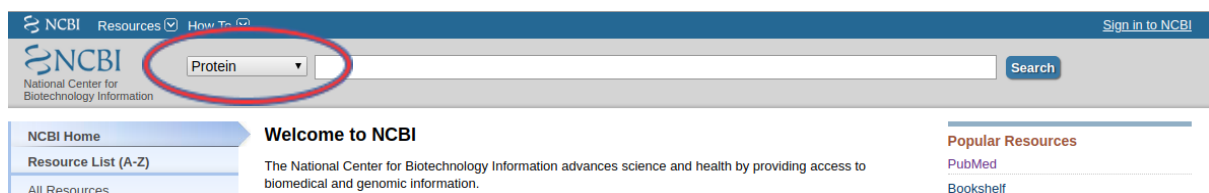


¿En qué se parecen una gallina y una mosca?

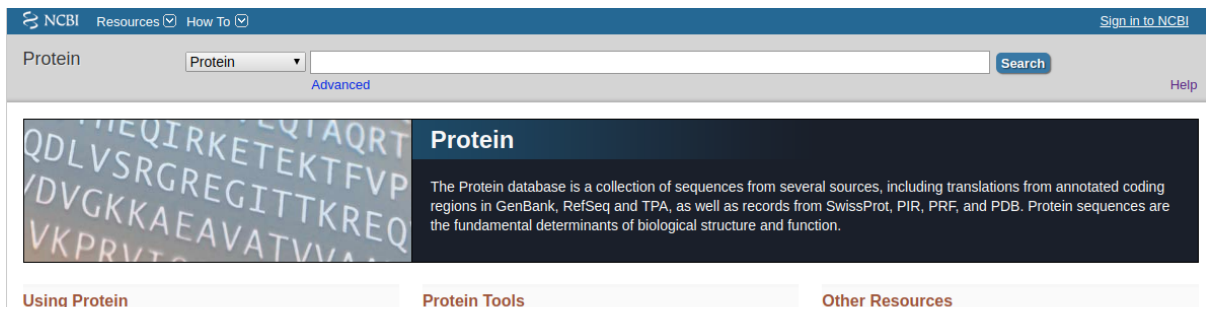
Uso y desusos de las Bases de Datos Biológicas

Hemos usado durante los talleres mucha información biológica, como secuencias de ADN, de ARN y de proteínas. ¿De dónde viene esa información? En términos generales los científicos de todo el mundo desarrollan diversos conocimientos relacionados con los seres vivos. Estos conocimientos se obtienen a base de observaciones y experimentación. Los datos y conclusiones obtenidos son compartidos entre científicos de forma organizada, ya sea por medio de publicaciones en revistas super-archi-nerds o, por ejemplo, a través de Bases de datos disponibles en internet. Una base de datos (BD) es una colección estructurada de datos; en particular, una base de datos biológica es una colección de información relacionada con seres vivos. Estos datos provienen de experimentos científicos, literatura publicada, análisis computacional, etc. La información contenida en bases de datos biológicas puede incluir, por ejemplo: funciones, estructura y localización de proteínas o genes, efectos clínicos de mutaciones, así como similitudes de secuencias o distancias evolutivas, etc. Entre las bases de datos más utilizadas por científicos de todo el mundo, bioinformáticos o no, se encuentran GenBank (colección de todas las secuencias biológicas estudiadas) y PDB (que guarda la información estructural disponible acerca de ácidos nucleicos y proteínas).

El *citocromo c* es una proteína presente en la mayoría de los organismos e involucrada en la respiración aeróbica. En humanos y chimpancés está formado por los mismos 104 aminoácidos, ordenados exactamente igual. ¿Podríamos ver entonces qué relación guarda nuestra proteína con la de una mosca o un pollo, comparando sus secuencias? El primer paso para nuestra comparación entre organismos es buscar nuestras secuencias en las bases de datos. En el buscador general del NCBI (<https://www.ncbi.nlm.nih.gov>) se pueden buscar, entre otras cosas, secuencias de ADN o proteínas. Como la comparación la haremos sobre la base de las proteínas de los organismos, debemos elegir en el menú desplegable la opción **Protein**; así estaremos indicando al buscador que acceda a la base de datos correspondientes a secuencias de proteínas.

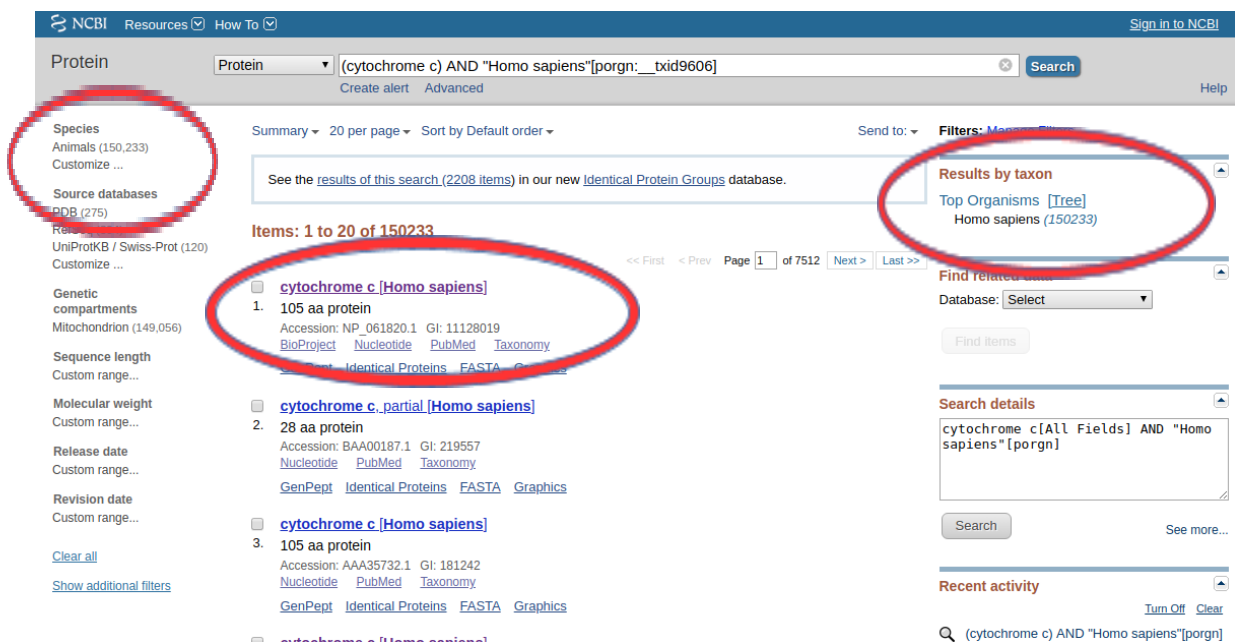


Alternativamente, podríamos acceder en forma directa a la página web de **NCBI Protein** (<https://www.ncbi.nlm.nih.gov/protein>).



Si buscamos el término **cytochrome c**, obtendremos un listado de secuencias del citocromo c en múltiples organismos. Si quisiéramos acceder a una secuencia en particular, para verla en pantalla o guardarla en un archivo en nuestra computadora, deberíamos hacer clic en el ítem que nos interese del listado de resultados. Podemos usar los filtros disponibles en los menús laterales para facilitar la búsqueda del citocromo c en un organismo en particular. Como alternativa, podríamos realizar una nueva búsqueda en el buscador de la página pero agregando el nombre del organismo en el que estamos interesados.

- Ingreseemos a la entrada correspondiente al citocromo c de **Homo sapiens**.



Al inspeccionar uno de los resultados de la búsqueda en detalle, observamos toda la información disponible acerca de la secuencia de esa proteína: de qué organismo proviene (**ORGANISM**), quiénes son los investigadores que la determinaron (**AUTHORS**), la referencia al artículo donde se publicó (**TITLE, JOURNAL, PUBMED**), algunas características de esa secuencia, etc. Al final de la página está la secuencia de aminoácidos de la proteína.

- Para conservar esta secuencia podríamos simplemente copiar, pegar y guardar la secuencia en un archivo de Word, el procesador de texto. Sin embargo, solemos trabajar con archivos de texto sencillos, que respetan un

formato estandarizado llamado **FASTA**. Podemos usar el botón “enviar a” (**Send to**) en la parte superior de la página y elegir descargar un archivo (**File**) que contendrá la secuencia deseada en formato FASTA.

GenPept

cytochrome c [Homo sapiens]

NCBI Reference Sequence: NP_061820.1

[Identical Proteins](#)
[FASTA](#)
[Graphics](#)

Go to:

LOCUS NP_061820 105 aa linear PRI 04-MAY-2019

DEFINITION cytochrome c [Homo sapiens].

ACCESSION NP_061820

VERSION NP_061820.1

DBSOURCE REFSEQ: accession [NM_018947.5](#)

KEYWORDS RefSeq; RefSeq Select.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Send to:

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Show in Genome Data Viewer

...

/db_xref="CCDS:[CCDS5393.1](#)"

/db_xref="GeneID:54205"

/db_xref="RSCC:[RSCC_43960](#)

/db_xref="MIM:123970"

ORIGIN

1 mgdvekgkki fimkcsqcht vekgghktg pnhlgfgrk tggapgsys aanknkgii

61 gedtlmleyle npkkyipgatk mifvgikkke eradliaylk katne

//

Para tener con qué comparar a la proteína humana, el primer paso sería repetir la búsqueda para el pollo, la mosca y cada uno de los organismos que nos interesen.

- Trabajaremos con las secuencias del citocromo c de nueve organismos, listados en la tabla de abajo. Además de su denominación taxonómica, deberíamos conocer su nombre común: intentemos completar la tabla.
- Podríamos descargar las nueve secuencias a nuestra computadora, luego abrir cada archivo, copiar la secuencia y pegarla en un nuevo archivo que finalmente contendrá todas las secuencias descargadas. Para facilitar el trabajo en este curso, les proveemos un archivo que ya contiene las secuencias a comparar (pueden descargarlo desde [aquí](#)).

Secuencia	Nombre taxonómico	Nombre común
NP_061820.1	<i>Homo sapiens</i>	Humano
NP_001072946.1	<i>Gallus gallus</i>	
NP_001065289.1	<i>Pan troglodytes</i>	
NP_001157486.1	<i>Equus caballus</i>	
NP_001183974.1	<i>Canis lupus familiaris</i>	
AEP27192.1	<i>Gorilla gorilla</i>	
XP_024245566.1	<i>Oncorhynchus tshawytscha</i>	
NP_001086101.1	<i>Xenopus laevis</i>	

NP_477164.1	<i>Drosophila melanogaster</i>	
-------------	--------------------------------	--

Si abrimos el archivo de secuencias guardadas, las dos primeras se ven así:

```
>NP_061820.1 cytochrome c [Homo sapiens]
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLE
NPKKYIPGTMIFVGIKKKEERADLIAYLKATNE
>NP_001072946.1 cytochrome c [Gallus gallus]
MGDIKGGKKIFVQKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAEFGFSYTDANKNKGITWGEDTLMEYLE
NPKKYIPGTMIFAGIKKSERVDLIAYLKDATSK
```

Para facilitar el trabajo en comunidad, los científicos suelen adoptar convenciones: formas típicas de hacer las cosas. El formato FASTA es una de ellas. En el formato FASTA cada secuencia se almacena, al menos, en dos líneas.

La primera línea de cada secuencia se llama encabezado. Esta línea, que siempre comienza con el símbolo mayor (>), describe a la secuencia: en nuestro caso incluye el código identificador de la secuencia en la base de datos original, el nombre de la proteína y el organismo al que pertenece.

En las siguientes líneas se encuentra la secuencia de aminoácidos correspondiente.

Para comparar las secuencias es necesario poder alinearlas. Realizar un **alineamiento de secuencias** consiste en escribirlas una debajo de otra y comparar, posición a posición, la identidad de los aminoácidos presentes en todas las secuencias, para ver cuáles y cuántos aminoácidos son iguales (o parecidos) en una misma posición. Idealmente, los aminoácidos equivalentes de las distintas secuencias deberían quedar *alineados* en la misma columna. Podemos incluir un guión (gap) en una posición dada de una secuencia para indicar que ésta carece de un aminoácido presente en otra(s) secuencias, dentro de la misma columna.

- 1) ¿Cuán sencillo será alinear dos o más secuencias a mano? ¿Cuánto influirán el número de secuencias a alinear, su longitud, y la similitud entre ellas?
- 2) ¿Son parecidos los citocromos c de humano y gallo?
- 3) ¿Qué teorías subyacen a este análisis? ¿Cómo nos ayuda la evolución a explicar sus similitudes y diferencias?

Cuando el alineamiento manual de secuencias es complicado, podemos usar programas que realizan alineamientos de una forma optimizada y rápida. Uno de estos programas se llama Clustal. Podemos acceder a éste de distintas maneras: por ejemplo, la última versión, llamada Clustal Omega, está disponible en el sitio del Laboratorio Europeo de Bioinformática (*European Bioinformatics Institute, EBI*) (<https://www.ebi.ac.uk/Tools/msa/clustalo/>).

- Clustal Omega recibirá nuestras secuencias sin alinear: podemos copiarlas desde un archivo y pegarlas en la ventana de búsqueda, o utilizar el botón de subida de un archivo (**upload a file**) para cargar el archivo desde nuestra computadora. Finalmente, presionaremos el botón de envío (**Submit**).

Clustal Omega

Input form

Web services

Help & Documentation

Bioinformatics Tools FAQ

Feedback

Share

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or, upload a file: Seleccionar archivo No se eligió archivo

Use a [example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

Luego de unos minutos obtendremos nuestro alineamiento.

- 4) Podemos elegir verlo en colores (**Show Color**). ¿Qué indican los colores?
- ¿Qué indican el guión (-), los dos puntos (:) y el asterisco (*)?
- 5) A simple vista, ¿se conserva la secuencia del citocromo c en los organismos?
- 6) ¿Creeríamos que todos los organismos se asemejan por igual al resto, o se pueden identificar grupos de mayor similitud? Si es así, ¿tienen sentido?
- 7) ¿Qué evidencias nos aportaría este análisis, a la luz de la evolución?

Alignments Result Summary Phylogenetic Tree Submission Details

Download Alignment File Hide Colors View result with Jalview Send to Simple Phylogeny Send to MView

CLUSTAL O(1.2.4) multiple sequence alignment

```

NP_477164.1      MGSQDAENGKKIFVQKCAQCHTYEVGGKHVGPNLGGVGRKCGTAAGYKYTDANIKKGV      60
XP_024245566.1  -MGDIKGGKAFVQKCAQCHTYEVGGKHVGPNLWGLFGRKTGQAEFSTYDANKKGI      58
NP_061820.1     -MGDVEKGGKIFIMKCSQCHTYEVGGKHVGPNLHGLFGRKTGQAPGYSYTAANKNGI      58
NP_001865289.1  -MGDVEKGGKIFIMKCSQCHTYEVGGKHVGPNLHGLFGRKTGQAPGYSYTAANKNGI      58
AEF27192.1     -MGDVEKGGKIFIMKCSQCHTYEVGGKHVGPNLHGLFGRKTGQAPGYSYTAANKNGI      58
NP_00186101.1  -MGDVEKGGKIFVQKCAQCHTYEVGGKHVGPNLHGLFGRKTGQAEFSTYDANKNGI      58
NP_001872946.1  -MGDIKGGKIFVQKCAQCHTYEVGGKHVGPNLHGLFGRKTGQAEFSTYDANKNGI      58
NP_001157486.1  -MGDVEKGGKIFVQKCAQCHTYEVGGKHVGPNLHGLFGRKTGQAPGYSYDANKNGI      58
NP_001183974.1  -MGDVEKGGKIFVQKCAQCHTYEVGGKHVGPNLHGLFGRKTGQAPGYSYDANKNGI      58
               ** * ** * : ** * ** * ** * ** * : ** * ** * ** *
NP_477164.1      TWTEGNLDVLPKKYIPGTMIFVAGLKKAEERADLIAYLKSNK--      105
XP_024245566.1  VMDTDTLMYLENPKKYPGTMIFVAGLKKKGERADLIAYLKSAIS-      104
NP_061820.1     IWGEDTLMEYLENPKKYPGTMIFVAGLKKKGERADLIAYLKATNE      105
NP_001865289.1  IWGEDTLMEYLENPKKYPGTMIFVAGLKKKGERADLIAYLKATNE      105
AEF27192.1     IWGEDTLMEYLENPKKYPGTMIFVAGLKKKGERADLIAYLKATNE      105
NP_00186101.1  VMDTDTLMYLENPKKYPGTMIFVAGLKKKGERADLIAYLKQSTSS      105
NP_001872946.1  TWGEDTLMEYLENPKKYPGTMIFVAGLKKKSERVDLIAYLKDATSK      105
NP_001157486.1  TWKEETLMEYLENPKKYPGTMIFVAGLKKKTEREDLIAYLKATNE      105
NP_001183974.1  TWGEETLMEYLENPKKYPGTMIFVAGLKKKTERADLIAYLKATKE      105
               *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *

```

Alinear secuencias nos permite ver las similitudes y diferencias que hay entre ellas. Este es el primer paso para intentar conocer su evolución. La reconstrucción de los vínculos evolutivos entre secuencias u organismos es una tarea ardua y difícil de validar, porque nos resulta imposible conocer la historia evolutiva real de las proteínas y los organismos. Sin embargo, existen **métodos computacionales que nos permiten inferir los vínculos evolutivos e inferir un árbol filogenético** a partir de un alineamiento de proteínas. Las herramientas de inferencia filogenética son sencillas de manejar pero complejas de dominar. Para comenzar, Clustal nos

provee un sencillo árbol filogenético (***Phylogenetic Tree***) para las secuencias alineadas.

- 8) A juzgar por los organismos participantes, ¿cuáles creería que deberían estar más agrupados en el árbol filogenético?
- 9) **Observemos el árbol filogenético. ¿Concuerda con lo esperado? ¿De qué organismos son los citocromos c más parecidos? ¿Cómo se explica?**

