Introducción a la Bioinformática

Práctica 1: BLAST y Recuperación de Secuencias

Recuperación de Secuencias

La recuperación de secuencias, es decir la búsqueda y obtención de secuencias de interés en bases de datos, es una de las tareas más comúnes en bioinformática. A primera vista puede parecer una tarea sencilla, pero llegar a hacerlo de una manera realmente efectiva requiere de cierto conocimiento y destreza.

Esta práctica cubrirá con cierta extensión esta labor, y al final de ella serémos capaces de extraer la información precisa de las bases de datos más comúnes, de una manera eficiente.

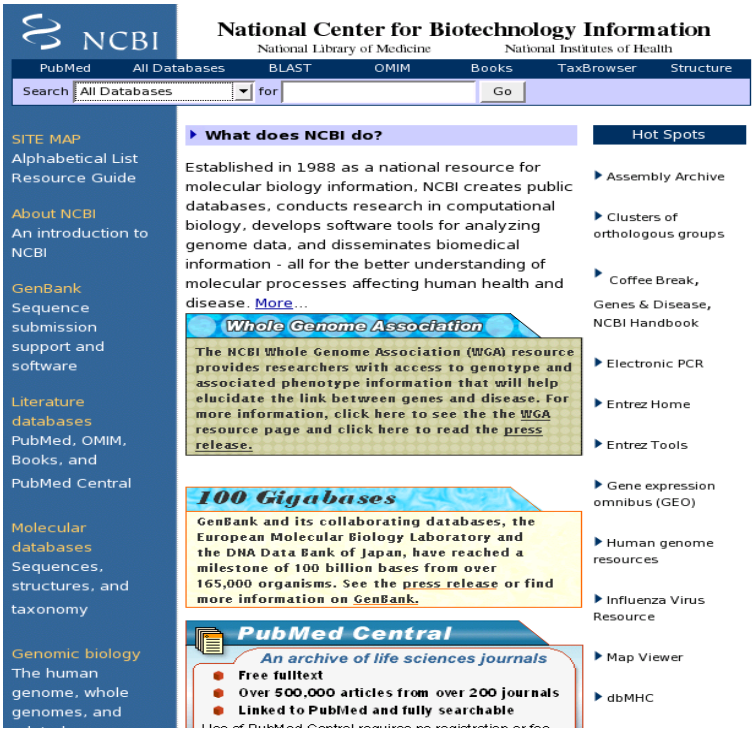
NCBI

Una de las bases de datos más conocidas presentes en el NCBI es el GenBank. Esta base de datos consta de 59,750,386,305 bases en 54,584,635 entradas en las divisiones más comúnes de GenBank (EST/UniGene, STS, GSS HTGS)y 63,183,065,091 bases en 12,465,546 entradas en la división WGS (Febrero 2006).

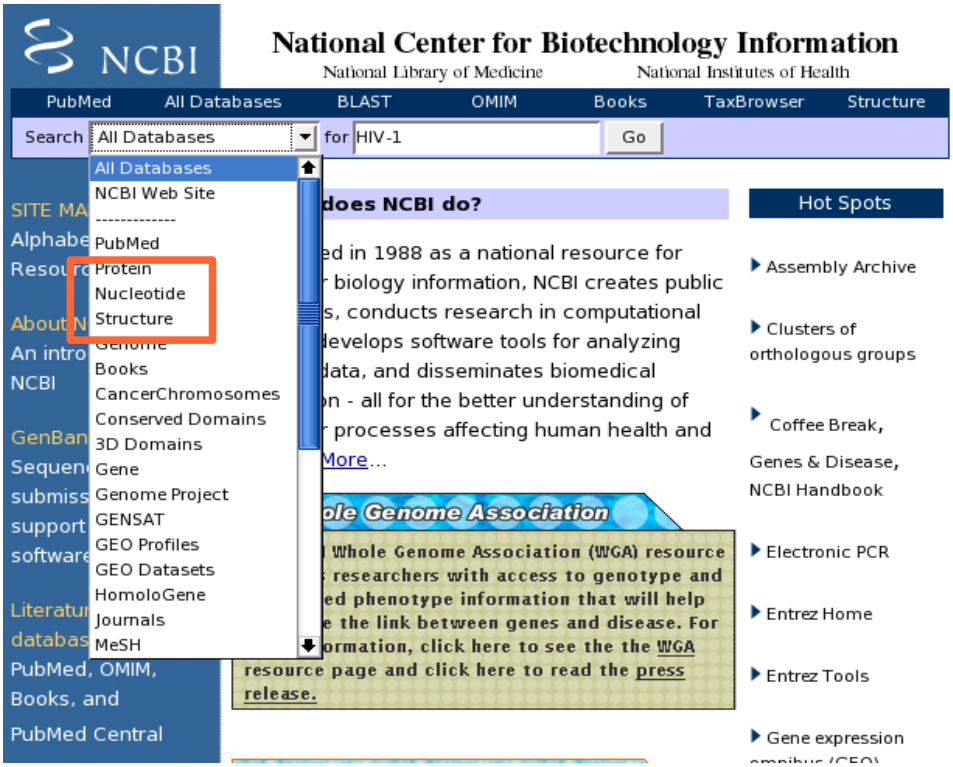
A continuación veremos una de las formas más sencillas de acceder a la información presente en GenBank y el NCBI en general.

Acceda al sitio web del NCBI ubicado en la siguiente dirección:

http://www.ncbi.nlm.nih.gov/



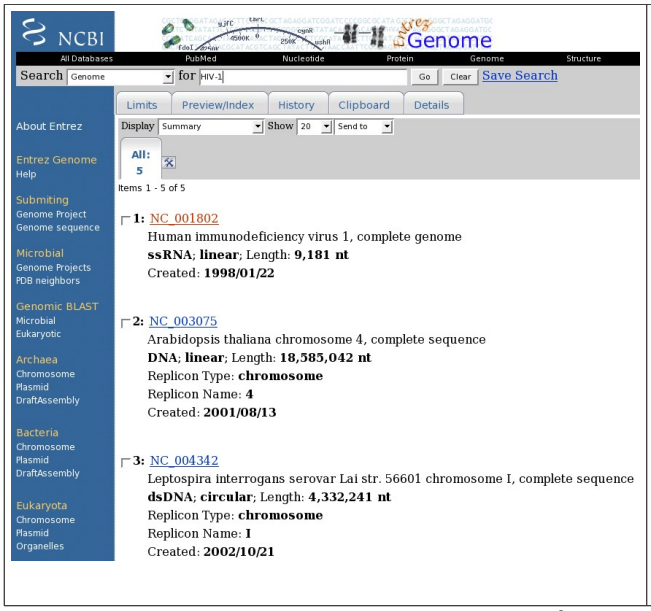
Realizaremos una búsqueda de HIV-1. Asegurese de que ha definido una búsqueda en todas las bases de datos en el menú desplegable ubicado en la esquina superior izquierda y digite el término "**HIV-1**". A continuación presione el botón "**go**".



Obviamente es posible escoger cualquiera de las posibilidades ofrecidas en el menú. Son de destacar Pubmed, Protein y Nucleotide, con las cuales buscamos directamente en la base de datos de bibliografía, DNA o proteínas respectivamente.
Unos segundos después seremos llevados a la página web del sistema **ENTREZ** del NCBI, desde donde tendremos una perspectiva general de la información relacionada con nuestra secuencia presente en el NCBI.

De esta manera es posible saber qué información existe para nuestro término de búsqueda en todo el sitio web del NCBI (ej, 168532 entradas de proteínas, 1 entrada en la sección de taxonomía y 167470 entradas de nucleótidos).
También es posible acceder directamente al sitio web de ENTREZ a través de la siguiente dirección: http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

ENTREZ es, de manera sencilla, el sistema que mantiene unida toda la información presente en el NCBI, algo así como el “GOOGLE” del NCBI, y es quien realiza la búsqueda de nuestro término a través de todas las bases de datos presentes en el NCBI.



A continuación presione el hipervínculo de la sección Genome. Espere unos segundos. Se encontrará con una página de resultados similar a la imagen a su izquierda.

Por el momento dejaremos esta búsqueda ahí y la retomaremos más adelante.

Regrese a la página principal del NCBI: http://www.ncbi.nlm.nih.gov/
Realice de nuevo la búsqueda por **HIV-1**, pero esta vez asegurese de escoger la sección "**genome**" y no "**all databases**". Espere unos segundos y analice la página de resultados que obtiene.

Seguramente ya se ha percatado de que la página de resultados es idéntica a la que se obtuvo mediante el vínculo "**genome**" de la primera búsqueda, hecha en el sistema ENTREZ.

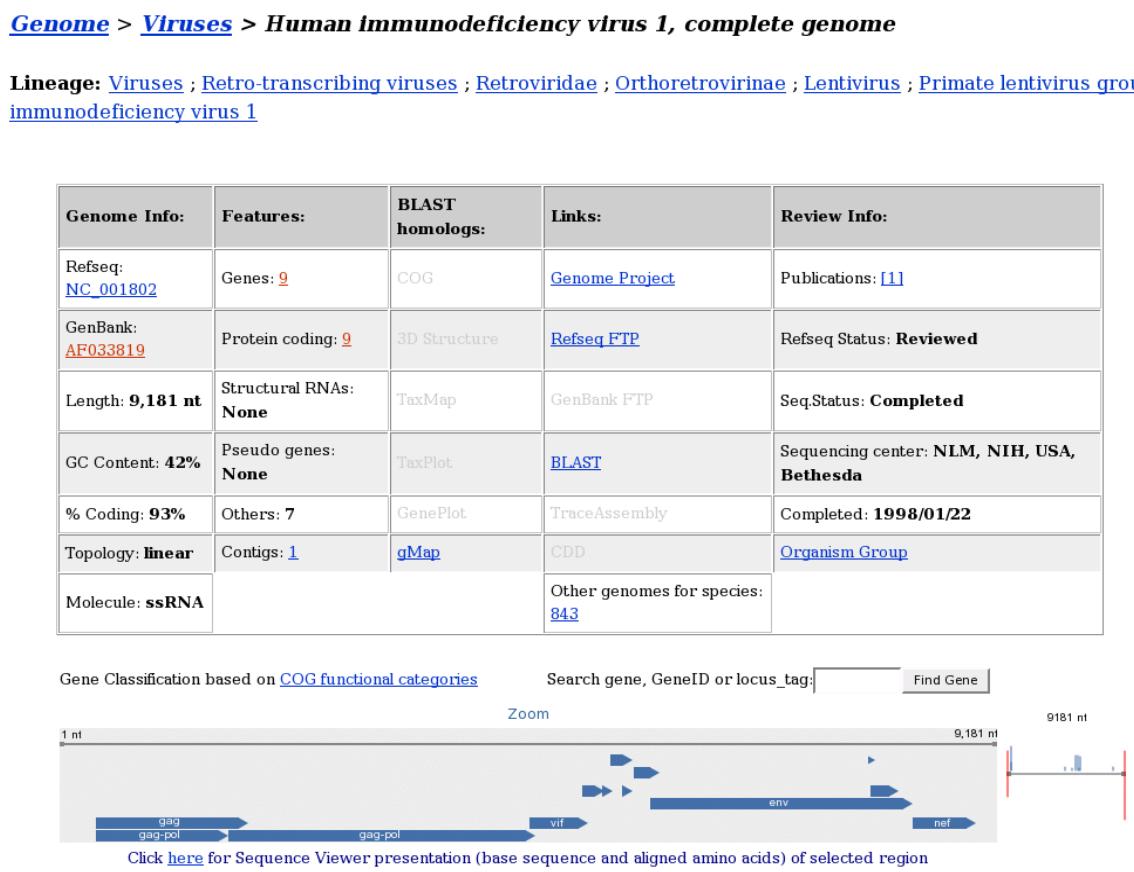
Realice nuevamente la búsqueda en el sistema ENTREZ y explore las diversas entradas que muestra la página de resultados (ej, Protein, UniGene, OMIM, Pubmed etc.). Corrobore dichos resultados con los que arrojan las búsquedas con las opciones en el menú desplegable del sitio web del NCBI.

Accediendo a las secuencias

Ya que ha experimentado con las diferentes bases de datos que ofrece el NCBI y la manera más común de realizar búsquedas en ellas, es momento de conocer la manera en que podemos acceder a los datos que queremos obtener con nuestra búsqueda.

Realice nuevamente la búsqueda de HIV-1 en la sección genome. Encontrará 5 entradas acompañadas de una breve descripción. Siga el hipervínculo para la entrada con identificador: **NC_001802**

En este momento debe encontrarse en una página web cuyo contenido es similar a la siguiente gráfica:



Genome **>** ***Viruses*** **> *Human immunodeficiency virus 1, complete genome***

Lineage: Viruses ; Retro-transcribing viruses ; Retroviridae ; Orthoretrovirinae ; Lentivirus ; Primate lentivirus group ; Human immunodeficiency virus 1

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_001802	Genes: 9	COG	Genome Project	Publications: [1]
GenBank: AF033819	Protein coding: 9	3D Structure	Refseq FTP	Refseq Status: **Reviewed**
Length: **9,181 nt**	Structural RNAs: **None**	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **42%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: **NLM, NIH, USA, Bethesda**
% Coding: **93%**	Others: **7**	GenePlot	TraceAssembly	Completed: **1998/01/22**
Topology: **linear**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **ssRNA**			Other genomes for species: 843	

Gene Classification based on COG functional categories Search gene, GeneID or locus_tag: Find Gene

Zoom

1 nt 9,181 nt 9181 nt

gag gag-pol gag-pol vif env nef

Click here for Sequence Viewer presentation (base sequence and aligned amino acids) of selected region

El cuadro que observa resume la información relacionada con el genoma que hemos buscado. Gracias a éste sabemos que cuenta con 9 genes, que codifican 9 proteínas y que su longitud es de 9181 nucleótidos. Entre otras cosas.

Este tipo de resúmen es necesario cuando tratamos de acceder a este tipo de información, es decir si lo que buscamos es simplemente una proteína o secuencia de ADN, por lo general no seremos llevados a un cuadro de resúmen como éste sino directamente a la entrada de dicha secuencia.

Estudiando la entrada para HIV-1

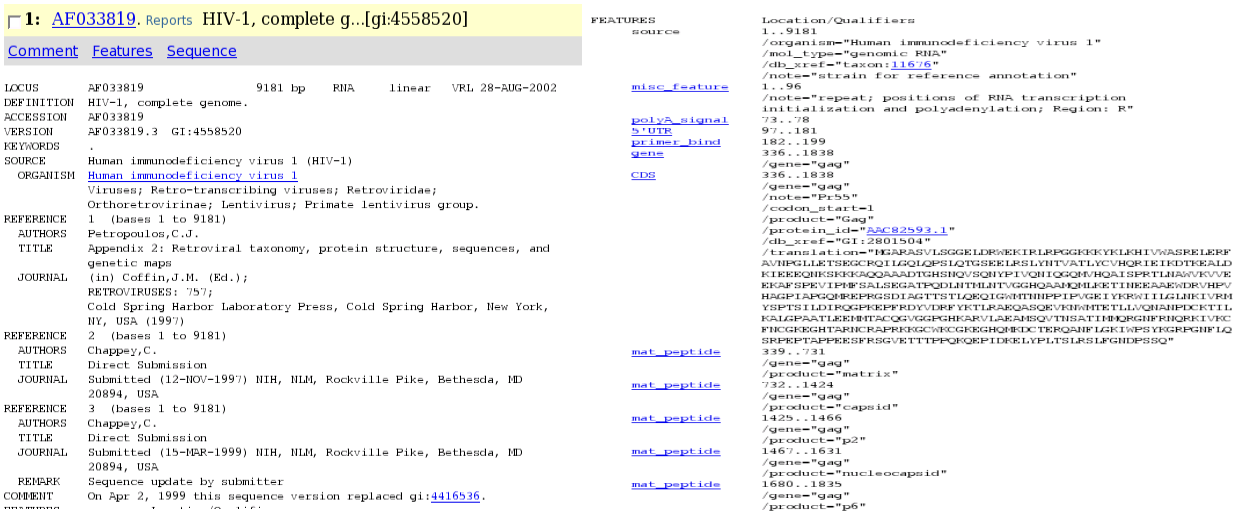
Además de nuestro interés por conocer algunas características del genoma que consultamos, nos resulta interesante obtener también su secuencia completa, para poder acceder a dicha información tenemos que consultar la entrada en genBank para dicho genoma.

Esto se hace siguiendo el hipervínculo al genBank: **AF033819**.

Es importante anotar la diferencia que existe entre el NCBI en general y el genBank, como podemos verlo el genBank es una de las bases de datos del NCBI, si siguieramos los enlaces a proteínas seríamos llevados a la sección de proteínas del NCBI, cuya información proviene de las bases de datos de UNIPROT.

Una vez hemos seguido el vínculo anterior somos llevados a una entrada típica del genBank. Esta consta básicamente de 3 grandes secciones:

1. **Información general**: encontramos el nombre de la entrada, su número de acceso, publicaciones, palabras clave etc.
2. **Anotación**: se encuentran, para este caso, las regiones codificantes de dicha entrada, así como otras anotaciones (ej., regiones repetitivas).
3. **Secuencia**: la secuencia de DNA (o proteína si es el caso) completa.



1: AF033819. Reports HIV-1, complete g...[gi:4558520]

Comment Features Sequence

```
LOCUS       AF033819               9181 bp    RNA     linear   VRL 28-AUG-2002
DEFINITION  HIV-1, complete genome.
ACCESSION   AF033819
VERSION     AF033819.3  GI:4558520
KEYWORDS    .
SOURCE      Human immunodeficiency virus 1 (HIV-1)
  ORGANISM  Human immunodeficiency virus 1
            Viruses; Retro-transcribing viruses; Retroviridae;
            Orthoretrovirinae; Lentivirus; Primate lentivirus group.
REFERENCE   1  (bases 1 to 9181)
  AUTHORS   Petropoulos,C.J.
  TITLE     Appendix 2: Retroviral taxonomy, protein structure, sequences, and
            genetic maps
  JOURNAL   (in) Coffin,J.M. (Ed.);
            RETROVIRUSES: 757;
            Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York,
            NY, USA (1997)
REFERENCE   2  (bases 1 to 9181)
  AUTHORS   Chappey,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-NOV-1997) NIH, NLM, Rockville Pike, Bethesda, MD
            20894, USA
REFERENCE   3  (bases 1 to 9181)
  AUTHORS   Chappey,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-MAR-1999) NIH, NLM, Rockville Pike, Bethesda, MD
            20894, USA
  REMARK    Sequence update by submitter
COMMENT     On Apr 2, 1999 this sequence version replaced gi:4416536.
```

```
FEATURES             Location/Qualifiers
     source          1..9181
                     /organism="Human immunodeficiency virus 1"
                     /mol_type="genomic RNA"
                     /db_xref="taxon:11676"
                     /note="strain for reference annotation"
     misc_feature    1..96
                     /note="repeat; positions of RNA transcription
                     initiation and polyadenylation; Region: R"
     polyA_signal    73..78
     5'UTR           97..181
     primer_bind     182..199
     gene            336..1838
                     /gene="gag"
     CDS             336..1838
                     /gene="gag"
                     /note="Pr55"
                     /codon_start=1
                     /product="Gag"
                     /protein_id="AAC82593.1"
                     /db_xref="GI:2801504"
                     /translation="MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERF
                     AVNPGLLETSEGCRQILGQLQPSLQTGSEELRSLYNTVATLYCVHQRIEIKDTKEALD
                     KIEEEQNKSKKKAQQAAADTGHSNQVSQNYPIVQNIQGQMVHQAISPRTLNAWVKVVE
                     EKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPV
                     HAGPIAPGQMREPRGSDIAGTTSTLQEQIGWMTNNPPIPVGEIYKRWIILGLNKIVRM
                     YSPTSILDIRQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTIL
                     KALGPAATLEEMMTACQGVGGPGHKARVLAEAMSQVTNSATIMMQRGNFRNQRKIVKC
                     FNCGKEGHTARNCRAPRKKGCWKCGKEGHQMKDCTERQANFLGKIWPSYKGRPGNFLQ
                     SRPEPTAPPEESFRSGVETTTPPQKQEPIDKELYPLTSLRSLFGNDPSSQ"
     mat_peptide     339..731
                     /gene="gag"
                     /product="matrix"
     mat_peptide     732..1424
                     /gene="gag"
                     /product="capsid"
     mat_peptide     1425..1466
                     /gene="gag"
                     /product="p2"
     mat_peptide     1467..1631
                     /gene="gag"
                     /product="nucleocapsid"
     mat_peptide     1680..1835
                     /gene="gag"
                     /product="p6"
```

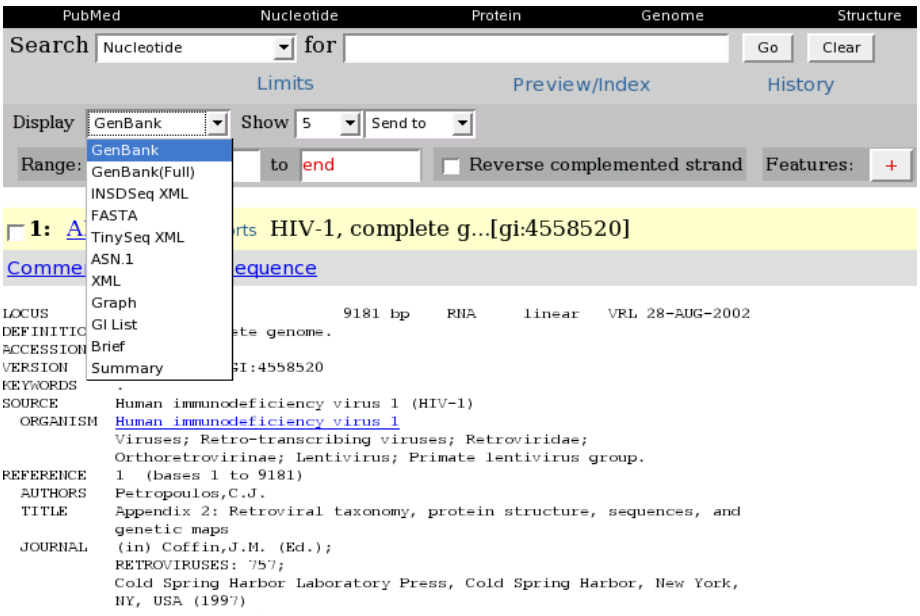
La información que se nos muestra por defecto en la entrada puede ser modificada de acuerdo a nuestras necesidades. Es decir, generalmente requerimos únicamente de la secuencia y no de las anotaciones o la información general, algunas otras veces necesitamos toda la información pero en un formato que sea más legible por un computador que por un ser humano.

Para cualquiera de estos casos es posible cambiar la forma en que se nos muestran los datos de una entrada.

Formatos de salida

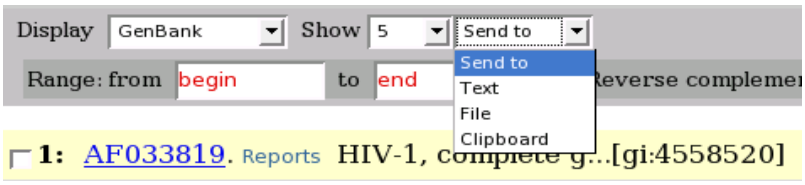
Los formatos de salida que ofrece el NCBI para los resultados pueden ser seleccionados en el menú desplegable ubicado en la región superior izquierda de la página, justo debajo del menú que nos permite realizar las búsquedas. De estos formatos los más relevantes son:

- Genbank
- FASTA
- XML
- ASN.1



Explore cada uno de los formatos del menú desplegable. ¿Qué diferencias y semejanzas encuentra en cada uno de ellos? Preste especial atención al formato FASTA. ¿Por qué razón cree que este es el formato más usado en bioinformática?

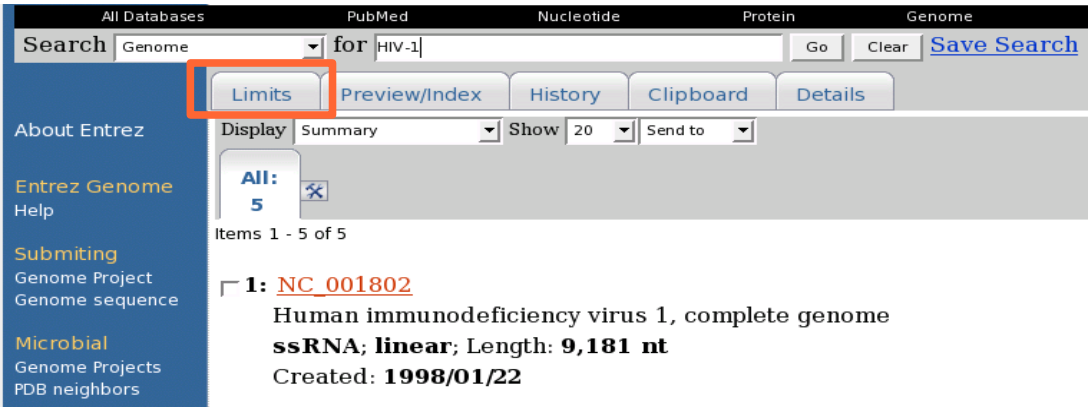
Justo al lado derecho del menú desplegable para los diferentes formatos existe otro menú, el cual nos permite elegir el lugar al que queremos enviar nuestros resultados.



Explore cada una de las opciones allí mostradas. Describa las diferencias que encuentra.

Sequence Retrieval System

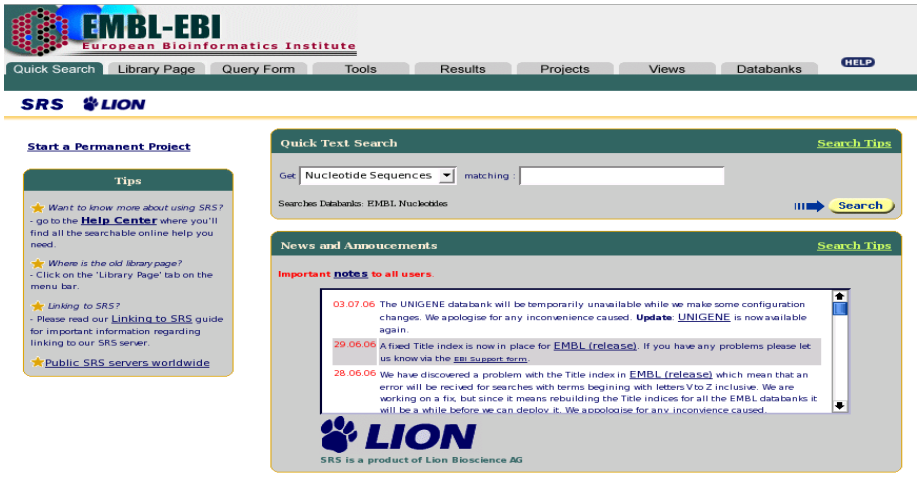
A pesar de todas las virtudes del sistema ENTREZ del NCBI, algunas veces la búsqueda de información allí puede tornarse tediosa y casi imposible, esto se debe principalmente a las pocas posibilidades que este sistema ofrece para filtrar los resultados, la cual se hace por medio de la opción "**limits**", cuyas opciones en realidad son un poco "limitadas".



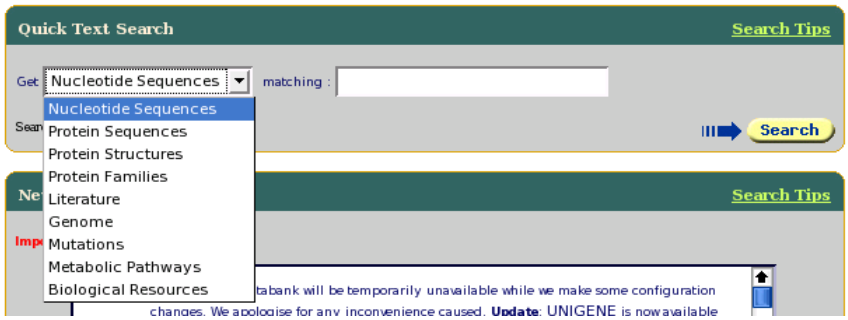
Existe sin embargo una alternativa excelente para la búsqueda de secuencias biológicas, que nos permite controlar casi todos los aspectos de nuestra búsqueda, esta alternativa es el **Sistema de Recuperación de Secuencias** (SRS). Este sistema fue desarrollado teniendo en mente precisamente esta labor de recuperar secuencias biológicas de una manera efectiva, de allí su diseño y sus capacidades.

En este taller trabajaremos con el SRS ofrecido por el **Instituto Europeo de Bioinformática** (EBI), cabe anotar que existen muchos servidores SRS alrededor del mundo que ofrecen sus servicios de manera gratuita (ej., el servidor srs del CBIB: http://srs.ibun.unal.edu.co:8080/srs81/).

Visite la siguiente URL: http://srs.ebi.ac.uk



Una manera sencilla de consultar el SRS es mediante la casilla Quick Text Search. En dicha casilla es posible realizar búsquedas en diversas bases de datos disponibles en el menú desplegable.

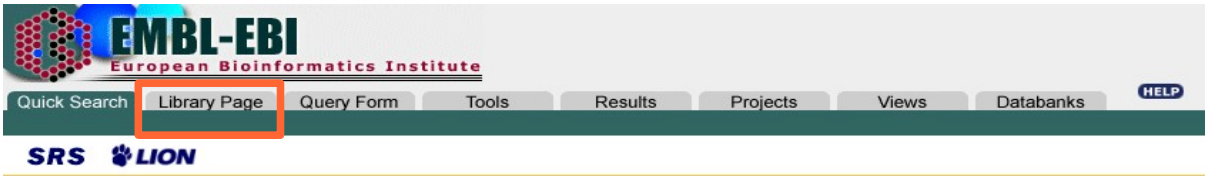


Por ejemplo seleccionando la opción "**Nucleotide Sequences**" realizaremos nuestra búsqueda en la base de datos de DNA EMBL (homóloga al genBank y al DDBJ).

Realice la búsqueda por HIV-1 con diferentes opciones del menú desplegable.

Hasta este momento el SRS parece ser bastante menos completo el sitio web del NCBI, pero ahora empezaremos a comprobar donde radica todo su potencial.

Seleccione la pestaña Library Page ubicada en la parte superior de su pantalla



A continuación será llevado a la sección del SRS donde se describen cada una de las bases de datos que componen el sistema. Como puede ver el SRS comprende muchas bases de datos a la vez y esa es una de sus principales virtudes, por esta razón al SRS se le conoce algunas veces como una "base de datos de bases de datos", pues a través de este sistema podemos consultar múltiples bases de datos al mismo tiempo, de acuerdo a nuestras necesidades particulares.

Como puede darse cuenta el SRS es similar al sistema ENTREZ del NCBI, en el sentido en que nos permite consultar muchas bases de datos al mismo tiempo, pero esta vez no restringidos únicamente a aquellas con las que cuenta el NCBI sino a virtualmente cualquier base de datos. El número de bases de datos con las que cuenta el SRS depende de cada implementación, es decir el administrador del SRS determina qué bases de datos quiere o no incluir en su sistema.

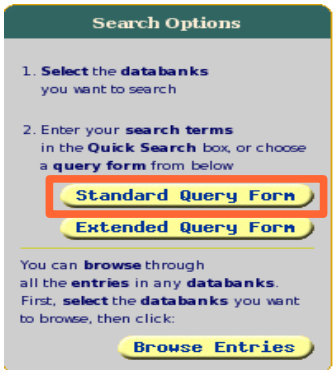
Póse el cursor del mouse por alguna de las entradas, depués de unos segundos una casilla de texto explicativo aparecerá. ¿Qué tipo de información proveen las bases de datos EMBL (Contig Updates), UniprotKB/Swissprot?

Al seguir el enlace a cualquiera de estas bases de datos obtendremos mayor información acerca de esta, como el número de entradas presentes, fecha de actualización etc. Sin embargo, por ahora nuestro interés es el de seleccionar algunas bases de datos para realizar nuestras búsquedas.

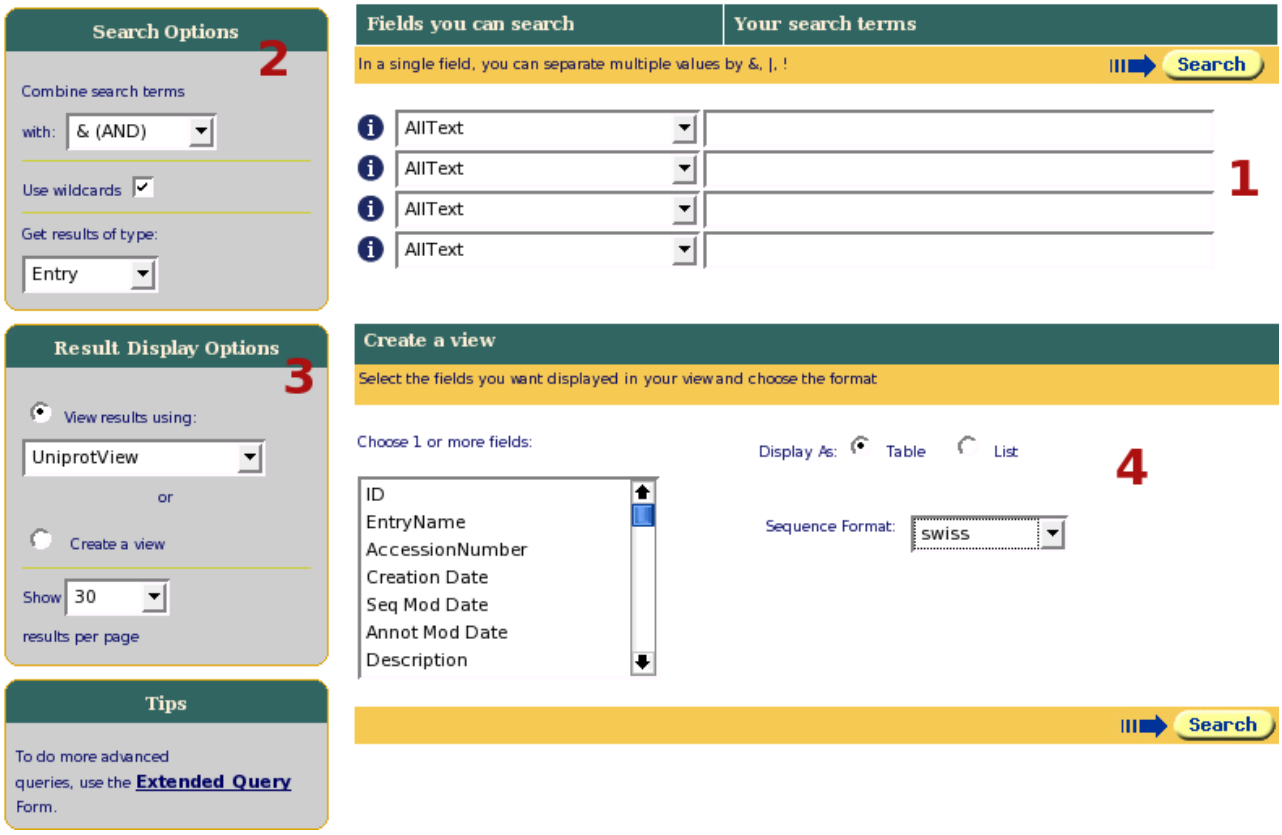
Seleccione las casillas pertenecientes a las bases de datos de "**UniprotKB/Swissprot**" y "**UniprotKB/TrEMBL**". Cerciorese de que estas sean las únicas bases de datos seleccionadas.

A la izquierda de su pantalla encontrará la casilla "Search Options" la cual nos permitirá seleccionar el nivel de profundidad de nuestra buśqueda, Por ser esta la primera vez que trabajamos con este sistema seleccionaremos la forma estándar de búsqueda.

Presione el botón "**Standard query Form**" de la casilla "**Search Options**".



Esta acción le llevará al formulario estándar de búsqueda en el SRS.

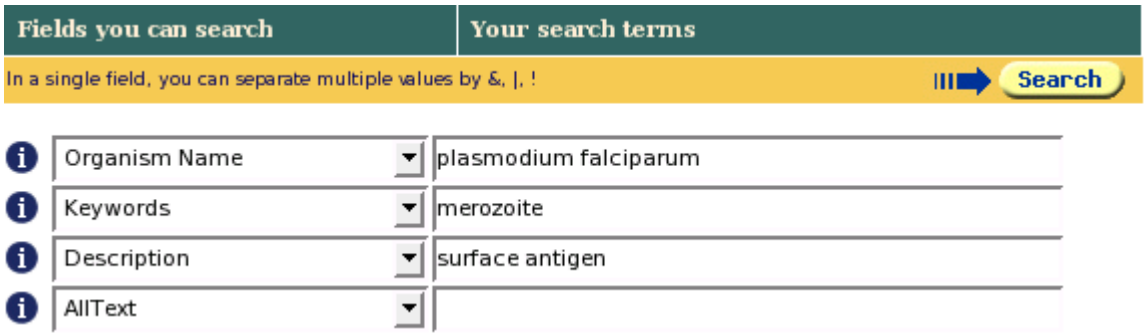


El cual consta de 4 partes fundamentales.

1. **Campos de búsqueda**, donde podemos entrar nustros términos de búsqueda de acuerdo a cualquiera de las opciones presentes en los respectivos menúes desplegables.
2. **Opciones de búsqueda**, donde podemos definir, entre otras cosas, el tipo de conector lógico (booleano) a utilizar para los términos definidos en **1**.
3. **Opciones para mostrar los resultados**, donde podemos definir el número de resultados que queremos por página, así como el formato de salida, ya sea alguno de los definidos en el menú desplegable o mediante la creación de una vista personalizada (opción "**create view**").
4. **Crear vista**, esta opción trabaja en conjunto con la opción **3**, y acá podemos definir el tipo de campos que queremos ver en nuestra página de resultados.

Para nuestro ejemplo, tenemos interés en seleccionar todas las proteínas de superficie conocidas de *Plasmodium falciparum* con actividad inmunogénica, relacionadas con el merozoito.

Defina estos criterios en la sección "**campos de búsqueda**" de acuerdo a la siguiente imágen:



A continuación presione el botón “**search**” ubicado en la parte superior de esta sección y espere unos segundos.

Seguramente en este momento ya tenga una visión más exacta de las posibilidades que ofrece el SRS y sus principales diferencias con el sistema ENTREZ. Primero, pudimos definir exactamente no solamente la base de datos que queríamos consultar, sino las secciones específicas de esta. Además de esto pudimos también definir exactamente los términos de búsqueda en secciones específicas de las entradas, lo cual nos da un completo control sobre los resultados que queremos obtener.

Cree usted que existe alguna manera de realizar esta misma consulta en el ¿sistema ENTREZ?

Juegue un poco con las diferentes opciones de formatos que ofrece el SRS en la sección **3**, del formulario de búsqueda. Intente también creando su propio formato de salida con la opción **create view** y la sección **4**.

Encuentre todas las proteínas nucleares hipotéticas de SACCHAROMYCES CEREVISIAE, y muestre la información en formato fasta.

BLAST: Basic local Aligment Search Tool

BLAST es un **algoritmo** para comparación (alineamiento) de secuencias. Más exactamente se encuentra clasificado dentro de los algoritmos para alineamiento local.

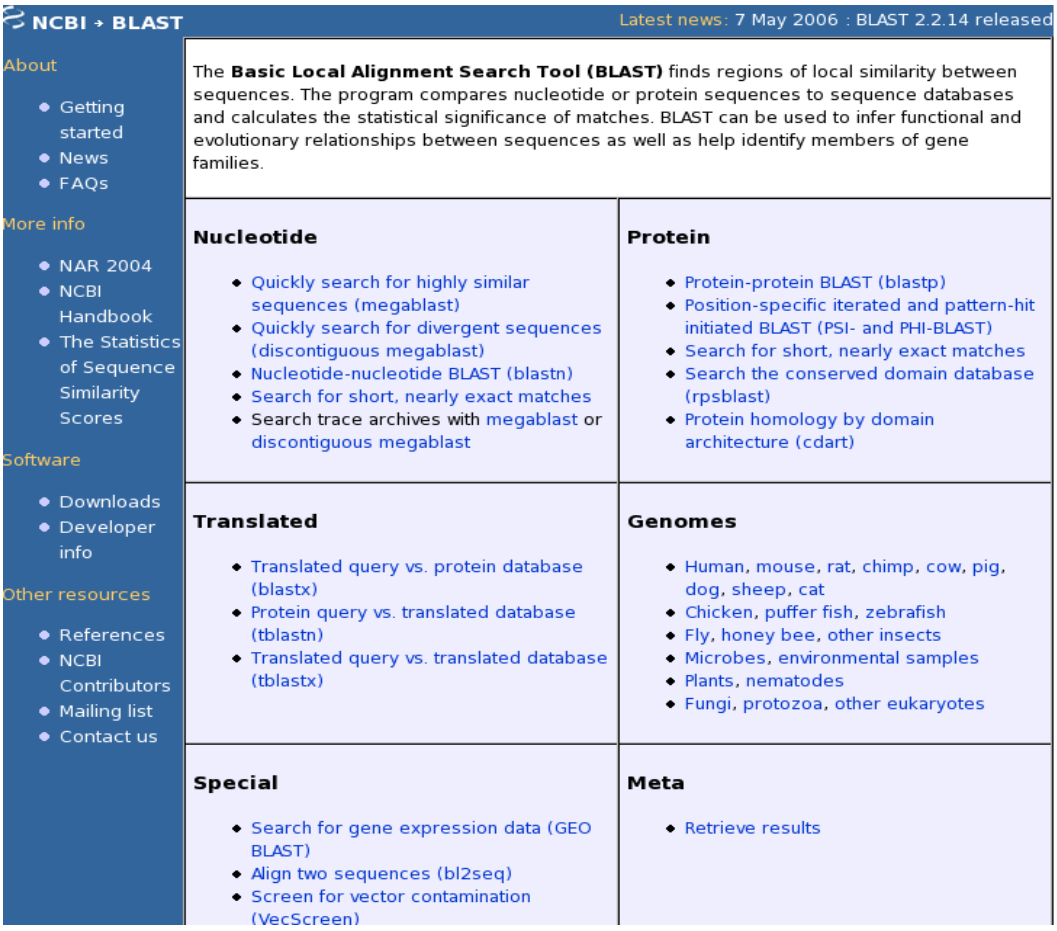
Existen varias “implementaciones” de este algoritmo, una de las más conocidas es la realizada por el NCBI, el **NCBI-BLAST**.

Otra implementación muy conocida de este algoritmo es la realizada por la Universidad de Washington el **WU-BLAST** (http://blast.wustl.edu/).

Es importante notar que el NCBI-BLAST **no** se utiliza cuando estamos realizando búsquedas de secuencias por palabras clave o términos de búsqueda. Este se utiliza cuando estamos buscando secuencias similares a la nuestra en las diferentes bases de datos del NCBI. Entonces, existe una clara diferencia entre consultar el NCBI para **obtener una o varias secuencias** requeridas y consultarlo para **buscar coincidencias** de nuestra secuencia con otras.

El NCBI-BLAST es el programa/algoritmo que usa por defecto el NCBI para realizar búsquedas de secuencias en sus bases de datos. En esta sección trabajaremos con dicha implementación.

Visite el sitio web del NCBI-BLAST: http://www.ncbi.nlm.nih.gov/BLAST/



realizar una búsqueda BLAST puede resumirse en 3 sencillos pasos:

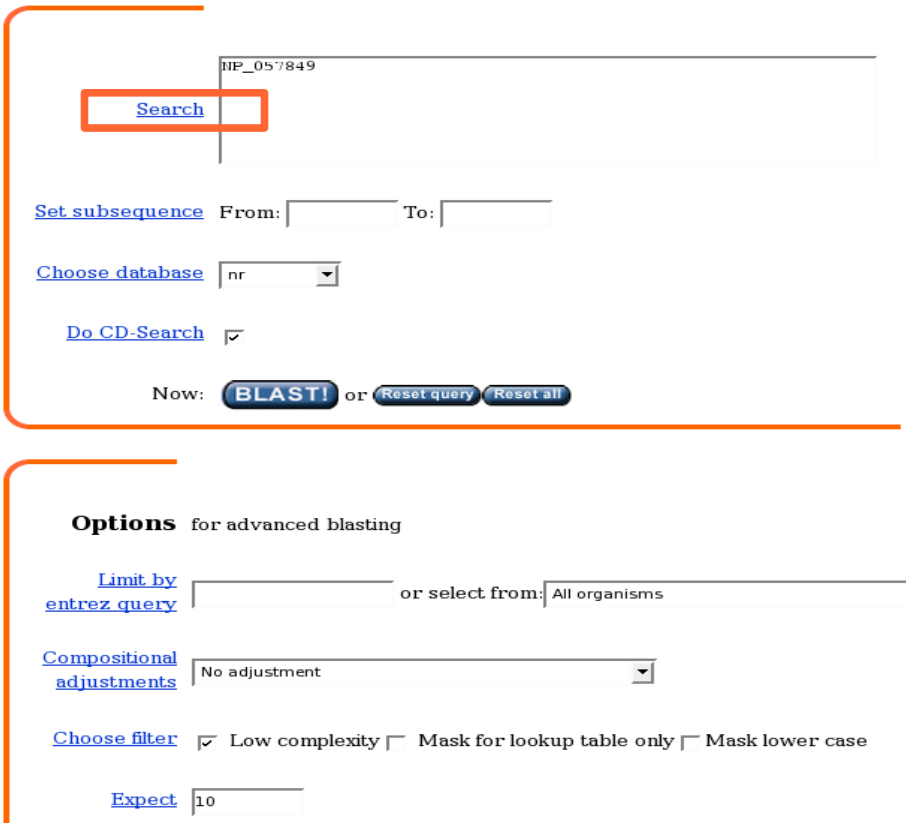
1. Seleccionar el tipo de programa BLAST a usar (blastp, blastn, blastx, tblastx,tblastn).
2. Introducir nuestra secuencia pregunta (**query sequence**, en términos BLAST).
3. Seleccionar la base de datos en la que queremos buscar.

Opcionalmente podemos controlar la salida de los resultados, modificando algunas de las opciones de salida.

Selección del programa BLAST

Revise cuidadosamente las tres primeras secciones de los programas BLAST (Nucleotide, Protein y Translated). Describa la funcionalidad de cada uno de ellos, de acuerdo a las descripciones de los mismos[1].

Siga el enlace a "**blastp**" y digite el identificador: **NP_057849** en la casilla de texto **Search**.



1 El siguiente enlace puede ayudarle a comprender más dichas descripciones: http://www.ncbi.nlm.nih.gov/blast/producttable.shtml#pstab

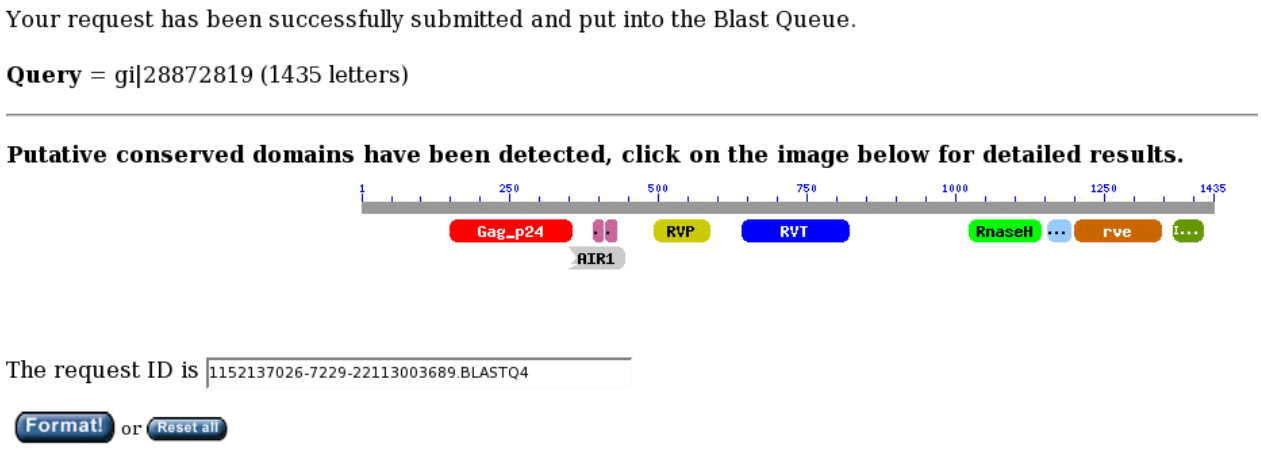
Existen diversas formas de ingresar nuestra secuencia para realizar búsquedas BLAST. Una de ellas es digitando un identificador conocido por el NCBI. Otra manera de hacerlo es ingresando directamente la secuencia en esta misma casilla, ya sea “cruda” o en formato FASTA.

El menú desplegable “**choose database**”, permite seleccionar alguna de las bases de datos permitidas para nuestro tipo de búsqueda. Seleccione **Swissprot**.

La opción “**Do-CD-Search**” que viene por defecto seleccionada le dice a BLAST que también realice una búsqueda de **Dominios Comunes** para dicha proteína. Utilizaremos las opciones (**Options**) por defecto.

Presione el Botón **Blast** y espere unos segundos.

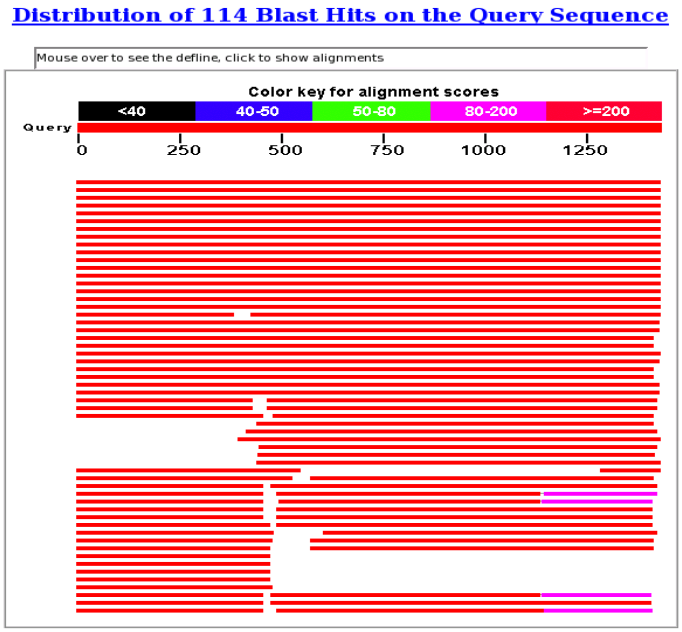
Esta acción le llevará a una página intermedia entre la página de resultados y el formulario de consulta, es una página de “formateo” de resultados, en la cual además es posible ver los dominios comúnes encontrados para nuestra secuencia. No nos preocuparemos mucho por el formato de salida, y dejaremos las opciones por defecto, así que...



...presione el botón **Format** y espere unos segundos (algunas veces es necesario esperar un poco más).

En este momento nuestra secuencia esta siendo comparada contra cada una de las entradas en las bases de datos que escogimos (3,782,570 secuencias). En unos momentos aparecerá una página de resultados con las secuencias que BLAST ha encontrado son muy similares (o idénticas) a nuestra secuencia.

La primera parte de la página de resultados muestra la siguiente gráfica:



Esta nos permite ver la distribución de los alineamientos. Cada banda debajo del mapa representa una secuencia de la base de datos que resultó ser muy similar a la secuencia de búsqueda. De esta manera podemos ver la extensión de los alineamientos.

Resulta evidente que para la secuencia que hemos elegido BLAST encontró varias secuencias idénticas a la nuestra, por esta razón predomina el color rojo (notar el código de colores en el mapa).

La segunda sección corresponde a las descripciones de los alineamientos.

gi|120836|sp|P18800|GAG_HV1ND	Gag polyprotein (Pr55Gag) [Cont...	786	0.0
gi|120829|sp|P12495|GAG_HV1Z2	Gag polyprotein (Pr55Gag) [Cont...	783	0.0
gi|120838|sp|P04594|GAG_HV1MA	Gag polyprotein (Pr55Gag) [Cont...	764	0.0
gi|120843|sp|P24736|GAG_HV1U4	Gag polyprotein (Pr55Gag) [Cont...	749	0.0
gi|120844|sp|P05889|GAG_HV1W2	Gag polyprotein (Pr55Gag) [Cont...	727	0.0
gi|120885|sp|P17282|GAG_SIVCZ	Gag polyprotein [Contains: Core...	689	0.0
gi|120890|sp|P19504|GAG_SIVSP	Gag polyprotein [Contains: Core pr	535	5e-151
gi|120853|sp|P04590|GAG_HV2RO	Gag polyprotein (Pr55Gag) [Cont...	529	3e-149
gi|120854|sp|P20874|GAG_HV2ST	Gag polyprotein (Pr55Gag) [Cont...	527	1e-148
gi|120889|sp|P12496|GAG_SIVS4	Gag polyprotein [Contains: Core pr	526	2e-148
gi|120848|sp|P24106|GAG_HV2CA	Gag polyprotein (Pr55Gag) [Cont...	524	8e-148
gi|120850|sp|P18041|GAG_HV2G1	Gag polyprotein (Pr55Gag) [Cont...	521	5e-147
gi|2495241|sp|Q74119|GAG_HV2KR	Gag polyprotein (Pr55Gag) [Con...	521	7e-147
gi|120849|sp|P15832|GAG_HV2D2	Gag polyprotein (Pr55Gag) [Cont...	518	4e-146
gi|120852|sp|P05891|GAG_HV2NZ	Gag polyprotein (Pr55Gag) [Cont...	517	1e-145
gi|120847|sp|P18095|GAG_HV2BE	Gag polyprotein (Pr55Gag) [Cont...	516	2e-145
gi|399525|sp|P31634|GAG_SIVMS	Gag polyprotein [Contains: Core pr	516	2e-145
gi|120846|sp|P17756|GAG_HV2D1	Gag polyprotein (Pr55Gag) [Cont...	516	2e-145
gi|120851|sp|P12450|GAG_HV2SB	Gag polyprotein (Pr55Gag) [Cont...	514	8e-145
gi|120887|sp|P05894|GAG_SIVM1	Gag polyprotein [Contains: Core pr	510	2e-143
gi|130597|sp|P16088|POL_FIVPE	Pol polyprotein [Contains: Prot...	507	1e-142
gi|120884|sp|P05892|GAG_SIVVT	Gag polyprotein [Contains: Core...	507	1e-142

Esta es una lista de las secuencias encontradas (ordenadas de a cuerdo a su valor E), en cuatro columnas: identificador, breve descripción de la secuencia, bit score y valor E.

La tercera sección corresponde a los alineamientos hechos por BLAST de nuestra secuencia pregunta (**query**) y la secuencia encontrada (**subject**).

Estos alineamientos son mostrados en un formato convencional de alineamiento pareado, mostrando a demás el porcentaje de identidad, el valor E del alineamiento, el número de gaps y el Score.

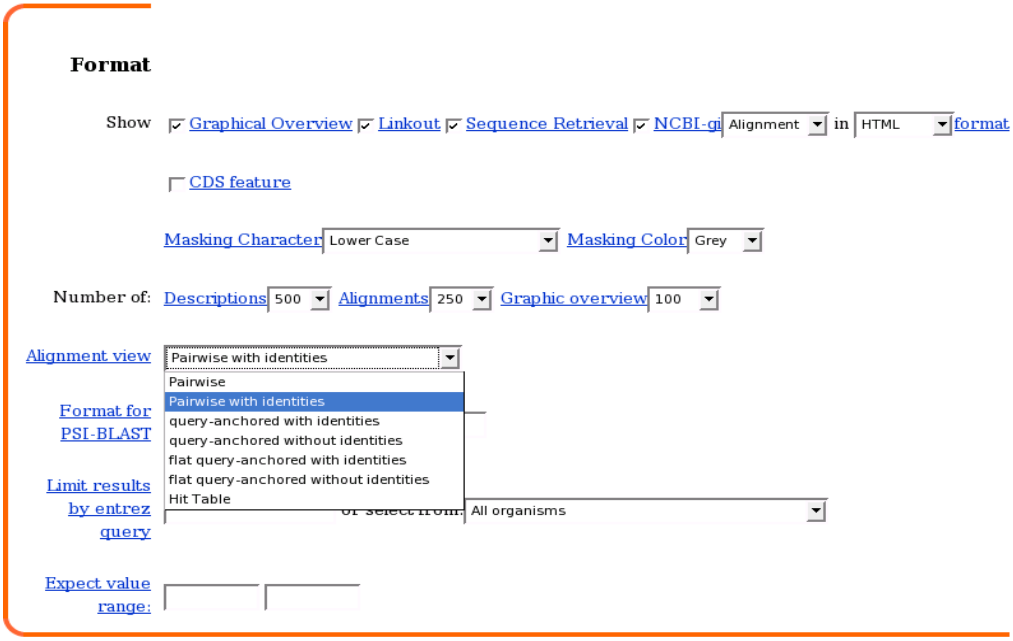
```
>gi|400822|sp|P31822|POL_FIVT2  Pol polyprotein [Contains: Protease (Retropeps
H (RT); Deoxyuridine 5'-triphosphate
nucleotidohydrolase (dUTPase); Integrase (IN)]
Length=1124

 Score =  494 bits (1273),  Expect = 7e-139
 Identities = 286/673 (42%), Positives = 404/673 (60%), Gaps = 39/673 (5%)

Query  492   TLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLP-----GRWKPKMIGGIGGFIKVRQ  546
             TL +RP + I + G    + LLDTGAD T+L               K  MIG +GG  +
Sbjct  46    TLERRPEIQIFVNGHPIKFLLDTGADITILNRKDFQIGNSIENGKQNMIG-VGGGKRGTN  104

Query  547   YDQILIEICGHKAIGTVLVGPTPV---------NIIGRNLLTQIGCTLNFP--ISPIETVP  596
             Y  + +EI          + G   V          ++GR+ + +    L         I  V
Sbjct  105   YINVHLEIRDENYRMQCIFGNVCVLEDNSLIQPLLGRDNMIKFNIRLVMAQISEKIPIVK  164

Query  597   VKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDST  656
             V++K    GP+VKQWPL+ EKI+AL +I   +E EGK+ +  P NP+NTPVFAIKKK S
Sbjct  165   VRMKDPTQGPQVKQWPLSNEKIEALTDIVERLESEGKVKRADPNNPWNTPVFAIKKK-SG  223

Query  657   KWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPLDEDFRKYTA  716
             KWR L+DFR LNK T    EVQLG+PHPAGL+ KK VTVLD+GDAYF++PLD D+  YTA
Sbjct  224   KWRMLIDFRVLNKLTDKGAEVQLGLPHPAGLQMKKQVTVLDIGDAYFTIPLDPDYAPYTA  283

Query  717   FTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQYMDDLYV  776
             FT+P  NN  PG RY +  LPQGW  SP I+QS++  IL+PF KQN ++ IYQYMDD+Y+
Sbjct  284   FTLPRKNNAGPGRRYVWCSLPQGWVLSPLIYQSTLNNILQPFIKQNSELDIYQYMDDIYI  343

Query  777   GSDLEIGQHRTKIEELRQHLLRWGLTTPDKKHQKEPPFLWMGYELHPDKWTVQPIVLPEK  836
             GS+L   +H+ K+EELR+ LL WG  TP+ K Q+EPP+ WMGYELHP  W++Q   L
Sbjct  344   GSNLNKKEHKQKVEELRKLLLAWGFETPEDKLQEEPPYKWMGYELHPLTWSIQQKQLEIP  403

Query  837   DSWTVNDIQKLVGKLNWASQIYPGIKVRQLCKLLRGTKALTEVIPLTeeaelelaeNREI  896
             +  T+N++QKL GK+NWASQ  P + +++L  ++RG + L  +   T EA+ E+ + +E
Sbjct  404   ERPTLNELQKLAGKINWASQTIPDLSIKELTNMMRGDQKLDSIREWTVEAKREVQKAKEA  463

Query  897   LKEPVHGVYYDPSKDLIAEIQKQGQGQWTYQIYQE-PFKNLKTGKYARMRGAHTNDVKQL  955
             ++       YYDP++ L A++   G  Q  YQ+YQ+ P    L  GK  R +    N
Sbjct  464   IETQAQLNYYDPNRGLYAKLSLVGPHQICYQVYQKNPEHILWYGKINRQKKKAENTCDIA  523

Query  956   TEAVQKITTESIVIWGKTPKFKLPIQKetwetwwteYWQATWI---------PEWEFVNT  1006
               A  KI  ESI+  GK P +++P  +E          W++  I          PE EF++
Sbjct  524   LRACYKIREESIIRIGKEPVYEIPASREA--------WESNLIRSPYLKAPPPEVEFIHA  575
```

Al seguir cada uno de los links de las secuencias alineadas seremos llevados a la página de resultados para dicha secuencia.

Seleccione las casillas de selección de las cuatro primeras secuencias mostradas y a continuación presione el botón “Get selected sequences” que se encuentre al inicio de los alineamientos. ¿Qué obtiene al realizar esta acción?

Explore las otras posibilidades allí ofrecidas. Especialmente la opción “**tree view**”. ¿Qué utilidad cree usted que puede tener esta opción?

Realice nuevamente esta búsqueda, pero esta vez modifique el formato en la página intermedia, seleccionando la opción "**pairwise with identities**" en el menú desplegable en "**aligment view**". ¿Qué diferencias y similitudes encuentra en el formato de salida con respecto al parámetro por defecto: **Pairwise**?



Bogotá Colombia - Julio de 2006.

Cualquier sugerencia o inquietud dirigirla a:

ampinzonv@unal.edu.co ó andrespinzon@gmail.com