_005359.1 </a> myoglobin [Homo sapiens] >gi 4495...	<a href="#">316</a>	6e-86	<a href="#">Gene Info</a>
<a href="#">gi 62511907 gb AA84516.1 </a> myoglobin transcript variant 1 [Homo	<a href="#">315</a>	1e-85	
<a href="#">gi 386872 gb AA59595.1 </a> myoglobin	<a href="#">315</a>	1e-85	
<a href="#">gi 229361 prf I716588</a> myoglobin	<a href="#">313</a>	4e-85	
<a href="#">gi 127683 sp P02145 MYG_PANTR</a> Myoglobin	<a href="#">312</a>	9e-85	<a href="#">Structures</a>
<a href="#">gi 51317141 sp P62735 MYG_HYLBY</a> Myoglobin >gi 51317413 sp P62734	<a href="#">311</a>	1e-84	
<a href="#">gi 127656 sp P02147 MYG_GORER</a> Myoglobin	<a href="#">311</a>	2e-84	
<a href="#">gi 229360 prf I716588</a> myoglobin	<a href="#">311</a>	2e-84	
<a href="#">gi 55728442 emb CAH90965.1 </a> hypothetical protein [Pongo pygmaeus	<a href="#">310</a>	5e-84	
<a href="#">gi 230638 pdb 2MM1 </a> Myoglobin Mutant With Lys 45 Replaced By...	<a href="#">309</a>	6e-84	
<a href="#">gi 127689 sp P02148 MYG_PONFY</a> Myoglobin >gi 229570 prf I761377A	<a href="#">308</a>	2e-83	
<a href="#">gi 62901707 sp P68086 MYG_BRYPA</a> Myoglobin >gi 62901706 sp P68...	<a href="#">300</a>	4e-81	

**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|47523546|ref|NP_999401.1| myoglobin [Sus scrofa]
gi|127688|sp|P02189|MYG_PIG Myoglobin
gi|164547|gb|AAA31073.1| myoglobin
length=154      normalized score
                  raw score
Score = 296 bits (758), Expect = 5e-80
Identities = 144/154 (93%), Positives = 148/154 (96%), Gaps = 0/154 (0%)

Query 1      MGLSDGEWQLVLNVWGKVEADIPGHGQEVLRIRLFKGHPTLEKFDKFKHLKSEDEMKASE 60
            MGLSDGEWQLVLNVWGKVEAD+ GHGQEVLRIRLFKGHPTLEKFDKFKHLKSEDEMKASE 60
Sbjct 1      MGLSDGEWQLVLNVWGKVEADVAGHGQEVLRIRLFKGHPTLEKFDKFKHLKSEDEMKASE 60


Query 61     DLKKHGATVLTALGGSILKKKGHHEAEIKPLAQSHATKKKIPVKYLEFISECIIQVLOSKEH 120
            DLKKHG TVLTALGGSILKKKGHHEAE+ PLAQSHATKKKIPVKYLEFISE IIOVLOSKEH 120
Sbjct 61     DLKKHGNVLTALGGSILKKKGHHEAELTPLAQSHATKKKIPVKYLEFISEAIIOVLOSKEH 120

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



Centro de Ciencias Genómicas

**Genómica Evolutiva I**  
**LCG-UNAM,**  
**Semestre 2022-1**



**Pablo Vinuesa** ([vinuesa@ccg.unam.mx](mailto:vinuesa@ccg.unam.mx))  
Centro de Ciencias Genómicas UNAM  
<http://www.ccg.unam.mx/~vinuesa/>

**Mini-tutorial de uso de [BLAST y] BLAST+ desde la línea de comandos:**

1. Generación de bases de datos (indexadas) mediante [formatdb y] makeblastdb
2. Interrogación de bases de datos mediante [blastall -p [blastn|blastp|blastx|tblastn|tblastx] y] blastn, blastp, blastx, delta-blast ...
3. Recuperación de secuencias de una base de datos usando Id's y fastacmd o blastdbcmd

Documentación de BLAST+ en NCBI  
<http://www.ncbi.nlm.nih.gov/books/NBK1762/>  
<http://www.ncbi.nlm.nih.gov/books/NBK52640/>  
<http://www.ncbi.nlm.nih.gov/books/NBK279690/>

**FORMATEO DE ARCHIVOS FASTA PARA  
GENERAR BASES DE DATOS INTERROGABLES CON BLAST+**

- BLAST usa **bases de datos indexadas** para acelerar la operación de búsqueda.
- Existen diversas bases de datos pre-compiladas y formateadas. La más general y extensa es la "nr" o no-redundante. Hay muchas más como: est, wgs, pat, pdb, microbial genomes o env\_nt.
- También es posible generar bases de datos propias usando el programa **formatdb** o **makeblastdb**. Descárgalo desde <ftp://ftp.ncbi.nih.gov/blast/> junto con los demás binarios de la suite de programas BLAST+. [en ubuntu: apt-get install ncbi-blast+ (blast2 es legacy-blast)]
- Para generar una base de datos se utilizan secuencias en formato FASTA, y con una **sintaxis de identificador NCBI canónica**. Por ejemplo:

Icd|integer  
Icd|string  
**gnl|yourDB|ID**

}

estos son los formatos de las cabeceras FASTA para generar bases de datos de secuencias localmente.  
Puedes ver más ejemplos aquí:

[http://ncbi.github.io/cxx-toolkit/pages/ch\\_demo#ch\\_demo.id1\\_fetch.html\\_ref\\_fasta](http://ncbi.github.io/cxx-toolkit/pages/ch_demo#ch_demo.id1_fetch.html_ref_fasta)

Este **identificador es esencial para un correcto indexado de la BD** y para así poder, por ejemplo, recuperar secuencias de la BD usando listas de identificadores.



### Introducción a BLAST+ desde la línea de comandos

**REFERENCIAS CLAVES:**  
1: Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WF, McGinnis SD, Merezuk Y, Raytselis Y, Sayers EW, Tao T, Ye J, Zaretskaya I.BLAST: a more efficient report with usability improvements. Nucleic Acids Res.2013 Jul;41(Web Server issue):W29-33. doi: 10.1093/nar/gkt282. Epub 2013 Apr 22. PubMed PMID: 23609542; PubMed Central PMCID: PMC3692093.  
2: Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PubMed Central PMCID: PMC2803857.

Para correr un programa de la suite BLAST+ necesitamos esencialmente 2 cosas:  
1. Una base de datos a interrogar, adecuadamente formateada  
2. Una o más secuencias problema con las que buscaremos homólogos en la base de datos llamando a los programas adecuados en función del tipo de secuencia problema (DNA o proteína) y de la base de datos.

**Ejemplos de uso de programas de la suite de programas BLAST+**  
**# 1) formateo de la base de datos**  
**makeblastdb -in** sequences4blastdb.fna **-dbtype** nucl **-parse\_seqids**  
**# 2) ejecutamos una búsqueda con blastn sobre la base de datos recién formateada**  
**blastn -query** query\_seqs.fas **-db** sequences4blastdb.fna **-out** 16S\_out.tab **-outfmt** 6 **-max\_target\_seqs** 1  
**# 3) recuperamos los hits usando blastdbcmd**  
**blastdbcmd -db** sequences4blastdb.fna **-entry** my\_hits.list

### Introducción a BLAST+ desde la línea de comandos

**Ayuda desde la línea de comandos:**  
**1. Ayuda en formato condensado:**  
Programa -h (por ejemplo: blastn -h)  
**2. Ayuda detallada**  
Programa -help (por ejemplo: blastp -help)

Conviene revisar además el **BLAST Command Line Applications User Manual** en: <http://www.ncbi.nlm.nih.gov/books/NBK1763/>  
Sigue un resumen de algunos comandos básicos y sus opciones, comparando el "blast viejo o legacy blast" con blast+ (actual)

BLAST	BLAST+	Descripción
formatdb	makeblastdb	
-i	-in	Archivo de entrada con secuencias
-p T/F	-dbtype prot/nucl	Mol type
-o T	-parse_seqids	Parsea e indexa seq IDs
-n	-out	Nombre de base para archivos de salida

### BLAST+ - el nuevo BLAST escrito en C++

Continuación (ver blast[np...]-h para despliegue de opciones)

BLAST	BLAST+	Descripción
blastall	blastn, blastp,...	
-p	No existe	blastn, blastp, blastx, tblastn, ...
-i	-query	Archivo de entrada
-d	-db	Base de datos de blast
-o	-out	Nobre de archivos de salida
-m	-outfmt	Formato salida; TAB: 6 == m 8
-e	-eval	Punto de corte para valor de Expectancia
-v	-num_descriptions	Máximo número de descripciones - hits
-b	-num_alignments	Número máximo de alineamientos
-a	-num_threads	No. de cores a usar
	-max_target_seqs	No. max. de secuencias y descripciones
-F F	-dust no   -seg no	Deshabilitar filtrado de regiones de baja complejidad; DNA:dust AA:seg

### BLAST+ - el nuevo BLAST escrito en C++

BLAST	BLAST+	Descripción
fastacmd	blastdbcmd	
-d	-db	Base de datos de blast
-s	-entry	Cadena de búsqueda
-D 1	-entry all	DB dump en formato FASTA

**Ejemplos de uso de programas de la suite de programas BLAST+**  
**# 1) formateo de la base de datos**  
**makeblastdb -in** sequences4blastdb.fna **-dbtype** nucl **-parse\_seqids**  
**# 2) ejecutamos una búsqueda con blastn sobre la base de datos recién formateada**  
**blastn -query** query\_seqs.fas **-db** sequences4blastdb.fna **-out** 16S\_out.tab **-outfmt** 6 **-max\_target\_seqs** 1  
**# 3) recuperamos los hits usando blastdbcmd**  
**blastdbcmd -db** sequences4blastdb.fna **-entry** my\_hits.list

**Ya es hora de hacer unos ejercicios con datos reales...**

Ejercicios: formato de bases de datos de nt y aa con blastdb y búsquedas locales con blastall

- I. Formateo de base de datos de secuencias 16S de *Mycobacterium* spp. y búsqueda en ella de homólogos mediante blastn
  - 1) Descargar el archivo [16S\\_4blastN.tgz](#) de la página del curso
  - 2) Descomprimirlo y abrir el tarro con: **tar -xvzf 16S\_4blastN.tgz**
  - 3) Construiremos la base de datos con las secuencias disponibles en el archivo 16S\_seqs4\_blastDB.fna. Primero que nada averigüen:
    - 3.1 ¿cuántas secuencias tiene; cuántas especies representa?
    - 3.2 ¿qué información contienen los identificadores (el *fasta header*) ?
    - 3.3 ¿es su formato adecuado para un indexado correcto?Usa la línea de comandos para dar respuesta a estas preguntas
  - 1) ¿Qué línea de comando usarías para un generar una base de datos con el archivo 16S\_seqs4\_blastDB.fna para que esté indexado?
  - 1) ¿Cómo clasificarías las secuencias contenidas en el archivo 16S\_problema.fna ?
  - 2) Recupera los 10 hits más próximos a cada secuencia problema de la base de datos para su posterior alineamiento; filtra aquellos hits con  $\geq 98.5\%$  de identidad

Ejercicios: continuación

- II. Formateo de base de datos de secuencias de integrones bacterianos y descubrimiento y anotación de genes (cassettes) amplificados de cepas de *E. coli* recuperadas por Jazmín Madrigal del río Apatlaco, Mor. México.
  - 1) Descargar el archivo [gene\\_discovery\\_and\\_annotation\\_using\\_blastx.tgz](#) de la página
  - 2) Descomprimirlo y abrir el tarro con:  
**tar -xvzf gene\_discovery\_and\_annotation\_using\_blastx.tgz**
  - 1) Construiremos la base de datos con las secuencias disponibles en el archivo `integron_cassettes4blastdb.faa`. Primero que nada averigüen:
    - 3.1 ¿cuántas secuencias tiene; cuántas especies representa?
    - 3.2 ¿qué información contienen los identificadores (el *fasta header*) ?
    - 3.3 ¿es su formato adecuado para un indexado correcto?Usa la línea de comandos (shell) para dar respuesta a estas preguntas
  - 1) ¿Qué comando usarías para un generar una base de datos con el archivo `*4blastdb.faa` para que esté indexado?
  - 1) ¿Qué comandos usarías para identificar y anotar los genes que pudieran estar codificados en las secuencias contenidas en el archivo `3cass_amplicons.fna`?
  - 6) Recupera los 10 hits más próximos a cada secuencia problema de la base de datos para su posterior alineamiento.

Campos del formato tabular -m 6 (-m 8 legacy) de NCBI-BLAST

- Como ya vimos, la [opción -m 8](#) de blastall especifica una salida en formato tabular, con los campos separados por tabuladores.
- Estos datos (líneas) se pueden parsear fácilmente usando Perl o comandos de UNIX como:  

```
# imprime sólo hits con %ID > 95% y aln_len > 500
perl -ane '{ print "$F[0]\tF[1]" if $F[2] > 95.0 && $F[3] > 500 }' blast_m8.out
# obtén una lista no redundante de hits
cut -f2 blast_output.txt | sort -u
```
- Los campos o columnas son las siguientes: (-m 9 los imprime como comentario)
  - 0: query name
  - 1: subject name
  - 2: percent identities
  - 3: alignment length
  - 4: number of mismatched positions
  - 5: number of gap positions
  - 6: query sequence start
  - 7: query sequence end
  - 8: subject sequence start
  - 9: subject sequence end
  - 10: e-value
  - 11: bit score