Introducción a la Filoinformática: LEG-UM5, Rabat. 10-14 Junio 2019.

Pablo Vinuesa (vinuesa@ccg.unam.mx)

Progama de Ingeniería Genómica, CCG-UNAM, México
http://www.ccg.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**

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...

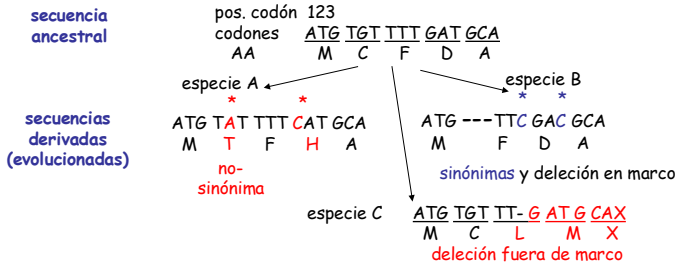
Análisis evolutivo del alineamiento y selección del modelo de sustitución más ajustado
⇩ · tests de saturación, modeltest, ...

Estima filogenética
⇩ · NJ, ME, MP, ML, Bayes ...

Pruebas de confiabilidad de la topología inferida
⇩ · proporciones de bootstrap probabilidad posterior ...

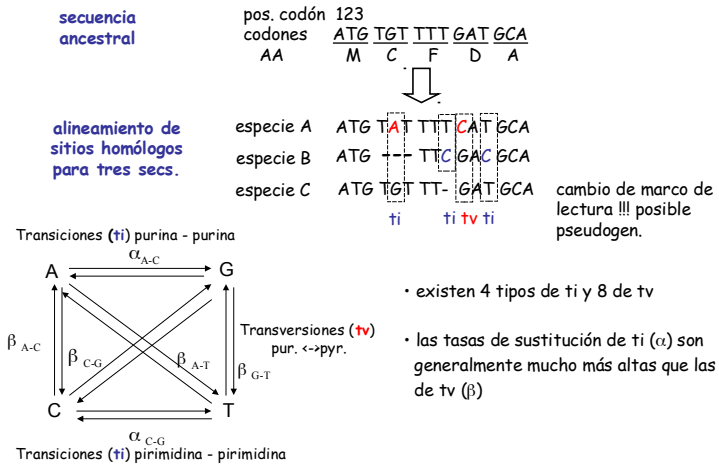
Interpretación evolutiva y aplicación de las filogenias

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



- Todas las mutaciones en 2as posiciones resultan en sustituciones no sinónimas
- 96% de mutaciones en 1as posiciones resultan en sustituciones no sinónimas
- Casi todas las sustituciones sinónimas ocurren en las 3as 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



- 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)

Alineamientos pareados y búsqueda de homólogos en bases de datos

Los alineamientos pareados son la base de lo métodos de búqueda 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 (cenancestro).
- 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|71548896|ref|ZP_00669120.1|  Translation elongation factor G:Small GTP-binding protein domain
[Nitrosomonas eutropha C71]
 gi|71486077|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    MTREFSLEKTRNIGIMAHIDAGKTTTERVLYYTGRIHKIGETHEGASQMDWMAQEQERG  60
            M++   LE+ RNIGIMAHIDAGKTTT+ER+L+YTG  HK+GE H+GA+ MDWM QEQERG
Sbjct  1    MSKRNPLERYRNIGIMAHIDAGKTTTSERILFYTGVSHKLGEVHDGAATMDWMEQEQERG  60

Query  61   XXXXXXXXXXXWN-------DHRINIIDTPGHVDFTVEVERSLRVLDGAVAVLDAQSGVE  113
            ITITSAATT  W        +HRIN+IDTPGHVDFT+EVERSLRVLDGA  V  +  GV+
Sbjct  61   ITITSAATTCFWKGMAGNYPEHRINVIDTPGHVDFTIEVERSLRVLDGACTVFCSVGGVQ  120
```

(... truncado)

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  MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNK---GI  58
             D  KG+ +F     QC T  +  K+  GP L G+ GRK G A G++Y+  N N    G+
P00090    1  Q-DAAKGEAVF----KQCMTCHRADKNMVGPALGGVVGRKAGTAAGFTYSPLNHNSGEAGL  56

P00001   59  IWGEDTLMEYLENPKKYIP---------------GTKMIFVGIKKKEERADLIAYLKKATNE  105
            +W ++ ++ YL +P  Y+                 TKM F  +   ++R D+ AYL  AT +
P00090   57  VWTQENIIAYLPDPNAYLKKFLTDKGQADKATGSTKMTF-KLANDQQRKDVAAYL--ATLK  114
```

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

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 discontínuos intrones-exones; barajado de exones** ...).
BLAST y FASTA buscan alineamientos locales con alta puntuacion (**HSPs** ó high-scoring pairs)

(b)

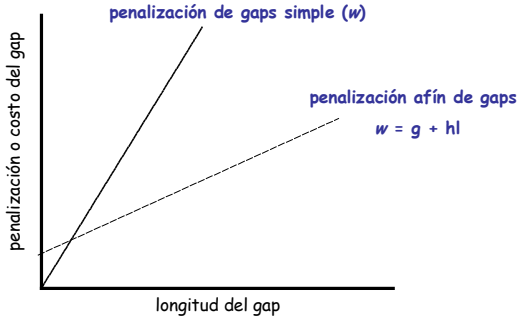
```
P13569   1221  EGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLL-----NTEGEIQIDGVS  1273
+     ++  +S ++  G+ + L+G +GSGKS   +A L +L       T GEI  DG
P33593     13  QAAQPLVHGVSLTLQRGRVLALVGGSGSGKSLTCAATLGILPAGVRQTAGEILADGKP    70

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

P13569   1323  GLRSVIEQFP-GKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPV  1379
 L + IE        VL       +S G   Q M +A +VL ++  ++ DEP+  LD V
P33593    117  TLTAAIEAVGLENAARVLKLYPFEMSGGMLQRMMIAMAVLCESPFIIADEPTTDLDVV   174
```

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

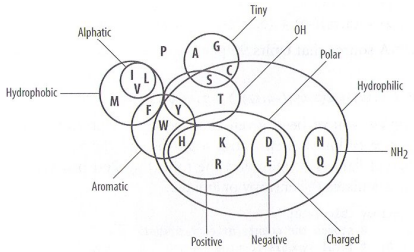
alineamientos pareados y factores de penalización afines para gaps

- 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.

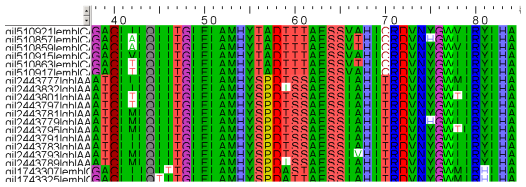


vinuesa[at]ccg[dot]unam[dot]mx;
http://www.ccg.unam.mx/~vinuesa/

Similitud entre pares de secuencias de AA

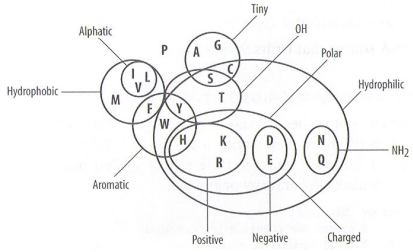


- Este diagrama muestra a los **aminoácidos** agrupados atendiendo a sus características químicas y físicas.
- Desde una perspectiva evolutiva **esperamos encontrar más sustituciones entre aas. similares que entre los menos relacionados.**
- Estos patrones puede observarse en alineamientos múltiples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

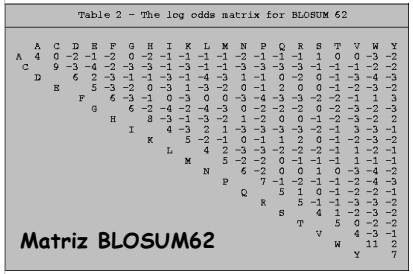


Segmento de un aln. múltiple de citocromos *C* de primates

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 púramente 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.

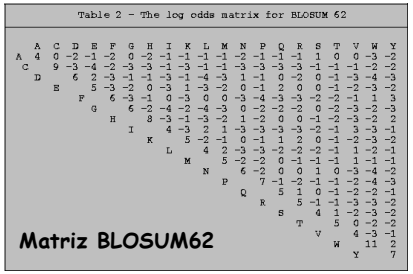
Es importante notar que los scores son evolutivamente simétricos al no conocerse la dirección del cambio evolutivo.

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



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

$f_a f_b$ = 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".

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 mutiplicarlos por λ, una constante específica para cada matriz.
λ es aprox. igual al inverso del factor de escalamiento (c).

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

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

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 resíduos de algunas proteínas difieren mucho de las frecuencias de resíduos empleadas para calcular las matrices PAM y BLOSUM, versiones recientes de BLASTP y PSI-BLAST incorporan una "**composition-based** λ" 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** (*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

BLAST: Basic Local Alignment Search Tool

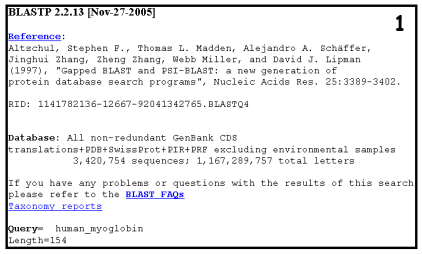
BLAST consta de una **familia de programas**. Los 5 ppales son:

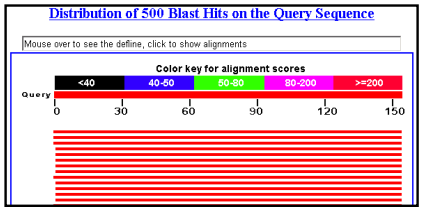
BLASTN (nt-nt), **BLASTP** (p-p), **BLASTX** (translated nt-p),
TBLASTN (p-translated nt), usado en mapeo de prots contra DNA genómico
TBLASTX (translated nt - translated nt) usado en la predicción de genes

y **variantes de BLASTP** como **PSI-** y **PHI-BLAST**

BLAST: Basic Local Alignment Search Tool

- **Anatomía de un reporte de NCBI-BLAST estándar**





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**.

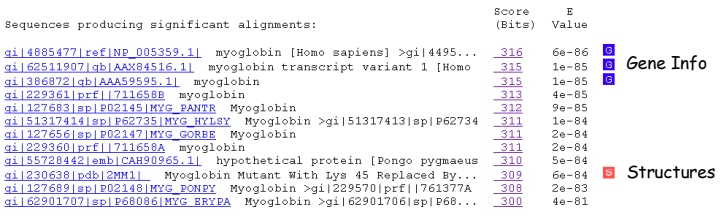
escala de color que indica el score de los HSPs

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 linea**. 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



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|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    MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE  60
            MGLSDGEWQLVLNVWGKVEAD+ GHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
Sbjct  1    MGLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE  60

Query  61   DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH  120
            DLKKHG TVLTALGGILKKKGHHEAE+ PLAQSHATKHKIPVKYLEFISE IIQVLQSKH
Sbjct  61   DLKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH  120

Query  121  PGDFGADAQGAMNKALELFRKDMASNYKELGFQG  154
            PGDFGADAQGAM+KALELFR DMA+ YKELGFQG
Sbjct  121  PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG  154
```

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/missmatch) y **gap penalties** empleados

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

```
  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                                     <- matriz de sustitución
Gap Penalties: Existence: 11, Extension: 1           <- gap penalties
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              <- E value umbral usado = 10; HSPs con gap
Number of HSP's better than 10 without gapping: 0    <- E value umbral usado = 10; HSPs no gap
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                                                <- neighborhood word threshold score
A: 40                                                <- two-hit distance
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)                                   <- extension attenuation parameter
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)                                   <- aln. threshold (ungapped)
S2: 66 (30.0 bits)                                   <- aln. threshold (gapped)
```

BLAST: Basic Local Alignment Search Tool

- **RESUMEN de gapped-BLAST**

- BLAST es un progrma 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 exprolada, 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*

vinuesa[at]ccg[dot]unam[dot]mx;
http://www.ccg.unam.mx/~vinuesa/

CCG Centro de Ciencias Genómicas

Genómica Evolutiva I

LCG-UNAM,

Semestre 2018-1

Pablo Vinuesa (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/

BASES DE DATOS PARA NCBI-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.
- Tién 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:

lcl|integer
lcl|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

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.

Uso de *formatdb* para generar bases de datos para NCBI-BLAST

- Por defecto, **formatdb** produce 3 archivos con el mismo nombre de base y con las extensiones base.nhr, base.nsq, y base.nin. Estos son archivos binarios, y en este caso se trata de bases de datos de secuencias de nucleótidos ya que la primera letra de la ext. comienza con n (p para proteína).
- Los tres parámetros más usados para correr **formatdb** son:
  - **-i** input data file (contiene una o más secuencias en formato FASTA)
  - **-n** output file base name (if this parameter is not set, the input file name is used as base)
  - **-p** type of file: T for protein, F for nucleic acid (True|False)
- La opción **-o** produce otro conjunto de archivos requeridos para el indexado. Esta opción es esencial si se van a formatear bases de datos grandes.
- La syntaxis básica para formatear una base de datos es:

```
formatdb -i mis_ESTs.fna -p F -n mis_ESTs_db -o T # para nucleótidos
formatdb -i mis_PROTs.faa -p T -n mis_PROTs_db -o T # para proteínas
```

el primer comando toma un archivo FASTA de nucleótidos y crea los 3 archivos de base de datos: mis_ESTs_db.nhr, mis_ESTs_db.nsq, y mis_ESTs_db.nin
y 2 archivos de indexado:
mis_ESTs_db.nsd, mis_ESTs_db.nsi

- Las opciones (ayuda) de formatdb se llaman así: **formatdb --help**

Uso de *blastall* desde la línea de comandos

- Los programas blastn blastp blastx tblastn y tblastx se especifican como parámetro del comando **blastall** en el "BLAST antiguo" (legacy BLAST).
- Las opciones básicas y esenciales son:
  - **-p** programa a ejecutar (blastn, blastp, blastx, tblastn, tblastx)
  - **-d** base de datos sobre la cual buscar homólogos (creada con formatdb)
  - **-i** input sequence file (una o más secuencias en formato FASTA)
  - **-o** output file name # si lo prefieren puden redirigir la salida > outputfile
- Ejemplos de sintaxis básica serían:

a) **un análisis de blastn**

```
blastall -p blastn -i my_query_file -d my_database -o \
  my_blast_output.txt
```

b) **un análisis de blastp**

```
blastall -p blastp -i my_query_file -d my_database -m 8 \
        -b 10 > my_blast_output.txt
```

Otras opciones muy útiles de *blastall*

- **-e [valor de expectancia de corte]**. El valor por defecto es 10.0. Este número tiene que especificarse en notación decimal, no exponencial, por ejemplo -e 0.001
- **-F** filter query sequence. Por defecto esta opción implementa el filtro " DUST" que enmascara regiones de baja complejidad
- **-m 8** produce formato tabular de salida, muy útil para grandes conjuntos de datos
- **-b [número]** trunca el reporte a un máximo de [número] alineamientos

- **-M protein substitution matrix**. La matriz por defecto es BLOSUM62. Se pueden especificar: BLOSUM45, BLOSUM80, PAM30 y PAM70.

- El resto de las opciones pueden consultarse tecleando "`blastall`" sin más argumentos en la línea de comandos

Campos del formato tabular –m 8|9 de NCBI-BLAST

- Como ya vimos, la **opción –m 8|9** 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

Recuperar secuencias de una base de datos usando fastacmd

- Para recuperar las secuencias especificadas en una lista de GIs a partir de una base de datos, se usa el comando **fastacmd** usando la siguiente sintaxis:

```
fastacmd -d mis_ESTs.fna -s AU108953,AU108955 -l 80
                    ó
fastacmd -d mis_ESTs.fna -i archivo_con_GIs_a_recuperar -l 80
                    ó
fastacmd -d mis_ESTs.fna -D
```

donde **-d** designa la base de datos, **-s** la cadena de identificador de secuencia a recuperar, y **-l** el no. de caracteres por línea de secuencia. Alternativamente podemos recuperar una serie de secuencias, cuyos IDs vienen especificados en un archivo (uno por línea) que se para como parámetro a la opción **-i**.

La opción **-D** hace un "*dump*" o vertido de toda la base de datos.

BLAST+ - el nuevo BLAST escrito en C++

REFERENCIAS CLAVES:
1: Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk 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.

Conviene revisar además el **BLAST Command Line Applications User Manual**

en: http://www.ncbi.nlm.nih.gov/books/NBK1763/

Aquí sólo va un resumen de algunos comandos básicos, comparando blast con blast+:

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

vinuesa[at]ccg[dot]unam[dot]mx;
http://www.ccg.unam.mx/~vinuesa/

BLAST+ - el nuevo BLAST escrito en C++

Continuación (ver blast[npx...] -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	-evalue	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** secuences4blastdb.fna **-out** 16S_out.tab **-outfmt** 6 \ **-max_target_seqs** 1

 # 3) recuperamos los hits usando blastdbcmd
blastdbcmd -db secuences4blastdb.fna **-entry** my_hits.list

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

Ejercicios: formateo 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 ¿ cuantas secuencias tiene; cuantas especies representa?
   3.2 ¿qué información contienen los idenditificadores (el *fasta header)* ?
   3.3 ¿ es su formato adecuado para un indexado correcto?

   Usa la línea de comandos para dar respuesta a estar 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 >= 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; cuantas especies representa?
   3.2 ¿qué información contienen los idenditificadores (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.

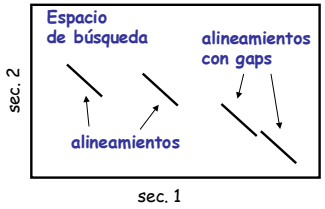
vinuesa[at]ccg[dot]unam[dot]mx;
http://www.ccg.unam.mx/~vinuesa/

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 = sumatoria de scores individuales p_{ab} – costo gaps.

BLAST no explora todo el espacio de búsqueda entre dos secuencias (es un heurístico).



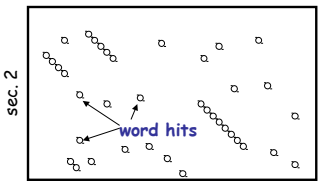
BLAST reporta todos los alns. pareados (HSPs) estadísticamente significativos encontrados en su búsqueda heurística del espacio de búsqueda. Hay que entender que **en las búsquedas BLAST siempre hay que hacer un compromiso entre velocidad** y **sensibilidad.** La velocidad se gana al no explorar toda la matriz, perdiéndose sensibilidad.

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**

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.



MPRDG { MPR, PRD, RDG } secuencia y palabras de 3 letras

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.

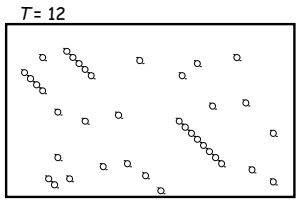
Vecinos aceptados de RDG serían:

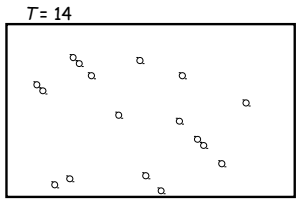
Palabra	Score (Blosum62)
RDG	17
KGD	14
QGD	13
RGE	13
EGD	12
...	

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.





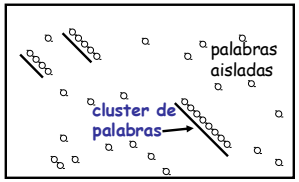
El **tamaño de palabra** *W* es otro parámetro que controla el número de word hits. *W*=1 producirá más hits que *W* = 5. Cuanto más chico sea *W* más sensible y lenta la búsqueda.
La interrelación entre *W*, *T* y la matriz de sustitución empleada es crítica, y su selección juiciosa es la mejor manera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

vinuesa[at]ccg[dot]unam[dot]mx;
http://www.ccg.unam.mx/~vinuesa/

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*, par 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.

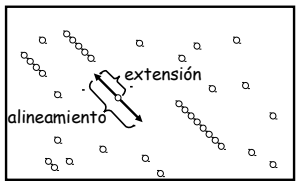
2. En **BLASTP** (y otros programas basados en aa) se usan valores de *W* de 2 ó 3. Para hacer las búsquedas más rápidas *W* = 3 y *T* = 999, que elimina todas las palabras vecinas. La distancia (*A*) entre vecinos del algoritmo two-hit es por defecto = 40 aas.

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

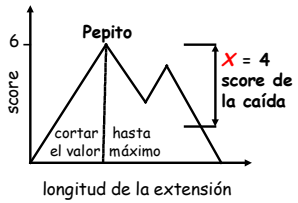
BLAST: Basic Local Alignment Search Tool

• **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



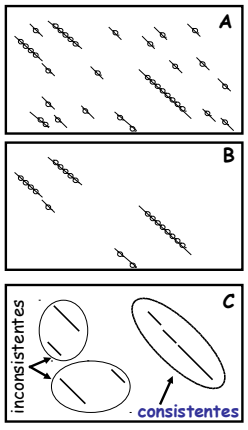
Ej. del control de extensión usando +1/-1 para match y mismatch respect., ***X*** = **4**, (no gaps)

```
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 *AT***), 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, segir 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**

vinuesa[at]ccg[dot]unam[dot]mx;
http://www.ccg.unam.mx/~vinuesa/