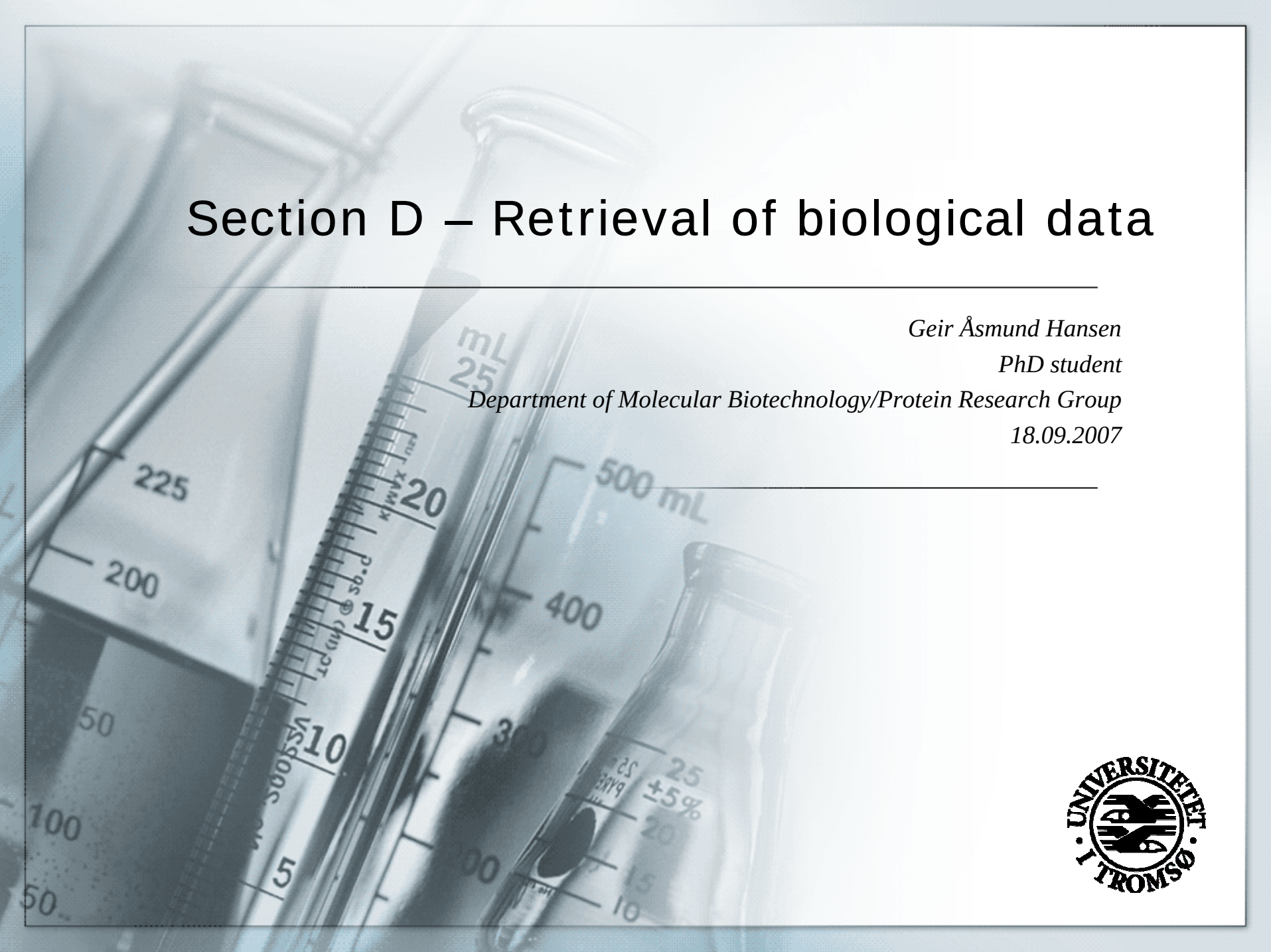
Section D – Retrieval of biological data

Geir Åsmund Hansen
PhD student
Department of Molecular Biotechnology/Protein Research Group
18.09.2007



Outline of lecture

- D1: Data retrieval with Entrez and DBGET/LinkDB
- D2: Data retrieval with SRS (Sequence retrieval system)

Section D1 –
Data retrieval with Entrez and DBGET/LinkDB

- Access to distributed data
- Entrez
- Getting started with NCBI and Entrez
- DBGET/LinkDB

Section D1 –
Access to distributed data

- A large amount of biological data is widely distributed on the internet
- The problem is to find the data you are looking for and to download it in a useable format
- Some information are difficult to find using standard search engines
- Therefore, some internet tools are dedicated to data retrieval and gathers available data from many different databases
  - Entrez
  - DBGET
  - SRS
- Each of the tools allow text or sequence based searches linked up to a number of different databases
- The data retrieval tools differ in the databases they cover and how the retrieved information is accessed and presented

Section D1 – Entrez

- Data retrieval system developed by the National Center for Biotechnology Information (NCBI)
- Integrates information from all the different NCBI databases
  - Nucleotide sequences (GenBank and its subsidiaries)
  - Protein sequences
  - Macromolecular structures
  - Whole genomes
- Entrez searches also other resources outside NCBI
  - OnLine Mendelian Inheritance in Man (OMIM)
  - Medline (through PubMed)
- In total, Entrez links to 11 databases

Section D1 – Entrez

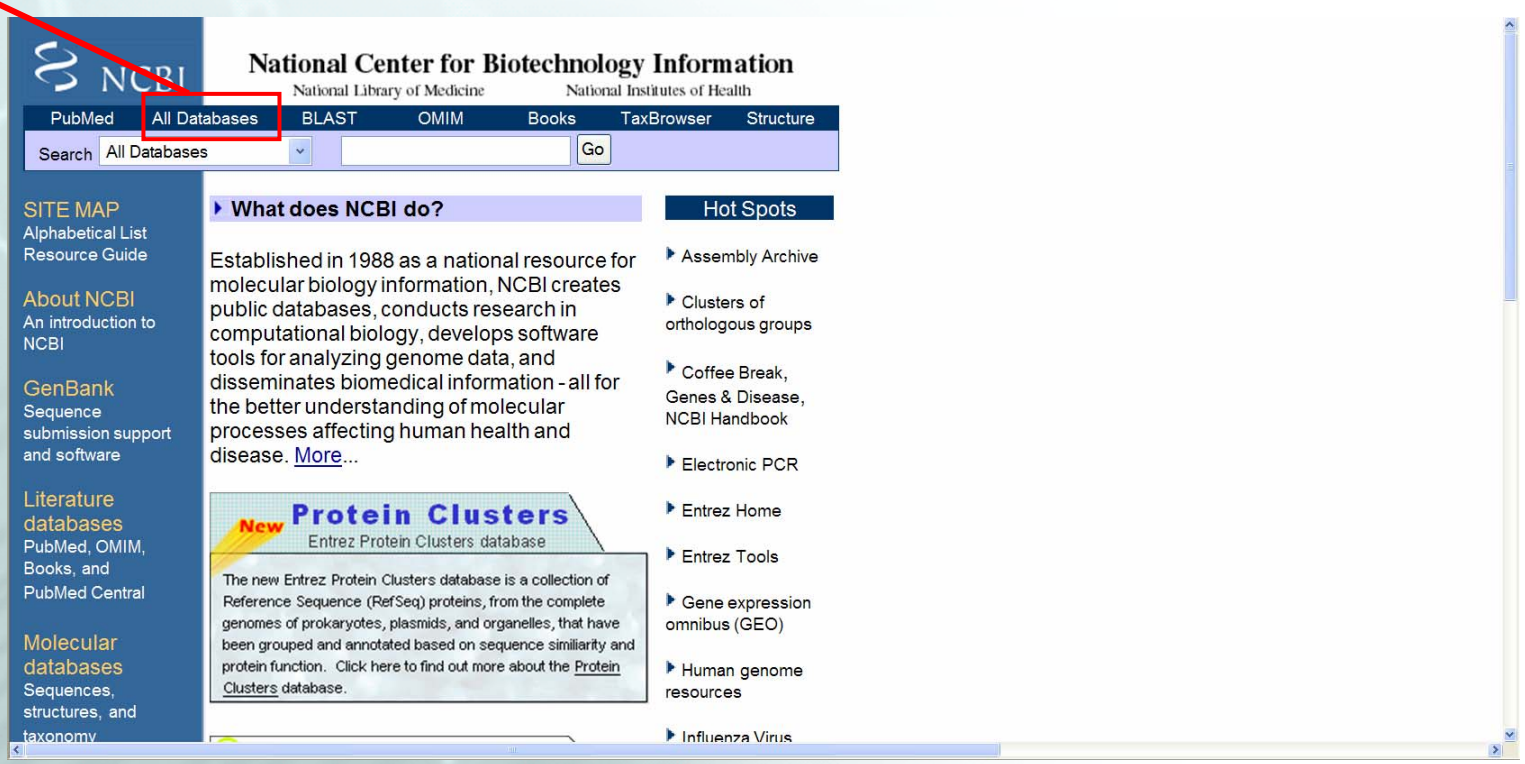
Category	Database
Nucleic acid sequences	Entrez nucleotides: sequences obtained from GenBank, RefSeq and PDB
Protein sequences	Entrez protein: Sequences obtained from SWISS-PROT, PIR, PRF, PDB and translations from annotated coding regions in GenBank and RefSeq
3D structures	Entrez Molecular Modelling Database (MMDB)
Genomes	Complete genome assemblies from many sources
PopSet	From GenBank, set of cDNA collected to analyse the evolutionary relatedness of a population
Omim	OnLine Mendelian Inheritance in Man
Taxonomy	NCBI Taxonomy Database
Books	Bookshelf
ProbeSet	Gene Expression Omnibus (GEO)
3D domains	Domains from Entrez Molecular Modelling Database (MMDB)
Literature	PubMed

Section D1 –
Getting started with NCBI and Entrez

- Entrez is accessed via the NCBI homepage
- Text search used to search individual databases
- Sequences used to search BLAST
- Hits listed in order of relevance or similarity

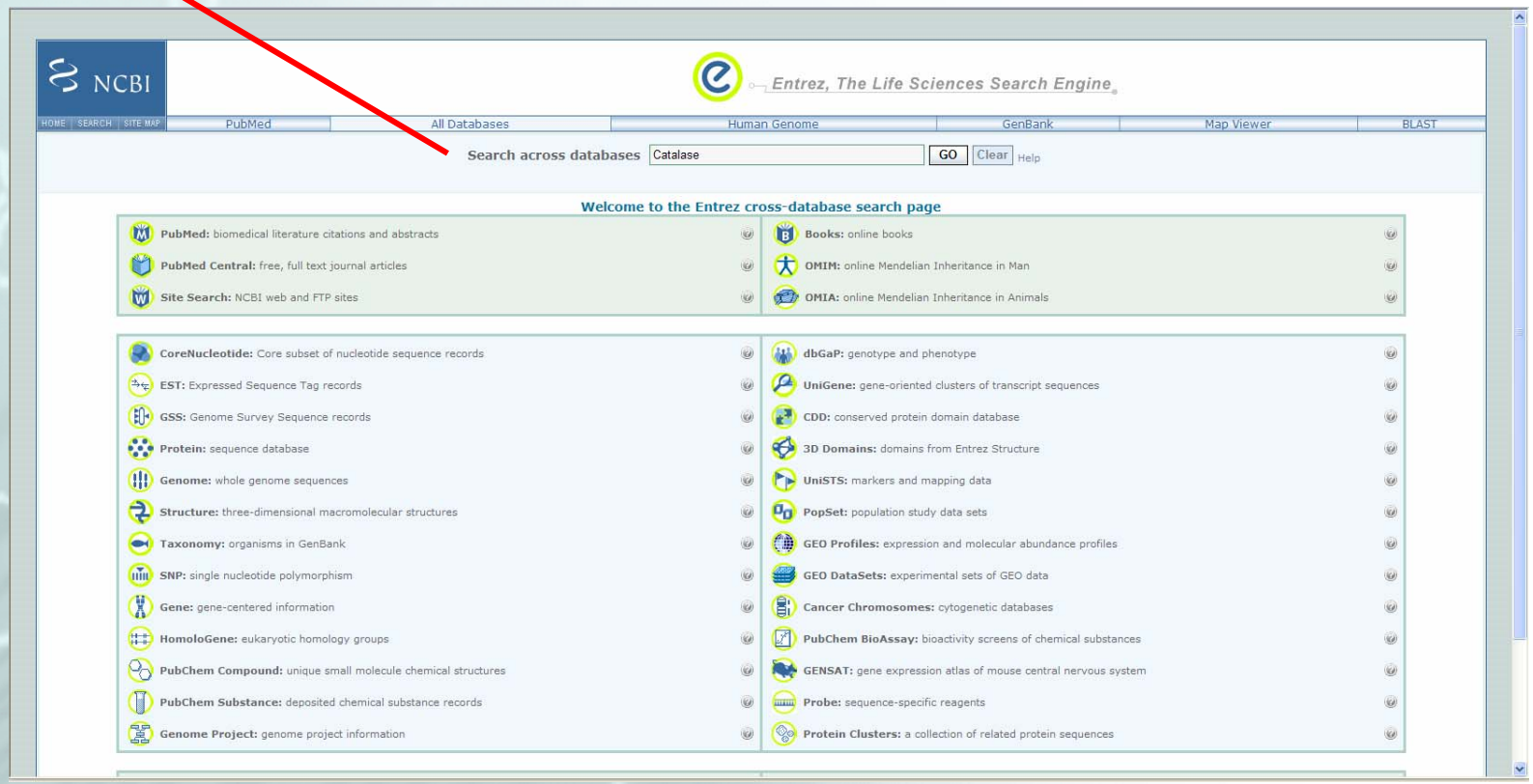
Section D1 – Entrez

Link to Entrez

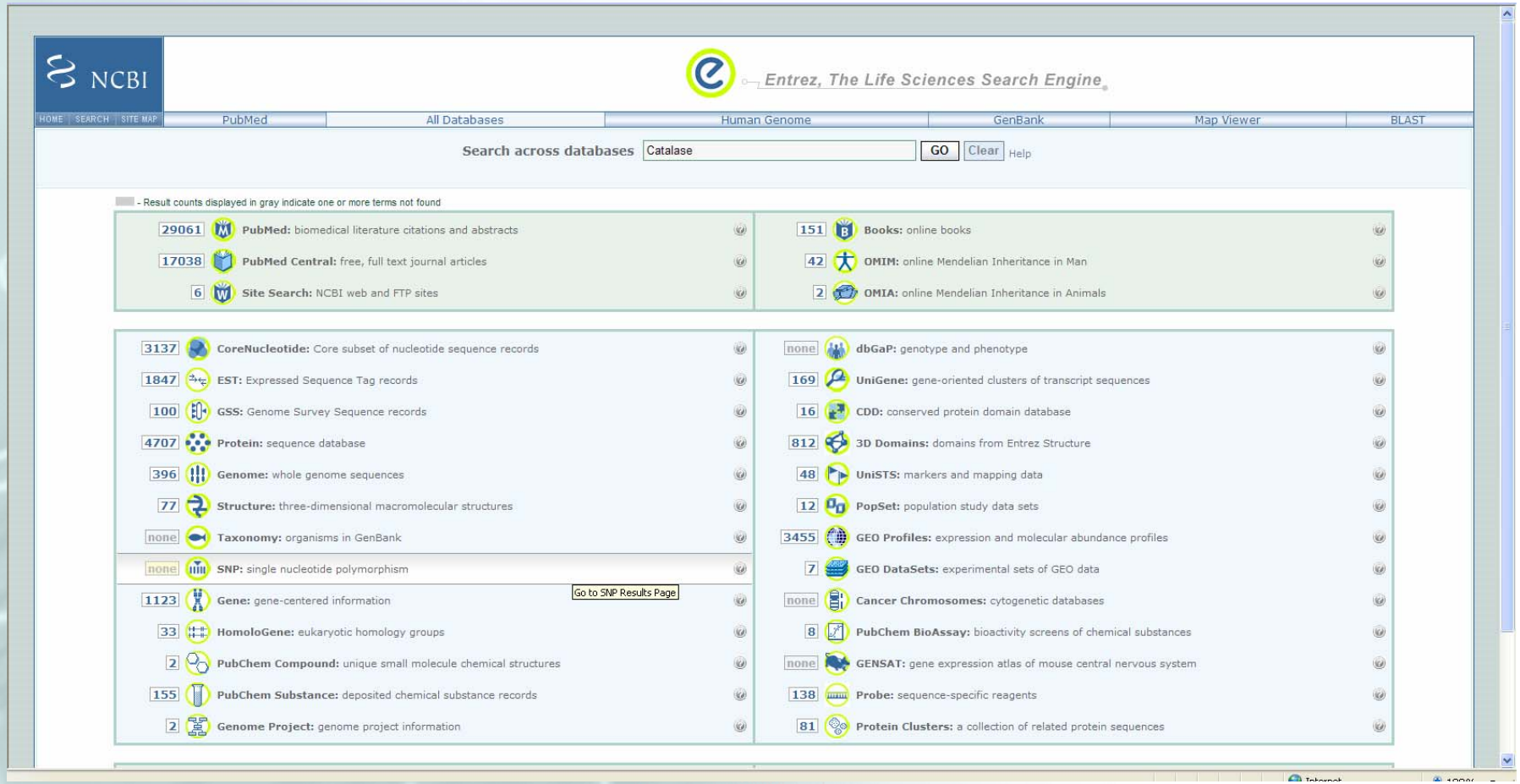


Section D1 –
Entrez – search across databases

Text search across databases

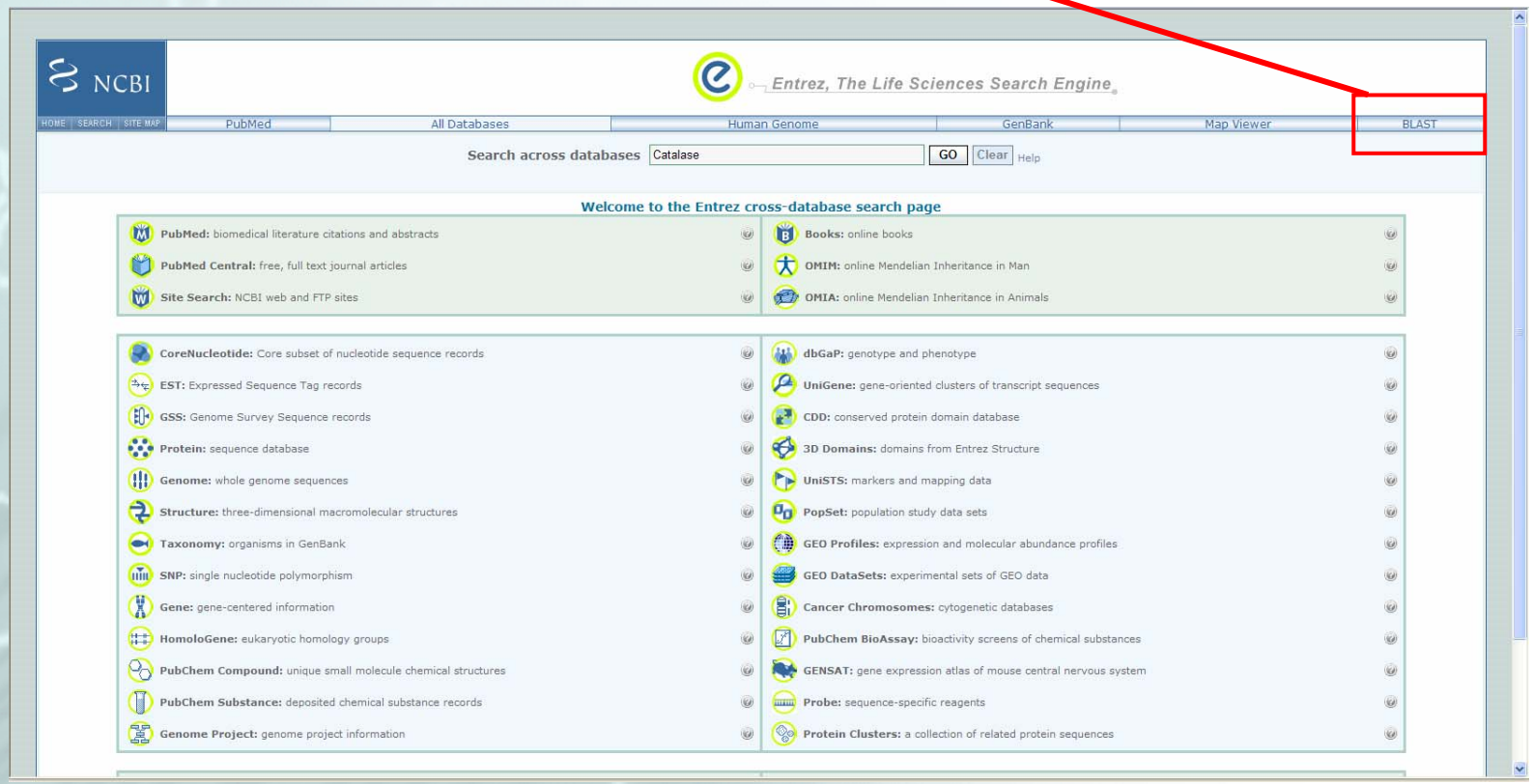


Section D1 –
Entrez – search across databases

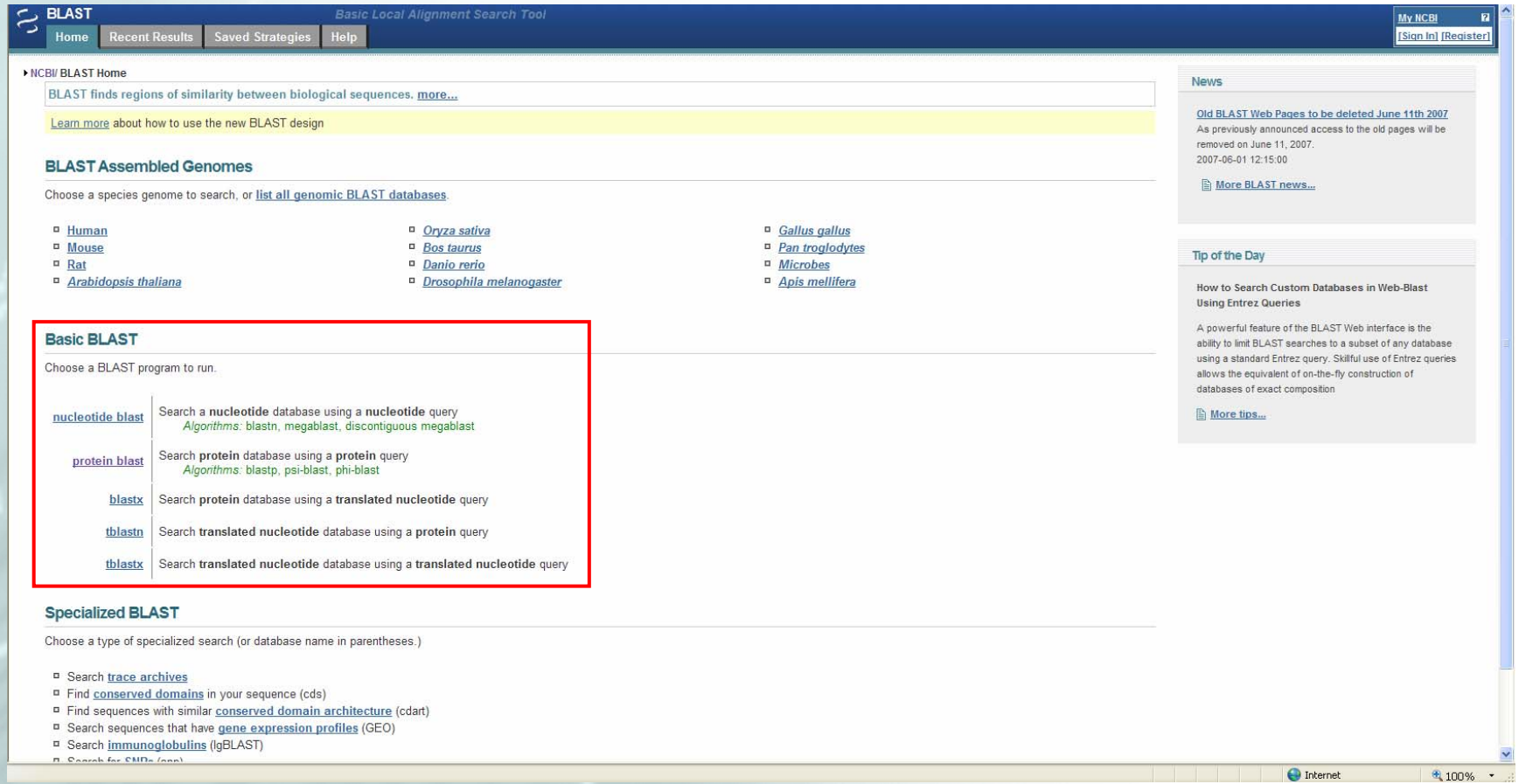


Section D1 –
Entrez – BLAST

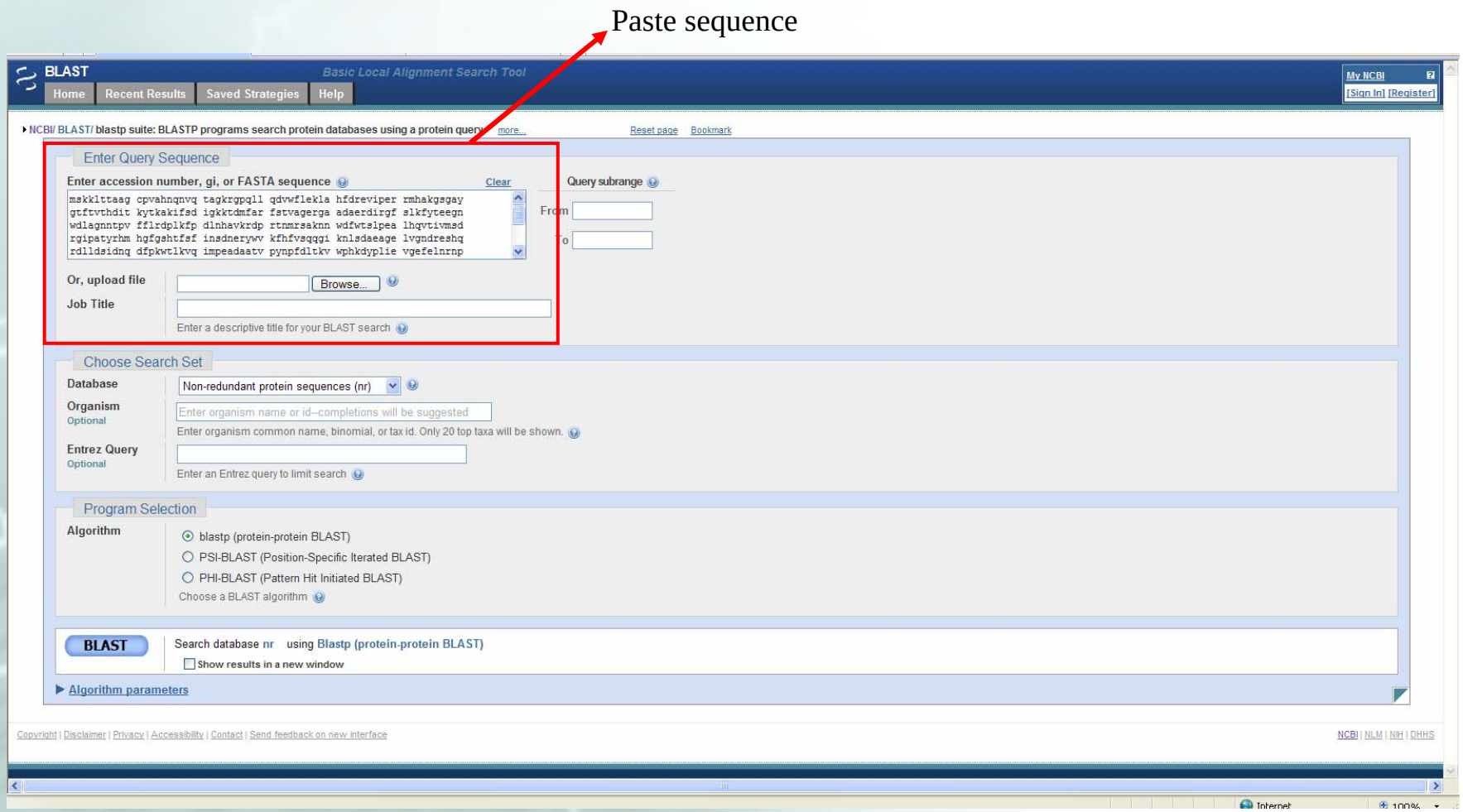
BLAST search



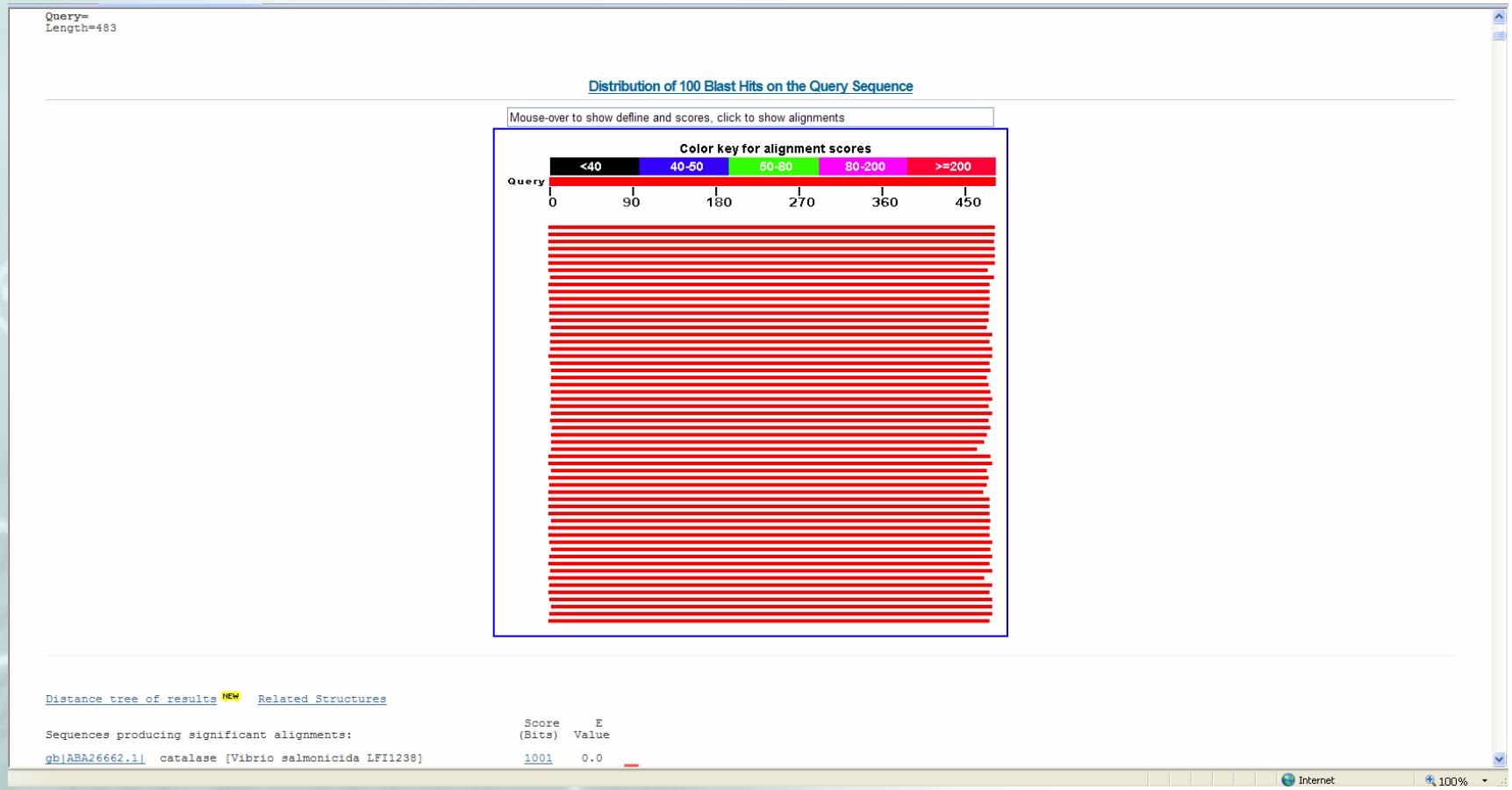
Section D1 –
Entrez – BLAST



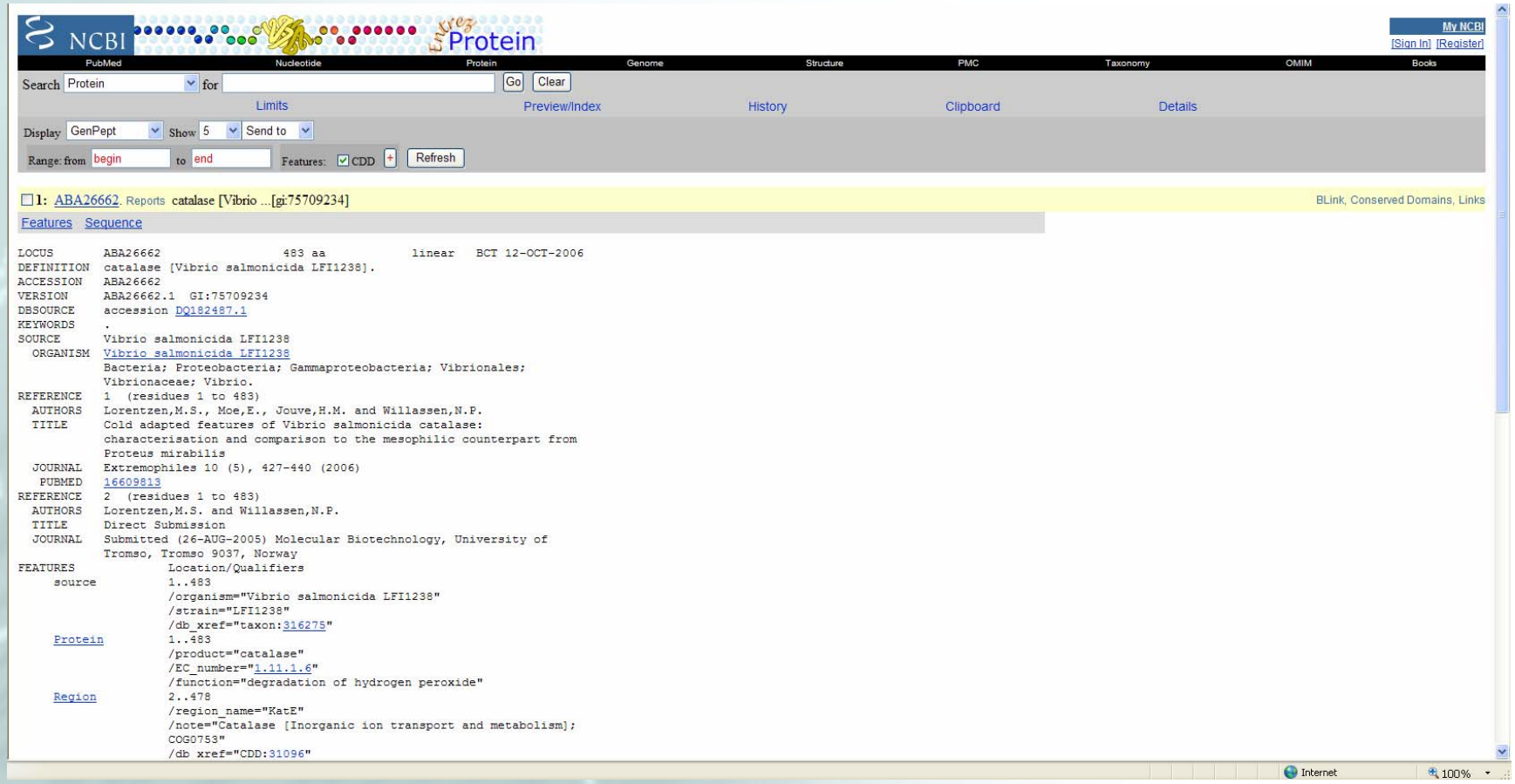
Section D1 –
Entrez – BLAST



Section D1 –
Entrez – BLAST



Section D1 –
Entrez – BLAST



Section D1 – DBGET/LinkDB

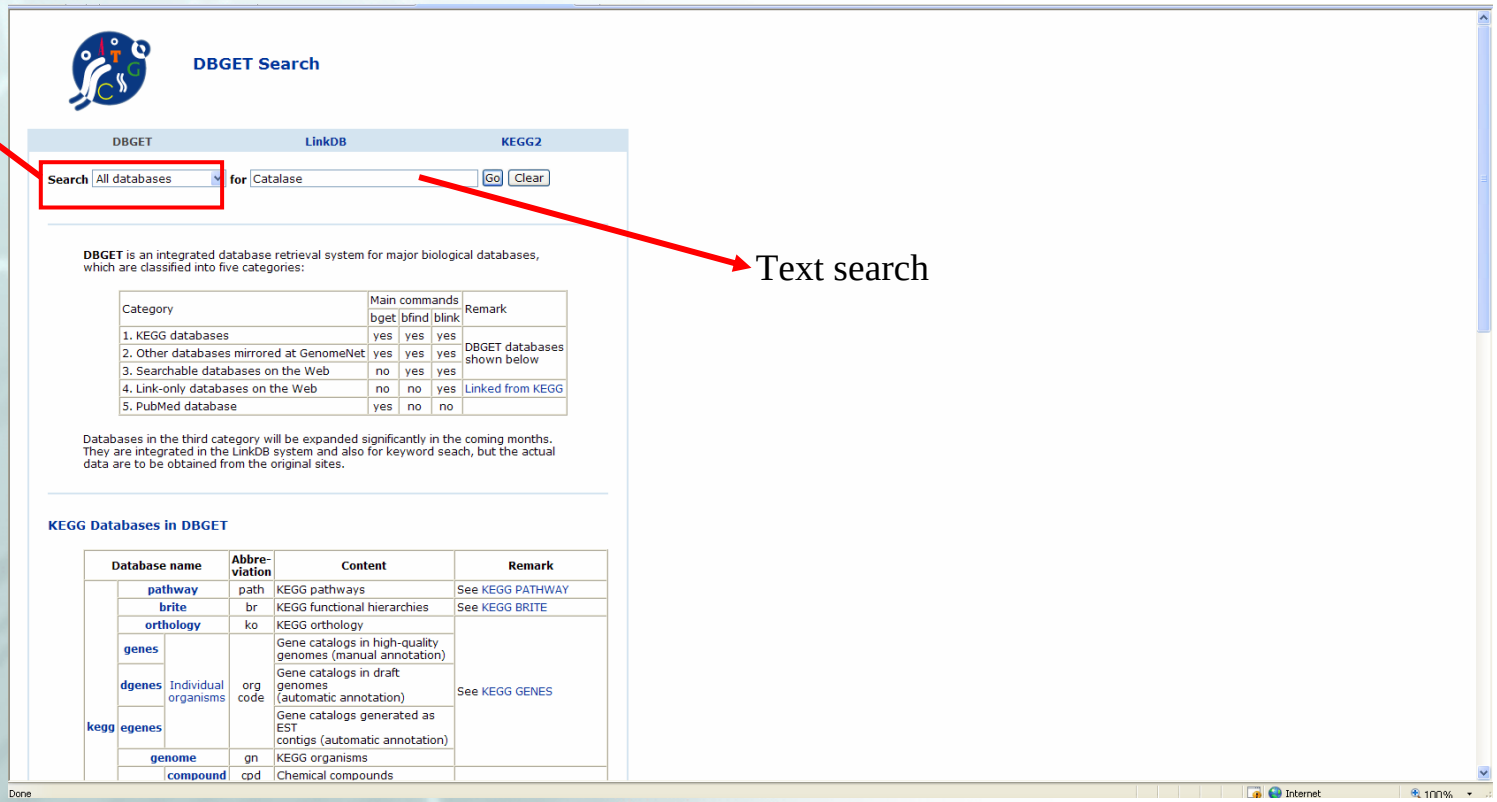
- DBGET is a data retrieval system developed and jointly maintained by the Institute for Chemical Research (Kyoto University) and the Human Genome Center (University of Tokyo)
  - Integrated with more than 20 databases
  - Search either by
    - Text searches
    - Search based accession number
  - Results presented as a list with all corresponding information
  - Closely associated with Kyoto Encyclopedia of Genes and Genomes (KEGG)
- LinkDB is an associated database of links between entries in the different databases available to DBGET, and further organism-specific databases
  - *C. elegans* Database (ACeDB)
  - Flybase
  - *Saccharomyces* Genome Database (SGD)

Section D1 – DBGET/LinkDB

Category	**Database**
Nucleic acid sequences	GenBank, EMBL
Protein sequences	SWISS-PROT, PIR, PRF, PDBSTR
3D structures	PDB
Sequence motifs	PROSITE, EPD, TRANSFAC
Enzyme reactions	LIGAND
Metabolic pathways	PATHWAY
Amino acid mutations	PMD
Amino acid indices	AAindex
Genetic diseases	OMIM
Literature	LITDB, Medline
Organism-specific gene catalogs	*E. coli, H. influenzae, M. genitalium, M. pneumoniae, M. jannaschii, Synechocystis, S. cerevisiae*

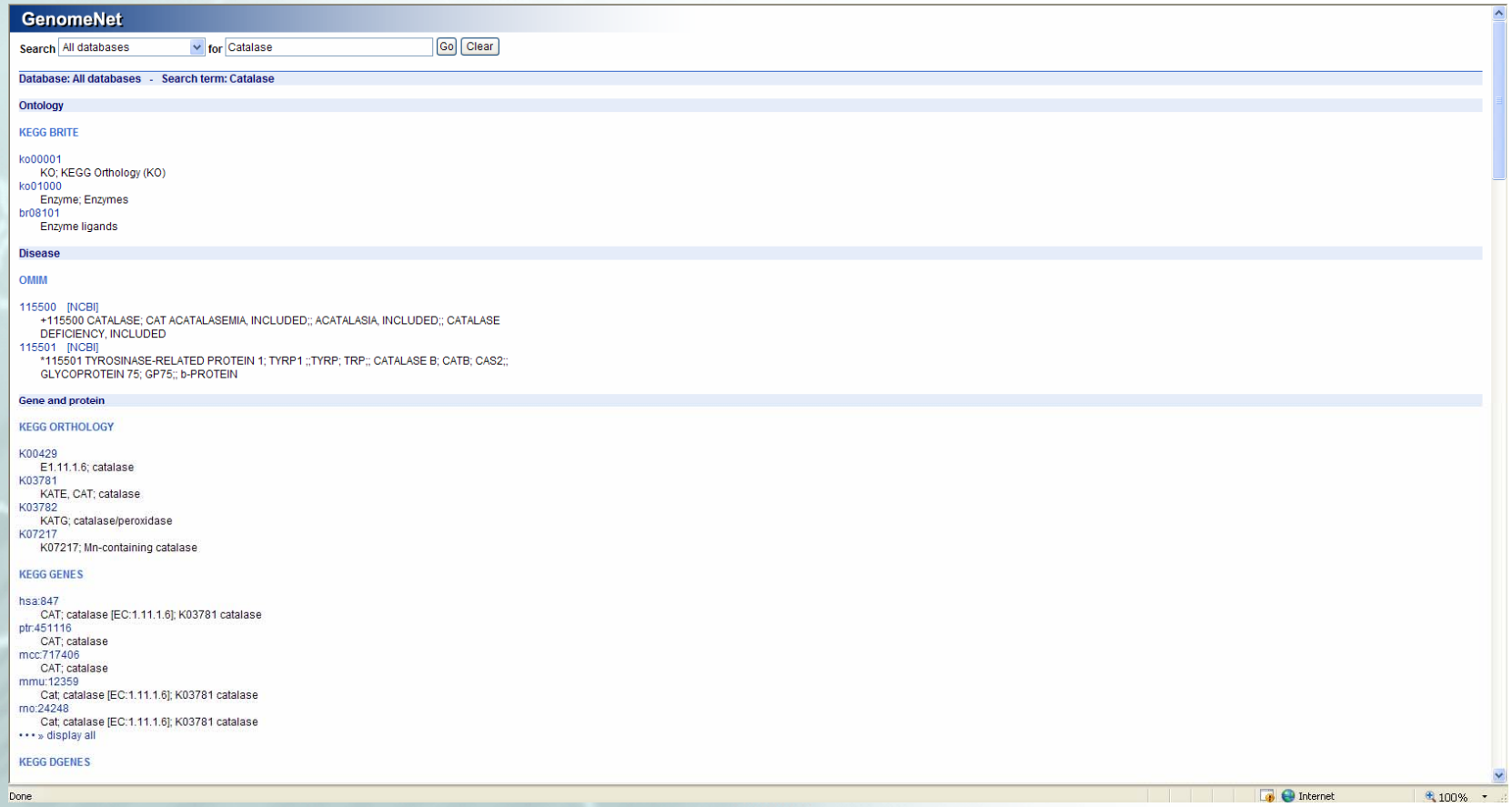
Section D1 – DBGET

Choose database

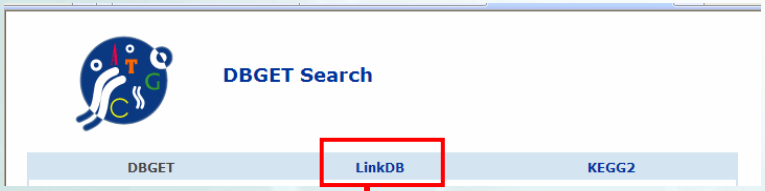


Text search

Section D1 – DBGET



Section D1 – LinkDB





Section D1 – LinkDB

LinkDB Search Result

Mode: Single Entry to Database
From: H.sapiens:847
To: All
Hits: 60 from 11+ databases

1. GenBank (21)
2. EMBL (21)
3. UniProt (1)
4. PROSITE (2)
5. OMIM (1)
6. Pfam (1)
7. ORTHOLOGY (1)
8. ENZYME (1)
9. PATHWAY (3)
10. BRITE (2)
11. GENOME (1)
12. OTHERS (5)
13. All databases (60)

DBGET integrated database retrieval system, GenomeNet

Section D2 – Data retrieval with SRS (Sequence retrieval system)

- SRS and Entrez/DBGET
- Using SRS
- Installing SRS

Section D2 –
SRS and Entrez/DBGET

- Sequence retrieval system (SRS) is a retrieval tool developed by the European Bioinformatics Institute (EBI)
- Integrates over 80 molecular biology databases
- Main difference between Entrez/DBGET and SRS is that SRS is an open source software that can be downloaded and installed locally
  - Availability not restricted by the activity of curator
- SRS are the standard for several large sites, including i.e. SWISS-PROT and ExPASy

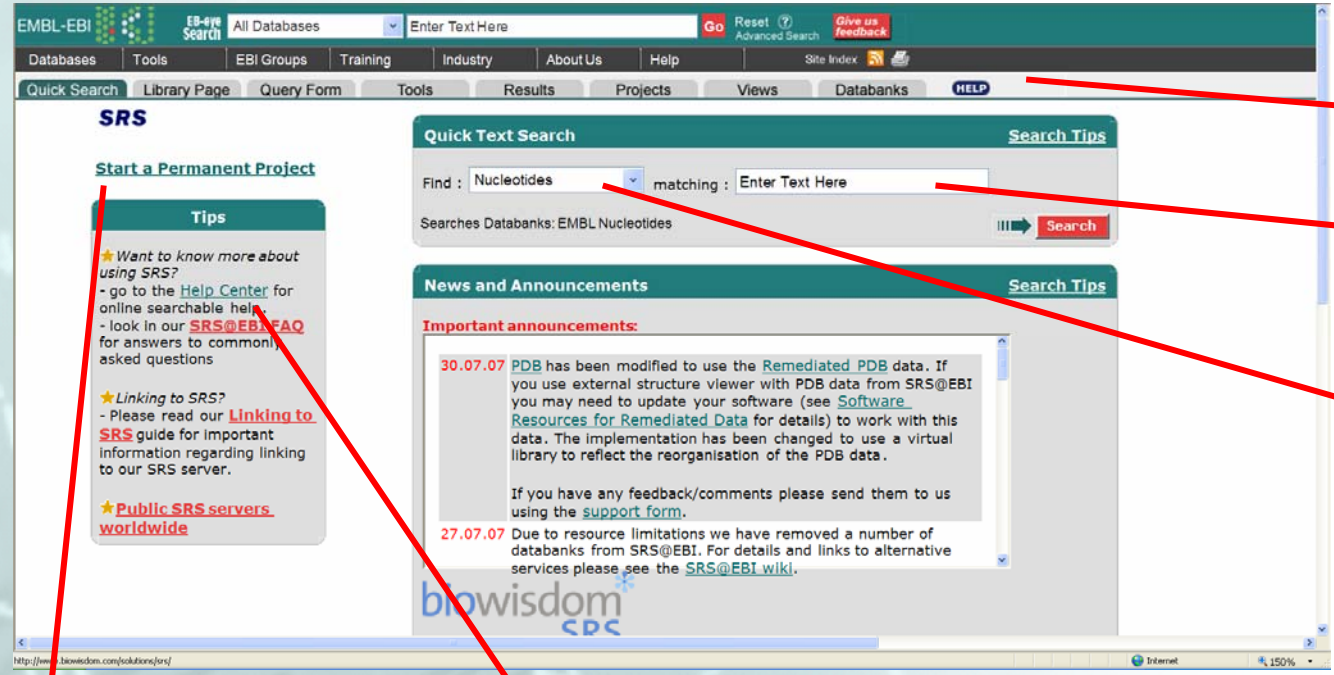
Section D2 – Using SRS

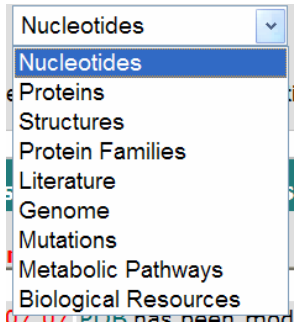
- SRS can either be run from their homepage or locally on a computer
- SRS classifications are, unlike Entrez, grouped by the type data and not type of molecule
- SRS *Sequence Libraries* classification covers all sequences (nucleotide and protein) whereas these data are separated in Entrez
  - InterPro&Related – Secondary databases of protein motifs
  - SeqRelated – Specialized sequence databases as UniGene
  - TransFac – Transcription factor databases
- To search in SRS, expand the relevant classifications and check boxes corresponding to the required databases
- After submitting your query, the results are displayed as a table with hyperlinks to files in the appropriate databases

Section D2 – Installing SRS

- SRS can either be run from their homepage or locally on a computer
- Work by constructing indexing files that have to be updated regularly
- Installation set comes includes configuration and modulation files for the most popular databases
  - SRS can be installed and operated with the minimum of effort
- Large indexing files and needs lot of processor time

Section D2 – SRS
Start page





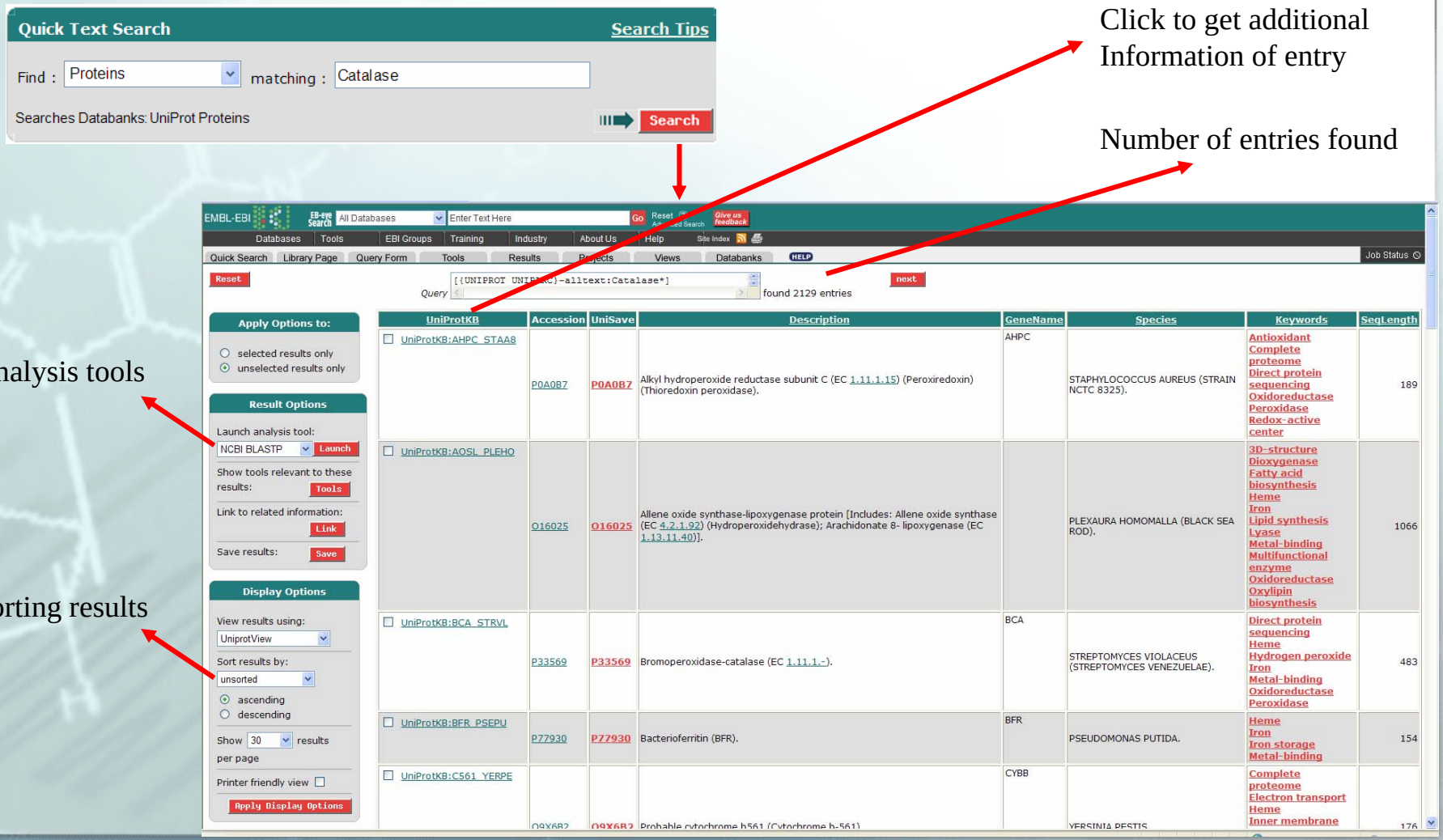
SRS toolbar

Query field

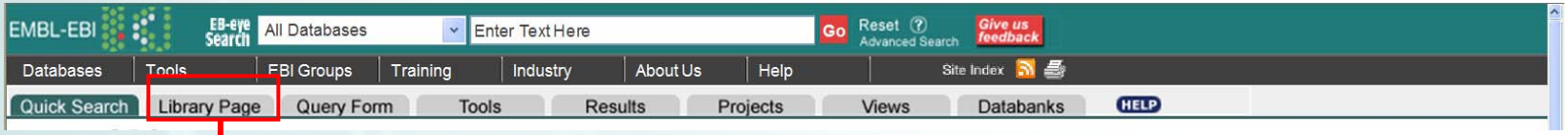
Start permanent project – Requires SRS user name

Help center – Tutorial and FAQ

Section D2 – SRS
Query using SRS quick search

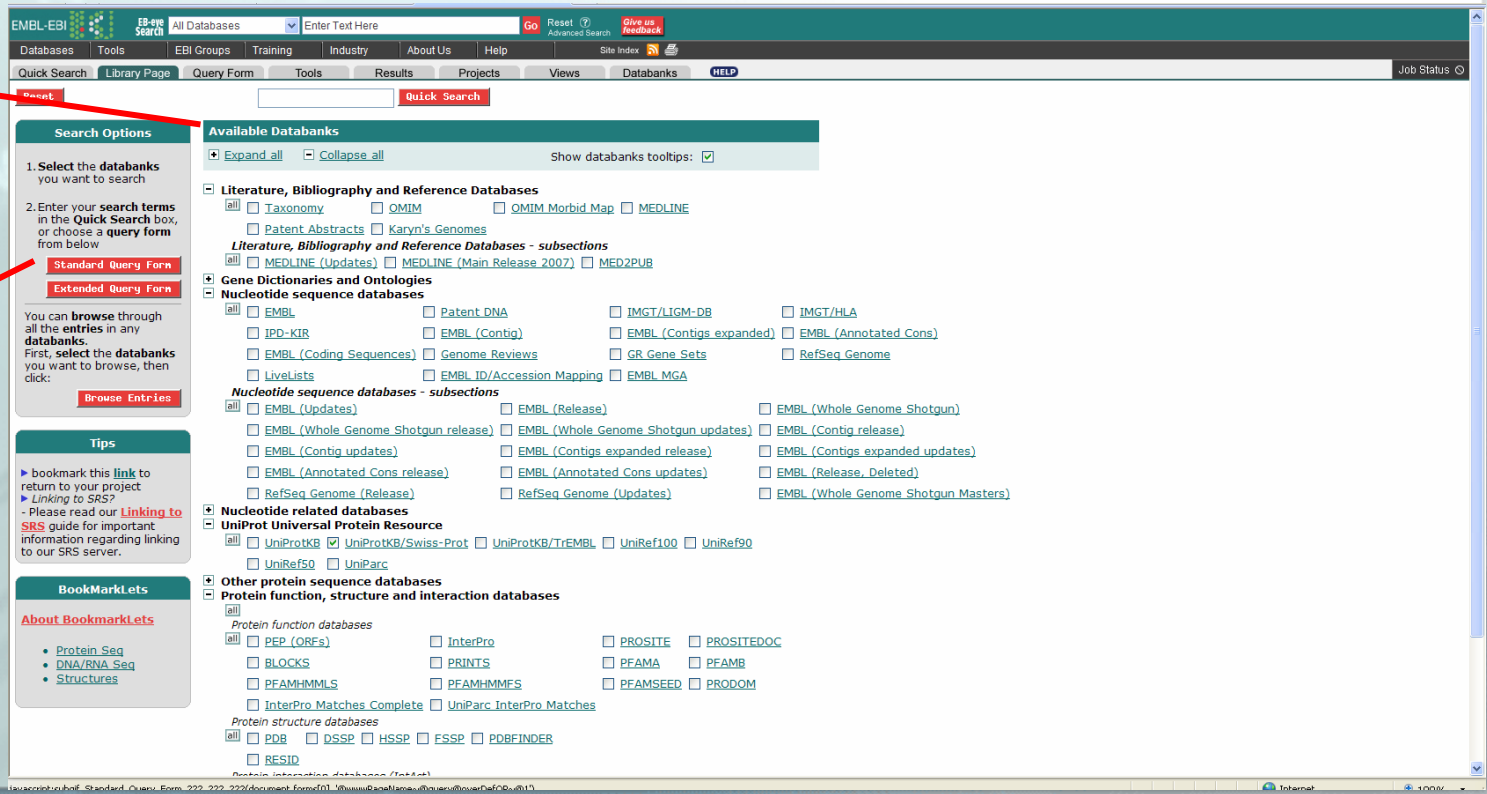


Section D2 – SRS
Query using SRS standard form



Available databanks

Standard query form



Section D2 – SRS
Query using SRS standard form

Search Options

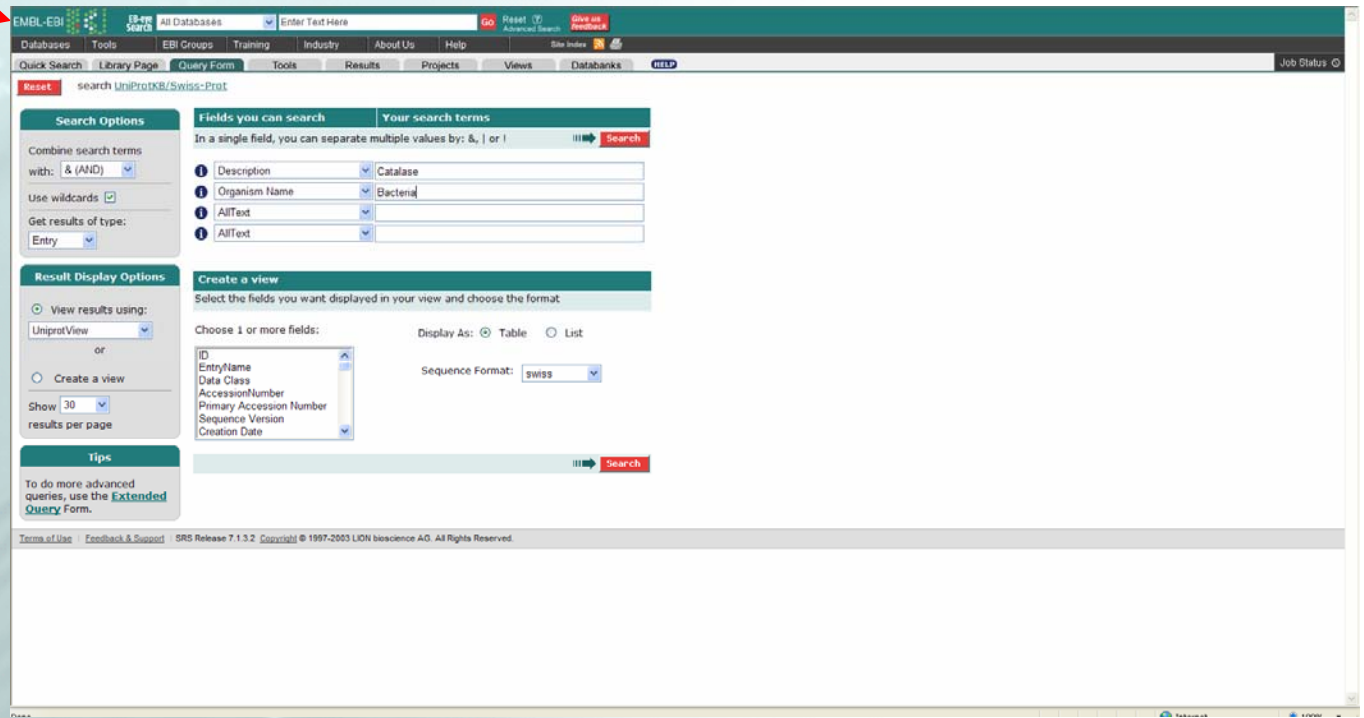
1. **Select** the **databanks** you want to search
2. Enter your **search terms** in the **Quick Search** box, or choose a **query form** from below

Standard Query Form

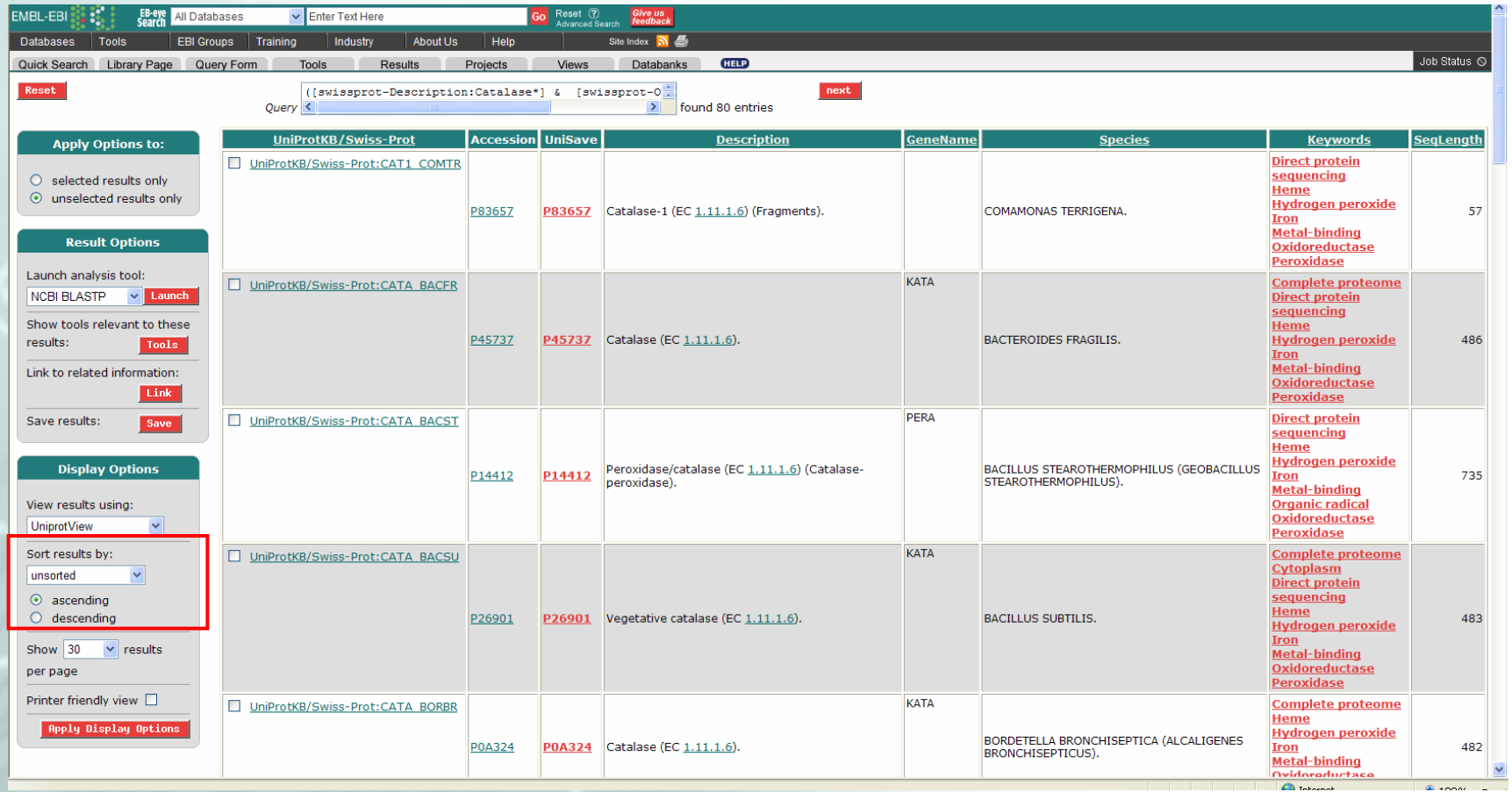
Extended Query Form

You can **browse** through all the **entries** in any **databanks**.
First, **select** the **databanks** you want to browse, then click:

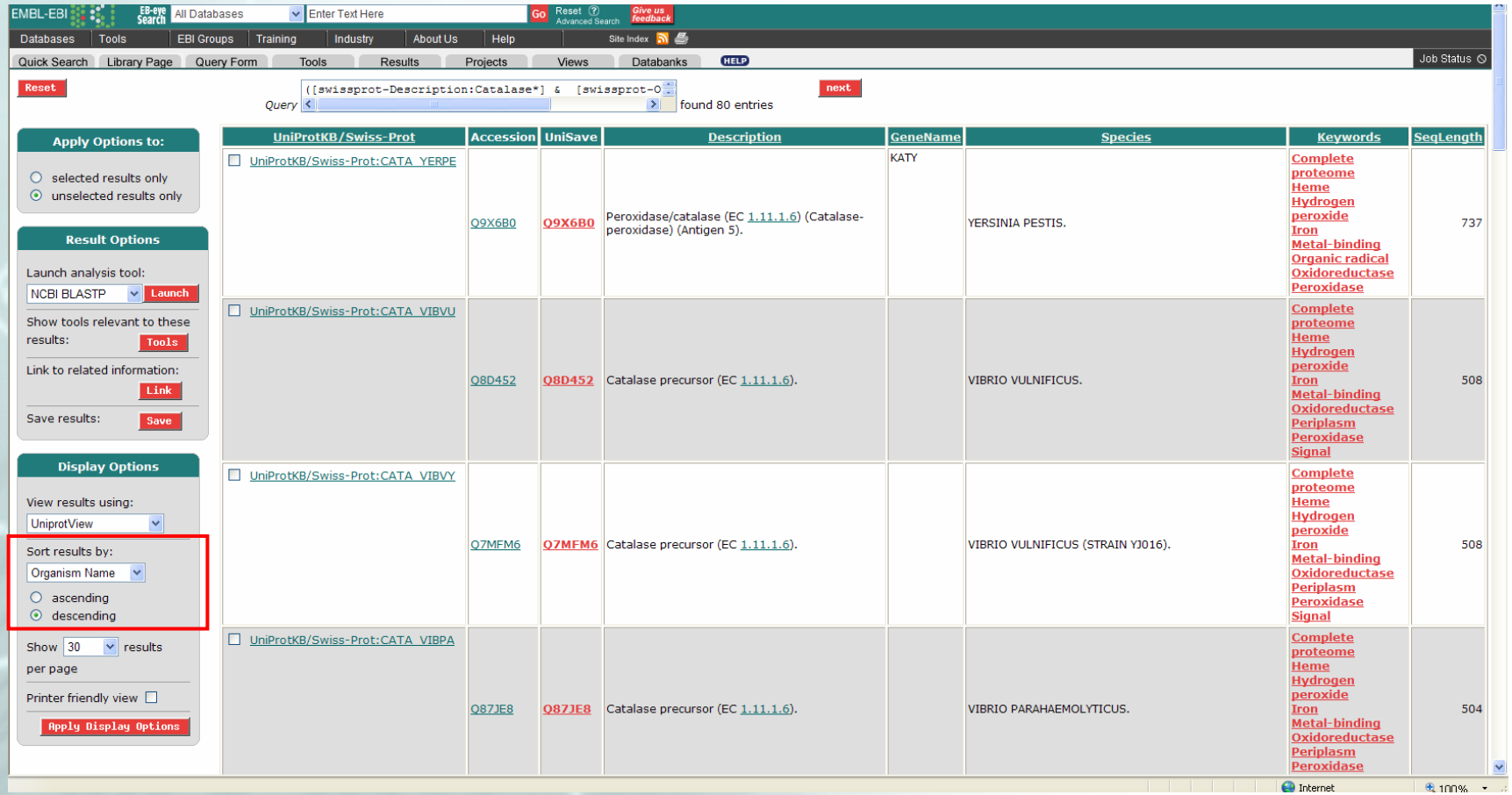
Browse Entries



Section D2 – SRS
Query using SRS standard form

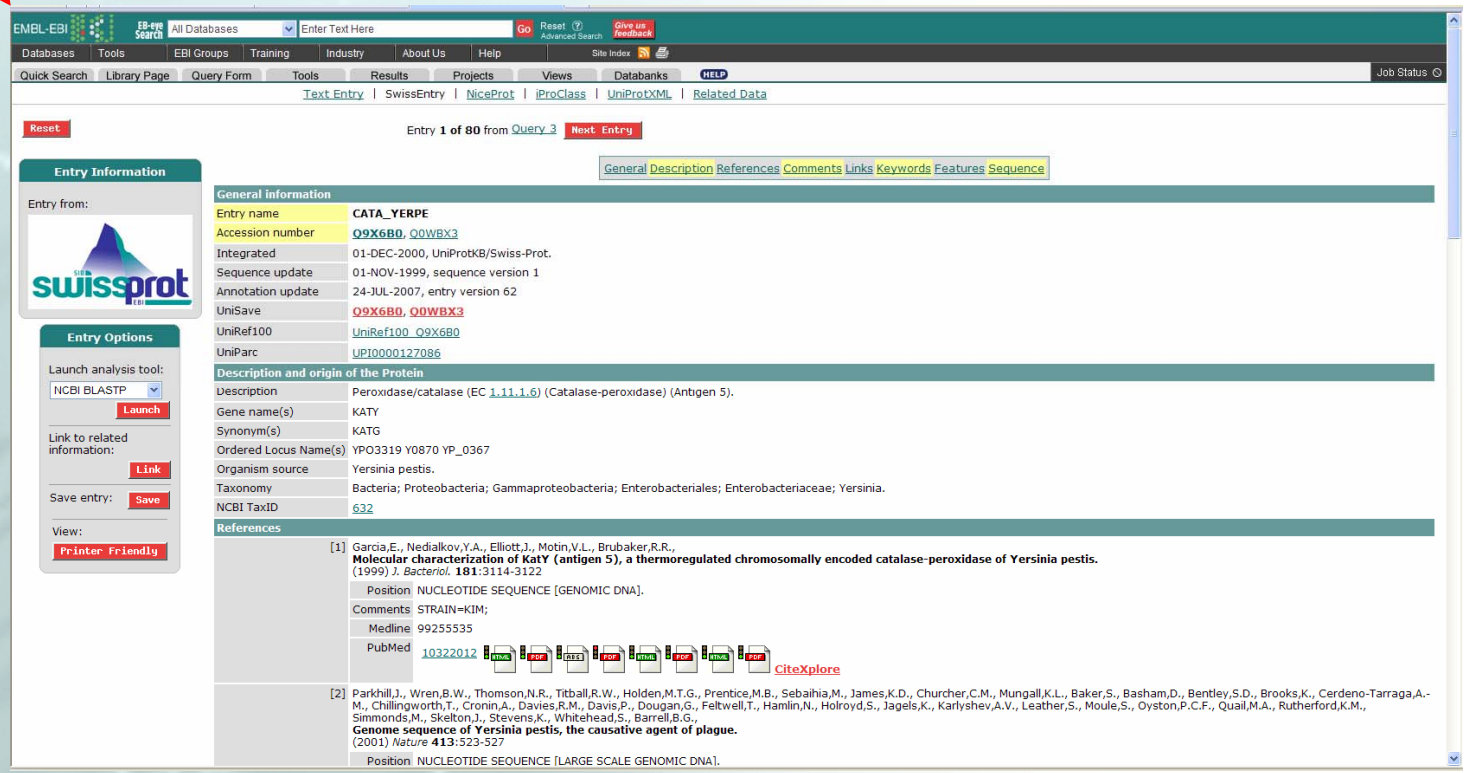


Section D2 – SRS
Query using SRS standard form

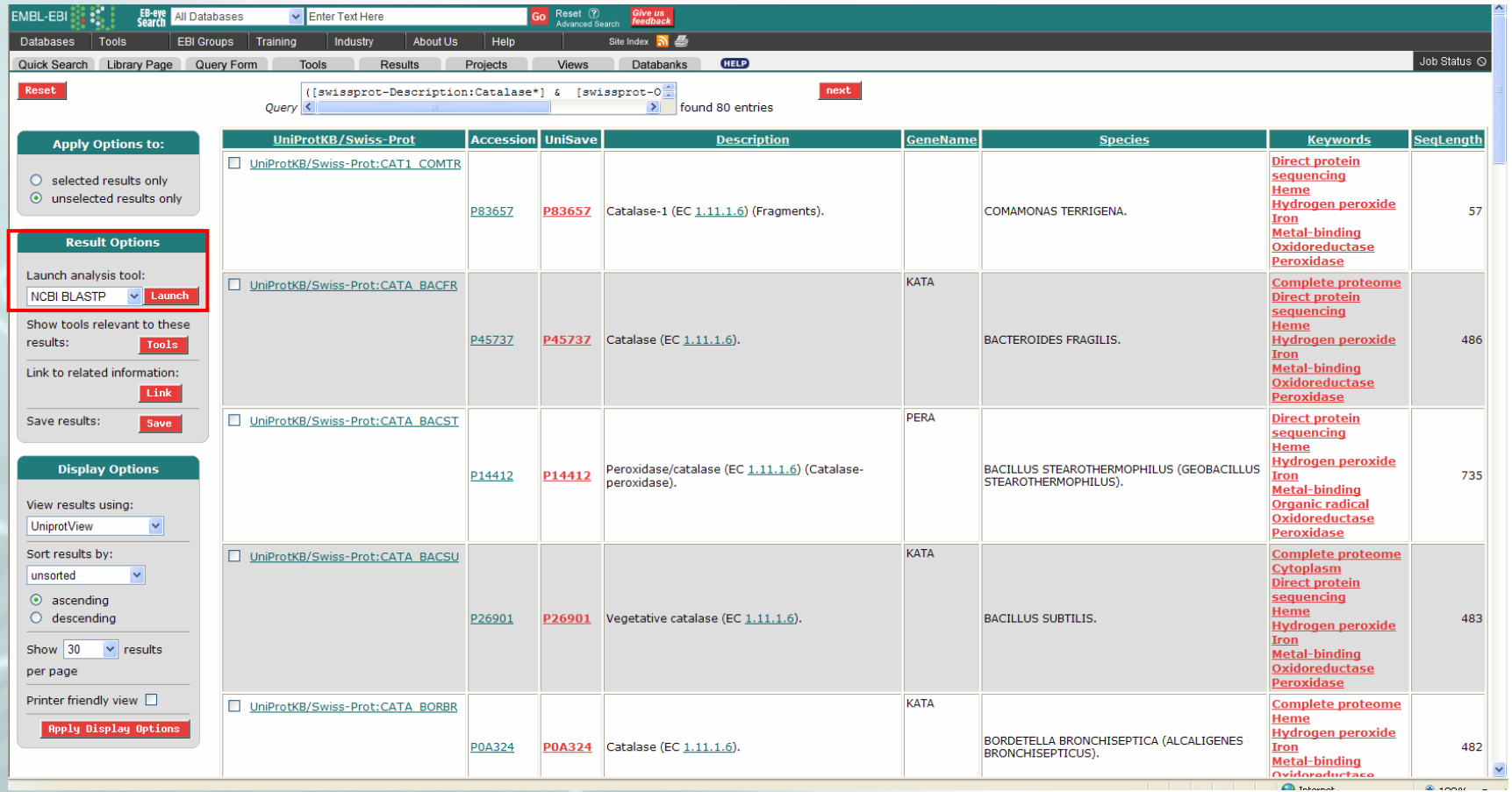


Section D2 – SRS
Query using SRS standard form

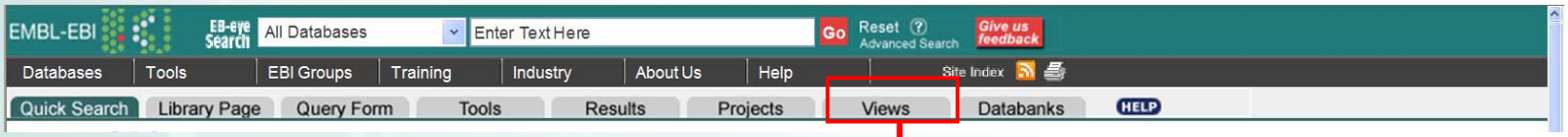
UniProtKB/Swiss-Prot	Accession	UniSave	Description	GeneName	Species	Keywords	SeqLength
☐ UniProtKB/Swiss-Prot:CATA_YERPE	Q9X6B0	Q9X6B0	Peroxidase/catalase (EC 1.11.1.6) (Catalase-peroxidase) (Antigen 5).	KATY	YERSINIA PESTIS.	Complete proteome Heme Hydrogen peroxide Iron Metal-binding Organic radical Oxidoreductase Peroxidase	737



Section D2 – SRS
Launching tools

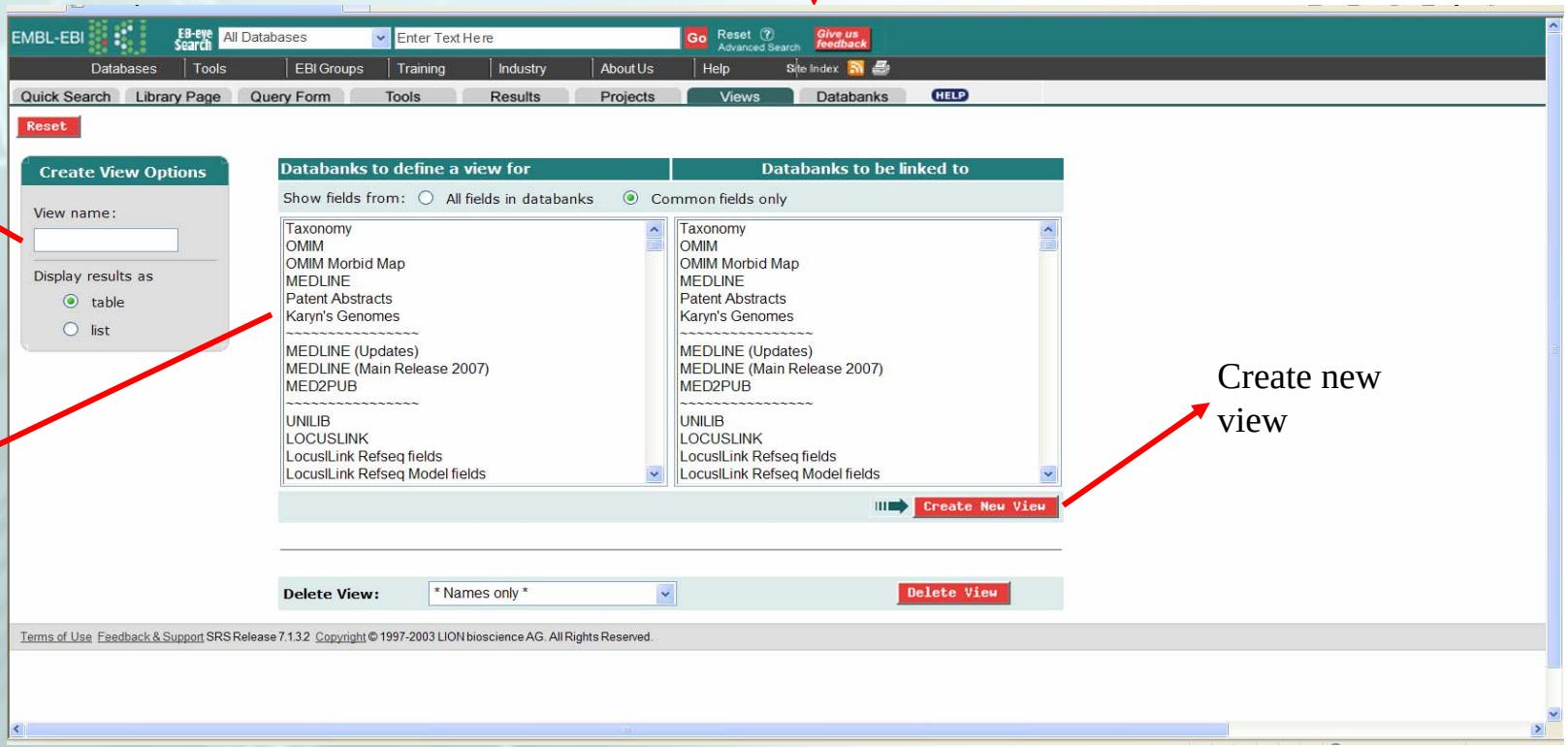


Section D2 – SRS
Creating user-defined views



Give your view a name

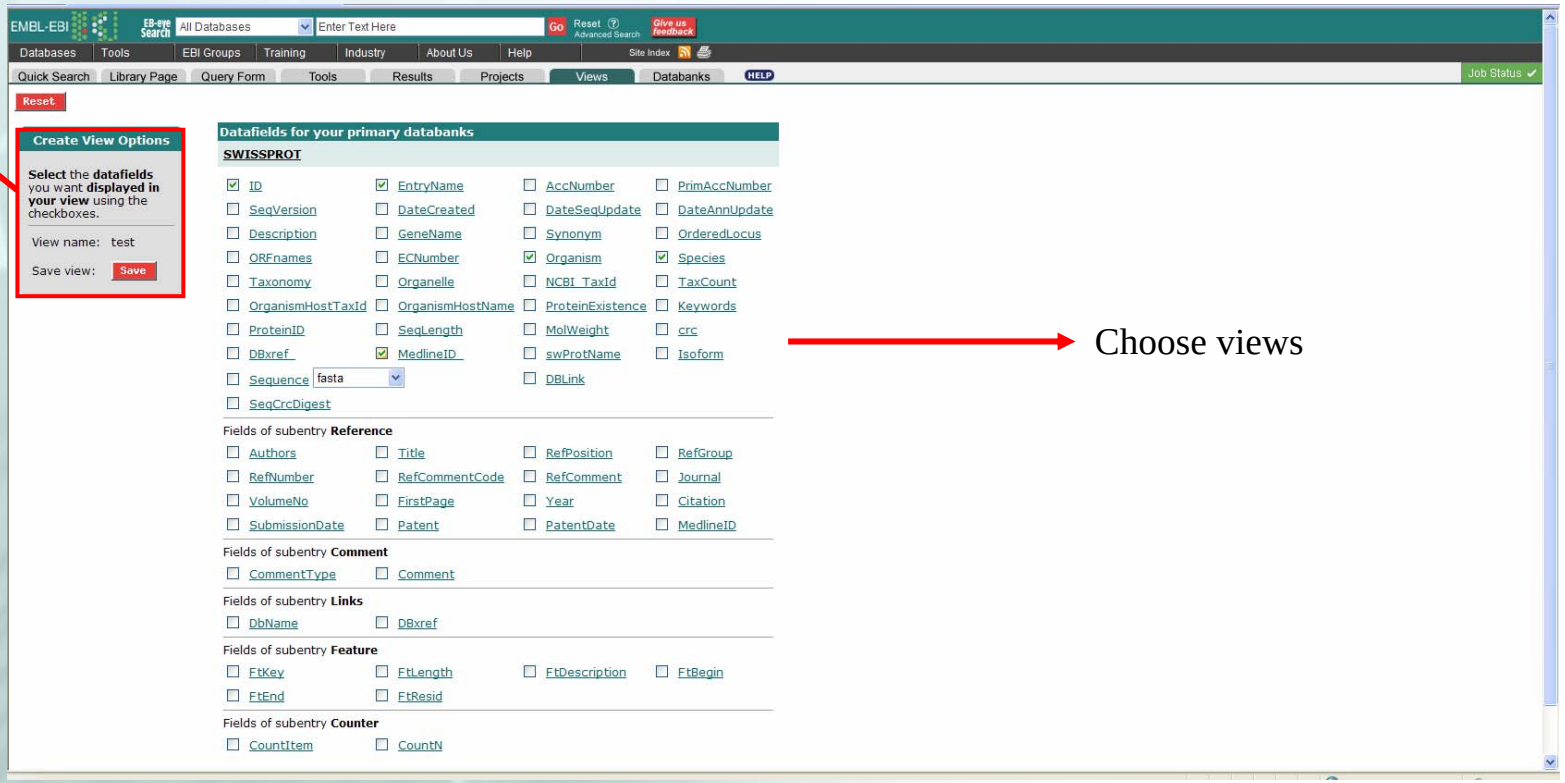
Choose databank to create view from



Create new view

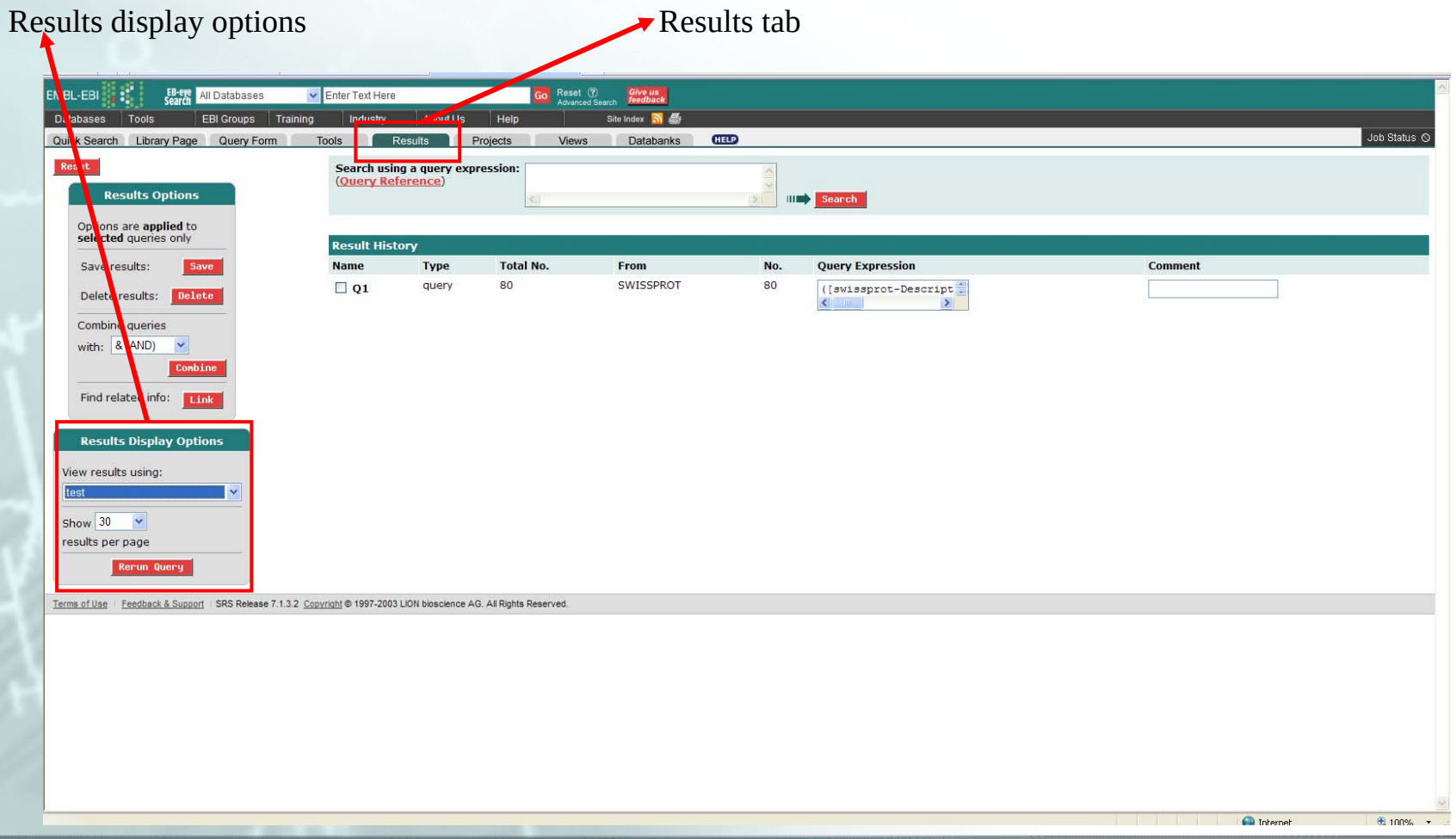
Section D2 – SRS
Creating user-defined views

Save your view



Choose views

Section D2 – SRS
Creating user-defined views



Section D2 – SRS
Saving your project

