

# 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.

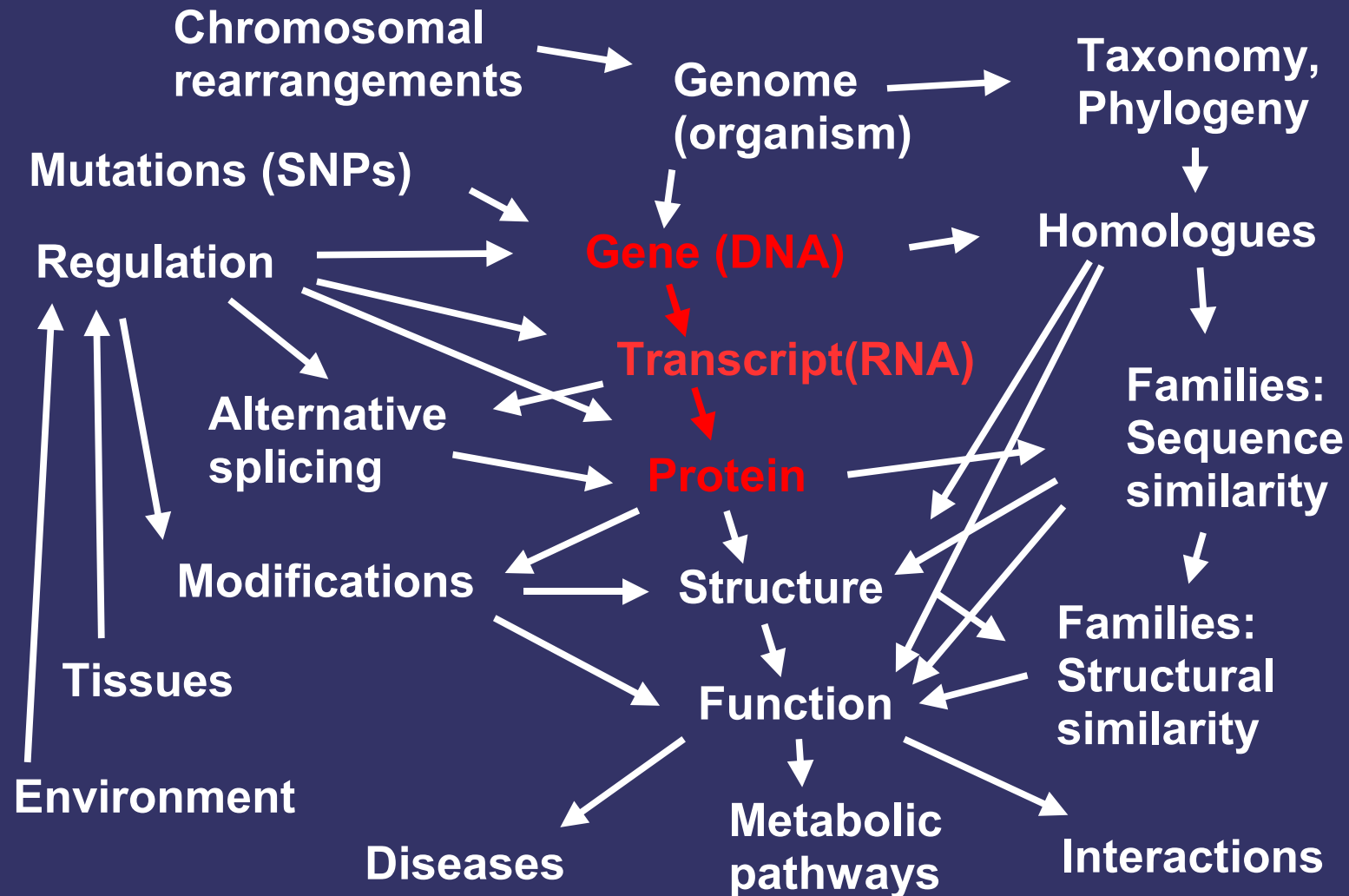
For example:

- 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 grouped 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.



Search GO:

[Terms](#) [Gene Products](#)

[Top Docs](#) [Gene Ontology](#) [GO Links](#) [GO Summary](#)

- [GO:0003673 : Gene Ontology \(46199\)](#)
- [GO:0008150 : biological process \(30188\)](#)
  - [GO:0007610 : behavior \(291\)](#)
    - [GO:0000004 : biological process unknown \(3665\)](#)
  - [GO:0007](#)
  - [GO:0008](#)
  - [GO:0016](#)
  - [GO:0007](#)
  - [GO:0008](#)
  - [GO:0007](#)
  - [GO:0016](#)
- [GO:0005575](#)
- [GO:0003674](#)

# Gene Ontology BROWSER

- [GO:0003673 : Gene Ontology \(46199\)](#)
  - [GO:0008150 : biological process \(30188\)](#)
    - [GO:0007610 : behavior \(291\)](#)
      - [GO:0000004 : biological process unknown \(3665\)](#)
    - [GO:0007154 : cell communication \(6212\)](#)
    - [GO:0008151 : cell growth and/or maintenance \(20547\)](#)
    - [GO:0016265 : death \(525\)](#)
    - [GO:0007275 : development \(3620\)](#)
    - [GO:0008371 : obsolete \(1640\)](#)
    - [GO:0007582 : physiological processes \(854\)](#)
    - [GO:0016032 : viral life cycle \(27\)](#)
  - [GO:0005575 : cellular component \(22371\)](#)
  - [GO:0003674 : molecular function \(37018\)](#)

# Gene Ontology

Mycoplasma pneumoniae proteome.  
Distribution of GO anotations.

GO Classification for <i>M. pneumoniae</i>		
Term	Proteins	
<b>GO:0003674</b> <b>molecular_function</b>	<b>431</b>	<b>62.7%</b>
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%
<b>GO:0008150</b> <b>biological_process</b>	<b>343</b>	<b>49.9%</b>
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%
<b>GO:0005575</b> <b>cellular_component</b>	<b>277</b>	<b>40.3%</b>
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 and 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.

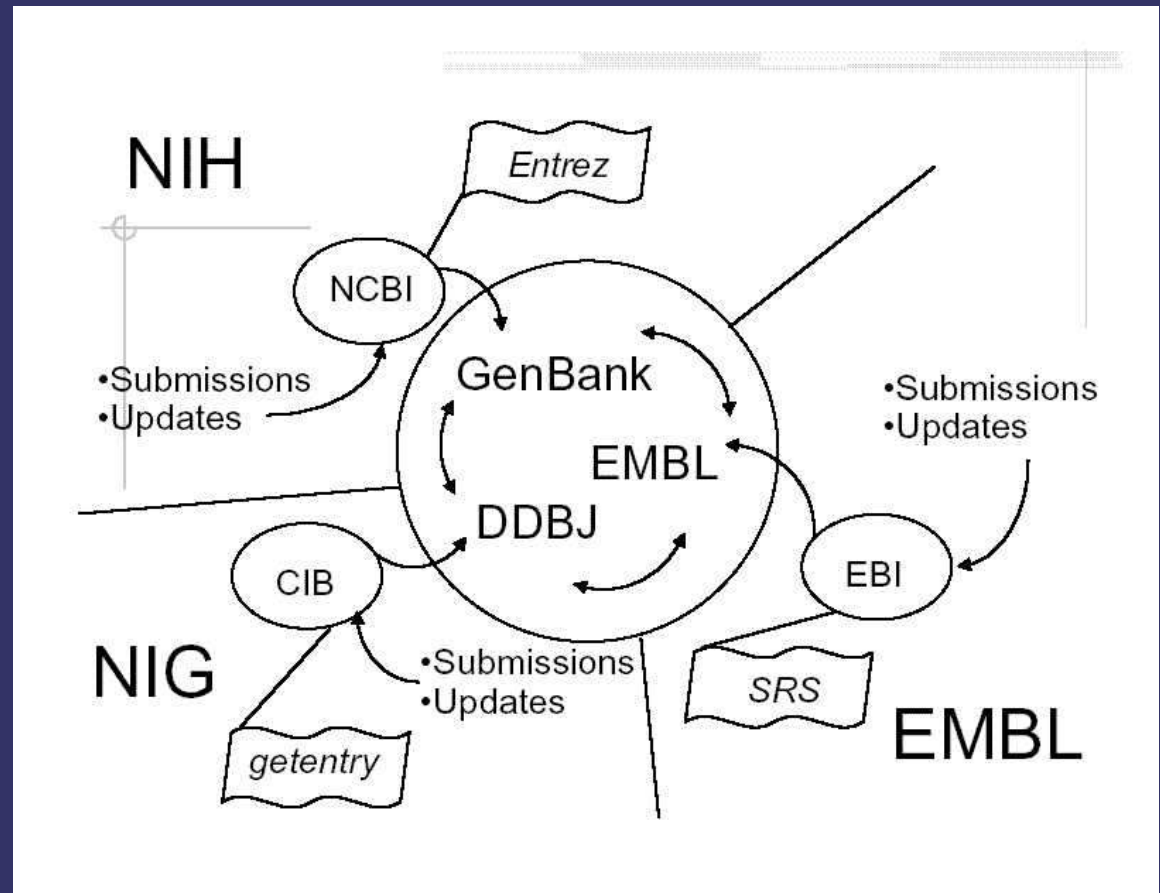
# GenBank, EMBL AND DDBJ

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](http://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

## Functional 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

The screenshot displays the NCBI main page with the following elements:

- Header:** NCBI logo and "National Center for Biotechnology Information" title, with "National Library of Medicine" and "National Institutes of Health" as sub-headers.
- Navigation Menu:** PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Structure.
- Search Bar:** A search input field with a dropdown menu set to "Nucleotide" and a "Go" button.
- Left Sidebar:**
  - SITE MAP:** Guide to NCBI resources
  - About NCBI:** The science behind our resources. An introduction for researchers, educators and the public.
  - GenBank:** Sequence submission support and software.
  - Literature databases:** PubMed, OMIM, Books and PubMed Central
  - Genomic biology:** The human genome, whole genomes and related resources
  - Tools:** Data mining
  - Research at NCBI:** People, projects and
- Main Content:**
  - What does NCBI do?:** Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information – all for the better understanding of molecular processes affecting human health and disease. [More...](#)
  - PubMed Central:** An archive of life sciences journals
    - Free fulltext
    - 80,000 articles from over 100 journals
    - Linked to PubMed and fully searchableUse of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.
  - Hot Spots:**
    - Cancer genome anatomy project
    - Clusters of orthologous groups
    - Coffee Break
    - Electronic PCR
    - Gene expression omnibus
    - Genes and disease
    - Human genome resources
    - Human/mouse homology maps
    - LocusLink

# 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

NCBI

Entrez  
search and retrieval system

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search PubMed for [ ] Go Clear

Limits Preview/Index History Clipboard

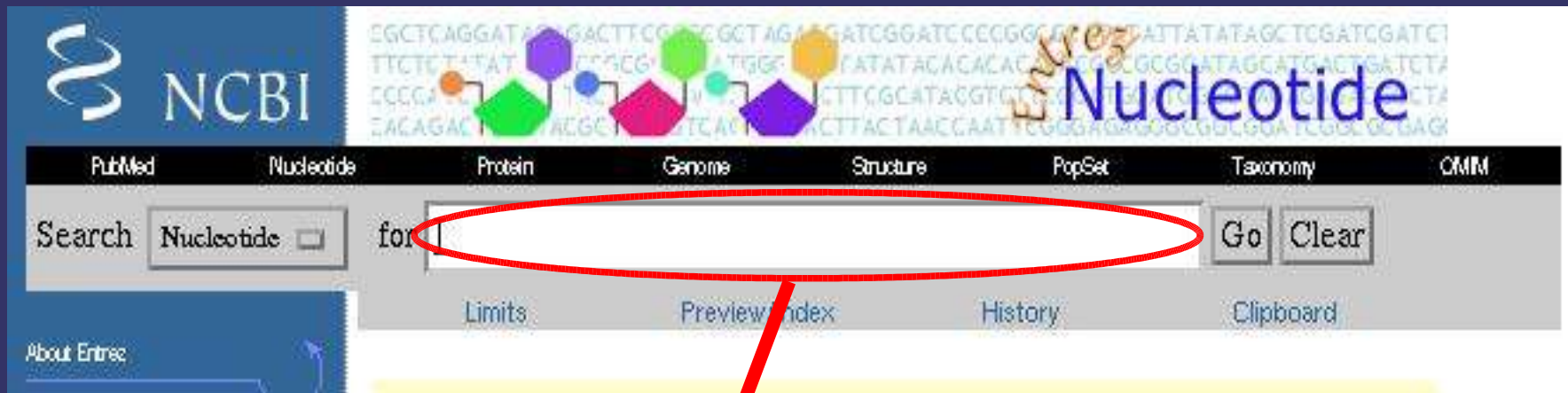
Entrez is a retrieval system for searching several linked databases. It provides access to:

- [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](#): BookShelf online books
- [ProbeSet](#): gene expression and microarray datasets
- [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
- NEW** [PMC](#): full-text digital archive of life sciences journal literature
- NEW** [NCBI Web Site](#): NCBI Web site search

Pre-computed similarity searches are available for most database records, which produce a list of related sequences, structure neighbors, as well as related articles.

[NCBI's Protein Sequence Information Survey Results](#)

# TEXT QUERIES



Word searches: RNA

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

Phrase searching: "16s RNA"

Field restriction: Protein [AUTH]

# QUERY BY CONTENT

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. At the top, the NCBI logo is on the left, and the text "National Center for Biotechnology Information" is centered, with "National Library of Medicine" and "National Institutes of Health" below it. A navigation bar contains links for PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Structure. The BLAST link is circled in red. Below the navigation bar is a search box with a dropdown menu set to "Nucleotide" and a "Go" button. The main content area is divided into several sections: "What does NCBI do?" (Established in 1988 as a national resource for molecular biology information...), "Hot Spots" (a list of projects like Cancer genome anatomy project, Clusters of orthologous groups, etc.), and "PubMed Central" (An archive of life sciences journals, featuring free fulltext, 80,000 articles from over 100 journals, and being linked to PubMed and fully searchable). A sidebar on the left contains links for "SITE MAP", "About NCBI", "GenBank", "Literature databases", "Genomic biology", "Tools", and "Research at NCBI".

**NCBI**  
National Center for Biotechnology Information  
National Library of Medicine National Institutes of Health

PubMed Entrez **BLAST** OMIM Books TaxBrowser Structure

Search Nucleotide for  Go

**SITE MAP**  
Guide to NCBI resources

**About NCBI**  
The science behind our resources. An introduction for researchers, educators and the public.

**GenBank**  
Sequence submission support and software

**Literature databases**  
PubMed, OMIM, Books and PubMed Central

**Genomic biology**  
The human genome, whole genomes and related resources

**Tools**  
Data mining

**Research at NCBI**  
People, projects and

**What does NCBI do?**

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information – all for the better understanding of molecular processes affecting human health and disease. [More...](#)

**Hot Spots**

- ▶ Cancer genome anatomy project
- ▶ Clusters of orthologous groups
- ▶ Coffee Break
- ▶ Electronic PCR
- ▶ Gene expression omnibus
- ▶ Genes and disease
- ▶ Human genome resources
- ▶ Human/mouse homology maps
- ▶ LocusLink

**PubMed Central**  
*An archive of life sciences journals*

- **Free fulltext**
- **80,000 articles from over 100 journals**
- **Linked to PubMed and fully searchable**

Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.

# HIT LIST

The screenshot shows the NCBI Entrez Nucleotide search interface. At the top, the NCBI logo is on the left, and a decorative banner with the word 'Entrez' and 'Nucleotide' is on the right. Below the banner is a navigation bar with tabs for PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, OMIM, and Books. The search bar contains the text 'bacillus subtilis bofC' and has 'Go' and 'Clear' buttons. Below the search bar are links for Limits, Preview/Index, History, Clipboard, and Details. The results section shows a 'Display' dropdown set to 'Summary', a 'Show:' dropdown set to '20', and a 'Send to' dropdown set to 'Text'. Below this is a summary bar indicating 'Items 1-4 of 4' and 'One page.' The results list contains four entries, each with a checkbox, an accession number, a description, and a 'Links' link.

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

Display Summary Show: 20 Send to Text

Items 1-4 of 4 One page.

- 1: [NC\\_000964](#) Links  
Bacillus subtilis, complete genome  
gil16077068|refl|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  
gil6977794|emb|Y15896.1|BSY15896[6977794]
- 3: [Z99118](#) Links  
Bacillus subtilis complete genome (section 15 of 21): from 2795131 to 3013540  
gil2635200|emb|Z99118.1|BSUB0015[2635200]
- 4: [X93081](#) Links  
B.subtilis bofC, orf1, csbX, and orf4 genes  
gil1941915|emb|X93081.1|BSBOFCGEN[1941915]

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 FORMATS

NCBI Entrez Nucleotide

Search  for

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

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

Reference sequence project

Submit to GenBank

Display	Summary	Show:	Send to	Text
	Brief	20		
	ASN.1			
	FASTA			
1: <a href="#">NC</a>	B: XML gi GenBank			One page.
	GI list			Links
2: <a href="#">Y15</a>	B: LinkOut yr Nucleotide Neighbors gi ProbeSet Links			Links
	OMIM Links			Links
3: <a href="#">Z99</a>	B: PMC Links gi PopSet Links			Links
	Protein Links			Links
4: <a href="#">X93</a>	B: PubMed Links gi SNP Links Structure Links Taxonomy Links UniGene Links UniSTS Links			Links

Revised: July 5, 2002.

# GenBank FORMAT

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Nucleotide for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Get Subsequence

1: X93081. *B.subtilis* bofC, ...[gi:1941915] Links

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.

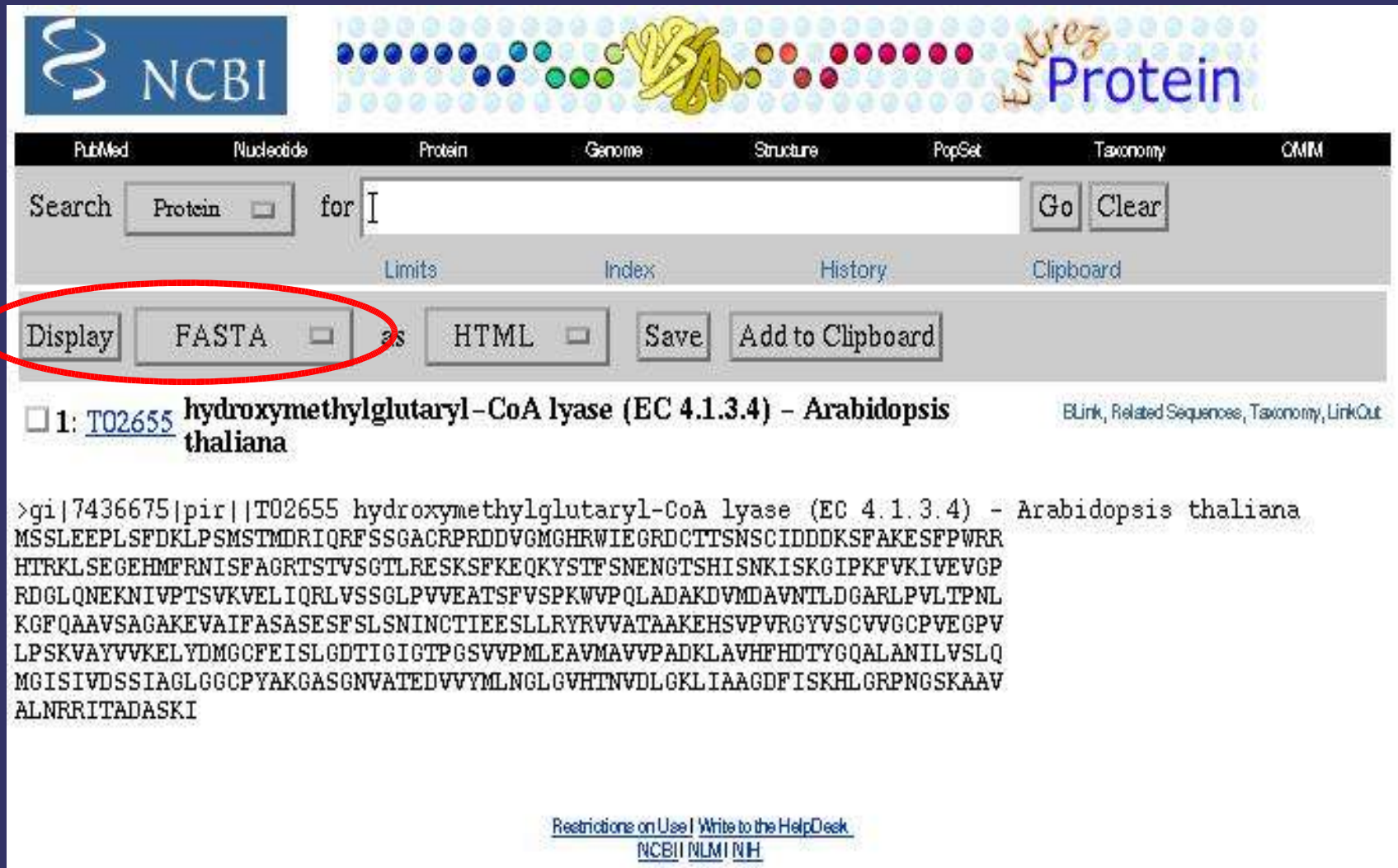
TITLE Direct Submission

JOURNAL Submitted (14-NOV-1995) S.M. Cutting, Dept. of Microbiology, University of Pennsylvania School of Medicine, 346 Johnson Pavillion, 3610 Hamilton Walk, Philadelphia, PA 19104-6076, USA

Fields



# DISPLAY RESULTS: FASTA format



NCBI Entrez Protein

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search Protein for [ ] Go Clear

Limits Index History Clipboard

Display FASTA as HTML Save Add to Clipboard

1: [T02655](#) **hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4) - Arabidopsis thaliana** [ELink](#), [Related Sequences](#), [Taxonomy](#), [LinkOut](#)

```
>gi|7436675|pir||T02655 hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4) - Arabidopsis thaliana
MSSLEEPLSFDKLPMSMTMDRIQRFSSGACRPRDDVGMGHRWIEGRDCITTSNSCIDDDKSFakesFPWRR
HTRKLEGEHMFRNISFAGRTSTVSGTLRESKSFKEQKYSTFSNENGTSHISNKISKGIPKFKIVEVGP
RDGLQNEKNIVPTSVKVELIQRLVSSGLPVVEATSFVSPKWVPLADAKDVM DAVNTLDGARLPVLT PNL
KGFQAAVSAGAKEVAIFASASESFLSNINCTIEESLLRYRVVATAAKEHSVPVVRGYVSCVVGCPVEGPV
LP SKVAYVVKELYDMGCFEISLGDITIGIGTPGSVVPML EAVMAVVPADKLA VHFHDTY GQALANILVSLQ
MGISIVDSSIAGLGGCPYAKGASGNVATEDVVYMLNGLGVHTNVDLGKLI AAGDFISKHLGRPNGSKAAV
ALNRRITADASKI
```


[Restrictions on Use](#) | [Write to the Help Desk](#)  
NCBI | NLM | NIH

# DISPLAY RESULTS: GRAPHICS

1: [T02655](#) **hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4) - Arabidopsis thaliana** [ELink](#), [Related Sequences](#), [Taxonomy](#), [LinkOut](#)

[Hide Toolbar](#)

CDS with gene and mRNA  Hide sequence



**Legend:**  
— - protein

**Sequence:**

```
1  MSSLEEPLSF  DKLPSMSTMD  RIQRFSSGAC  RPRDDVGMGH  RWIEGRDCTT  SNSCIDDDKS  hydroxymethylglutar
61  FAKESFPWRR  HTRKLSEGEH  MFRNISFAGR  TSTVSGTLRE  SKSFKEQKYS  TFSNENGTSH  hydroxymethylglutar
121 ISNKISKGIP  KFKVIVEVGP  RDGLQNEKNI  VPTSUKVELI  QRLVSSGLPV  VEATSFVSPK  hydroxymethylglutar
181  WVPQLADAKD  VMDAVNTLDG  ARLPVLTPNL  KGFQAAVSAG  AKEVAIFASA  SESFSLSNIN  hydroxymethylglutar
241  CTIEESLLRY  RVVATAAKEH  SVPVRGYVSC  VVGCPVEGPV  LPSKVAYVVK  ELYDMCCFEI  hydroxymethylglutar
301  SLGDTIGIGT  PGSVVPMLER  VMAVVPADKL  AVHFHDYGG  ALANILVSLQ  MGISIVDSSI  hydroxymethylglutar
361  AGLGGCPYAK  GASGNVATED  VVYMLNGLGV  HTNVDLGKLI  AAGDFISKHL  GRPNGSKAAV  hydroxymethylglutar
421  ALNRRITADA  SKI
      hydroxymethylglutaryl-CoA lyase
```

# LINKS TO OTHER NCBI Dbs

NCBI Entrez Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Nucleotide for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Get Subsequence

1: X93081. *B.subtilis* bofC, ...[gi:1941915]

LOCUS BSBFOFCGEN 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

JOURNAL *Microbiology* 143 (Pt 1), 157-170 (1997)

MEDLINE [97177783](#)

PUBMED [9025289](#)

REFERENCE 2 (bases 1 to 2664)

AUTHORS Cutting, S.M.

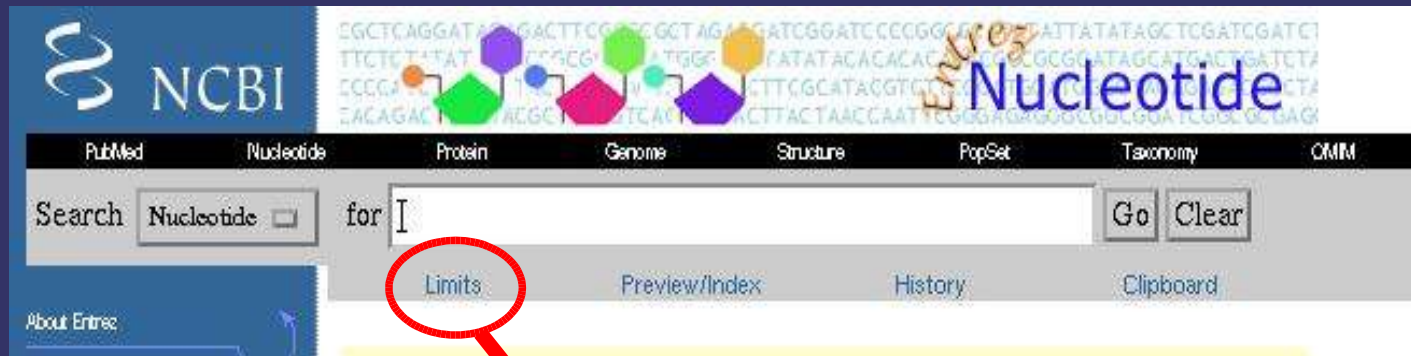
TITLE Direct Submission

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

Links

- ▶ Related Sequences
- ▶ Protein
- ▶ PubMed
- ▶ Taxonomy

# LIMITS



**Limited to:**

exclude ESTs  exclude STSs  exclude GSS  exclude working draft  exclude patents

From    To

Use the format YYYY/MM/DD; month and day are optional.

# RESULTS INDEX

The screenshot shows the NCBI Entrez Nucleotide search page. At the top, there is a navigation bar with links for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, Taxonomy, and OMM. Below this is a search bar with a dropdown menu set to 'Nucleotide' and a text input field containing 'for'. To the right of the search bar are 'Go' and 'Clear' buttons. Below the search bar is a secondary navigation bar with links for Limits, Preview/Index (circled in red), History, and Clipboard. The background features a colorful molecular structure and the text 'Entrez Nucleotide'.

## 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  Preview Index

Click    to add terms selected from Index to the query box.

The screenshot shows a list of search results for the term 'mycobacterium'. The results are displayed in a scrollable list with an 'Up' button at the top and a 'Down' button at the bottom. The results are as follows:

Term	Count
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

# SEARCH HISTORY

NCBI Entrez Nucleotide

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMM

Search Nucleotide for [ ] Go Clear

Limits Preview/Index **History** Clipboard

About Entrez

NCBI Entrez Nucleotide

PubMed Nucleotide Protein Genome Structure PopSet

Search Nucleotide for #1 AND #2 Go Clear

Limits Index **History** Clipboard

- Search History will be lost after one hour of inactivity
- To combine searches use # before search number, e.g., #2 AND #6

Search	Query	Time	Result
#2	Search <b>protease</b>	12:54:14	<a href="#">13061</a>
#1	Search <b>hiv</b>	12:54:06	<a href="#">37549</a>

Clear History

About Entrez

Entrez Nucleotide Help

Submit by Bankit

Check sequence revision history

How to create WWW links to nucleotides

# SEQUENCE FILES

## FLAT FILE FORMATS

For sequence databases, the main formats are:

- FASTA
- GenBank
- EMBL and SwissProt

# FASTA format

*> essential*      **GI**      **Accession.version**      **Locus**      **Additional information**

```
>gi|1941915|emb|x93081.1|BSBOFCGEN B.subtilis bofC gene
CTGCAGCGGCTGACAATAGCAGGCCGACAACGGTTGAGGTGTCAACAGCTGATTTTGTGATGAAGGATAA
ACCGCATTCTTTTTCTTTGAACGCTATAAGGATTCATATGAGGAGGAGATTCTCCGTTTTGCAGAAGCG
ATCGGCACAAACCAGGAGACTCCCTGCACCGGCAATGACGGTTTACAGGCCGGGAGGATCGCCAGAGCAG
CACAGCAATCGCTTGCTTTTTGGCATGCCTGTTAGCATTGAGCACACTGAAAAAATCGCTTTTTAATCTAA
CAGGATTACAATTCAGCAAGCTTGGGTATATACTCCATTGATACTTTAAGTAGGCGGTGGAGAAAATGAA
TACAGTACATGCTAAAGGAAATGTTTTGAACAAAATCGGAATTCCTTCTCACATGGTTTGGGGTTATATT
GGCGTTGTCATCTTTATGGTTGGAGACGGCCTCGAACAAGGCTGGCTGTCTCCTTTTTCTCGTTGATCATG
GTCTCAGTATGCAGCAATCCGCATCGTTATTTACCATGTACGGCATTGCTGTCACCATCTCAGCTTGGCT
TTCAGGAACGTTTGTGGAAACTTGGGGGCCGAGAAAAACGATGACTGTCGGATTGCTTGCATTTATCCTC
```

>

```
CTGCAGCGGCTGACAATAGCAGGCCGACAACGGTTGAGGTGTCAACAGCTGATTTTGTGATGAAGGATAA
ACCGCATTCTTTTTCTTTGAACGCTATAAGGATTCATATGAGGAGGAGATTCTCCGTTTTGCAGAAGCG
ATCGGCACAAACCAGGAGACTCCCTGCACCGGCAATGACGGTTTACAGGCCGGGAGGATCGCCAGAGCAG
CACAGCAATCGCTTGCTTTTTGGCATGCCTGTTAGCATTGAGCACACTGAAAAAATCGCTTTTTAATCTAA
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="AAADNSRPTTVEVSTADDFVMKDKPHFFFLERYKDSYEEEILRFA
EAIGTNQETPCTGNDGLQAGRIARAAQQSLAFGMPVSIHTEKIAF"
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 atcggcacia accaggagac tccctgcacc ggcaatgacg
181 gtttacaggc cgggaggatc gccagagcag cacagcaatc gcttgctttt ggcatgcctg
241 ttagcattga gcacactgaa aaaatcgctt tttaatctaa caggattaca attcagcaag
301 cttgggtata tactccattg aacttttaag taggcggtgg agaaaatgaa tacagtacat
361 gctaaaggaa atgttttgaa caaaatcgga attccttctc acatggtttg gggttatatt
421 ggcgttgtca tctttatggt tggagacggc ctggaacaag gctggctgtc tccttttctc
481 gttgatcatg gtctcagtat gcagcaatcc gcatcgttat ttaccatgta cggcattgct
541 gtcaccatct cagcttggct ttcaggaacg tttgtggaaa cttggggggcc gagaaaaacg
601 atgactgtcg gattgcttgc atttatcctc ggttcggccg cttttatcgg ctgggcgatt
661 cctcatatgt attatccggc tctcttgggc agctatgctc ttagaggctt gggatatccg
721 ctgtttgcat actcttttct cgtatgggtg tcatacagca cctctcaaaa tattcttggg
781 aaagccgtcg gctggttttg gtttatgttt acgtgcggcc ttaacgtgct cggtcctgctc
841 tattccagct atgcagtcc ggcctttgga gaaatcaata cgctttggag cgctttactg
901 tttgtggcgg caggcggaat tcttgcctta ttttttaaca aagataaatt tactccgata
961 caaaaacaag atcagccgaa atggaaagaa ctgtcgaagg catttacgat tatgtttgaa
1021 aaccctaagg taggcatcgg cggagtggtc aagacgatta atgcgatagg acaatttggg
1081 tttgccatct ttcttcctac ttatttagca cgatacgggt attcggtttc ggaatggctg
1141 caaatatggg ggactctggt 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; 005389.
DR      GOA; 005390.
DR      GOA; 005391.
DR      GOA; 005392.
DR      SWISS-PROT; 005389; YRBE_BACSU.
DR      SWISS-PROT; 005390; CSBX_BACSU.
DR      SWISS-PROT; 005391; BOFC_BACSU.
DR      SWISS-PROT; 005392; 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
QQSLAFGMPVSIHTEKIAF"
```

# 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 atcggcacia 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 ctccaacaag 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.

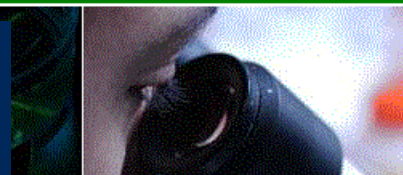


# EMBL

European Molecular Biology Laboratory

Search

## RESEARCH IN MOLECULAR BIOLOGY



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

### Services

EMBL services

Biological databases at EBI

Computational services

Core facilities

Szilárd Library

Technology Transfer

### Upcoming Events

## MOLECULAR BIOLOGY

© EMBL 2002

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

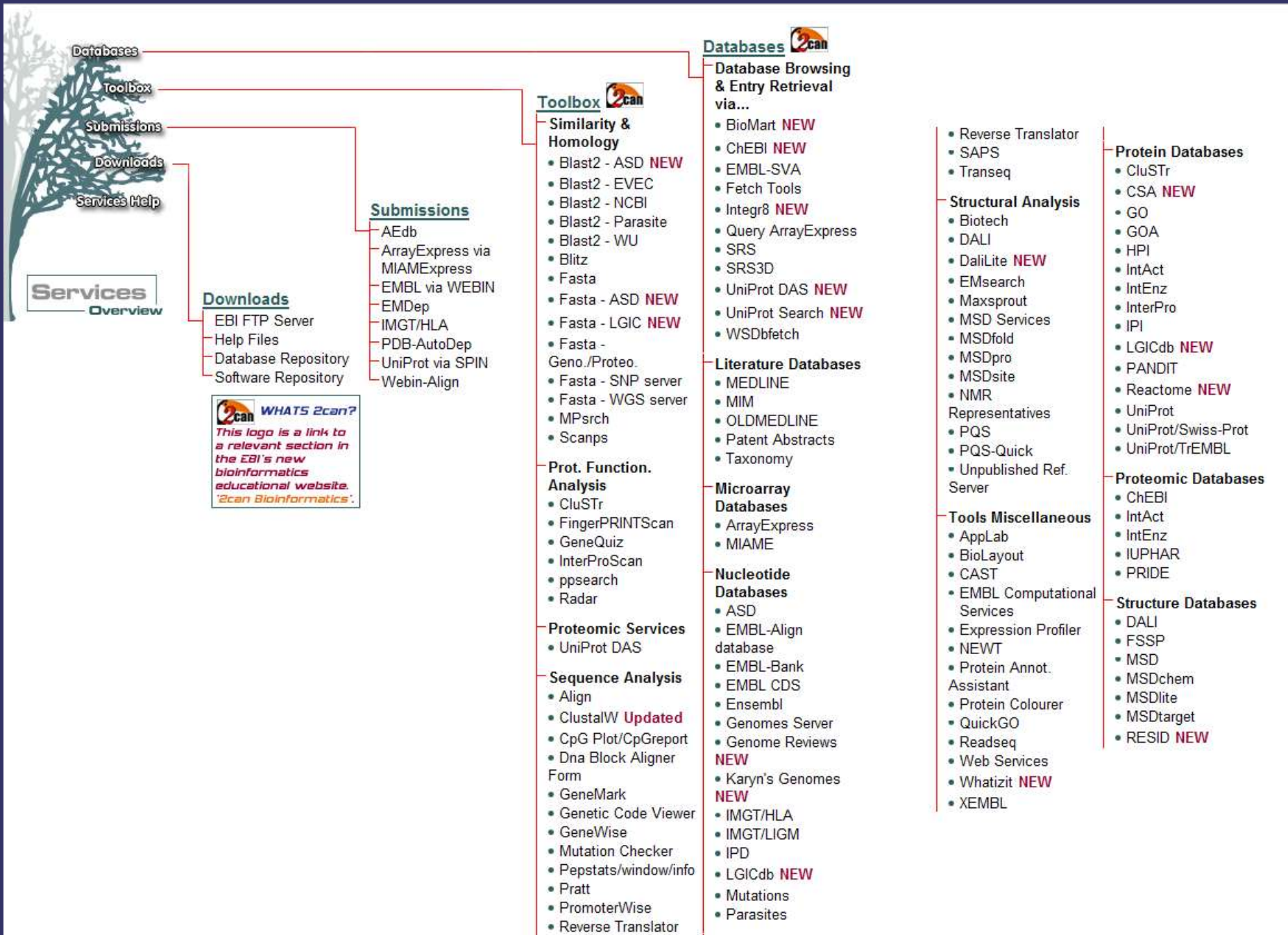


2can Bioinformatics - Training and Education at the EBI

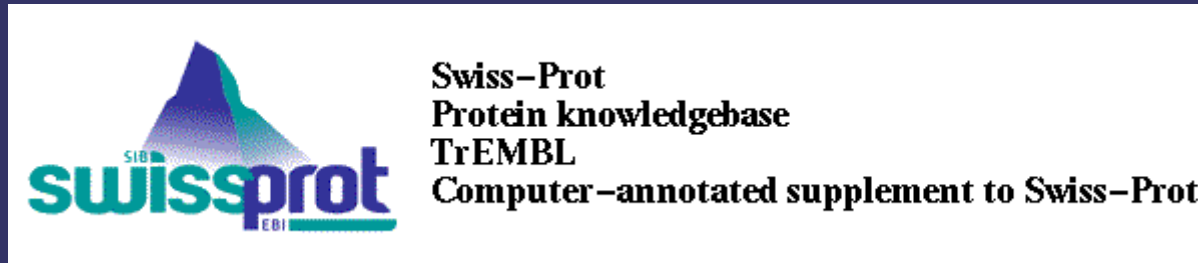
© Copyright European Bioinformatics Institute 2002-2004. All Rights and Trademarks Reserved.

# EMBnet





# 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 Dayhoff

pir.georgetown.edu **PIR** Protein Information Resource

About PIR Databases Search and Retrieval Download Support

AN INTEGRATED PUBLIC RESOURCE OF PROTEIN INFORMATICS TO SUPPORT GENOMIC AND PROTEOMIC RESEARCH AND SCIENTIFIC DISCOVERY

PIR produces the **Protein Sequence Database (PSD)** of functionally annotated protein sequences, which grew out of the *Atlas of Protein Sequence and Structure* (1965-1978) edited by Margaret Dayhoff and has been incorporated into an integrated knowledge base system of value-added databases and analytical tools.

PIR News Flash  
PIR featured in Georgetown Blue & Gray newsletter

Text Search Protein Databases:  GO!

Find an Exact Peptide Match:  GO!

Type in a string of single letter amino acid code (at least 3 letters)

UniProt  
the universal protein resource

Text Search UniProt Knowledgebase

Home About UniProt Getting Started Searches/Toc Databases Support/Documentation

Welcome to UniProt

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt is comprised of three components, each optimized for different uses. The **UniProt Knowledgebase (UniProt)** is the central access point for extensive curated protein information, including function, classification, and cross-reference. The **UniProt Non-redundant Reference (UniRef)** databases combine closely related sequences into a single record to speed searches. The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.

The sequences and information in UniProt are accessible via [text search](#), [BLAST similarity search](#), and [FTP](#).

European Bioinformatics Institute Swiss Institute of Bioinformatics Georgetown University

information,  
tion and  
quences,  
42.

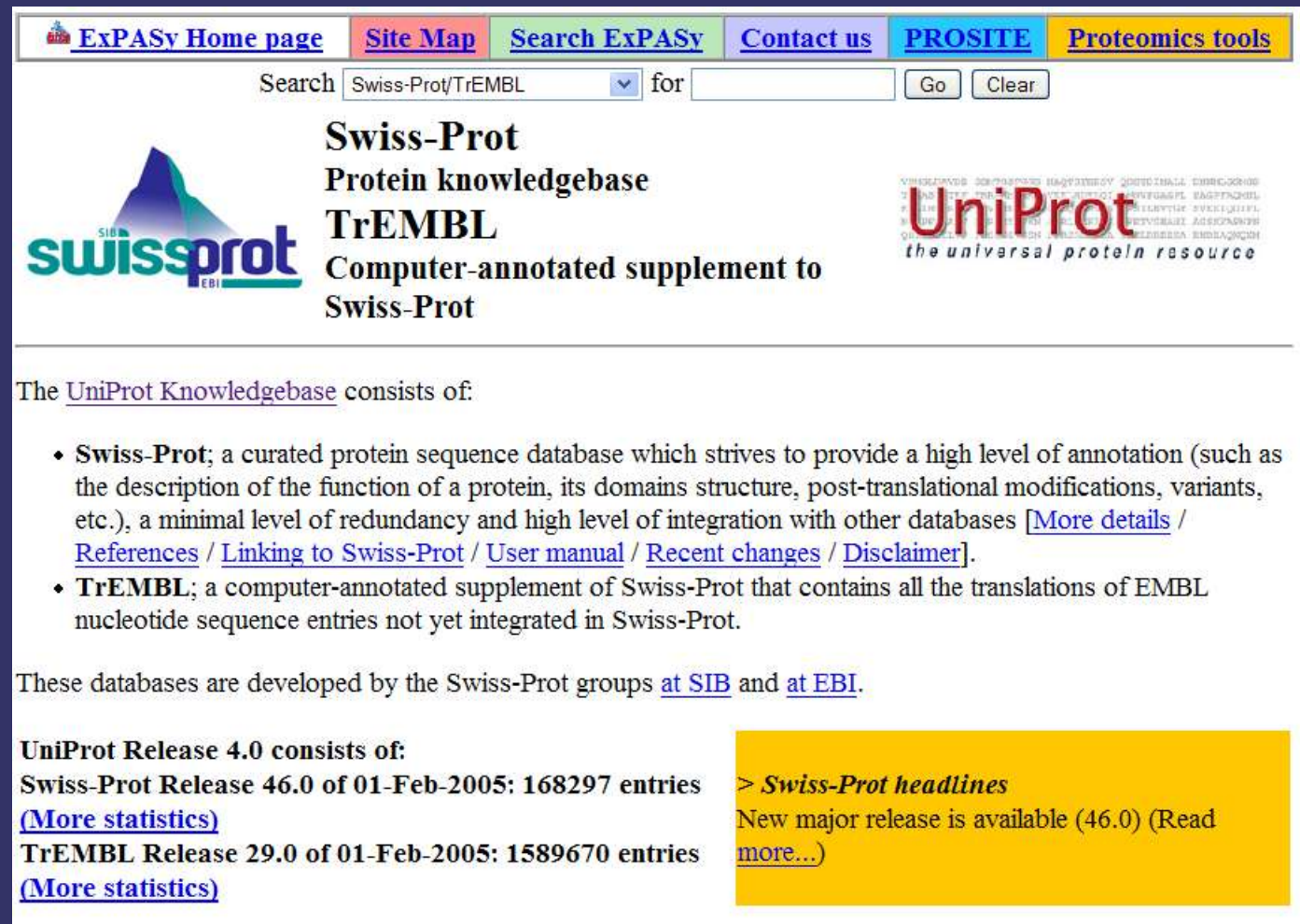
undant  
MBL,  
2004.

## 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.



The screenshot shows the ExPASy website interface. At the top, there are navigation links: ExPASy Home page, Site Map, Search ExPASy, Contact us, PROSITE, and Proteomics tools. Below these is a search bar with the text 'Swiss-Prot/TrEMBL' and a dropdown menu, followed by 'for' and a search button. The main content area features the Swiss-Prot logo (SIB EBI) and the UniProt logo (the universal protein resource). The text reads: 'Swiss-Prot Protein knowledgebase TrEMBL Computer-annotated supplement to Swiss-Prot'. Below this, it states: 'The UniProt Knowledgebase consists of:' followed by a list of two items: 'Swiss-Prot; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Disclaimer].' and 'TrEMBL; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.' Below the list, it says: 'These databases are developed by the Swiss-Prot groups at SIB and at EBI.' At the bottom, it lists: 'UniProt Release 4.0 consists of: Swiss-Prot Release 46.0 of 01-Feb-2005: 168297 entries (More statistics) TrEMBL Release 29.0 of 01-Feb-2005: 1589670 entries (More statistics)'. A yellow box on the right contains the text: '> Swiss-Prot headlines New major release is available (46.0) (Read more...)'.

# 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	<a href="#">Escherichia coli</a> [TaxID: 562]
Encoded on	Plasmid pBP311.
Taxonomy	<a href="#">Bacteria</a> ; <a href="#">Proteobacteria</a> ; <a href="#">Gammaproteobacteria</a> ; <a href="#">Enterobacteriales</a> ; <a href="#">Enterobacteriaceae</a> ; <a href="#">Escherichia</a> .

### 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

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

## 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	IPFFAAFCPLP	VFAHPETLVK	VKDAEDKLGA	RVGYIELDLN	SGKILESFRP
70	80	90	100	110	120
EERFPMSTF	KVLLCGAVLS	RVDAGQEQLG	RRIHYSQNDL	VEYSPVTEKH	LTDGMTVREL
130	140	150	160	170	180
CSAAITMSDN	TAANLLTTI	GGPKELTAFI	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

## Sequence

Link to FASTA format

# SRS

## Sequence Retrieval System

Version 6.1.3.11 | [List of Public SRS servers](#) | [EBI](#)

 **SRS**

  
European  
Bioinformatics  
Institute

**Permanent Session**

**Start**  **Databanks** **Information**

- **WARNING:** Temporary sessions last only 24 hours

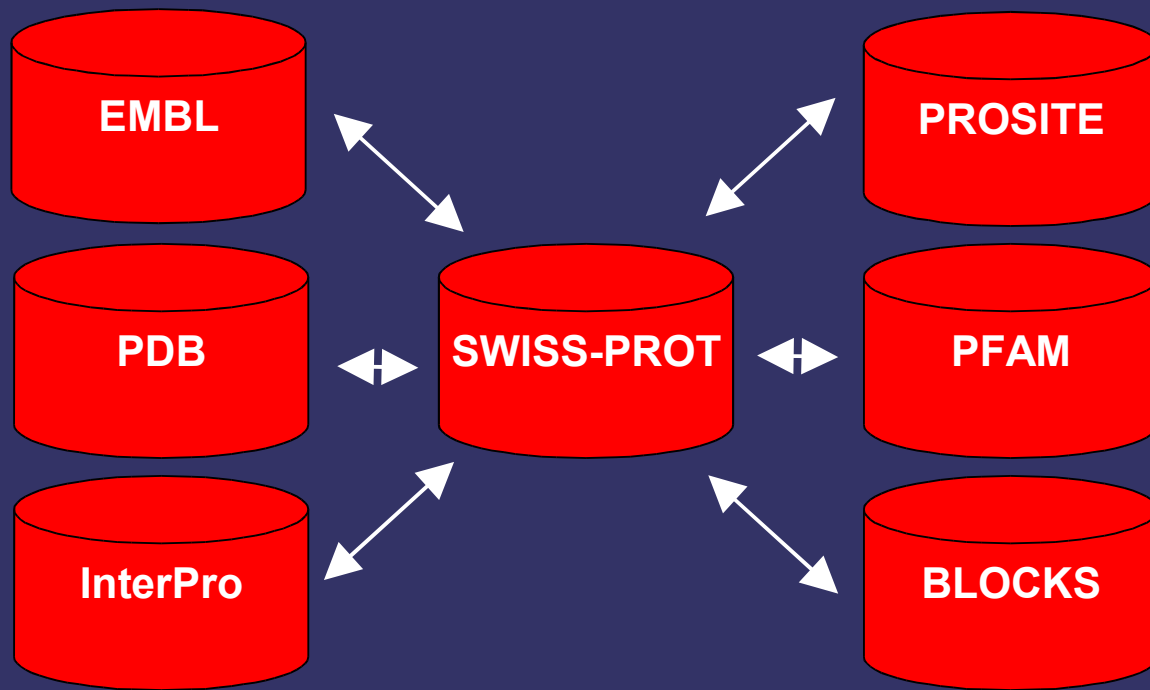


# SEQUENCE RETRIEVAL SYSTEM

- 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

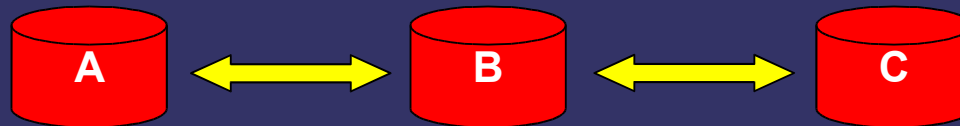
- Information from different databases is connected by means of cross-references.
- Cross-references are entry attributes that make reference to entries in other databases.
- Not all databases include cross-references.
  - SwissProt and EMBL have fields called DR, which includes cross references to most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

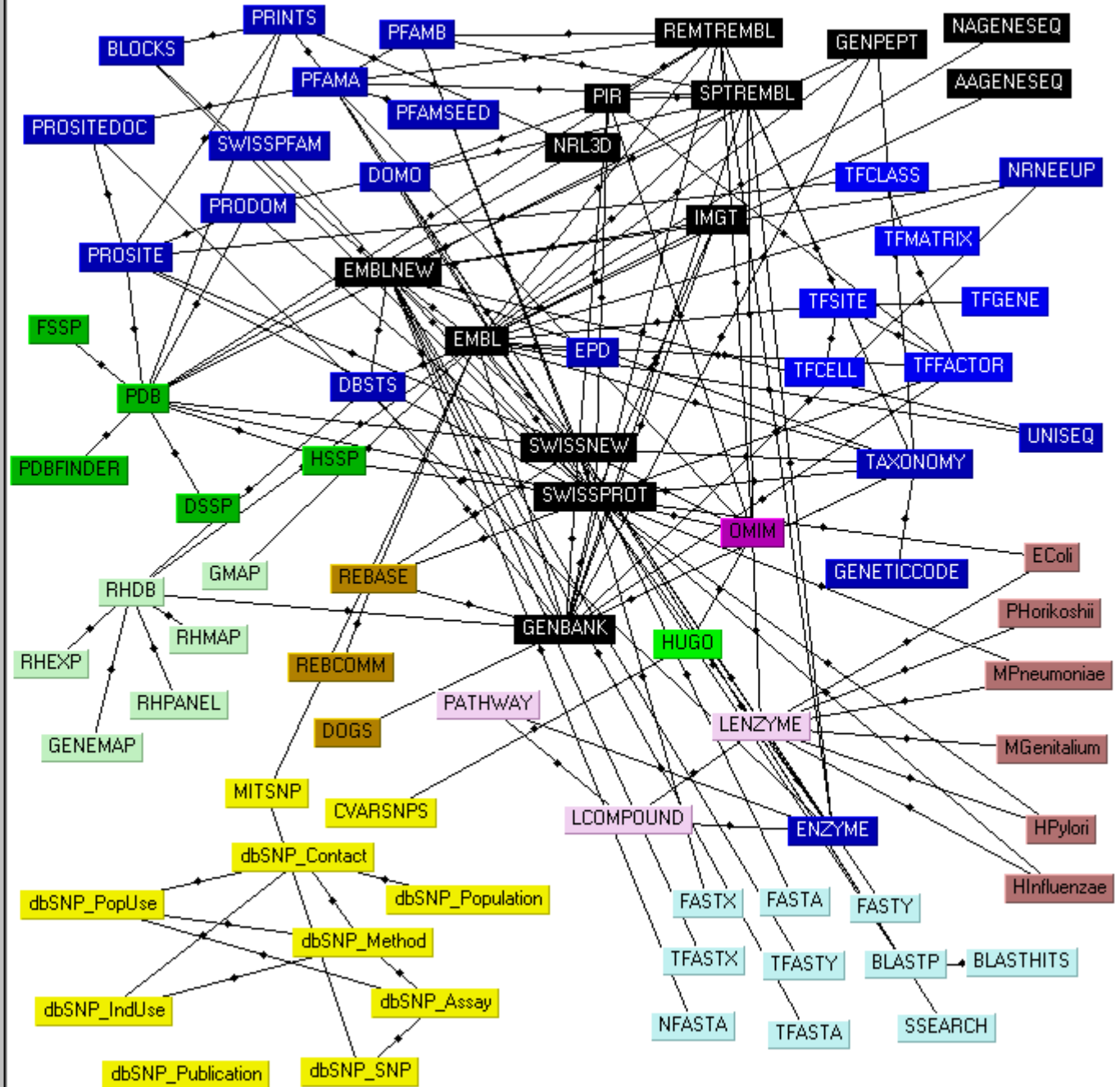


Indirect link between A and C

Links are established bidirectionally

# LIBRARY NETWORK

- EMBL  visible
- EMBLNEW  visible
- GENBANK  visible
- SWISSPROT  visible
- SWISSNEW  visible
- PIR  visible
- SPTREMBL  visible
- REMTREMBL  visible
- GENPEPT  visible
- SWALL  visible
- NRL3D  visible
- IMGT  visible
- AAGENESEQ  visible
- NAGENESEQ  visible
- OGLYC  visible
- SeqRelated  visible
- PROSITE  visible
- PROSITEDOC  visible
- BLOCKS  visible
- DOMO  visible
- PRINTS  visible
- PFAMA  visible
- PFAMB  visible
- SWISSPFAM  visible
- PFAMHMM  visible
- PFAMSEED  visible
- PRODOM  visible
- TAXONOMY  visible
- GENETICCODE  visible



# 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

**SRS**  **LION**

**Temporary Project**

1ZkK1PUAPR

**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  matching :

Searches Databanks: EMBL Nucleotides

 **Search**

**News and Announcements**

[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



SRS is a product of Lion Bioscience AG

**List Search**

[Search Tips](#)

Database ID:  ID:  File:  DATABASE ID:

# 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**

The screenshot shows the SRS@EMBL-EBI search interface. At the top, there are navigation tabs: Library, Query, Results, Projects, Views, and Databanks. The Library tab is active. Below the tabs, there is a search bar with a 'Quick Search' button and a 'Reset' button. The main content area is divided into several sections:

- Query forms:** Includes 'Standard' and 'Extended' buttons.
- Browse Databanks:** A section for navigating through different database categories.
- Quick Sequence Search:** Includes a 'FastA' button.
- Applications:** A section for additional search options.

The main search area displays a list of databases, organized into categories:

- Literature, Bibliography and Reference Databases:** MEDLINE, MEDLINE (Updates), MEDLINE (Main Release), OMIM.
- Nucleotide sequence databases:** EMBL, EMBL (Release), EMBL (Updates), EMBL (WGS), EMBL (TPA), EMBL (Contig), REFSEQ, ENSEMBL HUMAN, ENSEMBL MOUSE, ENSEMBL FLY, ENSEMBL FISH, IMGTHLA, IMGT/LIGM-DB, PATENT\_DNA.
- Protein sequence databases:** SWALL (SPTR), Swiss-Prot, SpTrEMBL, TrEMBL (Updates), IPI, RemTrEMBL, PIR, REFSEQP, PATENT\_PRT, JPO\_PRT, USPO\_PRT, MHCBN, SWISSCHANGE.
- Nucleotide related databases**
- Protein function databases**
- Protein structure databases**
- Enzymes, reactions and metabolic pathway databases**
- Mutation and SNP databases**
- Gene ontology resources**
- Mapping databases**
- Other databases**
- User owned databases**

At the bottom, there is a note: 'If you find problems or have suggestions please mail the SRS administrator'.

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.

Library Query Results Projects Views Databanks

Reset search [EMBL](#)

**Submit Query**

append wildcards to words

combine searches with **& (AND)**

Number of entries to display per page **30**

**Extended** query form

separate multiple values by & (and), |(or), !(but not) **Submit Query**

<b>i</b>	AllText	<input type="text"/>
<b>i</b>	AllText	<input type="text"/>
<b>i</b>	AllText	<input type="text"/>
<b>i</b>	AllText	<input type="text"/>

retrieve entries of type **Entry**

Use view **EMBLSeqSimpleView**

Create your own view

Select fields to display:  table

**Query  
Fields**

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

**Field Selection  
for displaying the results**

# RESULTS: HIT LIST

The screenshot displays the SRS@EMBL-EBI interface. At the top, there are navigation tabs: Library, Query, Results, Projects, Views, and Databanks. The 'Query' tab is active, showing a search query: "Query "[embl-AllText:narc\*]" found 343 entries". Below the query bar is a 'Reset' button and a 'next' link. On the left side, there is a control panel with the following elements:

- SRS Query**: Points to the search query bar.
- HITS**: Points to the 'unselected only' and 'selected only' radio buttons.
- Different options to View**: Points to the 'Link', 'Save', and 'View' buttons.
- Application Launcher**: Points to the 'Launch' button.

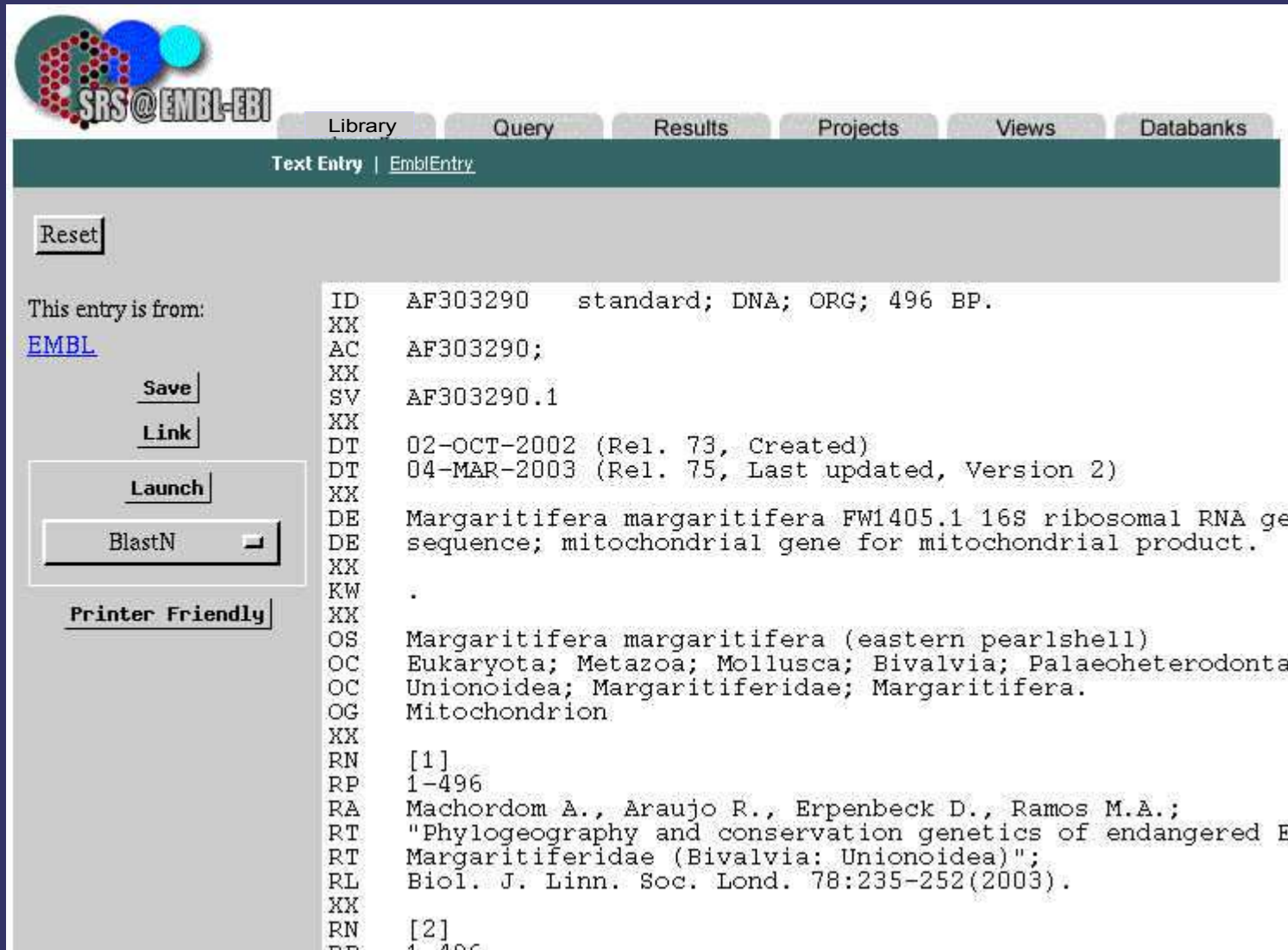
The main area displays a table of search results:

<a href="#">EMBL</a>	<a href="#">Accession</a> <small>(Link to SVA)</small>	<a href="#">Description</a>	<a href="#">SeqLength</a>
<input type="checkbox"/> <a href="#">EMBL:AF303290</a>	<a href="#">AF303290</a>	Margaritifera margaritifera FW1405.1 16S ribosomal RNA gene, partial sequence; mitochondrial gene for mitochondrial product.	496
<input type="checkbox"/> <a href="#">EMBL:AF303291</a>	<a href="#">AF303291</a>	Margaritifera margaritifera FW1405.2 16S ribosomal RNA gene, partial sequence; mitochondrial gene for mitochondrial product.	496
<input type="checkbox"/> <a href="#">EMBL:AF303292</a>	<a href="#">AF303292</a>	Margaritifera margaritifera FW1405.4 16S ribosomal RNA gene, partial sequence; mitochondrial gene for mitochondrial product.	496
<input type="checkbox"/> <a href="#">EMBL:AF303326</a>	<a href="#">AF303326</a>	Margaritifera margaritifera FW1405-1 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial gene for mitochondrial	657

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



The screenshot shows the SRS@EMBL-EBI interface. At the top, there are navigation tabs: Library, Query, Results, Projects, Views, and Databanks. Below these is a green bar with 'Text Entry | EmblEntry'. On the left side, there are several buttons: 'Reset', 'EMBL' (a link), 'Save', 'Link', 'Launch', 'BlastN' (with a dropdown arrow), and 'Printer Friendly'. The main content area displays the entry details for AF303290.1, including its ID, accession number, version, dates, description, keywords, organism, and references.

```
ID AF303290 standard; DNA; ORG; 496 BP.
XX
AC AF303290;
XX
SV AF303290.1
XX
DT 02-OCT-2002 (Rel. 73, Created)
DT 04-MAR-2003 (Rel. 75, Last updated, Version 2)
XX
DE Margaritifera margaritifera FW1405.1 16S ribosomal RNA ge
DE sequence; mitochondrial gene for mitochondrial product.
XX
KW .
XX
OS Margaritifera margaritifera (eastern pearlshell)
OC Eukaryota; Metazoa; Mollusca; Bivalvia; Palaeoheterodonta
OC Unionoidea; Margaritiferidae; Margaritifera.
OG Mitochondrion
XX
RN [1]
RP 1-496
RA Machordom A., Araujo R., Erpenbeck D., Ramos M.A.;
RT "Phylogeography and conservation genetics of endangered E
RT Margaritiferidae (Bivalvia: Unionoidea)";
RL Biol. J. Linn. Soc. Lond. 78:235-252(2003).
XX
RN [2]
RP 1-496
```

Entry View

# APPLICATIONS

The screenshot shows the SRS@EMBL-EBI web interface. At the top left is the logo with the text 'SRS@EMBL-EBI'. Below it are navigation tabs: 'Library', 'Query', 'Results', 'Projects', 'Views', and 'Databanks'. A search bar contains the query: "Query "[patent\_prt-AllText:spider\*]" found 48 entries". Below the search bar, there are controls: "show all [+]" and "collapse all [-]". A left sidebar titled 'Packages' lists: FASTA, BLAST, CLUSTAL, OTHER, EMBOSS, and TEST. The main content area shows a tree view of application categories: Alignment Applications (with sub-items: Alignment Differences, Alignment Global, Alignment Local, Alignment Multiple), Edit Applications, Information Applications, Nucleic Applications, Protein Applications, and Similarity Search Applications. Under 'Alignment Multiple', there are links for 'ClustalW' and 'Multiple Sequence Alignment'. At the bottom of the interface, it says 'SRS 7.0.2 | [feedback](#)'.

**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.

Reset

Expression

Perform operation on selected queries

Save

Delete

Link

View

default view

Combine

& (AND)

Number of entries to display per page 30

Name	Type	N Total	From Databank	N	Query Expression
<input type="checkbox"/> Q5	query	155	EMBL	155	[[embl-AllText:ponA*]
<input type="checkbox"/> Q4	query	343	EMBL	343	[[embl-AllText:narc*]
<input type="checkbox"/> Q3	query	48	PATENT_PRT	48	[[patent_prt-AllText:
<input type="checkbox"/> Q2	query	1	PATENT_PRT	1	[[patent_prt-AllText:
<input type="checkbox"/> Q1	query	69	EMBLRELEASE	69	[[emblrelease-AllText

My queries

# PROJECTS

SRS@EMBL-EBI

Library Query Results **Projects** Views Databanks

SRS project files

Save to desktop

Open from desktop

Browse...

This temporary project contains:

Queries	Views
<i>name query</i>	<i>No views</i>
Q5	[embl-AllText:ponA*]
Q4	[embl-AllText:narc*]
Q3	[patent_prt-AllText:spider*]
Q2	[patent_prt-AllText:leech*]
Q1	[emblrelease-AllText:cito*]

SRS 7.0.2 | [feedback](#)

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

The screenshot shows the EMBL-EBI website interface. At the top, there is a navigation bar with tabs for 'Quick Search', 'Library Page', 'Query Form', 'Tools', 'Results', 'Projects', 'Views', and 'Databanks'. The 'Views' tab is currently selected. Below the navigation bar, there is a 'Reset' button. The main content area is divided into two columns. The left column is titled 'Create View Options' and contains a 'View name:' field with the text 'PRUEBA' and a 'Display results as' section with radio buttons for 'table' (selected) and 'list'. The right column is titled 'Databanks to define a view for' and 'Databanks to be linked to'. Below these titles is a section 'Show fields from:' with radio buttons for 'All fields in databanks' and 'Common fields only' (selected). The 'Databanks to define a view for' list includes UniProt, UNIPROT\_reference, UNIPROT\_comment, UNIPROT\_links, UNIPROT\_features, UNIPROT\_counter, UniParc, DBCrossRef, UNIPARC\_counter, UniRef100, UniRef90, UniRef50, UniProt/Swiss-Prot, SWISSPROT\_reference, and SWISSPROT\_comment. The 'Databanks to be linked to' list includes MEDLINE, OMIM, TAXONOMY, GENETICCODE, Patent Abstracts, KarynsGenomes, MEDLINE (Main Release 2005), MEDLINE (Updates), MED2PUB, EMBL, EMBL\_reference, EMBL\_features, and EMBL\_counter. At the bottom of the right column, there is a 'Create New View' button. Below the main content area, there is a 'Delete View:' section with a dropdown menu showing '\* Names only \*' and a 'Delete View' button.

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 which data will be displayed; third (next slide), choose the fields that you want to have displayed; finally, save the new view.

# VIEWS



Reset

## Create View Options

Select the **datafields** you want displayed in your view using the checkboxes.

View name: PRUEBA

Save view:

## Datafields for your primary databanks

### UNIPROT

- |  |   |   |   |
|--|---|---|---|
| <input checked="" type="checkbox"/> <u>ID</u>        | <input type="checkbox"/> <u>EntryName</u>     | <input type="checkbox"/> <u>AccNumber</u>           | <input type="checkbox"/> <u>DateCreated</u> |
| <input type="checkbox"/> <u>DateSeqUpdate</u>        | <input type="checkbox"/> <u>DateAnnUpdate</u> | <input type="checkbox"/> <u>Description</u>         | <input type="checkbox"/> <u>GeneName</u>    |
| <input type="checkbox"/> <u>Synonym</u>              | <input type="checkbox"/> <u>OrderedLocus</u>  | <input type="checkbox"/> <u>ORFnames</u>            | <input type="checkbox"/> <u>Organism</u>    |
| <input checked="" type="checkbox"/> <u>Species</u>   | <input type="checkbox"/> <u>Taxonomy</u>      | <input type="checkbox"/> <u>Organelle</u>           | <input type="checkbox"/> <u>NCBI_TaxId</u>  |
| <input type="checkbox"/> <u>TaxCount</u>             | <input type="checkbox"/> <u>Keywords</u>      | <input type="checkbox"/> <u>ProteinID</u>           | <input type="checkbox"/> <u>SeqLength</u>   |
| <input checked="" type="checkbox"/> <u>MolWeight</u> | <input type="checkbox"/> <u>crc</u>           | <input type="checkbox"/> <u>DBxref</u>              | <input type="checkbox"/> <u>MedlineID</u>   |
| <input type="checkbox"/> <u>swProtName</u>           | <input type="checkbox"/> <u>Isoform</u>       | <input checked="" type="checkbox"/> <u>Sequence</u> | <input type="text" value="fasta"/>          |
| <input type="checkbox"/> <u>DBLink</u>               |   |   |   |

### Fields of subentry **Reference**

- |  |  |   |   |
|--|--|---|---|
| <input type="checkbox"/> <u>Authors</u>        | <input type="checkbox"/> <u>Title</u>          | <input type="checkbox"/> <u>RefPosition</u> | <input type="checkbox"/> <u>RefGroup</u>  |
| <input type="checkbox"/> <u>RefNumber</u>      | <input type="checkbox"/> <u>RefCommentCode</u> | <input type="checkbox"/> <u>RefComment</u>  | <input type="checkbox"/> <u>Journal</u>   |
| <input type="checkbox"/> <u>VolumeNo</u>       | <input type="checkbox"/> <u>FirstPage</u>      | <input type="checkbox"/> <u>Year</u>        | <input type="checkbox"/> <u>Citation</u>  |
| <input type="checkbox"/> <u>SubmissionDate</u> | <input type="checkbox"/> <u>Patent</u>         | <input type="checkbox"/> <u>PatentDate</u>  | <input type="checkbox"/> <u>MedlineID</u> |

### Fields of subentry **Comment**

- |   |   |
|---|---|
| <input type="checkbox"/> <u>CommentType</u> | <input type="checkbox"/> <u>Comment</u> |
|---|---|

### Fields of subentry **Links**

- |  |  |
|--|--|
| <input type="checkbox"/> <u>DbName</u> | <input type="checkbox"/> <u>DBxref</u> |
|--|--|

### Fields of subentry **Feature**

- |                                       |  |   |
|---------------------------------------|--|---|
| <input type="checkbox"/> <u>FtKey</u> | <input type="checkbox"/> <u>FtLength</u> | <input type="checkbox"/> <u>FtDescription</u> |
|---------------------------------------|--|---|

### Fields of subentry **Counter**

- |   |  |
|---|--|
| <input type="checkbox"/> <u>CountItem</u> | <input type="checkbox"/> <u>CountN</u> |
|---|--|

# DATABANKS

**EMBL-EBI**  
European Bioinformatics Institute

Quick Search Library Page Query Form Tools Results Projects Views **Databanks** HELP

**Display Options**

List databanks:  
by groups

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

## Protein structure analyses

- SCOP / CATH  
classification
- Functional sites mapped  
on structure

*Function prediction (molecular level)?*

## Secondary Database similarity search

- Prosite
- Pfam
- SMART
- PRINTS
- BLOCKS
- InterPro

*Protein motifs*

*Domain organization*

*Family assignment*

*Function prediction?*

## Protein structure prediction

- 1D features
- 3D structure / fold prediction

*Known / Predicted structure*

This presentation contains material from:

Rodrigo Lopez, EBI