Introduction to Molecular Biology Databases

Manuel J. Gómez
Laboratorio de Bioinformática
Centro de Astrobiología
INTA-CSIC

PRESENT BIOLOGY RESEARCH

Data sources

- Genome sequencing projects: genome sequences.
- Functional genomics: expression data.
- Protemics: protein catalogs, expression patterns.
- Structural genomics: protein structures.

- Functional interactions between cellular components: regulation networks.
- Physical interactions between proteins: Protein interaction networks.

- Accumulated experimental data: non structured information (or even not in electronic format) in the form of publications.

How much information is there?

- Nucleotide records
  - 9,102,634
- Nucleotides
  - 10,335,692,655
- Protein sequences
  - 1,183,833
- 3D structures
  - 12,863
- Expression data points
  - >20,000,000
- Human Unigene clusters
  - 84,130
- Maps and complete genomes
  - 11,166
- Different taxonomy nodes
  - 162,025
- dbSNP
  - 1,463,178
- Human Refgene records
  - 14,133
- Human contigs >500 kb (28,525 MB)
  - 257
- PubMed records
  - 10,965,353
- OMIM records
  - 11,950

Ouellette, 2000

Environmental genome shotgun sequencing of the Sargasso Sea.

Venter *et al.*
Science March 2004.

- **1.045 billion nucleotides**
- **1.2 million new genes**
- **782 new rhodopsin like photoreceptors**
- **1800 genomic species**

CLASSIFICATION OF MOLECULAR BIOLOGY Dbs

First classification scheme:

Based on the SOURCE of the CORE DATA, also referred to as DATABASE CONTENT.

- Primary DBs: the content consists in experimentally obtained data.
- Secondary DBs: the content is the result of analyses of data in Primary DBs.

THE CONTENT OF PRIMARY DATABASES

Experimentally derived information:

- Nucleic acid sequences: complete genomes, cloned genome fragments, cDNAs, ESTs, SNPs.
- Protein or Nucleic acid structures: atomic coordinates obtained by RMN, X-Ray crystallography.
- Transcript or Protein expression data: obtained in micro-array experiments or by proteomic approaches, respectively.
- Cellular processes, such as experimentally determined metabolic or regulatory pathways.

THE CONTENT OF SECONDARY DATABASES

Predictions or interpretations based on information contained in primary databases.

- Protein sequences, deduced from nucleotide sequences.
- Alignments of protein or nucleic acid sequences.
- Protein families, inferred by sequence similarity or by the presence of common Motifs or Domains.
- Protein families, inferred by structural similarity.
- Reconstructed (predicted) cellular processes, such as metabolic pathways.

CLASSIFICATION OF MOLECULAR BIOLOGY Dbs

Second classification scheme:

Follows the well known layout to describe the levels of organization of protein structures.

- Primary DBs: the content consists of SEQUENCES (primary structure of nucleic acids or proteins).
- Secondary DBs: the content consists of PATTERNS (regions of local regularity, for example, conserved motifs or domains).
- Structure DBs: the content consists of sets of ATOMIC COORDINATES (three-dimensional packing of secondary structure elements).

This scheme applies only to biological molecules (for example, a database of results from micro-array expression experiments would not fit).

DATABASE SEARCHES: CONTENT AND ANNOTATIONS

In any Primary Db (but also in Secondary Dbs) the content information is complemented with additional information, that is organized in what are known as ANNOTATIONS.

Annotations can refer to:

- Authorship of the entry.
- Experimental conditions.
- Source of the biological material.
- Subcellular location.
- Molecular function or cellular process.
- Bibliographic references.
- Cross-references: entry attributes that make reference to related entries in other databases.

Some annotations consist in information of Secondary type, in the sense that they are predictions, and some of them have been transferred from other databases.

DATABASE SEARCHES: QUERY TYPES

TEXT QUERIES

These type of fixed-form queries are searches performed against the ANNOTATIONS, which, almost by definition, consist of texts.

It is usual to allow the combination of words with Boolean operators (and, or, not) the use of wild cards (*), and also, to specify the search field, for example:

ponB and ayala* [AUTH]

would result in a search with the word *ponB*, in any field, combined with a search of the word *ayala** in the author field.

DATABASE SEARCHES: QUERY TYPES

QUERY BY CONTENT

This type of fixed-form query refers to searches performed against the CONTENT or CORE DATA, which, almost by definition, consists in abstract representations:

- Strings of characters that represent nucleotide or protein sequences.
- Tables of atomic coordinates that represent three dimensional objects
- Bitmap files that represent 2D gel images.

DATABASE SEARCHES: QUERY TYPES

QUERY BY CONTENT

For example, in the case of a sequence database, we may be asking:

Does a sequence exactly like:

LLLIHRLH,

or similar to it, exists in the database?

Because biological sequences change along time, or between individuals, this type of search is not a matter of finding exact matches in strings of characters. On the contrary, it must consider the principles of molecular evolution.

A number of algorithms have been developed to cope with that task, and BLAST is the most popular.

DATABASE SEARCHES: QUERY TYPES

QUERY BY CONTENT

Other types of "query by content", that make use specific algorithms to pose searches against the abstract representations in core data could be:

- Query by image content (QBIC), is the one used to search in image databases (check http://www.qbic.almaden.ibm.com/ as a source of information on QBIC).
- VAST, Vectorial alignment search tool, structural comparisons (NCBI).
- DALI, structural comparisons (EBI).

A POSSIBLE SELECTION OF ESSENTIAL MOLECULAR BIOLOGY DBs

GENES AND GENOMES

- GenBank and EMBL-nucleotide. Nucleotide databases maintained by the NCBI and the EMBL, respectively.
- NCBI Genomes. Complete genome sequences.
- Ensembl. Complete eukaryotic genomes, maintained by the Sanger Center.

PROTEINS

- SwissProt and TrEMBL. Protein sequence databases, maintained by the SIB and the EBI, in a server called ExPASy.
- Prosite. Protein Motifs and families database, at ExPASy.
- Pfam. Protein Domains and families, defined by HMM profiles.
- InterPro. Protein Motifs, Domains and Families, which integrates information from several other databases.

STRUCTURES

- PDB. Protein Data Bank, molecular structures database.
- SCOP. Hierarchical classification of proteins, based on structure comparisons

METABOLIC PATHWAYS

- KEGG. Kyoto Encyclopedia of Genes and Genomes.

BIBLIOGRAPHY

- PubMed. Biomedical literature, at the NCBI.

ONTOLOGIES

- Gene Ontology Consortium. Ontology of Molecular Biology terms.

MOLECULAR BIOLOGY Dbs: how many and where?

Most Molecular Biology Dbs are accessible through Internet. There are resources where extensive compilations of Molecular Biology databases are published.

For example:

- The Nucleic Acid Research journal publishes, every year, a Molecular Biology Db catalog, which is maintained also in electronic form. The list contained 339 databases in the year 2002. It contained 386 in 2003, and 548, in 2004.

- Deambulum Dabank Index is another database catalog, maintained by Infobiogen, a french institution focused in compiling and distributing information about molecular biology and biomedicine.

- The Weizmann Institute (Israel) also maintains a list of databases, organized by areas.

TOO MANY DATABASES?

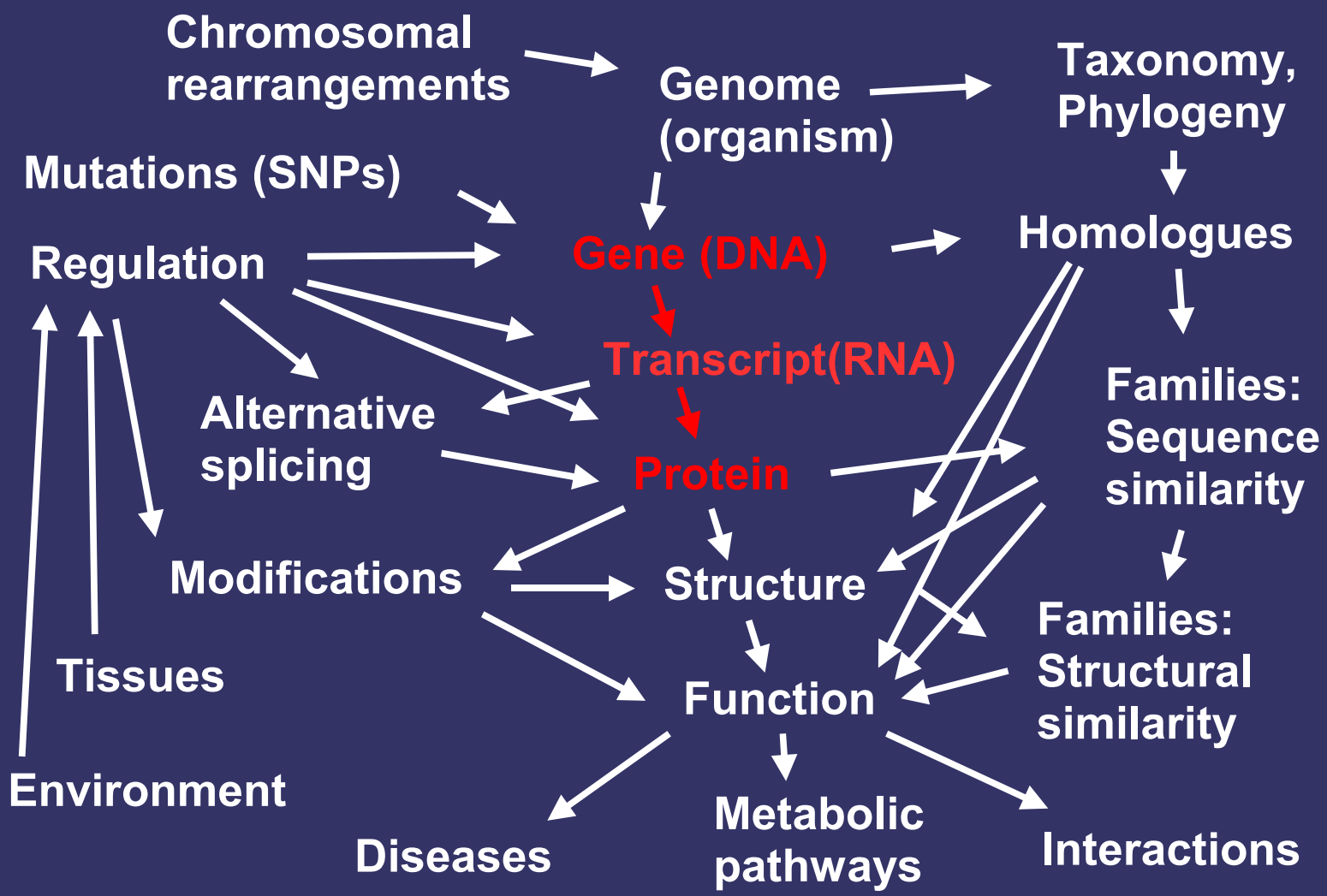
- Database diversity and specialization is unavoidable for operational (most databases originate from specific research projects) and sociological (everybody wants to make their own database) reasons.

- On the good side, the development of many specialized databases, may facilitate the task of finding information since search spaces are smaller.

- On the other hand, given the huge amount of information that has been, and is being generated by modern biological research, it is essential to devise systems that integrate the information of many databases.

- In order to do so, it is necessary to use standardized data formats and controlled vocabularies.

ONTOLOGIES IN MOLECULAR BIOLOGY



ONTOLOGIES IN MOLECULAR BIOLOGY

Controlled and structured vocabularies, constructed with TWO purposes: proposing standard collections of terms, and, organizing the knowledge of a given field around its language: the relations between the terms are supposed to reflect the biological reality.

- Enzyme Commission Nomenclature.

  EC 1. -. -.- Oxidoreductases.
  EC 1. 1. -.- Acting on the CH-OH group of donors.
  EC 1. 1. 1.- With NAD(+) or NADP(+) as acceptor.
  EC 1. 1. 2.- With a cytochrome as acceptor.

- MeSH (Medical Subject Headings) terms: NLM controlled vocabulary.

- Gene Ontology: developed and maintained by a consortium (GO Consortium) of laboratories and institutions involved in molecular biology database management.

Gene Ontology

Biological terms have been groupd in three ontologies:

- Molecular functions
- Biological processes
- Cellular components

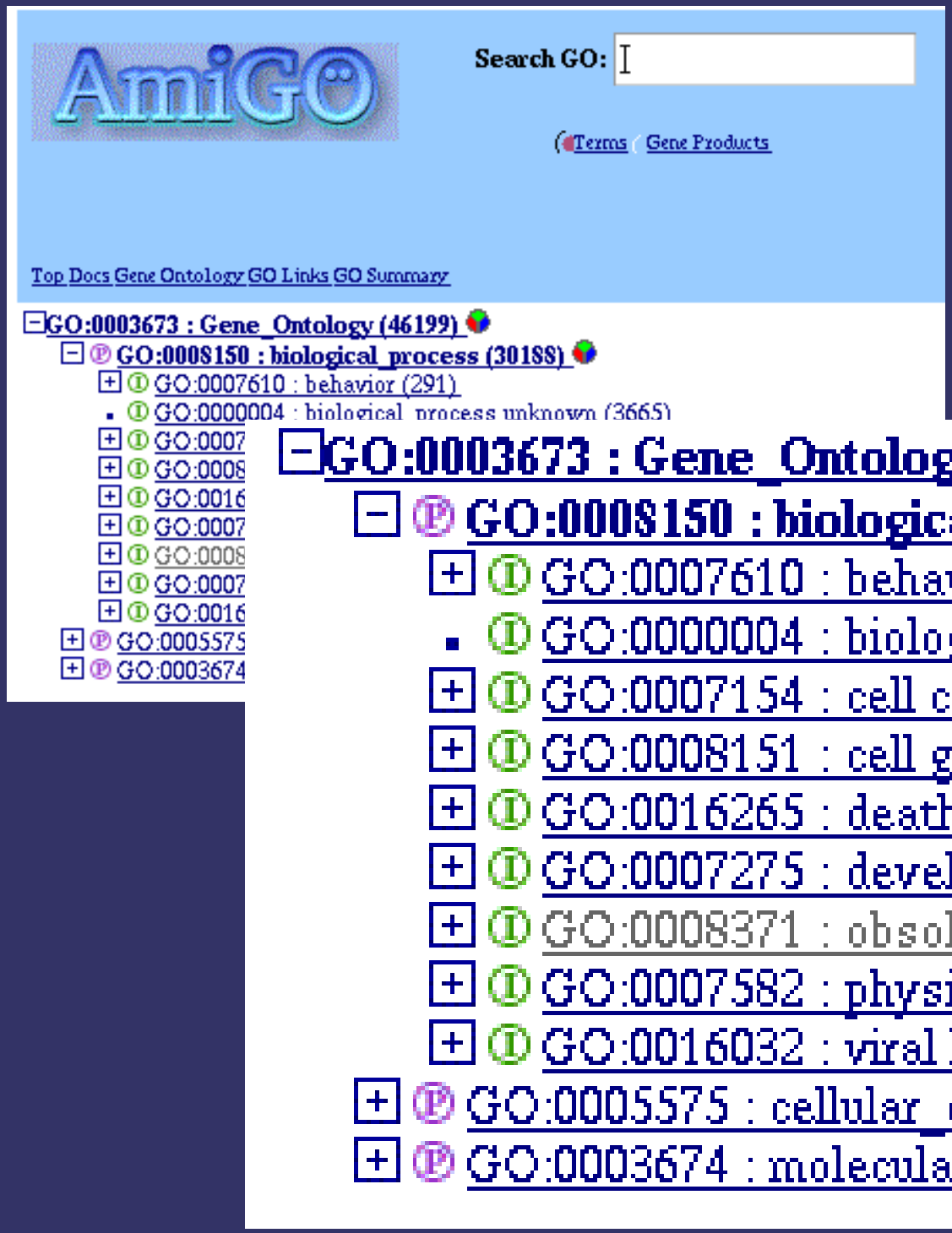
Within each ontology, the terms are related hierarchically.

The relation between parent and child terms can be of two types:

- Part of
- Instance of

Most molecular biology databases have joined this initiative, and have included annotations following this standard.

Gene Ontology BROWSER



Gene Ontology

Mycoplasma pneumoniae proteome.
Distribution of GO anotations.

GO Classification for *M. pneumoniae*			
Term		Proteins	
GO:0003674	**molecular_function**	**431**	**62.7%**
GO:0003676	nucleic acid binding	119	17.3%
GO:0030528	transcription regulator activity	5	0.7%
GO:0003754	chaperone activity	10	1.4%
GO:0003824	catalytic activity	245	35.6%
GO:0015070	toxin activity	1	0.1%
GO:0005194	cell adhesion molecule activity	1	0.1%
GO:0005198	structural molecule activity	56	8.1%
GO:0005215	transporter activity	64	9.3%
GO:0005488	binding	211	30.7%
GO:0005554	molecular_function unknown	65	9.4%
GO:0008150	**biological_process**	**343**	**49.9%**
GO:0008152	metabolism	280	40.7%
GO:0006810	transport	65	9.4%
GO:0006950	response to stress	16	2.3%
GO:0007049	cell cycle	22	3.2%
GO:0007154	cell communication	16	2.3%
GO:0007275	development	5	0.7%
GO:0007582	physiological process	335	48.7%
GO:0005575	**cellular_component**	**277**	**40.3%**
GO:0005576	extracellular	1	0.1%
GO:0005623	cell	274	39.8%
GO:0005941	unlocalized	5	0.7%

MOLECULAR BIOLOGY Dbs

TOPICS TO COME

- GenBank and EMBL nucleotide databases.
- ENTREZ
- EBI ans SIB databases
- SwissProt and TrEMBL.
- SRS: Systems that provide centralized access to multiple databases using a common query form.
- Exercises

GenBank AND EMBL

- GenBank and the EMBL nucleotide databases were founded in 1986.
- Contain ALL DNA sequences ever published .
- Submissions are made by the laboratories or centers that obtain them.
- Depositing sequences in GenBank / EMBL is a requisite of most journals to accept manuscripts in which new sequences are reported.
- In August 2003 they contained:
    - 18,197.000 sequence files
    - 22.617,000.000 nucleotides
- Every two months a new complete version of the database is published.

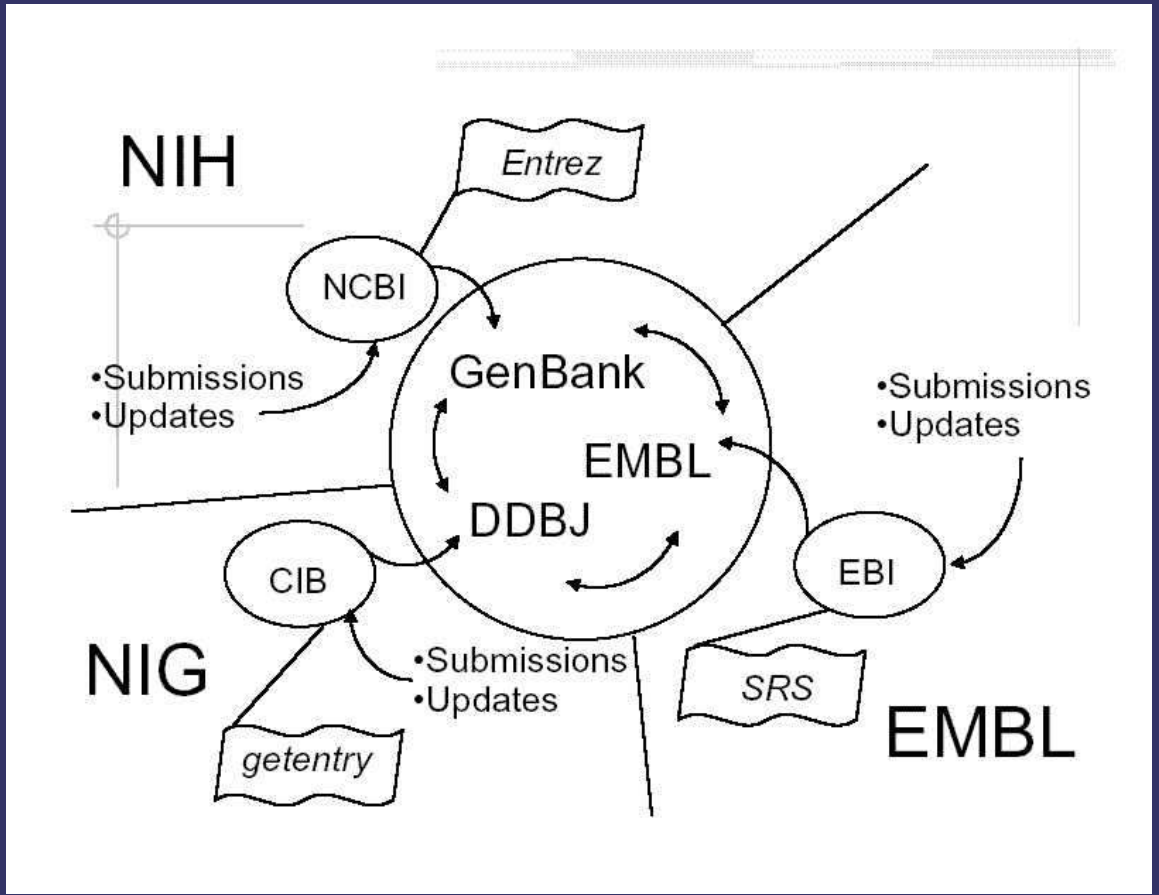
GenBank, EMBL AND DDBJ

GenBank and the EMBL nucleotide database, together with the DNA Data Bank of Japan, are part of the International Nucleotide Sequence Database Collaboration.

The three databases exchange information and update every 24 hours.

They contain the same type of primary information.

In addition, they store protein amino acid sequence, which is information of secondary type (predictions).



DATA RETRIEVAL FROM GenBank / EMBL / DDBJ

GenBank

- FTP ftp.ncbi.nih.gov
- WWW www.ncbi.nlm.nih.gov

EMBL

- FTP http://www.ebi.ac.uk/FTP/
- WWW http://www.ebi.ac.uk/Databases/nucleotide.html

DDBJ

- WWW http://www.ddbj.nig.ac.jp/Welcome-e.html

GenBank / EMBL / DDBJ FTP ACCESS:

Organismal Divisions in GenBank

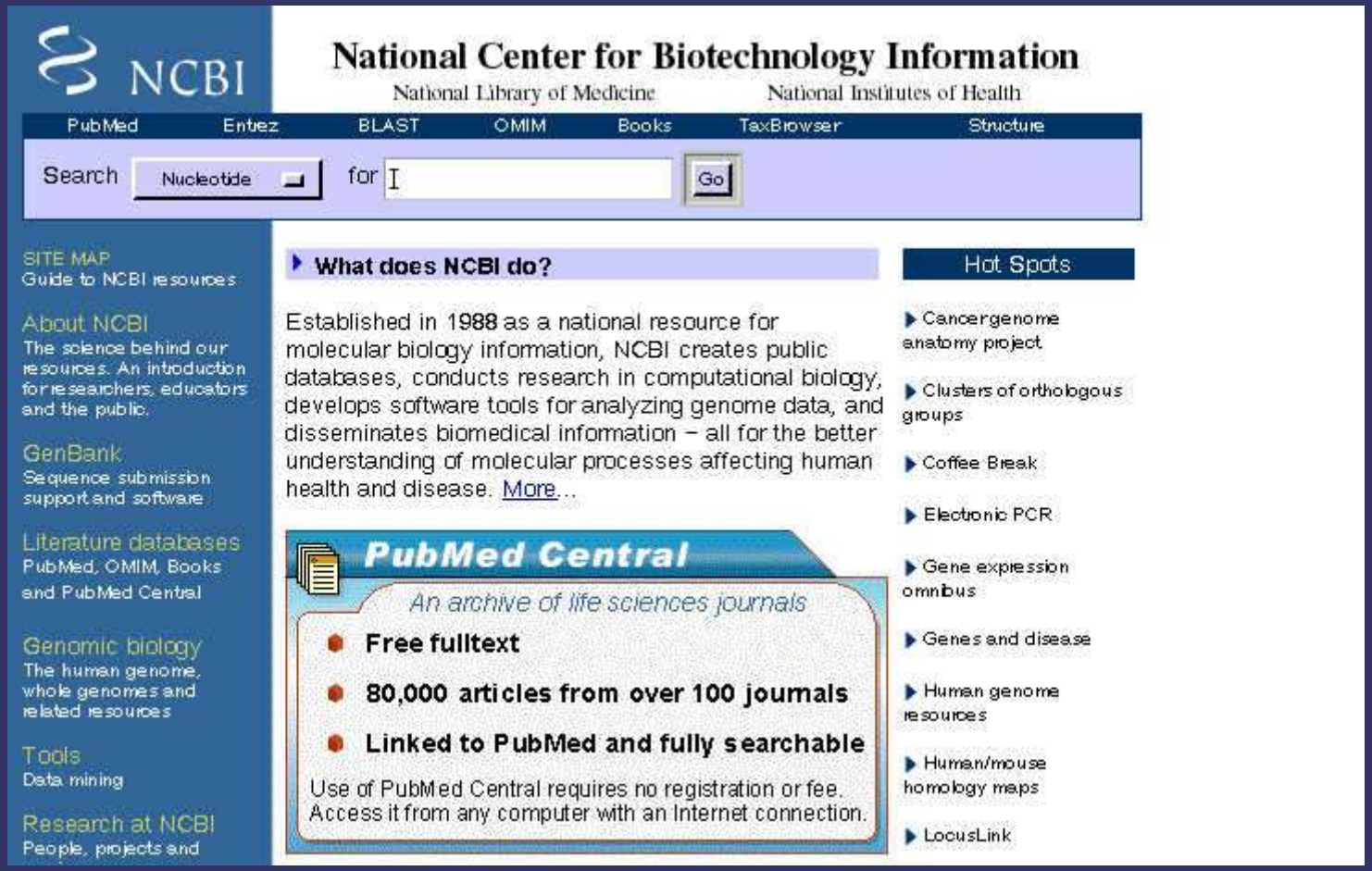
PRI	Primate	BCT	Bacterial
ROD	Rodent	RNA	Structural
MAM	Mammalian	VRL	Viral
VRT	Vertebrate	PHG	Phage
INV	Invertebrate	SYN	Synthetic
PLN	Plant	UNA	Unannotated

Funcional Divisions in GenBank

PAT	Patent
EST	Expressed Sequence Tags
STS	Sequence Tagged Sites
GSS	Genome Survey Sequences
HTG	High Throughput Genome

ENTREZ

Main page of the NCBI server



MAIN DATABASES ACCESIBLE WITH ENTREZ

PubMed: Bibliographic references in Molecular Biology and Medicine.

Nucleotide: Composite database of DNA sequences from Genbank, EMBL and DDBJ, plus other databases or projects such as RefSeq.

> RefSeq contains nucleotide sequences from the Nucleotide database that have been curated or re-annotated, by the NCBI.

Protein: Proteins sequences derived from translation of DNA sequences in GenBank, EMBL y DDBJ, plus sequences from PIR, SWISSPROT and Protein Data Bank (PDB).

Genome: Complete genomes, chromosomes, contig maps, physical maps.

MAIN DATABASES ACCESIBLE WITH ENTREZ

Entrez Gene: locus centered database that integrates information from other databases (replaces LocusLink).

Structure: (Molecular Modeling Database, MMDB) experimentally obtained structures from Protein Data Bank (PDB).

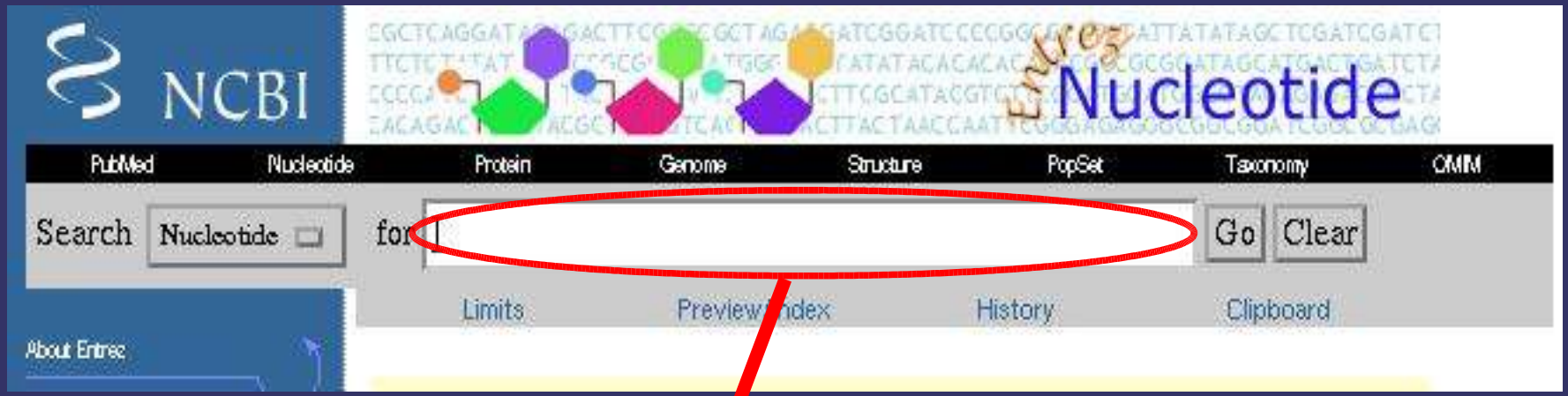
Taxonomy: Names and taxonomy of organisms that have at least one sequence at the NCBI databases.

OMIM: Online Mendelian Inheritance in Man, catalog of human mutations and associated diseases.

ENTREZ

Several menus to access the different databases

TEXT QUERIES



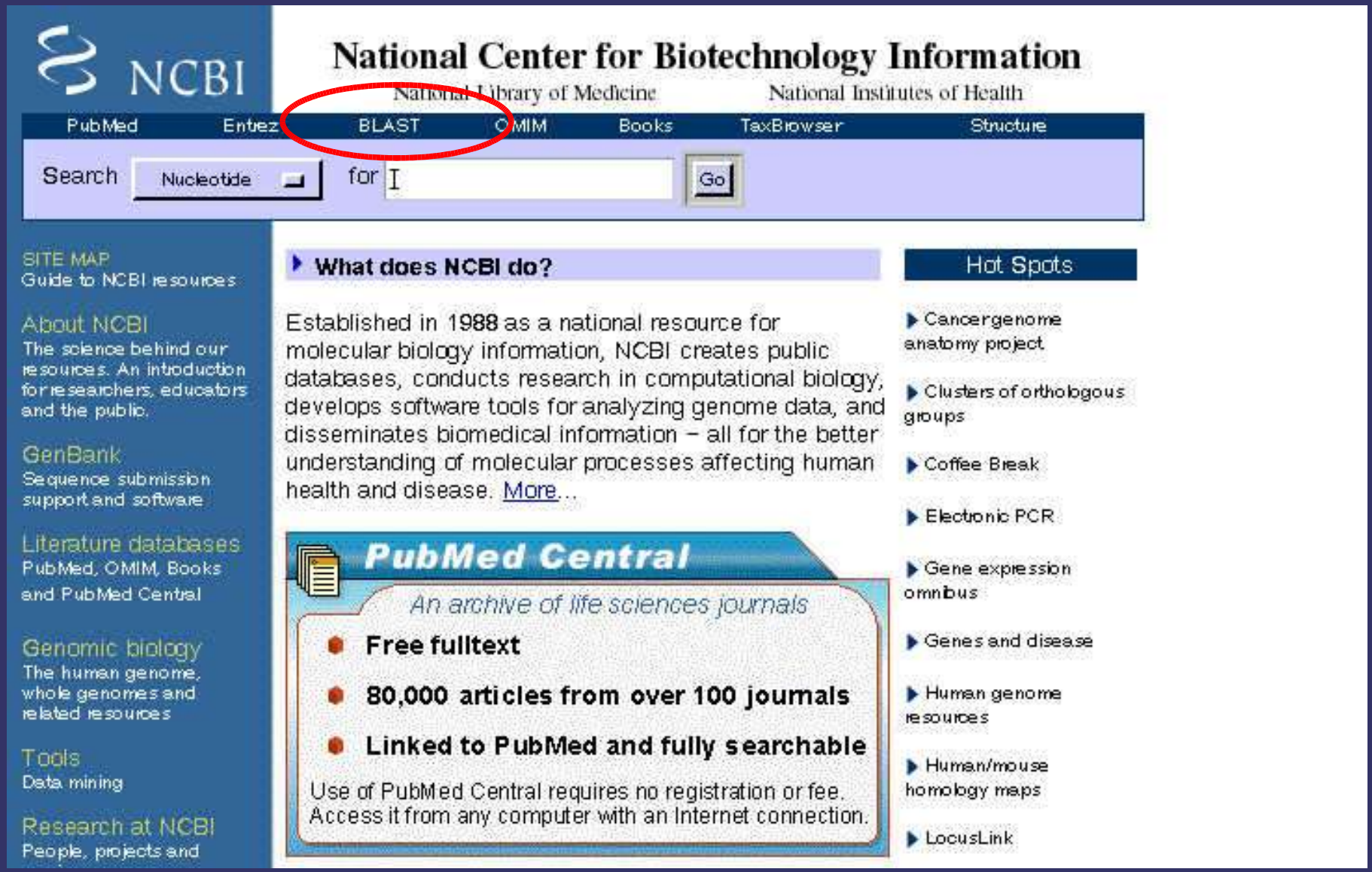
Word searches: RNA

Boolean operators : AND / OR / NOT : 16s AND RNA

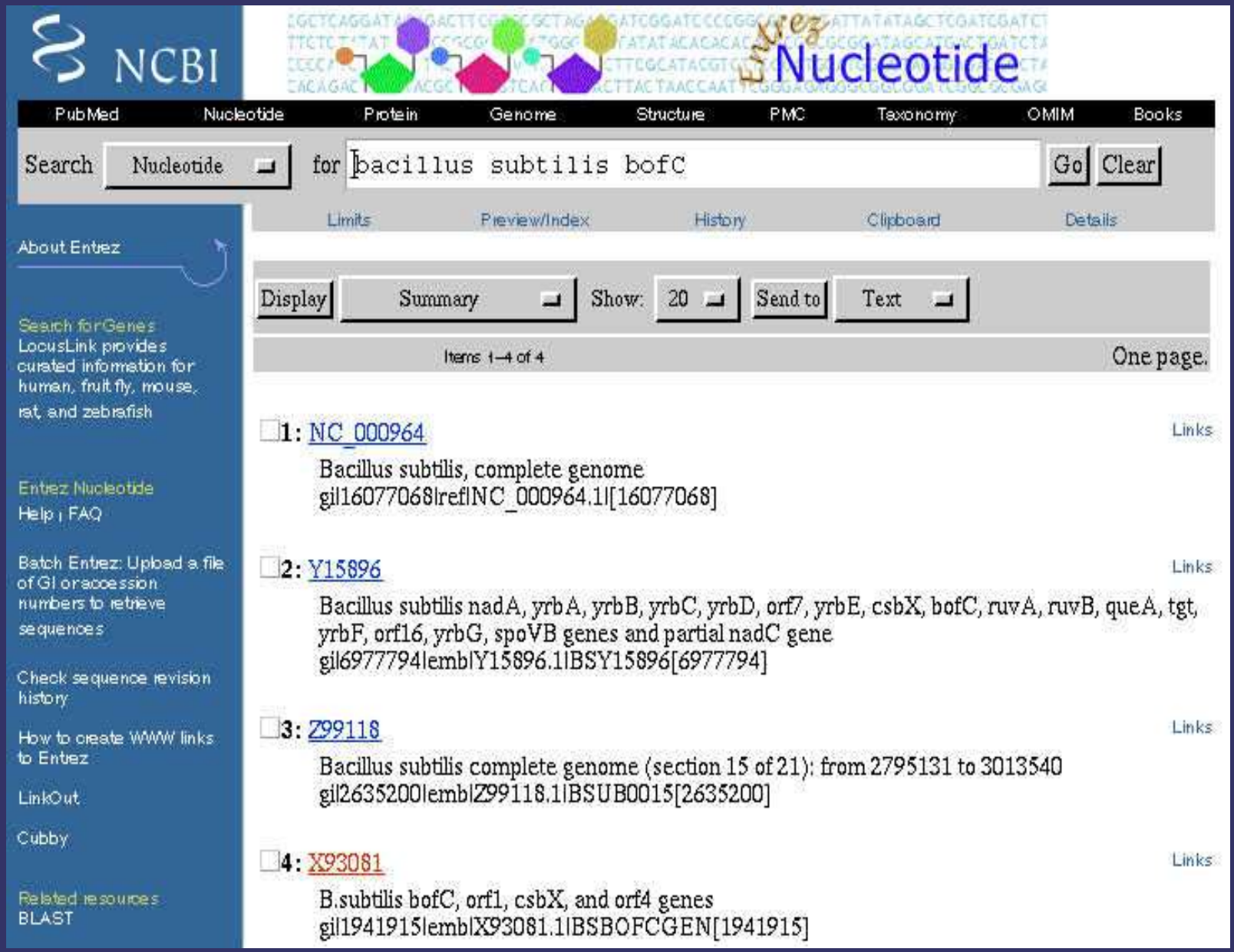
Phrase searching: "16s RNA"

Field restriction: Protein [AUTH]

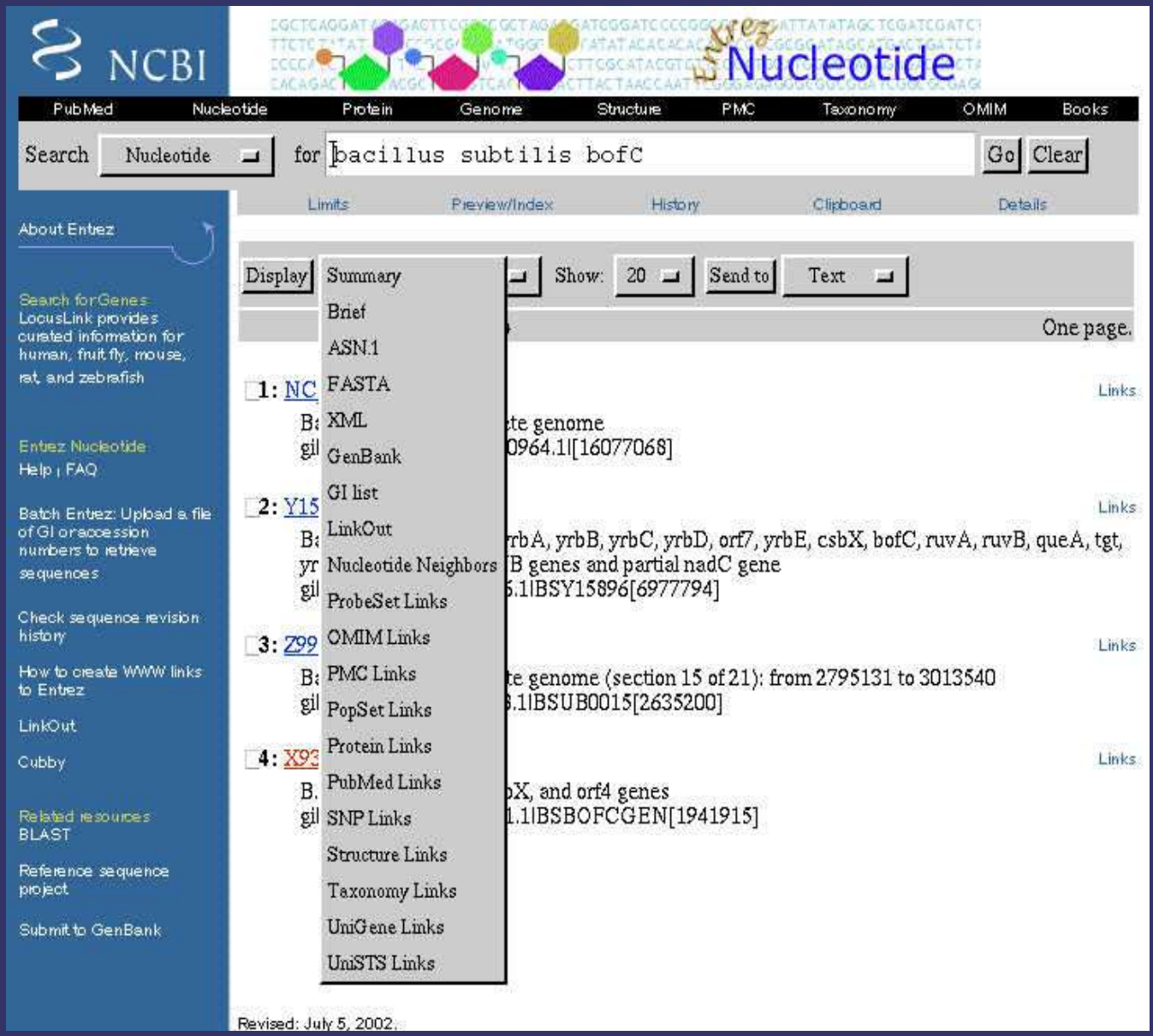
QUERY BY CONTENT



HIT LIST

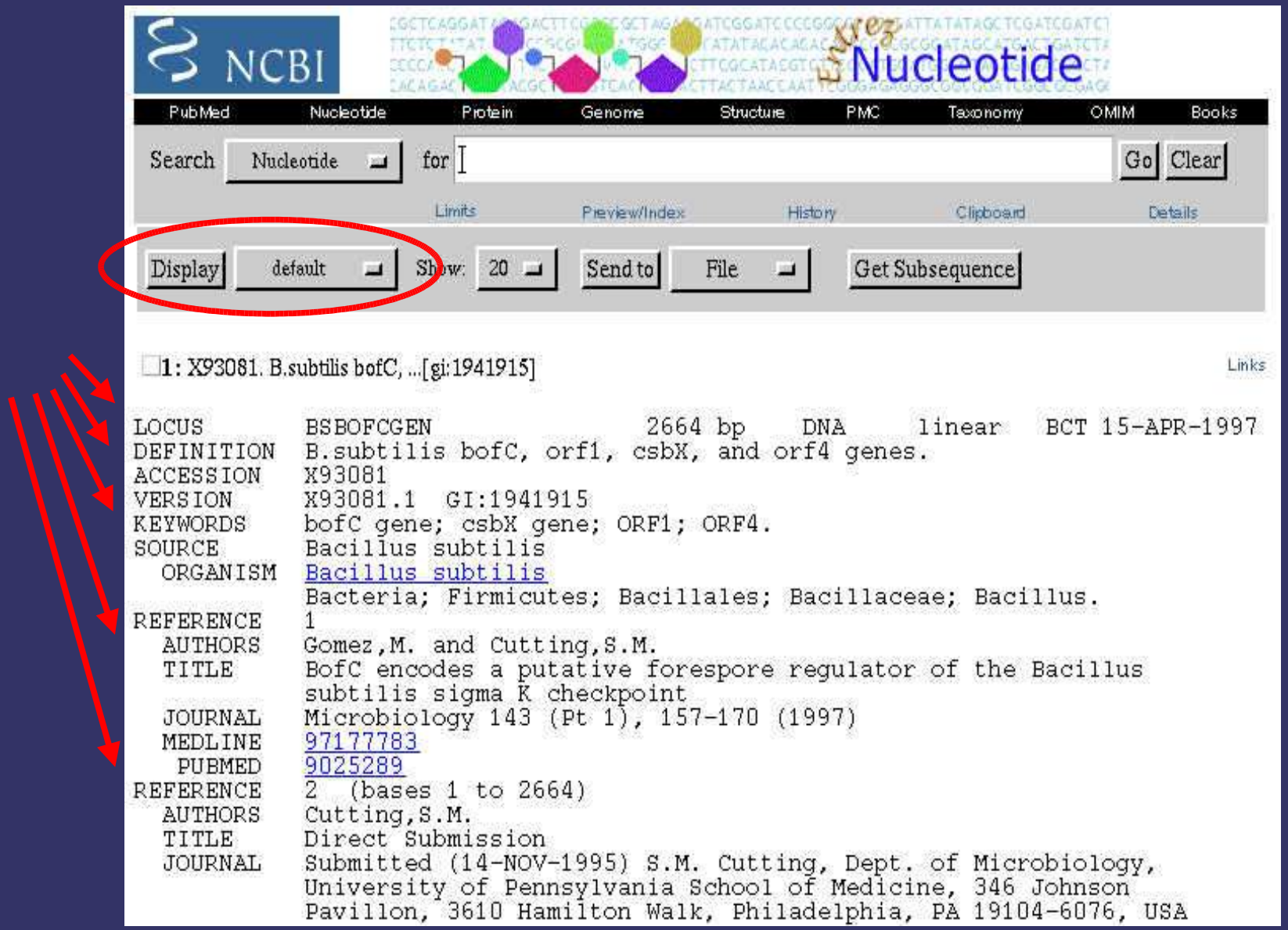

NCBI
Entrez Nucleotide
PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books
Search Nucleotide for bacillus subtilis bofC Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish
Entrez Nucleotide
Help | FAQ
Batch Entrez: Upload a file of GI or accession numbers to retrieve sequences
Check sequence revision history
How to create WWW links to Entrez
LinkOut
Cubby
Related resources
BLAST
Display Summary Show: 20 Send to Text
Items 1–4 of 4
One page.
1: NC_000964 Links
Bacillus subtilis, complete genome
gi|16077068|ref|NC_000964.1|[16077068]
2: Y15896 Links
Bacillus subtilis nadA, yrbA, yrbB, yrbC, yrbD, orf7, yrbE, csbX, bofC, ruvA, ruvB, queA, tgt, yrbF, orf16, yrbG, spoVB genes and partial nadC gene
gi|6977794|emb|Y15896.1|BSY15896[6977794]
3: Z99118 Links
Bacillus subtilis complete genome (section 15 of 21): from 2795131 to 3013540
gi|2635200|emb|Z99118.1|BSUB0015[2635200]
4: X93081 Links
B.subtilis bofC, orf1, csbX, and orf4 genes
gi|1941915|emb|X93081.1|BSBOFCGEN[1941915]

DISPLAY FORMATS

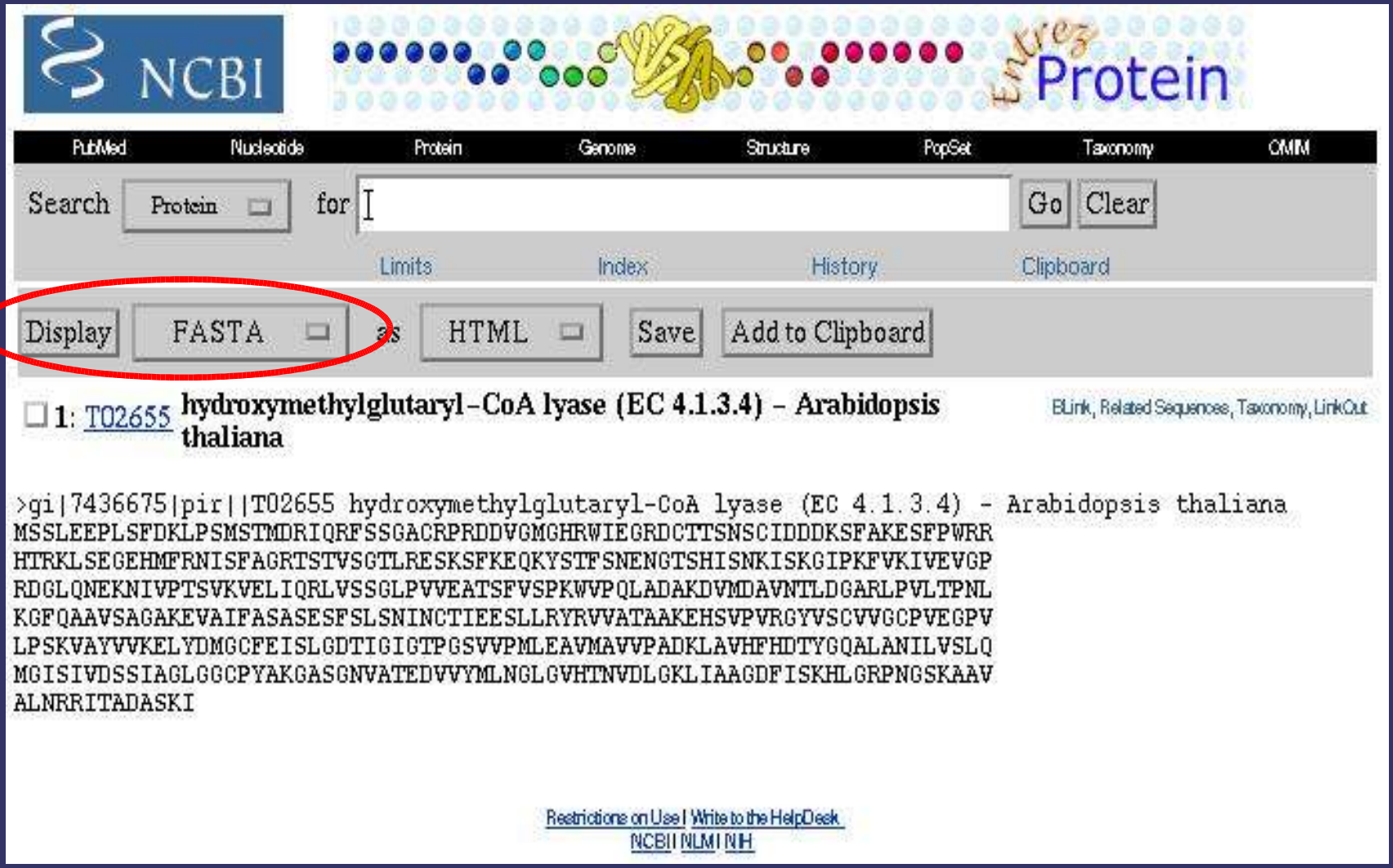


GenBank FORMAT

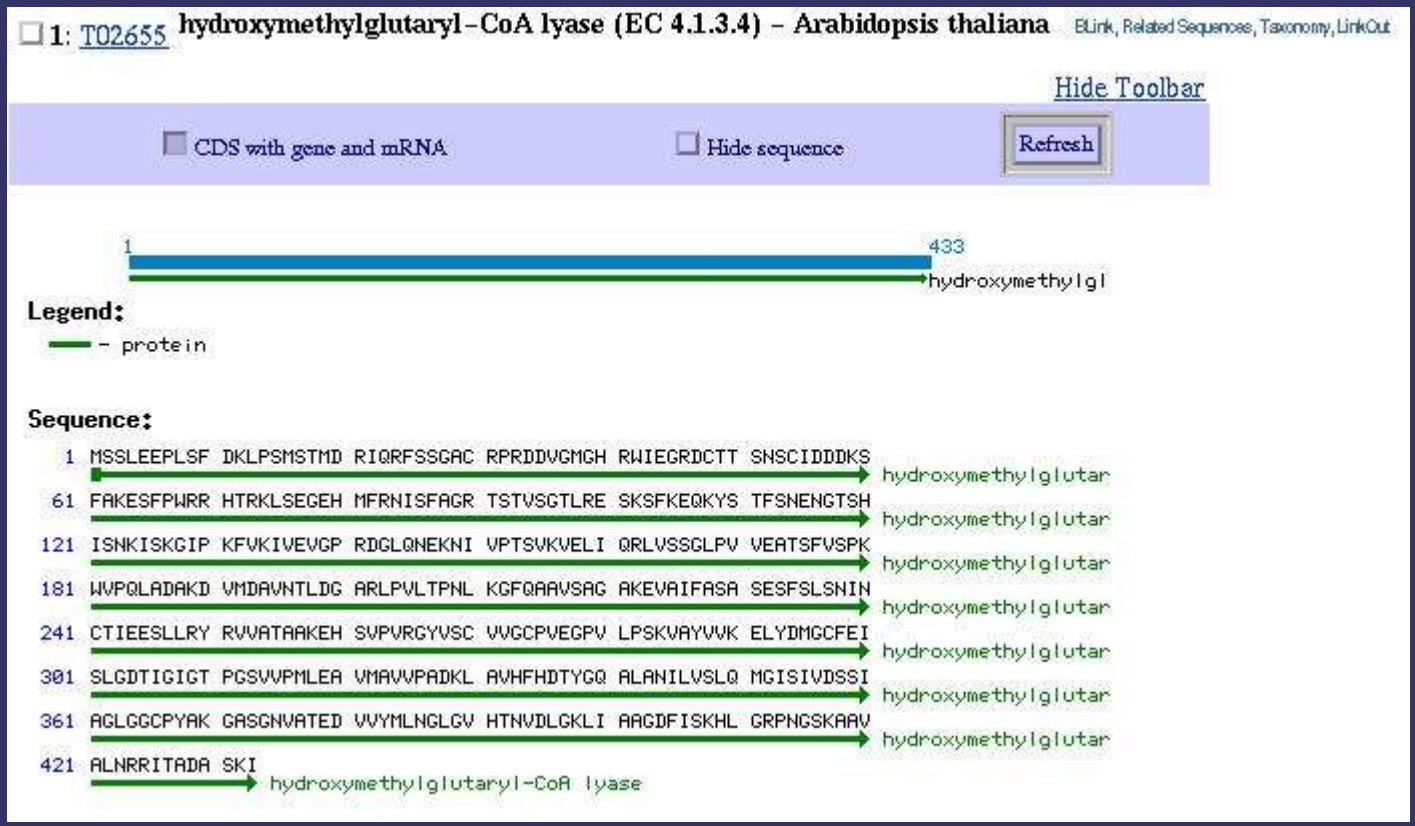
Fields



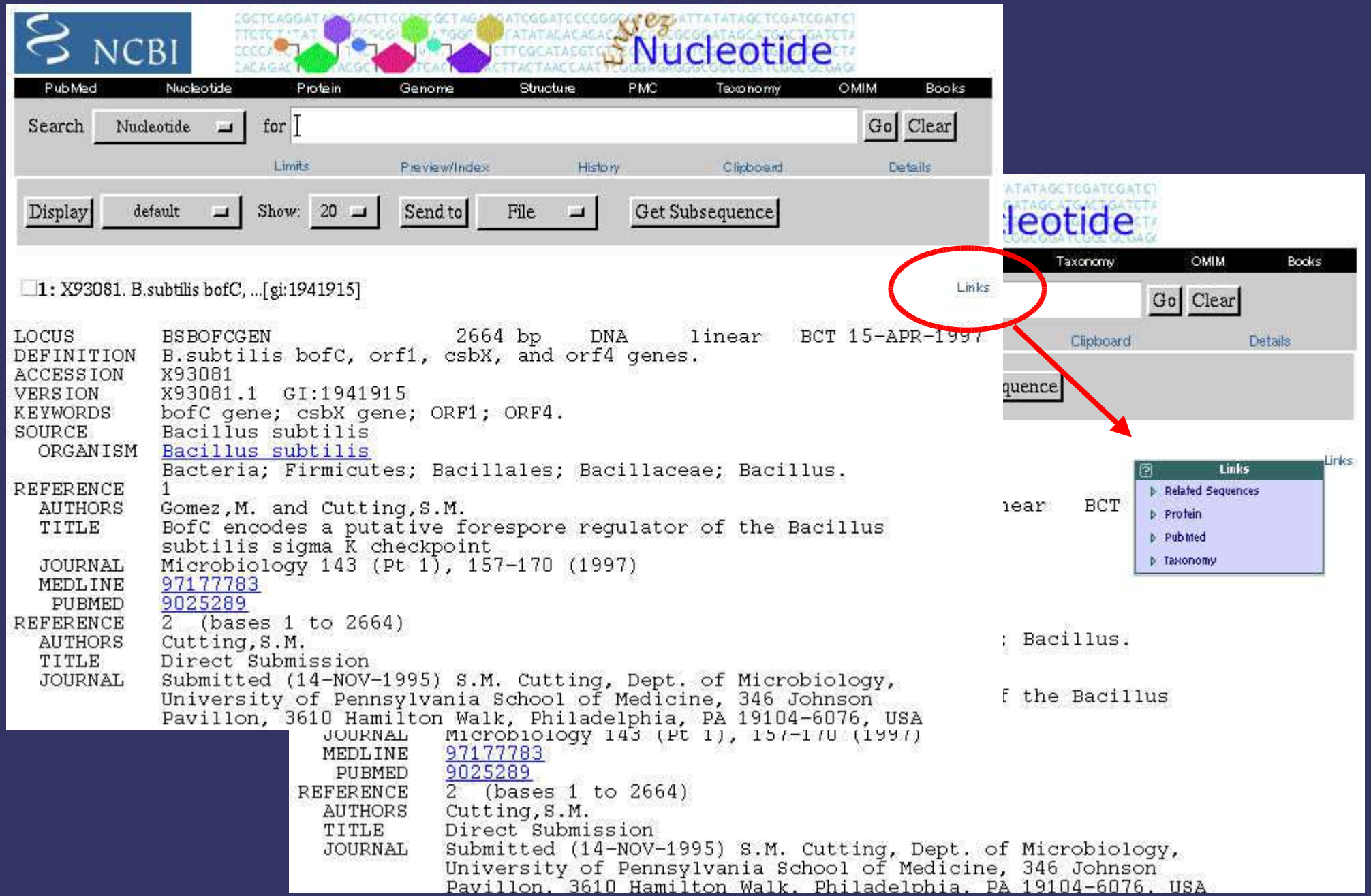
DISPLAY RESULTS: FASTA format



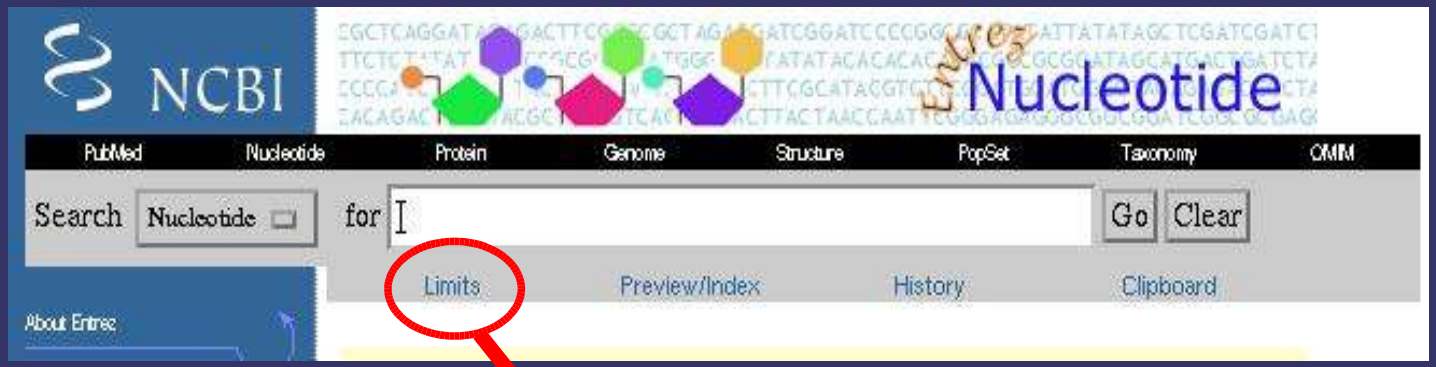
DISPLAY RESULTS: GRAPHICS

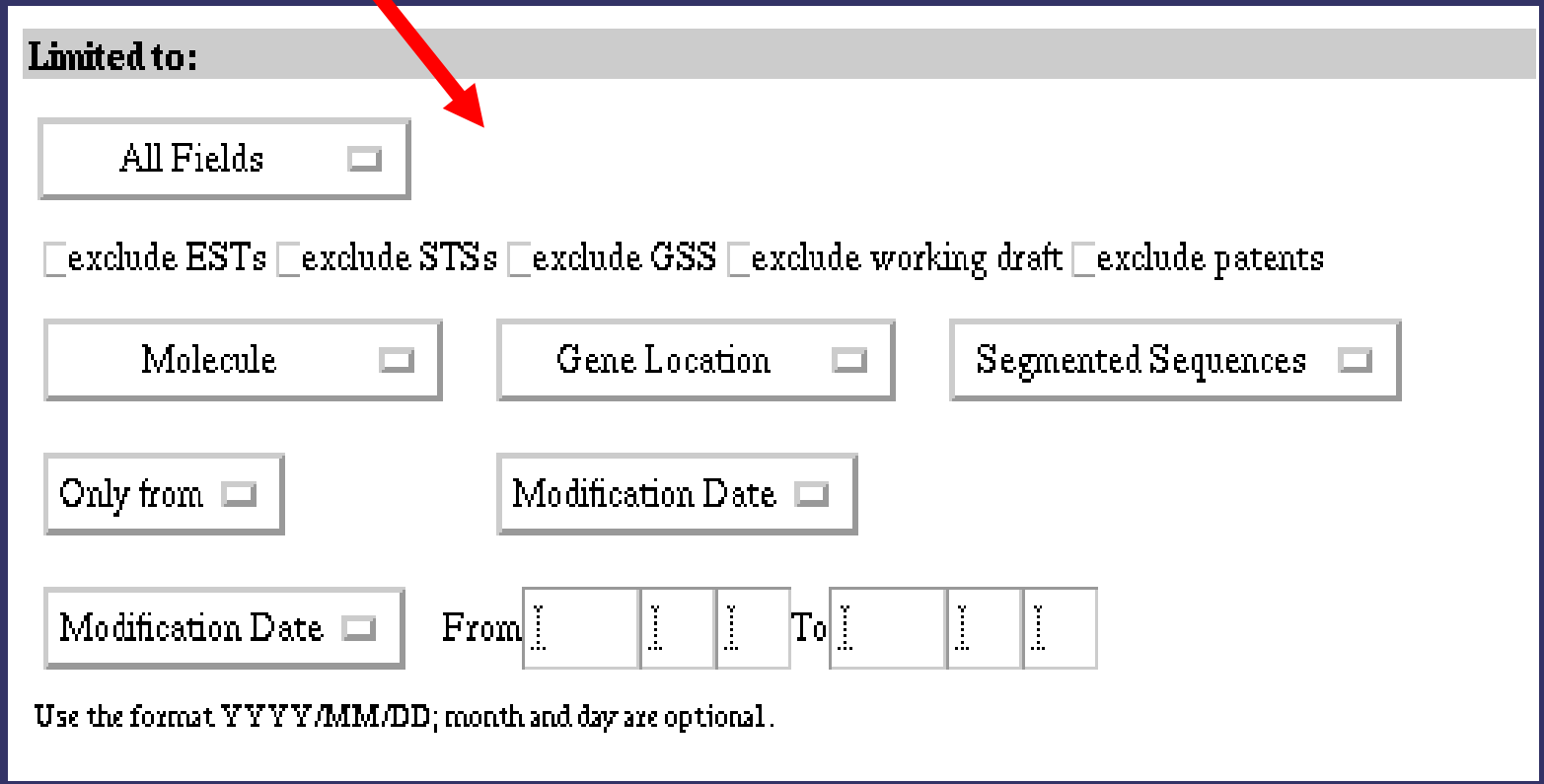


LINKS TO OTHER NCBI Dbs

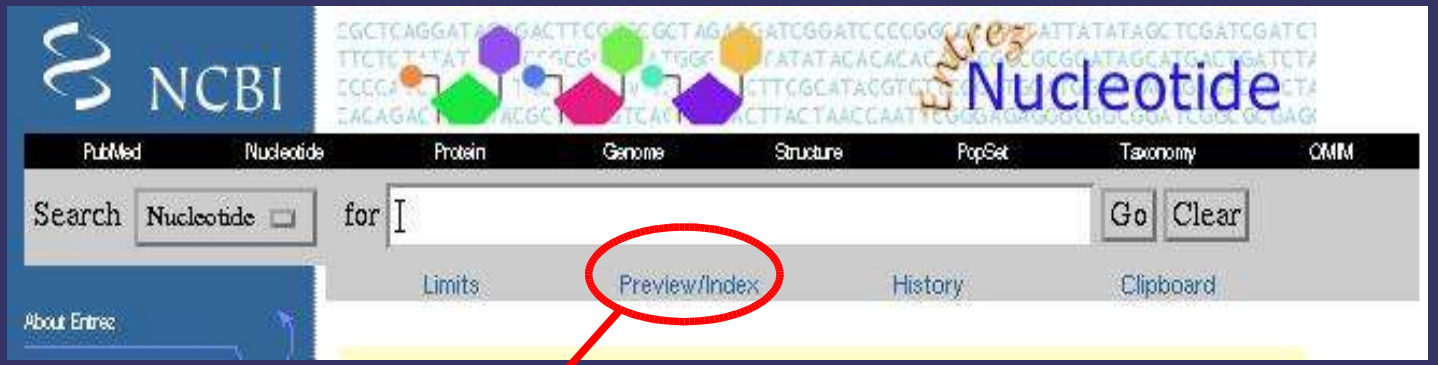


LIMITS


NCBI
Entrez Nucleotide
PubMed
Nucleotide
Protein
Genome
Structure
PopSet
Taxonomy
OMIM
Search
Nucleotide
for
Go
Clear
Limits
Preview/Index
History
Clipboard
About Entrez


Limited to:
All Fields
exclude ESTs
exclude STSs
exclude GSS
exclude working draft
exclude patents
Molecule
Gene Location
Segmented Sequences
Only from
Modification Date
Modification Date
From
To
Use the format YYYY/MM/DD; month and day are optional.

RESULTS INDEX



Add Term(s) to Query or View Index:

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.
- Multiple terms selected from Index will be ORed; click AND to add to search.

Title | mycobacterium | Preview | Index

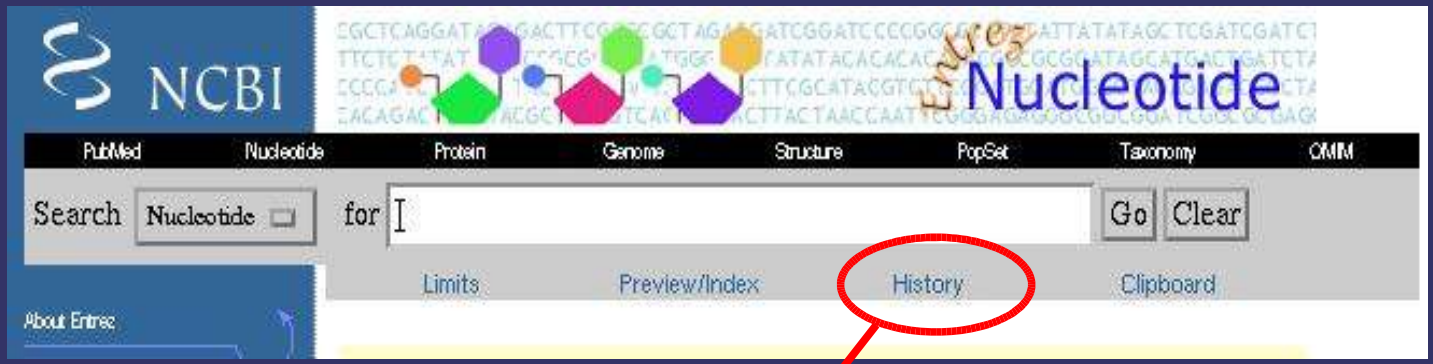
Click AND OR NOT to add terms selected from Index to the query box.

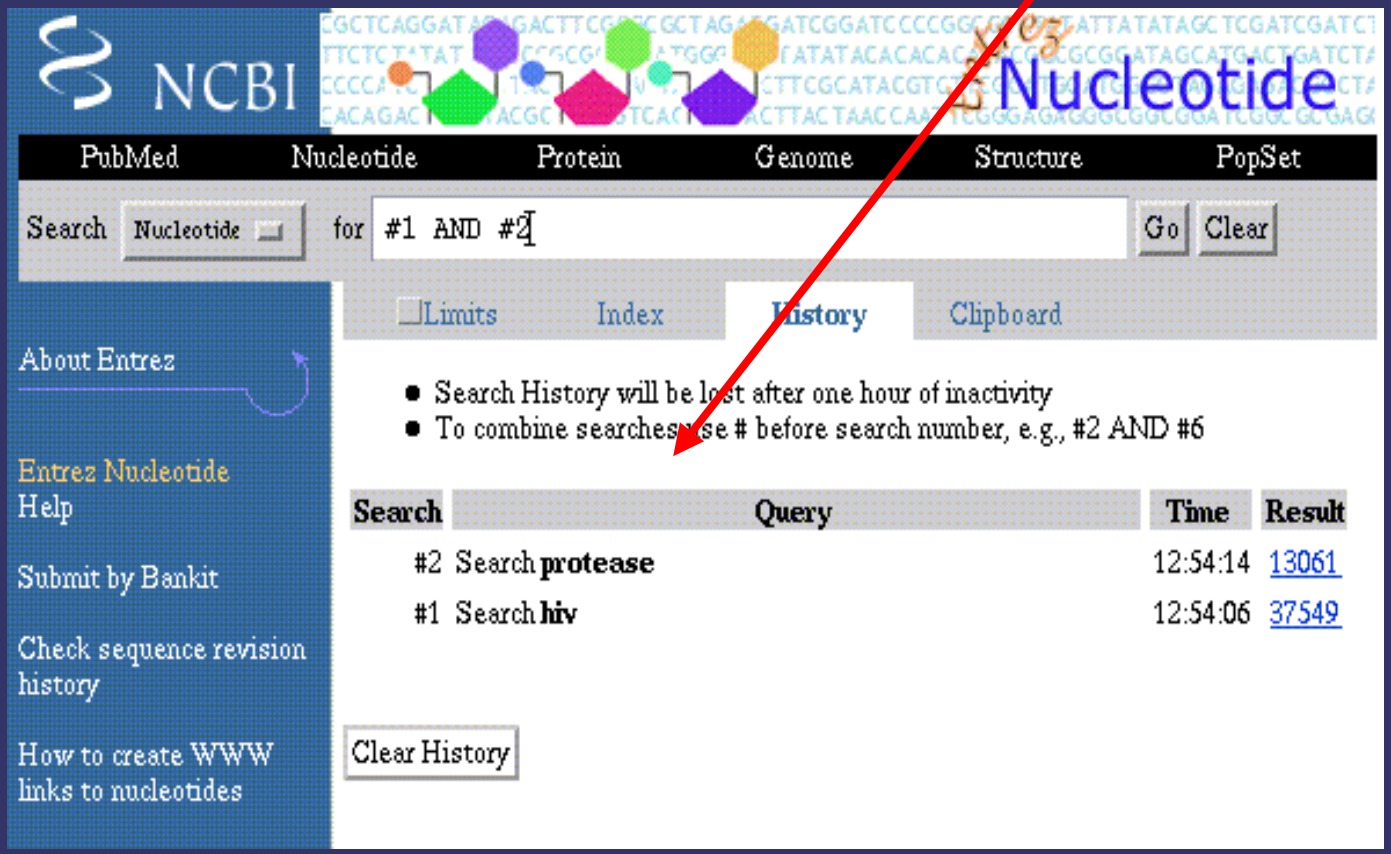
mycobacterium(41491)
mycobacterium/analysis(456)
mycobacterium/analysis@(2787)
mycobacterium/anatomy and histology@(293)
mycobacterium/chemistry(174)
mycobacterium/chemistry@(174)
mycobacterium/classification(1286)
mycobacterium/corynebacterium(1)
mycobacterium/cytology(156)
mycobacterium/cytology@(293)

Up

Down

SEARCH HISTORY





SEQUENCE FILES

FLAT FILE FORMATS

For sequence databases, the main formats are:

- FASTA
- GenBank
- EMBL and SwissProt

FASTA format


```
>gi|1941915|emb|X93081.1|BSBOFCGEN B.subtilis bofC gene
CTGCAGCGGCTGACAATAGCAGGCCGACAACGGTTGAGGTGTCAACAGCTGATTTTGTGATGAAGGATAA
ACCGCATTTCTTTTTCCTTGAACGCTATAAGGATTCATATGAGGAGGAGATTCTCCGTTTTGCAGAAGCG
ATCGGCACAAACCAGGAGACTCCCTGCACCGGCAATGACGGTTTACAGGCCGGGAGGATCGCCAGAGCAG
CACAGCAATCGCTTGCTTTTGGCATGCCTGTTAGCATTGAGCACACTGAAAAAATCGCTTTTTAATCTAA
CAGGATTACAATTCAGCAAGCTTGGGTATATACTCCATTGATACTTTAAGTAGGCGGTGGAGAAAATGAA
TACAGTACATGCTAAAGGAAATGTTTTGAACAAAATCGGAATTCCTTCTCACATGGTTTGGGGTTATATT
GGCGTTGTCATCTTTATGGTTGGAGACGGCCTCGAACAAGGCTGGCTGTCTCCTTTTCTCGTTGATCATG
GTCTCAGTATGCAGCAATCCGCATCGTTATTTACCATGTACGGCATTGCTGTCACCATCTCAGCTTGGCT
TTCAGGAACGTTTGTGGAAACTTGGGGGCCGAGAAAAACGATGACTGTCGGATTGCTTGCATTTATCCTC

>
CTGCAGCGGCTGACAATAGCAGGCCGACAACGGTTGAGGTGTCAACAGCTGATTTTGTGATGAAGGATAA
ACCGCATTTCTTTTTCCTTGAACGCTATAAGGATTCATATGAGGAGGAGATTCTCCGTTTTGCAGAAGCG
ATCGGCACAAACCAGGAGACTCCCTGCACCGGCAATGACGGTTTACAGGCCGGGAGGATCGCCAGAGCAG
CACAGCAATCGCTTGCTTTTGGCATGCCTGTTAGCATTGAGCACACTGAAAAAATCGCTTTTTAATCTAA
CAGGATTACAATTCAGCAAGCTTGGGTATATACTCCATTGATACTTTAAGTAGGCGGTGGAGAAAATGAA
```

GenBank Flat File (GBFF)

GBFF: HEADER

```
LOCUS       BSBOFCGEN               2664 bp    DNA     linear   BCT 15-APR-1997
DEFINITION  B.subtilis bofC, orf1, csbX, and orf4 genes.
ACCESSION   X93081
VERSION     X93081.1  GI:1941915
KEYWORDS    bofC gene; csbX gene; ORF1; ORF4.
SOURCE      Bacillus subtilis
  ORGANISM  Bacillus subtilis
            Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus.
REFERENCE   1
  AUTHORS   Gomez,M. and Cutting,S.M.
  TITLE     BofC encodes a putative forespore regulator of the Bacillus
            subtilis sigma K checkpoint
  JOURNAL   Microbiology 143 (Pt 1), 157-170 (1997)
  MEDLINE   97177783
   PUBMED   9025289
REFERENCE   2  (bases 1 to 2664)
  AUTHORS   Cutting,S.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (14-NOV-1995) S.M. Cutting, Dept. of Microbiology,
            University of Pennsylvania School of Medicine, 346 Johnson
            Pavillon, 3610 Hamilton Walk, Philadelphia, PA 19104-6076, USA
```

GBFF: FEATURES

```
FEATURES             Location/Qualifiers
     source          1..2664
                     /organism="Bacillus subtilis"
                     /strain="PY79"
                     /isolate="168"
                     /db_xref="taxon:1423"
                     /germline
     gene            1..275
                     /gene="orf1"
     CDS             <1..275
                     /gene="orf1"
                     /codon_start=3
                     /transl_table=11
                     /protein_id="CAA63619.1"
                     /db_xref="GI:1941916"
                     /db_xref="SPTREMBL:O05389"
                     /translation="AAADNSRPTTVEVSTADFVMKDKPHFFFLERYKDSYEEEILRFA
                     EAIGTNQETPCTGNDGLQAGRIARAAQQSLAFGMPVSIEHTEKIAF"
     gene            346..1650
                     /gene="csbX"
     CDS             346..1650
                     /gene="csbX"
                     /note="sigma B transcribed gene"
                     /codon_start=1
```

GBFF: SEQUENCE

```
BASE COUNT      670 a    518 c    690 g    786 t
ORIGIN
        1 ctgcagcggc tgacaatagc aggccgacaa cggttgaggt gtcaacagct gattttgtga
       61 tgaaggataa accgcatttc tttttccttg aacgctataa ggattcatat gaggaggaga
      121 ttctccgttt tgcagaagcg atcggcacaa accaggagac tccctgcacc ggcaatgacg
      181 gtttacaggc cgggaggatc gccagagcag cacagcaatc gcttgctttt ggcatgcctg
      241 ttagcattga gcacactgaa aaaatcgctt tttaatctaa caggattaca attcagcaag
      301 cttgggtata tactccattg atactttaag taggcggtgg agaaaatgaa tacagtacat
      361 gctaaaggaa atgttttgaa caaaatcgga attccttctc acatggtttg gggttatatt
      421 ggcgttgtca tctttatggt tggagacggc ctcgaacaag gctggctgtc tccttttctc
      481 gttgatcatg gtctcagtat gcagcaatcc gcatcgttat ttaccatgta cggcattgct
      541 gtcaccatct cagcttggct ttcaggaacg tttgtggaaa cttgggggcc gagaaaaacg
      601 atgactgtcg gattgcttgc atttatcctc ggttcggccg cttttatcgg ctgggcgatt
      661 cctcatatgt attatccggc tctcttgggc agctatgctc ttagaggctt gggatatccg
      721 ctgtttgcat actcttttct cgtatgggtg tcatacagca cctctcaaaa tattcttgga
      781 aaagccgtcg gctggttttg gtttatgttt acgtgcggcc ttaacgtgct cggtccgttc
      841 tattccagct atgcagttcc ggcctttgga gaaatcaata cgctttggag cgctttactg
      901 tttgtggcgg caggcggaat tcttgcctta ttttttaaca aagataaatt tactccgata
      961 caaaaacaag atcagccgaa atggaaagaa ctgtcgaagg catttacgat tatgtttgaa
     1021 aaccctaagg taggcatcgg cggagtggtc aagacgatta atgcgatagg acaatttgga
     1081 tttgccatct ttcttcctac ttatttagca cgatacgggt attcggtttc ggaatggctg
     1141 caaatatggg ggactctgtt ttttgtgaat
//
LOCUS       BSOTHERGENE               4356 bp    DNA     linear   BCT 15-APR-1997
```

EMBL AND SwissProt FORMAT

The field name is specified by a TWO CHARACTER keyword, in the beginning of each line, what makes easier to parse the file to extract specific information.

```
ID   BSBOFCGEN  standard; genomic DNA; PRO; 2664 BP.
XX
AC   X93081;
XX
SV   X93081.1
XX
DT   15-APR-1997 (Rel. 51, Created)
DT   15-APR-1997 (Rel. 51, Last updated, Version 12)
XX
DE   B.subtilis bofC, orf1, csbX, and orf4 genes
XX
KW   bofC gene; csbX gene; ORF1; ORF4.
XX
OS   Bacillus subtilis
OC   Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus.
XX
```

EMBL AND SwissProt FORMAT

```
RN   [1]
RX   MEDLINE; 97177783.
RX   PUBMED; 9025289.
RA   Gomez M., Cutting S.M.;
RT   "BofC encodes a putative forespore regulator
RT   of the Bacillus subtilis sigma K checkpoint";
RL   Microbiology 143:157-170(1997).
XX

XX
DR   GOA; O05389.
DR   GOA; O05390.
DR   GOA; O05391.
DR   GOA; O05392.
DR   SWISS-PROT; O05389; YRBE_BACSU.
DR   SWISS-PROT; O05390; CSBX_BACSU.
DR   SWISS-PROT; O05391; BOFC_BACSU.
DR   SWISS-PROT; O05392; RUVA_BACSU.
XX
```

EMBL AND SwissProt FORMAT

```
FT   source          1..2664
FT                   /db_xref="taxon:1423"
FT                   /germline
FT                   /mol_type="genomic DNA"
FT                   /organism="Bacillus subtilis"
FT                   /strain="PY79"
FT                   /isolate="168"
FT   CDS             <1..275
FT                   /codon_start=3
FT                   /db_xref="GOA:O05389"
FT                   /db_xref="SWISS-PROT:O05389"
FT                   /transl_table=11
FT                   /gene="orf1"
FT                   /protein_id="CAA63619.1"
FT                   /translation="AAADNSRPTTVEVSTADFVMKDKPHFFFL
                     ERYKDSYEEEILRFAEFTAIGTNQETPCTGNDGLQAGRIARAA
                     QQSLAFGMPVSIEHTEKIAF"
```

EMBL AND SwissProt FORMAT

```
XX
SQ   Sequence 2664 BP; 670 A; 518 C; 690 G; 786 T; 0 other;
     ctgcagcggc tgacaatagc aggccgacaa cggttgaggt gtcaacagct gattttgtga        60
     tgaaggataa accgcatttc tttttccttg aacgctataa ggattcatat gaggaggaga       120
     ttctccgttt tgcagaagcg atcggcacaa accaggagac tccctgcacc ggcaatgacg       180
     gtttacaggc cgggaggatc gccagagcag cacagcaatc gcttgctttt ggcatgcctg       240
     ttagcattga gcacactgaa aaaatcgctt tttaatctaa caggattaca attcagcaag       300
     cttgggtata tactccattg atactttaag taggcggtgg agaaaatgaa tacagtacat       360
     gctaaaggaa atgttttgaa caaaatcgga attccttctc acatggtttg gggttatatt       420
     ggcgttgtca tctttatggt tggagacggc ctcgaacaag gctggctgtc tccttttctc       480
//
```

GenBank IDENTIFIERS

Locus: unique , 10 character-long string used only by GenBank, and not maintained by EMBL. In the beginning, the strings had some type of meaning (for example, they referred to the source organism). Maintained only for historical reasons.

Accession: unique identifier in GenBank. It does not change when entries are updated. Ideal for publications. It is maintained in EMBL, in the lines identified with keyword AC. EMBL also adds its own identifier, that appears in lines labeled as ID.

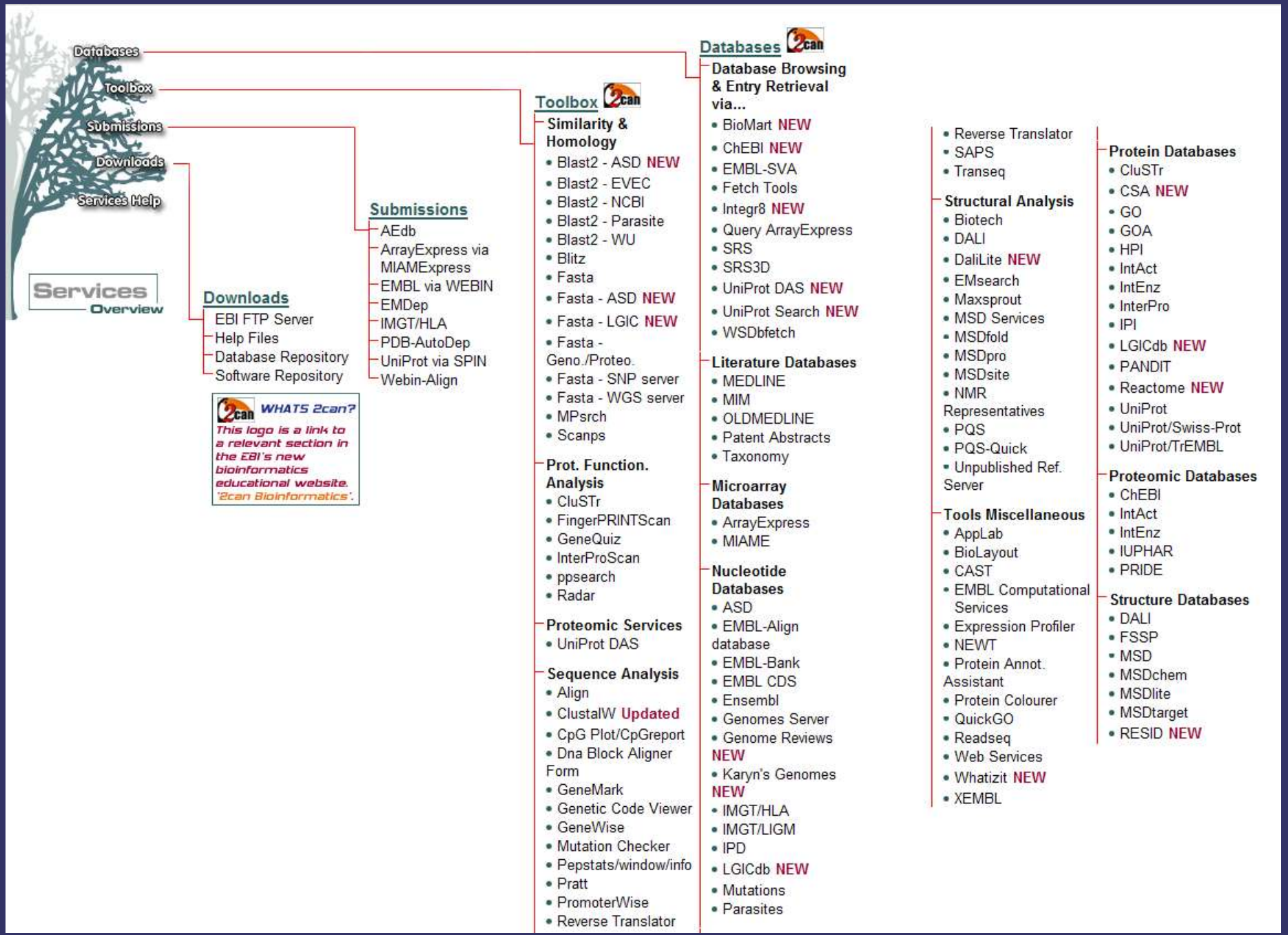
Nucleotide gi (gen identifier): unique identifier for each entry in GenBank, different for each entry update. In EMBL,the gi corresponds to the Nucleic Acid Identifier, which appears in lines NI.

Accession.version: the format accession.version combines the concepts of accession and gi. In EMBL, the corresponding identifier is at lines SV.

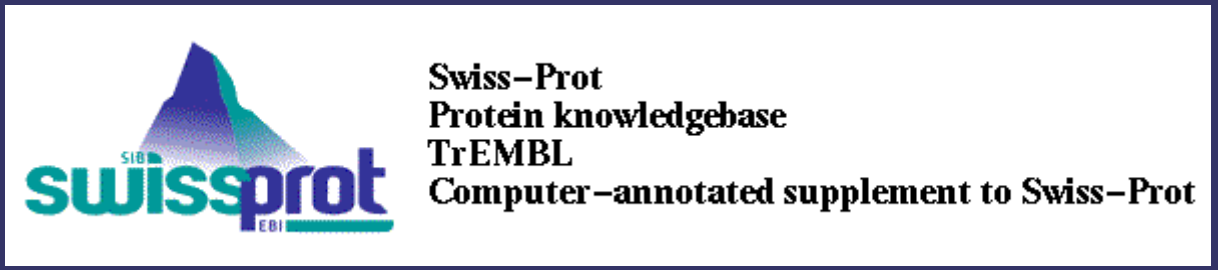

Database Search for
in
Nucleotide sequences
Go
Search EBI Website
Go
European Bioinformatics Institute
Research at the European Bioinformatics Institute
Bioinformatics Products and Services
FLASH Intro
European Bioinformatics Institute
a part of the European Molecular Biology Laboratory
the path to knowledge
About EBI | Funding | Whats New | Research Groups | Services | Toolbox | EBI Databases | Downloads | Submissions
BioMart Database Queries | SRS Database Queries | Site Search | Site Map | Services Map | PhD Studies | Contact Us | Terms of Use
2can Bioinformatics - Training and Education at the EBI
© Copyright European Bioinformatics Institute 2002-2004. All Rights and Trademarks Reserved.


GRENOBLE · HAMBURG · HEIDELBERG · HINXTON · MONTEROTONDO
EMBL
European Molecular Biology Laboratory
Search
RESEARCH IN MOLECULAR BIOLOGY
Services
EMBL services
Biological databases at EBI
Computational services
Core facilities
Szilárd Library
Technology Transfer
Upcoming Events
News
18 February 2004
EMBL researchers discover key molecular "switch" in eye development of medaka fish
13 February 2004
In Silico Research: First German Center for Modeling and Simulation in the Life Sciences Established in Heidelberg
12 February 2004
European researchers launch 10 Million Euro collaborative technology project
News archives
EMBL Publications
EMBL Internal
Local Information
EMBL Services
MOLECULAR BIOLOGY
© EMBL 2002


EMBnet


Databases
Toolbox
Submissions
Downloads
Services Help
Services
Overview
Downloads
EBI FTP Server
Help Files
Database Repository
Software Repository
Submissions
AEdb
ArrayExpress via MIAMExpress
EMBL via WEBIN
EMDep
IMGT/HLA
PDB-AutoDep
UniProt via SPIN
Webin-Align
WHATS 2can?
This logo is a link to a relevant section in the EBI's new bioinformatics educational website. '2can Bioinformatics'.
Toolbox
Similarity & Homology
Blast2 - ASD NEW
Blast2 - EVEC
Blast2 - NCBI
Blast2 - Parasite
Blast2 - WU
Blitz
Fasta
Fasta - ASD NEW
Fasta - LGIC NEW
Fasta - Geno./Proteo.
Fasta - SNP server
Fasta - WGS server
MPsrch
Scanps
Prot. Function. Analysis
CluSTr
FingerPRINTScan
GeneQuiz
InterProScan
ppsearch
Radar
Proteomic Services
UniProt DAS
Sequence Analysis
Align
ClustalW Updated
CpG Plot/CpGreport
Dna Block Aligner Form
GeneMark
Genetic Code Viewer
GeneWise
Mutation Checker
Pepstats/window/info
Pratt
PromoterWise
Reverse Translator
Databases
Database Browsing & Entry Retrieval via...
BioMart NEW
ChEBI NEW
EMBL-SVA
Fetch Tools
Integr8 NEW
Query ArrayExpress
SRS
SRS3D
UniProt DAS NEW
UniProt Search NEW
WSDbfetch
Literature Databases
MEDLINE
MIM
OLDMEDLINE
Patent Abstracts
Taxonomy
Microarray Databases
ArrayExpress
MIAME
Nucleotide Databases
ASD
EMBL-Align database
EMBL-Bank
EMBL CDS
Ensembl
Genomes Server
Genome Reviews NEW
Karyn's Genomes NEW
IMGT/HLA
IMGT/LIGM
IPD
LGICdb NEW
Mutations
Parasites
Reverse Translator
SAPS
Transeq
Structural Analysis
Biotech
DALI
DaliLite NEW
EMsearch
Maxsprout
MSD Services
MSDfold
MSDpro
MSDsite
NMR Representatives
PQS
PQS-Quick
Unpublished Ref. Server
Tools Miscellaneous
AppLab
BioLayout
CAST
EMBL Computational Services
Expression Profiler
NEWT
Protein Annot. Assistant
Protein Colourer
QuickGO
Readseq
Web Services
Whatizit NEW
XEMBL
Protein Databases
CluSTr
CSA NEW
GO
GOA
HPI
IntAct
IntEnz
InterPro
IPI
LGICdb NEW
PANDIT
Reactome NEW
UniProt
UniProt/Swiss-Prot
UniProt/TrEMBL
Proteomic Databases
ChEBI
IntAct
IntEnz
IUPHAR
PRIDE
Structure Databases
DALI
FSSP
MSD
MSDchem
MSDlite
MSDtarget
RESID NEW

SwissProt AND TrEMBL



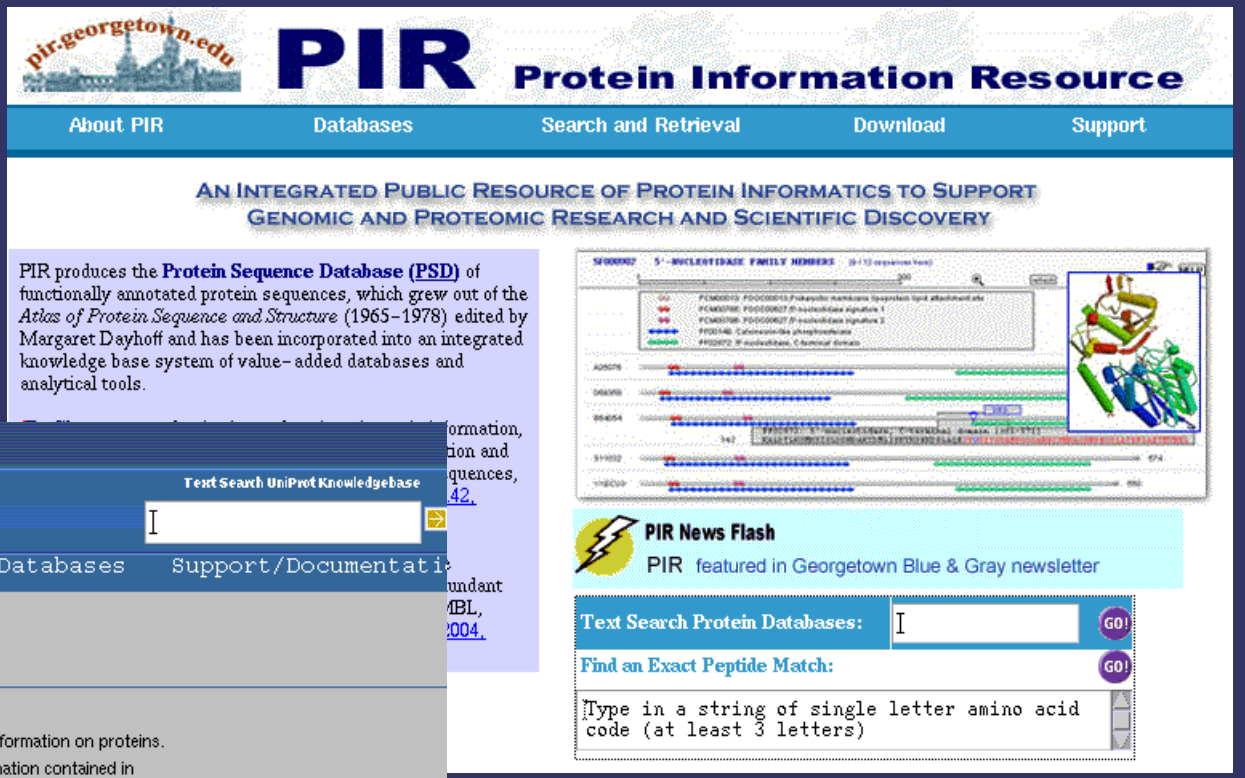
SwissProt: Developed first at the University of Geneva (Amos Bairoch, 1986), then at the Swiss Institute for Bioinformatics (SIB) and now maintained by SIB and EBI.

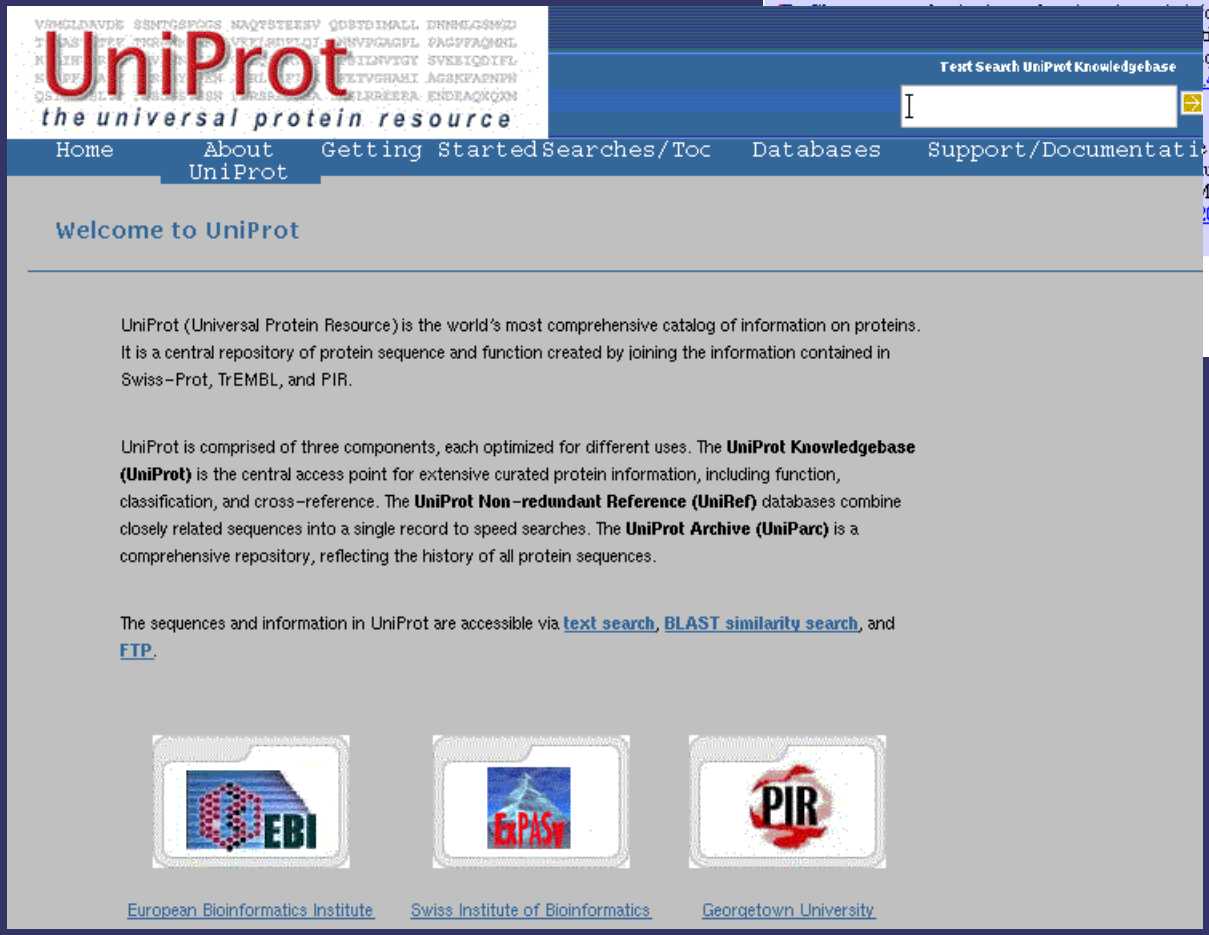
Protein database, non redundant, manually curated, with high quality annotations, and pioneer in recognizing the importance of collecting cross-references (in 2005 it contained references to 60 databases). Relatively low coverage (170.000 entries in 2005).

TrEMBL: automatically annotated protein sequences, which are generated by translating ORFs predicted at the EMBL nucleotide database (1.600.000 entries in 2005).

PIR: Protein Information Resource

PIR is the first database on Molecular Biology data ever created. Based on the Atlas of Protein Sequence and Structure of Margaret Daihoff



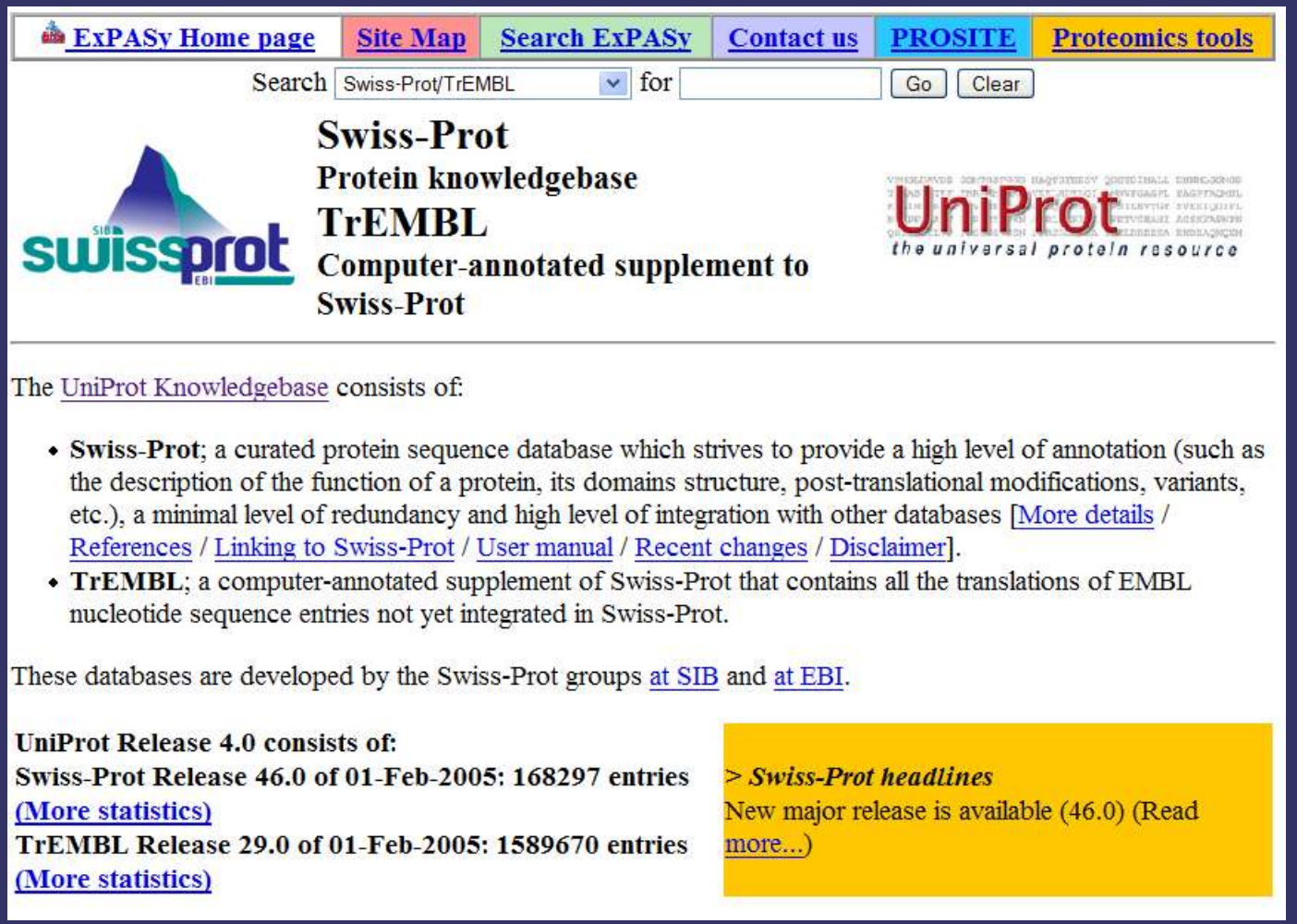


UniProt

Recently, PIR and SwissProt have joined forces to create the **UniProt** database. NIH is the main funding source.

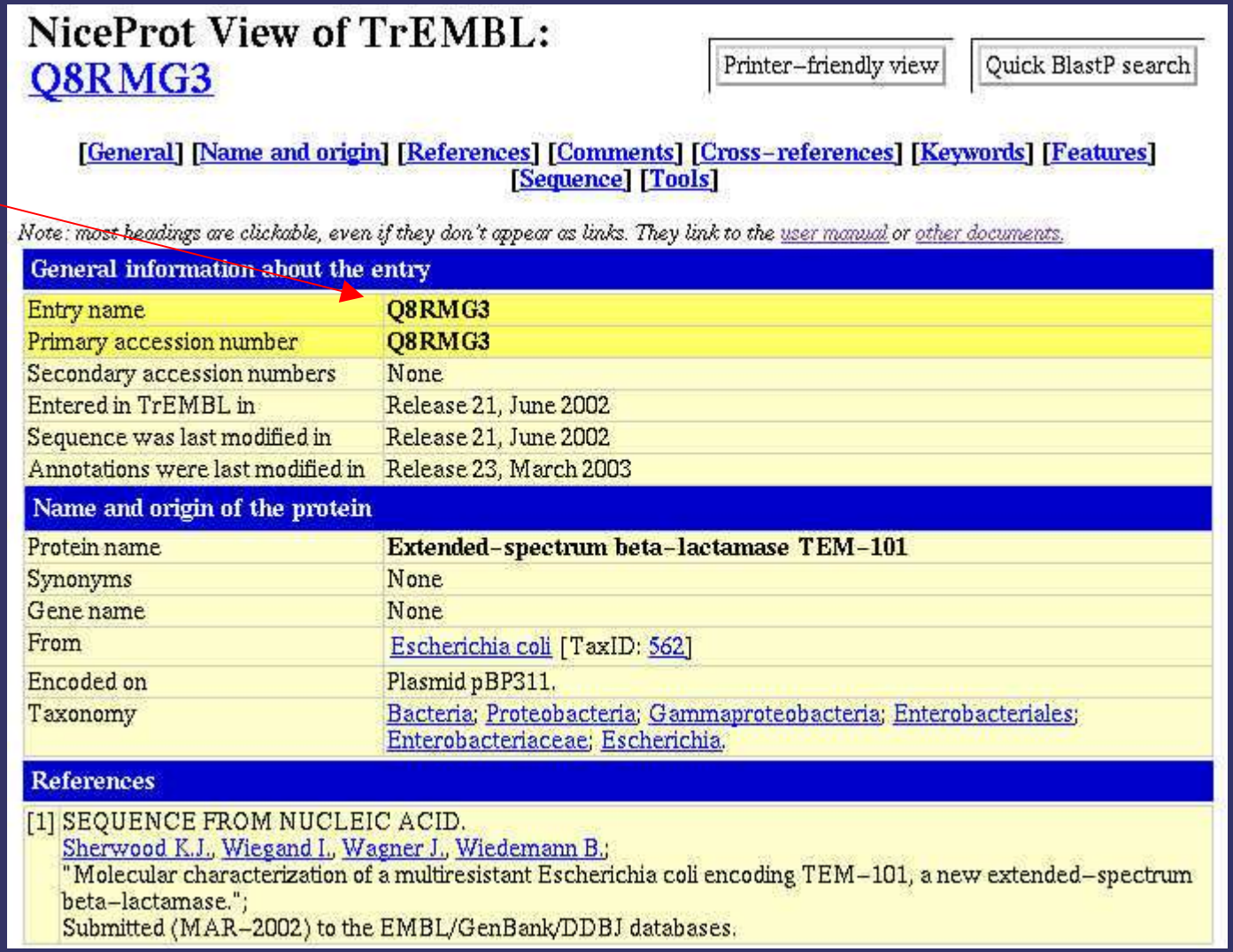
SwissProt AND TrEMBL

Both SwissProt and TrEMBL are part of a server called ExPASy, that also hosts other databases and tools.



Example of SwissProt entry: NiceView format

SwissProt ID



NiceProt View of TrEMBL: Q8RMG3

Printer-friendly view | Quick BlastP search

[General] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

General information about the entry	
Entry name	**Q8RMG3**
Primary accession number	**Q8RMG3**
Secondary accession numbers	None
Entered in TrEMBL in	Release 21, June 2002
Sequence was last modified in	Release 21, June 2002
Annotations were last modified in	Release 23, March 2003
Name and origin of the protein	
Protein name	**Extended-spectrum beta-lactamase TEM-101**
Synonyms	None
Gene name	None
From	Escherichia coli [TaxID: 562]
Encoded on	Plasmid pBP311.
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.

References

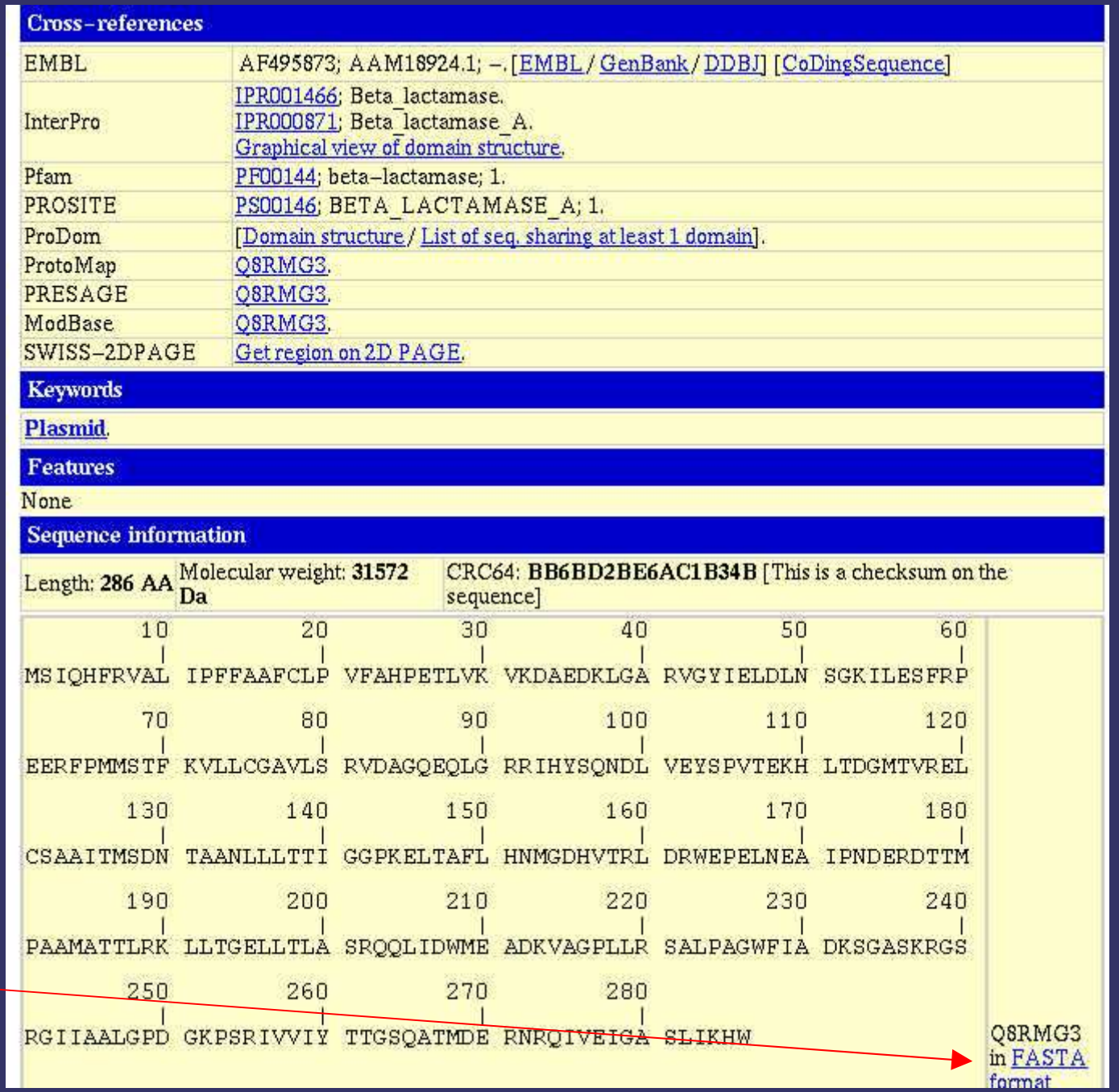
[1] SEQUENCE FROM NUCLEIC ACID.
Sherwood K.J., Wiegand I., Wagner J., Wiedemann B.;
"Molecular characterization of a multiresistant Escherichia coli encoding TEM-101, a new extended-spectrum beta-lactamase.";
Submitted (MAR-2002) to the EMBL/GenBank/DDBJ databases.

NiceView

Cross-references and links

Sequence

Link to FASTA format



Cross-references

EMBL	AF495873; AAM18924.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
InterPro	IPR001466; Beta_lactamase. IPR000871; Beta_lactamase_A. Graphical view of domain structure.
Pfam	PF00144; beta-lactamase; 1.
PROSITE	PS00146; BETA_LACTAMASE_A; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain].
ProtoMap	Q8RMG3.
PRESAGE	Q8RMG3.
ModBase	Q8RMG3.
SWISS-2DPAGE	Get region on 2D PAGE.

Keywords

Plasmid.

Features

None

Sequence information

Length: **286 AA** Molecular weight: **31572 Da** CRC64: **BB6BD2BE6AC1B34B** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MSIQHFRVAL IPFFAAFCLP VFAHPETLVK VKDAEDKLGA RVGYIELDLN SGKILESFRP

        70         80         90        100        110        120
         |          |          |          |          |          |
EERFPMMSTF KVLLCGAVLS RVDAGQEQLG RRIHYSQNDL VEYSPVTEKH LTDGMTVREL

       130        140        150        160        170        180
         |          |          |          |          |          |
CSAAITMSDN TAANLLLTTI GGPKELTAFL HNMGDHVTRL DRWEPELNEA IPNDERDTTM

       190        200        210        220        230        240
         |          |          |          |          |          |
PAAMATTLRK LLTGELLTLA SRQQLIDWME ADKVAGPLLR SALPAGWFIA DKSGASKRGS

       250        260        270        280
         |          |          |          |
RGIIAALGPD GKPSRIVVIY TTGSQATMDE RNRQIVEIGA SLIKHW
```

Q8RMG3 in FASTA format

SRS

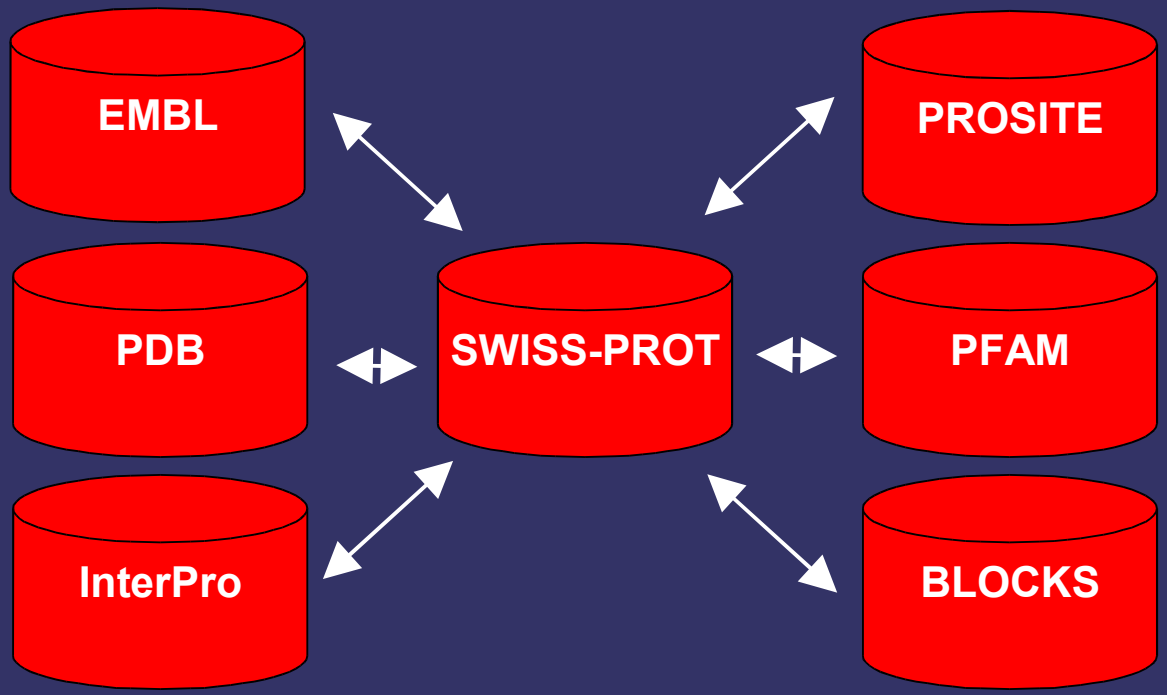
Sequence Retrieval System

SEQUENCE RETRIEVAL SYSTEM

- Developed at different successive institutions.
    - 1990 - Started by Thure Etzold, at the EMBL.
    - 1997 - EBI (Cambridge), financed in part by EMBnet.
    - 1998 - Lion Biosciences.
- It is not a database. It is a data warehouse. SRS uses flat text file versions of many databases: EMBL, Swiss-Prot, MEDLINE, etc., which are copied and indexed, to allow the connection of related information from several databases.
- There are many mirrors installed. The most popular is, probably, the one at the EBI.
- SRS offers access to more than 700 databases.

CONNECTIONS BETWEEN DATABASES

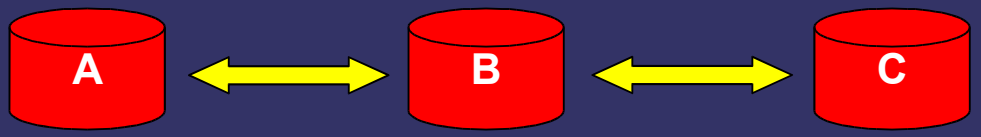
- Information from different databases is connected by means of cross-references.
- Cross-references are entry attributes that make reference to entries in other databases.
- Not all databases include cross-references.
  - SwissProt and EMBL have fields called DR, which includes cross references to most other databases.
  - The same SwissProt and EMBL include a field called RX, which includes cross-references to Medline and PUBMed.



CONNECTIONS BETWEEN DATABASES

DIRECT LINKING BETWEEN A AND B

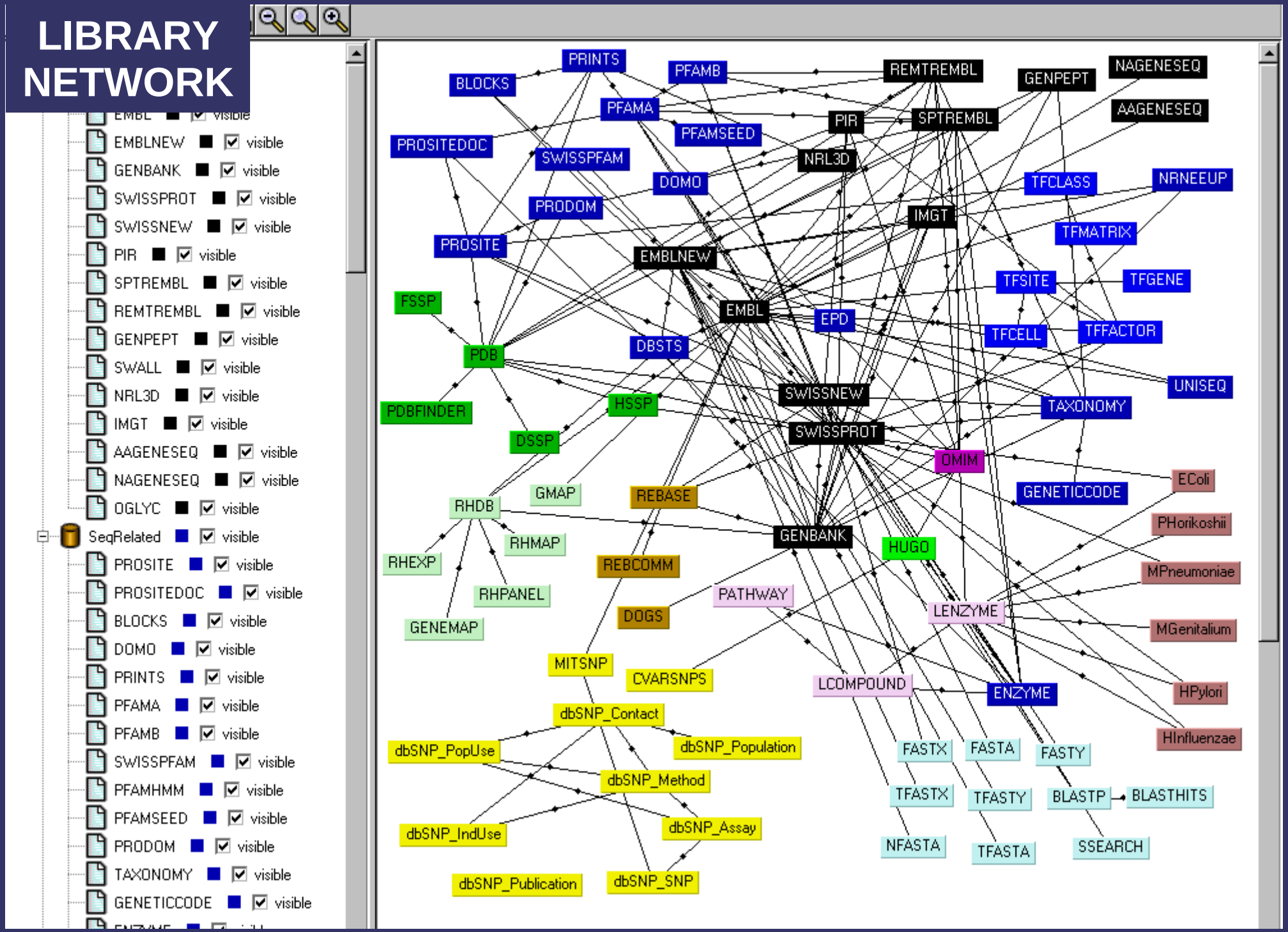
DIRECT LINKING BETWEEN C AND D



Indirect link betrween A and C

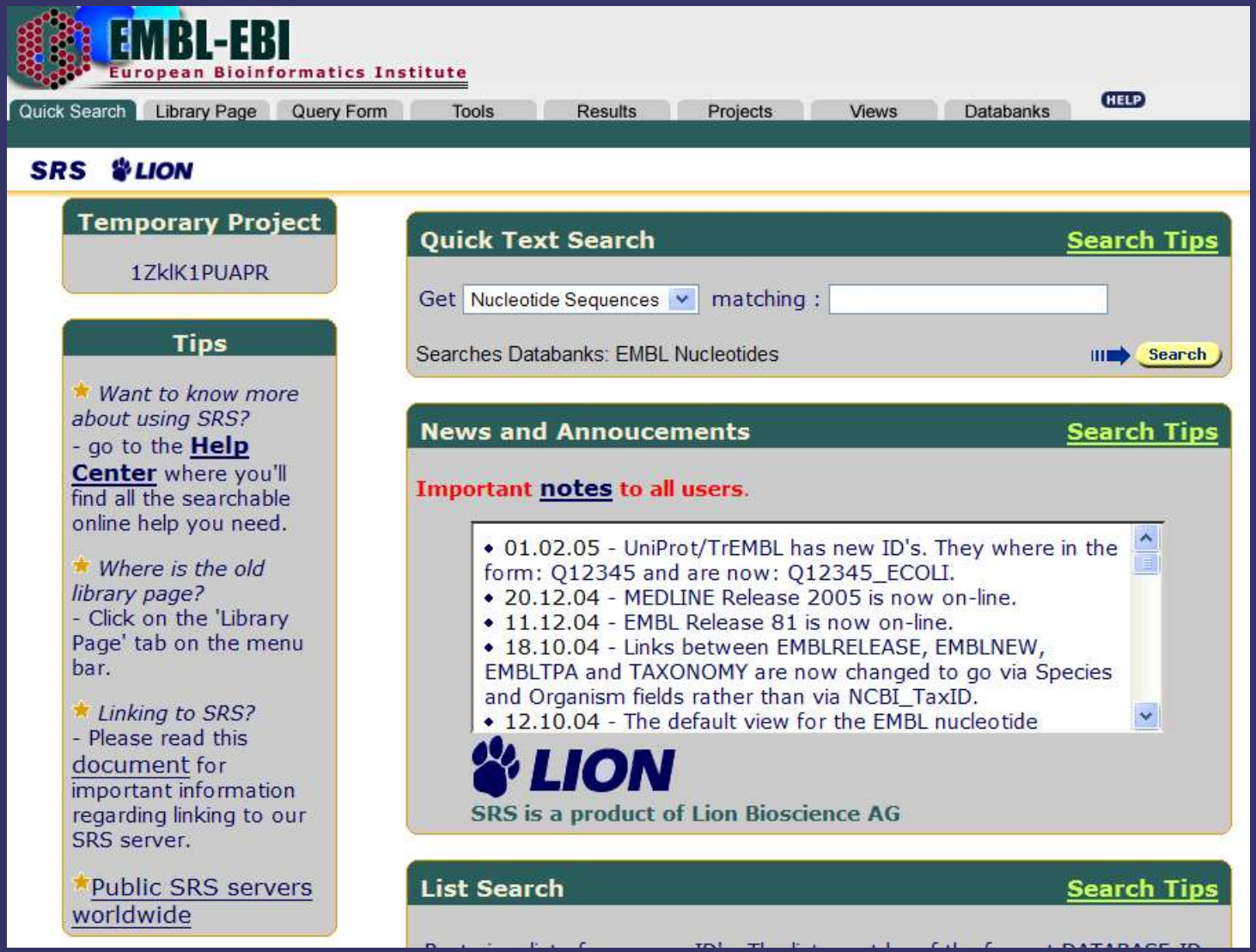
Links are established bidirectionally

LIBRARY NETWORK



APPLICATIONS IN SRS

- In addition to giving access to the information from many databases, SRS is also a platform executing bioinformatics applications. Among them:
  - BLAST,several versions.
  - FASTA, several versions.
  - PrositeSearch, motif search
  - ClustalW, multiple alignments


EMBL-EBI
European Bioinformatics Institute
Quick Search
Library Page
Query Form
Tools
Results
Projects
Views
Databanks
HELP
SRS
LION
Temporary Project
1ZklK1PUAPR
Tips
Want to know more about using SRS?
- go to the Help Center where you'll find all the searchable online help you need.
Where is the old library page?
- Click on the 'Library Page' tab on the menu bar.
Linking to SRS?
- Please read this document for important information regarding linking to our SRS server.
Public SRS servers worldwide
Quick Text Search
Search Tips
Get Nucleotide Sequences matching :
Searches Databanks: EMBL Nucleotides
Search
News and Annoucements
Search Tips
Important notes to all users.
• 01.02.05 - UniProt/TrEMBL has new ID's. They where in the form: Q12345 and are now: Q12345_ECOLI.
• 20.12.04 - MEDLINE Release 2005 is now on-line.
• 11.12.04 - EMBL Release 81 is now on-line.
• 18.10.04 - Links between EMBLRELEASE, EMBLNEW, EMBLTPA and TAXONOMY are now changed to go via Species and Organism fields rather than via NCBI_TaxID.
• 12.10.04 - The default view for the EMBL nucleotide
LION
SRS is a product of Lion Bioscience AG
List Search
Search Tips

SEARCHES: MAIN PAGE

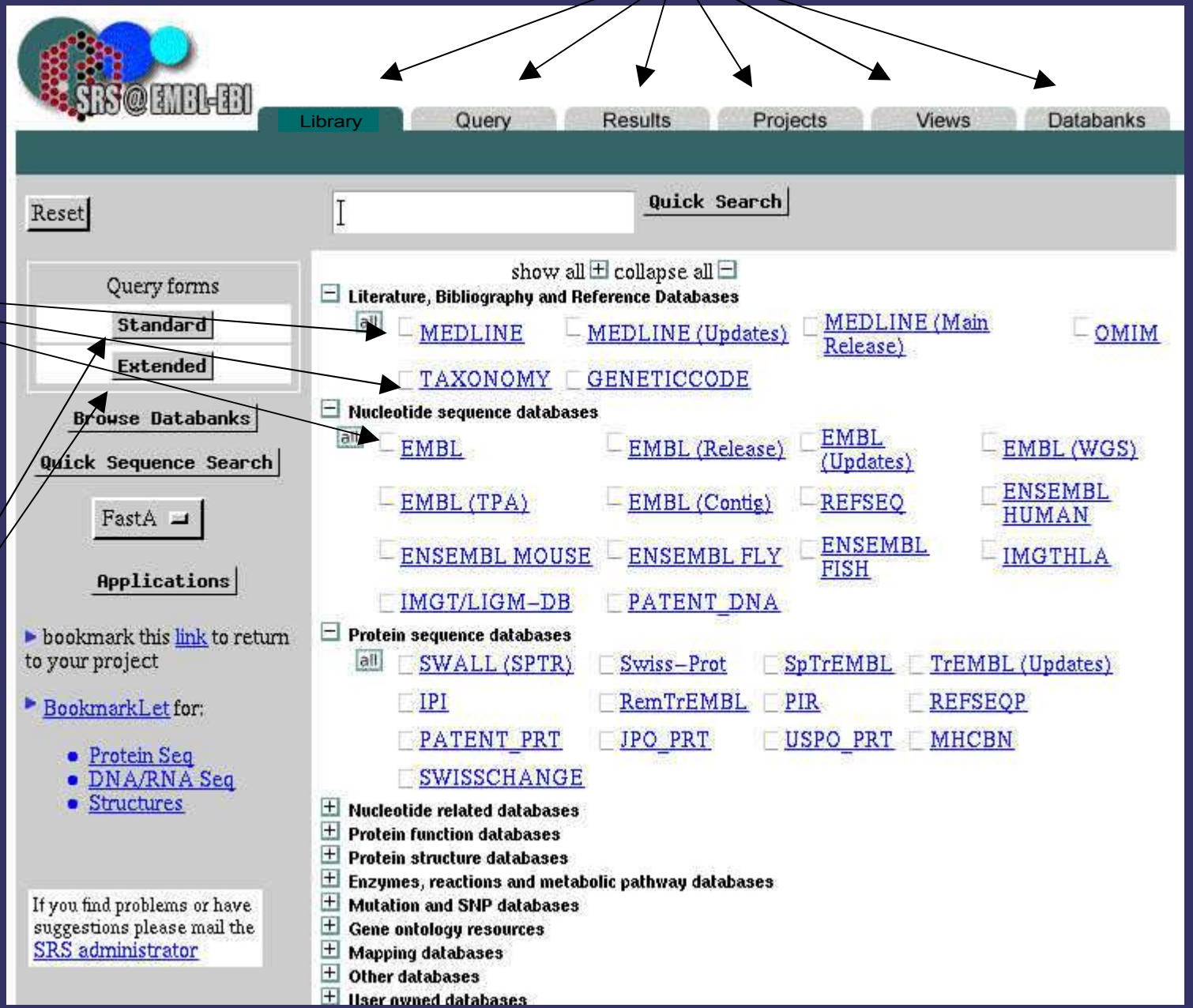
Workbenches

First step: choose the database(s) to search, at the Library tab

Libraries

Query Forms

Second step: load the form: standard or advanced

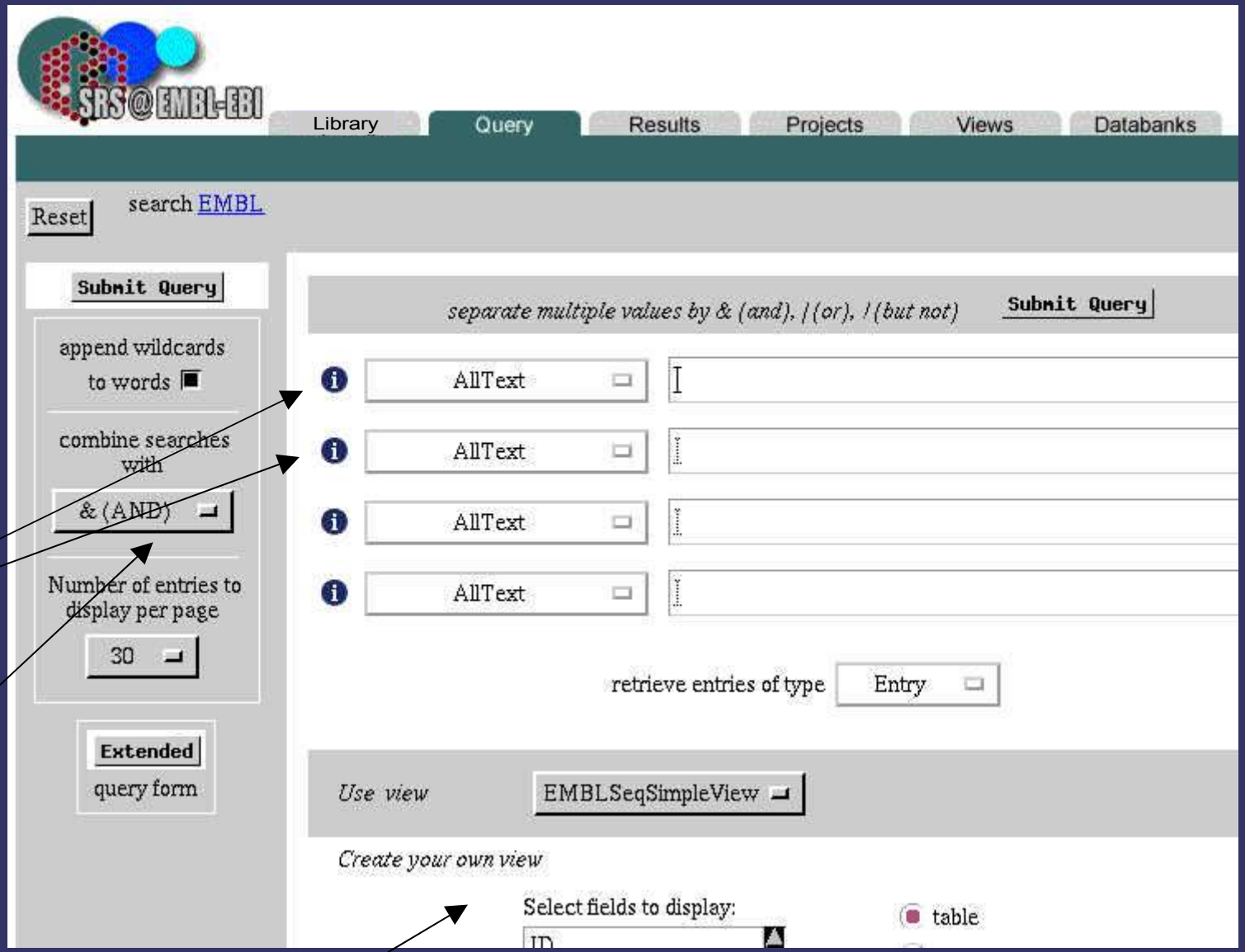


Library groups

SEARCHES: STANDARD QUERY FORM

Third step: enter the terms and specify the fields for the query, at the Query tab.

At this tab it is possible also to customize the output.



Query Fields

Operadores and (&), or (|), butnot (!)

Field Selection for displaying the results

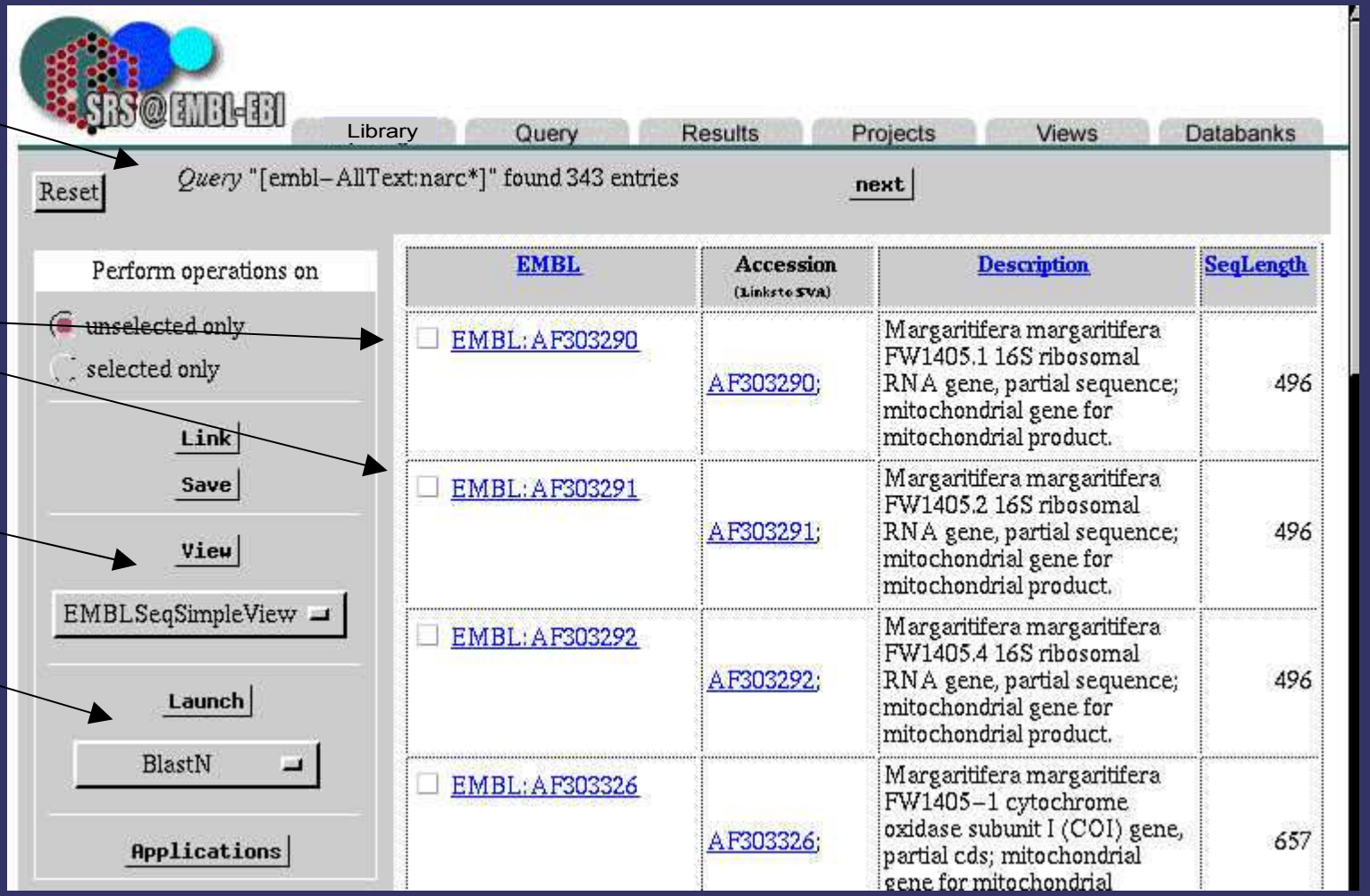
RESULTS: HIT LIST

SRS Query

HITS

Different options to View

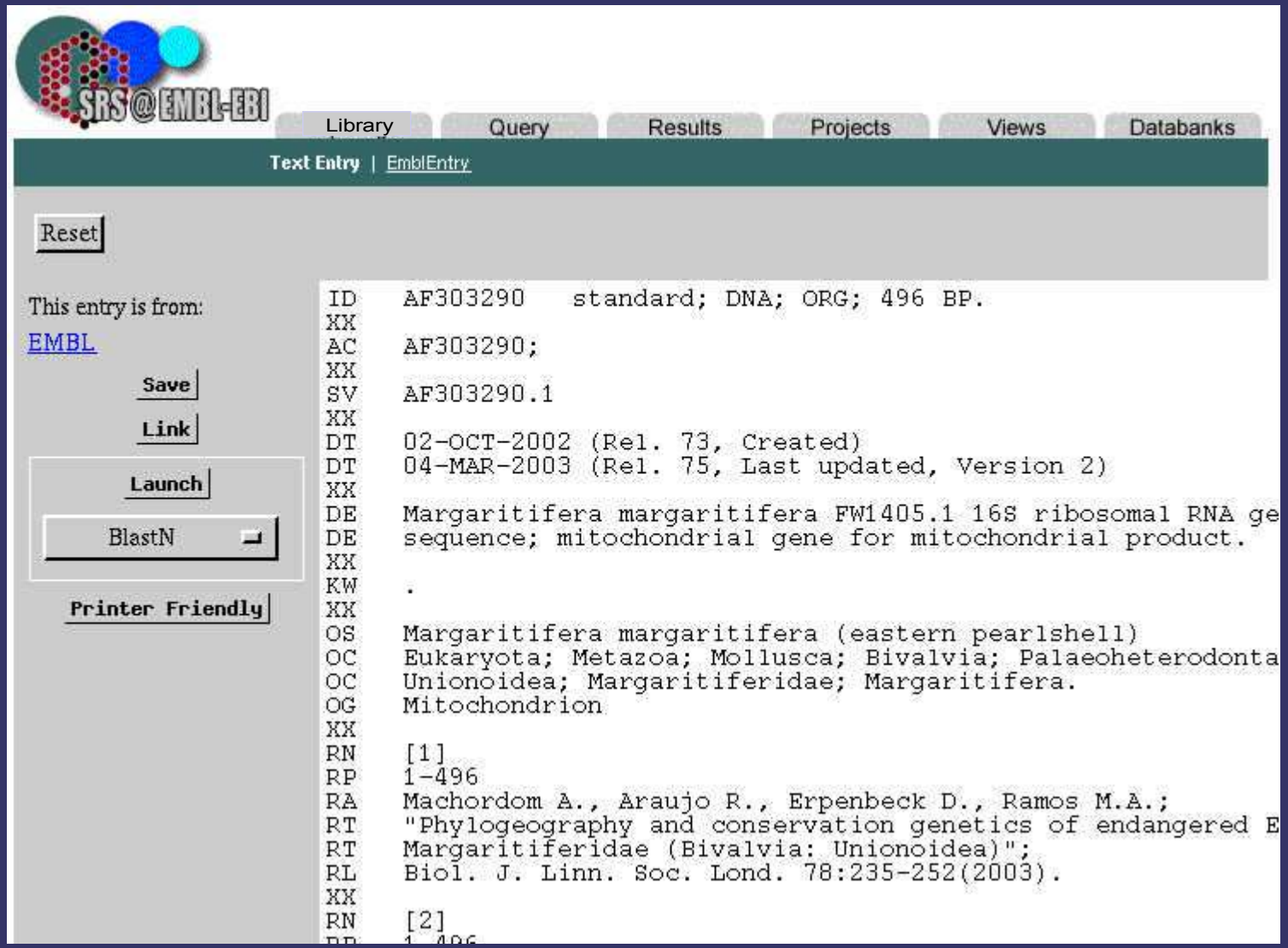
Application Launcher



To make new searches, without changing the database, go back to the tab Query.

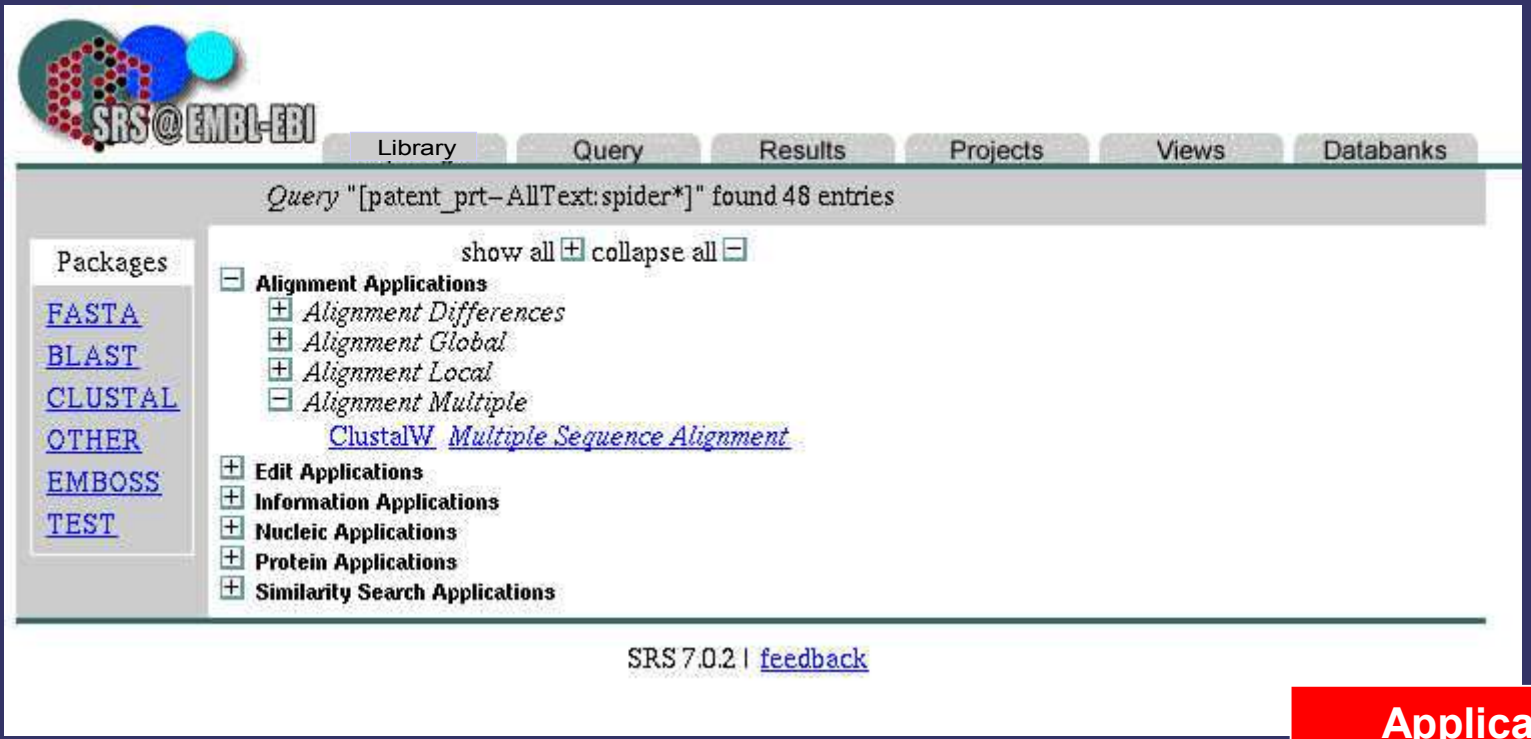
To change database, go to the tab Library.

RESULTS: COMPLETE ENTRY VIEW



Entry View

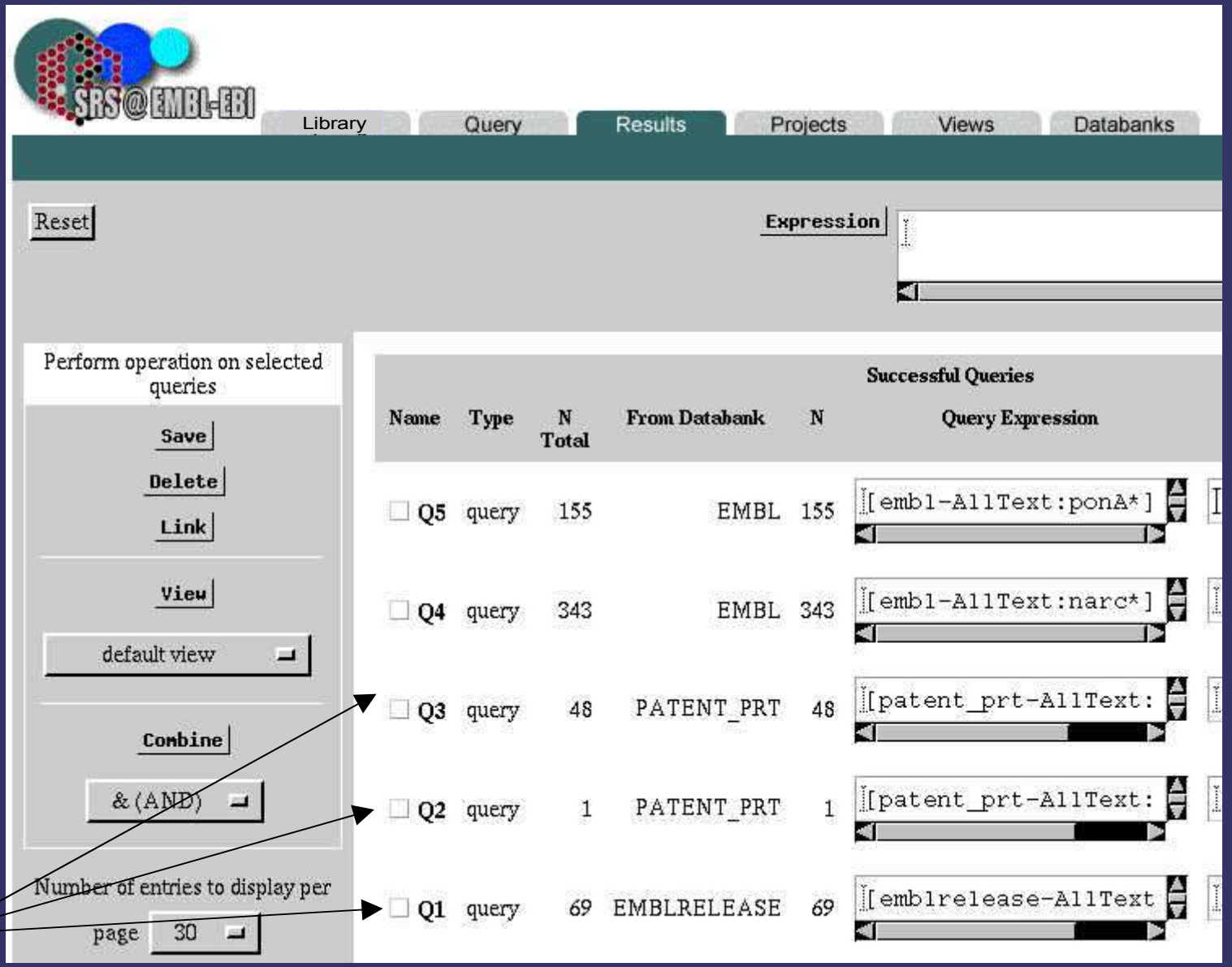
APPLICATIONS



Application Launcher

RESULTS

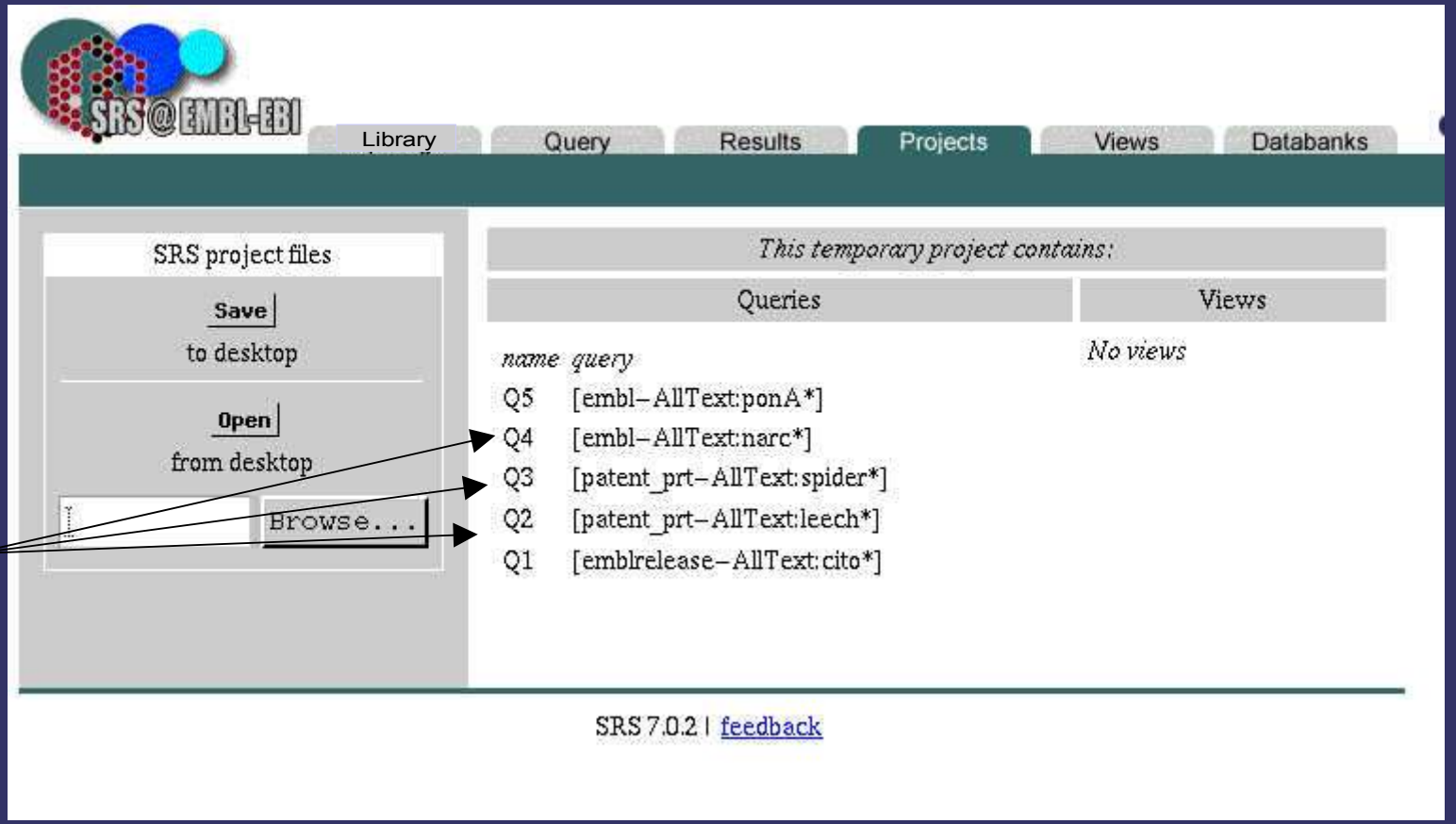
At the tab Results you will find a list of the queries and processes that you have been running in the current session. It will be possible to retrieve again the results, or to combine them with Boolean operators.



My queries

PROJECTS

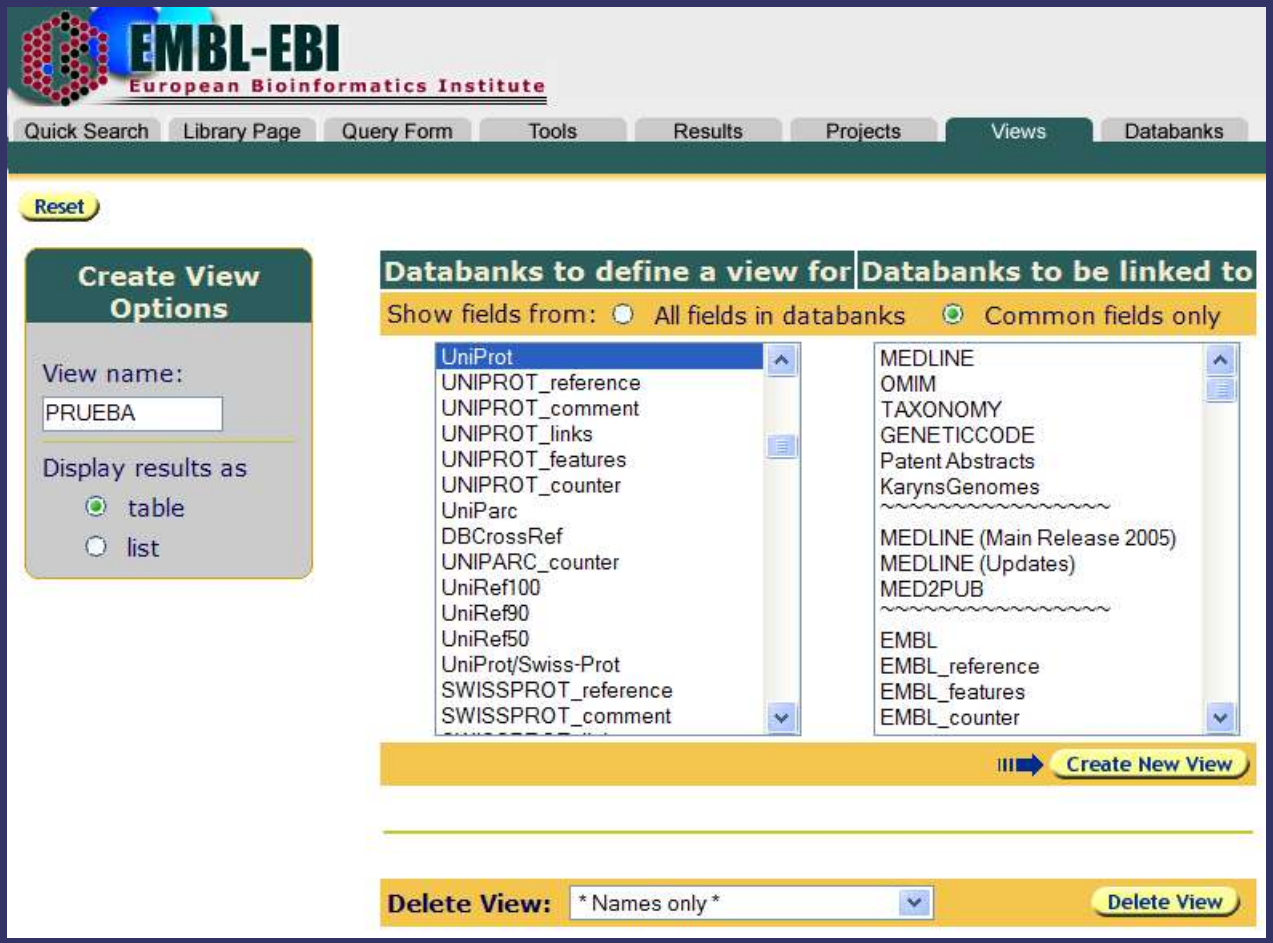
My queries



At the tab Projects you will have the option of saving the queries and processes that you have been running in the current session, that be will listed as an index. The project will be saved locally, as a compressed file.

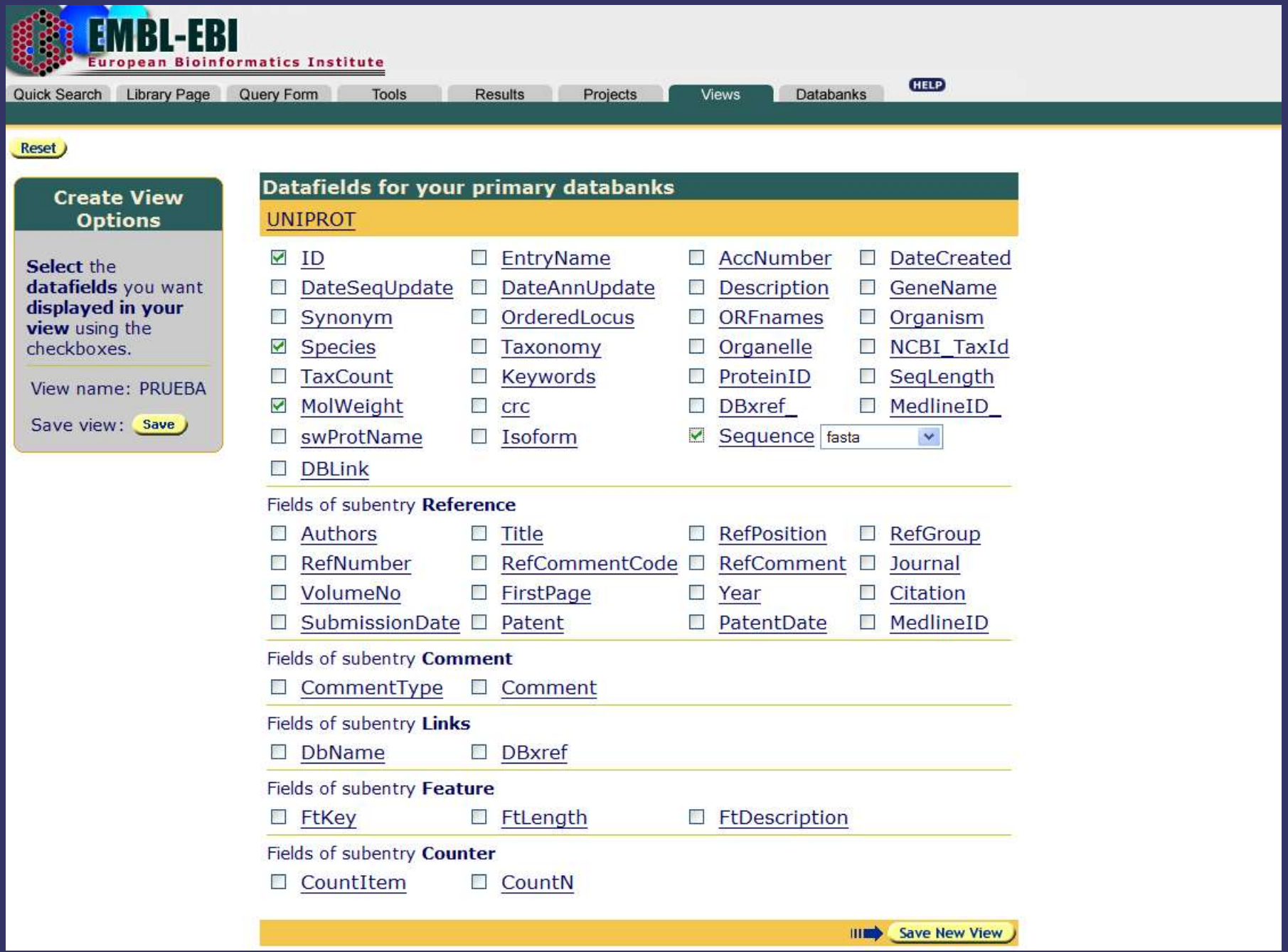
At this tab you have also the option of opening a previously saved project.

VIEWS

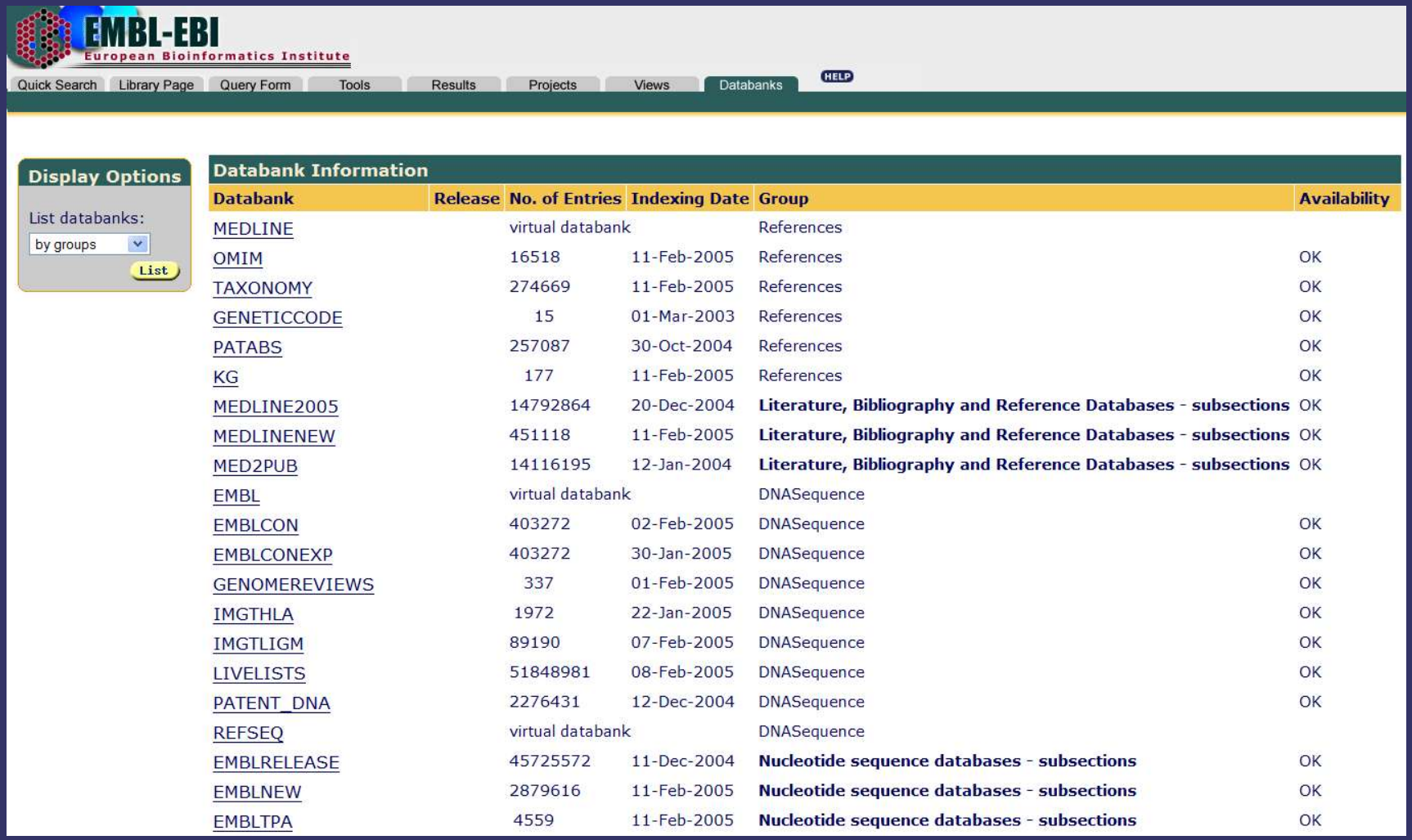


At the tab Views you have the option of creating new views for displaying the results. First, give a name to the new view; second, choose the database(s) from wich data will be displayed; third (next slide), choose the fields that you want to have displayed; finally, save the new view.

VIEWS


EMBL-EBI
European Bioinformatics Institute
Quick Search
Library Page
Query Form
Tools
Results
Projects
Views
Databanks
HELP
Reset
Create View Options
Select the datafields you want displayed in your view using the checkboxes.
View name: PRUEBA
Save view: Save
Datafields for your primary databanks
UNIPROT
ID
EntryName
AccNumber
DateCreated
DateSeqUpdate
DateAnnUpdate
Description
GeneName
Synonym
OrderedLocus
ORFnames
Organism
Species
Taxonomy
Organelle
NCBI_TaxId
TaxCount
Keywords
ProteinID
SeqLength
MolWeight
crc
DBxref_
MedlineID_
swProtName
Isoform
Sequence
fasta
DBLink
Fields of subentry Reference
Authors
Title
RefPosition
RefGroup
RefNumber
RefCommentCode
RefComment
Journal
VolumeNo
FirstPage
Year
Citation
SubmissionDate
Patent
PatentDate
MedlineID
Fields of subentry Comment
CommentType
Comment
Fields of subentry Links
DbName
DBxref
Fields of subentry Feature
FtKey
FtLength
FtDescription
Fields of subentry Counter
CountItem
CountN
Save New View

DATABANKS



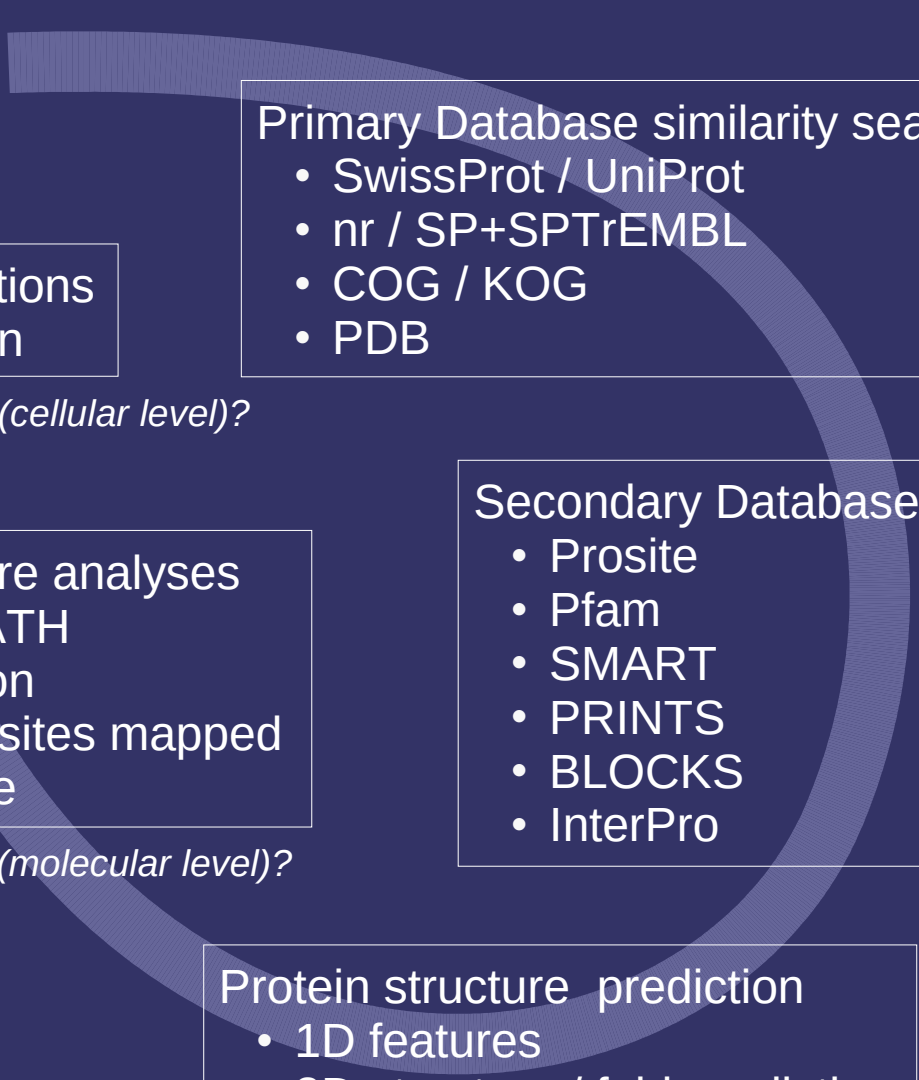
Databank Information

Databank	Release	No. of Entries	Indexing Date	Group	Availability
MEDLINE		virtual databank		References	
OMIM		16518	11-Feb-2005	References	OK
TAXONOMY		274669	11-Feb-2005	References	OK
GENETICCODE		15	01-Mar-2003	References	OK
PATABS		257087	30-Oct-2004	References	OK
KG		177	11-Feb-2005	References	OK
MEDLINE2005		14792864	20-Dec-2004	**Literature, Bibliography and Reference Databases - subsections**	OK
MEDLINENEW		451118	11-Feb-2005	**Literature, Bibliography and Reference Databases - subsections**	OK
MED2PUB		14116195	12-Jan-2004	**Literature, Bibliography and Reference Databases - subsections**	OK
EMBL		virtual databank		DNASequence	
EMBLCON		403272	02-Feb-2005	DNASequence	OK
EMBLCONEXP		403272	30-Jan-2005	DNASequence	OK
GENOMEREVIEWS		337	01-Feb-2005	DNASequence	OK
IMGTHLA		1972	22-Jan-2005	DNASequence	OK
IMGTLIGM		89190	07-Feb-2005	DNASequence	OK
LIVELISTS		51848981	08-Feb-2005	DNASequence	OK
PATENT_DNA		2276431	12-Dec-2004	DNASequence	OK
REFSEQ		virtual databank		DNASequence	
EMBLRELEASE		45725572	11-Dec-2004	**Nucleotide sequence databases - subsections**	OK
EMBLNEW		2879616	11-Feb-2005	**Nucleotide sequence databases - subsections**	OK
EMBLTPA		4559	11-Feb-2005	**Nucleotide sequence databases - subsections**	OK

At the tab Databanks you will visualize the databases that are currently indexed in SRS.

FUNCTION PREDICTION PROTOCOL

Based on sequence similarity, structural analyses and information about interacting partners.



Protein primary sequence

Primary Database similarity search
- SwissProt / UniProt
- nr / SP+SPTrEMBL
- COG / KOG
- PDB

Orthologs / paralogs
MSA
Family discovery or assignment
Functional residues
Phylogenetic profile
Gene neighbourhood
Function prediction?

Protein interactions characterization

Function prediction (cellular level)?

Secondary Database similarity search
- Prosite
- Pfam
- SMART
- PRINTS
- BLOCKS
- InterPro

Protein motifs
Domain organization
Family assignment
Function prediction?

Protein structure analyses
- SCOP / CATH classification
- Functional sites mapped on structure

Function prediction (molecular level)?

Protein structure prediction
- 1D features
- 3D structure / fold prediction

Known / Predicted structure

This presentation contains material from:

Rodrigo Lopez, EBI