

**MICB 405 Bioinformatics**  
**Lecture 2.2 – PART A**  
**Retrieving Biological Information with**  
**Entrez**

**FSC 1221**  
**September 16<sup>th</sup>, 2008**

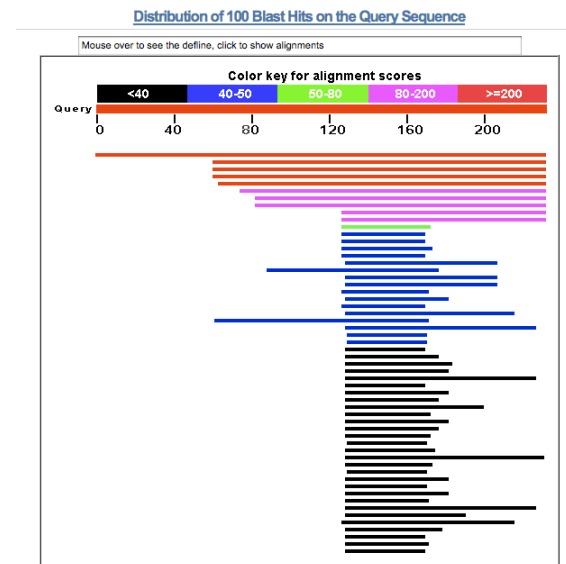
# Retrieving Biological Information

Entrez

BLAST



Search by name,  
identifier, feature



Search by  
sequence similarity

# Objectives

- By the end of the first part of today's lecture:
  - You will be able to describe the Entrez database retrieval system.
  - You will recognize links between different the different Entrez databases.
  - You will be able to describe how “neighboring” works in Entrez for three different databases.
  - You will be able to describe some advanced techniques for searching PubMed.

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

**NCBI**  
National Center for Biotechnology Information  
National Library of Medicine      National Institutes of Health

PubMed   **All Databases**   BLAST   OMIM   Books   TaxBrowser   Structure

Search All Databases ▼ for

**SITE MAP**  
Alphabetical List  
Resource Guide

**About NCBI**  
An introduction to NCBI

**GenBank**  
Sequence submission support and software

**Literature databases**

▶ **What does NCBI do?**

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

**Hot Spots**

- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home

**100 Gigabases**  
GenBank and its collaborating databases, the European Molecular Biology Laboratory and

# What is Entrez?

- integrated, text-based search and retrieval system used at NCBI for the major databases
  - includes PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others...

Search across databases

GO

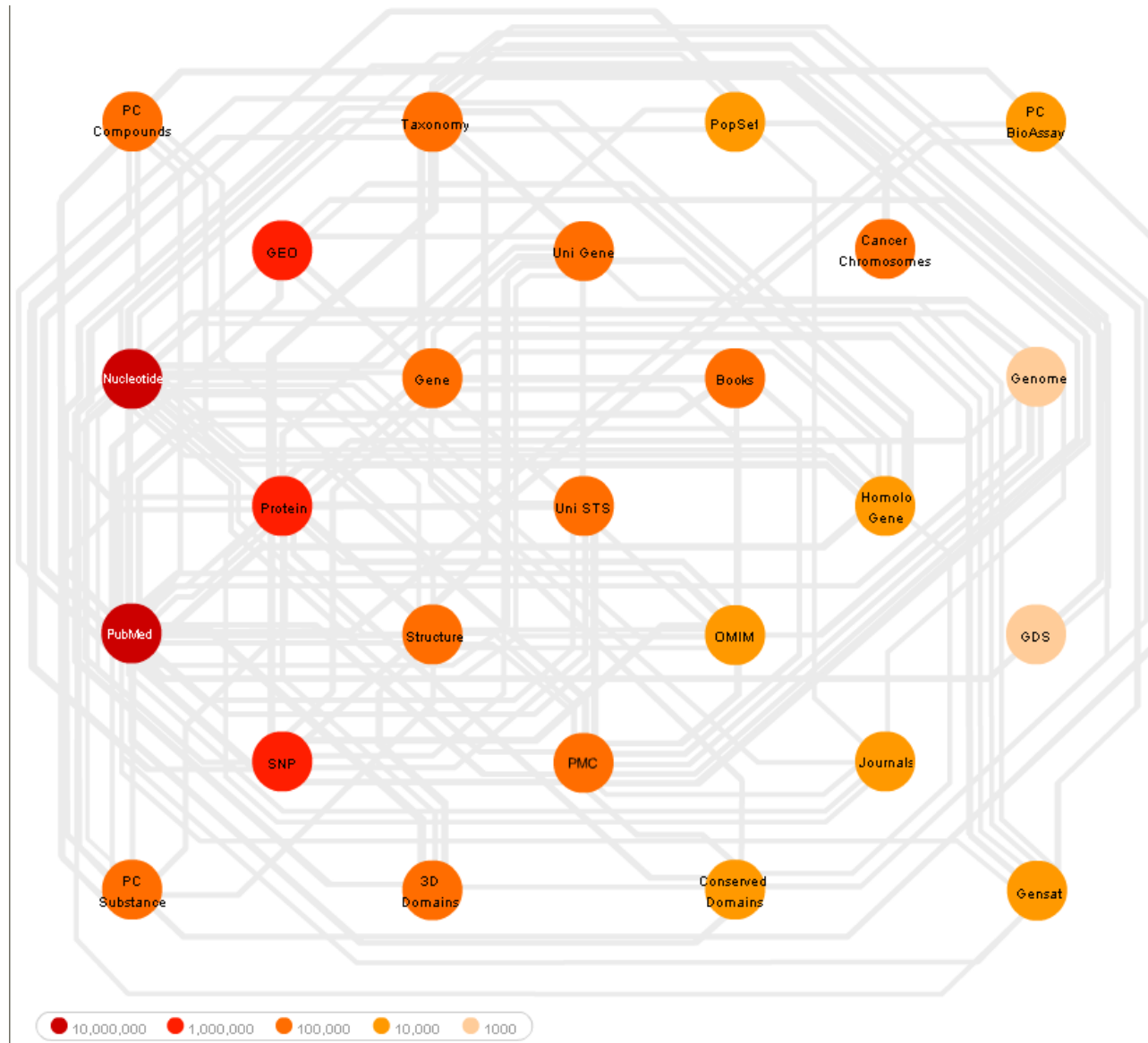
CLEAR

Help

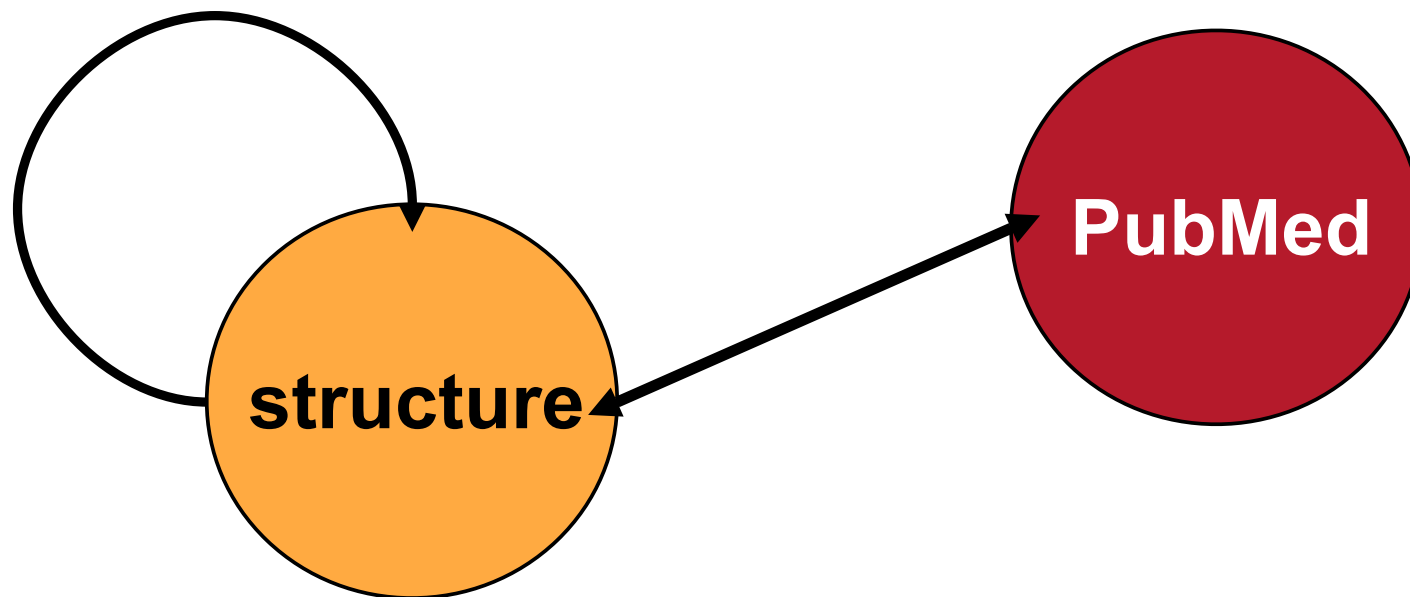
Welcome to the Entrez cross-database search page

- |   |  |
|---|--|
|  <b>PubMed:</b> biomedical literature citations and abstracts <a href="#">?</a>  |  <b>Books:</b> online books <a href="#">?</a>   |
|  <b>PubMed Central:</b> free, full text journal articles <a href="#">?</a>   |  <b>OMIM:</b> online Mendelian Inheritance in Man <a href="#">?</a>                       |
|  <b>Site Search:</b> NCBI web and FTP sites <a href="#">?</a>  |  <b>OMIA:</b> online Mendelian Inheritance in Animals <a href="#">?</a>                   |
|  <b>Nucleotide:</b> sequence database (GenBank) <a href="#">?</a>  |  <b>UniGene:</b> gene-oriented clusters of transcript sequences <a href="#">?</a>         |
|  <b>Protein:</b> sequence database <a href="#">?</a>   |  <b>CDD:</b> conserved protein domain database <a href="#">?</a>                          |
|  <b>Genome:</b> whole genome sequences <a href="#">?</a>   |  <b>3D Domains:</b> domains from Entrez Structure <a href="#">?</a>                       |
|  <b>Structure:</b> three-dimensional macromolecular structures <a href="#">?</a>   |  <b>UniSTS:</b> markers and mapping data <a href="#">?</a>                                |
|  <b>Taxonomy:</b> organisms in GenBank <a href="#">?</a>   |  <b>PopSet:</b> population study data sets <a href="#">?</a>                              |
|  <b>SNP:</b> single nucleotide polymorphism <a href="#">?</a>  |  <b>GEO Profiles:</b> expression and molecular abundance profiles <a href="#">?</a>       |
|  <b>Gene:</b> gene-centered information <a href="#">?</a>  |  <b>GEO DataSets:</b> experimental sets of GEO data <a href="#">?</a>                   |
|  <b>HomoloGene:</b> eukaryotic homology groups <a href="#">?</a>   |  <b>Cancer Chromosomes:</b> cytogenetic databases <a href="#">?</a>                     |
|  <b>PubChem Compound:</b> unique small molecule chemical structures <a href="#">?</a>                                    |  <b>PubChem BioAssay:</b> bioactivity screens of chemical substances <a href="#">?</a>  |
|  <b>PubChem Substance:</b> deposited chemical substance records <a href="#">?</a>  |  <b>GENSAT:</b> gene expression atlas of mouse central nervous system <a href="#">?</a> |
|  <b>Genome Project:</b> genome project information <a href="#">?</a>   |  |
|  <b>Journals:</b> detailed information about the journals indexed in PubMed and other Entrez databases <a href="#">?</a> |  <b>MeSH:</b> detailed information about NLM's controlled vocabulary <a href="#">?</a>  |
|  <b>NLM Catalog:</b> catalog of books, journals, and audiovisuals in the NLM collections <a href="#">?</a>               |  |

<http://www.ncbi.nih.gov/Database/datamodel>



# Entrez – Linking Data





# Entrez – Linking Databases

- Hard Links

- Direct connections between entries in two different databases

- Examples

- Link to paper describing a nucleotide sequence
- Link to taxonomy database for a protein sequence
- Link from nucleotide sequence to protein CDS
- Link from protein sequence to 3D structure entry

- Not all possible hard links are present

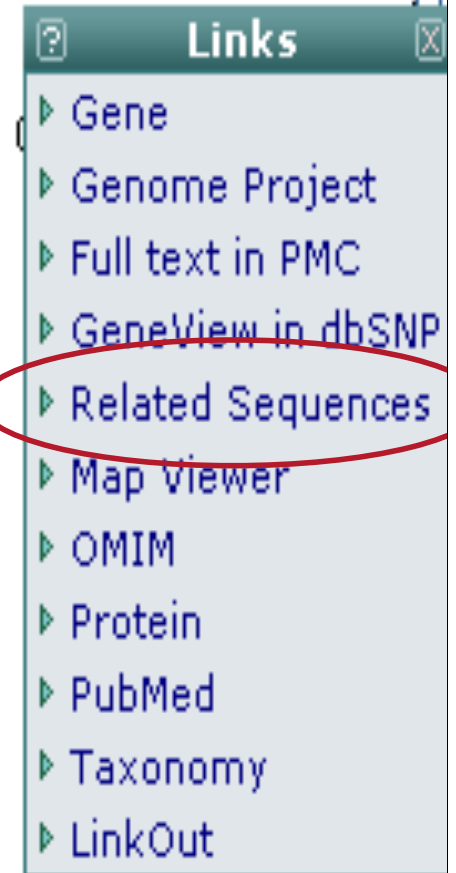
- Links depend on source of information

# Neighbors in Entrez

- Neighboring is another way to link entries
- Connections between entries within a database
  - Similar sequences
  - Related papers
  - Similarity in 3D structure
- Different definition of similarity for each database

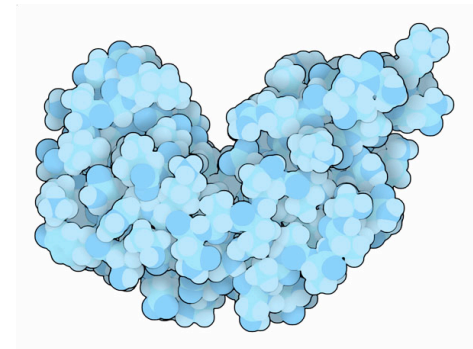
# Related Sequences

- Similar sequences identified using the BLAST program
  - Precomputed BLAST results for all sequences in GenBank
  - Sequence similarity meets a statistical criteria (cutoff)
  - Different list of neighbors for protein sequences vs. nucleotide sequences
  - Two sequences that have a high level of sequence similarity often have related biological functions



# Other Kinds of Neighbors in Entrez

- 3D structures
  - Similar structures



- Proteins with the same fold or arrangement of secondary structure elements
- Identified using a program called VAST
  - Vector Alignment Search Tool
  - Statistical criteria for similarity



# Related Articles

- Similar papers in PubMed

–For more information see:

<http://www.ncbi.nlm.nih.gov/entrez/query/static/computation.html>

–Measured by number of “words” that two papers have in common.

# Abstract Plus Display

The screenshot shows a Mozilla Firefox browser window displaying the Entrez PubMed website. The address bar shows the URL: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list\\_uids=1682344](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=1682344). The page header includes the NCBI logo, the PubMed logo, and the text "A service of the National Library of Medicine and the National Institutes of Health". There are navigation tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "OMIM", "PMC", "Journals", and "Books". The search bar contains "PubMed" and "for". Below the search bar are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The display settings are set to "AbstractPlus", "Show 20", and "Sort by". The results section shows one result: "1: [Nature](#). 2006 Jul 6; 442(7098):37." The abstract title is "Avian flu: multiple introductions of H5N1 in Nigeria." The authors listed are Ducatez MF, Olinger CM, Owoade AA, De Landtsheer S, Ammerlaan W, Niesters HG, Osterhaus AD, Fouchier RA, and Muller CP. The abstract text describes the emergence of the H5N1 virus in Nigeria. The PMID is 16823443. On the right side, there is a "Related Links" section with a "Links" dropdown menu showing options for "Nucleotide", "Protein", "Books", and "LinkOut". The "Related Links" section includes several links related to avian influenza and H5N1 spread.

Entrez PubMed - Mozilla Firefox

File Edit View Go Bookmarks Tools Help del.icio.us

[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list\\_uids=1682344](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=1682344)

The Weather Netwo... lucky.phpwebhostin... e-Learning at UBC She's Right Fit Joanne Fox

NCBI  
PubMed  
www.pubmed.gov

A service of the National Library of Medicine and the National Institutes of Health

My NCBI  
[Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display AbstractPlus Show 20 Sort by Send to

All: 1 Review: 0

1: [Nature](#). 2006 Jul 6; 442(7098):37.

**Avian flu: multiple introductions of H5N1 in Nigeria.**

[Ducatez MF](#), [Olinger CM](#), [Owoade AA](#), [De Landtsheer S](#), [Ammerlaan W](#), [Niesters HG](#), [Osterhaus AD](#), [Fouchier RA](#), [Muller CP](#).

Institute of Immunology, National Public Health Laboratory, 1950 Luxembourg.

As the avian influenza virus H5N1 swept from Asia across Russia to Europe, Nigeria was the first country in Africa to report the emergence of this highly pathogenic virus. Here we analyse H5N1 sequences in poultry from two different farms in Lagos state and find that three H5N1 lineages were independently introduced through routes that coincide with the flight paths of migratory birds, although independent trade imports cannot be excluded.

PMID: 16823443 [PubMed - indexed for MEDLINE]

**Related Links**

- ▶ Avian influenza. Evidence points to migration of H5N1 spread.
- ▶ Bird flu: on border patrol.
- ▶ Doubts hang over source of bird flu spread.
- ▶ Global patterns of influenza A virus in wild birds. [Science, 2006]
- ▶ Avian flu and the New World.

▶ See all Related Articles...

# PubMed Overview

- PubMed is a Web-based retrieval system developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine.
  - It is part of NCBI's vast retrieval system, known as **Entrez**.
- PubMed is a database of bibliographic information drawn primarily from the life sciences literature.

# PMID

- PubMed Unique Identifier = PMID

The screenshot displays the PubMed website interface. At the top, the NCBI logo is on the left, and the PubMed logo with the URL 'www.pubmed.gov' is in the center. To the right, it states 'A service of the National Library of Medicine and the National Institutes of Health'. Further right, there are links for 'My NCBI', 'Sign In', and 'Register'. Below this is a navigation bar with tabs for 'All Databases', 'PubMed', 'Nucleotide', 'Protein', 'Genome', 'Structure', 'OMIM', 'PMC', 'Journals', and 'Books'. The search bar contains the text 'Search PubMed for avian flu' with 'Go' and 'Clear' buttons. Below the search bar are tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The display settings show 'Display Summary', 'Show 20', 'Sort by', and 'Send to'. A summary bar indicates 'All: 3' and 'Review: 0'. The results section shows 'Items 1 - 3 of 3' and 'One page.' Three search results are listed, each with a checkbox, a document icon, a list of authors, a title, journal information, and the PMID. The first result is for a paper by Ducatez MF et al. (PMID: 16823443). The second is for a paper by Le QM et al. (PMID: 16228009). The third is for a paper by Chen H et al. (PMID: 16228009) with a large '6' next to the title 'Avian flu: H5N1 virus outbreak in migratory waterfowl'.

NCBI

PubMed  
www.pubmed.gov

A service of the National Library of Medicine  
and the National Institutes of Health

My NCBI  
[Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for avian flu Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Send to

All: 3 Review: 0

Items 1 - 3 of 3 One page.

1: [Ducatez MF, Olinger CM, Owoade AA, De Landtsheer S, Ammerlaan W, Niesters HG, Osterhaus AD, Fouchier RA, Muller CP.](#) Related Articles, Links  
Avian flu: multiple introductions of H5N1 in Nigeria.  
Nature. 2006 Jul 6;442(7098):37.  
PMID: 16823443 [PubMed - indexed for MEDLINE]


2: [Le QM, Kiso M, Someya K, Sakai YT, Nguyen TH, Nguyen KH, Pham ND, Ngyen HH, Yamada S, Muramoto Y, Horimoto T, Takada A, Goto H, Suzuki T, Suzuki Y, Kawaoka Y.](#) Related Articles, Links  
Avian flu: isolation of drug-resistant H5N1 virus.  
Nature. 2005 Oct 20;437(7062):1108. Erratum in: Nature. 2005 Dec 8;438(7069):754.  
PMID: 16228009 [PubMed - indexed for MEDLINE]

3: [Chen H, Smith GJ, Zhang SY, Qin K, Wang J, Li KS, Webster RG, Peiris JS, Guan Y.](#) Related Articles, Links  
Avian flu: H5N1 virus outbreak in migratory waterfowl | 6  
Nature. 2005 Jul 14;436(7048):191-2. No abstract available




## Anatomy of the Search Results Page


1: [Schwiebert LM](#) **Citation** [Related Articles](#), [Links](#)

**No Abstract**  Cystic fibrosis, gene therapy, and lung inflammation: for better or worse?  
Am J Physiol Lung Cell Mol Physiol. 2004 Apr;286(4):L715-6. Review. No abstract available.  
PMID: 15003935 [PubMed - indexed for MEDLINE]


2: [Pollard HB, Eidelman O, Jacobson KA, Srivastava M](#) [Related Articles](#), [Links](#)

**Abstract**  Pharmacogenomics of cystic fibrosis.  
Mol Interv. 2001 Apr;1(1):54-63. Review.  
PMID: 14993338 [PubMed - indexed for MEDLINE]

3: [Guenert DC, Bruscia E, Novelli G, Colosimo A, Dallepiccola B, Sanguolo F, Goncz KK](#) [Related Articles](#), [Links](#)

**Free in PMC**  Sequence-specific modification of genomic DNA by small DNA fragments.  
J Clin Invest. 2003 Sep;112(5):637-41. Review.  
PMID: 12952908 [PubMed - indexed for MEDLINE]

4: [Florea BL, Meaney C, Junginger HE, Borchard G](#) **Authors** [Related Articles](#), [Links](#)

**Free Full Text**  Transfection efficiency and toxicity of polyethylenimine in differentiated Calu-3 and nondifferentiated COS-1 cell cultures.  
AAPS PharmSci. 2002 4(3):E12 **Page number**  
PMID: 12423061 [PubMed - indexed for MEDLINE]

**Article title**

**Journal title abbreviation** **Date of publication** **Volume and issue number**

# PubMed Feature Tabs

The screenshot shows the PubMed website interface. At the top, the NCBI logo is on the left, and the PubMed logo with the URL [www.pubmed.gov](http://www.pubmed.gov) is in the center. To the right, it says "A service of the National Library of Medicine and the National Institutes of Health". In the top right corner, there is a "My NCBI" section with a user name "joannealisonfox" and a "Sign Out" link.

Below the header, there is a navigation bar with tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "OMIM", "PMC", "Journals", and "Books". The "PubMed" tab is selected.

The main search area has a search box with "PubMed" entered, a dropdown menu, and "Go" and "Clear" buttons. Below the search box, there are five feature tabs: "Limits", "Preview/Index", "History", "Clipboard", and "Details". These tabs are highlighted with a red box.

On the left side, there is a sidebar with links for "About Entrez", "Text Version", "Entrez PubMed", "Overview", "Help | FAQ", "Tutorials", "New/Noteworthy", "E-Utilities", "PubMed Services", "Journals Database", "MeSH Database", "Single Citation Matcher", "Batch Citation Matcher", "Clinical Queries", "Special Queries", "LinkOut", and "My NCBI".

In the center, there is a promotional box for "My NCBI" with the text "Set up an automated PubMed update in less than 5 minutes." and three steps: (1) Get a [My NCBI account](#), (2) Save your search, and (3) Your PubMed updates can be e-mailed directly to you. Below this, it says "Read the [My NCBI Help](#) material to explore other options, such as automated updates of other databases, setting search filters, and highlighting search terms."

At the bottom, there is a paragraph: "PubMed is a service of the [U.S. National Library of Medicine](#) that includes over 16 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources."

# PubMed Feature Tabs

- Limit
  - Limit searches to specific fields, age groups, gender, type of study, Entrez or publication date, a specific language, types of articles, or subsets.
- Preview/Index
  - Use the Preview/Index feature to preview the number of search results before displaying the results
- History
  - Use the History feature to view and combine your previous search queries.

# PubMed Feature Tabs

- Clipboard
  - Use the Clipboard feature to collect selected citations from one or several searches for further action.
- Details
  - Use the Details feature to view your search strategy as it was translated by PubMed.

# Entrez/PubMED

## Boolean Operators

- AND
  - Intersection of terms
  - Entry must have both terms
  - Default
- OR
  - Union of terms
  - Entry must have one of the terms
- NOT
  - Difference
  - Entry does not contain the term

# Reminders:

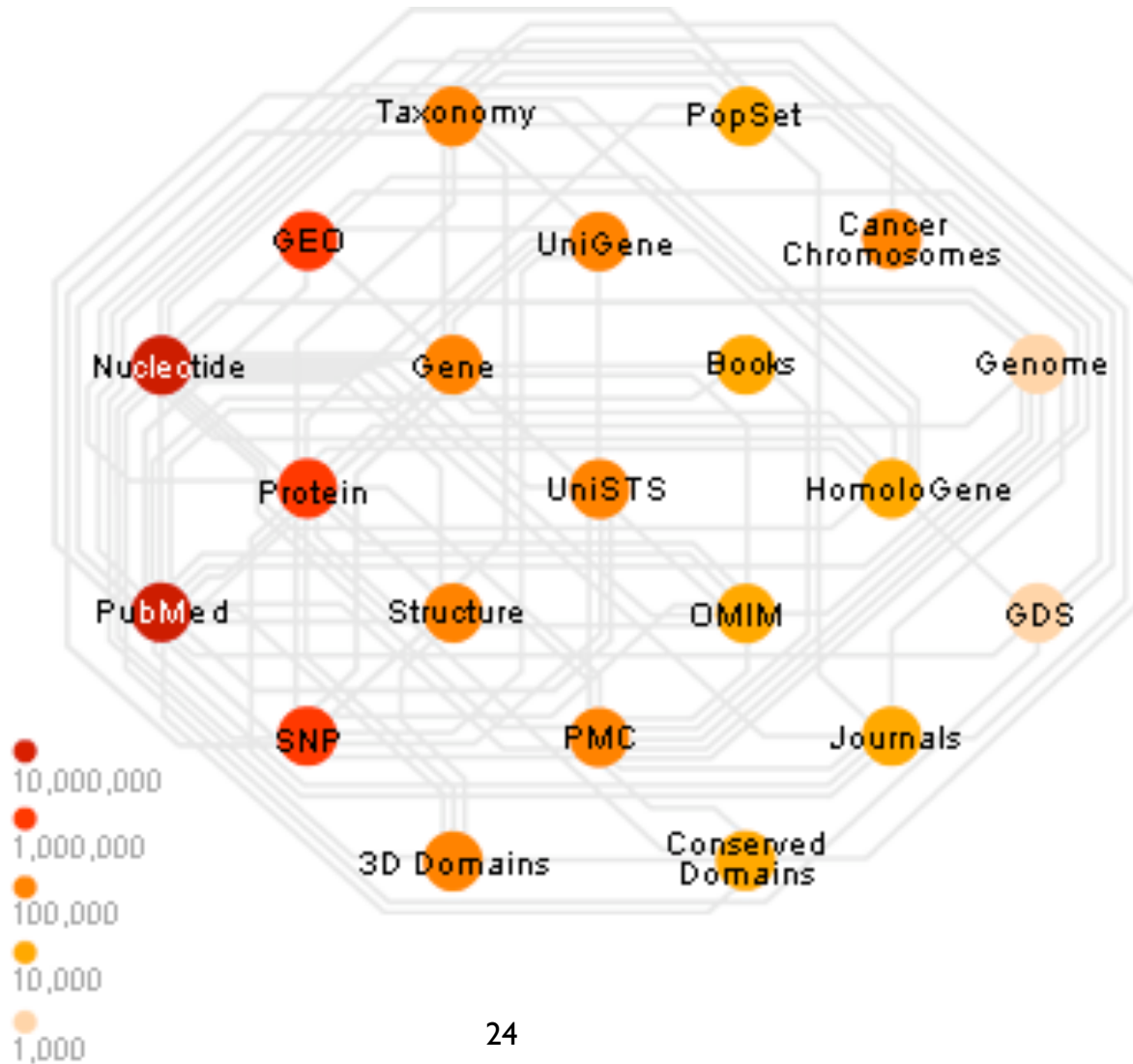
- Boolean operators -- AND, OR, NOT -- should be entered in uppercase letters.
- Boolean operators are processed from left to right.
- Use parentheses to nest terms together so they will be processed as a unit and then incorporated into the overall strategy.

# NCBI Bookshelf

- The **Bookshelf** is a growing collection of biomedical books that can be searched directly
  - free textbooks online

The screenshot shows the NCBI Bookshelf website. At the top left is the NCBI logo. The main header features the word "Bookshelf" in a large, bold, red font. To the right of the header is a "My NCBI" section with a welcome message for "joannealisonfox" and a "Sign Out" link. Below the header is a navigation bar with tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "PMC", "Taxonomy", "OMIM", and "Books". A search bar is located below the navigation bar, with the text "Search Books" and a dropdown menu set to "for bioinformatics". There are "Go" and "Clear" buttons next to the search bar. Below the search bar are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". On the left side of the page is a blue sidebar with a menu containing "About Entrez", "Books", "Overview", "Using the books", "Information for authors and publishers", and "Contact us". The main content area is enclosed in a red border and contains the following text: "The **Bookshelf** is a growing collection of biomedical books that can be searched directly by typing a concept into the textbox above and selecting 'Go'. Try one of these searches: ▶ [cell cycle control](#) ▶ [immunodeficiency](#) ▶ [protein evolution](#) Books are also linked to terms in PubMed abstracts: when viewing an abstract, select the 'Books' link to see phrases within the abstract hyperlinked to book sections. ▶ **New on the Bookshelf:** [Disease Control Priorities in Developing Countries](#) Dean T. Jamison, Joel G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, Prabhat Jha, Anne Mills, Philip Musgrove, editors Washington (DC) IBRD/The World Bank and Oxford University Press; 2006 [Global Burden of Disease and Risk Factors](#) Alan D. Lopez, Colin D. Mathers, Majid Ezzati, Dean T. Jamison, Christopher J. L. Murray, editors Washington (DC) IBRD/The World Bank and Oxford University Press; 2006

# Retrieving Biological Information with Entrez





# Links

- The **About Entrez** page at the NCBI  
<http://www.ncbi.nlm.nih.gov/Database/index.html>
- **Model of Entrez Databases** from NCBI  
<http://www.ncbi.nih.gov/Database/datamodel/index.html>
- **PubMed Tutorial** from NLM  
[http://www.nlm.nih.gov/bsd/pubmed\\_tutorial/m1001.html](http://www.nlm.nih.gov/bsd/pubmed_tutorial/m1001.html)

## Recommended Readings

- Lecture 2.2
  - *Baxevanis & Ouellette (3rd Edition)*
    - Chapter 3: p56 – p77
  - *Westhead, Parish & Twyman*
    - Sections DI