**Bitácora del curso**

Bioinformática Básica

2024 – 1

Docente: **Pedro E. Romero, *Dr. rer. nat.***

[promeroc@unmsm.edu.pe](mailto:promeroc@unmsm.edu.pe)

Enlace de la **carpeta compartida OneDrive**: [BB](https://unmsmmail-my.sharepoint.com/:f:/g/personal/promeroc_unmsm_edu_pe/EvYY6-VlmAdNnr52c3kgPtUB5hNGHt1x4HoKhQjc2DLh0g?e=K6VZYy)

Tabla con información de las y los estudiantes

[Información de estudiantes](https://unmsmmail-my.sharepoint.com/:x:/g/personal/promeroc_unmsm_edu_pe/EQIkpcz4TrlOiP5YS-RuPOUB1hiQuQmNzrsMgVOw2W3nCQ?e=GSdsYT)

**12 julio**

[**https://unmsmmail-my.sharepoint.com/:f:/g/personal/promeroc\_unmsm\_edu\_pe/EhpTqDiLDfJLuKkkF4dq67oB5pUXdsuUt\_BeO-4ohuZhSA?e=uCVGiT**](https://unmsmmail-my.sharepoint.com/:f:/g/personal/promeroc_unmsm_edu_pe/EhpTqDiLDfJLuKkkF4dq67oB5pUXdsuUt_BeO-4ohuZhSA?e=uCVGiT)

**09 julio**

**Proteínas**

**Uniprot**

<https://www.uniprot.org/>

**P38928**

Obtenga la siguiente información:

PROTEINA / GEN / ORGANISMO / FUNCIÓN / UBICACIÓN / TAMAÑO / SECUENCIA FASTA

**Protparam**

<https://web.expasy.org/protparam/>

Describir a la proteína

Comparar la proteina de levadura con la insulina humana

**Predicción del péptido señal**

SignalP

<https://services.healthtech.dtu.dk/services/SignalP-6.0/>

Si no funciona, usar:

<https://services.healthtech.dtu.dk/services/SignalP-5.0/>

Si no funciona el anterior

<https://services.healthtech.dtu.dk/services/SignalP-4.1/>

Buscar el péptido señal en dos proteínas

LISOZIMA DE HOMO SAPIENS

RODOPSINA DE DANIO RERIO

Buscar en Uniprot y en inglés

Buscar una proteína secretada y una proteína transmembrana en una bacteria. Cuál de ambas proteínas tiene péptido señal.

Verificar en cuatro proteinas dominios transmembrana.

Lisozima de *Homo sapiens*

Rodopsina de *Danio rerio*

Proteína secretada de una bacteria

Proteína transmembrana de una bacteria

<https://dtu.biolib.com/DeepTMHMM>

<https://services.healthtech.dtu.dk/services/TMHMM-2.0/>

**Estructura secundaria**

<http://bioinf.cs.ucl.ac.uk/psipred/>

Cn3D

<https://ftp.ncbi.nlm.nih.gov/cn3d/Cn3D-4.3.1_setup.exe?authuser=0>

**CN3D**

>NP\_001191615.1 insulin precursor [Aplysia californica]

MSKFLLQSHSANACLLTLLLTLASNLDISLANFEHSCNGYMRPHPRGLCGEDLHVIISNLCSSLGGNRRF

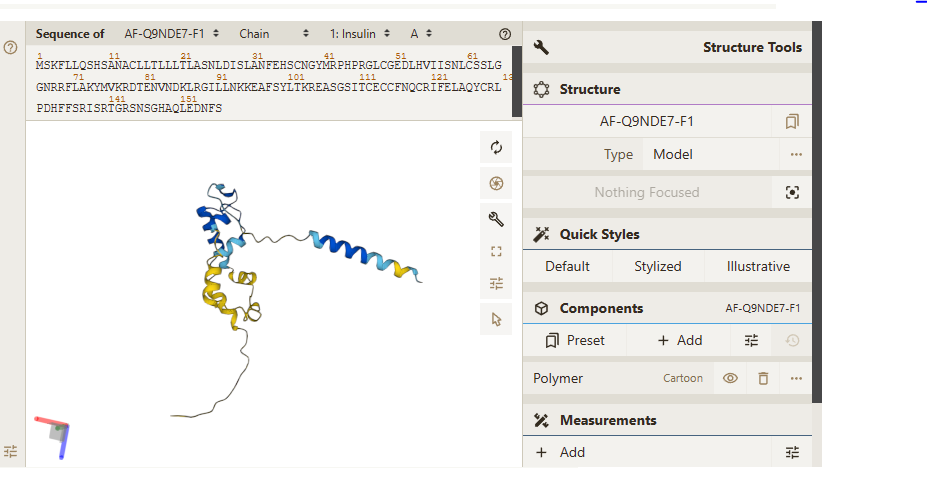
LAKYMVKRDTENVNDKLRGILLNKKEAFSYLTKREASGSITCECCFNQCRIFELAQYCRLPDHFFSRISR

TGRSNSGHAQLEDNFS



**FOLDSEEK**

[**https://search.foldseek.com/search**](https://search.foldseek.com/search)



Revisemos las estructuras secundarias de la insulina humana, GFP y rodopsina.

MIOGLOBINA HUMANA (O RATÓN)

PLASTOCIANINA (DE PLANTAS)

FLAVODOXINA (E. COLI)

**Material para seminarios**

Seminario 1: Uso de la terminal Linux

<https://swcarpentry.github.io/shell-novice-es/>

Seminario 2: Manejo de datos genómicos

<https://datacarpentry.org/wrangling-genomics/>

Seminario 3: R Studio para genómica

<https://datacarpentry.org/genomics-r-intro/>

Metabarcoding 16S

<https://training.galaxyproject.org/training-material/topics/microbiome/tutorials/mothur-miseq-sop/tutorial.html>

**02 julio 2024**

Proyecto personal: Búsqueda de bibliografía y esquema de metodología

<https://github.com/quipupe/metabarcoding/wiki/Microbial-diversity-in-the-Rimac-river>

**28 junio 2024**

**Artículo 1 (2 personas)**

Data mining of DNA sequences submitted by Peru-vian institutions to public genetic databases

<https://revistasinvestigacion.unmsm.edu.pe/index.php/rpb/article/view/17867/16385>

**Artículo 2 (2 personas)**

Lack of local genetic representation in one of the regions with the highest bird species richness, the Peruvian Amazonia

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0296305>

**Artículo 3 (3 personas)**

Representation and participation across 20 years of plant genome sequencing

<https://www.nature.com/articles/s41477-021-01031-8>

**Artículo 4 (3 personas)**

DNA barcoding in the Southeast Pacific marine realm: Low coverage and geographic representation despite high diversity

<https://www.nature.com/articles/s41477-021-01031-8>

**25 junio 2024**

**Carpeta con archivos** [Junio.25](https://unmsmmail-my.sharepoint.com/:f:/g/personal/promeroc_unmsm_edu_pe/EihcRoNH0wdJqzeK4YMbuqgBnzynzHCeesL2eVRJ2HSpkA?e=oJjlJd)

Primates.nex

MrBayes (Windows)

Tracer (Windows)

conda install bioconda::mrbayes

MrBayes (MAC) <https://github.com/NBISweden/MrBayes/tree/v3.2.7a>

Tracer (MAC) <https://github.com/beast-dev/tracer/releases/tag/v1.7.2>

**18 de junio 2024**

RAxML

Tutorial <https://cme.h-its.org/exelixis/web/software/raxml/hands_on.html>

Carpeta con archivos: [Junio.18](https://unmsmmail-my.sharepoint.com/:f:/g/personal/promeroc_unmsm_edu_pe/EheFEza1EnxOk3wsH51HYq0BV2gnXzIRU6p_7kBHrte-iQ?e=Pgqvmo)

|  |
| --- |
| conda activate bioinfo\_basic  conda install raxml -c bioconda  #Calcular el arbol  raxmlHPC -m GTRGAMMA -p 12345 -# 20 -s primates.phy -n T13  #Bootstrap  raxmlHPC -m GTRGAMMA -p 12345 -b 12345 -# 100 -s primates.phy -n T14  raxmlHPC-PTHREADS -T 30 -m GTRGAMMA -p 12345 -b 12345 -# 100 -s primates.phy -n pedro  raxmlHPC -m GTRCAT -p 12345 -f b -t RAxML\_bestTree.T13 -z RAxML\_bootstrap.pedro -n RESULTADO.BOOTSTRAP    raxmlHPC -m GTRGAMMA -p 12345 -q partition.simple -s dna.phy -n T22    raxmlHPC -m GTRGAMMA -p 12345 -q partition.12\_3 -s dna.phy -n T23 |

#Bipartition corresponde a los resultados de bootstrap se refiere a las bifurcaciones de grupos o linajes

raxmlHPC -m GTRGAMMA -p 12345 -q partition.simple -# 20 -s dna.phy -n output1

raxmlHPC -m GTRGAMMA -p 12345 -b 12345 -# 100 –q partition.simple -s dna.phy -n output2

raxmlHPC -m GTRCAT -p 12345 -f b -t RAxML\_bestTree.output1 -z RAxML\_bootstrap.output2 -n RESULTADO.tarea

#Partition en RAxML corresponde al numero de genes o regiones a utilizar

FigTree

<https://github.com/rambaut/figtree/releases>

**14 de junio 2024**

**Análisis filogenético**

KX259260

JN689230

GU572375

AY550080

EF158848

JQ956544

FJ848877

FJ648350

EU831208

EU831200

JN885091

HM236338

JX914665

KJ815049

Revisar el artículo de Medina et al. (2016)

<https://revistasinvestigacion.unmsm.edu.pe/index.php/rpb/article/view/12861>

Archivo MEG

<https://unmsmmail-my.sharepoint.com/:u:/g/personal/promeroc_unmsm_edu_pe/EWEI5GzolY9KrTOarSRW7KUBi02a9b1efgHQnfYrKH8dPw?e=hjsVTn>

Secuencia problema

>secuenciaXXX

TGAAACTGCGAATGGCTCATTAAAACAGTTATAGTTTATTTGATAATGGAAAGCTACATGGATAACCGTG

GTAATTCTAGAGCTAATACATGCTGTCAAACCCGACCTTTGGAAGGGTTGTATTTATTAGATATTAAGCC

AATATTCCTTCGGGTCTATTGTGGTGAATCATAGTAACTGATCGAATCTCTTCACGAGATAAATCATTCA

AGTTTCTGCCCTATCAGCTTTCGATGGTAGTGTATTGGACTACCATGGCAGTCACGGGTAACGGAGAATT

AGGGTTCGGTTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAAT

TACCCAATCCTGATTCAGGGAGGTAGTGACAAGAAATAACAACCTGGGGGCCTCACGGCCTTACGGGATT

GTAATGAGAACAATTTAAACGACTTAACGAGGAACAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGT

AATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAAAGCTCGTAGTTGAACTTCTGCATGTG

CCCAGTTCTGGCTTCGGTCAAGCTGTGGTGTATGCATCCGCTTGCAAAGCTAGACCGGTCTTCATTGATC

GACTAGTGGAGTAGGCTCTTTACCTTGAAAAAATTAGAGTGTTTCAGGCAGGCAATGGCTCGAATACATT

AGCATGGAATAATGGAATAGGACTTTTGTCCATTTGGTTGGTTATTGGACATAAGTAATGATTAAAAGGG

ACAGTTGGGGGCATTAGTATTTAATTGTCAGAGGTGAAATTCTTGGATTTATTAAAGACTAACTTATGCG

AAAGCATTTGCCAAGGATGTTTTCATTAATCAAGAACGAAAGTTAGGGGATCAAAGACGATCAGATACCG

TCCTAGTCTTAACTATAAACTATACCGACTCGGAATCGGACCGGCTTATAAAACTGGTTCGGCGCCGTAT

GAGAAATCAAAGTCTTTGGGTTCTGGGGGGAGTATGGTCGCAAGGCTGAAACTTAAAGGAATTGACGGAA

GGGCACCACCAGGCGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGGAAACTTACCAGGTCCAAACAT

GGGTGGGATTGACAGATTGAGAGCTCTTTCTTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGG

TGGAGTGATTTGTCTGGTTAATTCCGTTAACGAACGAGACCTTAACCTGCTAAATAGTACGTTGATGCAC

AATTGGCGTTACTTCTTAGAGGGACTATGCGCTTTGAAACGCATGGAAGTTTGAGGCAATAACAGGTCTG

TGATGCCCTTAGATGTCCTGGGCCGCACGCGCGCTACAATGACTCGCTCAGAAAGTACTTCCTGGTCCGG

AAGGATTCGGGTAATCTTTTAAATACGAGTCGTGTTAGGGATCGATCTTTGTAATTATGGATCTTGAACG

AGGAATGCCTAGTAAGTGCAAGTCATCAGCTTGTACTGATTACGTCCCTGCCCTTTGTACACACCGCCCG

TCGCTCCTACCGATTTCGAGTGATCCGGTGAACCTTCTGGACTGAGCACGCTTGCG

**28 de mayo 2024**

**Seminario 1**

Seminario 1: Uso de la terminal Linux

<https://swcarpentry.github.io/shell-novice-es/>

**24 mayo 2024**

Carpeta compartida [Mayo.24](https://unmsmmail-my.sharepoint.com/:f:/g/personal/promeroc_unmsm_edu_pe/EiwvZgy4Rt9MnY4YkM00wKcBWx-5XpFYnCNOX71jNu3Jxg?e=6sR1k6)

**Mobaxterm**

<https://mobaxterm.mobatek.net/download.html>

**Análisis de calidad en Linux**

Guía para hoy: <https://jshleap.github.io/bioinformatics/writting-jNGS_tutorial/>

Instalar fastqc, trimmomatic y cutadapt

conda **install bioconda::fastqc**  
conda **install bioconda::trimmomatic**  
conda **install bioconda::cutadapt**

Crear una carpeta llamada **calidad** y entrar a la carpeta

mkdir calidad  
cd calidad

Descargar dos archivos fastq

wget <https://github.com/jshleap/CristescuLab_misc/raw/master/Tutorials/NGS_QC/files/file1_R1.fastq.gz>

wget <https://github.com/jshleap/CristescuLab_misc/raw/master/Tutorials/NGS_QC/files/file1_R2.fastq.gz>

Utilizar fastqc para leer el archivo R1 y transferir el archivo R1.fastqc.html a su computador. Es más fácil salid del servidor para descargar el archivo.

fastqc file1\_R1.fastq.gz

scp [usuario@XX.XX.XX.XX:/home/calidad/file1\_R1.fastqc.html](mailto:usuario@XX.XX.XX.XX:/home/calidad/file1_R1.fastc.html) .

Evaluar el archivo R2.

Correr trimmomatic

**trimmomatic** **PE** **-threads** 10 **file1\_R1**.fastq.gz **file1\_R2**.fastq.gz **file1\_R1**.1P.fastq.gz **file1\_R1**.1U.fastq.gz **file1\_R2**.1P.fastq.gz **file1\_R2**.1U.fastq.gz **SLIDINGWINDOW**:4:20

Correr fastqc para el archivo **file1\_R1.1P.fastq.gz**, revisar si hay diferencias con el archivo original **file1\_R1.fastq.gz**

De preferencia los archivos generados con trimmomatic deben ser colocados en una nueva carpeta, el comando **mkdir** crea carpetas y el comando **mv** mueve archivos.

Correr otra vez Trimmomatic considerando la información de los adaptadores

**trimmomatic** **PE** **-threads** 4 **file1\_R1**.fastq.gz **file1\_R2**.fastq.gz **file1\_R1**.1P.fastq.gz **file1\_R1**.1U.fastq.gz **file1\_R2**.1P.fastq.gz **file1\_R2**.1U.fastq.gz **SLIDINGWINDOW**:4:20 **ILLUMINACLIP**:adapters.all.fa:2:40:15 **MINLEN**:100

Correr fastqc en el archivo file1\_R1.1P.fastq.gz

Revisar la presencia de adaptadores

zgrep **--color**=always "CAAACTGGGATTAGATACCCCACTATG" file1\_R1.1P.fastq.gz | head

Primers utilizados

Forward primer: **GTCGGTAAAACTCGTGCCAGC**

Reverse primer: **CATAGTGGGGTATCTAATCCCAGTTTG**

Obtener las secuencias reverse complement

GCTGGCACGAGTTTTACCGAC

CAAACTGGGATTAGATACCCCACTATG

Correr cutadapt

Adapter type Command-line option

3’adapter -a ADAPTER  
5’adapter -g ADAPTER  
5’ **or** 3’ (both possible) -b ADAPTER

cutadapt -g ^FWDPRIMER -G ^REVPRIMER -**a** ^FWDPRIMER...RCREVPRIMER -A ^REVPRIMER...RCFWDPRIMER -o out.1.fastq.gz -**p** out.2.fastq.gz **in**.1.fastq.gz **in**.2.fastq.gz

**Short Reads Scenario:**

If the reads are shorter than the targeted DNA fragment (amplicon), only the forward primer on R1 and the reverse primer on R2 need to be removed.

Cutadapt command:

cutadapt -g ^FWDPRIMER -G ^REVPRIMER --discard-untrimmed -o out.1.fastq.gz -**p** out.2.fastq.gz **in**.1.fastq.gz **in**.2.fastq.gz

**Long Reads Scenario:**

If the reads can be longer than the amplicon, a "linked adapter" approach is used.

The linked adapter is constructed to account for the possibility of reads extending into the primer on the opposite side.

Cutadapt commands for R1 and R2 respectively:

R1: -a ^FWDPRIMER...RCREVPRIMER  
R2: -A ^REVPRIMER...RCFWDPRIMER

Note: Reverse-complemented sequences are used for the reverse primer and the primer on the opposite side.

cutadapt -g **FORWARD** -a **REVERSE REV COMP** -G **REVERSE** –A **FORWARD REV COMP file1\_R1.1P.fastq.gz**

cutadapt -**a** AACCGGTT –g xxxx –G xxxxxx –A xxxxx -o output.fastq **file1\_R1.1P.fastq.gz**

**cutadapt -a AACCGGTT** –g xxxx –G xxxxxx –A xxxxx **-o output.fastq.gz input.fastq.gz**

cutadapt -a "CTGTCTCTTATACACATCT" -g "AGATGTGTATAAGAGACAG" -A "CTGTCTCTTATACACATCT" -G "AGATGTGTATAAGAGACAG" -o Read1 -p Read2 Read1 Read2

cutadapt -a "CTGTCTCTTATACACATCT" -g "AGATGTGTATAAGAGACAG" -A "CTGTCTCTTATACACATCT" -G "AGATGTGTATAAGAGACAG" -o Read1 -p Read2 Read1 Read2

Correr fastqc

**21 mayo 2024**

**Análisis de cromatogramas**

Instalar CutePeaks

<https://github.com/labsquare/CutePeaks>

Windows

<https://github.com/labsquare/CutePeaks/releases/download/0.2.3/CutePeaks-win32.exe>

Ejemplos de cromatogramas

<https://unmsmmail-my.sharepoint.com/:f:/g/personal/jsoliss_unmsm_edu_pe/EgPVeSOQg69Fjg6_8cKu2skB7IwdpIEHl_-PAfLCe7pSww?e=oZoaz3>

IUPAC codes

<https://www.bioinformatics.org/sms/iupac.html>

Demostración

Secuencia OK: L130

Secuencia not OK: L701

**Ensamblador CAP3**

<https://doua.prabi.fr/software/cap3>

**Informe**

[otras\_secuencias](https://unmsmmail-my.sharepoint.com/:f:/g/personal/jsoliss_unmsm_edu_pe/Egl70Ay2KlFLoa9vhgc2facBpcRdFOcGqgALPZhqferCWw?e=YoHFbO)

Información de los cortes

Tamaño del contig

Resultado de BLAST

**14 y 17 de mayo 2024**

Alineamientos múltiples

#Descargar secuencias en formato Genbank  
**for** i in $(cat lista); **do** esearch -db nucleotide -query $i | efetch -format gb > $i.gb; **done**

#Extraer las secuencias CDS  
for file in $(cat lista); **do** perl cds\_extractor.pl -i $file.gb -n; **done**

#Cambiar los nombres de las secuencias  
for i in $(cat lista); **do** cat $i.ffn | sed -r 's/.\*g=(\w+) .\*/>'$i'\_\1/g' > $i.names; **done**

#Revisar si los nombres de las secuencias son correctos y estandarizar los nombres de los genes  
sed -i 's/COX1/cox1/g' \*.names;  
**sed** -i 's/COX2/cox2/g' \*.names;  
**sed** -i 's/COX3/cox3/g' \*.names;  
**sed** -i 's/CYTB/cytb/g' \*.names;  
**sed** -i 's/ND1/nad1/g' \*.names;  
**sed** -i 's/ND2/nad2/g' \*.names;  
**sed** -i 's/ND3/nad3/g' \*.names;  
**sed** -i 's/ND5/nad5/g' \*.names;  
**sed** -i 's/ND6/nad6/g' \*.names;  
**sed** -i 's/ATP6/atp6/g' \*.names;  
**sed** -i 's/ND4L/nad4l/g' \*.names;  
**sed** -i 's/ND4/nad4/g' \*.names;  
**sed** -i 's/cob/cytb/g' \*.names;  
**sed** -i 's/Atp6/atp6/g' \*.names;  
**sed** -i 's/Cytb/cytb/g' \*.names;  
**sed** -i 's/nad4L/nad4l/g' \*.names;  
**sed** -i 's/\_coxIII\>/\_cox3/gI; s/\_coxII\>/\_cox2/gI; s/\_coxI\>/\_cox1/gI' \*.names;

#Extraer los genes por cada organismo  
**for** i in $(cat lista); **do** perl extractFromFasta.pl $i.names single ""$i"\_cox1" > $i.cox1.fa; **done**

#Agrupar las secuencias  
cat \*.cox1.fa > cox1.fa  
#Alinear  
mafft --auto cox1.fa > cox1.aln

Comando para una opción de mafft (EINSI)

mafft --genafpair --maxiterate 1000 input\_file > output\_file

<https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html>

Loop para alinear varios archivos

**for** i **in** \*.fasta; **do** mafft --auto $i > $i.aln; **done**

TranslatorX

<http://translatorx.co.uk/>

Pedro: Algoritmo de alineamiento de mapeo

**Ejercicio 21**

Alinear las secuencias del archivo **COI.fasta**. Utilizar una herramienta online e identificar si en su computador existe una herramienta para alineamientos múltiples

como ClustalX, MEGA, etc

**Ejercicio 25**

Alinear el archivo **16S.fasta**. Revisar la variabilidad del alineamiento con respecto al ejercicio anterior. Observar gaps y verificar opciones en MEGA de ClustalW vs MUSCLE

**Ejercicio 26**

Alinear el archivo fibrinogeno.fasta

Secuencias de aminoácidos

**Ejercicio 27**

Alinear el archivo **DENV.fasta**

Secuencias pueden provenir del mismo gen pero no de la misma zona por lo tanto no alinearán

**Ejercicio 28**

Alinear el archivo DENV\_complete.fasta

**Ejercicio 29**

Utilizar https://www.ebi.ac.uk/Tools/msa/ y contrastar los resultados de cuatro alineadores usando el archivo **DENV\_complete.fasta**

Clustal Omega

MUSCLE

MAFFT

T-Coffee

Este link no funciona tan bien

**Opcional**

https://mafft.cbrc.jp/alignment/server/

**Ejercicio 30**

Genera tu propio alineamiento múltiple con secuencias que desees investigar. Ten cuenta el problema te gustaría resolver, el gen que analizarás, si usarás varias especies o solo una especie. Obtén al menos 20 secuencias para alinear.

**10 mayo 2024**

efetch -db nucleotide -id KM233118 -format gb > KM233118.gb

 ssh [usuario@172.16.32.8](mailto:usuario@172.16.32.8)

Para bajarme los archivos del servidor

scp promeroc@172.16.32.8 /home/promeroc/KM233118.gb .

Para enviar los archivos al servidor

scp ./KM233118.gb promeroc@172.16.32.8:/home/promeroc/

**Activación de uso del servidor desde Linux.**

Ustedes deben haber recibido un correo electrónico por parte de la Maestría indicando cómo ingresar al servidor.

En mi caso, mi usuario es promeroc, y el comando es:

**ssh** nombre del usuario@**172**.**16**.**32**.**8**

la clave o password la deben tener en su correo.

Luego de activado el usuario, usar el comando

conda activate bioinfo\_basic

Comprobar si existen algunos programas instalados (revisar la clase del 03 de mayo)

Si el ambiente no ha sido creado, ejecutar (primero preguntarme y revisar si existe el archivo yml asociado a su usuario)

**conda** **env** **create** **-f** **bioinfo\_basic**.yml

**Actividad para hoy – Informe 05**

Utilizar las herramientas en Linux para diseñar una base de datos local de genomas mitocondriales. La idea es obtener un par de primers que puedan amplificar las especies mencionadas abajo.

Qué gen utilizaría para diseñar el primer.

Luego de obtenido el primer que pueda ser utilizado para todas estas especies, realizar un blast local.

*Mytilus edulis*

*Drosophila yakuba*

*Apis mellifera*

*Anopheles gambiae*

*Artemia franciscana*

*Ascaris suum*

*Caenorhabditis elegans*

*Strongylocentrotus purpuratus*

*Cyprinus carpio*

*Xenopus laevis*

*Gallus gallus*

*Mus musculus*

*Bos taurus*

*Balaenoptera physalus*

*Homo sapiens*

*ens*.

**07 mayo 2024**

**Diseño de primers**

**Ejercicio 20**

Utilizar Primer3 para diseñar un par de *primers* que amplifique parte del gen SRY del cromosoma Y humano.

Primer3: <https://primer3.ut.ee/>

NC\_000024.10:c2787682-2786855

**Ejercicio 21**

Usar el Primer3 para diseñar un par de *primers* que amplifique el gen SRY del cromosoma Y humano, **región 400-700**. La Tm está entre **50-55 grados**. El tamaño del fragmento debe ser entre **250-300 pb**.

NC\_000024.10:c2787682-2786855 (828 pb)

**Ejercicio 22**

Utilizar Primer3 para diseñar pares de *primers* que amplifiquen los genes E y N del SARS-CoV-2

**Ejercicio 22.1**

Utilizar Primer3 para diseñar pares de *primers* que amplifiquen los genes E y N del SARS-CoV-2. Que tengan la misma Temperatura de melting.

**Ejercicio 22.2**

Buscar que sus *primers* no se encuentren en el genoma humano. Tanto para los primers del gen N y E

**Ejercicio 23**

Utilizar **Primer-BLAST** para amplificar dos fragmentos, uno del gen E y otro del gen N del SARS–CoV-2, pero que no amplifiquen inespecíficamente a algún fragmento del genoma humano.

<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>

**Ejercicio 24**

Usar esta web para calcular propiedades de los oligonucleótidos y revisar la ocurrencia de homodímeros y heterodímeros.

<https://www.biosyn.com/gizmo/tools/oligo/oligonucleotide%20properties%20calculator.htm>

**Ejercicio 25**

Utilizar Primer-BLAST para diseñar un par de primers que amplifique el grupo de secuencias mencionadas en el archivo mioglobina\_ballenas.txt

**Ejercicio 26 – Informe 04**

Diseñar unos primers para las secuencias contenidas en el archivo Chara\_COI\_complete.fasta

1. Revisar qué **genes** se utilizan para **pruebas moleculares de monkeypox**.  
   2. Elegir **3 genes** para su prueba molecuar.  
   3. Disenhar 1 par de **primers** por cada gen.  
   Lo ideal sería, que los 3 pares puedan ser amplificados con una **misma temperatura** de melting.  
   La secuencia base es el linaje B1.6, típico de Perú

**03 mayo 2024**

Utilizar la terminal

**conda create -n blast**

**conda activate blast**

**(base) NO**

**(blast) SÍ**

conda **update** -**n** base -c defaults conda #Actualizar  
  
conda install blast

conda install bioconda::blast #Utilizar este comando por si no funciona la primera instalación

conda install emboss

conda install bioconda::emboss #Esta sí funciona

conda install bioconda::entrez-direct

**#Primer ejercicio**

**mkdir** **db #crear carpeta**  
  
efetch -**db** nucleotide -id KM233118 -**format** gb > **db**/KM233118.gb  
  
efetch -**db** nucleotide -id KM233118 -**format** fasta > **db**/KM233118.fa  
  
**cat** **db**/KM233118.gb | extractfeat -filter -**describe** **gene** > **db**/KM233118-features.fa

#cat KM233118-features.f  
makeblastdb -h  
  
makeblastdb -dbtype nucl -**in** **db**/KM233118-features.fa -**out** **db**/KM233118-features  
   
   
>**test** [source] Zaire ebolavirus isolate Ebola virus/**H**.sapiens-wt/SLE/2014/Makona  
AATCATACCTGGTTTGTTTCAGAGCCATATCACCAAGATAGAGAA  
   
>**test** [source] Zaire ebolavirus isolate Ebola virus/**H**.sapiens-wt/SLE/2014/Makona  
AATCATACCTGG  
   
blastn -**db** **db**/KM233118-features -**query** **query**.fa  
  
blastn -**db** **db**/KM233118-features -**query** **query**.fa -outfmt 7  
  
blastn -**db** **db**/KM233118-features -**query** **query**.fa -outfmt "6 qseqid sseqid pident"  
#aqui se eplico no se que   
blastn -**db** **db**/KM233118-features -**query** short.fa  
# este comando sirve para   
blastn -**db** **db**/KM233118-features -**query** short.fa -task blastn  
  
blastn -**db** **db**/KM233118-features -**query** mini.fa -task blastn-short

Para la secuencia chiquitita.fa usar las 7 primeras bases **AATCATA**

**#Segundo ejemplo**  
efetch -id NC\_001133 -**db** nucleotide -**format** fasta > **db**/NC\_001133.fa  
  
makeblastdb -dbtype nucl -**in** **db**/NC\_001133.fa -**out** **db**/NC\_001133  
  
head -2 **db**/NC\_001133.fa > start.fa  
  
blastn -**db** **db**/NC\_001133 -**query** start.fa  
  
blastn -subject **db**/NC\_001133.fa -**query** start.fa  
  
blastn -**query** start.fa -subject **db**/NC\_001133.fa -dust **no**

#Instalació**n** **de** miniconda  
**mkdir** -p ~/miniconda3  
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86\_64.sh -O ~/miniconda3/miniconda.sh  
bash ~/miniconda3/miniconda.**sh** -b -**u** -p ~/miniconda3  
**rm** -rf ~/miniconda3/miniconda.**sh**

qseqid means Query Seq-id

qgi means Query GI

qacc means Query accesion

qaccver means Query accesion.version

qlen means Query sequence length

sseqid means Subject Seq-id

sallseqid means All subject Seq-id(s), separated by a ';'

sgi means Subject GI

sallgi means All subject GIs

**Informe 03**

Elige una secuencia nucleotídica codificante conservada de un parásito

Busca 10 homólogos de este gen en especies distantes (Parte 1: lista de secuencias nucleotídicas)  
10 secuencias en FASTA  
  
Usando la secuencia aminoacídica del mismo gen conservado que encontraste

Busca 10 proteínas homólogas en especies distantes (Parte 2: Lista de secuencias aminoacídicas)10 secuencias en FASTA

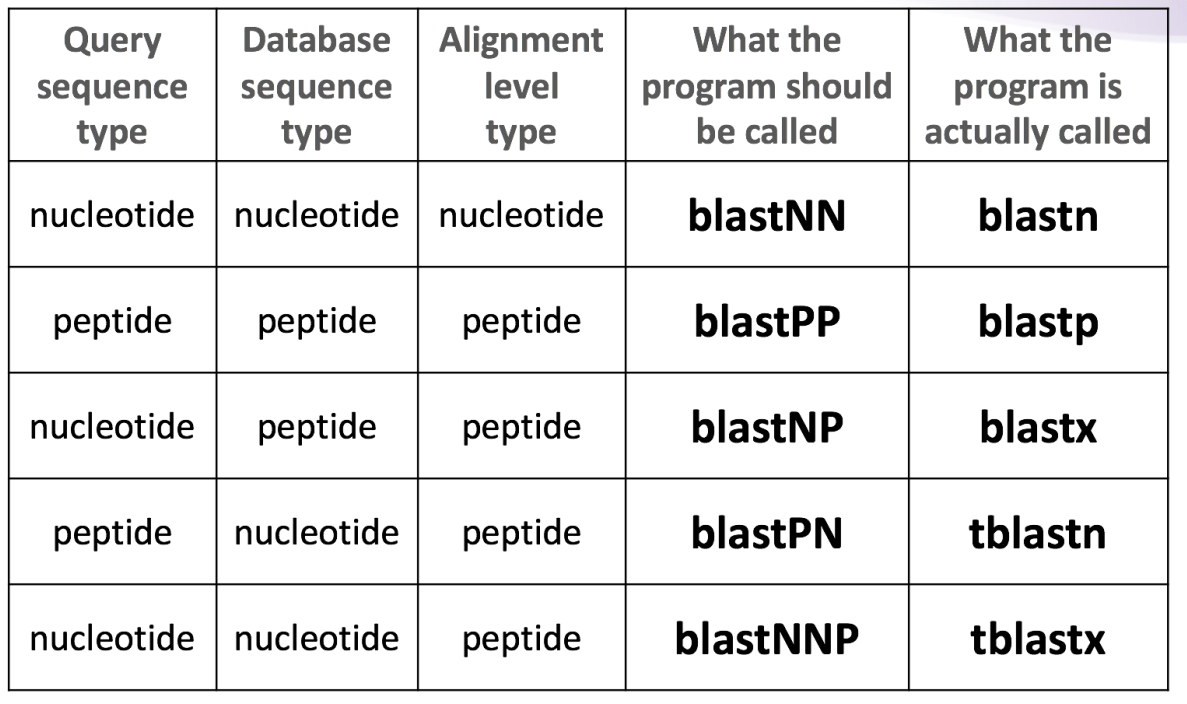
Realiza una comparación entre tus resultados generando un alineamiento local entre una secuencia nucleotídica y una secuencia aminoacídica (Parte 3, Pantallazos y texto de cómo compararía una secuencia de nt contra una de aa).  
  
Identifica a qué cromosoma del genoma humano (y que coordenadas) podríamos encontrar este gen/proteína. (Parte 4. Pantallazos y texto)

**30 abril 2024**

**BLAST**

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

**Tipos de BLAST**



**Terminología**

**Query**: La secuencia que enviamos a la base de datos

**Target**: La base de datos o colección que buscamos.

**Subject**: La secuencia resultante en la base de datos. Produce un alineamiento.

**Score**: El puntaje del alineamiento.

**E-Value**: El valor esperado (E) es el número de hits o resultados mejores que el resultado "esperado” por azar. Toma como referencia el tamaño de la base de datos.

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

**Ejercicio 008**

Descarguemos los FASTA de estas dos secuencias y comparémoslas

Ubicar el gen de la nucleoproteína o NP en ambos genomas

<https://www.ncbi.nlm.nih.gov/nuccore/AF086833.2>

<https://www.ncbi.nlm.nih.gov/nuccore/KY786027.1>

Comparar ambas secuencias, una contra la otra.

**Ejercicio 009**

Comparar la secuencia de la proteína VP35 (AAD14582) contra las 2240 proteínas presentes el BioProject PRJNA257197

**Ejercicio 010. Blastear la secuencia problema 1**

ATTCTTTTTGATTTCTCCGTATCAGGTATTTACGAAGGACAAGGCTTTCTACAATCTCATGGTAGATTGG

CAAATTATTGGGCCGAGCTGGATTTGAACCAGCGTAGACATATTGCCAACGAATTTACAGTCCGTCCCCA

TTAACCACTCGGGCATCGACCCAGGAAGAATAAATTTTAGGCTTATTTTATTGGTAATCCATGATCAACT

TCCTTTCGTAGTACCCTACCCCCAGGGGAAGTCGAATCCCCGCTGCCTCCTTGAAAGAGAGATGTCCTGA

ACCACTAGACGATGGGGGCATACTTACCCGACCGCCATTATACTATGATCATAGTATGATCAGTTTTTTG

AAATTGTCAATATAATGGAATGGTATGATTAGACCCAAGGTCTCCTTCTATTTTTGATGATTTAATCAAA

TTGTTTAATTCGTTAGTCATGAATTATTCATTCTAATCGCCATTTATTCGATTTCGATTCGCTGGATATA

TTCAATATATCGATAGAGTACTCTATTAAATTAGACTAGAATACTAAAAAAAAATAGAAATAGTACGAAA

GATAGCACTTTTTACAGGAAAAATTTGTCTGTTAGAAAATCAAAAGAAAAGGGGATTCAATTCCATTTCT

TTCACTTTCATTCATTGATTTATTCATTATTAAGATATGACTATCTCTATCTCCCACTAAGCTAGGAGAT

TAATAAAGGATAAATCCTGGAAGAGGGATCAAGAAGTTAGAGAATTTTTTCTCGAATATTAGACAAGTAG

AATCCATAATTCGAGAAAATAGCTTGAATCTATGTCGAATAGGTGTACAGATAGGTATCAATCAAGTGAA

TTTCGTTCTGATGGGACTAAAATAAAGCCAATACCGATAAAATAGGGTTGTTTTGAAAACAATTCATTGC

ATTCATTTTTCTAGACTTTCTAGGTAAATACATTTTTTATTCAAGAATAAGGCACTAACTAGCGACTATG

ATTCTACTGCATGGACTTGTGTATATAATAGATGTACATAG

**Ejercicio 011. Blastear la secuencia problema 1 contra *Arabidopsis thaliana* utilizando megablast y blastn**

**Ejercicio 013. Blastear la secuencia problema 2**

ATGGCTAAGTTTGCTTCTATCGTTCCCCTTCTCTTCGCTGCTCTTCTTCTCTTTGCTGCTTTTGCTAGTA

ATGATCTTACTCATATACATATGCAGGTGAAGTTTGTCTTTATCTCTTGATACTAAACTTGAGAAATGTG

TACAAATGCAGAATCACCAGCAATGGTGGAAGCACAAAAGCTGTGCGAGAGATCAAGTGGAACTTGGTCT

GGAGTTTGTGGAAACAATAACACATGCAAGAACCAATGCATTAACCTTGAGGGAGCACGACATGGATCAT

GCAACTATCGATTCCCATATCACAGATGCATCTGTTACGTCCCATGTTAA

**Ejercicio 014. Realice un blast discontinuo con esta secuencia**

ATGCAAAAAACAGCGCTTGATTTTGCAAACAATCCAAGCGATGAGGGTATTTTGCAAGTCTATCAGAAGA

CCGTGGAGCTGCAGGAGATTGGAATTAAAAATAATCGCACCAAGATGGATGTTTTAACTGAAATATACAA

CGACATGTATCAAGAGAAAGGCGAAATAACAGGAGTGGAAACTGGGCTAGCTGATCTAGACGCTATGACC

GGAGGCTGGCAGGATAGCGACTTAATTATCGTAGCTGCCCGGCCTTCCATGGGAAAAACAGCTTTCGCTC

TAAATTTGGCTCAGAATGCGGCGTTAAAAGGTGGAGTTGTGGATGTTTTTTCATTGGAGATGTCCGACCG

TCAACTTGTTAACCGGATGCTCAGCAATTTAGGTTCAATCGAAGGTACAAAATGGAGAAACCCGCACAAG

TATTTCAGTGAAAAAGACTATGAAAACGCGAATCGGGCCATCGGTGAATATGAAAAACTAGACATCTATA

TTCACGACAAACCGTCACAGTCAGTGGCTGATATTCGTTCTGCCATTCGGAAGACCACAAAGGAACACCC

GGATCAAAAGCATTTGGTTGTGATTGATTATCTTCAATTAATCCGACCTATTGGAAAATTTGAAACGAAA

AACTTAGAGGTCGCAAGCATAACAGGCGAATTGAAGAACATAGCCCGGACATTCAATATTCCAATCATTC

TGCTTTCTCAACTTTCTCGCGGTGTTGAGCAGCGACAAGATAAGCGGCCGATGATGTCTGATCTTCGAGA

TTCAGGCAGCATTGAACAAGATGCCGATATCGTTAGTTTTCTCTATAGGGATGACTATTACGATAAGCAA

AGCGACCTAAAAAACATAGTGGAGATCATTTTTGCAAAACAAAGAAACGGCTCGGTTGGCACTGTCATGG

CTGCCTTTATTAAAGAGTACGGCCGATTCTTAAACCTTGATAGACAAATTGAAGCGAAACTTGCATAG

**Ejercicio 015. Secuencia problema 3**

**Realice un blastp normal y luego filtre las regiones de baja complejidad**

MTTCSRQFTSSSSMKGSCGIGGGSSRISSVLAPSAYGGLSVTSSRFSSGGACGMGGSYGGGFSSSSSFGGGYGGFAGYGGGLGAGLGAGLGGGLGGGFGVGFGGGDGLLAGSEKVTMQNLNDRLASYLEKVRALEEANAELEVKIRDWYQRQRPVENRDYSSYYKTIEDLRNKILTATVDNANVVLQIDNARLAADDFRTKYETELNLRMSVEADINALRRVLDELTLARADLEMQIESLKEELAYLRKNHEEEMNTMRGQVGGDVNVEMDAAPGVDLSRILNEMRDQYEKMAEKNRRDAEEWFFSKTEELNREVATNTEALQSSKTEISELRRSVQNLEIELQSQLSMKASLENSLEETKGRYCMQLAQIQDMIGGVEEQLAQLRCEMEQQNQEYKILLDVKTRLEQEIATYRRLLEGEDAHLSSAQFSSGSQSSRDVTSSSRQVRTKVVDVHDGKVVSSHEQILRTKN

**Para Pedro Explicar estos algoritmos**

PSI-BLAST (Position-Specific Iterated BLAST)

AlgorithmPHI-BLAST (Pattern Hit Initiated BLAST)

AlgorithmDELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

**Ejercicio 016. Realice un alineamiento local blastp entre XP\_033618937.1 y KFO28518.1**

Observe el resultado llamado DotPlot

**Ejercicio 017. Realice un blastx de la siguiente secuencia:**

>XM\_004228267.1 Plasmodium cynomolgi strain B 40kDa heat shock protein (PCYB\_001170) mRNA, complete cds

ATGGCAACGTTTACGAATTATCTTAGGAGAGAAAAATTAAATATATTCCTCTTTTGTGTTAAAAGTAAGA

ATGGAGTAAATGGAGGCGGTGAGGATAATTTGGGAAAAACTTCAGATTCAGGATTTGACAGGTCGCTAGC

CGAAAAGTCAGAGAACTTCAGTTCGTCCTTTGGAGGATTTTCCGGAAAAACAAAGGTTGGATTAAAAGGA

AAAGGAAACAAAATTGATTATTACAATGTGTTAGGAGTTCCTAGAGATGCGACAGAAAATGATATAAAGA

AGGCCTATAAAAAGTTAGCCATGAAATGGCACCCAGATAAACACTTAGATGAGAAGGACAAAAAGGCTGC

GGAAGAAAAATTTAAGGTTATTTCTGAAGCTTATGATGTTTTGTCAGACCCAGATAAAAAAAACGTACGA

TTTGTATGGTGA

**Ejercicio 018. Realice un blastn de la siguiente secuencia versus el gen de esta proteína**

>XM\_004228267.1 Plasmodium cynomolgi strain B 40kDa heat shock protein (PCYB\_001170) mRNA, complete cds

ATGGCAACGTTTACGAATTATCTTAGGAGAGAAAAATTAAATATATTCCTCTTTTGTGTTAAAAGTAAGA

ATGGAGTAAATGGAGGCGGTGAGGATAATTTGGGAAAAACTTCAGATTCAGGATTTGACAGGTCGCTAGC

CGAAAAGTCAGAGAACTTCAGTTCGTCCTTTGGAGGATTTTCCGGAAAAACAAAGGTTGGATTAAAAGGA

AAAGGAAACAAAATTGATTATTACAATGTGTTAGGAGTTCCTAGAGATGCGACAGAAAATGATATAAAGA

AGGCCTATAAAAAGTTAGCCATGAAATGGCACCCAGATAAACACTTAGATGAGAAGGACAAAAAGGCTGC

GGAAGAAAAATTTAAGGTTATTTCTGAAGCTTATGATGTTTTGTCAGACCCAGATAAAAAAAACGTACGA

TTTGTATGGTGA

**Ejercicio 019. Realice un tblastn con la siguiente secuencia:**

ATGTCTGCAATTGCACTATATTTTAATGTTAATTTATATAGTAGAGAACGTAATTTAAATTTATTAGTAT

CTCTTAATAATACGTATAAACTATATTATAGTTATAAAGTAATAGTGGCTAATCTATATAAAAATATTAA

AGTAATAGAATACTTTAATAAATATTCTTTATTATCATCTAAACACTTAGATTTTTTAGATTGATCTAAA

TTAGTTATTTTAATTAATAATGAGGGTCAAAGTATAAAACTTAATGGTAGTTGAGAATTAGGTATAAATT

TACGTAAAGATTATAATAAAACTAGAACTACGTTTACTTGATCTCATTAA

**Buscar 1 secuencia de nucleotidos de un gen o fragmento génico que desee**

**Buscar 1 secuencia de aminoacidos de una proteina o fragmento proteíco que desee**

**Utilizar Blastn o Blastp y conseguir 20 secuencias homólogas al gen y la proteina escogidas.**

**Enviar el fasta de nucleotidos y de proteinas a mi correo**

**Título: Informe 2**

**26 abril 2024**

**Informe 1**

Chris

(\w)\w+ (\w)

\1. \2

Veronica

(\w)\w+ (\w+)

\1. \2

Guillermo

(\w)\w+

\1.

Luis

(\w)\w+\ (\w+)

\1. \2

Expresiones regulares

[Expresiones regulares.pdf](https://unmsmmail-my.sharepoint.com/:b:/g/personal/promeroc_unmsm_edu_pe/ETxtF498akhKhQdA3KRhjGYBPCutKjf2opHFd0gWMODrpw?e=cIsAEw)

Verificar si se tiene instalado el Notepad Plus

<https://notepad-plus-plus.org/downloads/v8.6.5/>

FPexamples.fta

[FPexamples.fta](https://unmsmmail-my.sharepoint.com/:u:/g/personal/promeroc_unmsm_edu_pe/EWUECqgWUD5KrW9HJg-6P0sBXaPYKlUpte5We8SCNHKeCg?e=yeaeDZ)

lutzomya.fasta

[lutzomya.fasta](https://unmsmmail-my.sharepoint.com/:u:/g/personal/promeroc_unmsm_edu_pe/EcwiDlc1UCJAuwvbcXZ9dZMBEfCV7oqkpBSiuLGVSXqcqg?e=w3NVmA)

**23 abril 2024**

Participación de hoy

<https://www.sciencedirect.com/science/article/pii/S0001706X14003167>

esearch -**db** nucleotide -**query** "AB921329:AB921431 [ACCN]" | efetch -**format** acc | epost -**db** nucleotide | efetch -**format** fasta

Para bajar las secuencias el archivo debe tener los codigos de acceso y llamarse “secuencias”

**Entrez Direct**

Si no reconoce nuccore colocar nucleotide

efetch –**db** nuccore –**format** gb –**id** AF086833 | head   
    
efetch –**db** nuccore –**format** gb –**id** AF086833 > AF086833.gb

cat AF086833.gb  
   
# Accession number AF086833 in Fasta format.   
   
efetch **-db**nuccore –**format f**asta –**id** AF086833 > AF086833.fa   
   
# efetch can take additional parameters and select a section of the sequence.   
   
efetch –**db** nuccore –**format** fasta –**id** AF086833 –**seq\_start** 1 –**seq\_stop** 3    
efetch –**db** nuccore –**format** fasta –**id** AF086833 –**seq\_start** 1 –**seq\_stop** 5 –**strand** 1    
   
efetch –**db** nuccore –**format** fasta –**id** AF086833 –**seq\_start** 1 –**seq\_stop** 5 –**strand** 2    
   
esearch -db nucleotide -query PRJNA257197    
   
esearch -db protein -query PRJNA257197    
   
esearch -db nucleotide -query PRJNA257197 | efetch –**format** fasta > genomes.fa    
   
esearch -db protein -query PRJNA257197 | efetch –**format** fasta > proteins.fa    
   
esearch -db sra -query PRJNA257197 | efetch -format runinfo > runinfo.csv 

#si no funciona no mirarlo : )  
cat runinfo.csv | cut -d , -f 1,2,16 | head -3    
   
efetch -db taxonomy -id 9606,7227,10090 -format xml | xtract -Pattern Taxon -first TaxId ScientificName GenbankCommonName Division

**19 abril 2024**

Probaremos dos formas de instalar Entrez Direct en Linux

*Más sencilla*

conda install bioconda::entrez-direct

(las computadoras deben tener instalada la aplicación “conda”)

*Menos sencilla*

sh -c "$(curl -fsSL <https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh>)"

export PATH=${HOME}/edirect:${PATH}

esearch -**db** pubmed -**query** "selective serotonin reuptake inhibitor"  
esearch -**db** pubmed -**query** "lycopene cyclase" | efetch -**format** medline  
esearch -**db** protein -**query** "lycopene cyclase" | efetch -**format** fasta

**Ejercicio 005**

Buscar una secuencia de nucleótidos en la base de datos Nucleotide y descargarla como fasta usando Entrez Direct

**Ejercicio 006**

<http://www.scielo.org.pe/pdf/rpb/v18n2/a12v18n2.pdf>

Descargar las 13 secuencias en formato fasta

epost -**db** nuccore -**input** secuencias -**format** acc | efetch -**format** fasta > seq.fa

JF414806

JF414807

JF414808

JF414809

JF414810

JF414811

JF414812

JF414813

JF414814

JF414815

JF414816

EU239241

GQ370443

Usamos los comandos

**ls**  
**cd**  
**cd** ..  
**clear**

cat

grep

Diferencia entre nuccore y nucleotide

<https://www.biostars.org/p/161430/>

**Ejercicio 007**

Descargar el archivo ejemplo.fastq de la carpeta compartida BB y observarlo

<https://unmsmmail-my.sharepoint.com/:u:/g/personal/promeroc_unmsm_edu_pe/EVP3cbso1jxKkb41SCr2g6oBdzRfr2kBBn-r9UitIzMmIA?e=leUjdp>

**16 abril de 2024**

Recordemos los conceptos de las siguientes bases de datos:

BioProject, BioSample, SRA, Nucleotide, Genome

**Ejercicio 001**

Busca una bacteria de ocurrencia en Perú en el NCBI usando la opción “All databases” ¿Qué resultados se encuentra?

**Ejercicio 002**

Busca un gen bacteriano en el GenBank ¿Qué se encuentra? ¿Cuántos resultados hay? ¿De qué tamaño son las secuencias?

**Divisiones de Genbank**

<https://www.ncbi.nlm.nih.gov/genbank/htgs/divisions/>

**Diferentes códigos genéticos**

<https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c#SG11>

**Ejercicio 003**

Buscar en la base de datos el genoma de una especie con ocurrencia en el Perú. ¿Qué característica tiene?

<https://www.ncbi.nlm.nih.gov/nuccore/NC_045370.1>

**Ejercicio 004**

Buscar un artículo científico en PMC, extraer los códigos de acceso o números de acceso y descargar las secuencias de dicho artículo (obtener secuencias en FASTA)

O usar este artículo para conseguir las secuencias

<http://www.scielo.org.pe/pdf/rpb/v18n2/a12v18n2.pdf>

**Me pidieron:**

GEO profile

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE251907>

Base de datos de data sets de expresión genética

Guías para ensamblar virus

<https://training.galaxyproject.org/training-material/topics/assembly/tutorials/assembly-with-preprocessing/tutorial.html>

Cómo descargar los fastq desde SRA

wget --output-document sratoolkit.tar.gz https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz

tar -vxzf sratoolkit.tar.gz  
export **PATH**=$PATH:$PWD/sratoolkit.3.0.0-mac64/bin  
which fastq-dump  
fastq-dump --stdout -X 2 SRR390728

Entrez Direct

[Entrez Direct: E-utilities on the Unix Command Line - Entrez Programming Utilities Help - NCBI Bookshelf (nih.gov)](https://www.ncbi.nlm.nih.gov/books/NBK179288/)

Blue mussel, *Mytilus edulis*; fruitfly, *Drosophila yakuba*; honeybee, *Apis mellifera*; mosquito, *Anopheles gambiae*; brine shrimp, *Artemia franciscana*; nematodes, *Ascaris suum* and *Caenorhabditis elegans*; sea urchin, *Strongylocentrotus purpuratus*; carp, *Cyprinus carpio*; frog, *Xenopus laevis*; chicken, *Gallus gallus*; mouse, *Mus musculus*; cow, *Bos taurus*; fin whale, *Balaenoptera physalus*; human,

zgrep --color=always "TGGAATTCTCGG" file1\_R1.1P.cutadapt.fastq.gz | head