

# **Database search engines**

**Shatakshi Kulkarni**

# Entrez

- ❑ <https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html>
- ❑ [Search NCBI databases - NLM \(nih.gov\)](#)
- ❑ Global Query Cross Database search system
- ❑ It is an integrated search and retrieval system and allows to search discrete health sciences databases
- ❑ Allows access to large database only by a single query string and Efficiently retrieves sequences, structures and references
- ❑ It allows users to search for and retrieve information from multiple NCBI databases through a single interface.
- ❑ Provides view of gene, gene sequences and chromosome maps
- ❑ Ability to integrate information by cross-referencing from NCBI based on pre-existing & logical relationships between individual entries
- ❑ Databases linked to Entrez retrieval system:
  - a) PubMed
  - b) Books (Online books)
  - c) Nucleotide sequence database
  - d) Protein databases
  - e) Genome databases
  - f) Structure databases

# Entrez

## ❑ Entrez search fields:

- Keyword allows to search a set of indexed terms
- Accession allows to search accession numbers
- Author Name
- Affiliations of authors
- Journal Title
- E.C. Numbers
- Feature Key searches for particular DNA feature
- SeqId is string identifier
- Title Words
- Text Words
- Organism
- PubMed ID
- Publication and modification date
- Protein Name

# Entrez

ncbi.nlm.nih.gov/Web/Search/entrezfs.html

NCBI

Entrez Molecular Sequence Database System

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI SITE MAP

### Introduction

Entrez is a molecular biology database system that provides integrated access to nucleotide and protein sequence data, gene-centered and genomic mapping information, 3D structure data, PubMed MEDLINE, and more. The system is produced by the National Center for Biotechnology Information (NCBI) and is available via the Internet.

### Entrez Databases and Retrieval System

Entrez covers over 20 databases including the complete protein sequence data from PIR-International, PRF, Swiss-Prot, and PDB and nucleotide sequence data from GenBank that includes information from EMBL and DDBJ.

The Entrez retrieval system uses an intuitive user interface for rapidly searching sequence and bibliographic data. A unique feature of the system is its use of precomputed similarity searches for each record to create links to "neighbors" or related records in other Entrez databases. These links facilitate integrated access across the various databases. An Entrez global query provides search capability for a subset of Entrez databases at one time. Results may be viewed in various formats including FlatFile, FASTA, XML, and others. A graphical interface provides easy visualization of complete genomes or chromosomes, as well as biological annotation on individual sequences. Entrez also allows Batch downloads of large search results.

### Internet Access to Entrez

Entrez is available via the World Wide Web at <http://www.ncbi.nlm.nih.gov/Entrez/>

## Homepage

ncbi.nlm.nih.gov/search/

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine National Center for Biotechnology Information

Search NCBI

### Literature

#### PubMed

PubMed® comprises more than 37 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

#### Featured Bookshelf titles

[Amyotrophic Lateral Sclerosis](#)  
Araki T, editor.

[Drug Therapy for Early Rheumatoid Arthritis](#)  
A Systematic Review Update  
Donahue KE, Gartlehner G, Schulman ER, et al.

#### Literature databases

**Bookshelf**  
Books and reports

**MeSH**  
Ontology used for PubMed indexing

**NLM Catalog**  
Books, journals and more in the NLM Collections

**PubMed**  
Scientific and medical abstracts/citations

**PubMed Central**  
Full-text journal articles

### Data

#### Genes

Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

**Gene**  
Collected information about gene loci

**GEO DataSets**  
Functional genomics studies.

**GEO Profiles**  
Gene expression and molecular abundance profiles

#### Proteins

Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

**Conserved Domains**  
Conserved protein domains

**Identical Protein Groups**  
Protein sequences grouped by identity

**Protein**  
Protein sequences

#### BLAST

A tool to find regions of similarity between biological sequences

**blastn**  
Search nucleotide sequence databases

**blastp**  
Search protein sequence databases

**blastx**  
Search protein databases using a translated nucleotide query

Search NCBI

Homo sapiens BRCA1

Search

Results found in 22 databases

GENE

**BRCA1 - BRCA1 DNA repair associated**

*Homo sapiens (human)*

Also known as: BRCA1, BRCC1, BROVCA1, FANCS, IRIS, PNC44, PPP1R53, PSCP, RNF53

Gene ID: 672

RefSeq transcripts (368) RefSeq proteins (368) RefSeqGene (1) PubMed (3,360)

Orthologs Genome Data Viewer BLAST

### RefSeq transcripts

BRCA1 - 3 of 368 transcripts

Transcript	Isoform	Len (nt)
NM_001408458.1	116	3,785
NM_001407967.1	85	6,390
NM_001407959.1	78	6,851

[View full table](#) NCBI Datasets

### RefSeq Sequences

#### Literature

Bookshelf	202
MeSH	0
NLM Catalog	1
PubMed	18,881
PubMed Central	92,860

#### Genes

Gene	1,818
GEO DataSets	10,702
GEO Profiles	185,122
PopSet	81

#### Proteins

Conserved Domains	2
Identical Protein Groups	223
Protein	3,508
Protein Family Models	7
Structure	199

#### Genomes

Assembly / Genome	NCBI Datasets
BioCollections	0
BioProject	423
BioSample	15,044

#### Clinical

ClinicalTrials.gov	0
ClinVar	87,717
dbGaP	0
dbSNP	41,177

#### PubChem

BioAssays	0
Compounds	0
Pathways	0
Substances	0

## After searching for a query

# SRS/Sequence Retrieval System

## ❑ Retrieval:

- It is an information indexing and retrieval system designed for libraries with a flat file format such as the EMBL nucleotide sequence databank, the SwissProt protein sequence databank or the Prosite library of protein subsequence consensus patterns.
- Retrieval system maintained by EBI
- The SRS search engine allows researchers to perform precise searches, retrieve specific sequences, and conduct comprehensive analyses.
- Not integrated as Entrez but allows to query multiple databases
- Offers direct access to sequence analysis applications
- Queries launched using Quick text search

## ❑ Input/Output Flexibility:

- Supports various formats, with particular adaptation to GCG/Genetics Computer Group programs, making it a versatile tool for bioinformatics research.
- ❑ SRS (Sequence Retrieval System) is a system designed to support the data structure of various biological databases (libraries) by creating special indices.
- ❑ These indices help manage and navigate complex data formats, including:
  - **Lists of sub-entities:** For example, feature tables in a genomic database that list genes, exons, or other annotations.
  - **Hierarchically structured data-fields:** For instance, taxonomic classification that organizes species into a hierarchy of kingdoms, phyla, classes, etc.