Taller01: Búsquedas y uso de Herramientas deln NCBI a través del Sistema Entrez

13 de agosto de 2018

1. Objetivo

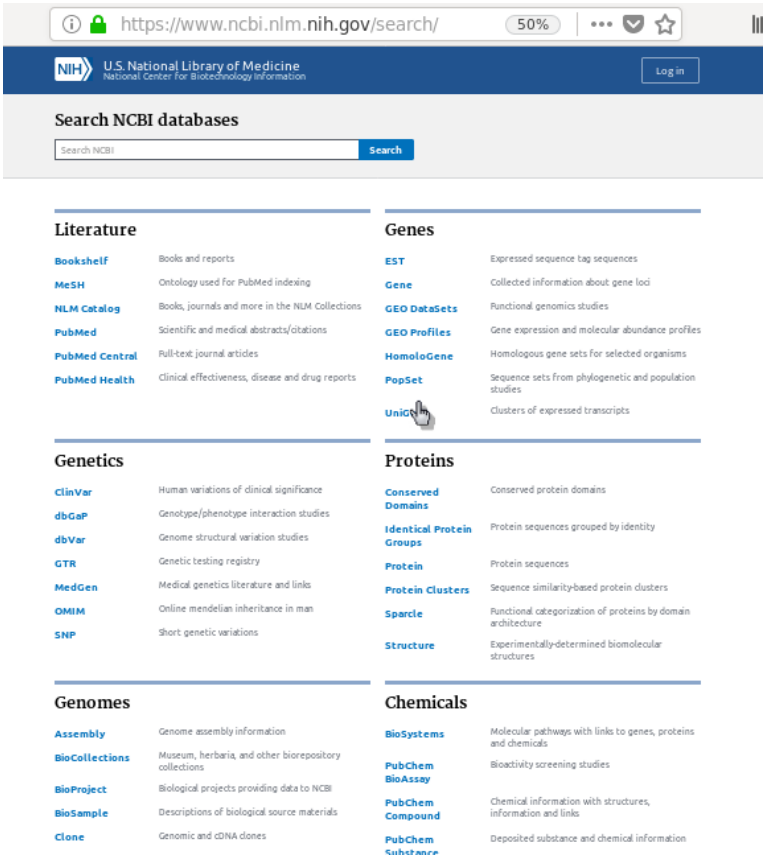
Utilizar las herramientas del NCBI para tratar de buscar posibles funciones de una secuencia anónima de nucleótidos y buscar si presenta alguna relación con otras secuencias depositadas en la base datos de genes.

2. Introducción

El término ORF se refiere a una parte de una secuencia de nucleótidos que tiene la potencialidad de codificar un péptido o una proteína; es decir, que debe contener un codón o triplete de iniciación y un codón de terminación.

La búsqueda de ORF's en una secuencia de nucleótidos tiene muchas utilidades en genética molecular. Por ejemplo, puede ayudar a la predicción de genes, a la determinación del origen de pseudogenes, etc.

Acceso a Entrez a través de: `http://www.ncbi.nlm.nih.gov/Entrez/`



3. Guía

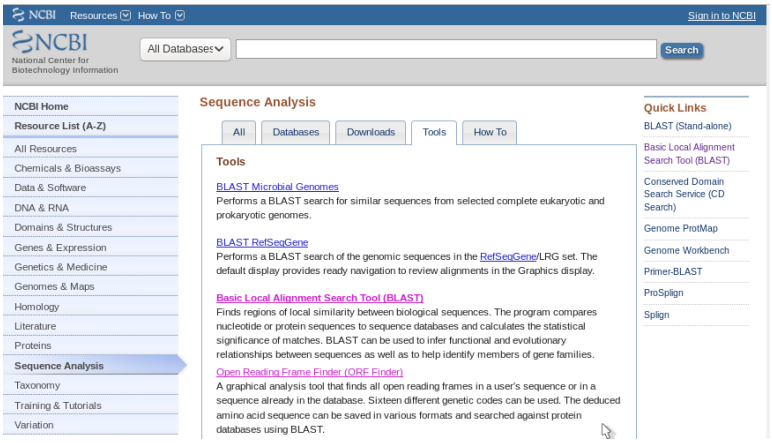
Se tiene una secuencia de cDNA (ver abajo) de la cual queremos saber si presenta alguna relación con secuencias de genes ya conocidos, y si es así, tratar de deducir su posible función fisiológica, metabólica, celular.

```
TTGCGCGAGGGCCGCCCCAGCCGCCCGATCAGCGTCGGCCCCTGCGCTGGCTGAGGGGGAATCACCCCCCGCTGCCAAAC
GCAGGCACGGACCCCAGTCCTGCGGCGCGCTGGCGGGCAGGTCAGACCGAGCCCTCCAGTGAATCAGGCGACAAATACAA
CGCCAGGGATCAGCGACGGCAGCGCTACAAGAATCGACGGCGTGTTGGAGGGACCCAGCTCTATGCCGAGTTCAGTCGCC
AGCAGAGGTACTGCCGTGACTGGGAAGGCGGACCCCAGTAGCGCCTGTGACGCGGGGACGGGTCAGCGTCTGCCGGTAGC
CGGACCGCCCAGGCAGGCTGACCAAGCGGCGCCTGTGGCAGTTGGCCCAACAAACGAGCACGTCCCCCGAACCACGACGC
TGAGCGACAGTCTAGGCCGTGCTAGTGAACGCACGTGAGCCACGAGACACTGCAAGTCAGCCGGACTCCTCGTCGGGTCG
CCCGCTAGCCCTACCTCGAGTGTCACCCCAGACCTTAAGCGCTGGCTCATTAGGGTAGGCTCAATCGGGCGGCGACGGCT
CTGGCCCCACAGACTGGGTAACCTTGCCAACCAGGCCTCGATGAAGTGGGTGTGGGCCCTGGCCCTGCTGGCCGCCTGGG
CCGCCGCCGAGCGCGACTGCCGCGTGAGCAGCTTCCGCGTGAAGGAGAACTTCGACAAGGCCCGCTTCAGCGGCACCTGG
TTCGCCTTGGCCAAGAAGGACCCCGAGGGCCTGTTCCTGCAGGACAACTTCGTGGCCGAGTTCAGCGTGGACGAGACCGG
CCAGATGAGCGCCACCGCCAAGGGCCGCGTGTGCCTGCTGAACAACTGGGACGTGTGCGCCGACAAGGTGGGCACCTTCA
CCGACACCGAGGACCCCGCCAAGTTCAAGATGAAGTACTGGGGCGTGGCCAGCTTCCTGCAGAAGGGCAACGACGACCAC
TGGATCGTGGACACCGACTACGACACCTACGCCGTGCAGTACAGCTGCCGCCTGCTGAACCTGGACGGCACCTGCGCCGA
CGACTACAGCTTCGTGTTCAGCCGCGACCCCAACGGCCTGCCCCCCGAGGCCCAGAAGATCGTGCGCCAGCGCCAGGAGG
AGCTGTGCCTGGCCCGCCAGTACCGCCTGATCGGCCACAACGGCTACTGCGACGGCCGCAGCGAGCGCAACCTGCTGTAA
GGTGCGGCGCGTCTACTCTCCCCCATCCCGCTAGGGCCTGCGGTGTCGCCGGACCAGGGTCACACCAACCATCTCACGTG
CGGCCCTGTCGTTCACCCTCCCATCCACTGACGGCGCGGCAGACCCGGGAGGCCGGCCGACAGCTGGCGGCGCGCGCGCC
AAGGGAGCTGGCGACGACTCCCGCCACCGGCTGGACGAAAATGGCAAGTCTAGGCGCCAGCGTCGCCGTAGAAGGGGTCG
TTTTCAGCCGGACATGGCCGGACACAGTGGCCAGTGGCCGATCGTCAAGGCCTGATGCCCGTAAATGGCTTCGTTAGTTT
GGCGCCCAGGCGCTCGGCCAGCCGGGCGTTAGAGCAACACAGAGTTCGCCGCACCGCCCGCCCGAACGCAGAGAAGGCGA
GAGCGAGTTTGGCCTATAGGAGCCGCCGCTCACGGCAACCGCGCACAAACGGGTAAGCCCTCGGCCCAACCCCGGAGTGC
TTAACCCGGCTTCGCGAAGGCCGAAGTCCGGAGAGGACAGCATGGGTCGTGCGTCAGGGCCTGGCGTTCGTCGCAAAAAC
```

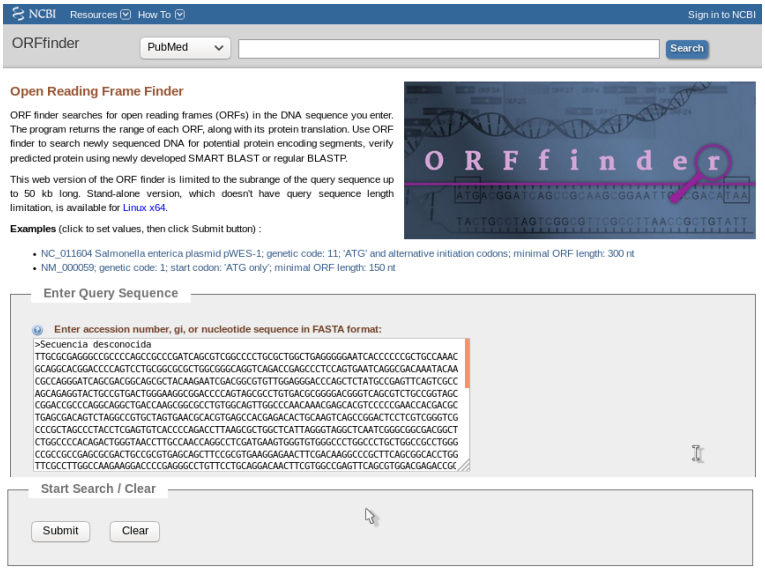
3.1. Búsquedas de “Open Reading Frames” (ORF’s).

1. **Herramientas**. Lo primero que vamos a hacer es tratar de ver si contiene algún marco abierto de lectura (Open Reading Frame–ORF), es decir, si contiene un conjunto de codones que son capaces de traducirse a proteína. Para ello vamos a utilizar la utilidad ***ORF Finder*** que se encuentra en el NCBI (`http://www.ncbi.nlm.nih.gov`).

   Hacemos clic en el vínculo correspondiente a esa utilidad, que se encuentra en la solapa “*Tools*” de la entrada “*Sequence analysis*” y entramos en la página correspondiente a la búsqueda de ORF’s.

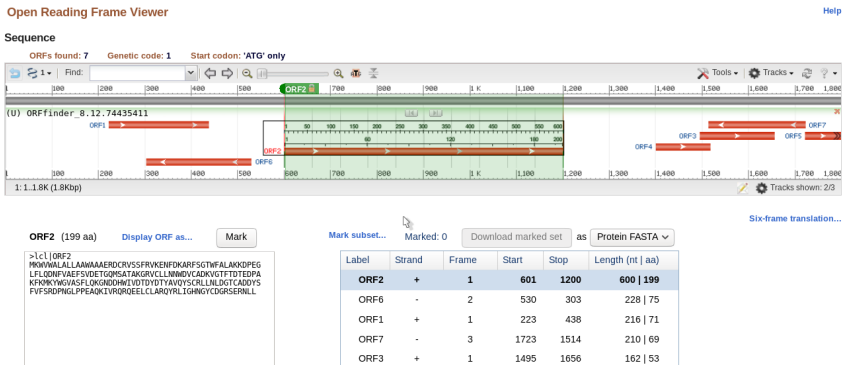


2. **Consulta**. La nueva página te presenta el programa, pudiendo introducir la clave de una de las secuencias ya contenidas en las bases de datos, o una propia. Esto último es lo que vamos a hacer nosotros. En el cuadro grande en blanco vamos a introducir la secuencia problema en formato FASTA (Formato muy utilizado en bioinformática, pues todos los programas bioinformáticos reconocen este formato). Para ello pegamos la secuencia en el cuadro de consulta y una vez que se haya pegado la secuencia hacemos click en OrfFind para ejecutar el programa.



3. **Inspecionar ORF.** El resultado del programa da los posibles ORF's en las dos cadenas (aparecen 3 posibilidades para una cadena y otras 3 para la otra). De todas las ORF's que aparecen en cada una de las 3 pautas de lectura de las hebras plus (+) y minus (-), empezaremos por investigar con la mayor de todas (presenta 600 nucleótidos).

   En la figura siguiente está marcado la ORF que nos interesa. Hacemos click en ella, y aparecerá una nueva pantalla con la ORF seleccionada, ya aislada y con su traducción a proteína.



   Nos quedaremos con la secuencia de la proteína (Parte derecha) que se codificaría a partir de este ORF. Para ello copiaríamos la secuencia y la editaríamos convenientemente utilizando un programa editor de texto (busque uno existente de algún lenguaje: python IDLE o R, en últimas use el bloc de notas), cuidando de ponerla en formato FASTA.

   Este archivo lo utilizaremos en un paso posterior, para ilustrar el uso de la herramienta BLAST

4. Seleccione el ORF más pequeño y responda:

   a) Cúal es la secuencia de aminoácidos de este ORF

   b) Cúal es su secuencia de nucleótidos?

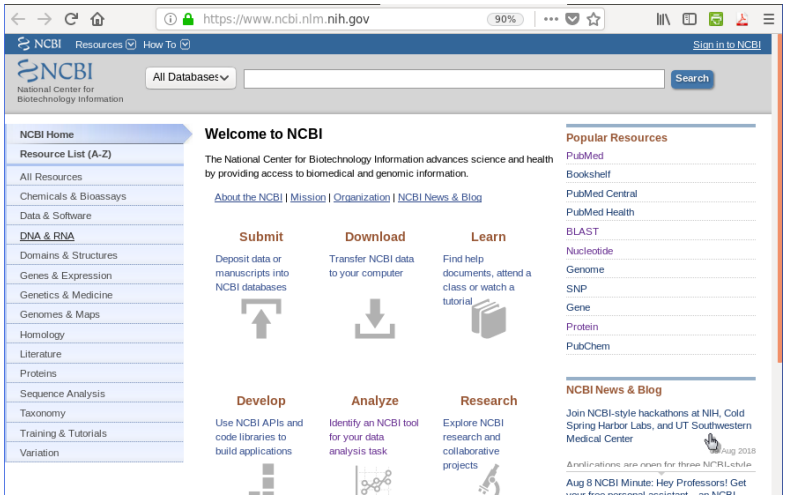
   c) Cúales son el codón de inicio y final que utiliza.

3.2. BÚSQUEDAS DE HOMOLOGÍAS

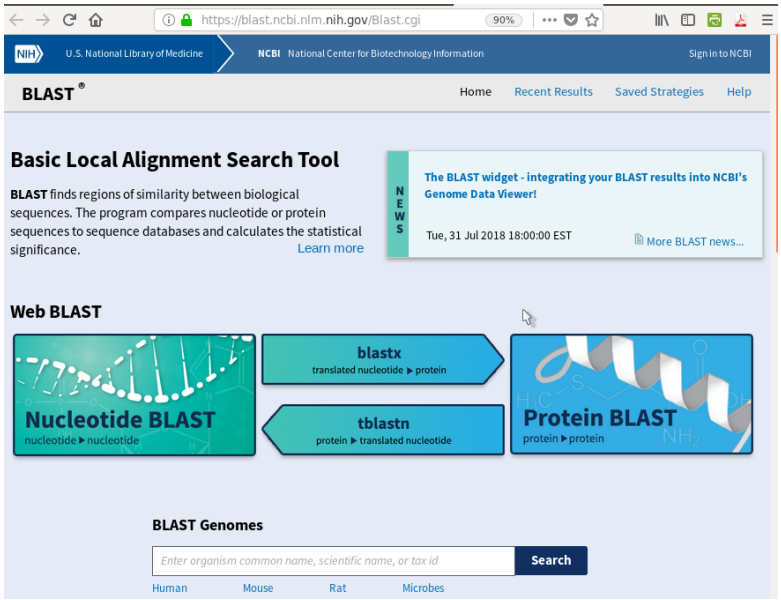
Hasta ahora lo que tenemos es una secuencia de proteína, pero no sabemos nada de ella, ni su función, ni su familia ni el parentesco que guarda con otras proteínas de la misma especie o de otras especies. Conocer la función de una proteína es un trabajo duro de laboratorio; una forma aproximada para saber algo de una proteína problema es buscar en las bases de datos, otras proteínas que tengan parecido (homología) con ella, es decir, tratar de deducir

en la medida de lo posible y por comparación, la familia de proteínas a la que pertenece y su posible función. Uno de los programas más utilizados para buscar parecidos u homologías es ***BLAST*** (Basic Local Alignment Search Tool). Este programa compara una secuencia de proteína o de nucleótidos con una base de datos (de proteínas o de nucleótidos). Nosotros vamos a utilizar la variante *BLASTP* que compara una proteína contra una base de datos de proteínas.

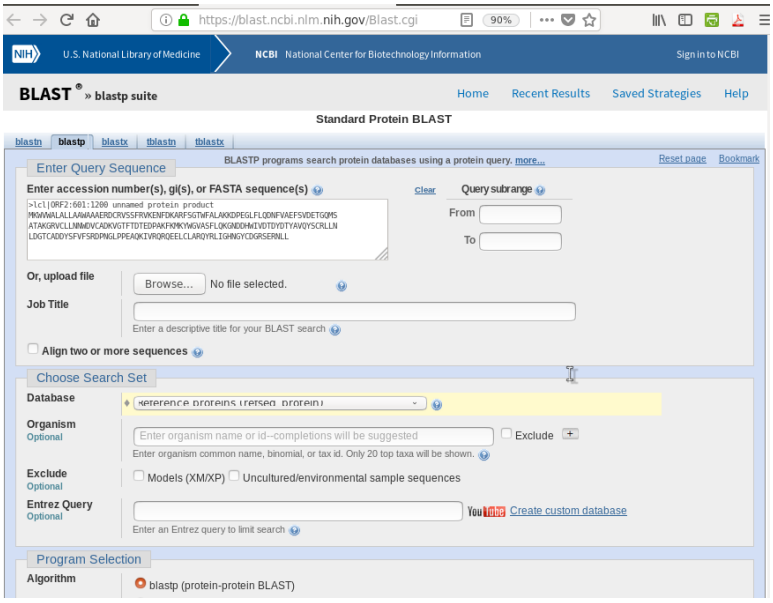
1. Nosotros utilizaremos directamente la herramienta BLAST desde su página de inicio. El enlace lo tenemos en la página inicial del NCBI, en la columna de la derecha (Recursos populares).



Puesto que se trata de una posible proteína, utilizaremos la opción “Protein BLAST”.

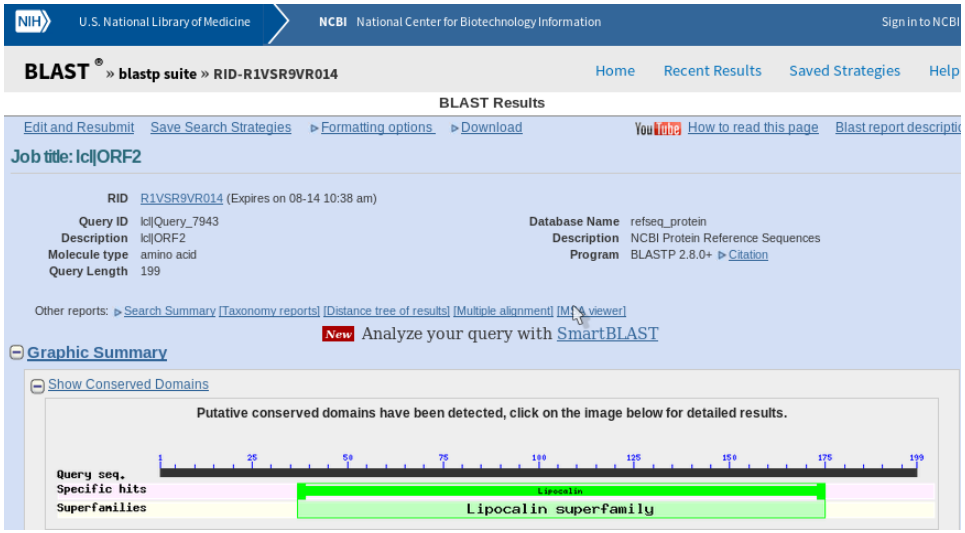


2. Copiamos la secuencia de la proteína problema en la ventana en blanco, y seleccionamos una base de datos de proteínas contra la que comparar (Buscar secuencias similares –homólogas- a la nuestra. En este caso hemos elegido la base de datos **Refseq** de proteínas, aunque podríamos haber utilizado otra distinta. **Refseq** tiene la ventaja de que se trata de una colección exhaustiva de secuencias de proteínas no redundantes y bien anotadas.



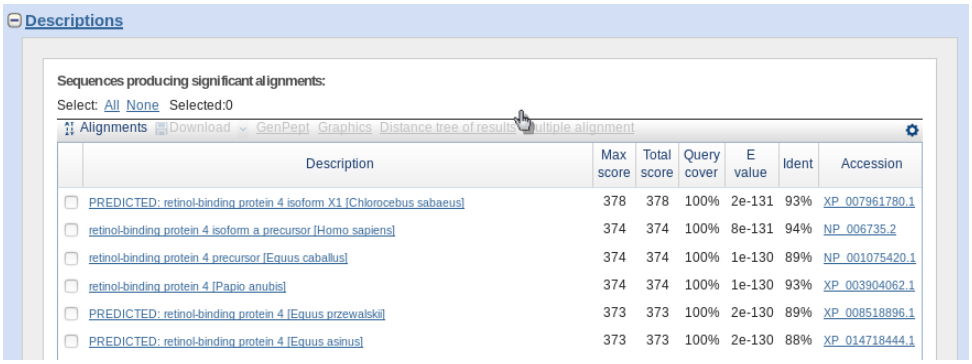
3. Una vez incluida la secuencia de trabajo presionamos en el botón BLAST que aparecerá más abajo en la misma página. Con ello se iniciará el proceso de búsqueda de secuencias similares a la nuestra.

   Durante el proceso de búsqueda de secuencias nos aparecen unas pantallas que ya nos indican de qué tipo de proteína se trata nuestra proteína problema. Una de esas pantallas tiene el siguiente aspecto:



   Como se puede ver, se ha detectado un dominio de Lipocalinas. Si hacemos click en el esquema que muestra el dominio de lipocalina podremos obtener información sobre esas proteínas, e incluso quizá su estructura en 3 dimensiones. Las lipocalinas son pequeñas proteínas con forma de cesta que portan en su interior moléculas hidrofóbicas, y sus funciones son muy variadas.

4. Una vez que esté terminada la búsqueda aparece una pantalla con los resultados.



Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: retinol-binding protein 4 isoform X1 [Chlorocebus sabaeus]	378	378	100%	2e-131	93%	XP_007961780.1
retinol-binding protein 4 isoform a precursor [Homo sapiens]	374	374	100%	8e-131	94%	NP_006735.2
retinol-binding protein 4 precursor [Equus caballus]	374	374	100%	1e-130	89%	NP_001075420.1
retinol-binding protein 4 [Papio anubis]	374	374	100%	1e-130	93%	XP_003904062.1
PREDICTED: retinol-binding protein 4 [Equus przewalskii]	373	373	100%	2e-130	89%	XP_008518896.1
PREDICTED: retinol-binding protein 4 [Equus asinus]	373	373	100%	2e-130	88%	XP_014718444.1

Cada proteína homóloga aparece marcada en azul, si hacemos click en los enlaces que aparecen bajo la columna "**Accession**" podremos ver la información sobre esa proteína, la secuencia, quién la secuenció, otras bases de datos que tengan información sobre esa proteína etc.

5. Responda qué podemos concluir de este análisis:

   a) A que proteína correspondería nuestra secuencia?

   b) A que grupo de proteínas correspondería y cual es la función de estas?

   c) Cuál sería la función de la proteína encontrada.

.

.