

**Curso fundamental de Inferencia Filogenética Molecular**



Pablo Vinuesa ([vinuesa@ccg.unam.mx](mailto:vinuesa@ccg.unam.mx))  
 Programa de Ingeniería Genómica, CCG, UNAM



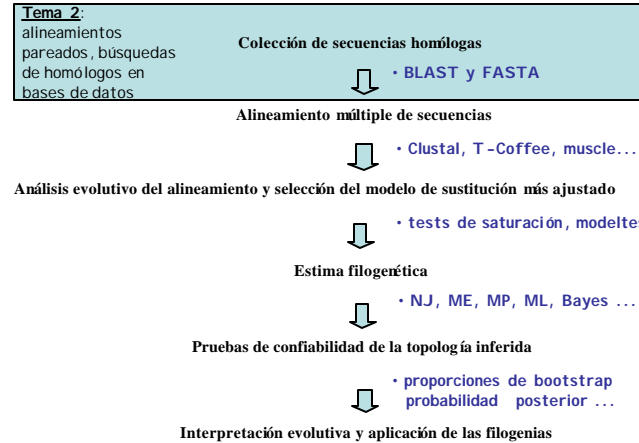
<http://www.ccg.unam.mx/~vinuesa/>

Tutor: PDCBM, Ciencias Biológicas, PDCBioq. y Profesor de la Lic. Ciencias Genómicas y posgrado

**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);
- **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**



**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**.

```
>F|g17548895|ccf1P_00669120.1| Translation elongation factor G:Small GTP-binding protein domain
[Mitrosomonas autotropha C72]
g113486711|g1P018536.21 Translation elongation factor G:Small GTP-binding protein domain
[Mitrosomonas autotropha C72]
length=69c

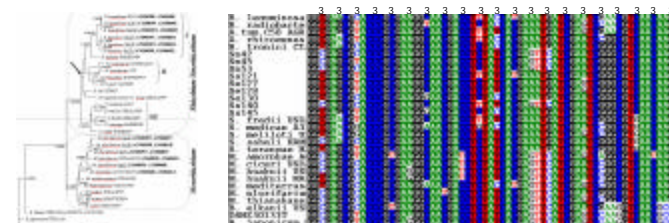
Score = 828 bits (2140), Expect = 0.0
Identities = 454/697 (65%), Positives = 541/697 (77%), Gaps = 3/697 (1%)

Query 1: MYREFFLEKTRNSIGDNRHIDAGKTTTENVLTFPRINKIGETHSDAQDMDNQEQEER 60
M++ LE+ NSIGDNRHIDAGKTTT+ER+L+TQ HE+SE H+GL+ MQAM QEQEER
Sbjct 1: MEERNPLEKTRNSIGDNRHIDAGKTTT+ER+L+TQ HE+SE H+GL+ MQAM QEQEER 60

Query 61: XXXXXXXXXXXXXN-----DNRINIIDTGHVDTVEVERSLRVLDGAVVLDGQGGVE 115
ITITSAAAT W +RRIN+IDTGHVDT+EVERSLRVLDGA Y + GV+
Sbjct 61: ITITSAAATPFRWNRHNTSHRINIVLDTGHVDTIEVERSLRVLDGACTYFCVVDVQY 120 (... truncado)
```

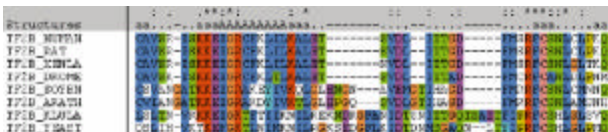
**Homología entre secuencias de DNA y proteína: conceptos y terminología básica**

- A lo largo de la evolución las secuencias descendientes de otra ancestral van acumulando diversos tipos de mutaciones. Estas son **mutaciones puntuales** o **reorganizaciones genómicas**, que pueden involucrar **inserciones**, **deleciones**, **inversiones**, **translocaciones** o **duplicaciones**, mediados por distintos mecanismos de recombinación (homóloga e ilegítima)
- Cualquier análisis filogenético y/o evolutivo de secuencias moleculares requiere de un **alineamiento** para poder comparar sitios homólogos entre las secuencias a estudiar. Para ello se escriben las secuencias en filas una sobre la otra, de modo que los sitios homólogos quedan alineados por columnas. Cada sitio o columna del alineamiento corresponde a un **carácter**, y los nt o aa que ocupan dichas posiciones representan los distintos **estados del carácter**



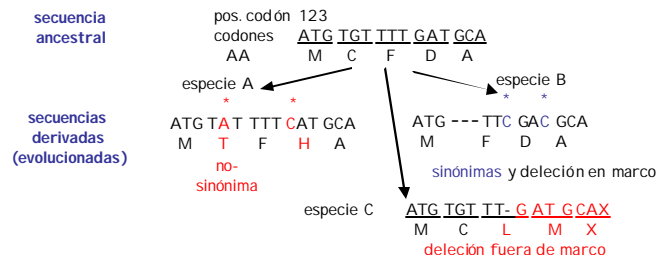
### Homología entre secuencias de DNA y proteína: conceptos y terminología básica

- Cuando por eventos de inserción o deleción (**indeles**) las secuencias homólogas presentan distintas longitudes, es necesario introducir "gaps" en el alineamiento para mantener la correspondencia entre sitios homólogos situados antes y después de las regiones afectadas por indeles. Estas regiones se identifican mediante guiones (-). **Los indeles no se distribuyen aleatoriamente en las secuencias codificadoras.** Casi siempre aparecen ubicados entre dominios funcionales o estructurales, preferentemente en bucles (loops) que conectan a dichos dominios. Esto vale tanto para RNAs estructurales (tRNAs y rRNAs) como para proteínas. No suelen interrumpir el marco de lectura.



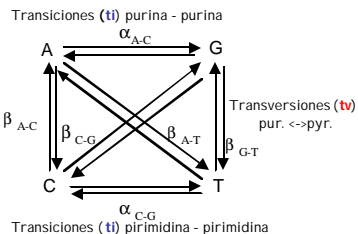
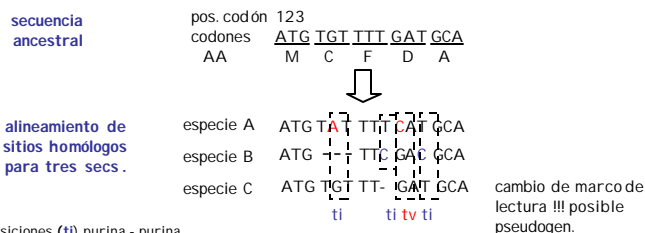
- A mayor **distancia genética** (evolutiva) entre un par de secuencias, mayor será el número de mutaciones acumuladas. Dependiendo del tiempo de separación de los linajes y la tasa evolutiva del locus, puede llegar a ser imposible alinear ciertas regiones debido a fenómenos de **saturación mutacional**. Las regiones de homología dudosa deben de ser excluidas de un análisis filogenético

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



- 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 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 ( $\alpha$ ) son generalmente mucho más altas que las de tv ( $\beta$ )

### Programación dinámica: algoritmo de Needleman-Wunsch y alineamientos pareados globales

Un valor de puntuación es escogido para cada tipo de sustitución (par de residuos o aln. de residuo contra un gap). El set completo de estas puntuaciones conforman una matriz de ponderaciones o puntuaciones (**scoring matrix**), de dimensiones  $S(i, j)$

Existen muchas definiciones del score de un alineamiento, pero la más común es simplemente la suma de scores o puntuaciones para cada par de letras alineadas y pares letra-gap, que conforman el alineamiento.

Así, para la matriz de sustitución siguiente y un  $w$  lineal de 5, calcula la puntuación del siguiente alineamiento

-	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

**AGACTAGTTAC**  
**CGA---GACGT**  
 Score =  $-3+7+10-3 \times 5 + 7-4+0-1+0 = 1$

**Programación dinámica y la generación de alineamientos pareados (globales y locales): dot plots y visualización de la similitud entre secuencias**

- las 2 secs. representan los dos ejes de la gráfica
- se pone un punto donde ambas coinciden
- la diagonal más larga representa la región de mayor identidad
- el camino 1 es el preferido al ser el más parsimonioso (implica menos cambios)
- la diagonal cruzada revela un **palíndromo**

Sec.2

Sec.1

**alineamiento diagonal 1**

```

secuencia 1: ATGCGTCGTT
              ||| ||| |||
secuencia 2: ATGCGT  GGT
            
```

**alineamiento diagonal 2**

```

              gap
secuencia 1: ATG---CGTCGTT
              ||| ||| |||
secuencia 2: ATGCGTCGTT
            
```

**Programación dinámica y la generación de alineamientos pareados (globales y locales): dot plots y visualización de la similitud entre secuencias**

secuencia 1: ATGCGTCGTT  
secuencia 3: ATCCGTCAT

secuencia 1: ATGTCGTT  
secuencia 3: ATTCGTCAT

- la diagonal cruza celdas vacías, correspondientes a posiciones con distintos estados de carácter
- se pueden alinear dos secuencias aleatorias postulando una combinación de sustituciones y gaps
- se puede calcular el "costo" de un alineamiento contando el número de sustituciones ( $s$ ) y gaps ( $g$ ), o una función de ellos: p. ej.:  $D = s + w$ , donde  $w$  es un factor de penalización (FP) para la creación de gaps (**gap penalty**) donde para  $w = 1$  abrir un gap cuesta igual que una sustitución  $w = 2$  cuesta el doble un gap que una sustitución. Se emplean valores bajos de  $w$  si pensamos abundaron indeles en la hist. evol. de las secs.
- **generalmente**  $w = g + hl$ , donde  $l$  es la longitud del gap,  $g$  es un FP de apertura del gap, y  $h$  es el FP para extender el gap. Estos son **FP afines**. La fórmula es muy flexible al permitir un control independiente del número y longitud ( $l$ ) de los gaps mediante  $g$  y  $h$

**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.

**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  NGDVERGKRIPIFKKCSQCFTVKEKGGKKTGPNLHGLPGRKTQQAPQYSYTAANEK---GI  58
          D  EG+ +F   QC T +  E+  GP L G+ GRK G A G++Y+ N N  G+
P00090  1  Q-DAARGEAVF----KQCMTCERADKNNVGPALGGVGRKAGTAAGPTTYPFLNHSGEAGL  56
P00001  59  IWGEDTUMKYLKPKYIP-----GEMIFVGIKKKKRADI.IAYLKKATWE  105
          +N ++ ++ YL +P Y+          TEM F +  ++R D+ AYL AT +
P00090  57  VWTQESIIAYLPDPHAYLKKFLTKGQADKATGSKMTF--KLANDQQRKDVAAYL--ATLK  114
            
```

**Alineamiento global** óptimo del citocromo C humano (105 residuos, SWISS-PROT acc. P00001) y citocromo C2 de Rhodospseudomonas 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**.

**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	RDGNAILENISFSISPOQRVGLLQHTGSGKSTLLSAFLKLL-----NYECEIQIDQVS	1273
		+ ++ +S ++ G+ + L+G +GSGXS +A L +L T GEI DG	
P33593	13	QAAQPLVBSVSLTLQGRVLAALVGGSGSGKSLNCAATLGLPAGVRRQTAGEIADGKP	70
P13569	1274	WDSITL-----GQNRKAFQVVPQKVFIPSGTYFRKNDPTEQWSDQEIWKVADEV	1322
		L Q R AF + + + + + R AD+	
P33593	71	VSPCALRGIKIATINQNRSAFRPL-----RPMHSHAKETCLADGFPADDA	116
P13569	1323	GRSVIHQFP-GKLDVFLVDGGCVLSEHGKQLMCLARSVLSEAKILLDDEPSRNLDPV	1379
		L + IE VL +S G Q M +A +YL ++ ++ DEP+ LD V	
P33593	117	TLTAATEAVGLENAAARVLLKLYPFMSGGMLQRMNIAANVLCSPPIIADEPTTLDLV	174

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

• **Saul Needleman** and **Christian Wunsch** (1970). *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, J Mol Biol. **48**(3):443-53.

[http://en.wikipedia.org/wiki/Needleman-Wunsch\\_algorithm](http://en.wikipedia.org/wiki/Needleman-Wunsch_algorithm)

Este algoritmo es un ejemplo de PD y **garantiza encontrar el alineamiento global de puntuación máxima**

• Smith TF, Waterman MS (1981). "**Identification of Common Molecular Subsequences**". *Journal of Molecular Biology* **147**: 195-197.

[http://en.wikipedia.org/wiki/Smith-Waterman\\_algorithm](http://en.wikipedia.org/wiki/Smith-Waterman_algorithm)

Algoritmo de PD que garantiza encontrar el alineamiento local de puntuación máxima

• **Ver material suplementario 1: algoritmos de programación dinámica de NW y SW**

**Programación dinámica: Notas prácticas sobre el uso de los algoritmos de Smith-Waterman y Needleman-Wunsh.**

**Alineamientos globales vs. locales**

- Aunque muy similares desde el punto de vista mecanístico, ambos tienen propiedades y aplicaciones muy diferentes. Por ejemplo, si queremos alinear dos genes eucarióticos muy divergentes esperamos que la estructura y secuencia de exones esté relativamente conservada, si bien los intrones habrán sufrido muchos eventos de indel.
- Los exones tal vez sólo representen el 1-5% de la secuencia de estos genes. Por ello si queremos usar una estrategia de alineamiento global el resultado seguramente será desastroso desde un punto de vista biológico. Muy posiblemente las regiones exónicas homólogas no se alineen. Ello se debe a que su contribución a la puntuación (score) del alineamiento es mínimo dado su reducido tamaño relativo.
- En cambio un algoritmo de aln. local sí podrá identificar y alinear correctamente a las regiones exónicas homólogas. Pero usando implementaciones como las vistas en el ejemplo sólo recuperaremos aquel aln. local con la puntuación más alta.
- Estas limitaciones de los algoritmos clásicos de SW y NW han sido eliminadas en las múltiples variantes que existen para distinto propósitos (BLAST, Clustal, etc).

**Similitud entre pares de secuencias de AA**

- El alineamiento de aa difiere del de nt en dos aspectos fundamentales:
  - 1.- Existen más "símbolos" en el alineamiento de aa (20) que de nt (4)
  - 2.- El alineamiento no consiste simplemente en alinear residuos de tal manera que la mayor cantidad coincida, ya que hay que considerar los posibles **caminos mutacionales** mediante los cuales un aa es sustituido por otro

Cys (UGU) → Tyr (UAU)    1 subst. en la 2a. pos del codón

Cys (UGU) → Met (AUG)    3 subst. Una en cada posición del codón

**Por lo tanto alinear Cys con Tyr es 3 veces menos costoso que alinearla con Met**

- En el **alineamiento de nt** generalmente se valora un "match" como +1 y un "mismatch" como -3 (en NCBI BLAST), o como +5/-4 en WU-BLAST, es decir, los nt se consideran idénticos o distintos). Esto, unido a las penalizaciones de gap, define el costo de un alineamiento de nt
- Los **alineamientos de proteínas** se basan generalmente en una **matriz empírica de costo de sustitución**, derivada de la comparación de secuencias alineadas. Estas matrices empíricas reflejan someramente los caminos mutacionales.



**Alineamiento de proteínas: selección de matrices de ponderación - consejos prácticos**

- Las matrices PAM fueron derivadas de las secuencias de proteínas disponibles a finales de los 60s y ppios. de los 70s. Era una base de datos muy reducida y estaba **sesgada a proteínas chicas, globulares e hidrofílicas**! Al carecer de suficientes homólogos con diversos niveles de divergencia evolutiva tuvieron que emplear supuestos teóricos (extrapolación) para obtener las matrices de sustitución para prots. más distantes (mediante exponenciación)
- las matrices PAM son una pobre elección para alinear (o buscar en las bases de datos) proteínas con dominios hidrofóbicos (p. ej. dominios transmembrana)
- Qué matriz escoger en función del nivel de divergencia esperada (**potencial de mira retrospectiva en tiempo evolutivo**)

% identidad	PAM	BLOSUM	mira retrospectiva en tiempo evolutivo
20- 50 %	250	45	homólogos en la zona de penumbra
50- 75 %	250	62	ortólogos y parálogos en superfamilias <sup>1</sup>
75- 90 %	160	80	ortólogos y parálogos en familias <sup>2</sup>
90- 99 %	40	90	ortólogos muy cercanos

<sup>1</sup>Superfamilias de proteínas contienen diversas familias de proteínas con = 30% identidad entre ellas  
<sup>2</sup>Familias de proteínas contienen secuencias con = 85% identidad entre ellas  
 Estas definiciones fueron acuñadas por Dayhoff et al. (1978)

**Estadísticas 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.

Los estadísticos de Karlin-Altschul asumen 5 supuestos:

- Un score positivo ha de ser posible
- El score esperado ha de ser negativo
- Los residuos de una secuencia son independientes y distribuidos idénticamente
- Las secuencias son infinitamente largas
- Los alineamientos no contienen gaps

Los primeros dos supuestos los cumple cualquier matriz estimada a partir de datos reales. Los tres supuestos finales son problemáticos. Se han solucionado en trabajos posteriores.

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

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**)

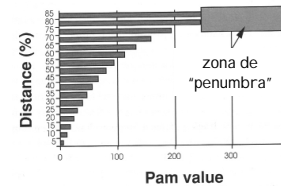
**E** Describe el ruido de fondo (por azar) presente en matches de dos secs.

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

**Alineamiento de proteínas: selección de matrices de ponderación - consejos prácticos para la identificación de homólogos**



Distancias observadas vs. evolutivas (PAM) entre prots.

Diferencia % obs.	Dist. evol. PAM
1	1
5	5
10	11
15	17
20	23
30	38
40	56
50	80
60	112
70	159
80	246
85	328 ← z. penumbra

- A medida que el nivel de divergencia entre pares de proteínas alcanza el valor de PAM250 (**- 20% identidad**), comienza a ser dudosa su relación de homología, pudiendo tratarse de secuencias que presentan cierto grado de similitud por azar, en base a composiciones de AAs similares en ambas secuencias !!!
- Al entrar en esta **zona de penumbra**, es esencial considerar información adicional, particularmente motivos estructurales, para validar o descartar una posible relación de homología

A medida que el nivel de divergencia evolutiva entre pares de proteínas incrementa (distancias PAM) disminuye el número de diferencias observadas, debido a fenómenos de **reversión (homoplasia)**. Por tanto, si no se cuenta con evidencia estructural, el análisis filogenético de proteínas debe restringirse a aquellas con = 20% de identidad. **Los alns. tampoco son confiables**

**BLAST: Basic Local Alignment Search Tool**

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. J Mol Biol 215: 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-402.

Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., and Altschul, S. F. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29: 2994-3005.



**BLAST: Basic Local Alignment Search Tool**

1. El primer paso para iniciar una búsqueda de BLAST es seleccionar la base de datos de secuencias en la que se quieren encontrar los homólogos (secuencias significativamente similares).

- Bases de datos
- Genómicas
  - Secuencias no redundantes
  - Expressed sequence tags
  - Protein Data Bank
  - Environmental samples
  - ...

2. El segundo paso es la selección del programa de búsqueda y parámetros del mismo

- Progr. de búsqueda
- BLASTN (nt-nt),
  - BLASTP (p-p),
  - BLASTX (translated nt-p),
  - TBLASTN (p-translated nt),
  - TBLASTX (translated nt-translated nt)
  - PSI y PHI BLAST (variantes de BLASTP)

**BLAST: Basic Local Alignment Search Tool**

1. Abajo una búsqueda sobre genomas microbianos.



**BLAST: Basic Local Alignment Search Tool**

1. Abajo (y páginas siguientes) una búsqueda BLASTP sobre la base de datos no redundante

**Basic BLAST**

Choose a BLAST program to run.

- [nucleotide blast](#) Search a **nucleotide** database using a **nucleotide** query  
*Algorithms:* blastn, megablast, discontinuous megablast
- [protein blast](#) Search **protein** database using a **protein** query  
*Algorithms:* blastp, psi-blast, phi-blast
- [blastx](#) Search **protein** database using a **translated nucleotide** query
- [tblastn](#) Search **translated nucleotide** database using a **protein** query
- [tblastx](#) Search **translated nucleotide** database using a **translated nucleotide** query

**BLAST: Basic Local Alignment Search Tool**

1. Abajo una búsqueda BLASTP sobre la base de datos no redundante con valores por defecto de los parámetros de búsqueda



**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 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

Gene Info

Structures

**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).

**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 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)

neighborhood word threshold score

two-hit distance

extension attenuation parameter

aln. threshold (ungapped)

aln. threshold (gapped)

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

matriz de sustitución

gap penalties

E value umbral usado = 10: HSPs con gap

E value umbral usado = 10: HSPs no gap

**BLAST: Basic Local Alignment Search Tool**

• Anatomía de un reporte de NCBI-BLAST estándar

6. Cladogramas o árboles de NJ o ME. Navegar por los hits en forma de árboles



**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

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

**BLAST: Basic Local Alignment Search Tool**

- Ver material suplementario 2: El algoritmo BLAST
- Ver material suplementario 3: PSI - BLAST