

Databases



Sequence Databases

Bibliographic Databases

Clinical Databases

Integrated Databases

Structural Databases

Sequence Databases



**Release/
Updates**

Nucleotide Databases:

EMBL: European Molecular Biology Laboratory

Genbank

DDBJ: DNA Data Bank of Japan

Current Release: 49,2 Million entries

**International
repository for all
nucleotide
sequences
submitted by
researchers**

Accession numbers are unique to each entry.

**One alphabetical character is followed by five digits, or
two alphabetical characters are followed by six digits.**

Sequence Databases

Nucleotide Databases:

RefSeq: Reference Sequence

NC_123456

Complete Prokaryote Genome

Complete Eukaryote Chromosome

NG_123456

Homo sapiens Genomic Region

A database of non-redundant reference sequences standards, including genomic DNA contigs, mRNAs and proteins for known genes. Contributions are taken from the NCBI and collaborative sequencing efforts

NM_123456

mRNA of several organisms, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*

Those accession numbers beginning with X indicate model entries produced as a result of the Genome Annotation process.

Sequence Databases

Protein Databases:

SwissProt: Swiss Protein

Entry names are often the name of the gene followed by the species. Accession numbers are of the following format:

[O,P,Q] [0-9] [A-Z, 0-9] [A-Z, 0-9] [A-Z, 0-9] [0-9],

e.g. P26367 (PAX6_HUMAN)

Contains translated sequences from EMBL, adaptations from PIR, extracted from the literature and directly submitted by researchers. Annotation is high quality and the data is cross-referenced to other databases.

Sequence Databases

Protein Databases:

TrEMBL: Translated EMBL

SpTrEMBL & RemTrEMBL

Acts as a supplement to SwissProt and contains translated EMBL sequences with automatic annotation. TrEMBL entries are manually annotated before being entered into SwissProt.

Remaining TrEMBL contains entries that will never be incorporated into SwissProt. These include: immunoglobulins; T-cell receptors; small fragments; synthetic sequences; CDS not coding for real proteins; patent application sequences

SwissProt TrEMBL contains entries which will eventually be integrated into the SwissProt database. SwissProt accession numbers have been assigned.



Sequence Databases

Protein Databases:

PIR: Protein Information Resource

The PIR is a computer system offering both peptide and nucleotide sequences designed to aid protein identification.

Although much of the protein information in the PIR has been integrated into SwissProt, it may contain some unique sequences.



Sequence Databases

Protein Databases:

SwissProt + TrEMBL + PIR = UNIPROT

Current Release: 1,9 Million entries

UniParc – protein archive database

Released as of January 2005
without restriction

Sequence Databases

Protein Databases:

RefSeqP: Reference Sequence Proteins

RefSeqP provides a protein reference standard for the central dogma. It is used, as is RefSeq, to provide a foundation for the functional annotation of the human genome.

Accession numbers for all proteins are of the format: NP_123456

Sequence Databases



Searching for a sequence:

Text Search:

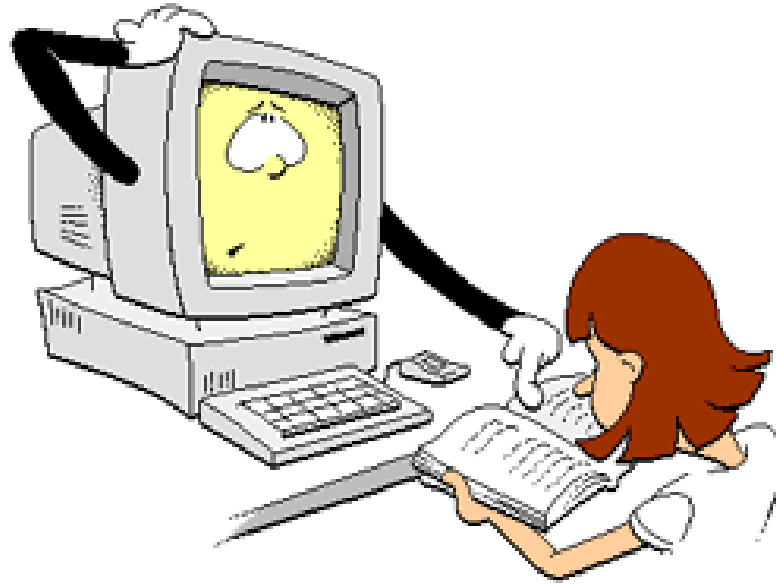
Use text with a boolean operator

BRCA1 & BRCA2 – searches for BRCA1 **AND** BRCA2

BRCA1 | BRCA2 – searches for one gene **OR** the other

BRCA1 ! BRCA2 – searches for BRCA1 **BUT NOT** BRCA2

Computers are **THICK!**



Database entries often presented as **flatfiles**

Each piece of information is on a separate line, distinguished by a code. Computers index this code, so they can search for the relevant entry.



EMBL entry for a sequence fragment implicated in Human Breast Cancer

Identification	ID	AY144588 standard; DNA; HUM; 68 BP.
Accession	AC	AY144588;
Sequence Version	SV	AY144588.1
Date	DT	23-SEP-2002 (Rel. 73, Created)
	DT	23-SEP-2002 (Rel. 73, Last updated, Version 1)
Description	DE	Homo sapiens truncated breast and ovarian cancer susceptibility protein
Keyword	DE	(BRCA1) gene, partial cds.
Organism Source	KW	.
Organism Classification	OS	Homo sapiens (human)
	OC	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
	OC	Eutheria; Primates; Catarrhini; Hominidae; Homo.



Reference Number

RN [1]

Reference Position

RP 1-68

Reference Author

RA Rajkumar T., Soumittra N., Nirmala Nancy K., Shanta V.;

Reference Title

RT "Novel 5bp deletion in BRCA1 gene in South Indian family";

Reference Location

RL Unpublished.

RN [2]

RP 1-68

RA Rajkumar T., Soumittra N., Nirmala Nancy K., Shanta V.;

RT ;

RL Submitted (27-AUG-2002) to the EMBL/GenBank/DDBJ databases.

RL
India

Molecular Oncology, Cancer Institute (WIA), Canal Bank Road, Adyar, RL Chennai, TN 600020,



Feature Table
Header

Feature Table
Data

FH	Key	Location/Qualifiers
FH		
FT		source 1..68
FT		/country="India: South India"
FT		/db_xref="taxon:9606"
FT		/note="identical sequence found in daughter with breast cancer"
FT		/sex="female"
FT		/organism="Homo sapiens"
FT		/isolation_source="mother with breast cancer"
FT		/dev_stage="adult"
FT		/mRNA 68
FT		/gene="BRCA1"
FT		/product="truncated breast and ovarian cancer susceptibility protein"

```

FT CDS <1..68
FT /codon_start=3
FT /note="contains premature stop codon due to
frameshift
FT caused by deletion"
FT /product="truncated breast and ovarian cancer
susceptibility protein"
FT /protein_id="AAN10167.1"
FT /translation="EAASGCETSVSEDCSGLSE"
FT exon 1..68
FT /number=12
FT /gene="BRCA1"
FT misc_feature 61..62
FT /note="site of deletion"
FT /gene="BRCA1"
SQ Sequence 68 BP; 19 A; 12 C; 23 G; 14 T; 0 other;
gtgaagcagc atctgggtgt gagagtgaaa caagcgtctc tgaagactgc tcagggctat 60
cagagtga
//

```

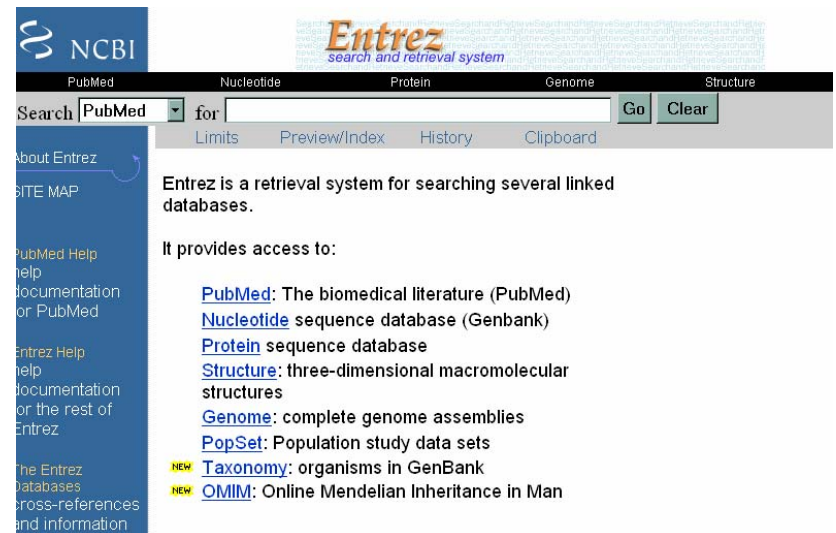
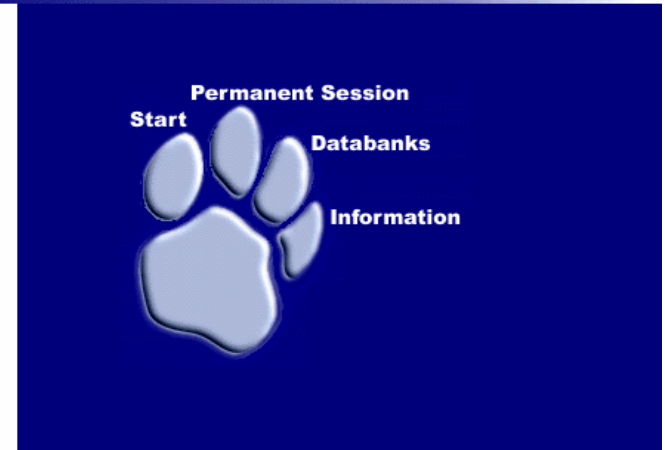
Sequence Header

Searching the databases with a “search engine”:

The Sequence Retrieval System (SRS) from LION Bioscience AG is a very common search tool

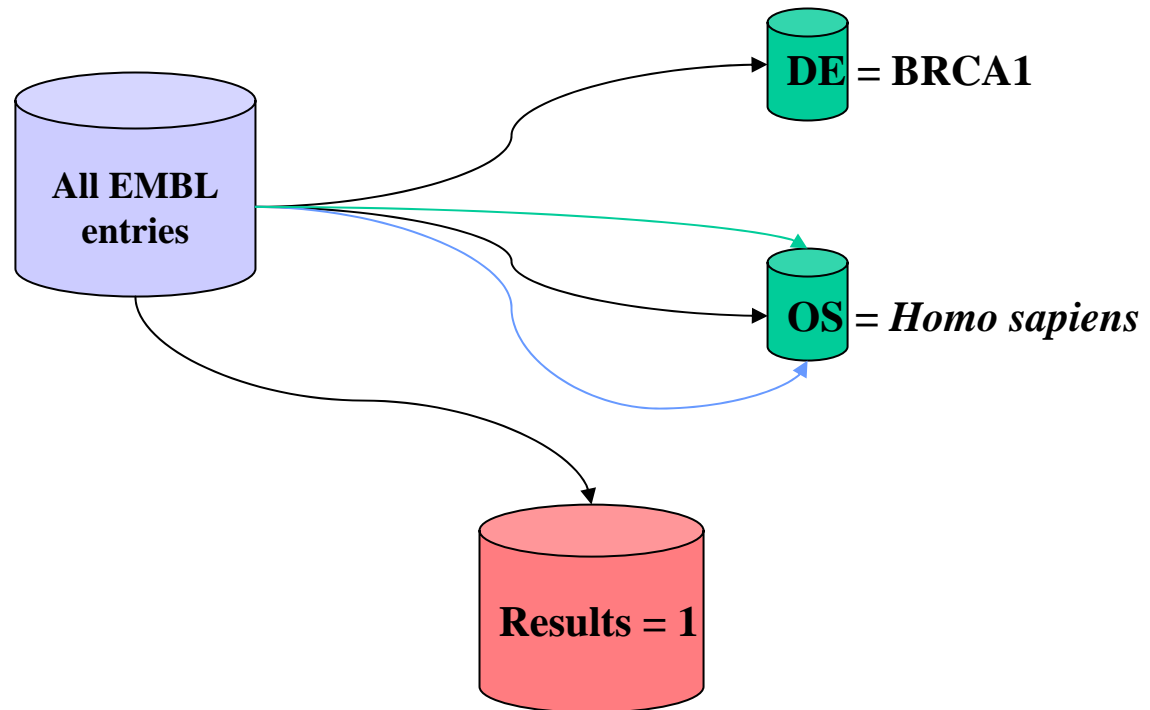
The NCBI in the USA has its own search engine called Entrez.

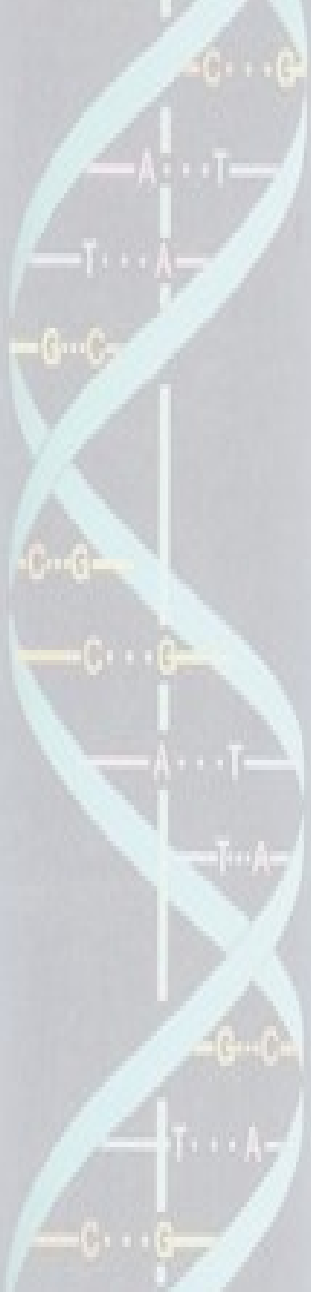
Version 6.1



To search for the BRCA1 gene in Homo sapiens in the EMBL database:

BRCA1 [DE] & Human [OS]





Bibliographic Databases

Used for searching for reference articles

*For all (loosely) medically related papers, use **PubMed** from the NCBI*

Currently holds over 12 million MEDLINE entries.

The screenshot shows the Entrez-PubMed website in a Microsoft Internet Explorer browser window. The address bar displays the URL: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>. The page features the NCBI logo, the PubMed logo, and the National Library of Medicine (NLM) logo. A navigation bar includes tabs for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, Taxonomy, OMIM, and Books. Below this is a search bar with the text "Search PubMed for" and buttons for "Go" and "Clear". A secondary navigation bar includes "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area contains a list of search instructions: "Enter one or more search terms, or click [Preview/Index](#) for advanced searching", "Enter [author names](#) as smith jc. Initials are optional.", and "Enter [journal titles](#) in full or as MEDLINE abbreviations. Use the [Journals Database](#) to find journal titles." Below this is a yellow box with the text: "PubMed, a service of the National Library of Medicine, provides access to over 12 million MEDLINE citations back to the mid-1960's and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources." There are also sections for "Bookshelf Additions" and "New Journals Database". The left sidebar contains links for "About Entrez", "Text/Version", "Entrez PubMed", "PubMed Services", and "Related Resources".



Bibliographic Databases

Other scientific databases may include:

Web of Knowledge: <http://wok.mimas.ac.uk>

Web of Knowledge

Free to academics, but requires username and password

PubCrawler: <http://www.pubcrawler.ie>

Free to academics, will search journals and sequences daily, weekly or monthly and alert the user when results are found corresponding to their search



Clinical Databases

Generally contain information from the Human.

Human Gene Mutation Database, Cardiff, UK:

<http://www.hgmd.org>

Registers known mutations in the human genome and the diseases they cause.

dbSNP, Bethesda, USA:

<http://ncbi.nlm.nih.gov/SNP/>

The largest database for single nucleotide polymorphisms. Accession numbers used in dbSNP are not compatible with other SNP databases.



Integrated Databases

These contain overview information garnered from a variety of different databases, and then offer links to further information.

GeneCards: <http://bioinformatics.weizmann.ac.il/cards>

An extremely thorough overview of a particular gene, with links to various other integrated and clinical databases.

Interpro: <http://www.ebi.ac.uk/interpro>

Integration of individual protein resources PRINTS; PROSITE; SMART; ProDom; Pfam; TIGRfam into one database. A search will scan entries of each and output results.

Integrated Databases

Ensembl: <http://www.ensembl.org>

A joint project by EBI and Sanger to annotate all the information currently known about the human genome in one larger database

e!
Ensembl

Browsing Contig Display

In addition to sequence displays a map of DNA fragments is shown giving the location of genes.

Each display is a magnified view of the red window in the display above.

Location on the chromosome

1Mb overview of the region

Landmark map markers

Genes positions are shown under the map

Adjacent contigs are shown in alternating blue

Use these buttons to move and resize your view

Use these menus to reconfigure your view and access advanced features.





Structural Databases

Tertiary protein structure prediction is possibly the Holy Grail of bioinformatics.

PDB: Protein DataBank, New Jersey, USA

<http://www.rcsb.org/>

EMSD: EBI Macromolecular Structure Database

<http://www.ebi.ac.uk/msd/index.html>

Management and distribution of data on macromolecular structures in close collaboration with the PDB.

This houses a collection of 3D coordinates of each atom in a protein, allowing the structure to be displayed by viewing software. Protein structures are submitted by individual researchers and have been determined by x-ray diffraction, or NMR.



Structural Databases

SCOP: Structural Classification of Proteins

<http://scop.mrc-lmb.cam.ac.uk/scop/>

**Current Release: 686 folds; 1073 Superfamilies; 1827 Families
representing 15,979 PDB entries**

CATH: Classification, Architecture, Topology, Homology

http://www.biochem.ucl.ac.uk/bsm/cath_new/

Current Release: 36,480 Domains