MICB 405 Bioinformatics

Lecture 2.2 – PART A
Retrieving Biological Information with Entrez

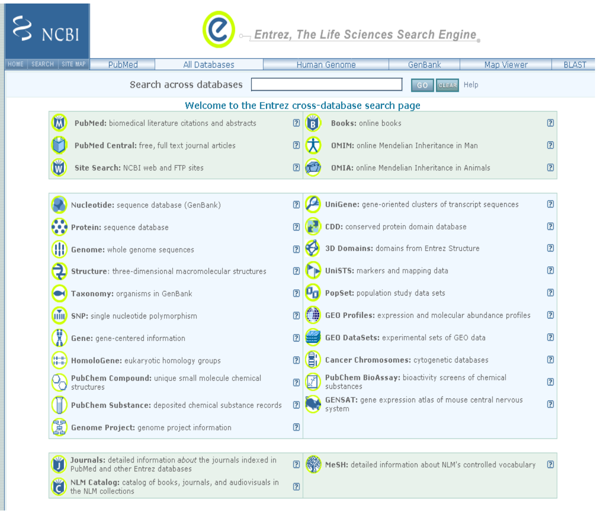
FSC 1221

September 16th, 2008

Retrieving Biological Information

Entrez

BLAST



Entrez, The Life Sciences Search Engine

Search across databases

Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts

PubMed Central: free, full text journal articles

Site Search: NCBI web and FTP sites

Books: online books

OMIM: online Mendelian Inheritance in Man

OMIA: online Mendelian Inheritance in Animals

Nucleotide: sequence database (GenBank)

Protein: sequence database

Genome: whole genome sequences

Structure: three-dimensional macromolecular structures

Taxonomy: organisms in GenBank

SNP: single nucleotide polymorphism

Gene: gene-centered information

HomoloGene: eukaryotic homology groups

PubChem Compound: unique small molecule chemical structures

PubChem Substance: deposited chemical substance records

Genome Project: genome project information

UniGene: gene-oriented clusters of transcript sequences

CDD: conserved protein domain database

3D Domains: domains from Entrez Structure

UniSTS: markers and mapping data

PopSet: population study data sets

GEO Profiles: expression and molecular abundance profiles

GEO DataSets: experimental sets of GEO data

Cancer Chromosomes: cytogenetic databases

PubChem BioAssay: bioactivity screens of chemical substances

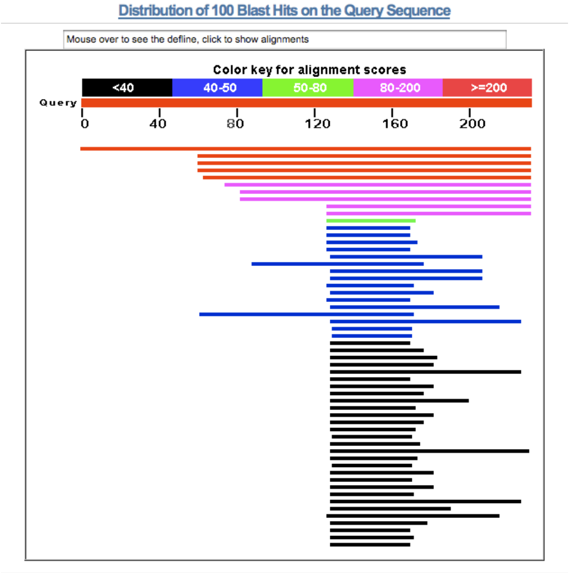
GENSAT: gene expression atlas of mouse central nervous system

Journals: detailed information *about* the journals indexed in PubMed and other Entrez databases

NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections

MeSH: detailed information about NLM's controlled vocabulary

Search by name, identifier, feature



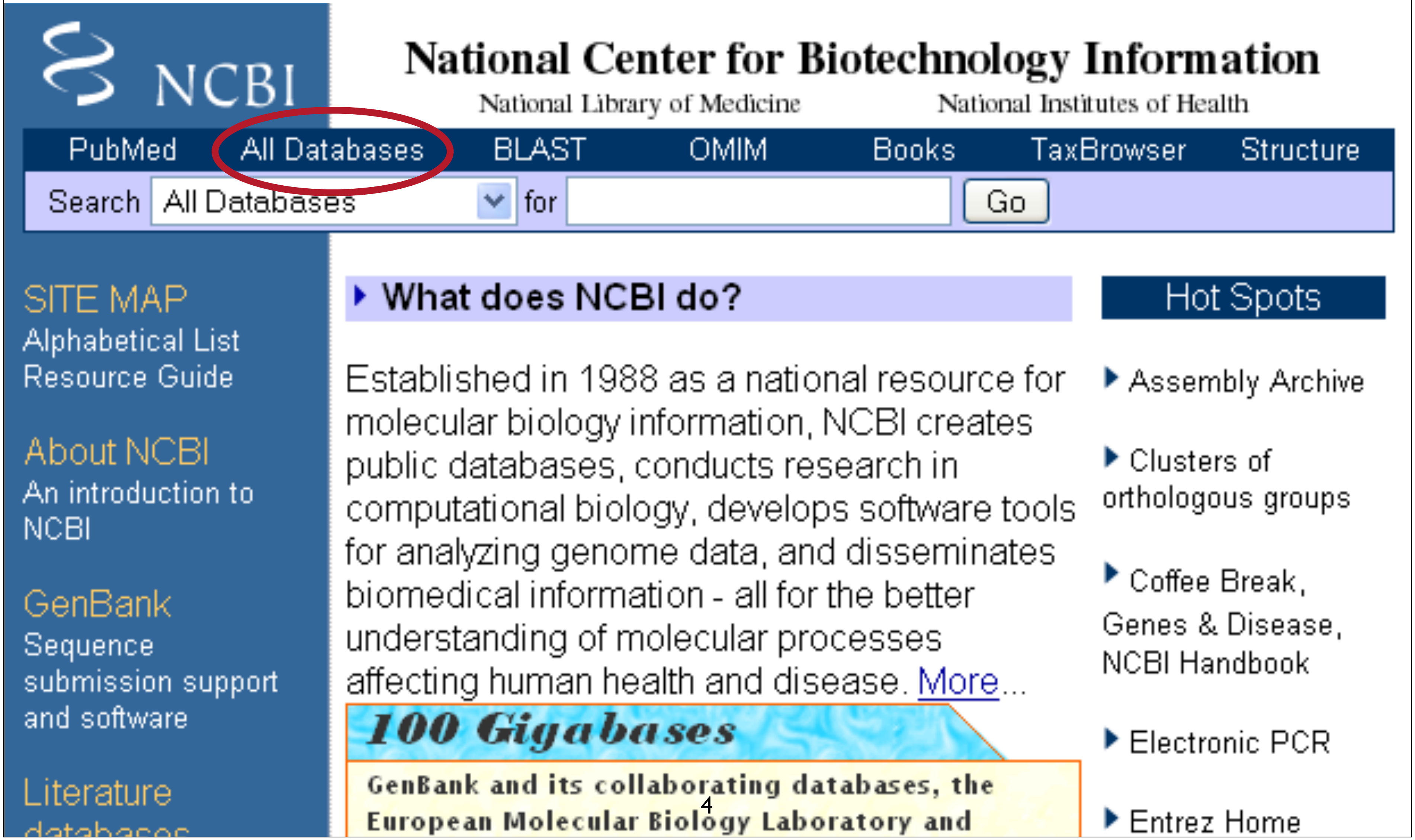
Mouse over to see the defline, click to show alignments

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/



National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

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...

100 Gigabases

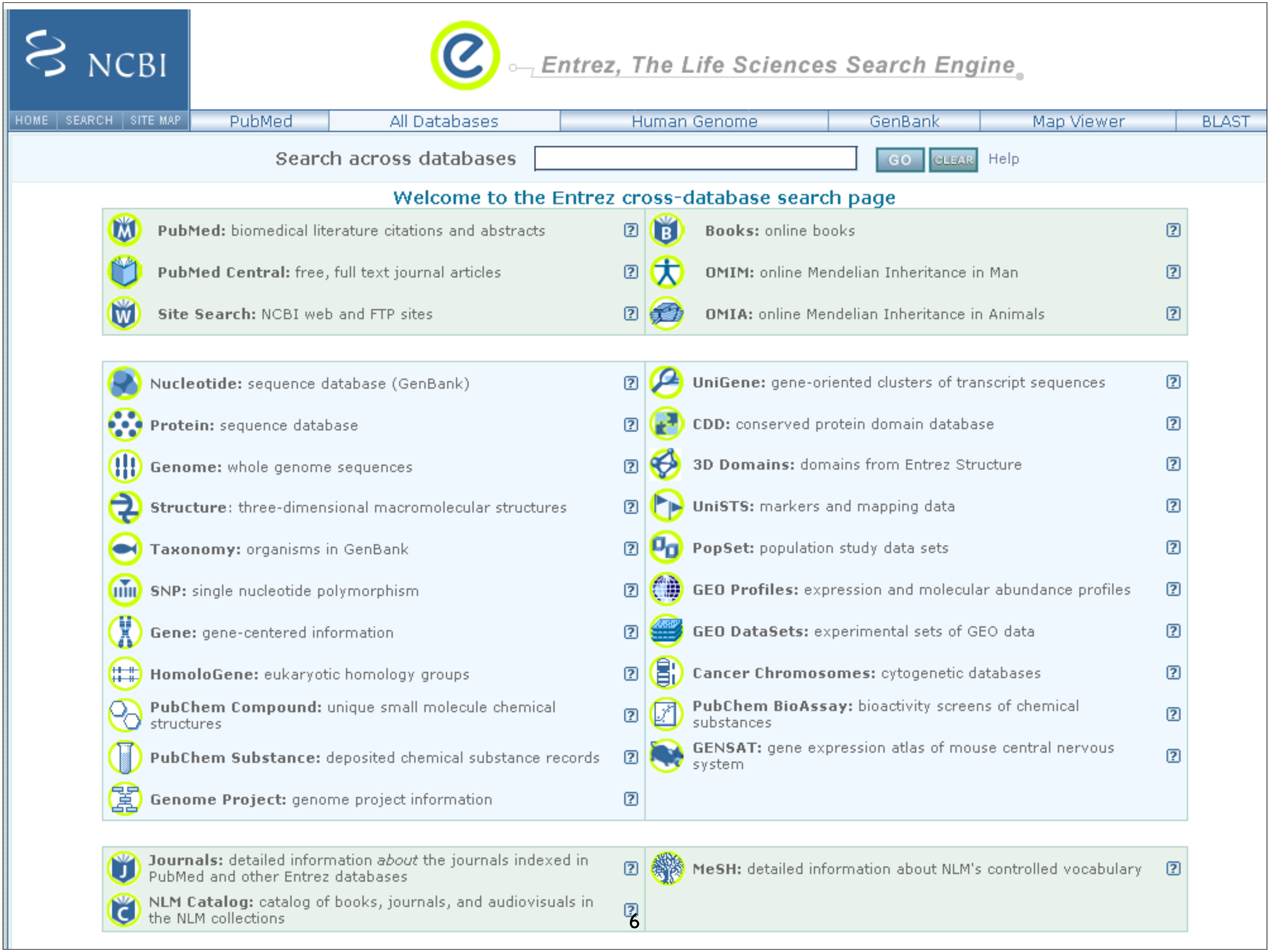
GenBank and its collaborating databases, the European Molecular Biology Laboratory and

Hot Spots

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

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…



Entrez, The Life Sciences Search Engine

Search across databases

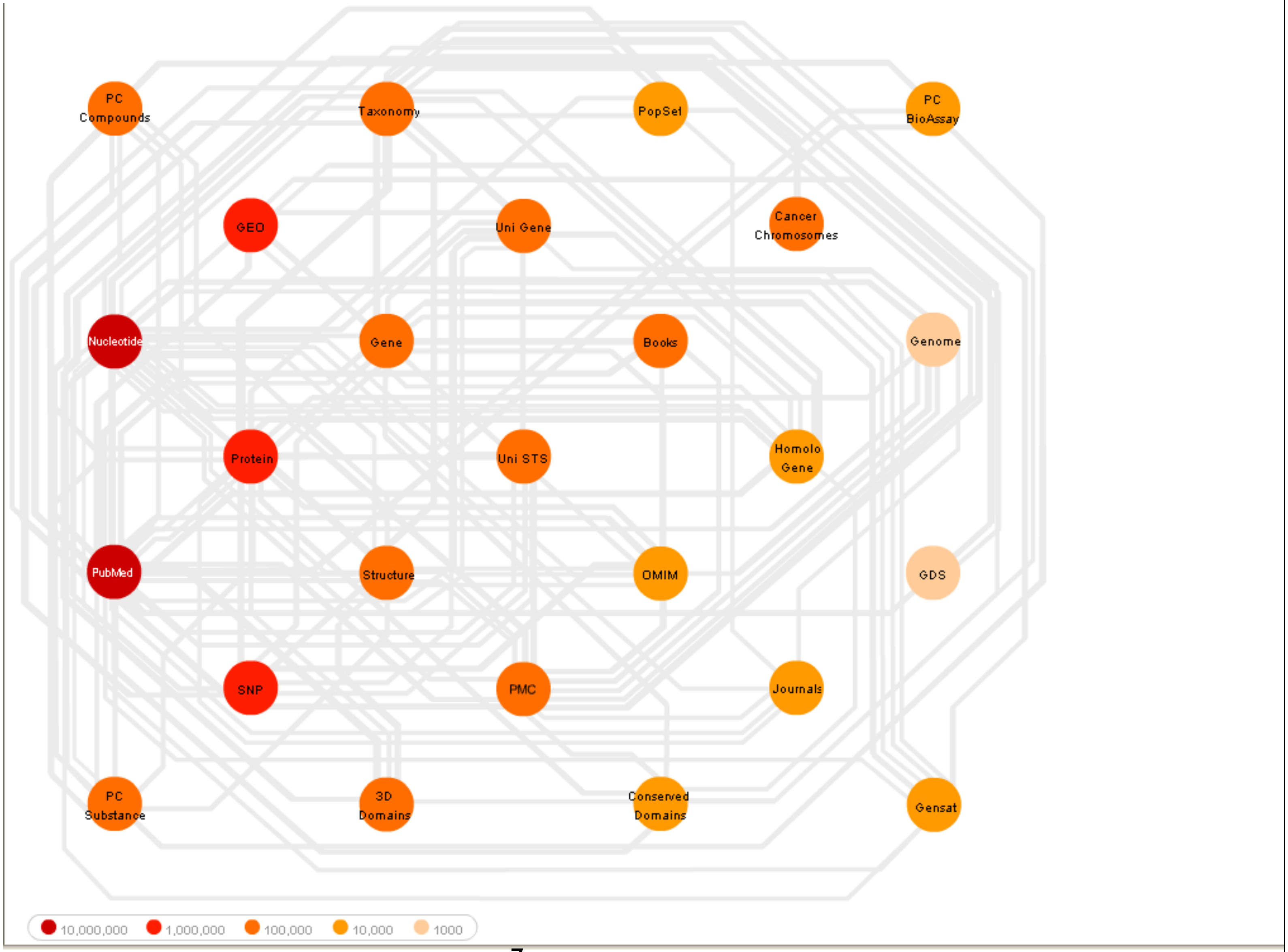
Welcome to the Entrez cross-database search page

- **PubMed:** biomedical literature citations and abstracts
- **PubMed Central:** free, full text journal articles
- **Site Search:** NCBI web and FTP sites
- **Books:** online books
- **OMIM:** online Mendelian Inheritance in Man
- **OMIA:** online Mendelian Inheritance in Animals

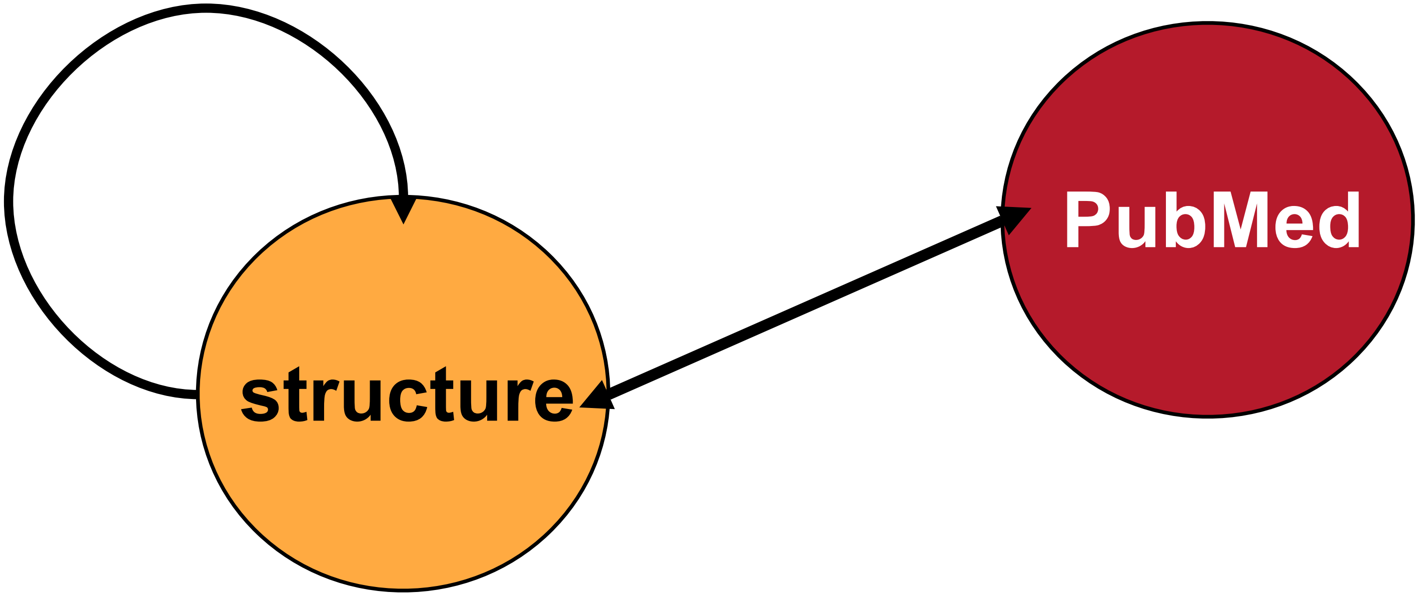
- **Nucleotide:** sequence database (GenBank)
- **Protein:** sequence database
- **Genome:** whole genome sequences
- **Structure**: three-dimensional macromolecular structures
- **Taxonomy:** organisms in GenBank
- **SNP:** single nucleotide polymorphism
- **Gene:** gene-centered information
- **HomoloGene:** eukaryotic homology groups
- **PubChem Compound:** unique small molecule chemical structures
- **PubChem Substance:** deposited chemical substance records
- **Genome Project:** genome project information
- **UniGene:** gene-oriented clusters of transcript sequences
- **CDD:** conserved protein domain database
- **3D Domains:** domains from Entrez Structure
- **UniSTS:** markers and mapping data
- **PopSet:** population study data sets
- **GEO Profiles:** expression and molecular abundance profiles
- **GEO DataSets:** experimental sets of GEO data
- **Cancer Chromosomes:** cytogenetic databases
- **PubChem BioAssay:** bioactivity screens of chemical substances
- **GENSAT:** gene expression atlas of mouse central nervous system

- **Journals:** detailed information *about* the journals indexed in PubMed and other Entrez databases
- **NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections
- **MeSH:** detailed information about NLM's controlled vocabulary

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



Entrez – Linking Data



Entrez – Linking Databases

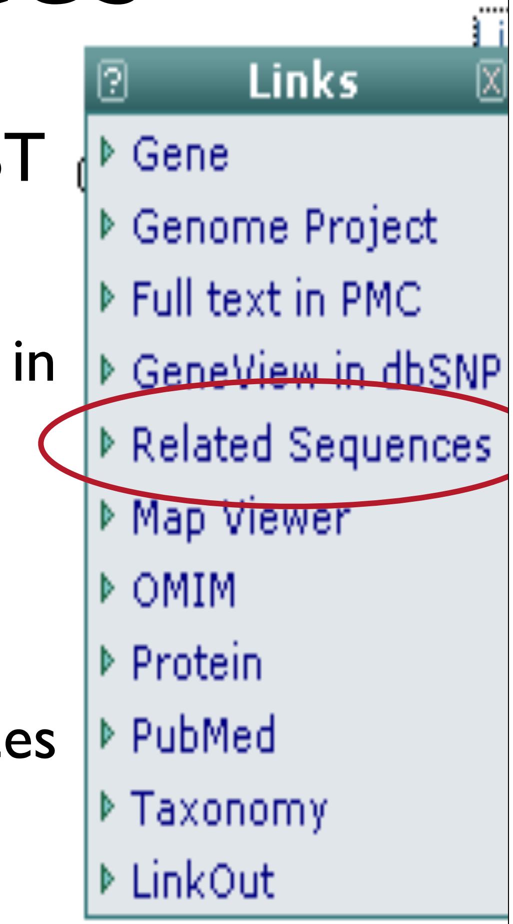
- Hard Links
  - Direct connections between entries in two different databases
    - Examples
      - Link to paper describing a nucleotide sequence
      - Link to taxonomy database for a protein sequence
      - Link from nucleotide sequence to protein CDS
      - Link from protein sequence to 3D structure entry
  - Not all possible hard links are present
    - Links depend on source of information

Neighbors in Entrez

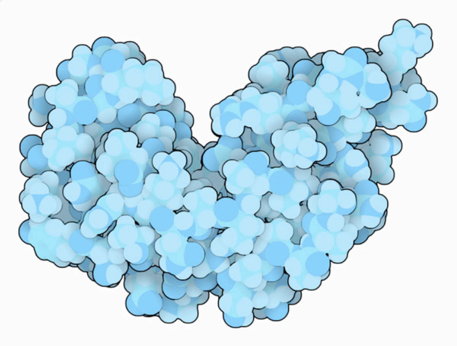
- Neighboring is another way to link entries
- Connections between entries within a database
  - Similar sequences
  - Related papers
  - Similarity in 3D structure
- Different definition of similarity for each database

Related Sequences

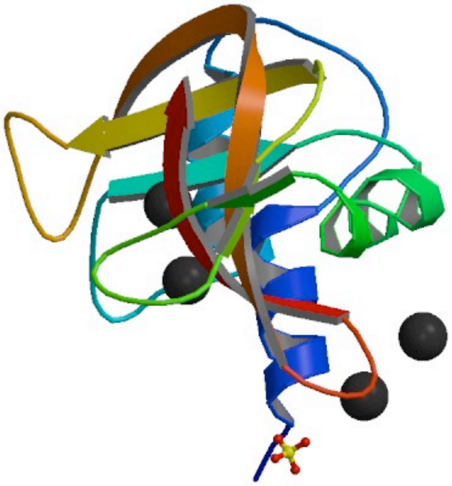
- Similar sequences identified using the BLAST program
  - Precomputed BLAST results for all sequences in GenBank
  - Sequence similarity meets a statistical criteria (cutoff)
  - Different list of neighbors for protein sequences vs. nucleotide sequences
  - Two sequences that have a high level of sequence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



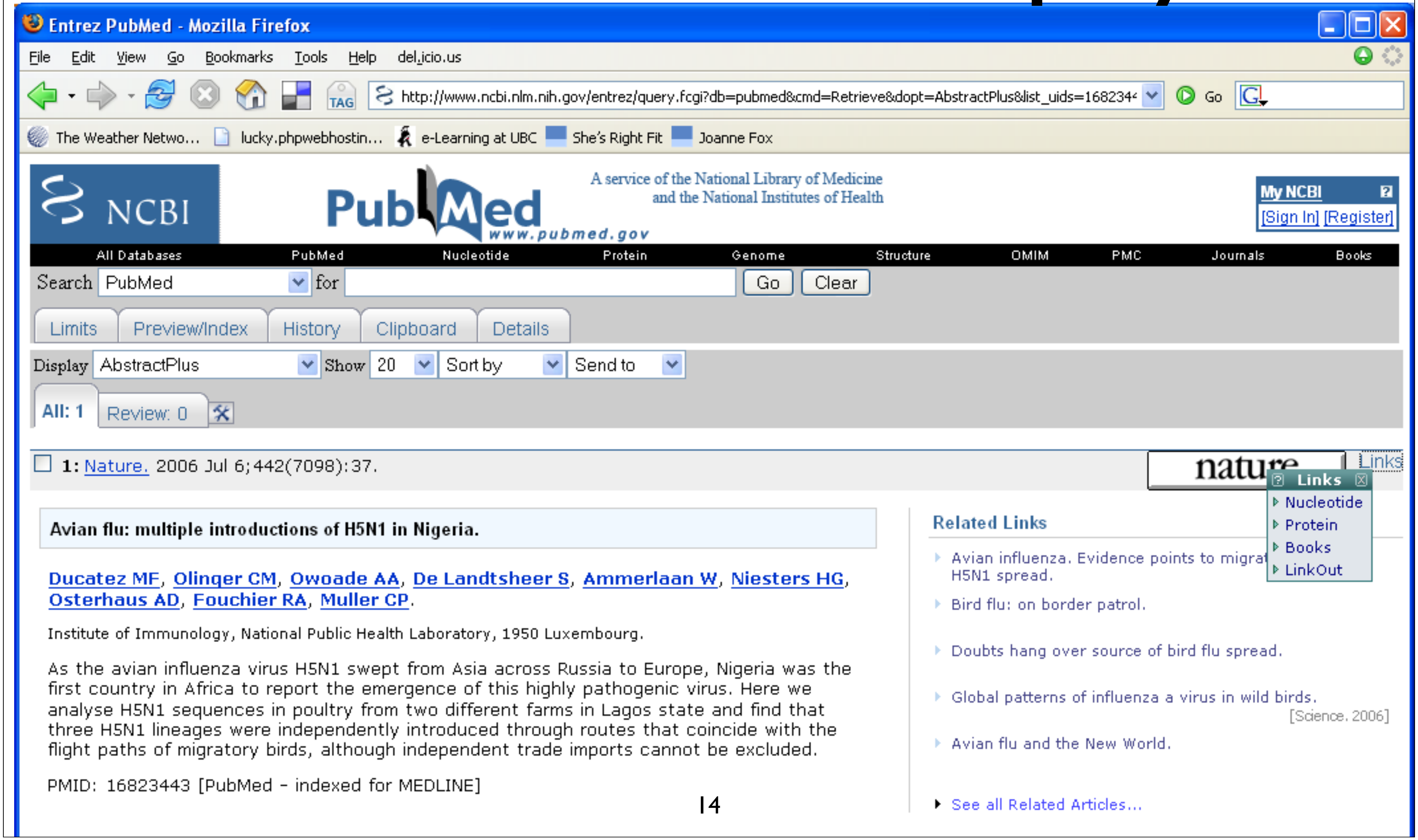
Related Articles

- 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



http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=168234

Search PubMed for

Display AbstractPlus Show 20

☐ **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

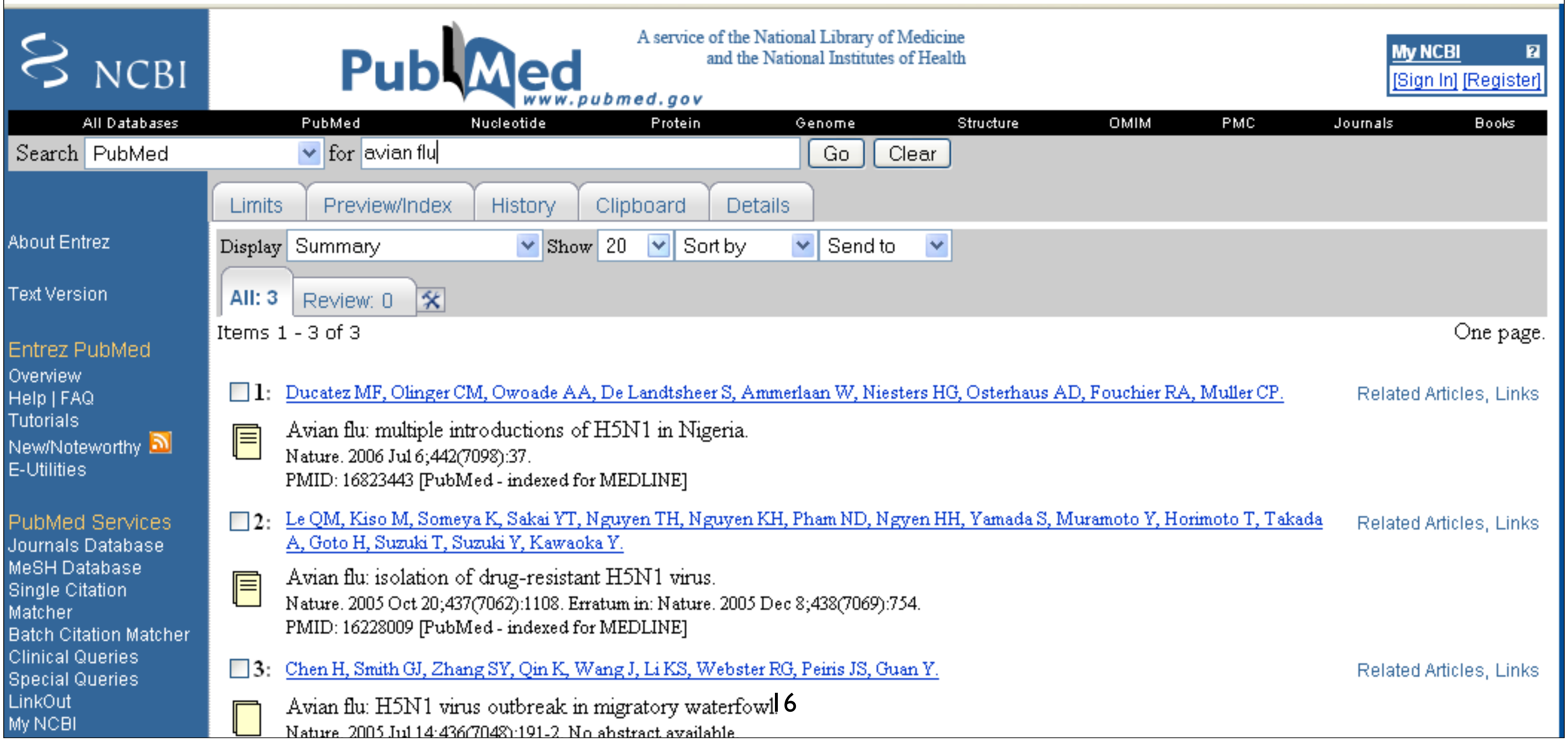
- Avian influenza. Evidence points to migrat H5N1 spread.
- Bird flu: on border patrol.
- Doubts hang over source of bird flu spread.
- Global patterns of influenza a virus in wild birds. [Science. 2006]
- Avian flu and the New World.
- See all Related Articles...

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



Search PubMed for avian flu

Display Summary Show 20 Sort by Send to

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]

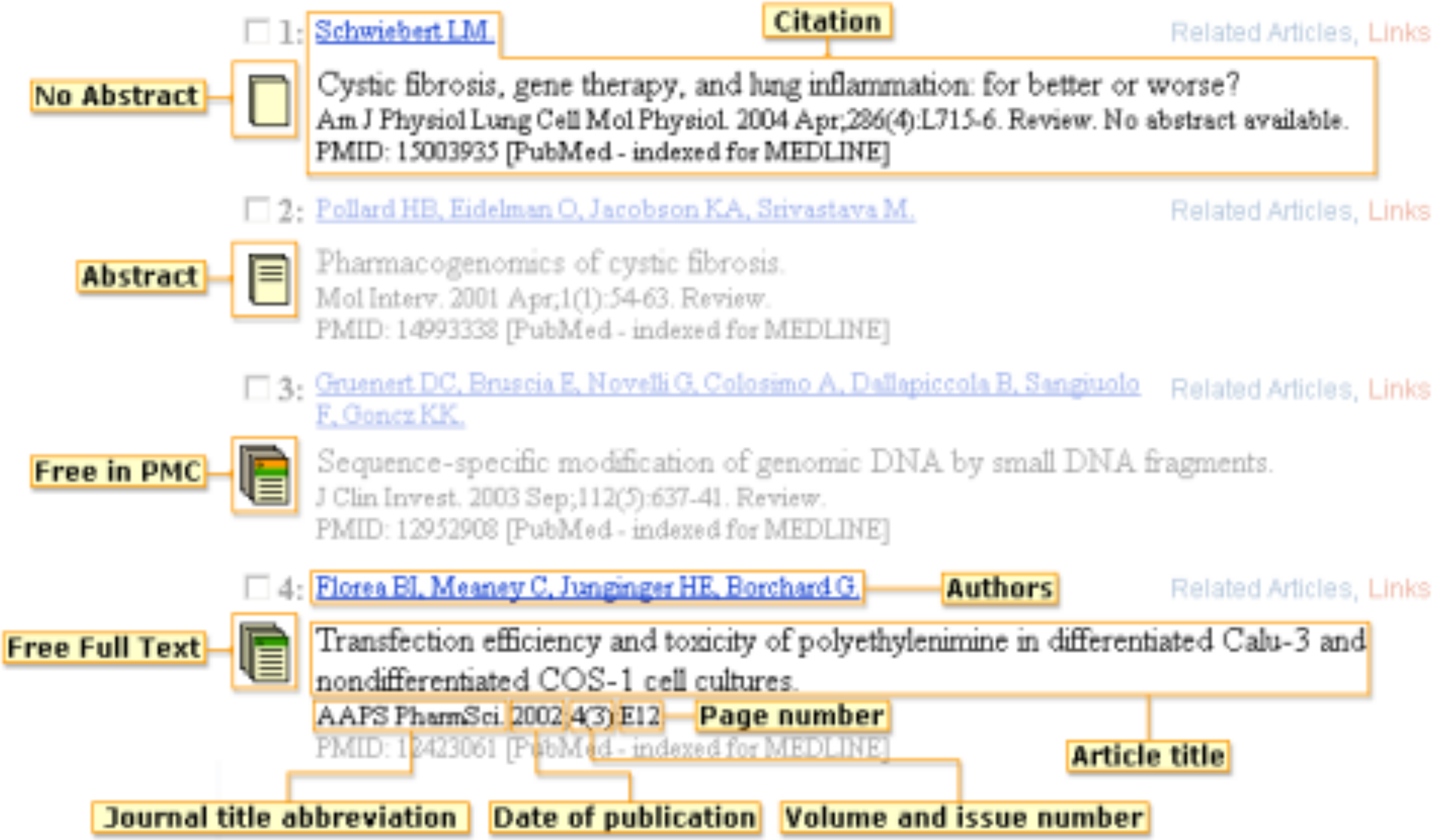
□ 2: Le QM, Kiso M, Someya K, Sakai YT, Nguyen TH, Nguyen KH, Pham ND, Ngyen HH, Yamada S, Muramoto Y, Horimoto T, Takada A, Goto H, Suzuki T, Suzuki Y, Kawaoka Y. Related Articles, Links

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.
Nature. 2005 Jul 14;436(7048):191-2. No abstract available.

Anatomy of the Search Results Page



□ 1: Schwiebert LM. Related Articles, Links

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

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

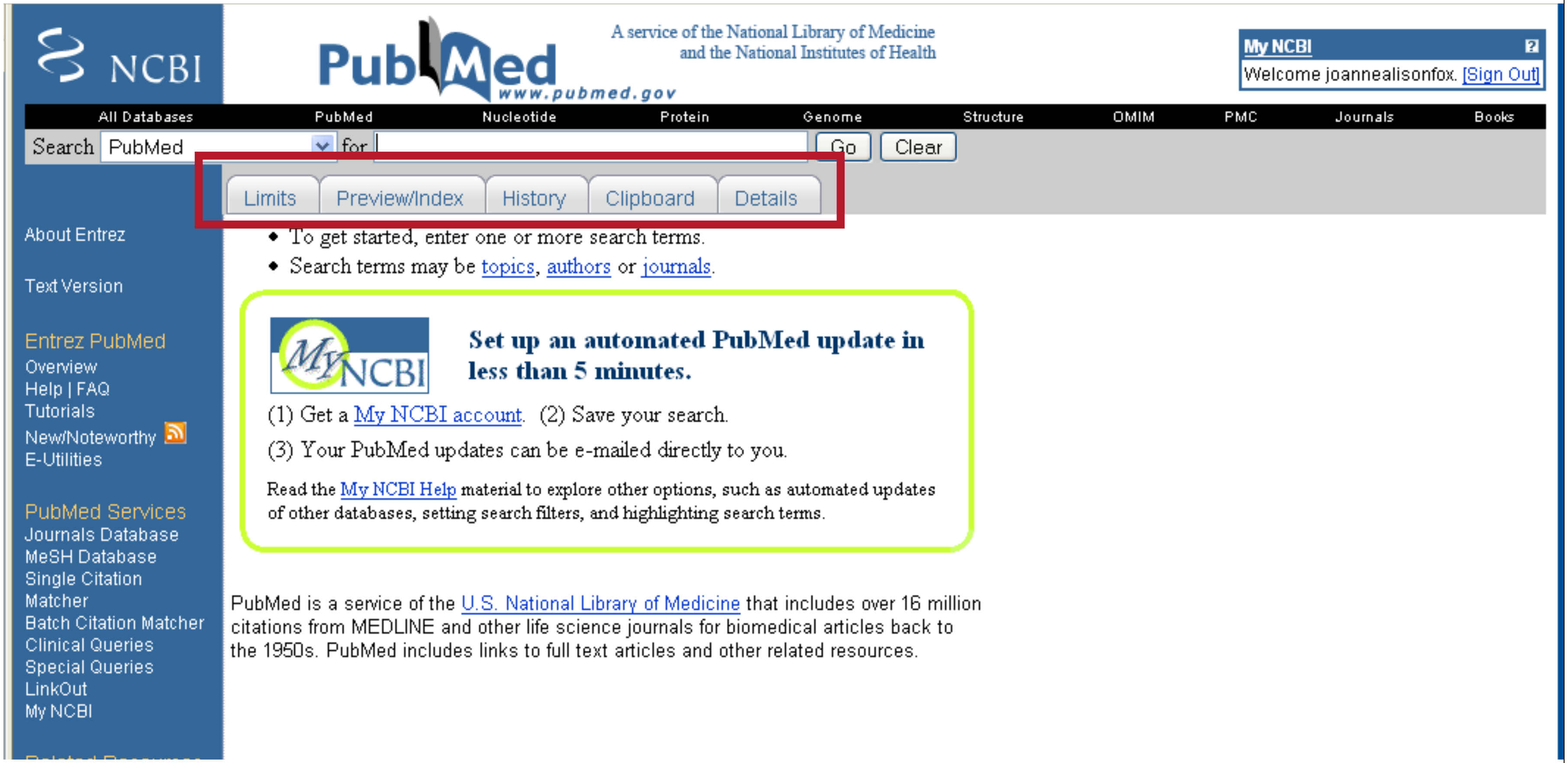
□ 3: Gruenert DC, Bruscia E, Novelli G, Colosimo A, Dallapiccola B, Sangiuolo F, Goncz KK. Related Articles, Links

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 BI, Meaney C, Junginger HE, Borchard G. Related Articles, Links

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

PubMed Feature Tabs



Welcome joannealisonfox. [Sign Out]

- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

Set up an automated PubMed update in less than 5 minutes.

(1) Get a My NCBI account. (2) Save your search.

(3) Your PubMed updates can be e-mailed directly to you.

Read the My NCBI Help material to explore other options, such as automated updates of other databases, setting search filters, and highlighting search terms.

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

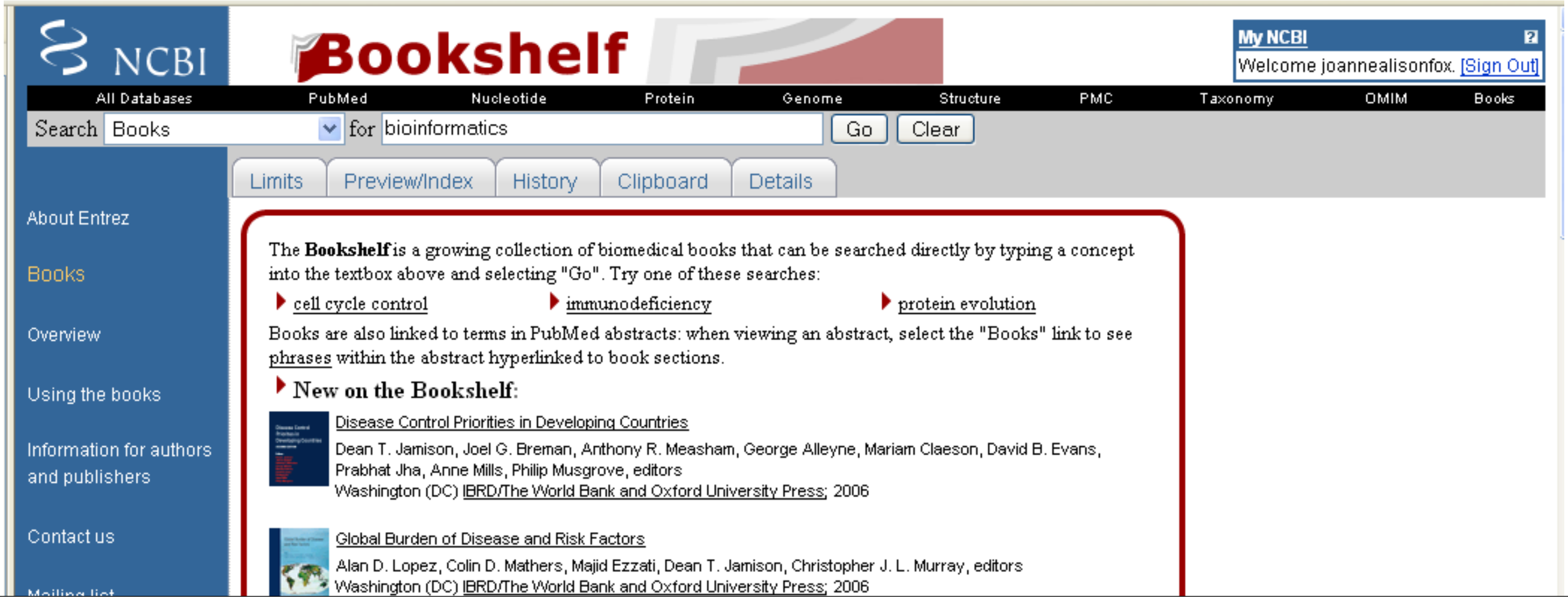
- AND
  - –Intersection of terms
  - –Entry must have both terms
  - –Default
- OR
  - –Union of terms
  - –Entry must have one of the terms
- NOT
  - –Difference
  - –Entry does not contain the term

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 **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

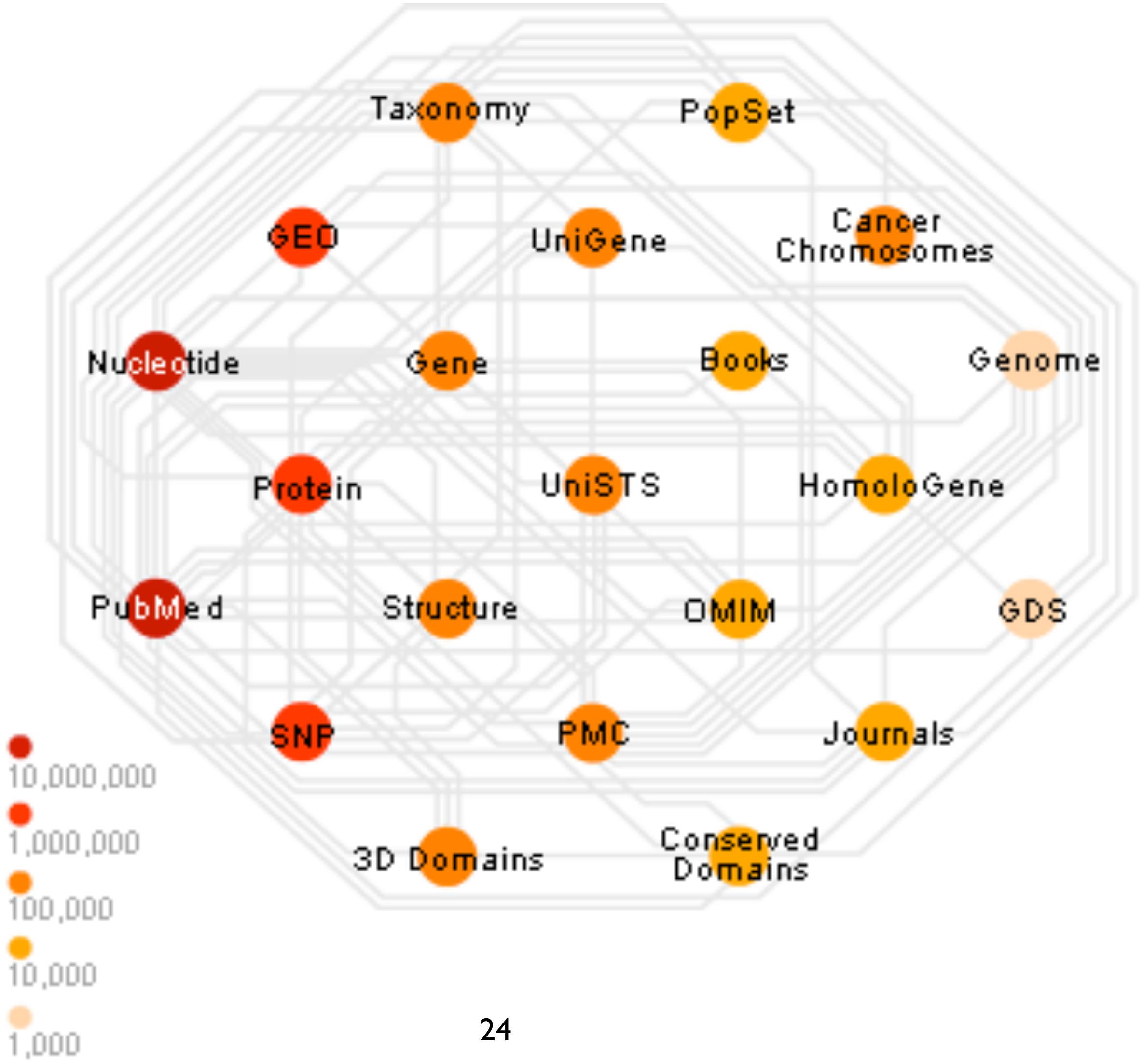
Books are also linked to terms in PubMed abstracts: when viewing an abstract, select the "Books" link to see phrases within the abstract hyperlinked to book sections.

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

Recommended Readings

- Lecture 2.2
  - *Baxevanis & Ouellette (3rd Edition)*
    - Chapter 3: p56 – p77
  - *Westhead, Parish & Twyman*
    - Sections D1