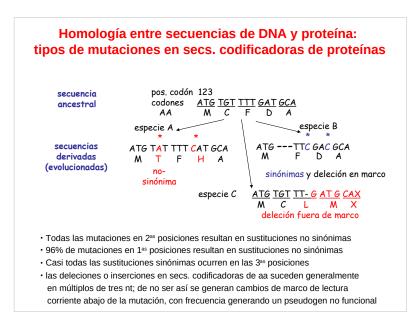
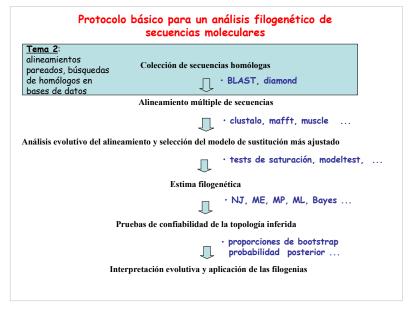
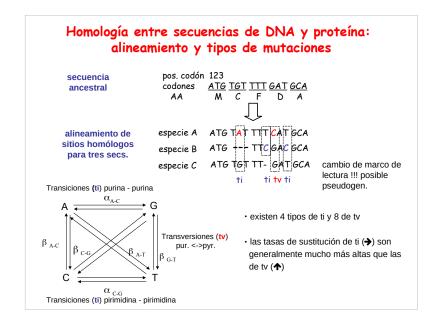
Introducción a la Filoinformática: Pan-genómica y Filogenómica -Diplomado en Bioinformática, Univ. Autónoma de Sinaloa, Oct. 2021 http://econtinua.uas.edu.mx/diplomados/2510-2700-18 001.htm Pablo Vinuesa (vinuesa[at]ccg . unam . mx; @pvinmex ) Centro de Ciencias Genómicas, CCG-UNAM, Cuernavaca, México http://www.ccg.unam.mx/~vinuesa/ Todo el material del curso (presentaciones, tutorales y datos) lo encontrarás en: https://github.com/vinuesa/intro-filoinfo\_UAS · 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



© Pablo Vinuesa 2021, @pvinmex vinuesa[at]ccg[dot]unam[dot]mx; http://www.ccg.unam.mx/~vinuesa/

Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021





### Alineamientos pareados y búsqueda de homólogos en bases de

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.

```
> | Cgi171548986|ref|ZP 00669120.1| Translation elongation factor G:Small GTP-binding protein domain [Nitrosomonas eutropha C71]
gi171486977|gb|ZR018626.1| 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 MTREFSLEKTRNIGIMAHIDAGKTTTTERVLYTTGRIKKIGETHBGASQMDWAAQEQERG 60
M++ LE+ RNIGIMAHIDAGKTTTTERVLYTTGRIKKIGETHBGASQMDWAAQEQERG 60
Spict 1 MSKRMPLEBERVRNIGIMAHIDAGKTTTSENLIFTGVSHKIGEVHBGAATMDWAGEQERG 60

Query 61 XXXXXXXXXXXXWN------DHRINIIDTPGHVDFTVEVERSLRVLDGAVAVLDAQQSVE 113
ITTSAATT W +HRIN+IDTPGHVDFTVEVERSLRVLDGAVAVLDAQQSVE 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 o 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)

(5)			
P13569	1221	EGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLLNTEGEIQIDGVS + ++ +S ++ G+ + L+G +GSGKS +A L +L T GEI DG	1273
P33593	13	QAAQPLVHGVSLTLQRGRVLALVGGSGSGKSLTCAATLGILPAGVRQTAGEILADGKP	70
P13569	1274	WDSITLQQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEV L O R AF + + + + + + K AD+	1322
P33593	71	VSPCALRGIKIATIMONPRSAFNPLHTMHTHARETCLALGKPADDA	116
P13569	1323	GLRSVIEQFP-GKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPV	1379
P33593	117	L + IE VL +S G Q M +A +VL ++ ++ DEP+ LD V TLTAAIEAVGLENAARVLKLYPFEMSGGMLQRMMIAMAVLCESPFIIADEPTTDLDVV	174

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

© Pablo Vinuesa 2021, @pvinmex vinuesa[at]ccg[dot]unam[dot]mx; http://www.ccg.unam.mx/~vinuesa/

# Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021

### 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.

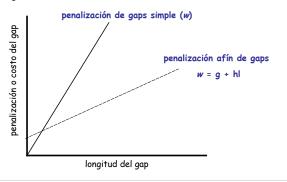
(a)

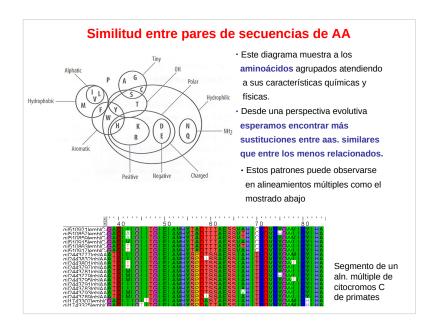
Alineamiento global óptimo del citocromo C humano (105 resíduos, SWISS-PROT acc. P00001) y citocromo C2 de Rhodopseudomonas palustris (114 resíduos, SWISS-PROT acc. P00090).

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.

#### 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.



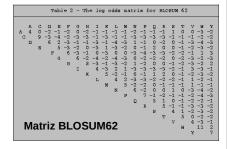




 Matrices de sustitución de AAs log-odds scores

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

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

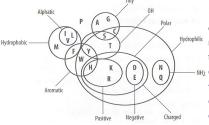


- = verosimilitud de la hipótesis a testar; frecuencia esperada o diana, probabilidad con la que esperamos encontrar a y b apareados en un alineamiento múltiple
- f<sub>a</sub>f<sub>b</sub> = verosimilitud de la hipótesis nula; frecuencia de fondo, probabilidad con la que esperamos encontrar a y b en 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".

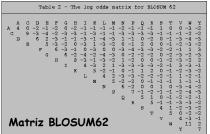
© Pablo Vinuesa 2021, @pvinmex vinuesa[at]ccg[dot]unam[dot]mx; http://www.ccg.unam.mx/~vinuesa/

Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021





 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.

### 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  $\lambda$ , una constante específica para cada matriz.  $\lambda$  es aprox. igual al inverso del factor de escalamiento (c).

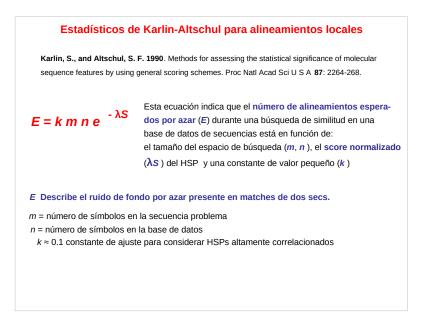
$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b} \qquad p_{ab} = f_a f_b e^{\lambda S_{ab}} = \text{score normalizado}$$

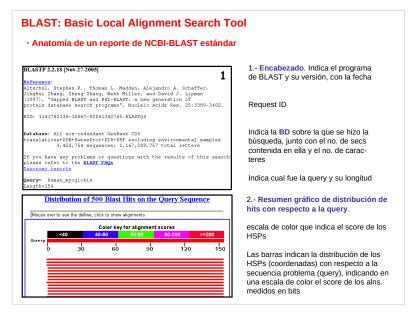
por tanto, para despejar  $\lambda$  necesitamos  $f_af_b$  y encontrar el valor de  $\lambda$  para el que la suma de las frecuencias diana implícitas valga 1.

$$\sum_{a=1}^{n} \sum_{b=1}^{a} p_{ab} = \sum_{a=1}^{n} \sum_{b=1}^{a} f_{a} f_{b} e^{\lambda S_{ab}} = 1$$

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

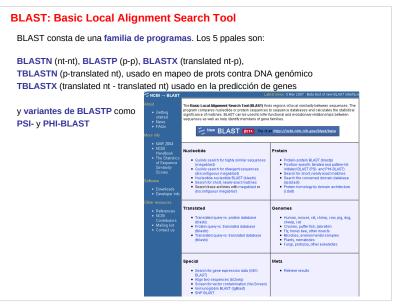
Dado que las  $f_a f_b$  de los resíduos de algunas proteínas difieren mucho de las frecuencias de resíduos empleadas para calcular las matrices PAM y BLOSUM, versiones recientes de BLASTP y PSI-BLAST incorporan una "composition-based  $\lambda$ " que es "hit-específica"

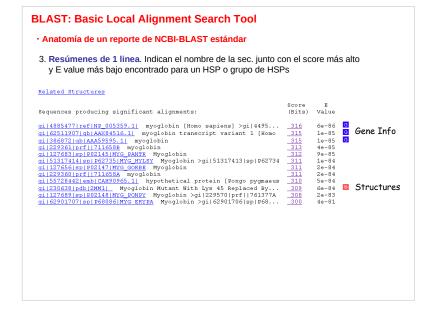




© Pablo Vinuesa 2021, @pvinmex vinuesa[at]ccg[dot]unam[dot]mx; http://www.ccg.unam.mx/~vinuesa/

# Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021





#### **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). > Tqi|47523546|ref|NP 999401.1| myoglobin [Sus scrofa] qi|127688|sp|P02189|MYG PIG Myoglobin qi|164547|qb|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%) MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE MGLSDGEWQLVLNVWGKVEAD+ GHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE MGLSDGEWOLVLNVWGKVEADVAGHGOEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE DIKKHGATVITALGGTIKKKGHHEARTKPIAOSHATKHKTPVKYLERTSECTTOVLOSKH 120 Query 61 DIKKHG TVLTALGGTIKKKGHHEAE+ PLAOSHATKHKTPVKYLEFISE TIOVLOSKH Sbjct 61 DLKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154 PGDFGADAQGAM+KALELFR DMA+ YKELGFQG PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG



### Genómica Evolutiva I LCG-UNAM, Semestre 2019-1



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
- 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/

© Pablo Vinuesa 2021, @pvinmex vinuesa[at]ccg[dot]unam[dot]mx; http://www.ccg.unam.mx/~vinuesa/

Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021

### 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.
- 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:

Icl|integer Icl|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.

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

#### REFERENCIAS CLAVES:

1: Boratyn CM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TI, Matten M, McGinnis SD, Merezhuk Y, Raytselik Y, Sayers EN, Tao T, Ve J, Zaretskaya I, ELAST: a Mode efficient port with stability improvements. Nucleic Acids Res. 2013 Jul;41 (Web Server issue):W29-33. doi: 10.1093/nar/gkt282. Epub 2013 Apr 222. PubMed Central PMCID: 23609422: PubMed Central PMCID: 32609426.

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: PMC2030857.

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 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

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

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

BLAST	BLAST+	Descripción
blastall	blastn, blastp,	
-р	No existe	blastn, blastp, blastx, tblastn,
-i	-query	Archivo de entrada
-d	-db	Base de datos de blast
-0	-out	Nobre de archivos de salida
-m	-outfmt	Formato salida; TAB: 6 == m 8
-е	-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

© Pablo Vinuesa 2021, @pvinmex vinuesa[at]ccg[dot]unam[dot]mx; http://www.ccg.unam.mx/~vinuesa/

Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021

#### 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. Avuda detallada

Programa -help (por ejemplo: blastp -h)

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++

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

- 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 segs4 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

- ¿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?
- 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

© Pablo Vinuesa 2021, @pvinmex vinuesa[at]ccg[dot]unam[dot]mx; http://www.ccg.unam.mx/~vinuesa/

# Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021

#### Eiercicios: 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

- 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

- ¿Qué comando usarías para un generar una base de datos con el archivo \*4blastdb.faa para que esté indexado?
- ¿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.