

# Entrez

---

Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>) is a retrieval system for searching several linked databases It provides access to:

- PubMed: The biomedical literature (PubMed)
- Nucleotide sequence database (Genbank)
- Protein sequence database
- Structure: three-dimensional macromolecular structures
- Genome: complete genome assemblies
- PopSet: population study data sets
- OMIM: Online Mendelian Inheritance in Man
- Taxonomy: organisms in GenBank
- Books: online books
- ProbeSet: Gene Expression Omnibus (GEO)

## **Entrez (Cont'd)**

---

- 3D Domains: domains from Entrez Structure
- UniSTS: markers and mapping data
- SNP: single nucleotide polymorphisms
- CDD: conserved domains
- Journals: journals in Entrez
- UniGene: gene-oriented clusters of transcript sequences
- PMC: full-text digital archive of life sciences journal literature

## PubMed and Medical Subject Heading (MeSH)

---

Look at PubMed at <http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=P>  
(p139 in the book)

<http://www.nlm.nih.gov:80/mesh/meshhome.html>

You can use *Limits* link to narrow down your search scope. In general, you can use this link for all Entrez search.

Use *Clipboard* to collect from multiple searches.

Save after clicking *details* link as URL.

Just try it on the web.

## Biological Data on the Web

---

See the first issue of *NAR* journal every year.

1. Sequence databases such as GenBank at <http://www.ncbi.nlm.nih.gov> and Swiss-Prot at <http://www.expasy.ch/spro>
2. Sequence family databases such as Pfam <http://pfam.wustl.edu> and PIR at <http://pir.georgetown.edu>.
3. Biochemical pathway databases such as KEGG <http://www.genome.ad.jp/kegg> and WIT <http://wit.mcs.anl.gov/WIT2>. See also EcoCyc and others at <http://www.ecocyc.org>
4. Structure databases such as PDB <http://www.pdb.org>

# Sequence Data Types at GenBank

---

1. mRNA
2. cDNA  
reverse transcript of mRNA
3. Genomic DNA
4. EST (expressed sequence tag)  
partial cDNA sequences from mRNA extracted from a cell under particular conditions or in specific developmental phases.
5. GSS (genome survey sequence)

# **Genbank flat file format and ASN.1**

---

on the web or p 149

look at NCBI toolkit

## GenBank search result in various formats

---

Changing *Summary* menu to various formats.

1. Brief
2. GenPept
3. ASN.1
4. FASTA
5. LinkOut
6. PubMed Links
7. Protein Neighbors
8. Nucleotide Links

Use *Batch Entrez* to download a large number of sequences.

<http://www.ncbi.nlm.nih.gov/Entrez/batch.html>

## Lab with Batch Entrez

---

1. Go to Nucleotide at <http://www.ncbi.nlm.nih.gov/entrez/>
2. type **oxidoreductase** and limit with **Limits** links.
3. Perform search and change **Summary** to **GI list** and then click **Display**
4. Then change **File** to **Text** and click **Send to**.
5. save the result on your local computer.
6. Then go to <http://www.ncbi.nlm.nih.gov/Entrez/batch.html>
7. Locate the saved GI-list file on your computer by clicking **Browse** and then click **Retrieve**.
8. Then change **Summary** to **FASTA** and then click **Display**.
9. Then you have a list of seqs for your analysis!

## Browsing Genomes at NCBI

---

Try `http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome`

You may try all bacterial genomes by following **Microbial** or  
at `http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html`

[G] - GenBank format; [F] - FASTA format; [T] - TaxTable; [P] - ProtTable;  
[C] - COG Table; [D] - 3-D neighbors; [L] - BLAST; [S] CDD search

Try [D] **3-D neighbors** and retrieve related (similar) PDB entries.

## Protein Data Bank (PDB)

---

PDB is a database of sequence structures and their raw data.

Search can be done

1. with PDB identifiers, for instance **1A00**
2. using the *SearchLite* or
3. using the *SearchFields*

Visit <http://www.pdb.org>.

The raw data is available in PDB format or mmCIF (MAcromolecular Crystallographic Information File) format.

The structures can be viewed using plug-in such as RasMol or Chime. (p154)