Scuola di Ingegneria Industriale e dell'Informazione

Laurea Magistrale in Ingegneria dell'Informazione
Laurea Magistrale in Ingegneria Biomedica
Master of Science in Bioinformatics for Computational Genomics

POLITECNICO DI MILANO

Dipartimento di Elettronica, Informazione e Bioingegneria





Bioinformatics and Computational Biology

Marco Masseroli, PhD *marco.masseroli@polimi.it*

Biomolecular Databank Survey

Index

- **Biomolecular databank survey**
  - Databank main features to be considered
  - Selected biomolecular databanks
    - Selected databank URLs, Example IDs
    - EMBL-EBI
    - UniGene
    - Entrez Gene
    - UniProt
      - Swiss-Prot, TrEMBL, PIR
    - PDB
    - KEGG, Reactome
    - GOA
    - OMIM
    - GEO
    - SOURCE, GeneCards, Harvester

Databank main features to be considered

- Scientific community acknowledgment
- Building procedures, components
  - Curated vs. computationally inferred
- Content provided:
  - Data:
    - Semantic types, organisms
    - Annotations
    - Cross-references
    - Updating frequency
    - Statistics
  - Query and analysis services:
    - Query options and response time
- Access (Web, FTP, Web service)
  - Data format and dimension

Biomolecular Databanks

Selected databanks

Primary DBs

- EMBL-EBI
- GenBank
- DDBJ

Sequence DBs

- UniGene
- RefSeq
- UCSC
- Ensemble

Genomic DBs

- GDB

Gene DBs

- **Entrez Gene**
- OmoloGene

Protein DBs

- **UniProt**
- Swiss-Prot
- TrEMBL
- PIR

Protein 3D structure DBs

- PDB

Protein domain DBs

- InterPro

Patway DBs

- **KEGG**
- Reactome

Gene Ontology Annot. DBs

- GOA

Disorders DBs

- **OMIM**
- GAD

Mutation DBs

- **dbSNPs**

Microarray DBs

- SMD
- GEO
- **Array Express**

Integrative DBs

- SOURCE
- GeneCards

Literature DBs

- **PubMed**

Biomolecular Databanks
Selected databanks

- For each considered databank, the following points are discussed:
  - Databank description
  - Type of data included
  - Query options
  - Updating
  - Statistics
  - FTP access

Biomolecular Databanks

Selected biomolecular databank URLs

- **EMBL**: http://www.ebi.ac.uk/embl/
- **GenBank**: http://www.ncbi.nlm.nih.gov/GenBank/index.html
- **DDJB**: http://www.ddbj.nig.ac.jp/
- **UniGene**: http://www.ncbi.nlm.nih.gov/UniGene/
- **RefSeq**: http://www.ncbi.nlm.nih.gov/RefSeq/
- **UCSC**: http://genome.ucsc.edu/
- **GDB**: http://www.gdb.org/
- **Ensemble**: http://www.ensembl.org/
- **Entrez Gene**:
  http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene
- **HomoloGene**: http://www.ncbi.nlm.nih.gov/HomoloGene/

Biomolecular Databanks
Selected biomolecular databank URLs

- **Swiss-Prot**: http://www.expasy.ch/sprot/
- **TrEMBL**: http://www.ebi.ac.uk/trembl/
- **PIR**: http://www-nbrf.georgetown.edu/pirwww/search/textpsd.shtml
- **UniProt**: http://www.pir.uniprot.org/
- **InterPro**: http://www.ebi.ac.uk/interpro/
- **PDB**: http://www.rcsb.org/pdb/
- **KEGG**: http://www.genome.ad.jp/kegg/
- **Reactome**: http://www.reactome.com/
- **GOA:** http://www.ebi.ac.uk/GOA/
- **OMIM**: http://www.ncbi.nlm.nih.gov/Omim/

Biomolecular Databanks
Selected biomolecular databank URLs

- **SNPs**: http://snp.cshl.org/
- **SMD**: http://genome-www5.stanford.edu/Microarray/
- **GEO**: http://www.ncbi.nlm.nih.gov/geo/
- **SOURCE**: http://source.stanford.edu/
- **GeneCards**: http://bioinformatics.weizmann.ac.il/cards/
- **Harvester:** http://harvester.embl.de/index.html

Of each main databank, at least you should know the building procedures (curated vs. computationally inferred) and the content provided (data types, main organisms, updating frequency)

Example IDs - Nucleotide sequence IDs

GenBank accession number	UniGene cluster ID	Entrez Gene ID
H59260	Hs.1634	993
H72122	Hs.104925	8507
H87471	Hs.169139	8942
R43509	Hs.75251	8554
W96134	Hs.78465	3725
AA039640	Hs.75188	7465
AA047413	Hs.55606	7571
AA158990	Hs.80680	9961
AA399473	Hs.295944	7980
AA447393	Hs.75890	8720

Example IDs - Amino acid sequence IDs

Swiss-Prot / UniProt accession number / ID	PIR accession	PDB ID
Q16719	A41648	1C25
P30304	A48157	1AH9
P09581	I38238	1C04
P30291	I53908	2RGF
Q14703	JC5517	3EZA
O95644	S10404	4HHB
P28352	S12008	5TMP
P48307	S51342	7ENL
P48431	S55048	9INS
P05412	T04859	13PK

EMBL-EBI Nucleotide Sequence Databank

EMBL-EBI
European Bioinformatics Institute

EBI Home | About EBI | Research | Services | Toolbox | Databases | Downloads | Submissions

EMBL-NUCLEOTIDE SEQUENCE DATABASE

- Index
- Access
- Documentation
- News
- Submission
- Group Info
- Contact

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 77, December 2003), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in Nucl. Acids Res., 2004, Vol. 32, D27-D30 provides further information and details.

The EMBL nucleotide sequence database group is headed by:
Rolf Apweiler.

Link	Explanation
Access	Completed Genomes Webserver, database queries (SRS) and FTP archives (EMBL release, alignments etc)
Submission	Primary sequence submissions, third party annotation, updates and alignment submissions.
Documentation	Release notes, user manual, database statistics, FAQ, EMBL Features and Qualifiers, Feature Table Document, Annotation Examples, Nucleotide Sequence Database Policies
Group info	Group members and publications
Contact	How to contact the EMBL Nucleotide Sequence Database

EMBL Fetch

Fetch an EMBL record by accession number

Go

TPA

Users can now submit re-annotations/ re-assemblies of sequences already present in EMBL and owned by other groups.

NCBI

The Nucleotide Sequence Database is produced in collaboration with GenBank (USA).

DDBJ

EMBL-EBI Nucleotide Sequence Databank

(*http://www.ebi.ac.uk/embl/index.html*)

Biomolecular Databanks
EMBL-EBI databank

- The EMBL-EBI Nucleotide Sequence Databank (EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications (http://www.ebi.ac.uk/embl/Submission/index.html)

- The database is produced in an international collaboration (http://www.ebi.ac.uk/embl/Contact/collaboration.html) with GenBank (US) and the DNA Data Base of Japan (DDBJ). All new and updated databank entries are exchanged on a daily basis

Biomolecular Databanks
EMBL-EBI databank - Updating

- EMBL-EBI databank releases are produced quarterly
  - The latest data collection can be accessed via FTP and WWW interfaces, or through Web services
  - The EBI's Sequence Retrieval System (SRS) integrates and links the main nucleotide and protein databanks as well as many other specialist molecular biology databanks
  - For sequence similarity searching, many tools (e.g. FASTA and BLAST) are available that allow to compare specific sequences against all data in the EMBL-EBI Nucleotide Sequence Databank, the complete genomic component subsection, or the Whole Genome Shotgun data sets

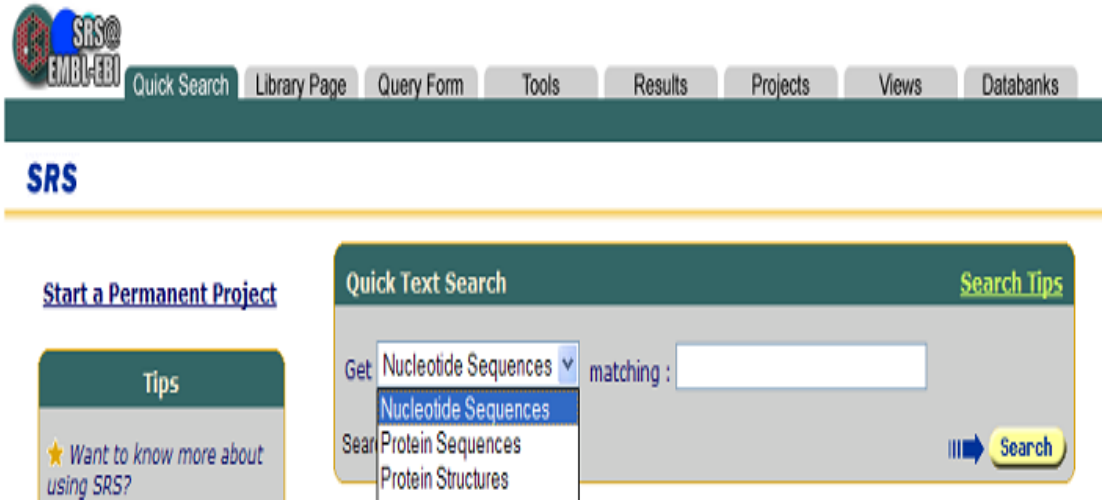
Biomolecular Databanks
EMBL-EBI databank - Services

- Services (http://www.ebi.ac.uk/services/):
  - Databases (http://www.ebi.ac.uk/Databases/):
    - Nucleotide, Protein, Structure, Microarray and Literature databases
    - Use SRS and SRS3D to search and retrieve data.
  - Tools (http://www.ebi.ac.uk/Tools/):
    - Homology & Similarity Detection (BLAST, FASTA)
    - Protein Function Analysis (InterProScan)
    - Structural Analysis (MSDfold, DALI)
    - Sequence Analysis (ClustalW)
    - Other (Expression Profiler)
  - Submissions, Downloads, Bioinformatics Educational Resources (2can) (http://www.ebi.ac.uk/2can/home.html)

EMBL-EBI databank - Access

- Access to Completed Genome server (http://www.ebi.ac.uk/genomes/):
  - Viruses (http://www.ebi.ac.uk/genomes/virus.html)
  - Phages (http://www.ebi.ac.uk/genomes/phage.html)
  - Organelles (http://www.ebi.ac.uk/genomes/organelle.html)
  - Archaea (http://www.ebi.ac.uk/genomes/archaea.html)
  - Bacteria (http://www.ebi.ac.uk/genomes/bacteria.html)
  - Eukaryota (http://www.ebi.ac.uk/genomes/eukaryota.html)

SRS6: Query all databases (http://srs.ebi.ac.uk/)



EMBL-EBI databank - Access

- FTP access (ftp://ftp.ebi.ac.uk/):
  - Sub-directories related to the EMBL database (/pub/databases/embl/)
    - Finished genomes, chromosomes and contigs (/pub/databases/embl/genomes/)
    - Complete latest full release of the EMBL Nucleotide Sequence Database (/pub/databases/embl/release/)
    - Complete list of sequence alignment data (/pub/databases/embl/align/)

EMBL-EBI databank - Access

- Web Services (http://www.ebi.ac.uk/Tools/webservices/):
  - EBI provides programmatic access to various data resources and analysis tools via Web Services, for:
    - Data retrieval
    - Analysis tool usage
    - Similarity search
    - Multiple alignment
    - Structural analysis
    - Literature search and bio-ontology usage
  - Warning: remember to submit a few jobs at a time!

EMBL-EBI databank - Access

- Example Web services:
  - EB-Eye: EMBL database search using the EB-eye search engine
  - WSDbfetch: implementation of Dbfetch, a generic DB retrieval system (http://www.ebi.ac.uk/Tools/ webservices/ WSDbfetch.html)
  - Soaplab: Includes most EMBOSS applications for launching through programmatic access
  - WSWUBlast: Compare a novel sequence with those contained in nucleotide and protein databases using WU-BLAST
  - WSClustalW2: Latest version of the ClustalW global multiple sequence alignment tool
  - WSSSM: Comparing protein structures in 3D
  - PICR: Protein Identifier Cross-Reference Service

EMBL-EBI databank - Access - XML format

- The EMBL Nucleotide Sequence Database has initiated efforts to produce an XML format for the distribution of its entries

- The development of this format will be carried out in collaboration with DDBJ and GenBank with the aim of developing a common representation for the distribution of data

- **SRS**:
  - The Sequence Retrieval System (SRS) can be used to browse the various biological sequence and literature databases the EBI has available (http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession)

- **SRS3D**:
  - SRS3D is an integrated environment that allows the end-user to quickly and easily retrieve/visualize sequence structure and also feature data from primary, secondary and tertiary protein databases (http://srs3d.ebi.ac.uk/)

EMBL-EBI databank - Query options

- **Fetch Tools**:
  - Dbfetch - allows to retrieve up to 50 entries at a time from various up-to date biological databases (http://www.ebi.ac.uk/cgi-bin/emblfetch)
  - Medlinefetch - allows to retrieve one entry at a time from the MEDLINE literature reference database (http://www.ebi.ac.uk/cgi-bin/medlinefetch)

- **Query ArrayExpress**: Search the ArrayExpress microarray database (http://www.ebi.ac.uk/microarray-as/ae/)

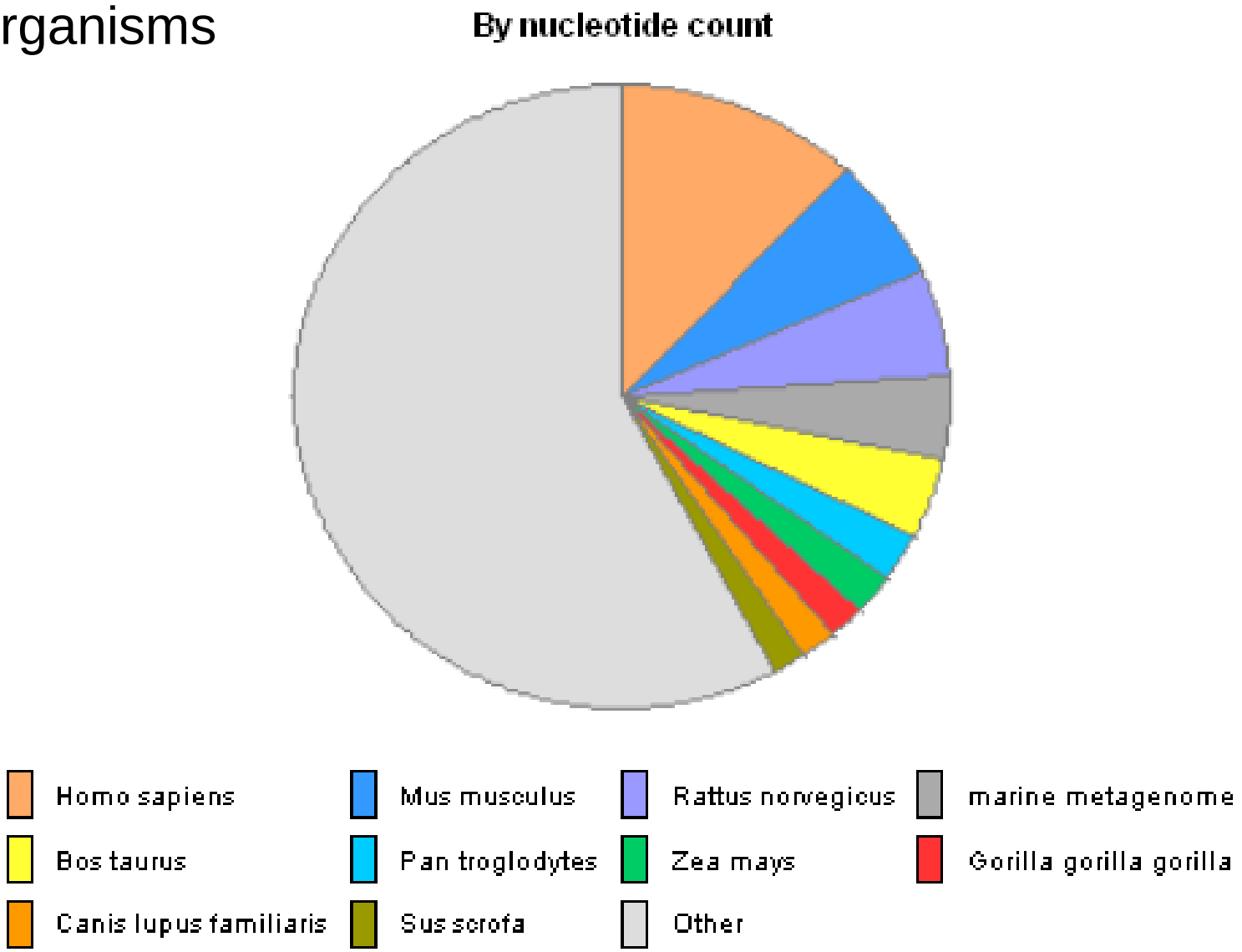
EMBL-EBI databank - Statistics

- To November 22, 2009, the EMBL-EBI Databank contained **266,218,636,744** nucleotides in **164,403,232** entries (http://www.ebi.ac.uk/embl/Services/DBStats/)

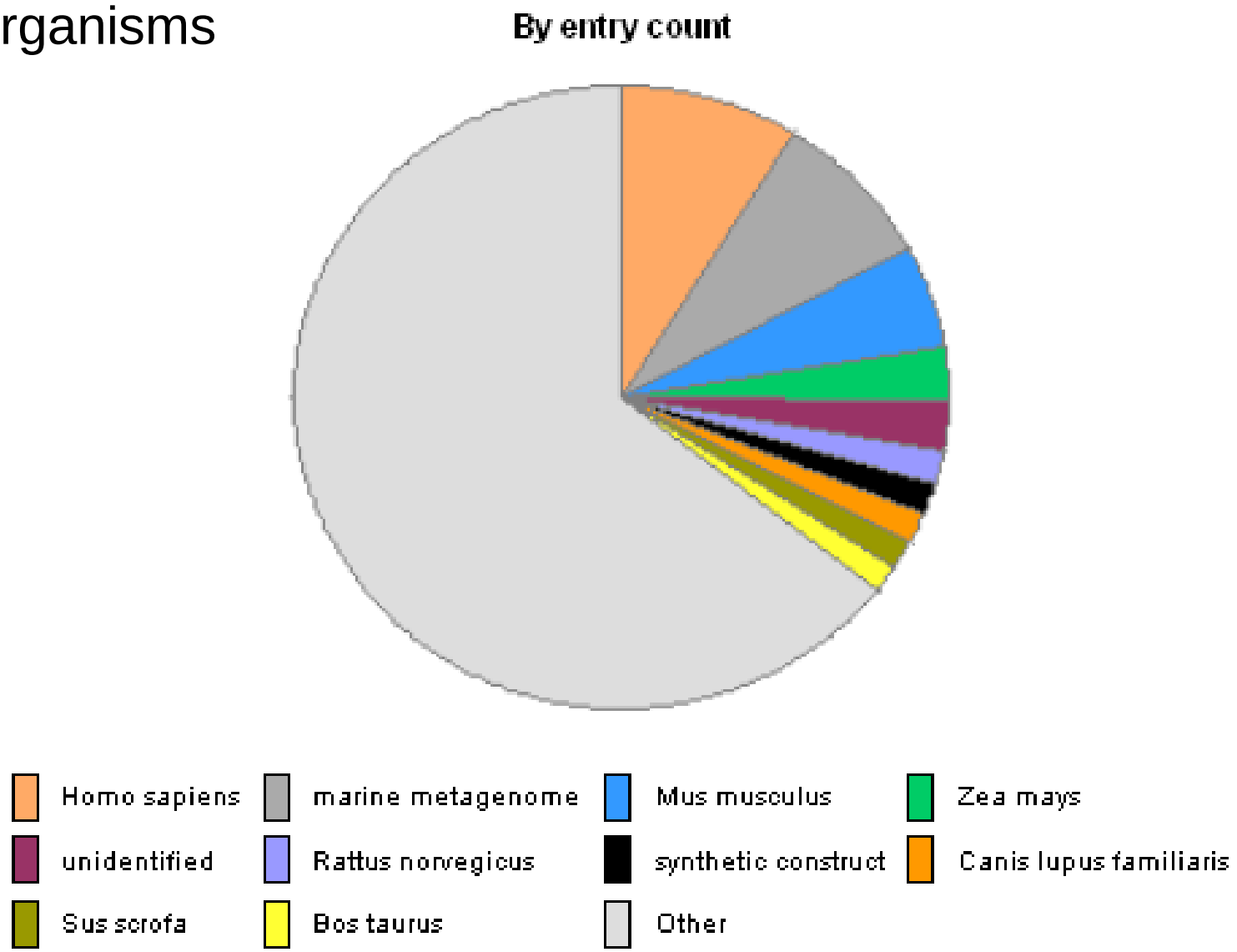
Breakdown by entry type:

Entry Type	Entries	Nucleotides
Standard	112,242,621	109,579,288,879
Constructed (CON)	3,116,508	n/a
Third Party Annotation (TPA)	6,566	372,876,438
Whole Genome Shotgun (WGS)	49,037,537	156,266,471,427

Top organisms

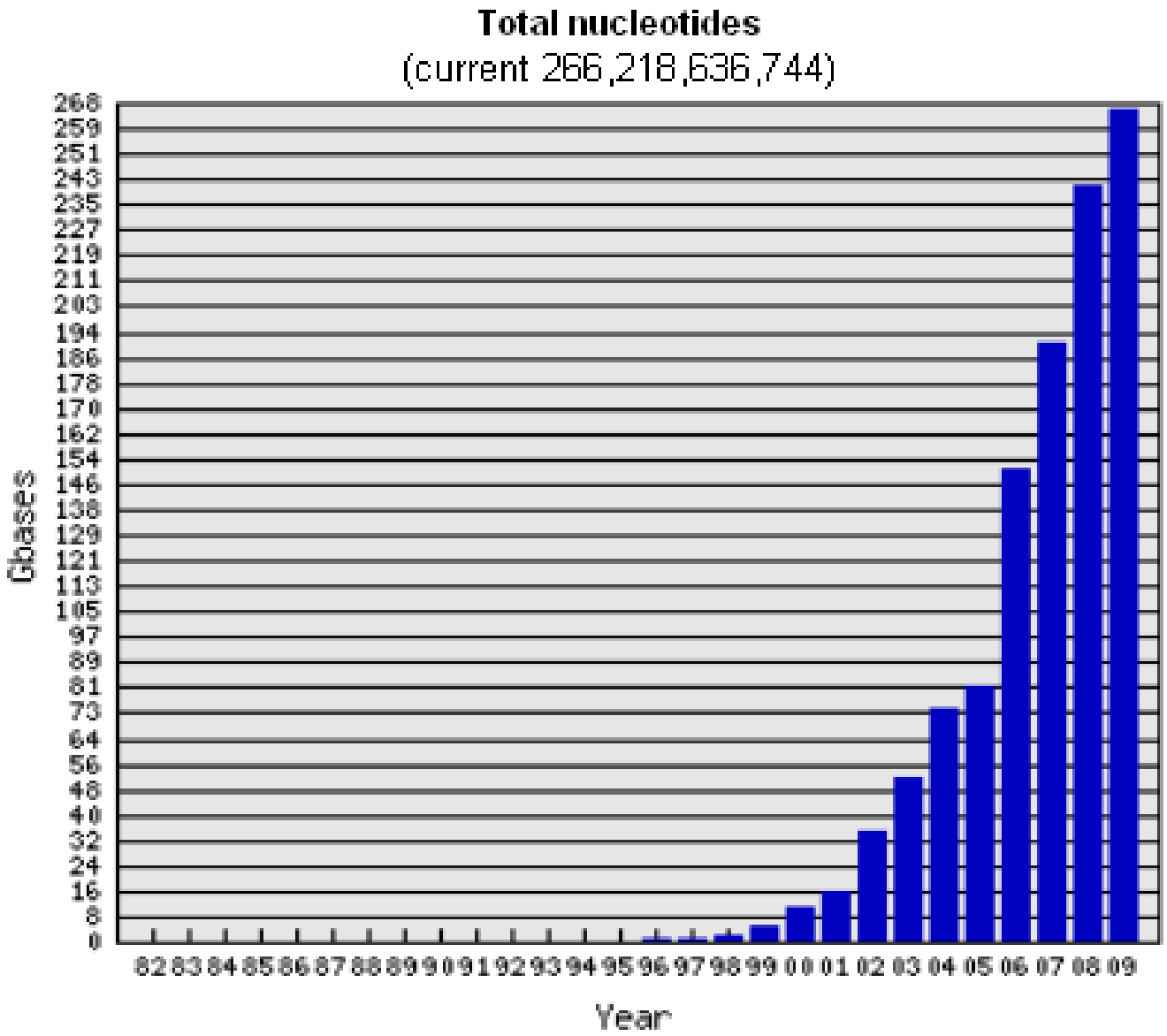


Top organisms



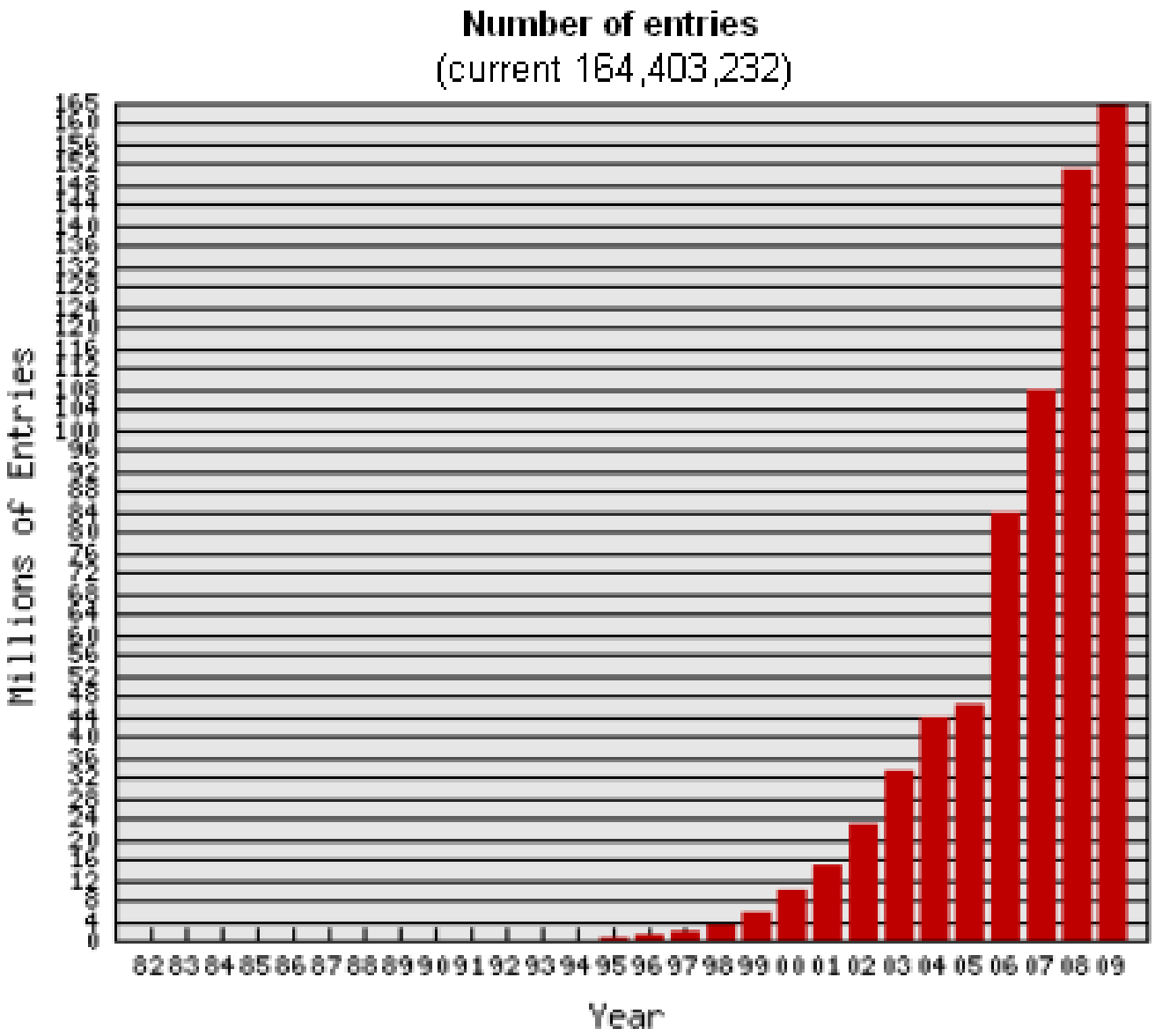
Biomolecular Databanks

EMBL-EBI databank - Statistics - Growth



Biomolecular Databanks

EMBL-EBI databank - Statistics - Growth



UniGene databank

UniGene databank

(*http://www.ncbi.nlm.nih.gov/UniGene/*)

NCBI-UniGene - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indirizzo http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&CID=1634 Vai

NCBI UniGene

PubMed Nucleotide Protein Genome Structure Popset Taxonomy

Search UniGene Go Clear

Limits Preview/Index History Clipboard Details

NCBI

UniGene
Homepage
Query Tips
FAQs
DDD
Download UniGene

Related Resources
LocusLink
HomoloGene
dbEST
Trace Archive
BLAST
CGAP

UniGene Cluster Hs.1634 *Homo sapiens*

CDC25A: Cell division cycle 25A Links

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES

organism, protein and percent identity and length of aligned region

H.sapiens:	pir:A41648 - A41648 protein-tyrosine-phosphatase	98.66 % / 524 aa (see ProtEST)
M.musculus:	sp:P48964 - MPI1_MOUSE M-phase inducer phosphatase 1	81.59 % / 524 aa (see ProtEST)
R.norvegicus:	sp:P48965 - MPI1_RAT M-phase inducer phosphatase 1	84.88 % / 524 aa (see ProtEST)
C.elegans:	ref:NP_491862.1 - protein phosphatase [Caenorhabditis elegans]	36.05 % / 162 aa (see ProtEST)
D.melanogaster:	sp:P20483 - MPIP_DROME M-phase inducer phosphatase	35.34 % / 327 aa (see ProtEST)
S. cerevisiae:	sp:P23748 - MPIP_YEAST M-phase inducer phosphatase	28.74 % / 307 aa (see ProtEST)

MAPPING INFORMATION

Genome View:	3		[MapView]
Genetic map:	3p21 cM		
UniSTS entry:		D3S4246	[Map View]
UniSTS entry:	Chr 3	RH11459	[Map View]
UniSTS entry:	Chr 3	D3S4167	[Map View]

Internet

Biomolecular Databanks
UniGene databank

- The UniGene databank has been created for automatically partitioning the genetic sequences stored in the GenBank primary databank into a non-redundant set of gene-oriented clusters
- Each UniGene cluster represent a unique gene and contains different information:
  - The sequences representing that gene
  - Position of the sequences in the chromosomic map
  - Information correlated to the tissues in which that gene has been found expressed and map location

UniGene databank

- Identification of the human protein codified by that gene and the homologous proteins in other organisms (protein similarity)
- Identification of the ortholog genes of that gene, i.e. the homologous genes in the other species in which they are known. For these ortholog genes, the cluster UniGene Number, Accession Number, and GeneID (if present) of the homologous gene are provided

Biomolecular Databanks
UniGene databank

- The UniGene databank contains the codes of hundred of thousands of Expressed Sequence Tag (EST) sequences, whose attribution to a specific gene is assigned on a statistical basis and has not been proved experimentally yet
- The UniGene databank is generally used by the researcher community as a resource for discovering new genes, or selecting reagents to use in gene mapping projects and large-scale gene expression analyses

UniGene databank - Building procedure

- UniGene datasets are automatically built using several subsequent stages of clustering procedures, with each stage adding less reliable data to the results of the preceding stage
- The used clustering procedures convert sequence discrete similarity scores to boolean links between sequences
- These procedures are still under development and results may change from time to time as improvements are made
- No attempt has been made to produce contigs or consensus sequences. There are several reasons why the sequences of a set may not actually form a single contig:
  - All splicing variants for a gene are put into the same set
  - EST-containing sets often contain 5' and 3' reads from the same cDNA, but such sequences do not always overlap

UniGene databank - Updating

- UniGene clustering results are updated as often as weekly to include GenBank changes
- The new resulting clusters are compared with the preceding week's build and renumbered to maintain continuity
- Since the sequences that make up a cluster may change from week to week, and since the cluster identifier may disappear (typically when two clusters merge), using the UniGene Cluster Identifier as a reference is ill-advised. Using the GenBank accession numbers of the sequences that comprise the cluster is a safe alternative

UniGene databank - Query options

- At UniGene Web site, searches can be performed using the GenBank accession number, or cluster number (UniGene ID), or one or more textual terms
  - Examples of GeneBank Accession Numbers are: AA485353, AA663986, H59260, R435099.
  - UniGene ID must be in the form Xx.#, where # is the cluster number and Xx represents the organism (e.g. Hs.79339, Hs.171995, Rn.43299)
  - Searched terms are extracted from various "plain text" fields, such as definition lines, gene symbols, and protein names

UniGene databank - Content

- UniGene contains sequences and information from several different organisms, as detailed at its home page (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene), including all mayor animal and plant model organisms
- The considered species were chosen because they have the greatest amounts of EST data available and represent a variety of species. Additional organisms will be added in the future

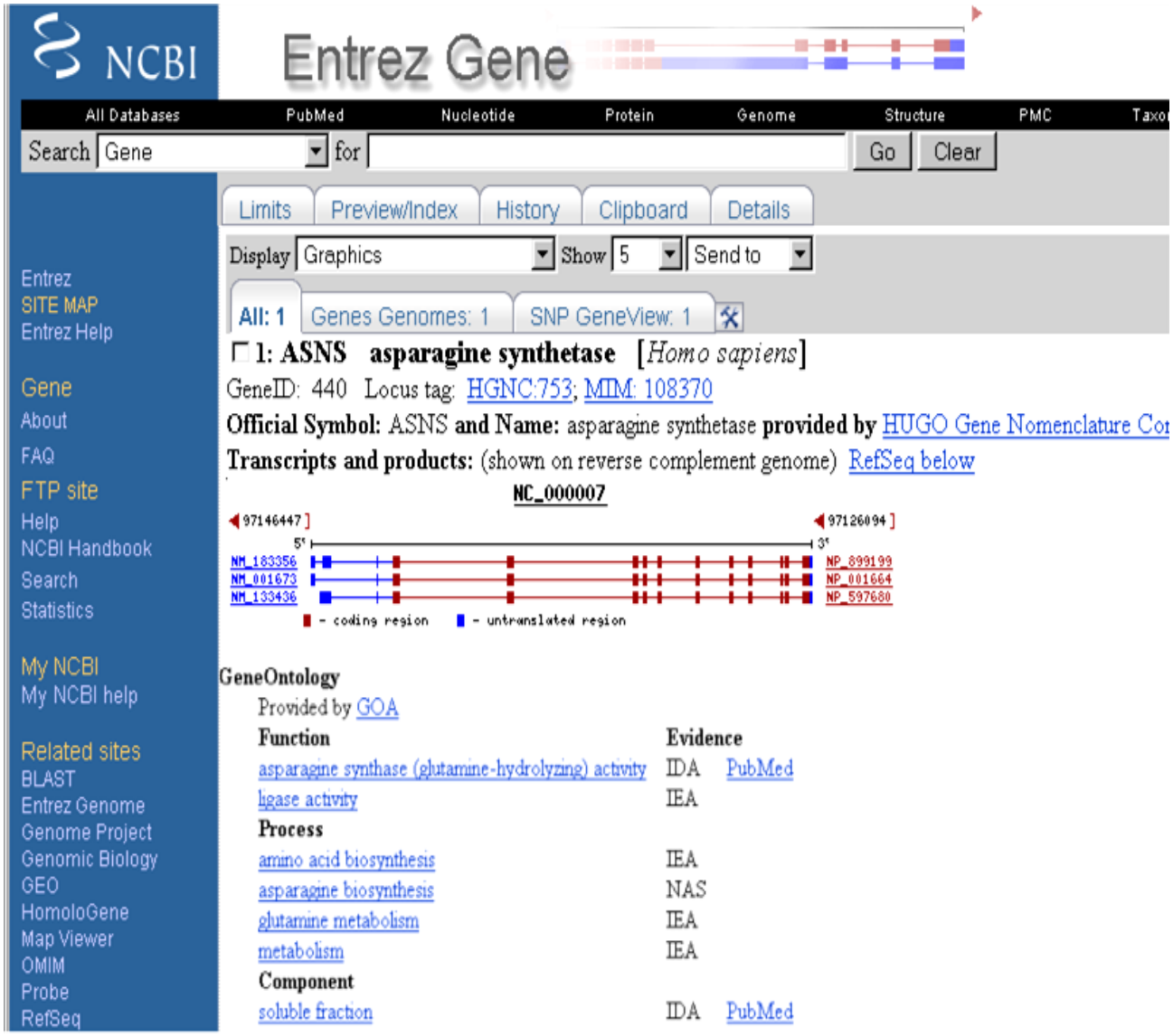
UniGene databank - FTP access

- At ftp://ftp.ncbi.nih.gov/repository/UniGene/ text datasets for each considered organism are available. For Homo sapiens:
  - Hs.info, statistics for the current build
  - Hs.data.gz, complete text of UniGene data
  - Hs.lib.info.gz, information on Library IDs
  - Hs.retired.lst.gz, list of the previous release UniGene clusters for comparison with the current release
  - Hs.seq.all.gz, human transcript sequences derived from both known genes and ESTs
  - Hs.seq.uniq.gz, the one sequences with the longest region of high-quality sequence for each cluster
  - Hs.profiles.gz, expression profile summaries of ESTs in each cluster from libraries with curated controlled vocabulary tissue, organ, or developmental stage of origin

Entrez Gene databank

Entrez Gene databank

(*http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene*)



Biomolecular Databanks
Entrez Gene databank

- Entrez Gene integrates information from the previous LocusLink databank and on genes annotated on Reference Sequences (http://www.ncbi.nlm.nih.gov/RefSeq/) from completely sequenced genomes
- It provides a unified look for gene-specific information independent of the specie(s) of origin
- It also provides a foundation for other functions, namely linkouts from BLAST results and GeneRIFs (http://www.ncbi.nlm.nih.gov/projects/GeneRIF/)

Entrez Gene databank - Query options

- Entrez Gene provides a unified query environment for genes defined by sequence and/or in NCBI's Map Viewer
- It can be searched by:
  - names
  - symbols
  - accessions
  - publications
  - GO terms
  - chromosome numbers
  - EC numbers
  - many other attributes associated with genes and the products they encode

Entrez Gene databank - FTP access

- At ftp://ftp.ncbi.nlm.nih.gov/gene/ a comprehensive extraction of Entrez Gene databank will be provided in tab-delimited text files matching Gene IDs to citation, accession, and name information
- The comprehensive extraction will be formatted in ASN.1, most likely with tools to convert the ASN.1 to XML

Entrez Gene databank - Statistics

- Statistics about records in Entrez Gene are available as:
  - A current snapshot by taxonomic node
  - A history for a single species (NCBI taxonomy ID)
- On November 20, 2009, available entries of main taxa were a total of about 6,613:
  - Archea 145
  - Bacteria 1812
  - Eukaryota 2,300
  - Viroids 1
  - Viruses 2,317
  - Other sequences 38

UniProt databank - Universal Protein Resource

Universal Protein Resource (UniProt) databank

(*http://www.pir.uniprot.org/*)

Biomolecular Databanks
UniProt databank

- Opened on-line on December 15, 2003, the Universal Protein Resource (UniProt) is the world's most comprehensive catalog of information on proteins
- It is a non redundant central repository of protein sequences and functions created by joining the information contained in Swiss-Prot, TrEMBL, and PIR databanks
- The UniProt Consortium is comprised of the EBI - European Bioinformatics Institute (http://www.ebi.ac.uk/), the SIB - Swiss Institute of Bioinformatics (http://www.isb-sib.ch/), and the PIR - Protein Information Resource (http://pir.georgetown.edu/)

UniProt databank - Components

- UniProt is comprised of three components, each optimized for different uses:
  - The **UniProt Archive** (**UniParc**) is a stable, comprehensive sequence collection without redundant sequences reflecting the history of all protein sequences
  - The **UniProt Knowledgebase** (**UniProt**) is the central access point for extensive accurate 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

Biomolecular Databanks
UniProt databank - UniProt archive

- In **UniProt Archive** new and updated protein sequences are loaded daily from public databases including Swiss-Prot, TrEMBL, PIR-PSD, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, and European, American, and Japanese Patent Office proteins
- To avoid redundancy, each unique sequence is stored only once and assigned a unique UniParc identifier. A cross-reference to the database from which the protein sequence has been loaded is created in UniParc
- When different sequence versions exist for the same protein, they are stored in UniParc and a sequence version is made available as part of each database cross-reference

UniProt databank - UniProt knowledgebase

- The **UniProt Knowledgebase** consists of two parts:
  - a section containing fully manually-annotated records resulting from information extracted from literature and curator-evaluated computational analyses
  - a section with computationally-analyzed records awaiting full manual annotation
- For the sake of continuity and name recognition, the two sections are referred to as "Swiss-Prot" and "TrEMBL" respectively

- **UniProt Non-redundant Reference** is composed of three databases, UniRef100, UniRef90 and UniRef50 (which merge all records from all source organisms with mutual sequence identity of 100%, > 90%, or > 50%, respectively, into a single record)
- The three databases provide complete coverage of sequence space while hiding redundant sequences from view
- The non-redundancy allows faster sequence similarity searches by using UniRef90 and UniRef50

UniProt databank - Access

- Protein sequences and annotations in UniProt are accessible via:
  - text search, on numerous database fields (http://www.pir.uniprot.org/search/textSearch.shtml)
  - BLAST similarity search (http://www.pir.uniprot.org/search/blast.shtml)
  - FTP (http://www.pir.uniprot.org/database/download.shtml)

    However, UniProt Archive protein sequences are not available via FTP
- Information is updated daily

UniProt databank - Data format

- **UniProt Knowledgebase** protein annotations are available in XML, FASTA, and Flat File formats. The Flat File format is identical with the former Swiss-Prot and TrEMBL format

- **UniProt UniRef** protein similarity data are available in XML and FASTA formats

Swiss-Prot databank

POLITECNICO DI MILANO

Swiss-Prot databank

(*http://www.expasy.ch/sprot/*)

NiceProt View of SWISS-PROT: Q16719 - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indietro Cerca Preferiti Multimedia

Indirizzo http://www.expasy.org/cgi-bin/niceprot.pl?Q16719

ExPASy Home page | Site Map | Search ExPASy | Contact us | SWISS-PROT

Hosted by SIB Switzerland | Mirror sites: Canada China Korea Taiwan USA

NiceProt View of SWISS-PROT: Q16719

Printer-friendly view | Quick BlastP search

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

General information about the entry	
Entry name	KYNU_HUMAN
Primary accession number	Q16719
Secondary accession numbers	None
Entered in SWISS-PROT in	Release 37, December 1998
Sequence was last modified in	Release 37, December 1998
Annotations were last modified in	Release 40, October 2001
Name and origin of the protein	
Protein name	Kynureninase
Synonyms	EC 3.7.1.3 L-kynurenine hydrolase
Gene name	KYNU
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

References

[1] SEQUENCE FROM NUCLEIC ACID, AND CHARACTERIZATION.
TISSUE=Hepatoma;

Internet

Biomolecular Databanks
Swiss-Prot databank

- Swiss-Prot is a curated and annotated protein sequence databank created in 1986 by the University of Geneve - Swiss Institute of Bioinformatics (SIB) in collaboration with the EMBL - European Bioinformatics Institute (EBI)
- Main characteristics of the Swiss-Prot databank are:
  - High level of annotations (protein functions, domains, post-traslational modifications, variants, etc.)
  - A minimal level of sequence data redundancy
  - High level of integration with other databanks
  - Broad documentation in form of index files and specialized documentation files

Biomolecular Databanks
Swiss-Prot databank

- Data in Swiss-Prot are primarily derived from coding sequence annotations in EMBL- EBI (GenBank/DDBJ) nucleic acid sequence data
- Format of the sequence entries in Swiss-Prot follows as closely as possible that of the EMBL Nucleotide Sequence Databank. As EMBL, Swiss-Prot is a Flat File databank
- For each sequence entry the core data are:
  - Amino acid sequence data
  - Citation information (bibliographical references)
  - Taxonomic data (description of the biological source of the protein)

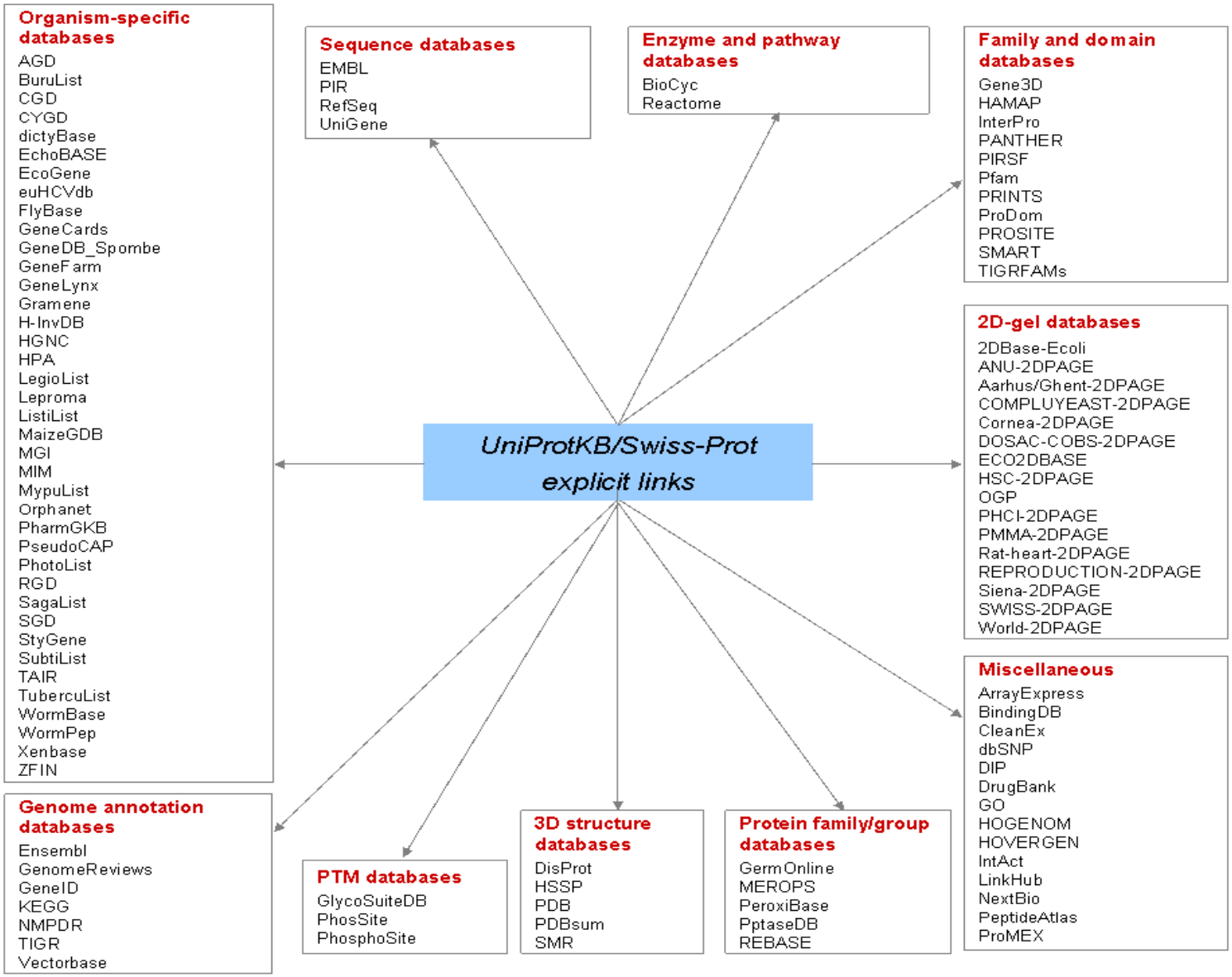
Swiss-Prot databank - Annotations

- For each of the contained proteins, Swiss-Prot provides also many **annotations** regarding:
  - Function/s
  - Post-translational modification/s (PTM) (e.g. phosphorylation)
  - Domains and sites (e.g. ATP-binding sites, zinc fingers)
  - Secondary structure (e.g. alpha helix, beta sheet)
  - Quaternary structure (e.g. homodimer, heterotrimer)
  - Similarity to other proteins
  - Disease/s associated with deficiencies in the protein
  - Sequence variants (e.g. alternative splicing)
  - Sequence conflicts (between papers)

Swiss-Prot databank - Cross-references

- High degree of integration with other databanks is provided between the three types of sequence-related databases (nucleic acid sequences, protein sequences, and protein tertiary structures), as well as with specialized data collections
- Currently, Swiss-Prot is cross-referenced with more than 100 different databanks (i.e. entries have pointers to related information found in other data collections), including:
  - EMBL-EBI, GenBank, DDBJ
  - PIR, PDB
  - OMIM
  - …

Biomolecular Databanks
Swiss-Prot databank - Cross-references



Swiss-Prot databank - Query options

- Swiss-Prot can be interrogated through:
  - The Sequence Retrieval System (SRS)
  - Full text search
  - Taxonomy browser
  - Advanced search by:
    - accession number, or ID
    - description, gene name, and organism
    - author
    - citation

Swiss-Prot databank - Statistics

- On November 22, 2009 (release 57.10) Swiss-Prot contained:
  - Sequence entries: 512,205
  - Amino acids: 180,277,873
  - References: 184,439
  - Represented species: 11,986
    [most represented: Homo sapiens (Human), Mus musculus (Mouse), Arabidopsis thaliana (Mouse-ear cress), Rattus norvegicus (Rat), Saccharomyces cerevisiae (Baker's yeast), Bos taurus (Bovine), Schizosaccharomyces pombe (Fission yeast)]
  - Shortest sequence (GWA_SEPOF): 2 amino acids
  - Longest sequence (TITIN_MOUSE): 35,213 amino acids

Swiss-Prot databank - FTP access and updating

- At ftp://ftp.expasy.org/databases/swiss-prot/ Swiss-Prot text datasets are available.
- **Weekly updates** are also separately available.
- Swiss-Prot is copyright. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified. Usage by and for commercial entities requires a license agreement

TrEMBL databank - Translated EMBL

EMBL-EBI
European Bioinformatics Institute

Get Nucleotide sequences for Go ? Site search Go

Site Map EBI Databa Querie

EBI Home | About EBI | Research | Services | Toolbox | Databases | Downloads | Submissions

UniProt/TrEMBL DATABASE

TrEMBL TRANSLATED EMBL

- TrEMBL Home
- Information
- Access
- Tools
- FTP
- People
- Projects
- Publications
- Documents
- Contact

UniProt/TrEMBL

UniProt/TrEMBL is a computer-annotated protein sequence database complementing the UniProt/Swiss-Prot Protein Knowledgebase.

UniProt/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and also protein sequences extracted from the literature or submitted to UniProt/Swiss-Prot.
The database is enriched with automated classification and annotation.

The UniProt/TrEMBL group is headed by: **Rolf Apweiler**.

The current TrEMBL Release is version 29.1 as of 15-Feb-2005, and contains 1614107 entries... more stats

TrEMBL Release 29.0 of 01-Feb-2005 contained 1589670 entries... more stats

Note: TrEMBL and Swiss-Prot have been incorporated into the UniProt (Universal Protein Resource). The UniProt Release 4.1 consists of: Swiss-Prot Protein Knowledgebase Release 46.1 of 15-Feb-2005 and TrEMBL Protein Database Release 29.1 of 15-Feb-2005.

Access the UniProt/TrEMBL Database

UniProt/Swiss-Prot

The UniProt/Swiss-Prot protein knowledge-base is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases.

Nucleotide DB

The EMBL Nucleotide Sequence Database constitutes Europe's primary nucleotide sequence resource.

Translated EMBL (TrEMBL) databank

(*http://www.ebi.ac.uk/trembl/*)

Biomolecular Databanks
TrEMBL databank

- UniProt/TrEMBL is a computer-annotated protein sequence database complementing the UniProt/Swiss-Prot Protein Knowledgebase
- UniProt/TrEMBL contains the translations of all coding sequences (CDS) in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and also protein sequences extracted from the literature or submitted to UniProt/Swiss-Prot
- The database is enriched with automated classifications and annotations

TrEMBL databank

- Two main sections of the database:
  - **SP-TrEMBL** (Swiss-Prot TrEMBL) contains the entries that will eventually be incorporated into UniProt/Swiss-Prot and can be considered as a preliminary section of UniProt/Swiss-Prot
  - **REM-TrEMBL** (REMaining TrEMBL) contains the entries which will not be included in UniProt/Swiss-Prot; REM-TrEMBL entries have no accession numbers

Biomolecular Databanks
TrEMBL databank

- The main species included are:
  - Homo sapiens
  - Viruses
  - Phages
  - Organelles
  - Archaea
  - Bacteria
  - Eukaryota

TrEMBL databank - Query options

- Query options:
  - Text
  - Accession number
- Search tools:
  - SRS - also used for more complex or multiple database queries
  - UniProt Power Search – It provides full text, advanced search, set manipulation and search filtering on the Universal Protein Resource
  - The ExPASy Server in Geneva - It offers the choice of full-text search or of individual lines
  - SP-ML - the UniProt/Swiss-Prot/TrEMBL in XML format

TrEMBL databank - Updating

- TrEMBL data are updated weekly, and the release is quarterly
- The November 2009 TrEMBL release was Release 40.10 (http://www.ebi.ac.uk/swissprot/sptr_stats/index.html)
- Release 40.10 contained 9,696,103 sequence entries comprising 3,128,947,785 amino acids:
  - 2,403,251 sequences were added since release 40 and the annotations of 5,642,746 entries were revised, with an increase of 32%

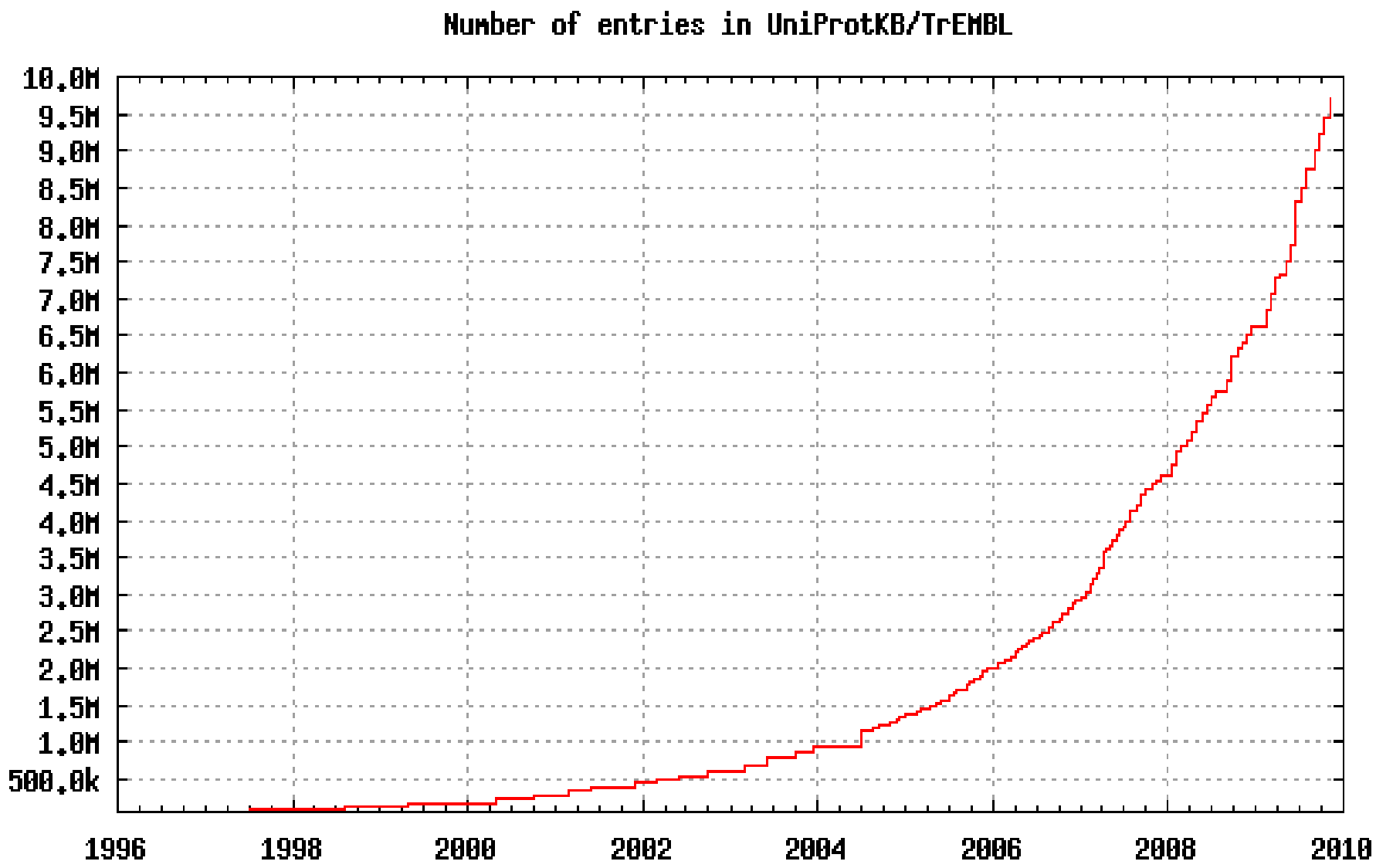
TrEMBL databank - FTP access and formats

- The data are all available for FTP download from the directory: ftp://ftp.ebi.ac.uk/pub/databases/trembl/

- The data are provided in the following format:
  - Xml
  - .*dat.gz
  - Swiss-Prot flat file
  - FASTA

Biomolecular Databanks

TrEMBL databank - Statistics - Growth



PIR databank - Protein Information Resource

Protein Information Resources (PIR) databank

(*http://www-nbrf.georgetown.edu/pirwww/*)

PIR Entry Request - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indietro Cerca Preferiti Multimedia

Indirizzo http://www-nbrf.georgetown.edu/cgi-bin/nbrfget Vai

pir.georgetown.edu

PIR Entry A41648

Site Map Site Search

Text Search Protein Databases: GO!

About PIR | Databases | Search & Retrieval | Download | Support

Annotation || Sequence || Composition Table || Download: (CODATA || FASTA || XML)

Create a BLAST (at PIR) submission form for A41648 Submit

```
ENTRY            A41648  #type complete    iProClass View of A41648
TITLE            protein-tyrosine-phosphatase (EC 3.1.3.48) cdc25A - human
ALTERNATE_NAMES  cell division cycle protein cdc25A
ORGANISM         #formal_name Homo sapiens #common_name man
   #cross-references taxon:9606
DATE             28-May-1992 #sequence_revision 25-Apr-1997 #text_change
                 11-Jun-1999
ACCESSIONS       A41648
REFERENCE        A41648
   #authors      Galaktionov, K.; Beach, D.
   #journal      Cell (1991) 67:1181-1194
   #title        Specific activation of cdc25 tyrosine phosphatases by
                 B-type cyclins: evidence for multiple roles of mitotic
                 cyclins.
   #cross-references MUID:92103683; PMID:1836978
   #accession    A41648
      ##molecule_type mRNA
      ##residues 1-523 ##label GAL
      ##cross-references GB:M81933; NID:g180170; PIDN:AAA58415.1;
                 PID:g180171
COMMENT          This protein can be activated by association with cyclin
                 B.
GENETICS
   #gene         GDB:CDC25A
      ##cross-references GDB:133773; OMIM:116947
```

Internet

Biomolecular Databanks
PIR databank

- The Protein Information Resource (PIR) is a division of the National Biomedical Research Foundation (NBRF) (http://www-nbrf.georgetown.edu/nbrf/) which is affiliated with Georgetown University Medical Center
- The Resource was established in 1984 to assist researchers in the identification and interpretation of protein sequence information and to support genomic/proteomic research on molecular evolution, functional genomics, and computational biology

Biomolecular Databanks
PIR databank

- The mission of PIR is to provide an integrated public resource of functional annotated protein sequences, non redundant, complete and cross-referenced, where entries are organized in “superfamilies”
- It is empowered with analysis tools for identifying and analyzing protein sequences and their nucleotide correspondence
- PIR is a system composed by several databases of nucleotide and amino acidic sequences

PIR databank - Components

- PIR-International maintains into Oracle object-relational DBMS a set of related protein sequence databases:
  - The PIR **Protein Sequence Database (PSD)** of functionally annotated protein sequences at http://www-nbrf.georgetown.edu/pirwww/search/textpsd.shtml
  - the PIR **Non-Redundant Reference** Sequence Database (PIR-NREF) for protein sequence identification at http://www-nbrf.georgetown.edu/pirwww/search/pirnref.shtml
  - the **International Protein Classification** Database (iProClass) at http://www-nbrf.georgetown.edu/iproclass for comprehensive structural/functional features and family relationships of proteins

Biomolecular Databanks
PIR databank - Components

- the PIR Sequence-Structure database (PIR-NRL3D) at http://www-nbrf.georgetown.edu/pirwww/search/textnrl3d.html
- the PIR Alignment database (PIR-ALN) at http://www-nbrf.georgetown.edu/pirwww/search/textpiraln.html
- the PIR database of amino acid modifications (PIR-RESID) http://www-nbrf.georgetown.edu/pirwww/search/textresid.html

Biomolecular Databanks
PIR databank - PIR-PSD

- The **PIR PSD** (Protein Sequence Database), distributed also in XML format, is the most comprehensive and expertly annotated protein sequence database in the public domain
- Its mission is to achieve the properties of comprehensiveness, timeliness, non-redundancy, quality annotation, and full classification of amino acid sequences
- PSD was updated biweekly. Release 80.00 (31 December, 2004) was the final release; it contained:
  - 283’416 sequences
  - 96’134’583 residues
  - 36’287 superfamilies

Biomolecular Databanks
PIR databank - PIR-NREF

- The **PIR-NREF** (Non-redundant REFerence), a comprehensive database for sequence searching and protein identification, contains non-redundant protein sequences from PIR-PSD, Swiss-Prot, PDB, TrEMBL, RefSeq, and GenPept
- Identical sequences from the same source organism (species) reported in different databases are presented as a single NREF entry with protein IDs and names from each underlying database, in addition to protein sequence, taxonomy, and composite bibliography
- It is updated biweekly.

PIR databank - Query options

- The PIR PSD databank can be interrogated by:
  - text searching of selected database fields
  - several identifiers including:
    - PIR unique ID (e.g. CCHU)
    - PIR accession or reference number (e.g. A41648)
    - GenBank accession number (e.g. M64864)
    - Protein identifier or protein_id (e.g. AAA17758.1)
    - Protein Data Bank (PDB) identifier
    - TIGR identifier (e.g. MG022)
    - Genome Data Bank (GDB) accession, PubMed ID

Biomolecular Databanks
PIR databank - Annotations

- PIR provides a **batch search** option of sequences and complete annotations for PSD and NREF
- In PIR-NREF, sequence search, based on BLAST, is also available and alignments of the results are provided
- PIR retrieved data can be displayed either in XML, FASTA, CODATA, CODATA/HTML, or NBRF/PIR format and include information on protein superfamily, title, species, taxonomy group, and sequence similarity
- In PIR, lists of complete genomes, species, keywords, superfamilies, homology domains, gene names, or journal names are also available

PIR databank - FTP access and data formats

- The releases of the PIR PSD, PIR-NREF and other databases (PIR-NRL3D, PIR-ALN, PIR-RESID) are available for downloading from the PIR anonymous FTP server at ftp://ftp.pir.georgetown.edu/pir_databases/ using:

  Login: anonymous Password: email address

- Downloading available formats are:
  - for PIR PSD: XML, FASTA, CODATA, NBRF/PIR
  - for PIR-NREF: XML, FASTA

PDB databank - Protein Data Bank

Protein Data Bank (PDB)

(*http://www.rcsb.org/pdb/*)

Structure Explorer - 1C25 - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indietro Cerca Preferiti Multimedia

Indirizzo http://www.rcsb.org/pdb/cgi/explore.cgi?job=summary&pdbId=1C25&page=&pid=212711036003923 Vai

PDB PROTEIN DATA BANK

Structure Explorer - 1C25

Summary Information

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details

Explore
SearchLite SearchFields

Title: Human Cdc25A Catalytic Domain

Compound: Mol_Id: 1; Molecule: Cdc25A; Chain: Null; Fragment: Catalytic Domain; Synonym: M-Phase Inducer Phosphatase 1; Ec: 3.1.3.48; Engineered: Yes; Mutation: Ins(M335)

Authors: E. B. Fauman, J. P. Cogswell, B. Lovejoy, W. J. Rocque, W. Holmes, W. Holmes, V. G. Montana, H. Piwnica-Worms, M. J. Rink, M. A. Saper

Exp. Method: X-ray Diffraction

Classification: Hydrolase

EC Number: 3.1.3.48

Source: Homo sapiens

Primary Citation: Fauman, E. B., Cogswell, J. P., Lovejoy, B., Rocque, W. J., Holmes, W., Montana, V. G., Piwnica-Worms, H., Rink, M. J., Saper, M. A.: Crystal structure of the catalytic domain of the human cell cycle control phosphatase, Cdc25A. *Cell* 93 *pp.* 617 (1998) [Medline]

Deposition Date: 17-Apr-1998 *Release Date:* 19-Aug-1998

Resolution [Å]: 2.30 *R-Value:* 0.227

Space Group: P 41

Unit Cell: *dim [A]:* *a* 43.51 *b* 43.51 *c* 117.10

angles [°]: *alpha* 90.00 *beta* 90.00 *gamma* 90.00

Polymer Chains: 1C25 *Residues:* 161

Internet

Biomolecular Databanks
PDB databank

- PDB is the single freely accessible worldwide repository for the processing and distribution of the 3-D structure data of biological macromolecules, such as:
  - Proteins
  - Nucleic acids
  - Protein-nucleic acid complexes
  - Viruses
- The PDB contents are primarily experimental data derived from X-ray crystallography and NMR experiments

Biomolecular Databanks
PDB databank

- For each contained structure, they are provided:
  - Sequence details
  - Atomic coordinates
  - Crystallization conditions
  - 3-D structure neighbors computed with various methods
  - Derived geometric data
  - Structure factors
  - 3-D images
  - Several links to other resources

PDB databank - Main features

- The **primary goals** of PDB are:
  - To enable locating structures of interest
  - To perform simple analyses on one or more structures
  - To act as a portal to additional information available on the Internet
  - To enable downloading information on a structure, notably the Cartesian atomic coordinates, for further analysis

Biomolecular Databanks
PDB databank - Annotations

- The PDB supports several data formats for representing structures, sequences, and graphical displays
- Single structure ASCII text files are available compressed for download in PDB format or in mmCIF (macromolecular Crystallographic Information File) format
- Protein sequences in FASTA format for display and download
- Default graphics format is the structure PDB format. The produced view can be modified, both in appearance and orientation of the molecule using Molscript and RasMol Scripting languages. Virtual Reality Modeling Language (VRML) is used for some display purposes

PDB databank - Components

- The PDB requires additional free tools to be installed beyond a Web browser to take full advantage of the PDB interface:
  - A Java capable and enabled Web browser. Without Java enabled, the QuickPDB option under "View Structure" can not be used
  - A VRML plug-in to be used with "View Structure"
  - The RasMol molecular display program to be used with "View Structure" and "Geometry"
  - A Chime plug-in to be used with the "First Glance" and "Protein Explorer" options under "View Structure" (requires Netscape Web browser)

PDB databank - Query options

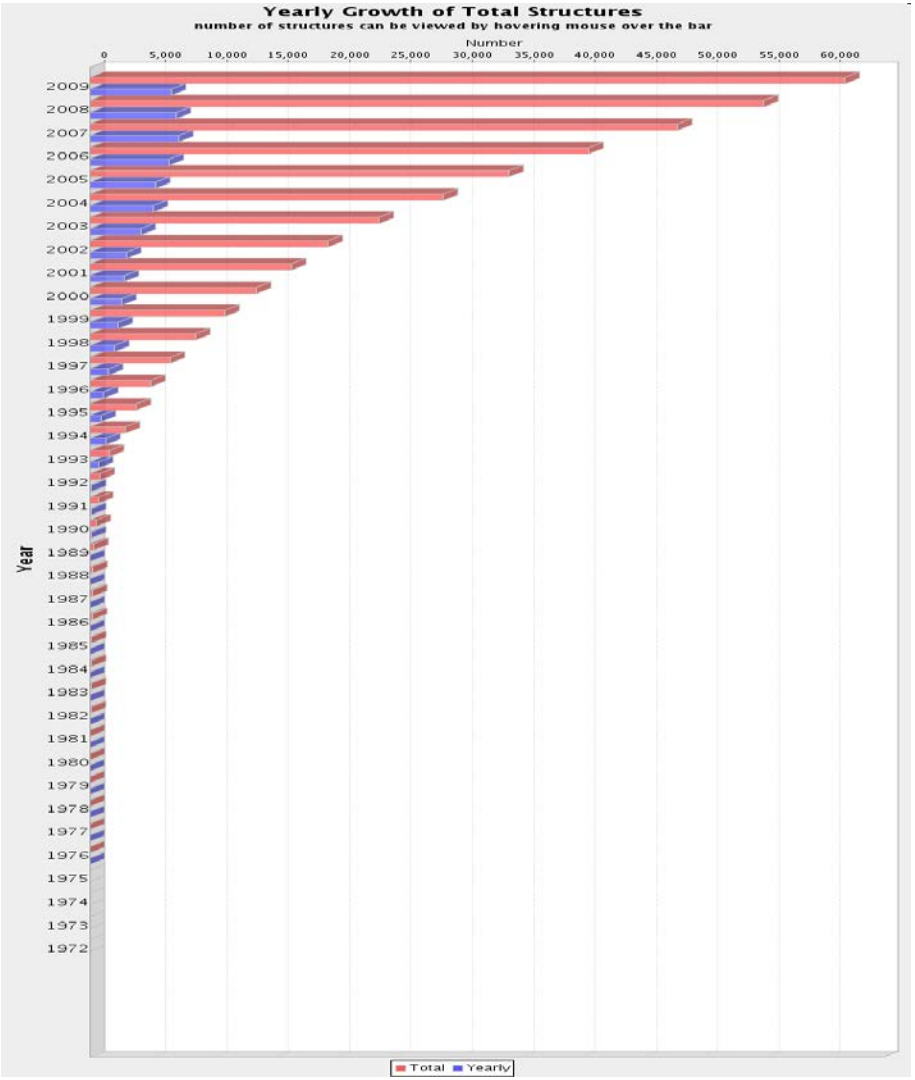
- The PDB can be interrogated by searching:
  - PDB identification code (e.g. 4hhb, 9ins, 1aha)
  - the text in PDB files (e.g. protein kinase, ribosome)
  - the text of both mmCIF files and the Web pages
  - against specific fields of information (e.g. author, deposition date)
  - on an entry status (e.g. processing, on hold, released)
  - iteratively on a previous search

PDB databank - Query options

- Search results: when multiple structures are returned, useful options are available:
  - Download results as a single compressed file containing the PDB files of all the returned structures
  - Summarize results in a variety of tabular reports based on: structure identifiers, sequence, experimental techniques, crystallographic cell dimensions, data collection methods, refinement details, primary citation information
- Combining text searching of multiple PDB ID and multiple result options, a kind of batch search can be performed

PDB databank - Statistics

- On November 17, 2009 PDB held 61,567 molecular structures:
  - Proteins: 56,951
  - Protein complexes: 2,515
  - Nucleic acids: 2,074
- Of these, 8,118 were defined by NMR and 53,020 by X-ray diffraction and other techniques



Biomolecular Databanks
PDB databank - FTP access

- PDB data and structure files can be obtained via the FTP server at ftp://ftp.wwpdb.org/pub/pdb/data/
- Software provided by PDB can be downloaded at ftp://ftp.wwpdb.org/pub/pdb/software/

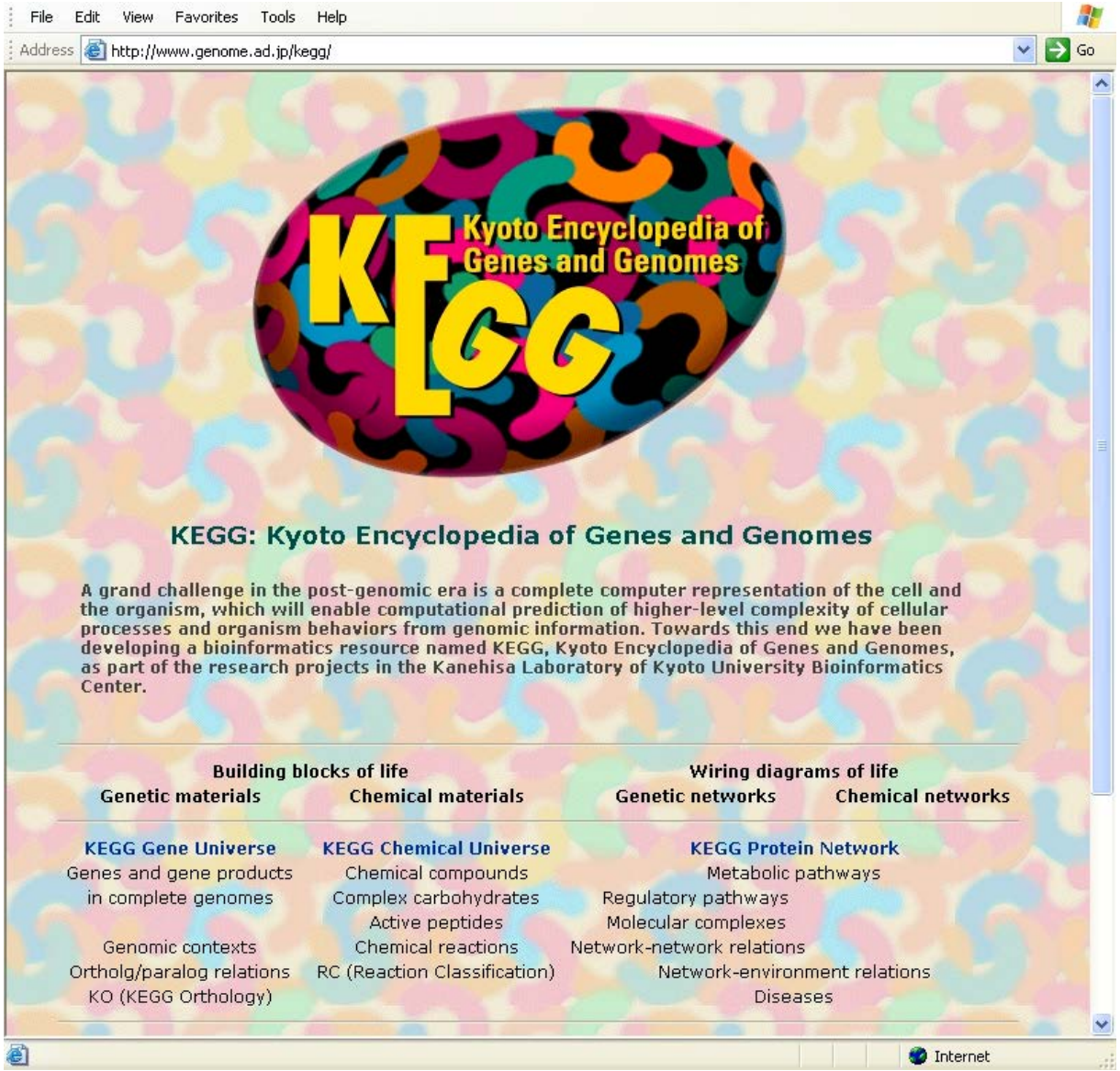
Biomolecular Databanks
PDB databank - Data updating

- Approximately 50-100 new structures are deposited each week by the international user community
- They are annotated by the Research Collaboratory for Structural Bioinformatics (RCSB) and released upon the depositor's specifications
- The PDB databank is constantly updated as new structures are deposited by the international scientific community
- Data files in the FTP site are updated quarterly

KEGG databank - Kyoto Encyclopedia of Genes and Genomes

Kyoto Encyclopedia of Genes and Genomes (KEGG) databank

(*http://www.genome.ad.jp/kegg/*)



Biomolecular Databanks
KEGG databank

- A grand challenge in the post-genomic era is a complete computer representation of the cell and the organism, which will enable computational prediction of higher-level complexity of cellular processes and organism behavior from genomic information
- Toward this end KEGG has been developing a knowledge-based approach for network prediction, which is to predict, given a complete set of genes in the genome, the protein interaction networks that are responsible for various cellular processes

Biomolecular Databanks
KEGG databank

- KEGG is the reference knowledge base that integrates current knowledge on molecular interaction networks such as pathways and complexes (PATHWAY database), information about genes and proteins generated by genome projects (GENES/SSDB/KEGG Orthology databases), and information about biochemical compounds and reactions (COMPOUND/GLYCAN/REACTION databases)
- New efforts are being made to abstract knowledge, both computationally and manually, about ortholog clusters in the KEGG Orthology database, and to collect and analyze carbohydrate structures in the GLYCAN database

Biomolecular Databanks
KEGG databank - Access

- The primary access mode to KEGG is through the GenomeNet Website at http://www.genome.ad.jp/kegg/
- Different KEGG resources can be accessed from KEGG table of contents at http://www.genome.ad.jp/kegg/kegg2.html
- KEGG graph objects are available in XML KEGG Markup Language (KGML) at http://www.genome.ad.jp/kegg/xml/
- FTP access is available at: http://www.genome.ad.jp/anonftp/
- For computerized access to KEGG, the SOAP server is open to academic users at http://www.genome.ad.jp/kegg/soap/

KEGG databank - Query options

- A gene in the following queries must be specified by the GENES entry identifier in the form of org:gene, where org is the three-letter KEGG species code and gene is the accession number, such as hsa:3096

Find GENES entry identifier: (enter keywords)

Go Clear

Search against: All species

KEGG species (three-letter code such as hsa) Help

A gene in the following queries must be specified by the GENES entry identifier in the form of **org:gene** where **org** is the three-letter KEGG species code and **gene** is the accession number, such as eco:b0015.

Biomolecular Databanks
KEGG databank - GENES database

- On November 22, 2009 the GENES database contained information about individual 5,135,391 genes in 1,120 organisms (121 eukaryotes + 931 bacteria + 68 archaea)
- GENES entries are created semi-automatically by selecting and joining various sources including authors' submissions to GenBank (ftp://ftp.ncbi.nih.gov/genbank/genomes/), the RefSeq database (ftp://ftp.ncbi.nih.gov/genomes/), the EMBL database (ftp://ftp.ebi.ac.uk/pub/databases/embl/genomes/), and publicly available organism-specific databases.

  They are then subjected to internal re-annotation, in which KEGG curators assign KEGG numbers for the KEGG Orthology grouping of genes without updating the description of the genes

KEGG databank - Statistics

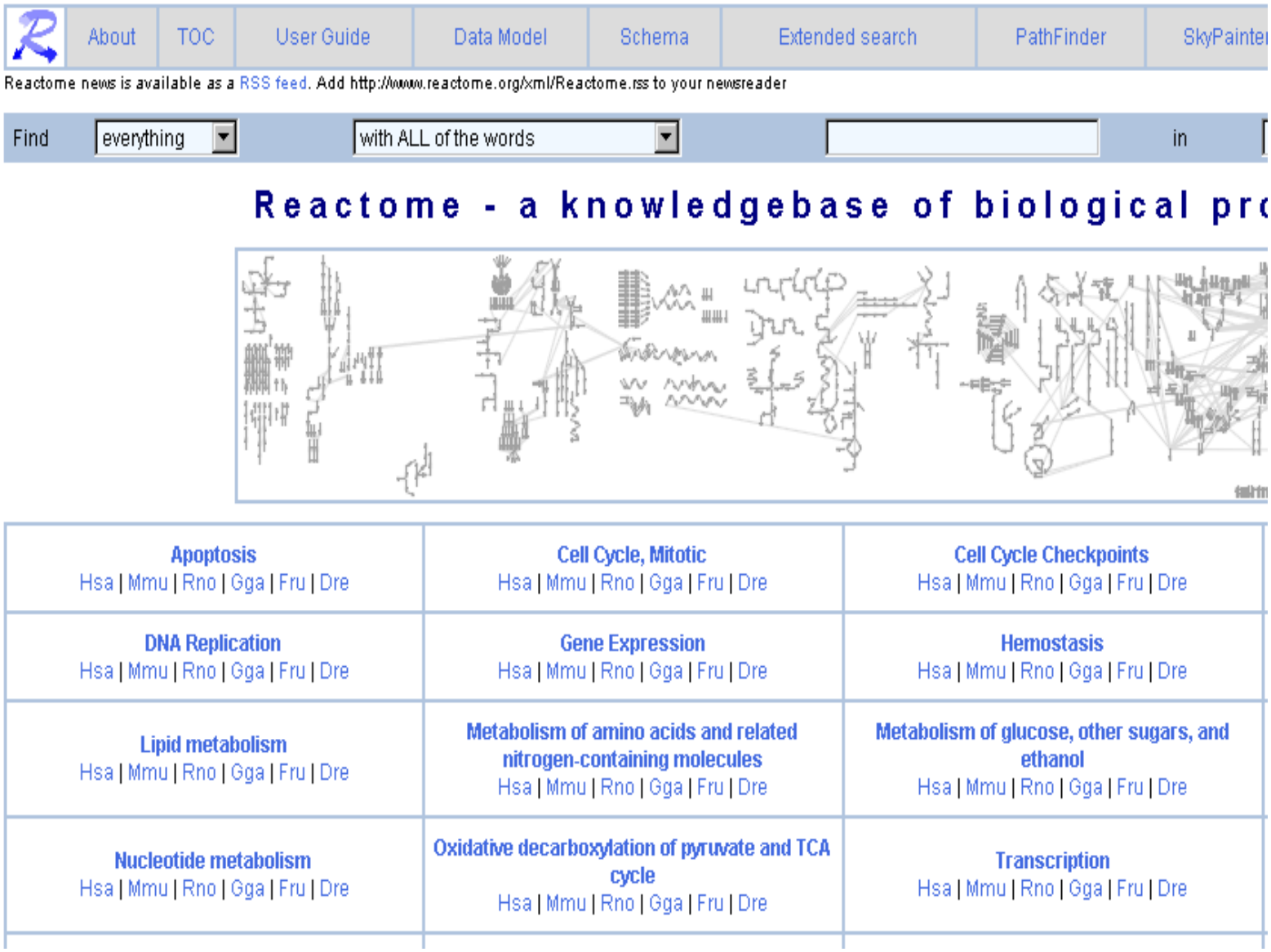
- On November 22, 2009 the data collection was as it follows:

Number of pathways	97,355 (PATHWAY database)
Number of reference pathways	342 (PATHWAY database)
Number of organisms	1,215 (GENOME database)
Number of genes	5,135,391 (GENES database)
Number of ortholog groups	12,833 (ORTOLOGY database)
Number of chemical compounds	16,054 (COMPOUND database)
Number of glycans	18,969 (GLYCAN database)
Number of chemical reactions	8,052 (REACTION database)
Number of reactant pairs	11,941 (RPAIR database)
Number of pathway modules	704 (MODULE database)
Number of human diseases	114 (DISEASE database)
Number of drugs	9,148 (DRUG database)
Number of enzyme terms	5,074 (ENZYME database)

Reactome databank - A knowledbase of biological processes

Reactome databank

(*http://www.reactome.com/*)



Biomolecular Databanks
Reactome databank

- Reactome is a curated database of biological processes in humans
- It covers biological pathways ranging from the basic processes of metabolism to high-level processes such as hormonal signalling
- While Reactome is targeted at human pathways, it also includes many individual biochemical reactions from non-human systems such as rat, mouse, fugu fish and zebra fish: this makes the database relevant to the large number of researchers who work on model organisms

Biomolecular Databanks
Reactome databank

- All the information is backed up by its provenance (a literature citation or an electronic inference based on sequence similarity)
- The basic information is provided by bench biologists who are experts in that domain of biology
- The entire set of human pathways known to the database are represented as a series of constellations in a "*starry sky*", which can be used to navigate through the universe of human reactions and is invaluable to visualize connections between pathways, some of which will be surprising to biologists who are not familiar with pathways outside their domain of research

Reactome databank - Query options

- Query keywords:
  - All text
  - Accession numbers
  - E.C. numbers
  - Swiss-Prot IDs

- Provided data format:
  - *.smbl.gz
  - Text

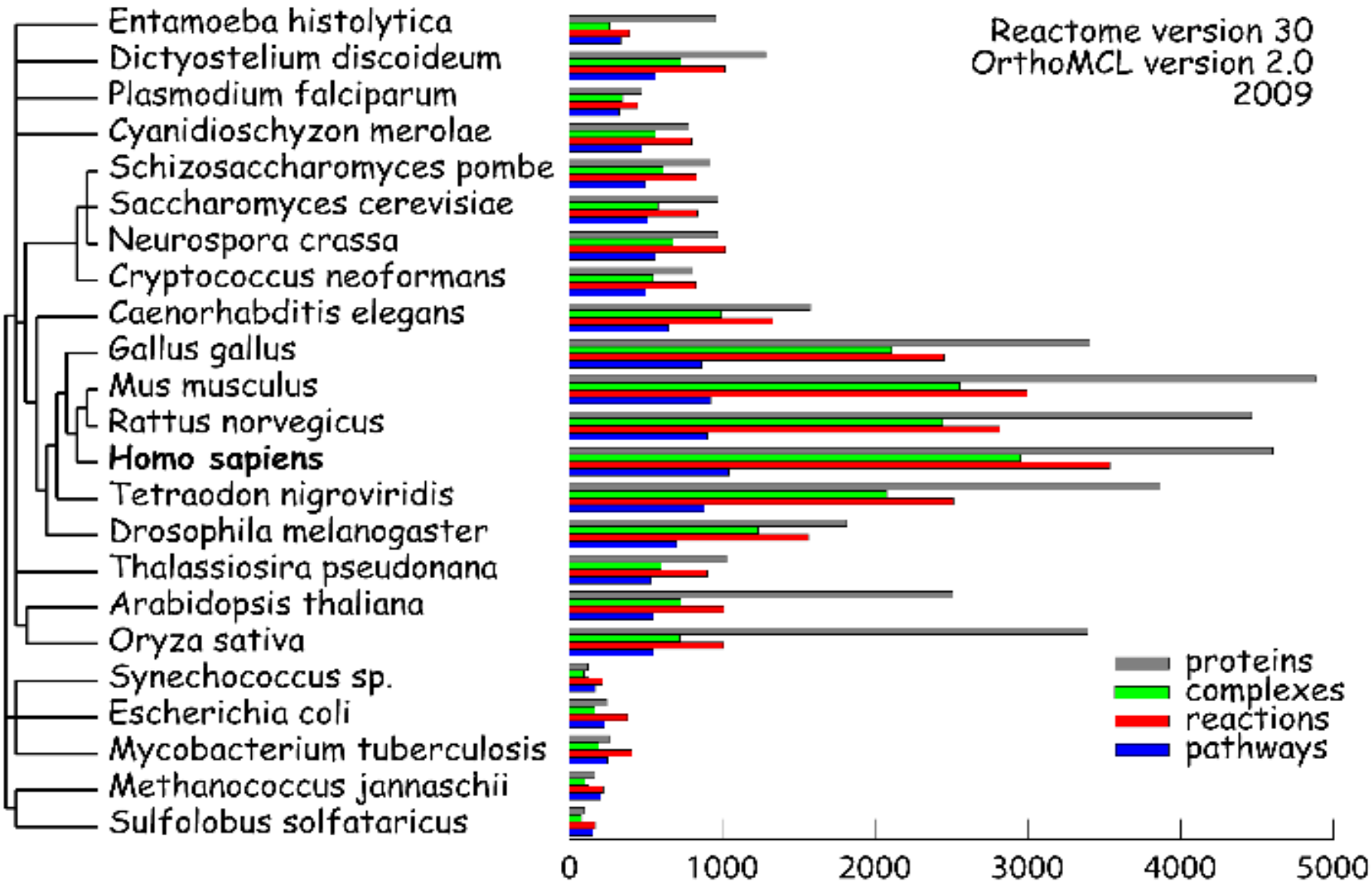
- Some data are available for downloading from: http://www.reactome.com/download/index.html

Reactome databank - Statistics

Reactome statistics (version 30):

Species	PROTEINS	COMPLEXES	REACTIONS	PATHWAYS
E. histolytica	887	264	394	338
D. discoideum	1191	732	1021	563
P. falciparum	446	349	447	328
C. merolae	722	564	802	476
S. pombe	854	617	834	497
S. cerevisiae	922	582	841	510
N. crassa	898	678	1022	562
C. neoformans	747	548	829	498
C. elegans	1462	995	1336	649
G. gallus	3109	2110	2456	867
M. musculus	4844	2557	3001	927
R. norvegicus	4378	2444	2819	904
H. sapiens	3916*	2955	3541	1045
T. nigroviridis	3740	2079	2517	883
D. melanogaster	1733	1237	1567	702
T. pseudonana	987	600	904	534
A. thaliana	2330	731	1014	549
O. sativa	3148	723	1009	549
S. sp	118	95	220	170
E. coli	240	167	382	230
M. tuberculosis	258	193	409	251
S. solfataricus	161	100	223	207
M. jannaschii	94	80	170	156

Biomolecular Databanks
Reactome databank - Statistics



GOA databank - Gene Ontology Annotation

EMBL-EBI
European Bioinformatics Institute

Get Nucleotide sequences for Go ? S

EBI Home | About EBI | Research | Services | Toolbox | Databases | Downloads | Submissions

GOA DATABASE

- GOA Home
- Introduction
- Contents of Current Release
- Data Searching and Retrieval
- Forthcoming Changes
- GOA News
- Feedback

GOA @EBI

GOA is a project run by the European Bioinformatics Institute that aims to provide assignments of gene products to the Gene Ontology (GO) resource.

The goal of the Gene Ontology Consortium is to produce a dynamic controlled vocabulary that can be applied to all organisms, even while knowledge of gene and protein roles in cells is still accumulating and changing. In the GOA project, this vocabulary will be applied to a non-redundant set of proteins described in the UniProt Resource (Swiss-Prot/TrEMBL/PIR-PSD) and Ensembl databases that collectively provide complete proteomes for Homo sapiens and other organisms.

In the first stage of this project, GO assignments have been applied to a data set representing the human proteome by a combination of electronic mappings and manual curation. Subsequently, GO assignments for all complete and incomplete proteomes that exist in UniProt have been provided. GOA will be updated monthly in accordance with the latest data released by the primary data sources.

- Detailed project outline
- What can I do with GOA?

The GOA Project is headed by **Rolf Apweiler**.

Access GOA

Search GOA via SRS

Go

GO

The EBI's Gene Ontology Consortium pages. GO is an international consortium of scientists with the editorial office based at the EBI.

UniProt/ Swiss-Prot

Gene Ontology Annotation (GOA) databank

(*http://www.ebi.ac.uk/GOA/*)

Biomolecular Databanks
GOA databank

- GOA is a project run by the EBI - European Bioinformatics Institute that provides assignments of gene products to the Gene Ontology (GO) resource
- The goal is to produce a dynamic controlled vocabulary that can be applied to all organisms
- This vocabulary will be applied to a non-redundant set of proteins described in the UniProt Resource and Ensembl databases that collectively provide complete proteomes for Homo sapiens and other organisms

GOA databank

- GOA allows to:
  - access functional information for the human proteome (GOA-Human) or for any protein in EBI's protein databases (GOA-UniProt)
  - ask complex questions such as '"find all proteins involved in apoptosis(GO:0006915) but not involved via death domain receptors (GO:0008625), and then find their coding sequences"
  - use GO-Slim to summarize the biological attributes of a proteome, compare proteomes, or find out what proportion of a proteome is involved

Biomolecular Databanks
GOA databank

- incorporate manual annotation into customer databases to enhance datasets, or use it to validate automated way of deriving information about gene function
- map GO terms to customer datasets
- find the location of human genes mapped to a particular GO term using Ensembl GO-View

GOA databank - FTP site and included species

- The FTP site is ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/

- The main species included are:
  - Homo sapiens
  - Rattus norvegicus
  - Mus musculus
  - Arabidopsis thaliana
  - Zebrafish

GOA databank - Query options

- Query options:
  - Text
  - Go terms
  - GO annotations
  - Keywords

- Downloadable data format:
  - Text or databases export to GO a tab delimited file
  - *.goa.gz

GOA databank - Statistics

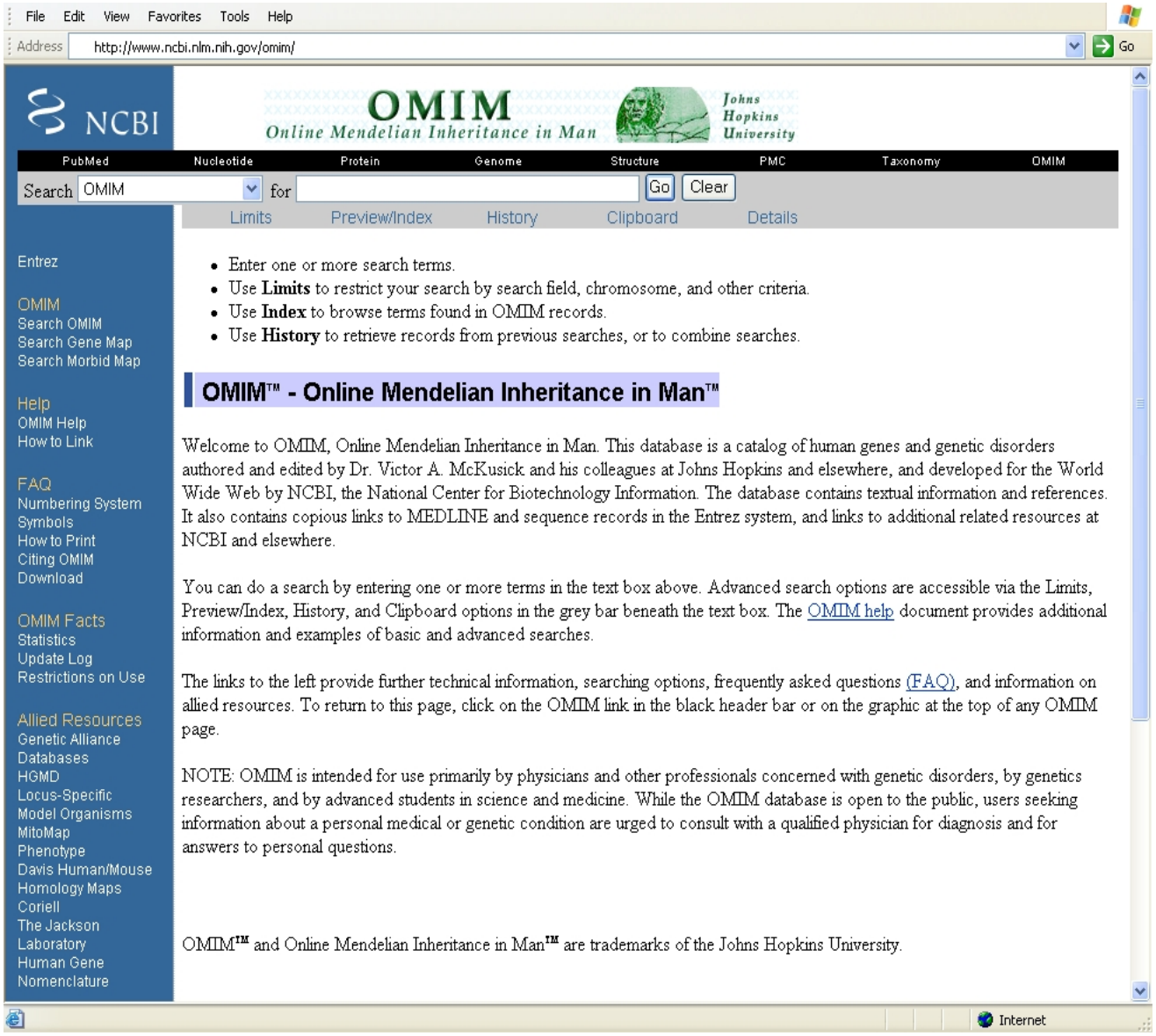
- Statistics of GOA Human 78, released on 8 October, 2009 were:

GO Annotation Source	Number of Associations	Number of Distinct Proteins
Electronic GO annotation using InterPro to GO mapping	29512	12202
Electronic GO annotation using Swiss-Prot keyword to GO mapping	41358	15276
Electronic GO annotation using UniProt Subcellular Location to GO mapping	2287	1915
Electronic GO annotation using EC to GO mapping	707	638
Electronic GO annotation using Compara projections	22829	4702
Total Electronic GO annotation	96693	17468
Manual GO annotation by Proteome Inc. extracted from Locus Link	14332	5394
Manual GO annotation by UniProt	30856	6364
Manual GO annotation by MGI	1224	624
Manual GO annotation by BHF-UCL	5393	817
Manual GO annotation by HGNC	2354	556
Manual GO annotation by GDB	111	45
Manual GO annotation by Reactome	5782	1991
Manual GO annotation by IntAct	6307	2292
Manual GO annotation by LIFEdb	280	279
Manual GO annotation by Human Protein Atlas	2437	1222
Total Manual GO annotations	69076	11890
Total GOA annotations	165769	18587
Number of distinct Pubmed references	16886	
Total number of Pubmed references	62691	

OMIM databank - Online Mendelian Inheritance in Man

Online Mendelian Inheritance in Man (OMIM) databank

(*http://www.ncbi.nlm.nih.gov/Omim/*)



Biomolecular Databanks
OMIM databank

- Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders, with links to literature references, sequence records, maps, and related databases
- Each OMIM entry has a full-text summary of a genetically determined phenotype, and has numerous links to other genetic databases such as DNA and protein sequence, PubMed references, general and locus-specific mutation databases, approved gene nomenclature, and the highly detailed MapViewer

OMIM databank

- OMIM includes also:
  - The **OMIM Gene map**, which presents the cytogenetic map location of disease genes and other expressed genes described in OMIM. It lists the chromosomal location, gene symbol(s), method(s) of mapping, and disorder(s) related to each specific gene
  - The **OMIM Morbid Map**, which lists in alphabetical order all disorders mapped in OMIM
  - Links to the human/mouse homology maps
- Information is updated daily

OMIM databank

- Each OMIM entry is assigned a unique six-digit MIM number whose first digit indicates whether its inheritance is autosomal (dominant or recessive), X-linked, Y-linked, or mitochondrial [**Autosomal**: not from a sex chromosome]
- Most MIM numbers are preceded by a symbol, e.g.:

  (*) indicates a separate locus and a proven mode of inheritance (in the judgment of the authors and editors)

  (#) indicates a descriptive entry of a phenotype or gene family
- The absence of a symbol means that the mode of inheritance has not been proven or that the distinction between this locus and another is uncertain

OMIM databank - Query options

- OMIM can be searched from its homepage or from any page in the NCBI Entrez suite of database by:
  - MIM number
  - disorder or gene name and/or symbol
  - plain English (e.g. ‘cryptorchidism webbed neck’)
- The limits function allows performing a restricted search
- The search engine ranks the entries matching the query so that the entry(ies) most relevant to the question are in the top 10 retrievals

OMIM databank - Statistics

- To November 22, 2009, the total OMIM entries were of 19,783, subdivided as it follows:
  - Gene with known sequence 12,972
  - Gene with known sequence and phenotype 349
  - Phenotype description, molecular basis known 2, 658
  - Mendelian phenotype or locus, molecular basis unknown 1,793
  - Other, mainly phenotypes with suspected mendelian basis 2,011
- Many loci (genes) are the site of more than one mutation causing phenotypically distinct disorders

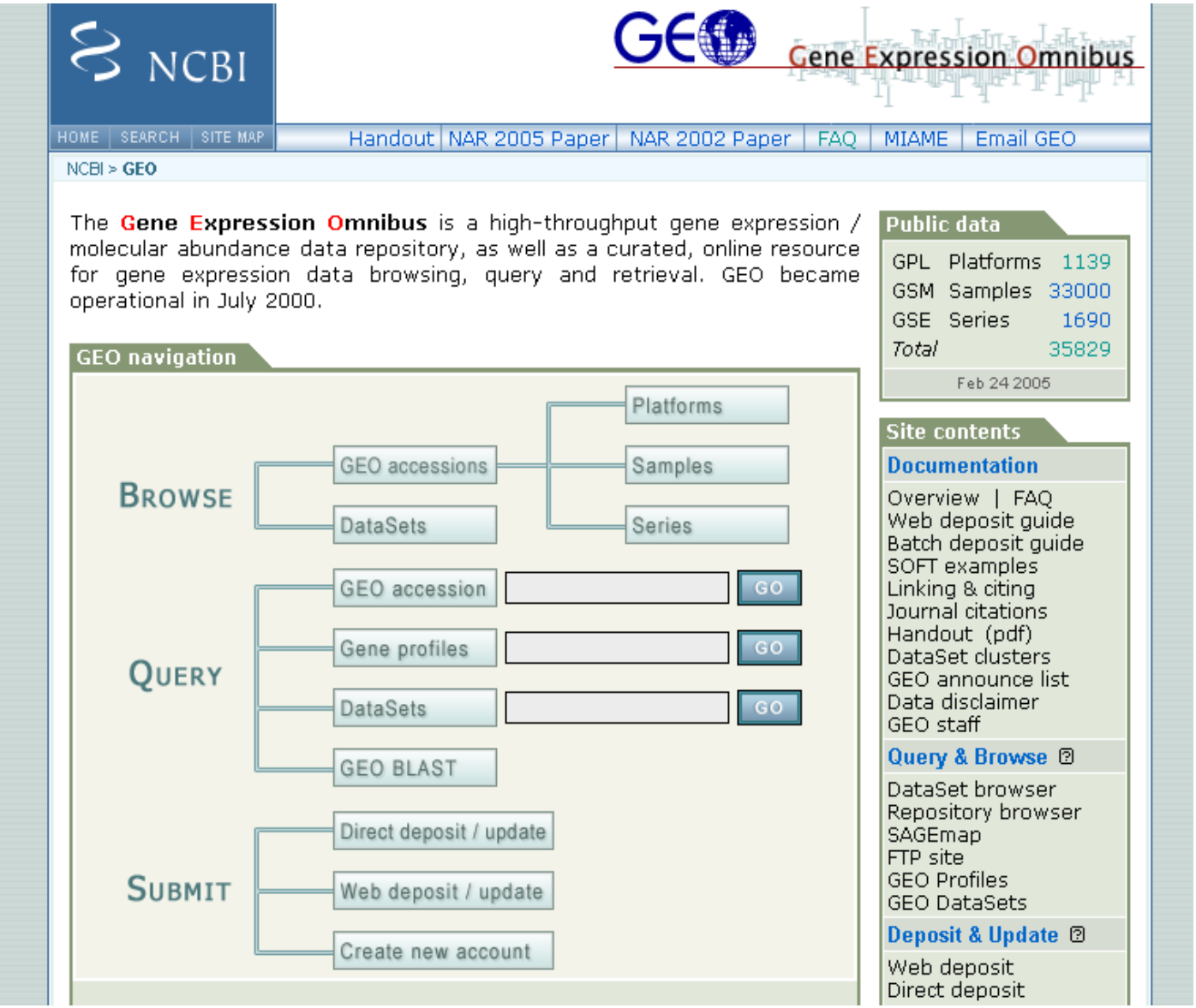
OMIM databank - FTP access

- At ftp://ftp.ncbi.nih.gov/repository/OMIM/ the following files are available for downloading:
  - omim.txt.Z, the complete text of OMIM
  - genemap, the OMIM Gene Map
  - genemap.key, the OMIM Gene Map key explaining symbols and columns in the genemap file
  - morbidmap, the OMIM Morbid Map
- The OMIM Gene Table, alphabetically listing gene symbols and their corresponding MIM numbers, is available at http://www.ncbi.nlm.nih.gov/Omim/Index/genetable.html

GEO databank - Gene Expression Omnibus

Gene Expression Omnibus (GEO) databank

(*http://www.ncbi.nlm.nih.gov/geo/*)



Biomolecular Databanks
GEO databank

- The Gene Expression Omnibus is a high-throughput gene expression / molecular abundance data repository, as well as a curated, online resource for gene expression data browsing, query and retrieval
- GEO serves as a public repository for a wide range of high-throughput experimental data, including single and dual channel microarray-based experiments measuring mRNA, genomic DNA, and protein abundance, as well as non-array techniques such as serial analysis of gene expression (SAGE), and mass spectrometry proteomic data

GEO databank

- To retrieve a particular GEO record for which you have the accession number, use the Accession Display bar, a tool with several options:
  - To query all GEO submissions in a specific field, or over all fields, use either the Entrez GDS or Entrez GEO interfaces:
    - Entrez GDS queries all GEO DataSet annotation, allowing identification of experiments of interest
    - Entrez GEO queries precomputed gene expression / molecular abundance profiles, allowing identification of genes or sequences or profiles of interest
  - To browse lists of GEO data and experiments, use either the GDS browser or view the list of current GEO repository contents

GEO databank

- GEO data can be viewed and downloaded in several formats:
  - HTML
  - SOFT format (Simple Omnibus Format in Text), an ASCII text format that was designed to be a machine readable representation of data retrieved from, or submitted to, GEO
    - The complete SOFT document contains all information for that dataset, including dataset description, type, organism, subset allocation, as well as a data table containing identifiers and values
  - The full text tab-delimited data tables provided may prove suitable for upload into personal microarray analysis software package or database/spreadsheet application

GEO databank - FTP site and included species

- The FTP site is ftp://ftp.ncbi.nih.gov/pub/geo/

- The main species included are the following:
  - Homo sapiens
  - Rattus norvegicus
  - Mus musculus
  - C.elegans
  - D.melanogaster
  - Saccharomyces cerevisiae
  - Escherichia coli
  - Arabidopsis thaliana

POLITECNICO DI MILANO

Stanford Online Universal Resource for Clones and ESTs (SOURCE) databank

(*http://source.stanford.edu/*)

SOURCE Search : (GenePage for Gene H59260) - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indirizzo http://genome-www5.stanford.edu/cgi-bin/SMD/source/sourceResult Vai

SOURCE GeneReport *H. sapiens*

CDC25A

cell division cycle 25A

UniGene, LocusLink, OMIM, GenAtlas, GeneCard, Ensembl, MapView

SwissProt Information

SwissProt Accession No.	P30304 M-PHASE INDUCER PHOSPHATASE 1 (Homo sapiens); 100% similarity over 522 a.a.
Function	this protein functions as a dosage-dependent inducer in mitotic control. it is a tyrosine protein phosphatase required for progression of the cell cycle. it may directly dephosphorylate p34 (cdc2) and activate the p34(cdc2) kinase activity. dephosphorylates p33(cdk2) in complex with cyclin e, in vitro.
Catalytic Activity	protein tyrosine phosphate + h(2)o = protein tyrosine + orthophosphate.
Enzyme Regulation	stimulated by cyclins b.
Similarity	strong, to other species m-phase inducer phosphatase and in general to protein-tyrosine phosphatases.
SwissProt Copyright	This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch).

Annotations

Summary Function	Protein tyrosine and threonine phosphatase (PTPase); acts as a protein tyrosine and threonine phosphatase (PTPase); mediates cell cycle progression by activating mitotic CDKs			
Gene Ontologies	Ontology	Annotation	Evidence	Source
	Biological Process	Cell cycle control	P	Proteome
		Regulation of CDK activity	P	Proteome
Other Ontologies	Ontology	Annotation	Evidence	Source
	organismal role	Control of Cell Proliferation	NR	Proteome

Internet

Biomolecular Databanks
SOURCE databank

- The Stanford Online Universal Resource for Clones and ESTs (SOURCE) is an integrational databank
- SOURCE compiles information collected from some of the most important publicly accessible gene and protein databanks, including:
  - UniGene
  - Entrez Gene
  - Swiss-Prot
  - dbEST
  - GeneMap99
  - RHdb

SOURCE databank - Mission

- The mission of SOURCE is to provide a unique scientific resource that pools publicly available data commonly sought after for any clone, GenBank accession number, or gene
- It has been designed specifically to facilitate the analysis of large sets of data produced by researchers using genome-scale experimental approaches
- SOURCE includes information on the following organisms:
  - Homo sapiens
  - Mus musculus (mouse)
  - Rattus norvegicus (rat)

SOURCE databank - Main features

- To November 2009, significant features were:
  - Direct links to MapView and Genome Browser for human genes
  - Direct retrieval of microarray gene expression (SMD) and Serial Analysis of Gene Expression (SAGE) data
  - Retrieval of upstream genomic sequences of human genes from the Transcript Sequence Retriever (TRASER) databank
  - Normalized gene expression distribution for tissue type
  - Gene Ontology
  - Information about codified protein/s and function/s

SOURCE databank - Main features

- **BatchSource**: a Web extraction interface allowing retrieval of a subset of the data available in SOURCE for multiple genes at once

  This function is very useful to users who are interested in large sets of genes or clones (such as those present on DNA microarrays)

  Batch available information include:

  - UniGene Cluster ID, Name, Symbol, and aliases
  - Representative mRNA and protein accessions
  - Gene Ontology data
  - Summary function

SOURCE databank - Query options

- SOURCE can be interrogated by:
  - Clone ID
  - GenBank accession number
  - UniGene cluster ID
  - Entrez Gene ID
  - gene name or symbol

- For the SOURCE databank neither updating time, statistics and dimension of contained data, or FTP access are available

GeneCards databank

GeneCards databank

(*http://bioinformatics.weizmann.ac.il/cards/*)

GeneCard for CDC25A - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indirizzo http://genome-www.stanford.edu/cgi-bin/genecards/carddisp?CDC25A&search=cdc25a&suff=txt

GeneCards™ — an academic web site of the WEIZMANN INSTITUTE OF SCIENCE

Terms of Use | GeneCards Homepage | Search Examples | Comment Form

Notice - Please read carefully prior to linking to any third-party site.

GeneCard for gene CDC25A

Approved UCL/HGNC/HUGO Human Gene Nomenclature database symbol
CDC25A (cell division cycle 25A)

Aliases and Additional Descriptions (According to GDB, HUGO, and/or SWISS-PROT)	• cell division cycle 25A • M-phase inducer phosphatase 1 (EC 3.1.3.48) (Dual specificity phosphatase Cdc25A).
Chromosomal Location (According to LocusLink and/or UDB and/or HUGO, Genomic Views According to UCSC and Ensembl)	*Chromosome*: **3** *LocusLink cytogenetic band*: **3p21** *Ensembl cytogenetic band*: Chr 3 *Unified DataBase coordinate (from pter)*: 44,845 ± 108 mega bases *Genomic View*: UCSC Golden Path
Proteins (According to SWISS-PROT and/or MIPS)	MPI1_HUMAN • **Size**: 523 amino acids; 58796 Da • **Function**: Functions as a dosage-dependent inducer in mitotic control. It is a tyrosine protein phosphatase required for progression of the cell c dephosphorylates CDC2 and activate its kinase activity. It also dephosphorylates CDK2 in complex with cyclin E, in vitro. • **Catalytic activity**: Protein tyrosine phosphate + H(2)O = protein tyrosine + phosphate. • **Enzyme regulation**: STIMULATED BY CYCLINS B. • **Alternative products**: 2 isoforms; 1/CDC25A1 (shown here) and 2/CDC25A2; are produced by alternative splicing. • **Similarity**: BELONGS TO THE MPI PHOSPHATASE FAMILY. • **Similarity**: CONTAINS 1 RHODANESE DOMAIN. • **3D structure:** PDB id 1C25 (3D) **MIPS Pedant Viewer:** 561

Internet

Biomolecular Databanks
GeneCards databank

- GeneCards is a copyrighted integrational databank of human genes, their products, and their involvement in different pathologies, with a major focus on medical aspects
- This databank, established in 1998, is very rich in information and provides data on the functionality of human genes with an approved symbol (known genes), as well as selected others
- For each gene contained, GeneCards provides links to the related scientific publications stored in the MedLine bibliographic databank
- It was developed at the Crown Human Genome Center and the Bioinformatics Unit at the Weizmann Institute of Science (http://www.weizmann.ac.il/)

Biomolecular Databanks
GeneCards databank

- GeneCards is particularly useful for people who wish to find information about genes of interest in the context of functional genomics and proteomics
- GeneCards is used to study small sets of genes of which is wanted to be retrieved as much as possible of the information available
- One of the fundamental aspects of GeneCards is the use of a standard nomenclature, whose diffusion is promoted

GeneCards databank - Sources

- In GeneCards are present data and automatically generated knowledge based on data automatically extracted from, or linked to, several databanks among which:
  - GenBank
  - UniGene
  - Entrez Gene
  - OMIM, Online Mendelian Inheritance in Man
  - SOURCE, the Stanford Online Universal Resource for Clones and ESTs
  - Swiss-Prot
  - PubMed

Biomolecular Databanks
GeneCards databank - Sources

- HUGO, Human Gene Nomenclature Committee
- SNP Database, Single Nucleotide Polymorphisms databank
- EuGene, Genomic Information for Eukaryotic Organisms
- GDB, Genome DataBase
- MGD, Mouse Genome Database
- FlyBase, a database of the Drosophila genome
- WormBase, the genome and biology of C.elegans
- The Tumor Gene Database
- The Breast Cancer Gene Database
- The Mammary Transgene Database

GeneCards databank - Annotations

- Main information included in GeneCards for each gene is (http://www.genecards.org/GuideGeneCard.shtml#content):
  - the official name and a list of synonyms
  - a list of the gene IDs in other gene-based resources
  - the (cytogenetic) locus of the gene
  - the name of its product/s (i.e. the protein/s), main features of this/these product/s, like cellular functions, expression patterns, similarities with other proteins, involvement in diseases
  - the UniGene cluster of sequences related to the gene
  - a list of disorders and mutations in which the gene is involved according to genetic evidence
  - Titles of related research articles

GeneCards databank - Annotations

- Medical applications, like new therapies and diagnoses, that are based on knowledge about this gene
- homologous genes in the mouse and worm
- a list of disorders and mutations in which the gene is involved according to genetic evidence
- the coordinates as distance from the p terminus of the chromosome (in megabases)
- titles of related research articles with links to the abstract and full citation in PubMed

GeneCards databank - Query options

- Information search in GeneCards can be performed by:
  - accession number and UniGene cluster ID
  - gene symbol (e.g. BRCA1)
  - keywords (e.g. apolipoprot*, Alzheim*)
  - SNP id (e.g. SNP and 762667)
  - clone identifier (e.g. p53, ATCC:106253, image:303124)
  - chromosome (e.g. chromosome:22)
  - locus (e.g. locus:20p*)

GeneCards databank - Statistics

- GeneCards statistics are available at http://www.genecards.org/index.shtml
- Version: 2.41.1 Release: November 1, 2009
  - Entries: 55,546
  - Entries with HUGO-approved symbols: 28,139
  - Protein-coding genes: 21,909
  - Pseudogenes: 10,363
  - RNA genes: 10,385
  - Genetic loci: 1,416
  - Gene clusters: 46
  - Uncategorized: 11,427

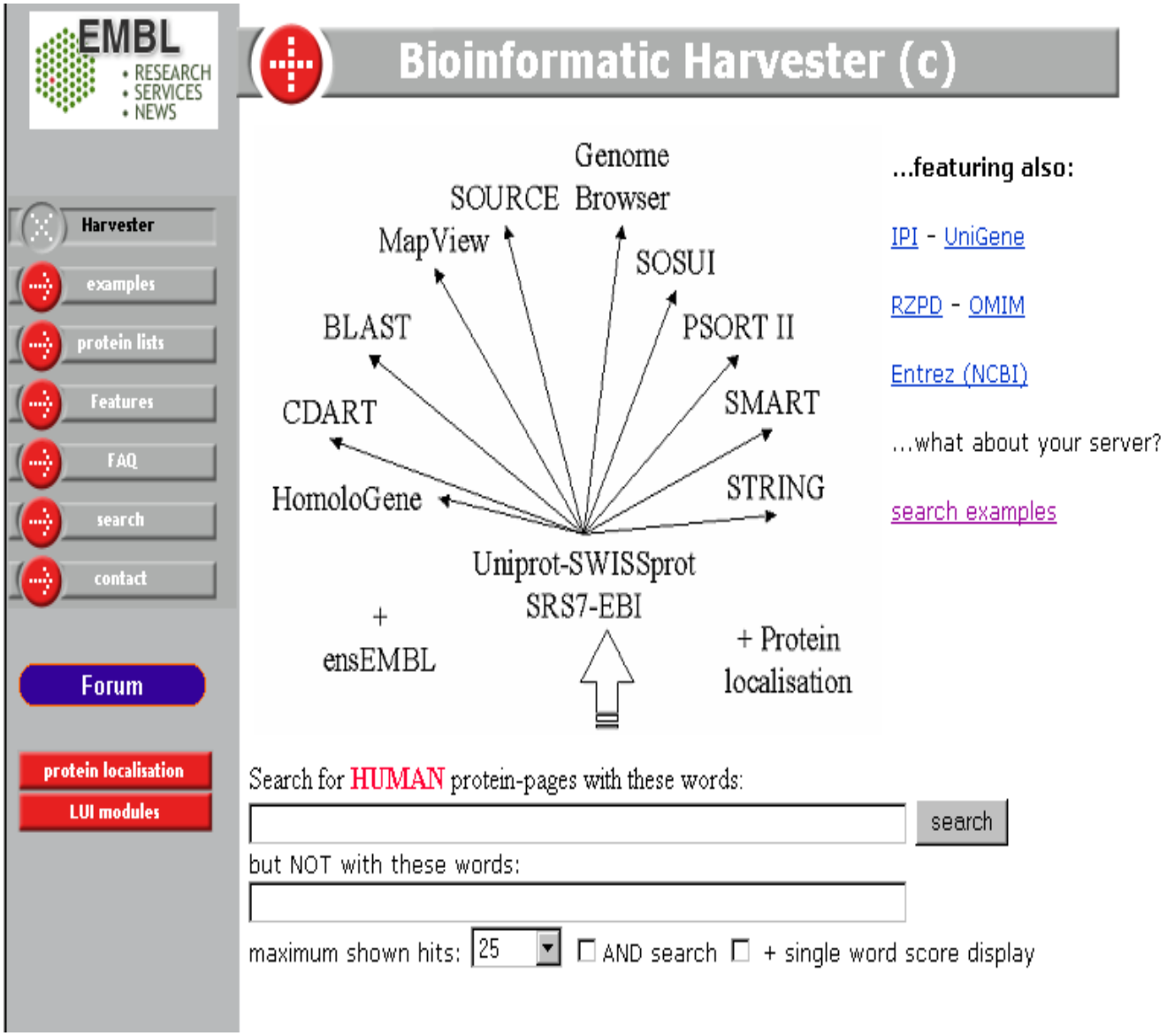
GeneCards databank - Mirror sites

- Mirror sites (http://www.genecards.org/mirror.shtml) of GeneCards databank can be established after signing a license agreement for the entire package
- The GeneCards package consists in the GeneCards database and Perl scripts to provide and support the functions for web user interface, database search, query reformulation support and navigation guidance system, including a spell correction system
- Many public mirror sites already exist world wide
- For the GeneCards databank neither updating time, dimension of contained data, or FTP access are available

Bioinformatic Harvester databank

Bioinformatic Harvester databank

(*http://harvester.embl.de/index.html*)



Biomolecular Databanks
Harvester databank

- Harvester collects information from selected public databanks
- The flexible crawler modules save databank entries either as text block (for search engine indexing) or provide “iframe” crosslinks (for databases rich in graphical information e.g. ensEMBL, BLAST, CDART, Genome Browser)
- “iframes” provide the user the latest information from the original database server

Biomolecular Databanks
Harvester databank

- "Text blocks" and "iframes" along with the protein specific "iframe" links are presented on a single HTML page for convenient study
  - Each "iframe" can be manipulated individually
- Various analysis methods, as PSORT II, SOSUI, SMART, Homologene, have been applied to the collected sequences
  - New server or analysis methods can be implemented as needed

Harvester databank

- Harvester allows a combined full text and protein sequence search
  - The full text search can be used for:
    - literature
    - protein function (SOURCE)
    - protein domain analysis (SMART)
    - predicted or evaluated protein localization (PSORT II, Uniprot)
    - annotations
    - database cross-links (BLAST-NCBI, CDART, ensEMBL, Genome Browser, GO, HSSP, InterPro, MapView, PFAM, Prosite, SMART, SOSUI, STRING, Uni-Gene)

Biomolecular Databanks
Harvester databank

- Harvester allows comparison of different prediction algorithms on a single HTML page
- Harvester search results, including all links and result scores, can be saved via the Internet browser used. Saving the results in ".XLS" format will allow subsequent dealing of the saved results within Excel

Harvester databank - Updates

- Information provided in "iframes" (active boxes within the page) are loaded from the particular server: Uniprot and Source database information is updated every 21 days, a frequency similar to that of the appearance of updates by the public databases it relates to
- Harvester also updates when the algorithms underlying the programs of the prediction servers have changed
- FTP site is not available