

# Fuentes de datos en Bioinformática

## Consultas al NCBI-ENTREZ

### Introducción

Con el desarrollo de esta guía podrá familiarizarse con algunas de las fuentes de datos en el NCBI, así como con el sistema ENTREZ y desarrollará ciertas habilidades para realizar búsquedas más efectivas en este sistema.

### Consultando el sistema ENTREZ

Visite el sitio web del sistema ENTREZ:

<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>

Después de unos segundos se encontrará con la página principal de ENTREZ:

The screenshot shows the NCBI Entrez search engine homepage. At the top, there is the NCBI logo and the text "Entrez, The Life Sciences Search Engine". Below this is a navigation bar with links for "HOME", "SEARCH", "SITE MAP", "PubMed", "All Databases", "Human Genome", "GenBank", "Map Viewer", and "BLAST". A search bar is located below the navigation bar, with the text "Search across databases" and a "GO" button. The main content area is titled "Welcome to the Entrez cross-database search page" and contains a grid of database icons and descriptions. The databases listed include PubMed, PubMed Central, Nucleotide (GenBank), Protein, Genome, Structure, Taxonomy, SNP, Gene, HomoloGene, PubChem Compound, PubChem Substance, Genome Project, Books, OMIM, Site Search, UniGene, CDD, 3D Domains, UniSTS, PopSet, GEO Profiles, GEO DataSets, Cancer Chromosomes, PubChem BioAssay, GENSAT, Journals, and NLM Catalog. Each database entry includes a small icon, a brief description, and a question mark icon for help.

Consultar este sistema es tan fácil como digitar nuestro término de búsqueda en el campo de texto (1) y oprimir el botón “go” (2). En esta oportunidad aprenderemos a realizar búsquedas más complejas mediante la utilización de operadores booleanos<sup>1</sup>.

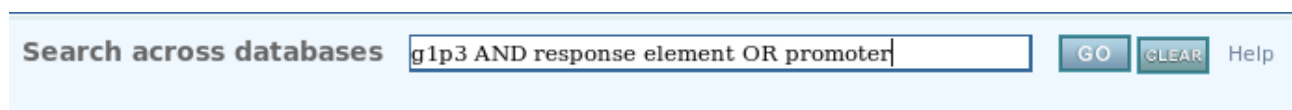
¿Qué es un operador booleano?

El sistema ENTREZ (como la mayoría de sistemas de búsqueda) cuenta con una serie de operadores booleanos que nos permiten hacer consultas más complejas. Estos operadores son: **AND**, **OR** y **NOT**.

Realice la siguiente búsqueda en el sistema ENTREZ:

**g1p3 AND response element OR promoter**

¿Cuántas entradas encontró para la sección Nucleotide?



Search across databases    [Help](#)

Ahora realice una nueva búsqueda:

**g1p3 AND (response element OR promoter)**

¿Cuántas entradas encontró para la sección Nucleotide?

Como puede observar, el uso de paréntesis en la búsqueda tiene un fuerte efecto en el número de resultados arrojados.

De acuerdo a las dos búsquedas hechas anteriormente y a los resultados que de estas obtuvo. ¿De qué manera cree usted que el sistema ENTREZ realizó dichas búsquedas, para cada uno de los dos casos?

Realice cada una de las siguientes búsquedas y describa la manera en que operan:

johnson a[AUTHOR]

johnson a[AUTHOR] AND 1998/02:2000/01/25[MDAT]

---

<sup>1</sup> Estos mismos operadores pueden ser usados no solamente en el sistema ENTREZ sino en cualquiera de las secciones del NCBI.

## La opción History

El NCBI mantiene un registro de las diferentes búsquedas que realizamos en una sesión, estas son guardadas temporalmente en una sección llamada “History”. A continuación aprenderemos a utilizar esta opción para realizar búsquedas combinadas.

Visite la página principal del NCBI: <http://www.ncbi.nlm.nih.gov/>

Siga el enlace “**Alphabetical list**” y busque allí el enlace “**Protein Sequences (entrez)**”, presione este último.

Después de unos segundos se encontrará en la página principal de la sección de proteínas.

The screenshot shows the NCBI Entrez Protein search interface. The search bar contains the word "Protein" and a dropdown menu. Below the search bar are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". A yellow box highlights the text: "The protein entries in the Entrez search and retrieval system have been compiled from a variety of sources, including SwissProt, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq." Below this is a "Human Genome" section with a link to "human genome resources". At the bottom, there are two columns: "Additional protein information" and "Retrieve taxonomy information".

Busque por el término: **streptomyces**, presione el botón “go”. Una vez la página halla terminado de cargar presione el botón “clear”.

Realice las mismas acciones con los términos : **pseudomonas** y **glucanase**.

Estas 3 búsquedas que acaba de realizar están ahora guardadas en la sección “history”.

Presione la pestaña “history”.

Después de unos segundos se encontrará con una página similar a la siguiente:

Search	Most Recent Queries	Time	Result
<a href="#">#35</a> Search <b>glucanase</b>		14:16:10	<a href="#">3891</a>
<a href="#">#34</a> Search <b>pseudomonas</b>		14:15:56	<a href="#">209793</a>
<a href="#">#33</a> Search <b>streptomyces</b>		14:15:34	<a href="#">61750</a>

Describe las diferentes columnas de los resultados que observa.

Ahora aprenderemos a combinar estas búsquedas en una sola, usando operadores booleanos.

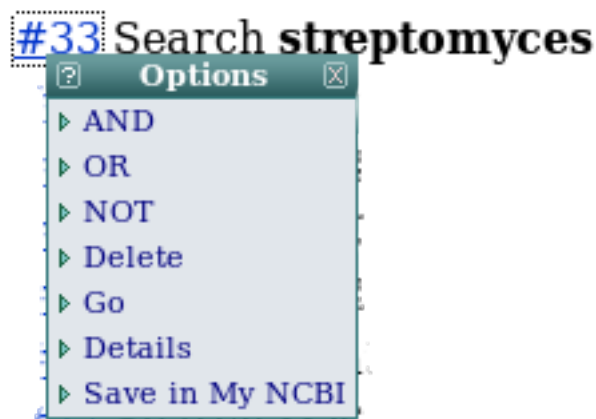
Es posible que la casilla de búsqueda se encuentre aún con el último término digitado, como en este caso, con la palabra “glucanase”.

The screenshot shows the NCBI search interface. At the top, there are tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. Below these is a search bar with a dropdown menu set to 'Protein' and the text 'glucanase' entered. To the right of the search bar are buttons for 'Preview', 'Go', and 'Clear'. Below the search bar is a table with the same structure as the one above, showing the most recent search: '#35 Search glucanase' with a time of '14:16:10' and a result of '3891'.

Asegúrese de que la casilla se encuentre limpia, presionando el botón “clear”. Como verá este paso es de suma importancia para alcanzar el objetivo que perseguimos.

Presione el número en la columna **search** a la izquierda del término **streptomyces** (en este caso es #33, pero puede ser diferente en su caso específico).

Encontrará una serie de opciones en un menú desplegable, como se muestra a continuación:

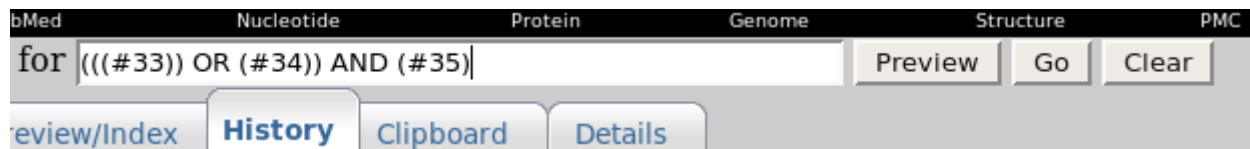


Presione el operador **OR**.

Haga exactamente lo mismo con la entrada para el término **pseudomonas**.

Haga lo mismo para el término **glucanase**, pero esta vez use el operador **AND**.

En la casilla de búsqueda debe encontrarse ahora con lo siguiente (o muy similar, de acuerdo al número de búsqueda en su caso específico):



Presione el botón “**go**”. ¿Cuántos resultados ha encontrado?

Realice esta misma búsqueda en el sistema ENTREZ. ¿Cuántos resultados aparecen esta vez en la sección **protein**?

¿Qué expresión booleana utilizó para realizar esta búsqueda?

El NCBI cuenta con otra serie de opciones interesantes para filtrar nuestras búsquedas. Una de las más interesantes es la opción “LIMITS”. Explore esta opción, ¿Que ventajas trae la utilización de esta opción?

Ahora que sabe de que manera realizar búsquedas más complejas en este sistema, tal vez le resulte más sencillo responder la siguiente pregunta:

En un análisis de secuencias promotoras en *Xanthomonas*, se necesita conocer cuántas publicaciones existen para *Xanthomonas Campestris* hasta la fecha. También es necesario conocer cuántas de estas publicaciones fueron hechas entre febrero del 2006 y el primero de enero del 2007.

Escriba las consultas booleanas que ha realizado para llevar a cabo estas búsquedas.

Guía elaborada por Andrés M. Pinzón V., del **Centro de Bioinformática** del Instituto de Biotecnología en la Universidad Nacional de Colombia y del **Laboratorio de Micología y Fitopatología** de la Universidad de los Andes, y está distribuida bajo licencia:



[Creative Commons](https://creativecommons.org/licenses/by-nc-sa/4.0/)

Bogotá Colombia – Febrero de 2007.

Cualquier sugerencia o inquietud dirigirla a:

[ampinzonv@unal.edu.co](mailto:ampinzonv@unal.edu.co) ó [andrespinzon@gmail.com](mailto:andrespinzon@gmail.com)