



Biological Databases

September, 2002

CRC-CGF Bioinformatics Course, 2002



Outline

- Database overview
- Finding biological databases
- Highlight some useful databases

CRC-CGF Bioinformatics Course, 2002



What is a database?

- A representation of some aspect of the real world
- A collection of facts about biological systems
- Coherent, consistent and designed for a specific purpose
- A system for managing and querying the database


CRC-CGF Bioinformatics Course, 2002



What is a good database?

- Broad coverage of the chosen topic
- Up-to-date information gathering
- Support staff
- Commitment to the future
- Good query interface


CRC-CGF Bioinformatics Course, 2002



What databases are there?

5S Ribosomal RNA Sequence Repositories
 Major Sequence Repositories
 Site Asthma Gene Database Asthma and Allergy Database Atlas of Genetics and Genomics in Oncology and Hematology Axeldb BLOCKS BTKbase
 BioImage CATH COMPEL CSD CUTG Clusters of Orthologous Groups (COG) Collection of mRNA-like non-coding RNAs Cre
 Comparative Genomics
 Transgenic Database ExpressionBase Cytokine Gene Polymorphism Database Data Decat DEXH/D Family Database DIP DNA Data Bank of Japan (DDBJ)
 Gene Expression
 DPInteract DRP5H Database of Germ-line p53 Mutations Database of Macromolecular Complexes Database of Ribosomal Crosslinks (DRC) Database on
 the Structure of Small Subunit Ribosomal RNA Decoys 'R' Us DrugDB EID EMBL Nucleotide
 Gene Identification and Structure
 Sequence Database of Small Subunit Ribosomal RNA Decoys 'R' Us EcoCyc EcoGene Endogenous GPCR List EpubDB ExInt FIMM FUNPEP FlyBase FlyNets FlyView G3-RH G84-RH
 GDB GOBASE GPCEDB GPRF Mutant Databases GenAtlas GenBank GenProtEC Gene Expression Database (GXD) GeneMap '99 Genome Sequence Database (GSDB)
 GDB GOBASE GPCEDB GPRF Mutant Databases GenAtlas GenBank GenProtEC Gene Expression Database (GXD) GeneMap '99 Genome Sequence Database (GSDB)
 Guide RNA Library
 Genomic Databases
 Database Hierarchy
 Homeobox Page Homeodomain Resource HuGeneMap Human BAC Ends Database Human Gene Mutation Database HvrBase IDB/IEDB
 Human PAX2 Allelic Variant Database Human PAX6 Allelic Variant Database Human Pdx1 and Type III Collagen Mutation Database HvrBase IDB/IEDB
 IMB Jena Image
 Metabolic Pathways and Cellular Regulation
 Keynet KMDb Kabacov Development Database KinMutBase Klotho Kyoto
 Encyclopedia of Genes and Genomes (KEGG) LGIC LIGAND LPFC LocusLink/RefSeq MAGEST MEROPS MHCPEP MITOMAP MITONUC/MITOALN MITOP MMDb MODBASE
 MPDB Membrane Protein Database Mendel Database MitBASE MitDat MmtDB Molecular Probe Database Mouse Atlas and Gene Expression Database Mouse
 Genome Database Mouse Tumor Biology Database (MTB) Munich Information Center for Sequences (MIPS) Mutation Spectra Database NCBI
 Taxonomy NDB NER NRSUB Non-canonical Base Pair Database O-Linked N-Glycanase O-Glycanase O-Glycanase Receptor Database Online Mendelian
 Protein Databases
 Inheritance in Man (OMIM) PDB PDB-REPRDB PDB PEDB PIR-ALN PKR PLACE PLM1/ENA PMD PPMdb PRESAGE PRINTS PROMISE PROSITE Peptaibol Pfam
 PhosphoBase PhosphoSitePlus
 Protein Sequence Motifs
 Pseudobases
 Protein Resources
 Mutation Database REBASE RESID RNA Modification Database Radiation Hybrid Database Receptor Database (RDP)
 RegulonDB Ribonuclease P Database Ribosomal Database Project (RDP) Ribosomal RNA Mutational Database RsgDB SBASE SCOPE SELEX_DB SENTRA SMART
 SRPDB SV40 Large T-Antigen Mutant Database SWISS-2DPAGE SWISS-PROT/TrEMBL SYSTEMS
 RNA Sequences
 Gene Indices
 Retrieval Systems and Database Structure
 Transgenic/Taxonomic Mutation Database Tree of Life UM-BBD UTRdb UniGene Vectordb Virgil Viroid and Viroid-Like RNA Database WIT2 Wnt Database
 XREFdb YDB Yeast Proteome Database (YPD) Yeast snoRNA Database ZFIN Zmdb dbMIF dbMIF-2 dbMIF-3 dbMIF-4 dbMIF-5 dbMIF-6 dbMIF-7 dbMIF-8 dbMIF-9
 Transgenics
 tRNA Website
 Miscellaneous
 Varied Biomedical Content

CRC-CGF Bioinformatics Course, 2002



Locating that database!

- Ask a colleague
- Bibliographic citation
- The Web as a database
- Nucleic Acids Research, Database Issue, January 1
- Web-page listings
- Searchable databases of databases

CRC-CGF Bioinformatics Course, 2002



The Web as a database

- WWW is a very large database of text, images, programs, maps, ...
- Google (<http://www.google.com>)
- AltaVista (<http://www.altavista.com>)
- WWW Virtual Library (<http://www.vlib.org>)
- SCIRUS (<http://www.scirus.com>)
- BioHunt (<http://au.expasy.org/BioHunt/>)
- BioWurld (<http://search.ebi.ac.uk:8888/compass>)

CRC-CGF Bioinformatics Course, 2002



WWW Virtual Library

The WWW Virtual Library

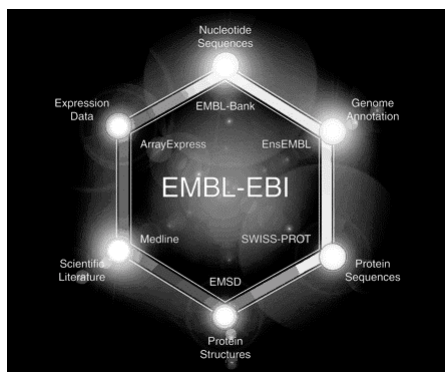
- | | |
|--|--|
| <ul style="list-style-type: none">• Agriculture
<i>Agriculture, Gardening, Forestry, Irrigation...</i>• Business and Economics
<i>Economics, Finance, Marketing, Transportation...</i>• Computing
<i>Computing, E-Commerce, Languages, Web...</i>• Communications and Media
<i>Communications, Telecommunications, Journalism...</i>• Education
<i>Education, Applied Linguistics, Linguistics...</i>• Engineering
<i>Civil, Chemical, Electrical...</i>• Humanities
<i>Anthropology, History, Museums, Philosophy...</i> | <ul style="list-style-type: none">• Information & Libraries
<i>General Reference, Information Quality, Libraries...</i>• International Affairs
<i>International Society, Sustainable Development, UN...</i>• Law
<i>Arbitration, Law, Legal History...</i>• Recreation
<i>Recreation and Games, Gardening, Sport...</i>• Regional Studies
<i>African, Asian, Latin American, West European...</i>• Science
<i>Biosciences, Health, Earth Science, Physics, Chemistry...</i>• Society
<i>Political Science, Religion, Social Sciences...</i> |
|--|--|

Search the WWW VL:

CRC-CGF Bioinformatics Course, 2002



Nucleic Acids Research



Annual Database Issue

Issue no 1

January 1

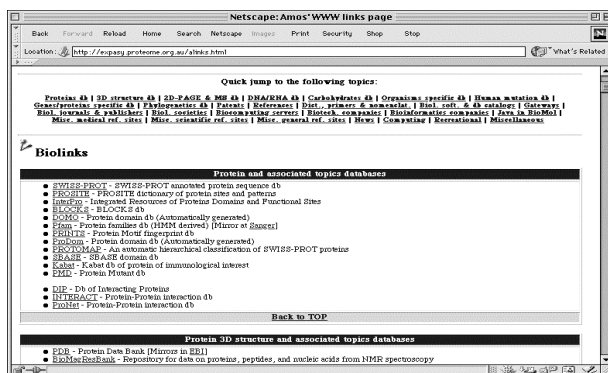
every year

<http://nar.oupjournals.org>

CRC-CGF Bioinformatics Course, 2002



ExPASy Listing



<http://au.expasy.org/> Links to lists of resources

CRC-CGF Bioinformatics Course, 2002



Searchable Databases

- NAR Database Issue [336]
 - <http://nar.oupjournals.org/>
- DBCAT [511]
 - <http://www.infobiogen.fr/services/dbcat>
- SRS (LionBiosciences) [655, 31 centres]
 - <http://downloads.lionbio.co.uk/publicsrs.html>



CRC-CGF Bioinformatics Course, 2002



DBCAT

CRC-CGF Bioinformatics Course, 2002



SRS

Public SRS Servers (07 Aug 2002)

If you maintain a publicly accessible srs6 server, and would like to have it added to this list, please send a message to support@ionbio.co.uk

ABC: Hungarian EMENET node	8 libraries	sbase.abc.hu/srs6bin/cgi-bin/wgetz	6.1.3.5
BCGSC: British Columbia Genome Sequence Center, Vancouver, Canada	57 libraries	srs6.bcgsc.bc.ca/bin/wgetz	6.1.3.10
BEN: Belgian EMENET node, Belgium	49 libraries	www.be.embnnet.org/srs6bin/wgetz	6.1.3.10
BioBase: EMENET Danmark, Aarhus, Denmark	27 libraries	www.dk.embnnet.org/srs6bin/cgi-bin/wgetz	6.0.7.3
Biocenter: Vienna Biocenter, EMENET Austria	38 libraries	emb1.bcc.univie.ac.at/5000/cgi-bin/wgetz	6.0.5.1
CABRI: Common Access to Biological Resources and Information, International	49 libraries	srs.cabri.org/srs6bin/cgi-bin/wgetz	6.1.3
CBR: Canadian Bioinformatics Resource, Halifax, Canada	94 libraries	www.cbr.nrc.ca/srs6bin/cgi-bin/wgetz	6.1.3.5
CDFD: Indian EMENET node, Hyderabad, India	29 libraries	www.cdfi.org.in/srs6bin/cgi-bin/wgetz	6.1.0.1
CIGB: EMENET Cuba Havana	110 libraries	bio.cigb.edu.cu/srs6bin/cgi-bin/wgetz	6.1.3.10
CMBI: Centre for Molecular and Biomolecular Informatics, Nijmegen, NL	67 libraries	www.cmbi.kun.nl/srs6bin/cgi-bin/wgetz	6.0.7.3

EBIND	EBI (5843)
EOCAT	BEN (607) Biocenter (579) CMBI (607) DKFZ (607) HGMP (579) Infobiogen (607)
EOCOTAL	DIC (607) EBI (607)
EOMDB	EBI (78)
ELOCKS	BCGSC (4071) Biocenter (11117) CDFD (8909) CIGB (4071) CMBI (11853) CMB (11117) CSC-Finland (11853) DDBJ (11853) DIC (11117) DKFZ (4071) EBI (4071) HGMP (4071) IBC (4071) IM (11853) IUBIO (11853) Infobiogen (11853) Oxford (4071) PKU (11853) Pasteur (4071) Sanger (4071)

CRC-CGF Bioinformatics Course, 2002



Useful Databases

- NCBI
- EBI
- Genomes
- Pathways
- GeneOntology
- On-line news services

CRC-CGF Bioinformatics Course, 2002



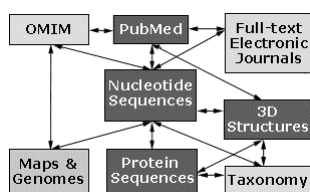
NCBI Databases

- National Center for Biotechnology Information
- <http://www.ncbi.nlm.nih.gov>
- Genbank (Divisions), dbEST, dbSTS, dbGSS, dbHTG, dbSNP, Taxonomy, LocusLink, RefSeq, Unigene, GeneMap, Davis Human-Mouse Homology Maps, CCAP, Entrez Genomes, COGS, CGAP, SAGE, OMIM, PUBMED, Entrez

CRC-CGF Bioinformatics Course, 2002



NCBI Entrez



Entrez is a search and retrieval system that integrates information from databases at NCBI.

<http://www.ncbi.nlm.nih.gov/Entrez/>

Entrez databases

PubMed: The biomedical literature (PubMed)
Protein sequence database
Genome: complete genome assemblies
OMIM: Online Mendelian Inheritance in Man
Books: online books
3D Domains: domains from Entrez Structure
SNP: single nucleotide polymorphisms

Nucleotide sequence database (Genbank)
Structure: three-dimensional macromolecular structures
PopSet: population study data sets
Taxonomy: organisms in GenBank
ProbeSet: gene expression and microarray datasets
UniSTS: markers and mapping data
CDD: conserved domains

CRC-CGF Bioinformatics Course, 2002



NCBI RefSeq

NCBI Reference Sequences

<http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html>

Definition: The NCBI Reference Sequence project (RefSeq) provides reference sequence standards for the naturally occurring molecules of the central dogma, from chromosomes to mRNAs to proteins. Toward this goal, intermediate larger genomic regions, instantiated as accessions of the format NG_123456 (genomic sequence with curated annotation) or NT_123456 (computed assembly and annotation) are also produced. RefSeq standards provide a foundation for the functional annotation of the human genome. They provide a stable reference for gene characterization, mutation analysis, expression studies, and polymorphism discovery.

NC_123456	Complete Genomes
NG_123456	Genomic Region <i>Homo sapiens</i>
NM_123456	mRNA
NP_123456	Protein
NT_123456	Genomic Contig
XM_123456	mRNA <i>Homo sapiens</i> model mRNA from the Genome Annotation process
XR_123456	RNA <i>Homo sapiens</i> model non-coding transcripts from the Genome Annotation process
XP_123456	Protein <i>Homo sapiens</i> model proteins from the Genome Annotation process

CRC-CGF Bioinformatics Course, 2002



EBI

- European Bioinformatics Institute
- <http://www.ebi.ac.uk>
- EMBL Nucleotide, Ensembl, Genomes, Align, dbSTS, Parasites, Mutations, IMGT, SwissProt, TrEMBL, InterPro, CluSTr, IPI, GOA, HPI, Intenz, NEWT, MSD, 3Dseq, FSSP, DALI, 3Dee, Array Express, Medline, Software Biocatalog, Flybase

CRC-CGF Bioinformatics Course, 2002



EBI - Swissprot

<http://www.ebi.ac.uk/swissprot/>

SWISS-PROT

The SWISS-PROT Protein Knowledgebase is an annotated protein sequence database established in 1986.



The SWISS-PROT Protein Knowledgebase 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.

It is maintained collaboratively by the [Swiss Institute for Bioinformatics \(SIB\)](#) and the [European Bioinformatics Institute \(EBI\)](#).

The SWISS-PROT group is headed by: **Rolf Apweiler**.

The current SWISS-PROT Release is version 40.26 as of 13-Aug-2002, and contains 112894 entries... [more stats](#)

CRC-CGF Bioinformatics Course, 2002



EBI - Swissprot

```
CC -!- FUNCTION: CYTOKINE. INHIBITS INFLAMMATORY CYTOKINE PRODUCTION.
CC SYNERGIZES WITH IL2 IN REGULATING INTERFERON-GAMMA SYNTHESIS.
CC MAY BE CRITICAL IN REGULATING INFLAMMATORY AND IMMUNE RESPONSES.
CC -!- SUBCELLULAR LOCATION: SECRETED.
CC -!- POLYMORPHISM: GLN AT POSITION 130 IS A SIGNIFICANT RISK FACTOR FOR
CC ASTHMA DEVELOPMENT.
CC -!- SIMILARITY: BELONGS TO THE IL-4 / IL-13 FAMILY.

DR EMBL; L06801; AAA36107.1; -.
DR EMBL; X69079; CAA48823.1; ALT_INIT.
DR EMBL; U10307; AAA83738.1; -.
DR EMBL; AF377331; AAK53823.1; -.
DR PDB; 3ITR; 15-JAN-95.
DR Genew; HGNC:5973; IL13.
DR InterPro; IPR003634; Interleukin_13.
DR Pfam; PF03487; Interleukin_13; 1.
DR SMART; SM00190; IL4_13; 1.

DR EMBL; X69079; CAA48824.1; -.
DR EMBL; U31120; AAB01681.1; -.
DR EMBL; AF043334; AAC03535.1; -.
DR PIR; A47481; A47481.
DR PDB; 3ITS; 26-JAN-95.
DR MIM; 147683; -.
DR InterPro; IPR001325; Interleukin_4_13.
DR ProDom; PD015987; Interleukin_13; 1.
DR PROSITE; PS00838; INTERLEUKIN_4_13; 1.

KW Cytokine; Glycoprotein; Signal; 3D-structure; Polymorphism.

FT SIGNAL 1 20
FT DISULFID 48 76 FT DISULFID 64 90
FT CARBOHYD 49 49 N-LINKED (GLCNAC) (POT.).
FT CARBOHYD 72 72 N-LINKED (GLCNAC) (POT.).
FT CONFLICT 45 45 A -> R (IN REF. 4).
FT CONFLICT 98 98 MISSING (IN REF. 4).

FT CHAIN 21 132 INTERLEUKIN-13.
FT CARBOHYD 38 38 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 57 57 N-LINKED (GLCNAC...) (POTENTIAL).
FT VARIANT 130 130 R -> Q.
FT CONFLICT 87 87 S -> G (IN REF. 5).

SQ SEQUENCE 132 AA; 14319 MW; 123F1DCAB87FD78B CRC64; IL13_HUMAN
```

CRC-CGF Bioinformatics Course, 2002



EBI - InterPro

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

InterPro is a useful resource for whole genome analysis and has already been used for the proteome analysis of a number of completely sequenced organisms including *preliminary* analyses of the mouse and human genomes.

Secondary protein databases on functional sites and domains like PROSITE, PRINTS, SMART, Pfam, ProDom, etc. are vital resources for identifying distant relationships in novel sequences, and hence for predicting protein function and structure. Unfortunately, these signature databases do not share the same formats and nomenclature, and each database has its own strengths and weaknesses.

To capitalise on these, the following partners: EBI, SIB, University of Manchester, Sanger Institute, GENE-IT, CNRS/INRA, LION bioscience AG and University of Bergen unified PROSITE, PRINTS, ProDom and Pfam into InterPro (Integrated resource of Protein Families, Domains and Sites). The latest databases to join the project were SMART, and more recently, TIGRFAMs.

CRC-CGF Bioinformatics Course, 2002



EBI - InterPro

General information about the entry	
Entry name	C_TYPE_LLECTIN_1
Accession number	PS00615
Entry type	PATTERN
Date	APR-1990 (CREATED); JUL-1998 (DATA UPDATE); JUL-1998 (INFO UPDATE)
PROSITE documentation	PDOC00537
Name and characterization of the entry	
Description	C-type lectin domain signature.
Pattern	C-[LIVMFYATG]-x(5,12)-[WL]-x-[DNSR]-x(2)-C-x(5,6)-[FYWLIVSTA]-[LIVMSTA]- C
Numerical results	
<ul style="list-style-type: none"> • SWISS-PROT release number: 40.7, total number of sequence entries in that release: 103373. • Total number of hits in SWISS-PROT: 137 hits in 128 different sequences • Number of hits on proteins that are known to belong to the set under consideration: 122 hits in 113 different sequences • Number of hits on proteins that could potentially belong to the set under consideration: 0 hits in 0 different sequences • Number of false hits (on unrelated proteins): 15 hits in 15 different sequences • Number of known missed hits: 50 • Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because of truncation: 0 • Precision (true hits / (true hits + false positives)): 89.05 % • Recall (true hits / (true hits + false negatives)): 70.93 % 	

CRC-CGF Bioinformatics Course, 2002



Sequence Database Entry

LOCUS HUMGMCSF 1480 bp mRNA PRI 27-APR-1993
 DEFINITION Human GM-CSF receptor mRNA, complete cds.
 ACCESSION M64445
 VERSION M64445.1 GI:183361
 KEYWORDS GM-CSF receptor; hematopoietin receptor; transmembrane protein.
 SOURCE Homo sapiens (library: cDNA) cDNA to mRNA.
 ORGANISM Homo sapiens
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
 REFERENCE 1 (bases 1 to 1480)
 AUTHORS Crosier,K.E., Wong,G.G., Mathyey-Prevot,B., Nathan,D.G. and
 Sieff,C.A.
 TITLE A functional isoform of the human granulocyte-macrophage
 colony-stimulating factor receptor contains a unique cytoplasmic
 domain
 JOURNAL Proc. Natl. Acad. Sci. U.S.A. 88, 7744-7748 (1991)
 MEDLINE 91352066

CRC-CGF Bioinformatics Course, 2002



Sequence Database Entry

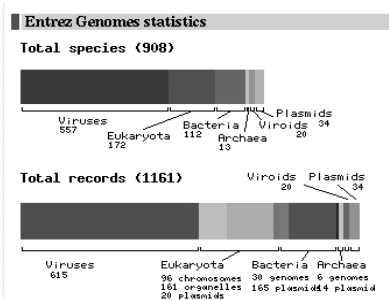
FEATURES	Location/Qualifiers
source	1..1480 /organism="Homo sapiens" /db_xref="taxon:9606" /cell_line="M-07" /tissue_lib="cDNA"
gene	163..1395 /gene="GM-CSF receptor"
CDS	163..1395 /gene="GM-CSF receptor" /codon_start=1 /product="GM-CSF receptor" /protein_id="AAA35908.1" /db_xref="GI:183362" /translation="MLLLVTSLLLCELPHPAPFLLIPEKSDLRTVAPASSLNVRFDSRT MNLSDQCENITFSKCFLTDKKNRVVEPRLSNNECSCTFREICLHEGVTPEVHVNTSQ RGFQKLLYPNSGREGTAAQNFSCFIYNADLMNCTWARGPTAPRDVQVFLYIRNSKRR REIRCPYYIQDSGTHVGCCHLDNLSGLTSRNYFLVNGTSREIGIQFFDSLDDTKKIERF NPPSNVTRCNTTHCLVRWKQPRTYQKLSYLDYQYQLDVHRKNTQPGTENLLINVSGD LENRYNFPSSSEPRAKHSVKIRAADVRIINWSSWSEAEIEFGSDDNLGVSYYVYVLLIVG TLVCGIVLGFLEKRFRLIQRLFPVPPVQIKDKLNDNHEVEDEMGPPQRHRCGWNLYPTP GPSPGSGSSPRLGSESSL"

CRC-CGF Bioinformatics Course, 2002



Genomes

- Microbial, Archaea, Bacteria, Eukaryota, Viruses, Phages, Viroids, Plasmids
- A feature of genome sequencing is its multi-centred, international nature
- Great value can be gained from genome comparison



CRC-CGF Bioinformatics Course, 2002



NCBI Human Genome Tools

Blast	Compare your sequence to the genome or its gene products.
Cytogenetics	Cytogenetic resource of FISH-mapped, sequence-tagged clones.
dbSNP	Database of SNPs and other genetic variations.
e-PCR	Check your sequence for STSs and view in genomic context.
GEO	Gene Expression Omnibus, public repository for expression data.
HomoloGene	Putative homologies among human, mouse, rat, and zebrafish.
Homology Maps	Blocks of conserved synteny between mouse and human.
LocusLink	Focal point for genes and associated information.
OMIM	Guide to genes and inherited disorders, JHU and collaborators.
RefSeq	Reference sequences of genomic contigs, mRNAs, and proteins.
SAGEmap	Gene expression results from SAGE tags mapped to sequences.
Sequencing	Summary of human genome sequencing progress.
MapViewer	Interactive viewer for genome maps, sequence, and genes.
Unigene	Organization of transcribed sequences into gene-based clusters.
UniSTS	A non-redundant collection of STSs with links to maps and sequence.

CRC-CGF Bioinformatics Course, 2002



NCBI - COGS

- attempt to reconstruct the phylogenetic relationships between organisms by comparing the protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages,
- a COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.
- COGs applied to *D.melanogaster* and *C.elegans*
- search for COGs by sequence, functional group, phylogenetic relationship or pathway

CRC-CGF Bioinformatics Course, 2002



TIGR Microbial Genomes

Multi-Genome Applications:

- [Build a Multi-Genome Query](#): Queries across all genomes in the database based on various types of search criteria.
- [Align Whole Genomes](#): Aligns any two genomes in the database based on exact DNA sequence matches.
- [Genome Browser](#): Linear Display of Genes of any genome in the database.
- [Restriction Digest](#): Displays restriction digest information for a genome.
- [Third Position GC Skew Display](#): Predicts genes by comparing possible open reading frames to a third position GC plot.
- [Batch Download](#): Retrieves to your local disk a large number of sequences.
- [Gene Attribute Download](#): Retrieves to your local disk selected gene attributes for either a user-generated list of genes or all genes from a specific organism and/or role category.
- [Download TIGR Genomes](#): Retrieve a TIGR genome using ftp.

Multi-Genome Analyses:

- [Genome v/s Genome Protein Hits](#): Shows the number of proteins in each organism that have hits to proteins in every other organism in the CMR.
- [COG/TIGRFAM/PFAM Comparison](#): Shows the percentage of genes in the CMR assigned a COG, a TIGRFAM, a PFAM or any combination of the three.
- [High/Average Protein Attribute Graph](#): Shows the high and average numbers from each genome for a variety of protein attributes.
- [Role Category Survey](#): Shows the number and percentage of genes in each genome for every role category in the CMR.
- [Role Category Graph](#): Shows the number and percentage of genes in the CMR assigned to each of the role categories. Unlike the Role Category Survey, this page does not break the genes down by organism.
- [Terminator Survey](#): Shows the number and percentage of genes in each genome with terminators.
- [Gene Ontology \(GO\) Terms Association](#): Shows the number of GO terms assigned to each genome.
- [DNA Molecule Information](#): Provides detailed information on each of the DNA molecules found in the organism (chromosomes, plasmids) including the topology (linear or circular), length, %A, T, G, C and number of genes.
- [Operon Predictions](#): Use this page to view TIGR's predictions of operons in microbial genomes.

<http://www.tigr.org>

Comprehensive
Microbial Resource

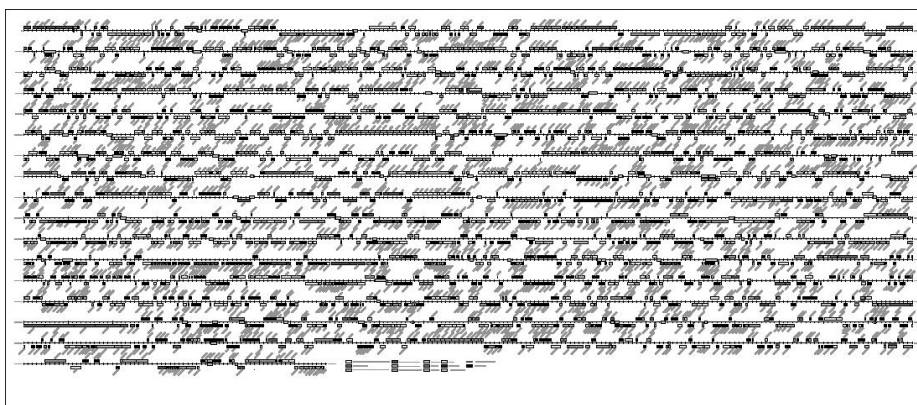
CRC-CGF Bioinformatics Course, 2002



Sanger - *M.leprae*

Complete gene map

http://www.sanger.ac.uk/Projects/M_leprae/gene_map.pdf



CRC-CGF Bioinformatics Course, 2002



Ensembl

e! project **Ensembl**

The Wellcome Trust
Sanger Institute EBI

Ensembl Genome Browser

About Ensembl

Ensembl is a joint project between EMBL, EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust. Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints.

Ensembl presents up-to-date sequence data and the best possible automatic annotation for eukaryotic genomes. Available now are **human**, **mouse**, **zebrafish**, and **mosquito**. Others will be added soon.

For an introduction to the Ensembl project, take the [Ensembl tour](#), and then go through a step-by-step [worked example](#) which introduces Ensembl's main functions. For more information read this short [paper](#) in Nucleic Acids Research.

For all enquiries, please contact the Ensembl [Help Desk](mailto:helpdesk@ensembl.org) (helpdesk@ensembl.org).

Ensembl provides

- ▶ Easy access to sequence data
- ▶ For known genes, predicted structure and location in the genome sequence
- ▶ Prediction of novel genes, all with supporting evidence
- ▶ Annotation of other features of the genome
- ▶ Targetted connections to other genome resources worldwide

Ensembl Species

Human	v. 7.29a.3	12 July 2002
Mouse	v. 7.3b.3	12 July 2002
Zebrafish	v. 7.0b.3	8 Aug 2002
Fugu	v. 7.1.3	11 Feb 2002
Mosquito	v. 7.1b.3	30 May 2002

Access to whole genome shotgun data (includes additional species) [Trace Server](#)

Help and documentation

- ▶ Species-specific documentation is available via the species home pages above.
- ▶ Take the [Ensembl tour](#), go through a step-by-step [worked example](#), or read this short [paper](#) in Nucleic Acids Research.
- ▶ For context-sensitive help on any web page click: [Help](#)
- ▶ There is also an [index](#) of context-sensitive help pages, and a set of guided [how do I...?](#) trails.

Recent Ensembl news

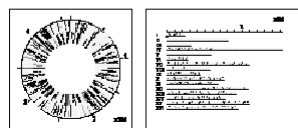
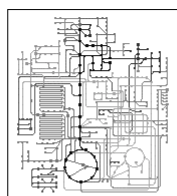
- ▶ Multi-species data retrieval
- ▶ Display your own data in Ensembl
- ▶ Ensembl developer resources
- ▶ Apollo genome browser
- ▶ Questions or suggestions? Try the [Help Desk](#)
- ▶ Documentation (includes tutorial on direct [Documentation](#))

<http://www.ensembl.org>

A system to produce and maintain automatic annotations of selected genomes

CRC-CGF Bioinformatics Course, 2002

Pathways - KEGG



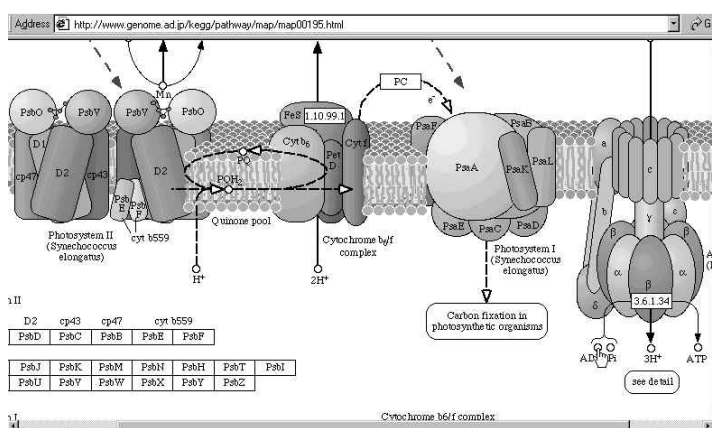
Kyoto Encyclopedia of Genes and Genomes

<http://www.genome.ad.jp/kegg/>

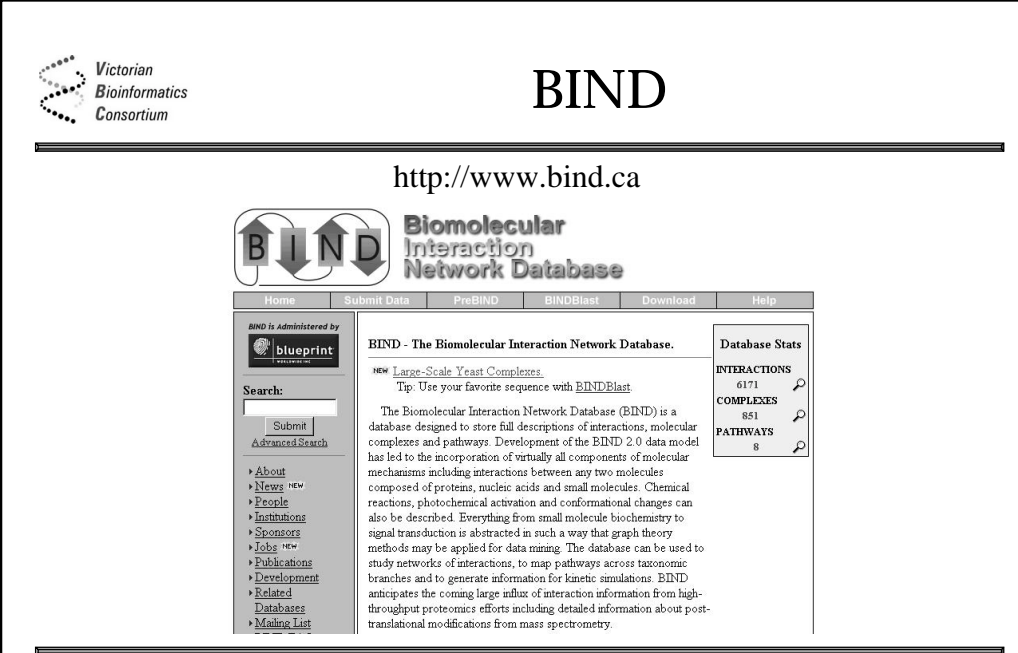
- pathways, sequences, genomic maps
- search for objects in pathways
- generate possible reaction pathways
- compare two genome maps
- infer presence of pathway in organism 2 by comparison with organism 1

CRC-CGF Bioinformatics Course, 2002

KEGG



CRC-CGF Bioinformatics Course, 2002



Victorian Bioinformatics Consortium

BIND

<http://www.bind.ca>

BIND Biomolecular Interaction Network Database

Home | Submit Data | PreBIND | BINDblast | Download | Help

Search: [Advanced Search](#)

• [About](#)
• [News](#) NEW
• [People](#)
• [Institutions](#)
• [Sponsors](#)
• [Jobs](#) NEW
• [Publications](#)
• [Development](#)
• [Related Databases](#)
• [Mailing List](#)

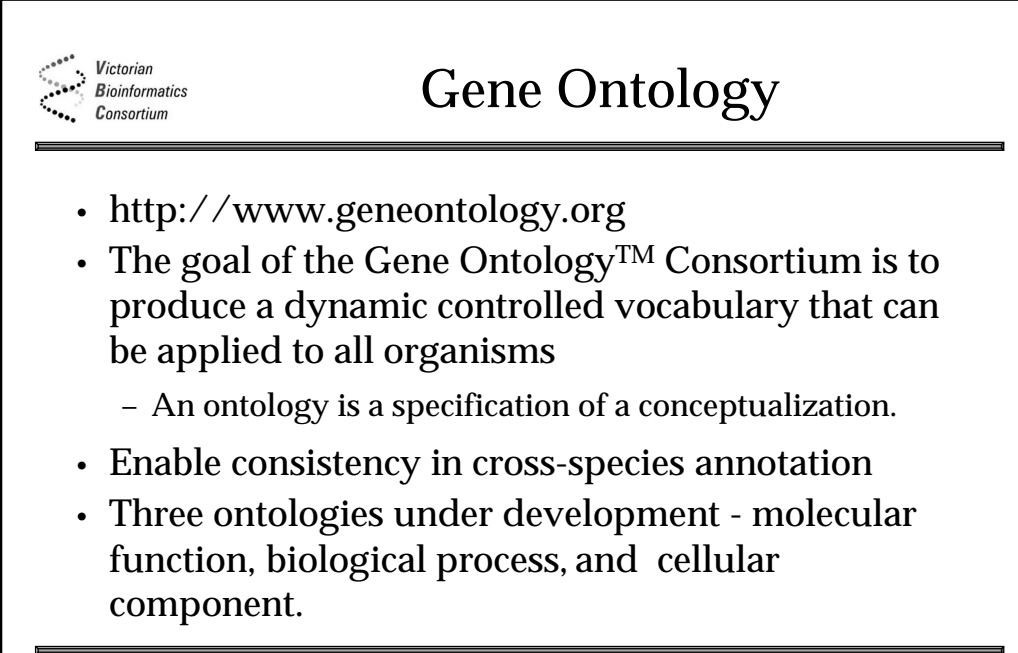
BIND - The Biomolecular Interaction Network Database.

NEW [Large-Scale Yeast Complexes](#)
Tip: Use your favorite sequence with [BINDBlast](#).

The Biomolecular Interaction Network Database (BIND) is a database designed to store full descriptions of interactions, molecular complexes and pathways. Development of the BIND 2.0 data model has led to the incorporation of virtually all components of molecular mechanisms including interactions between any two molecules composed of proteins, nucleic acids and small molecules. Chemical reactions, photochemical activation and conformational changes can also be described. Everything from small molecule biochemistry to signal transduction is abstracted in such a way that graph theory methods may be applied for data mining. The database can be used to study networks of interactions, to map pathways across taxonomic branches and to generate information for kinetic simulations. BIND anticipates the coming large influx of interaction information from high-throughput proteomics efforts including detailed information about post-translational modifications from mass spectrometry.

Database Stats	
INTERACTIONS	6171
COMPLEXES	851
PATHWAYS	8

CRC-CGF Bioinformatics Course, 2002



Victorian Bioinformatics Consortium

Gene Ontology

- <http://www.geneontology.org>
- The goal of the Gene Ontology™ Consortium is to produce a dynamic controlled vocabulary that can be applied to all organisms
 - An ontology is a specification of a conceptualization.
- Enable consistency in cross-species annotation
- Three ontologies under development - molecular function, biological process, and cellular component.

CRC-CGF Bioinformatics Course, 2002



On-Line News Services

- BioMedNet <http://www.bmn.com>
 - The Scientist <http://www.the-scientist.com>
 - Nature <http://www.nature.com>
 - Science <http://www.sciencemag.org>
 - Aust Biotech News <http://www.biotechnews.com.au>
 - Bio-IT <http://www.bio-itworld.com>
-
- Free private user registration
 - Scientific news
 - Weekly citation abstract service

CRC-CGF Bioinformatics Course, 2002



www.nature.com e-alerts

Nature	Nature Science Update	Naturejobs Newsletter
Nature Biotechnology	Nature Cell Biology	Nature Genetics
Nature Immunology	Nature Materials	Nature Medicine
Nature Neuroscience	Nature Structural Biology	
Nature Reviews Cancer	Nature Reviews Drug Discovery	Nature Reviews Genetics
Nature Reviews Immunology	Nature Reviews Molecular Cell Biology	
Nature Reviews Neuroscience		
British Journal of Cancer	Cell Death and Differentiation	Genes and Immunity
Heredity	Journal of Industrial Microbiology & Biotechnology	
Leukemia	Oncogene	
Drug Discovery@nature.com	Materials Update	Nature Cancer Update
Physics Portal	Signaling Update	
Special Offers	Third Party Promotions	

CRC-CGF Bioinformatics Course, 2002

