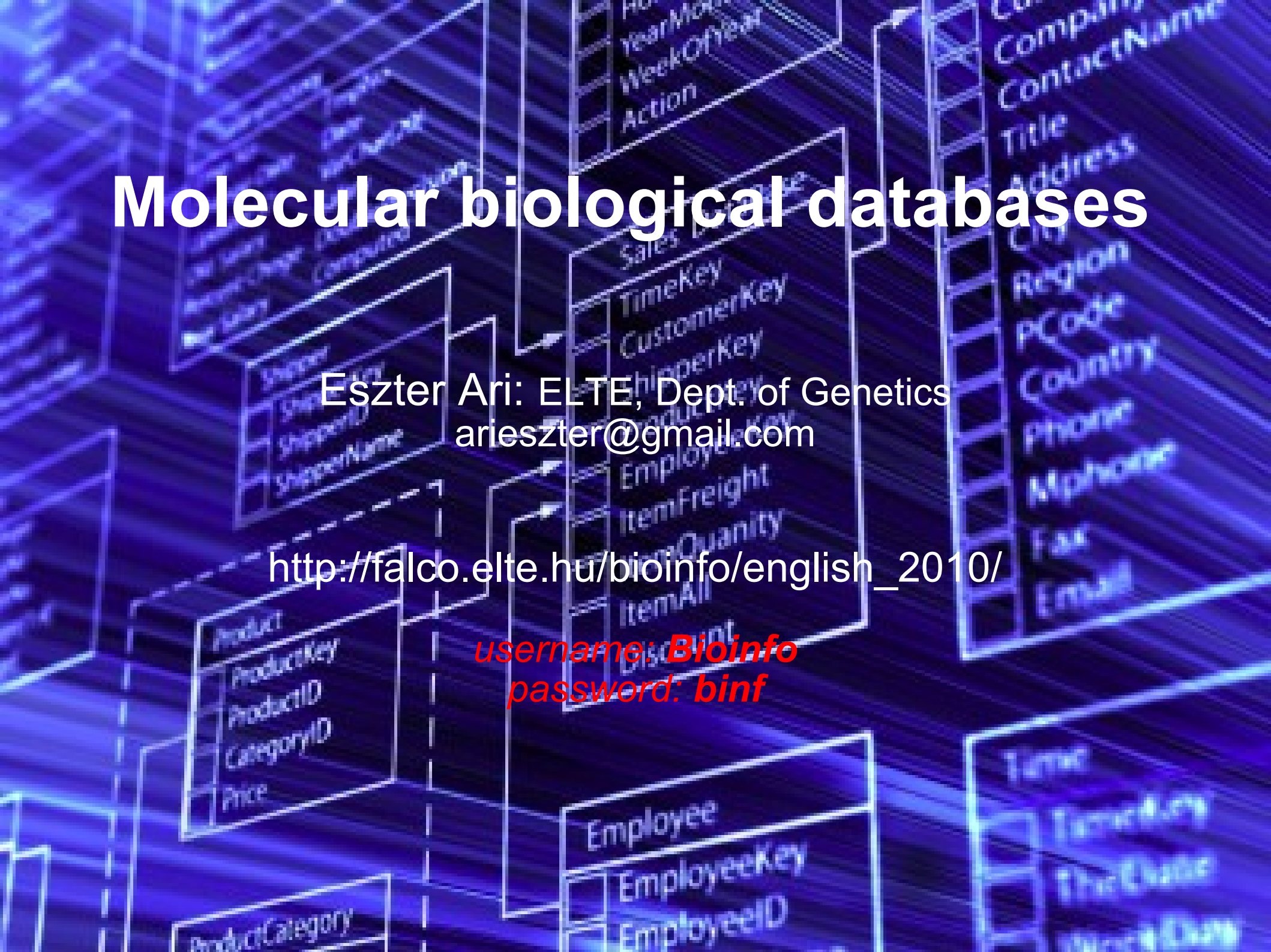


Molecular biological databases

Eszter Ari: ELTE, Dept. of Genetics
arieszter@gmail.com

http://falco.elte.hu/bioinfo/english_2010/

username: ***Bioinfo***
password: ***binf***

Overview

- About databases in general
- Types of molecular biological databases
  - Primary databases: Nucleotide databases
    - The structure and content of databases
  - Protein databases
  - Gene databases
  - Protein structure databases
  - Gene-ontology databases
  - Genome browsers
  - Evolutionary databases

What is a database?

- **Def.:** A database consists of an organized collection of data for one or more uses, typically in digital form. They are managed using database management systems, which store database contents, allowing data creation and maintenance, and search and other access.
- **Goals:** store database contents, allowing data creation and maintenance, and search and other access.
- *Life is easier with databases* ☺

- ~ Like arrays in programming
  - one-, two or more dimensional, e.g.:
    - name 1 → phone No. 1
    - name 2 → phone No. 2
    - .
    - name n → phone No. n

	field 1	field 2	field 3
record 1	value 11	value 12	value 13
record 2	value 21	value 22	value 23
record 3	value 31	value 32	value 33

Database implementations

- One way of classifying databases involves the type of their contents:
- Text-based databases:
  - contain human readable flat files
- Not text-based databases:
  - divergence of data storage and visualization
  - can be read only with a database software
    - e.g. XML (eXtensibe Markup Language),
    - ASN.1 (Abstract Syntax Notation 1): data exchange standard of NCBI

Flatfile example

```
LOCUS       AB107031                 302 bp    DNA     linear   PRI 10-JUL-2009
DEFINITION  Hylobates agilis DRD4 gene for dopamine receptor D4, partial cds,
            haplotype: Hag100.
ACCESSION   AB107031
VERSION     AB107031.1  GI:38141857
KEYWORDS    .
SOURCE      Hylobates agilis (agile gibbon)
  ORGANISM  Hylobates agilis
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hylobatidae; Hylobates.
REFERENCE   1
  AUTHORS   Shimada,M.K., Inoue-Murayama,M., Ueda,Y., Maejima,M., Murayama,Y.,
            Takenaka,O., Hayasaka,I. and Ito,S.
  TITLE     Polymorphism in the second intron of dopamine receptor D4 gene in
            humans and apes
  JOURNAL   Biochem. Biophys. Res. Commun. 316 (4), 1186-1190 (2004)
   PUBMED   15044110
REFERENCE   2  (bases 1 to 302)
  AUTHORS   Shimada,M.K. and Inoue-Murayama,M.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-MAR-2003) Contact:Makoto K Shimada Fujita Health
            University, Institute for Comprehensive Medical Science; 1-98
            Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan
FEATURES             Location/Qualifiers
     source          1..302
                     /organism="Hylobates agilis"
                     /mol_type="genomic DNA"
                     /isolate="Hagil#2480"
                     /db_xref="taxon:9579"
```

XML example

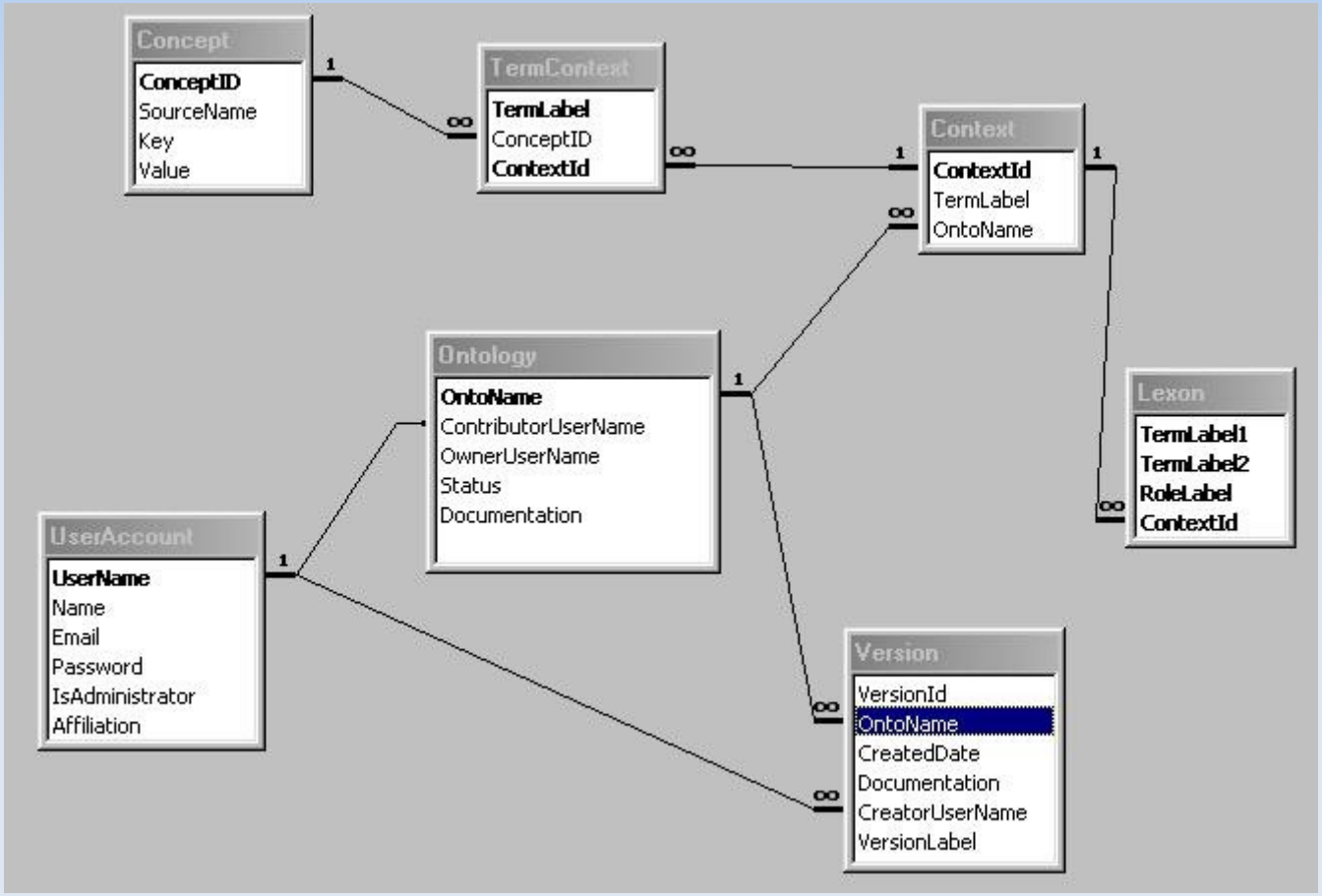
```
<?xml version="1.0" ?>
<!DOCTYPE TIGR (View Source for full doctype...)>
- <TIGR>
  - <PSEUDOCHROMOSOME>
    - <SCAFFOLD>
      - <SCAFFOLD_COMPONENT>
          <ASMBL_ID CLONE_NAME="NOR_4">68405</ASMBL_ID>
          <CHR_LEFT_COORD>1</CHR_LEFT_COORD>
          <CHR_RIGHT_COORD>1000</CHR_RIGHT_COORD>
          <ASMBL_LEFT_COORD>1</ASMBL_LEFT_COORD>
          <ASMBL_RIGHT_COORD>1000</ASMBL_RIGHT_COORD>
          <ORIENTATION>+</ORIENTATION>
          <DATE />
        </SCAFFOLD_COMPONENT>
      - <SCAFFOLD_COMPONENT>
          <ASMBL_ID CLONE_NAME="T15P10">67032</ASMBL_ID>
          <CHR_LEFT_COORD>1001</CHR_LEFT_COORD>
          <CHR_RIGHT_COORD>7001</CHR_RIGHT_COORD>
          <ASMBL_LEFT_COORD>1</ASMBL_LEFT_COORD>
          <ASMBL_RIGHT_COORD>6001</ASMBL_RIGHT_COORD>
          <ORIENTATION>+</ORIENTATION>
          <DATE />
        </SCAFFOLD_COMPONENT>
      </SCAFFOLD>
    - <ASSEMBLY CLONE_ID="0" DATABASE="ATH1" CHROMOSOME="4" CURRENT_DATE="Wed Apr 16 19:43:21 EDT 2003">
      <ASMBL_ID CLONE_NAME="CHR4v03212003">68411</ASMBL_ID>
      - <COORDSET>
          <END5>1</END5>
          <END3>18585042</END3>
        </COORDSET>
      - <HEADER>
          <CLONE_NAME>CHR4v03212003</CLONE_NAME>
        - <SEQ_LAST_TOUCHED>
            <DATE>Mar 22 2003 6:16PM</DATE>
          </SEQ_LAST_TOUCHED>
          <GB_ACCESSION />
          <ORGANISM>Arabidopsis thaliana</ORGANISM>
          <LINEAGE>Eukaryota ; Viridiplantae ; Embryophyta ; Tracheophyta ; Spermatophyta ; Magnoliophyta ; eudicotyledons ; core eudicots ; Rosidae ; eurosids II ;
            Brassicales ; Brassicaceae ; Arabidopsis</LINEAGE>
          <SEQ_GROUP>none</SEQ_GROUP>
          <AUTHOR_LIST CONTACT="cdtown@tigr.org" />
        </HEADER>
      - <GENE_LIST>
        - <PROTEIN_CODING>
          - <TU>
              <FEAT_NAME>68411.t02010</FEAT_NAME>
              <CHROMO_LINK>68169.t00463</CHROMO_LINK>
              <DATE>Nov 15 2001 7:47PM</DATE>
```

Database implementations

- Relational databases
  - cross references (links)
  - logical connections
  - multiple indexing
  - minimal redundancy
  - complexity
  - fast searching
- Database program: (e.g. MySQL, PostgreSQL, Oracle, MS SQL)
  - SQL: Structured Query Language

Structure of relational databases

- Table 1
  - Field 1
  - Field 2
  - Field n
- Table 2
  - Field 3
  - Field 4
  - Field n



Cross references
(connecting tables)

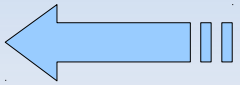
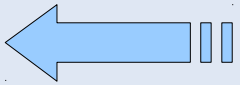
- Table 1 (GenBank)
  - Field 1 (LOCUS)
  - …
  - Field n taxid

    e.g. 3702

- Table 2 (Taxonomy)
  - Field 1 (taxid, e.g. 3702)
  - …
  - Field n (species name) *Arabidopsis thaliana*

More record can be connected with the same species

The structure of sequence databases

- Table (e.g. GenBank)
  - Record X
    - Annotation  text searching
      - Field 1 (e.g. Locus)
      - Fiels 2 (e.g. Definition)
      - Etc.
    - Sequence  similarity searching
      - Field n (e.g. cgagcatgcatctagtagcagcgtactac)

Types of molecular biology databases

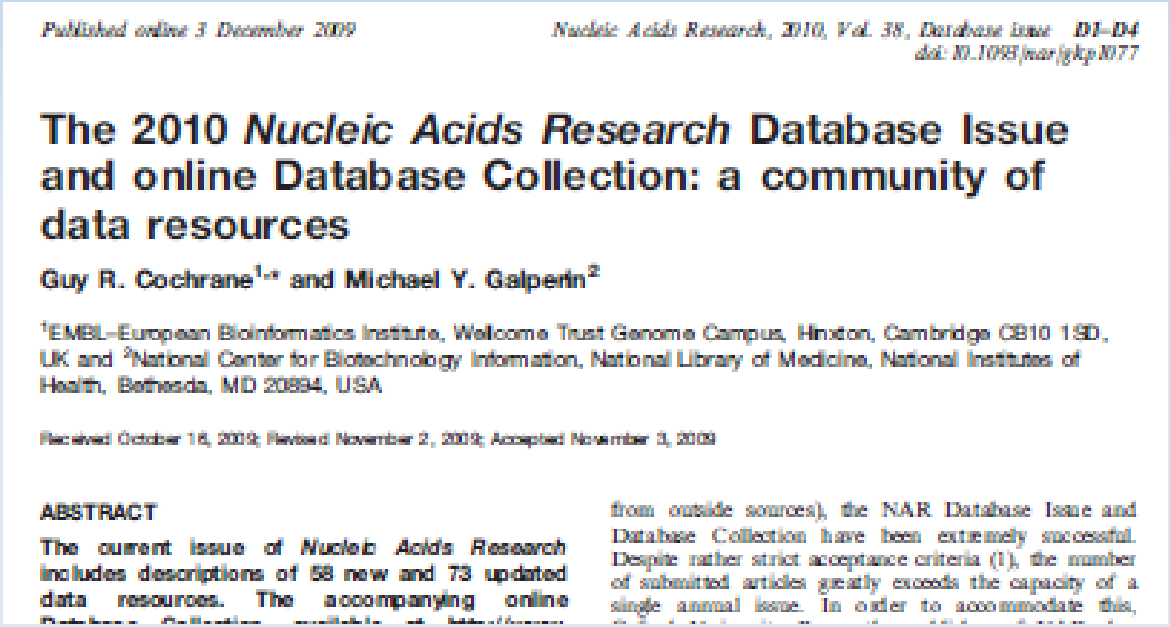
1. Primary databases
   - Nucleotide sequence databases
   - (Other: e.g. structural databases ← NMR data)
2. Secondary or derived databases
   - Protein databanks ← translated from coding DNA
   - Motive databanks (e.g: promoters)
3. Tertiary databanks:
   - connections or network batabases
   - connections of nucleotides and mainly proteins

Other, not sequence databases

- evolutionary, publication, …

Databases in science literature

Nucleic Acids Research: „Database Issue" first issue in every year



Published online 3 December 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue D1–D4*
doi: 10.1093/nar/gkp1077

The 2010 *Nucleic Acids Research* Database Issue and online Database Collection: a community of data resources

Guy R. Cochrane[1,*] and Michael Y. Galperin[2]

[1]EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received October 16, 2009; Revised November 2, 2009; Accepted November 3, 2009

ABSTRACT

The current issue of *Nucleic Acids Research* includes descriptions of 58 new and 73 updated data resources. The accompanying online

from outside sources), the NAR Database Issue and Database Collection have been extremely successful. Despite rather strict acceptance criteria (1), the number of submitted articles greatly exceeds the capacity of a single annual issue. In order to accommodate this,

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases

DATABASE The Journal of Biological Databases and Curation

Database part: e.g. *Bioinformatics*, *BMC Bioinformatics*

Great nucleotide sequential databases

- EMBL - European Molecular Biology Laboratory
  - Primary database in Europe
  - operated by **EBI** (European Bioinformatics Institute)
  - founded in 1980 Heidelbergben (D)
  - today: Hinxton (UK)
  - www.ebi.ac.uk/embl/
- GenBank
  - Primary database of the USA
  - operated by **NCBI** (National Center for Biotechnology Information)
  - founded in 1979 Los Alamosban (New Mexico, USA)
  - till 1992: Bethesda, Maryland
  - www.ncbi.nlm.nih.gov/Genbank/
- DDBJ - DNA Database of Japan
  - operated by CIB - Center for Information Biology, Mishima, Japan
  - www.ddbj.nig.ac.jp

Whats common in this 3 databases?

- Close cooperation, everyday data exchange → both contains the same sequences
  - same AC or accesion number for every submitted sequences → individual registrational number, eg: AY226138
  - common „feature table” (description of seq. features)
  - common taxonomy project
- It is enought to use one of them.
- Different database structure, sequence format
  - sequence format convertion: *readseq* (UNIX), *seqret* (EMBOSS), *SeqVerter* (Windows)

Submitting sequences

- Submitted sequences: different research groups, genome sequencing projectes
  - *sequence submission* → individual accesion number for every sequences
  - the submitted sequence can be partail too (*partial cds*), or non coding region, etc...
  - sci. papers about sequenced data can be published just after sequence submission

- EMBL: *WEBin*
  http://www.ebi.ac.uk/submission/webin.html
- GenBank: *New BankIt*
  http://www.ncbi.nlm.nih.gov/WebSub/index.cgi?tool=genbank

The size of EMBL and the top organisms

2010. Sep. 25.

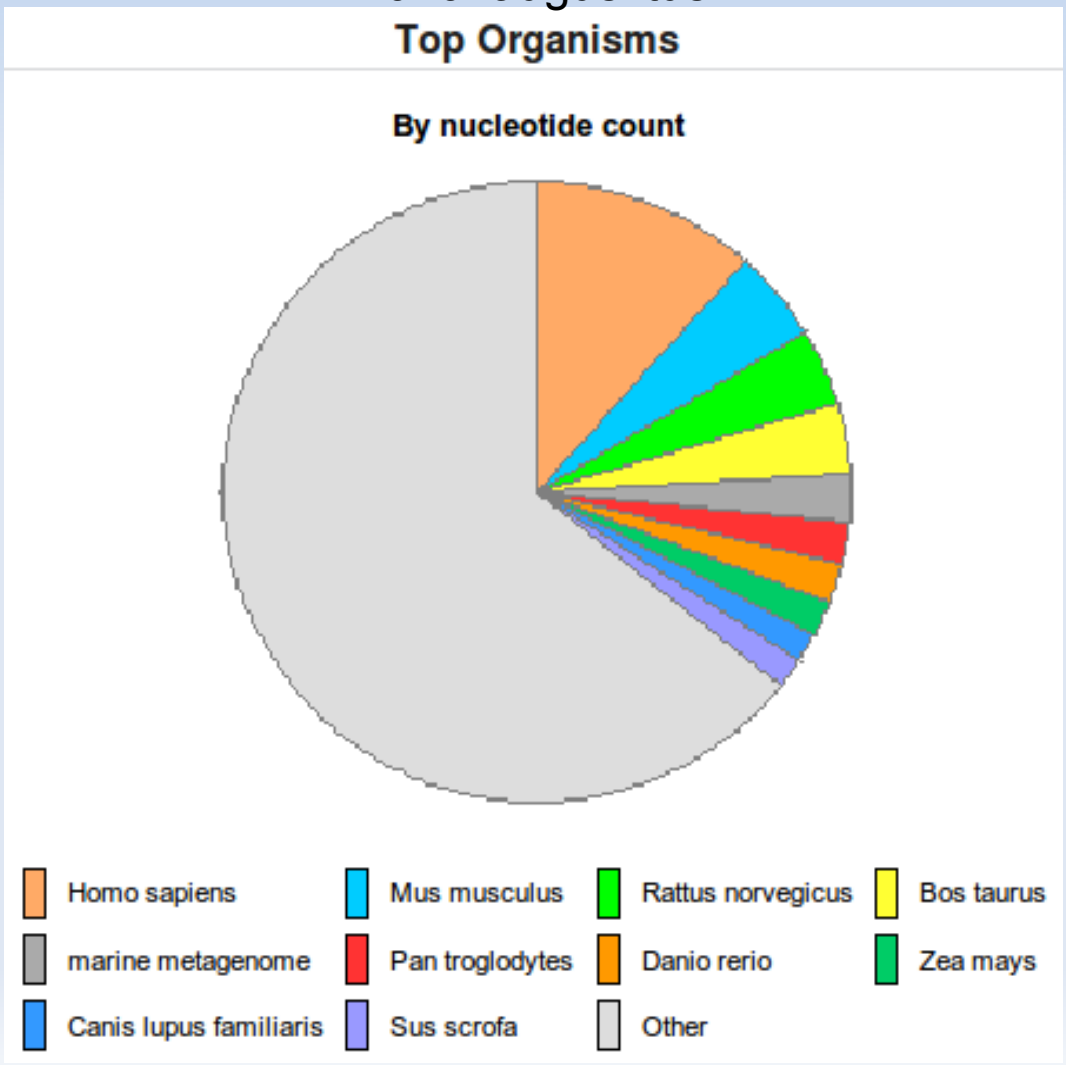
298 billion (giga)

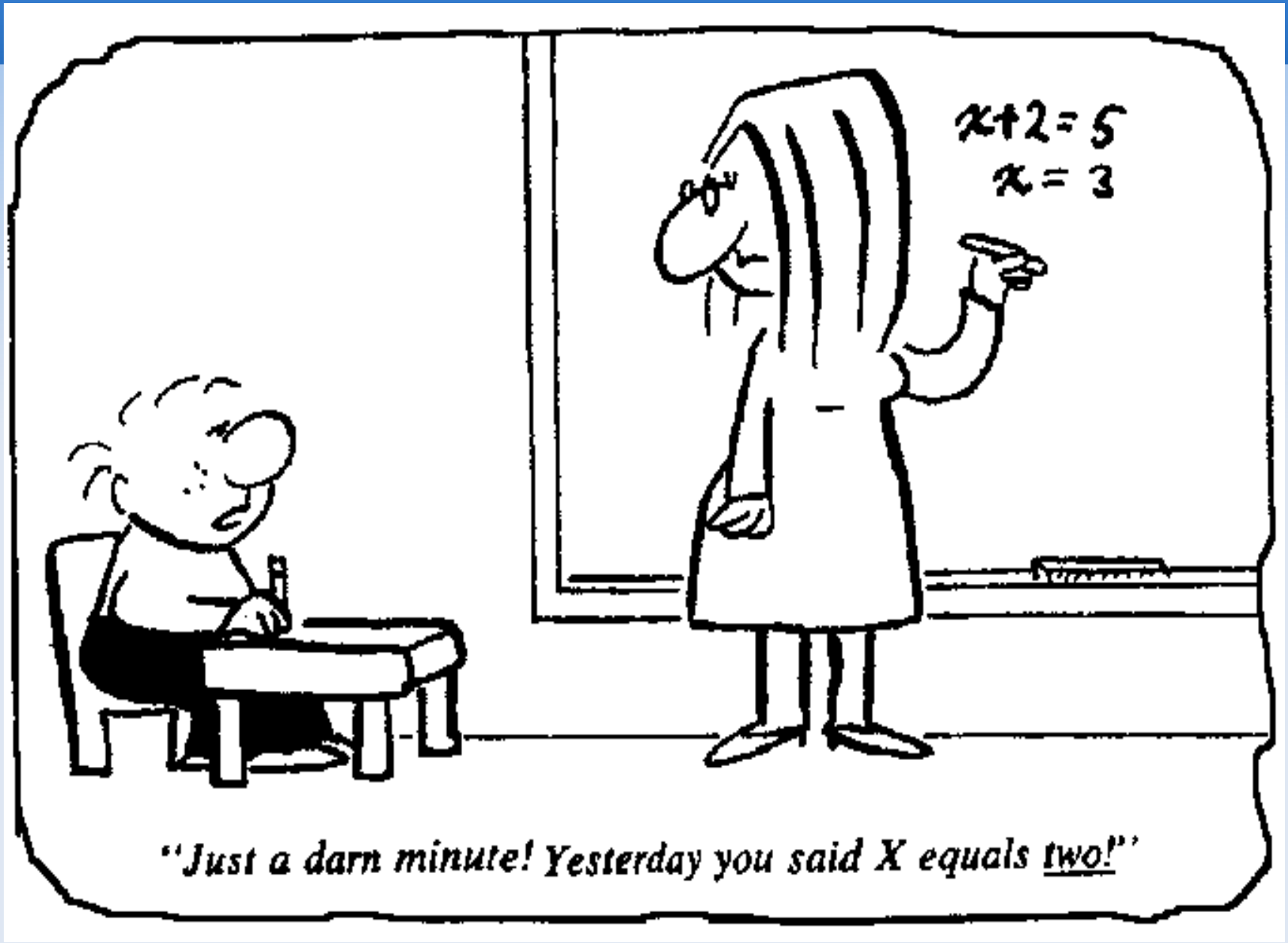
196 million (mega)

This week the EMBL Database contained **298,166,804,584** nucleotides in **195,945,264** entries.

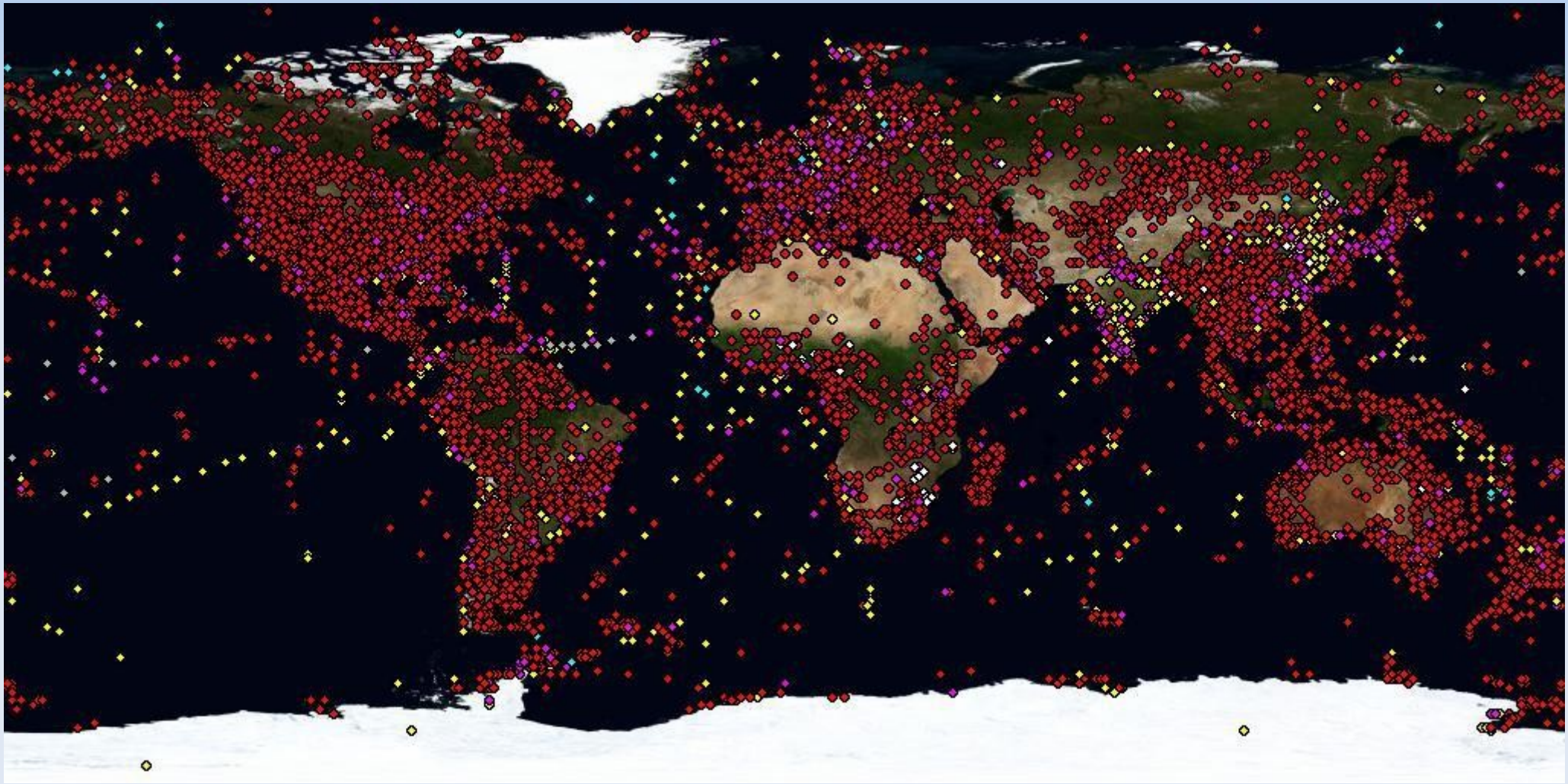
record

2010. augusztus



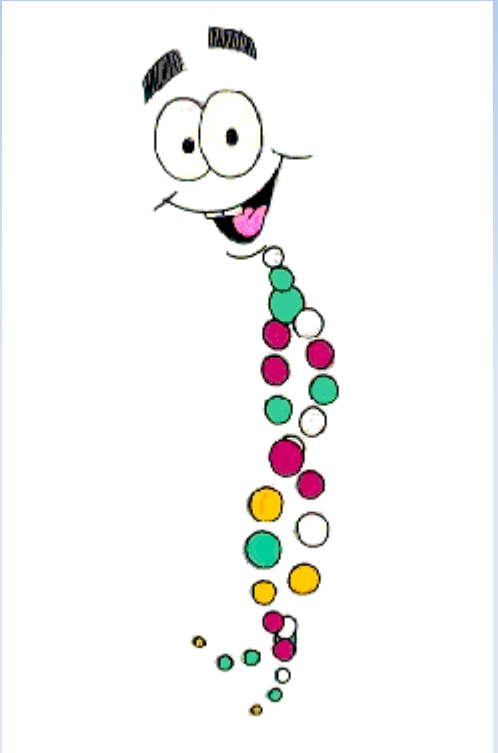

x+2=5
x=3
"Just a darn minute! Yesterday you said X equals two!"

Distribution of sequenced sequences



The structure of databases

- „flatfile” format
- Records (or entrys)
- Field
  - Annotation
  - Sequence
- Sections / Divisions
  - Mainly according to taxonomy
  - it has been changed over time



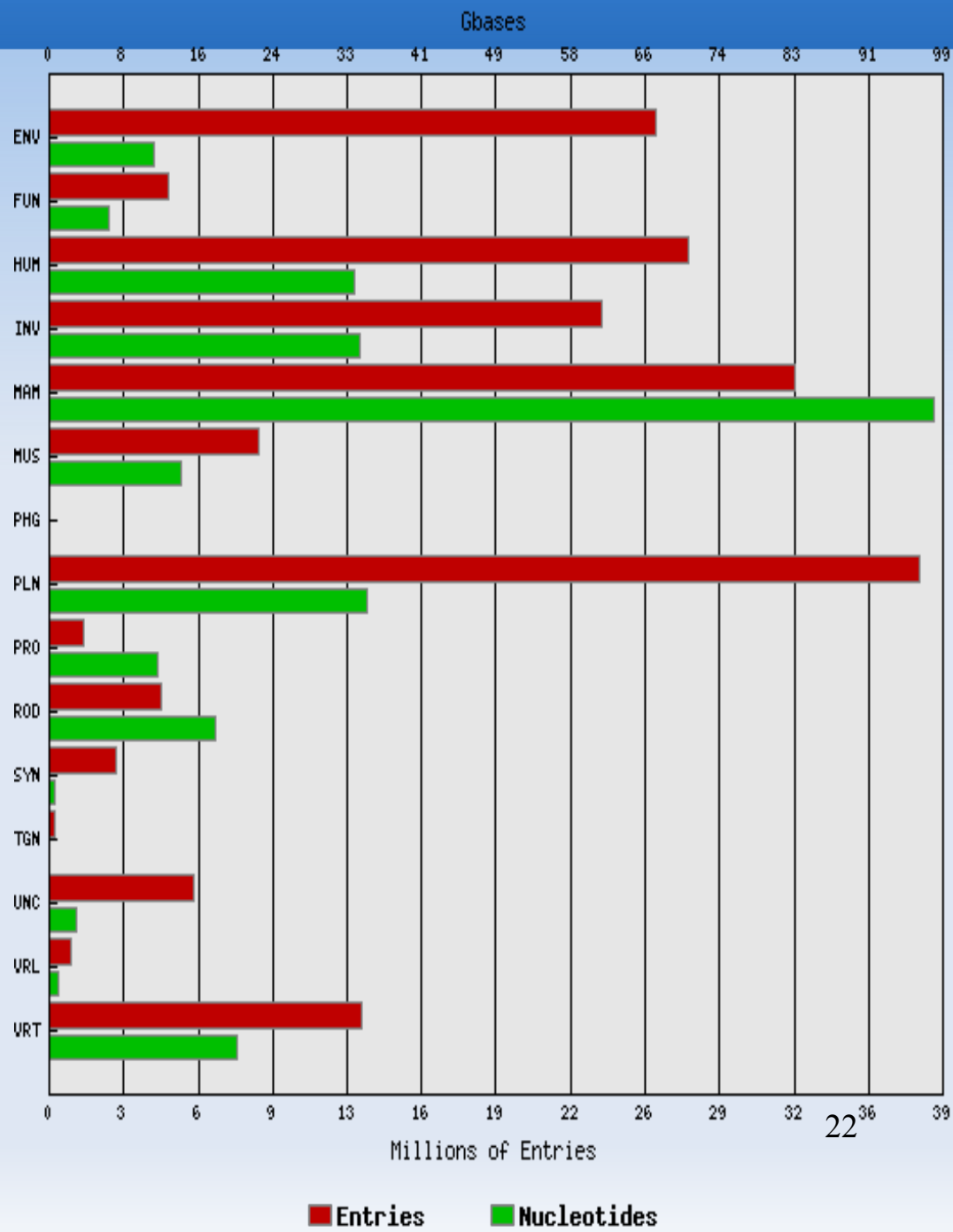
Main EMBL data classes

- **STD: Standard** (conventional sequences)
- STS: Sequence Tagged Site (PCR)
  - Short (200 to 500 base pair) DNA sequence that has a single occurrence in the genome and whose location and base sequence are known.
- EST: Expressed Sequence Tag
  - Short sub-sequence of a transcribed cDNA sequence. They may be used to identify gene transcripts, and are instrumental in gene discovery and gene sequence determination.
- WGS: Whole Genome Shotgun
- HTG: High Throughput Genome sequencing (unfinished)
- HTC: High Throughput cDNA sequencing (unfinished)
- GSS: Genome Sequence Scan (random genomic)
  - like ESTs, but from genomes
- TSA: Transcriptome Shotgun Assembly
- PAT: Patents


Savage Chickens
by Doug Savage
DO YOU THINK YOU'LL EVER HAVE KIDS?
I CAN'T. MY DNA SEQUENCE HAS BEEN PATENTED
©2005 BY DOUG SAVAGE
www.savagechickens.com

EMBL divisions

- ENV: Environmental Samples
- FUN: Fungi
- HUM: Human
- INV: Invertebrates
- MAM: Other Mammals
- MUS: Mus musculus
- PHG: Bacteriophage
- PLN: Plants
- PRO: Prokaryotes
- ROD: Rodents
- SYN: Synthetic
- TGN: Transgenic
- UNC: Unclassified
- VRL: Viruses
- VRT: Other Vertebrates



Main fileds in an EMBL record

- ID: (identification) individual identification string

  (entryname dataclass; molecule; division; sequencelength BP.)
- **AC**: (accession number): invariable → refered by this
- SV: (sequence version)
- DT: (date) of creation and modification
- DE: (description) short
- KW: (keyword)
- OS: (organism species); OC: classification; OG: (organelle)
- R?: references: RN (reference number), RC (reference comment), RP (reference positions), RX (reference cross-reference), RA (reference authors), RT (reference title), RL (reference location)
- DR: (databese cross-references)
- CC: (comments)
- FT: (feature table)
- XX: empty space
- SQ: (sequence header): the length of sequence, base content, sequence
- //: end of record

An EMBL record (part 1)

```
ID   HSCYCLOX   standard; mRNA; HUM; 3387 BP.
XX
AC   M90100;
XX
SV   M90100.1
XX
DT   30-MAR-1992 (Rel. 31, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 7)
XX
DE   Homo sapiens cyclooxygenase-2 (Cox-2) mRNA, complete cds.
XX
KW   cyclooxygenase-2; prostaglandin synthase.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RP   1-3387
RX   MEDLINE; 92366465.
RX   PUBMED; 1380156.
RA   Hla T., Neilson K.;
RT   "Human cyclooxygenase-2 cDNA";
RL   Proc. Natl. Acad. Sci. U.S.A. 89(16):7384-7388(1992).
XX
DR   GOA; P35354.
DR   SWISS-PROT; P35354; PGH2_HUMAN.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..3387
FT                   /db_xref="taxon:9606"
FT                   /mol_type="mRNA"
FT                   /organism="Homo sapiens"
FT                   /cell_type="endothelial"
FT                   /tissue_type="umbilical vein"
```

An EMBL record (part 2)

```
FT   5'UTR           1..97
FT                   /gene="Cox-2"
FT   CDS             98..1912
FT                   /codon_start=1
FT                   /db_xref="GOA:P35354"
FT                   /db_xref="SWISS-PROT:P35354"
FT                   /gene="Cox-2"
FT                   /EC_number="1.14.99.1"
FT                   /product="cyclooxygenase-2"
FT                   /protein_id="AAA58433.1"
FT                   /translation="MLARALLLCAVLALSHTANPCCSHPCQNRGVCMSVGFDQYKCDCT
FT                   RTGFYGENCSTPEFLaTRIKLFLKPTPNTVHYILTHFKGFWNVVNNIPFLRNAIMSYVLT
FT                   ...
FT                   KGLMGNVICSPAYWKPSTFGGEVGFQIINTASIQSLICNNVKGCPFTSFSVPDPELIKT
FT                   VTINASSSRSGLDDINPTVLLKERSTEL"
FT   sig_peptide     98..148
FT                   /gene="Cox-2"
FT   mat_peptide     149..1909
FT                   /gene="Cox-2"
FT                   /EC_number="1.14.99.1"
FT                   /product="cyclooxygenase-2"
FT   3'UTR           1913..3387
FT                   /gene="Cox-2"
FT   polyA_signal    3369..3374
FT                   /gene="Cox-2"
XX
SQ   Sequence 3387 BP; 1010 A; 712 C; 633 G; 1032 T; 0 other;
     gtccaggaac tcctcagcag cgcctccttc agctccacag ccagacgccc tcagacagca        60
     aagcctaccc ccgcgccgcg ccctgcccgc cgctgcgatg ctcgcccgcg ccctgctgct       120
     ...
     tacctgaact tttgcaagtt ttcaggtaaa cctcagctca ggactgctat ttagctcctc      3360
     ttaagaagat taaaaaaaaa aaaaaag                                          3387
//
```

Annotation: EMBL vs. GenBank

EMBL

- ID – individual identification
- AC – accession No.
- = GenBank ACCESSION
- SV – entry version
- DE – description
- OS – species
- OC – taxonomy
- FT – „feature table”:

GenBank

- LOCUS – kihalóban? a formátum miatt marad
- ACCESSION – accession No.
- = EMBL AC
- VERSION – entry verion, Accession.Version
- DEFINITION – description
- SOURCE – trivial species name
- ORGANISM – species, taxonomy
- FEATURES – „feature table”

A GenBank record (part 1)

```
LOCUS       HUMCYCLOX                3387 bp    mRNA    linear   PRI 31-DEC-1994
DEFINITION  Homo sapiens cyclooxygenase-2 (Cox-2) mRNA, complete cds.
ACCESSION   M90100
VERSION     M90100.1  GI:181253
KEYWORDS    cyclooxygenase-2; prostaglandin synthase.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 3387)
  AUTHORS   Hla,T. and Neilson,K.
  TITLE     Human cyclooxygenase-2 cDNA
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 89 (16), 7384-7388 (1992)
  MEDLINE   92366465
   PUBMED   1380156
COMMENT     Original source text: Homo sapiens umbilical vein cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..3387
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /cell_type="endothelial"
                     /tissue_type="umbilical vein"
     gene            1..3387
                     /gene="Cox-2"
     5'UTR           1..97
                     /gene="Cox-2"
```

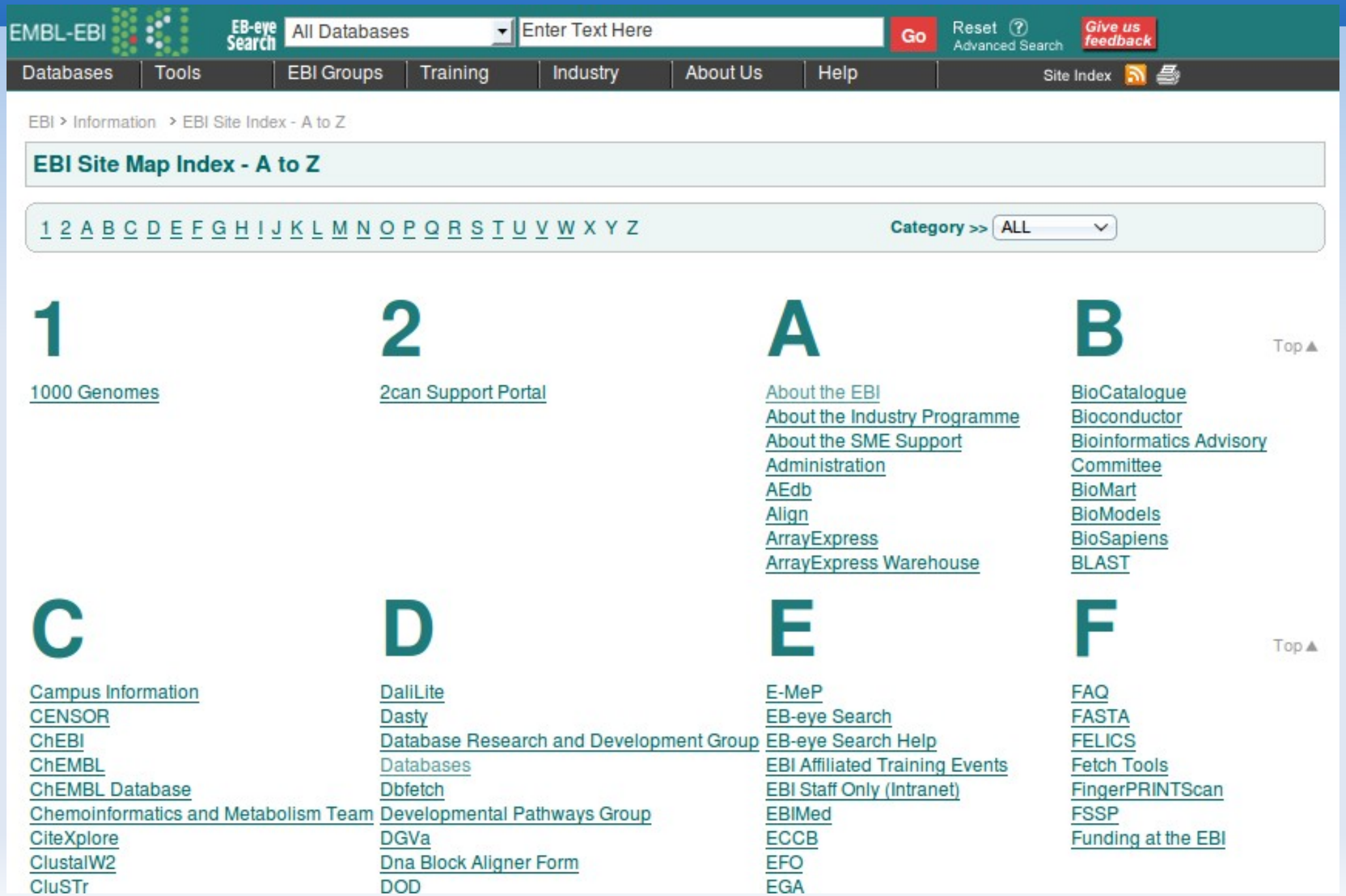
A GenBank record (part 2)

```
CDS             98..1912
                     /gene="Cox-2"
                     /EC_number="1.14.99.1"
                     /codon_start=1
                     /product="cyclooxygenase-2"
                     /protein_id="AAA58433.1"
                     /db_xref="GI:181254"
                     /translation="MLARALLLCAVLALSHTANPCCSHPCQNRGVCMSVGFDQYKCDC
                     TRTGFYGENCSTPEFLTRIKLFLKPTPNTVHYILTHFKGFWNVVNNIPFLRNAIMSYV
                     ...
                     VEVGAPFSLKGLMGNVICSPAYWKPSTFGGEVGFQIINTASIQSLICNNVKGCPFTSF
                     SVPDPELIKTVTINASSSRSGLDDINPTVLLKERSTEL"
     sig_peptide     98..148
                     /gene="Cox-2"
     mat_peptide     149..1909
                     /gene="Cox-2"
                     /product="cyclooxygenase-2"
                     /EC_number="1.14.99.1"
     3'UTR           1913..3387
                     /gene="Cox-2"
     polyA_signal    3369..3374
                     /gene="Cox-2"
BASE COUNT     1010 a    712 c    633 g   1032 t
ORIGIN
        1 gtccaggaac tcctcagcag cgcctccttc agctccacag ccagacgccc tcagacagca
       61 aagcctaccc ccgcgccgcg ccctgcccgc cgctgcgatg ctcgcccgcg ccctgctgct
          ...
     3301 tacctgaact tttgcaagtt ttcaggtaaa cctcagctca ggactgctat ttagctcctc
     3361 ttaagaagat taaaaaaaaa aaaaaag
//
```

Reliability

- It is important to know: there are some obligatory fields in a record – that must be filled in by the researcher who is submitting a sequence but a lot depends on the researcher!
- Sometimes it happenes that a record contains false or out-of-date information.
- → Check the important sequences from more source!

EBI site index



EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

EBI > Information > EBI Site Index - A to Z

EBI Site Map Index - A to Z

1 2 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Category >> ALL

1

1000 Genomes

2

2can Support Portal

A

About the EBI
About the Industry Programme
About the SME Support
Administration
AEdb
Align
ArrayExpress
ArrayExpress Warehouse

B

Top▲

BioCatalogue
Bioconductor
Bioinformatics Advisory Committee
BioMart
BioModels
BioSapiens
BLAST

C

Campus Information
CENSOR
ChEBI
ChEMBL
ChEMBL Database
Chemoinformatics and Metabolism Team
CiteXplore
ClustalW2
CluSTr

D

DaliLite
Dasty
Database Research and Development Group
Databases
Dbfetch
Developmental Pathways Group
DGVa
Dna Block Aligner Form
DOD

E

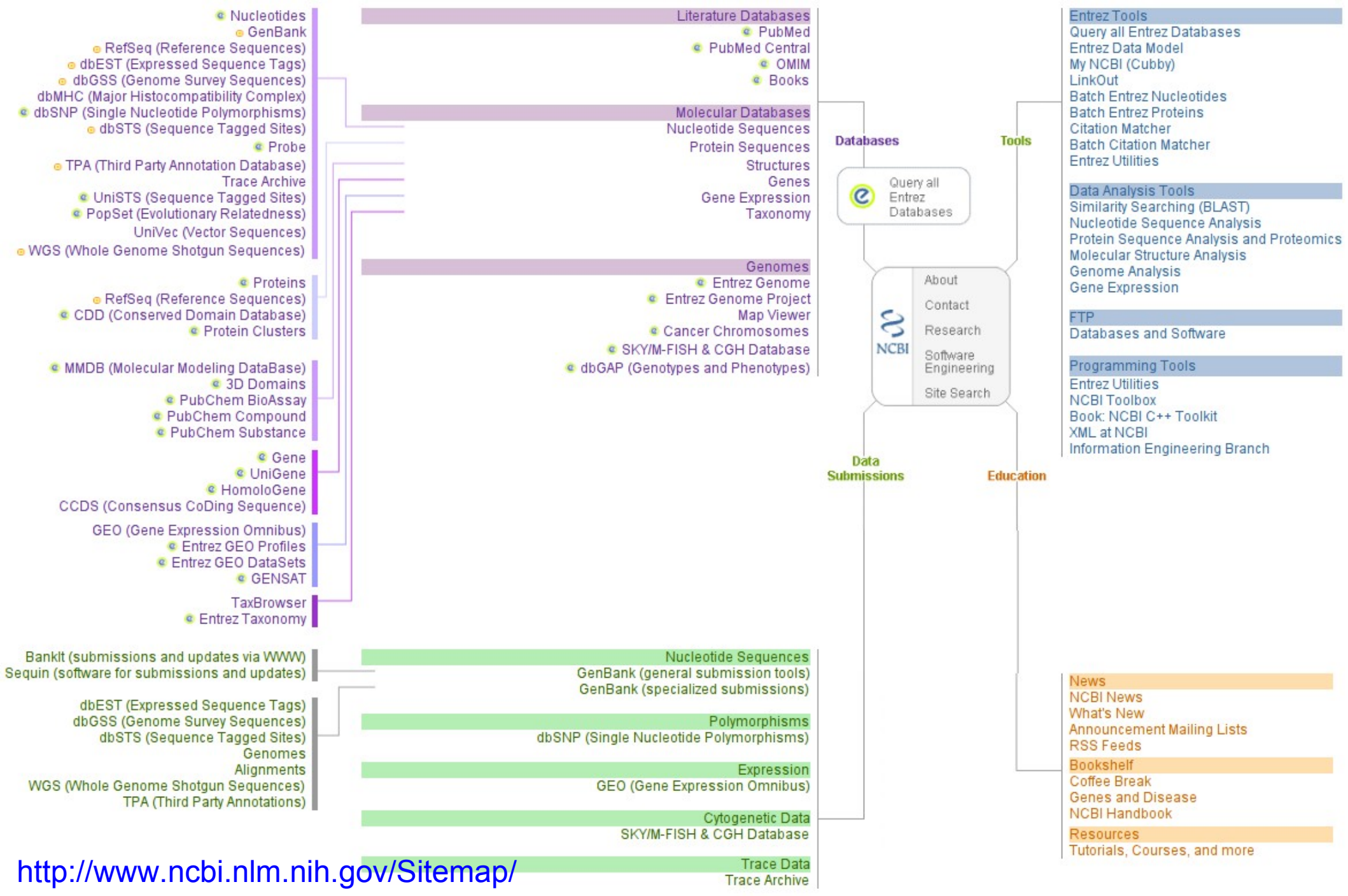
E-MeP
EB-eye Search
EB-eye Search Help
EBI Affiliated Training Events
EBI Staff Only (Intranet)
EBIMed
ECCB
EFO
EGA

F

Top▲

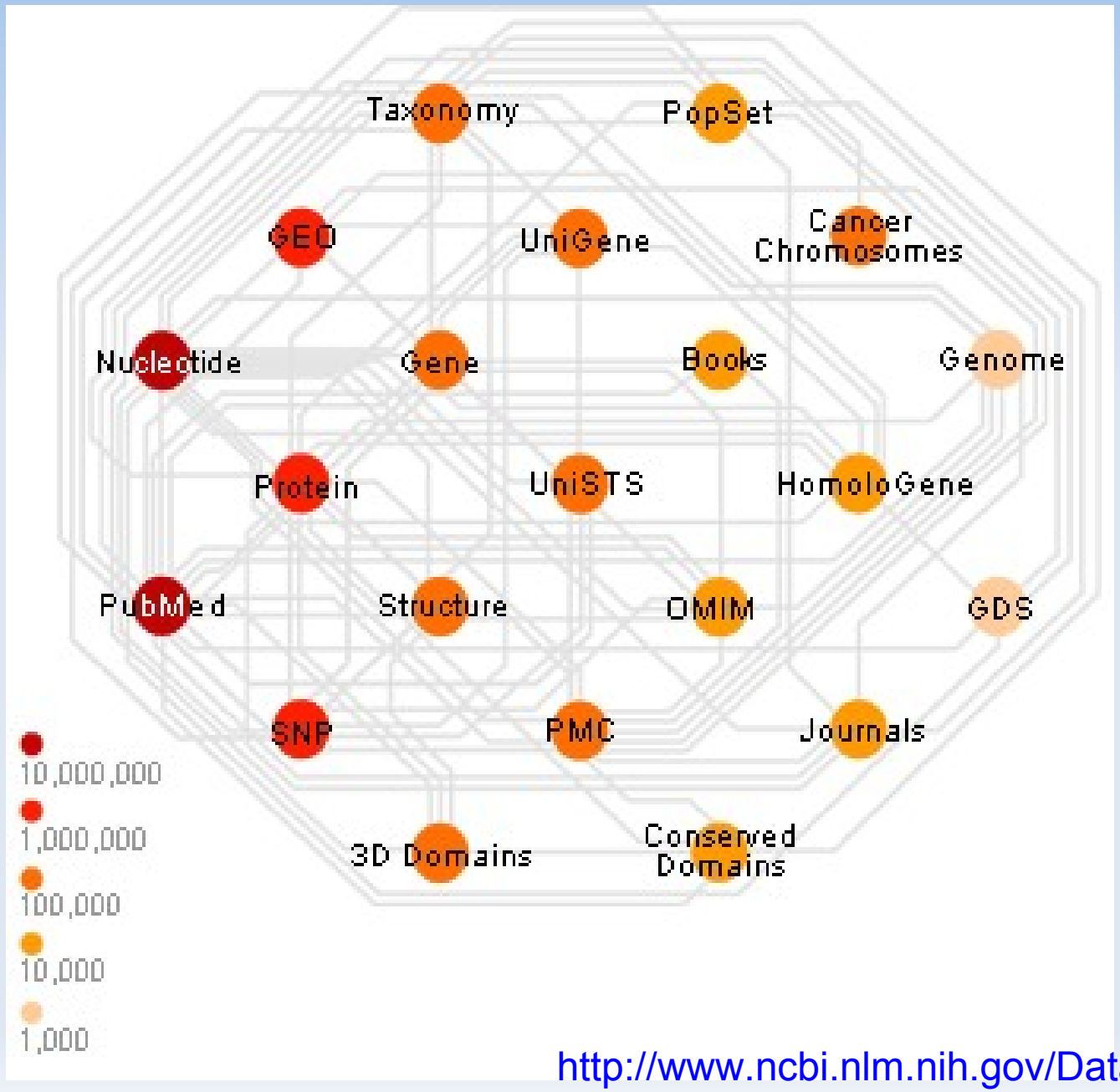
FAQ
FASTA
FELICS
Fetch Tools
FingerPRINTScan
FSSP
Funding at the EBI

NCBI site map



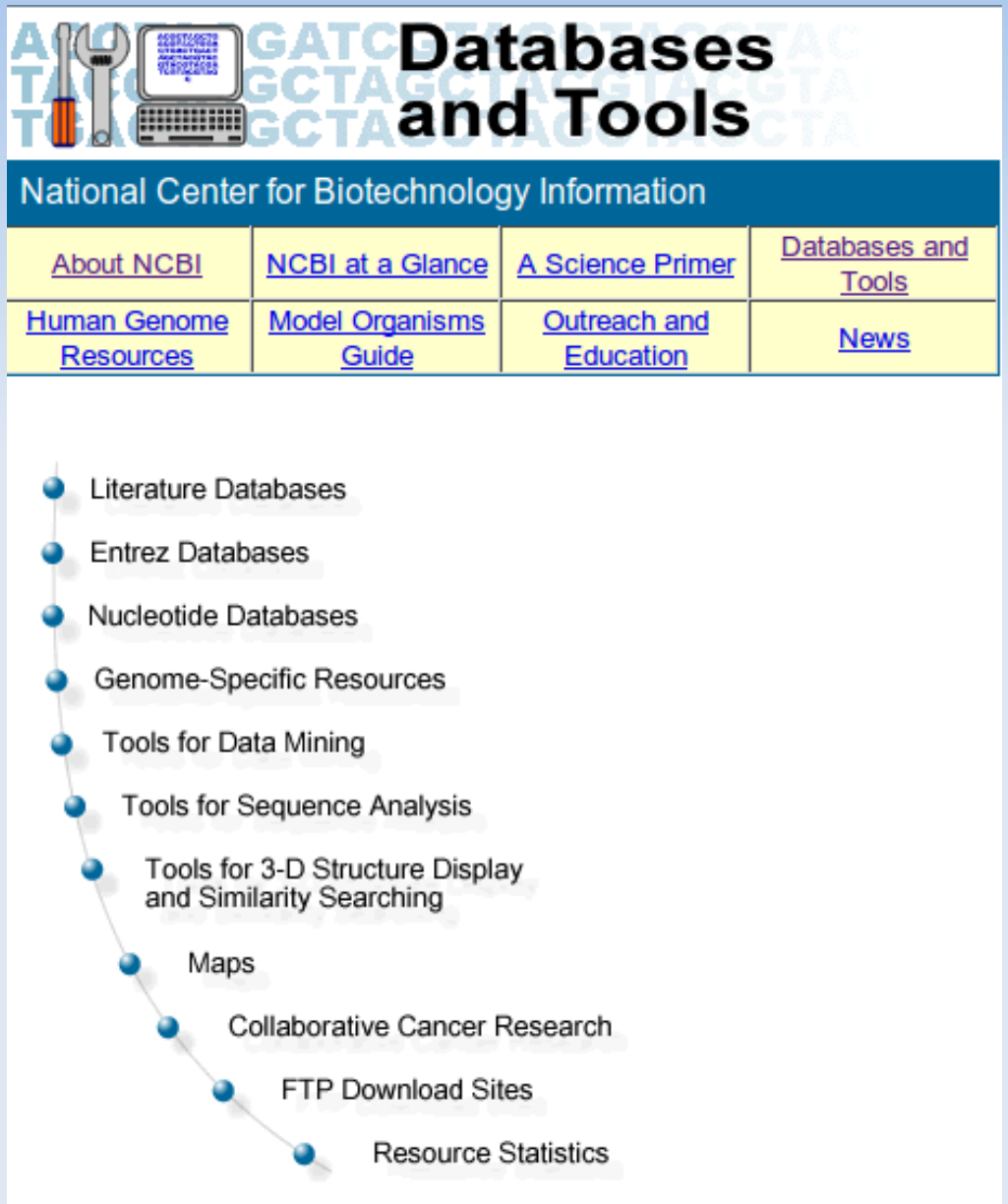
http://www.ncbi.nlm.nih.gov/Sitemap/

NCBI Entrez databases

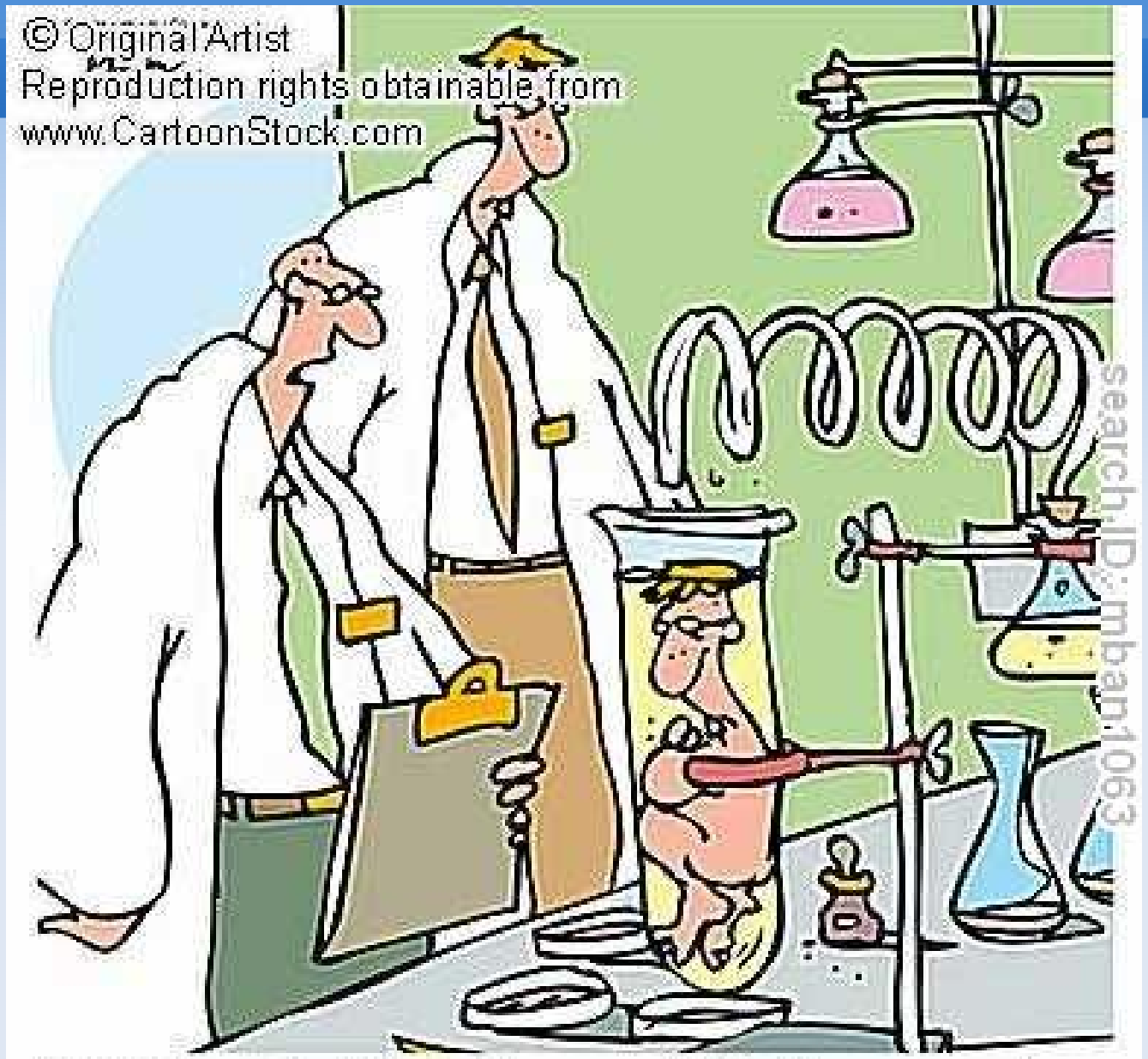


http://www.ncbi.nlm.nih.gov/Database/index.html

Learn about NCBI databases and tools



http://www.ncbi.nlm.nih.gov/About/tools/index.html


© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
search ID: mbap1063
"Well, it certainly looks like your DNA. How many times have I told you to wear gloves before touching anything?"

Protein sequence databanks I.: History

- Swiss-Prot
  - http://www.expasy.ch/sprot/
  - SIB (Swiss Institute of Bioinformatics) and EBI collaboration
  - Protein knowledgebase (ExPASy = Expert Protein Analysis System)
  - Best annotated database (annotated by hand)
  - Minimal redundancy
  - Good cross references
  - EMBL-like record format
  - Slow sequence dislpaying
  - → merged to ***UniProt*** database
- TrEMBL
  - Translated EMBL
  - automatized annotation
  - SP-TrEMBL: checked part
  - REM-TrEMBL: unchecked part – there is a possibility for unrealistic proteins
  - → merged to ***UniProt*** database

Protein sequence databanks II.

- PIR (Protein Identification Resource)
  - http://pir.georgetown.edu/
  - From *Margaret Dayhoff's* protein atlas, 1960s, National Biomedical Research Foundation (USA)
  - annotated by hand
  - Better cross references
  - Superfamily classification
  - 4 section: PIR1, PIR2, PIR3, PIR4 (best annotated: PIR1)
  - → merged to ***UniProt*** database
- Genpept
  - On the NCBI: http://www.ncbi.nlm.nih.gov/
  - Translated GenBank cDNAs (NCBI), like TrEMBL

PIR

A UniProt CONSORTIUM MEMBER

Protein Information Resource

TLALPN----RKAVADHLLM
LIGCLRNCSAVTAAAKQLAE
VTGFSN----AKTTAQHVKK

Protein Search | Site Search

About PIR | Databases | Search/Analysis | Download | Support

INTEGRATED PROTEIN INFORMATICS RESOURCE FOR GENOMIC, PROTEOMIC AND SYSTEMS BIOLOGY RESEARCH

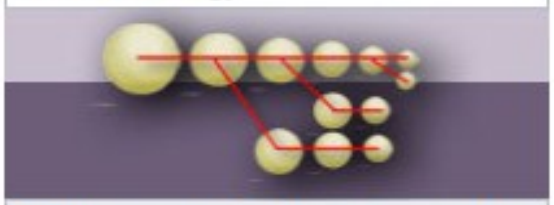
UniProt

The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information.

UniProtKB | UniRef | UniParc

Current release: 2010_09

PRO
Protein Ontology



- Representation of protein objects with descriptions and relationships
- Browse PRO
- Annotate with RACE-PRO

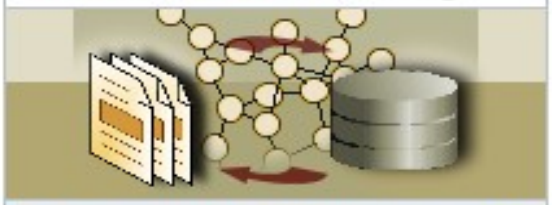
Sample PRO report

iProClass
Integrated Protein Knowledgebase

- Value-added reports for UniProtKB and unique UniParc proteins
- Functional analysis and protein ID mapping

Sample protein report

iProLINK
Literature Information & Knowledge



- Source for text mining and ontology development
- RLIMS-P text mining tool, BioThesaurus
- Bibliography mapping

Sample Biblio. report

O OTHER RESOURCE

- Proteomics: NIAID Biodefense Proteomics Admin. Center
- PIR Grid-Enablement: Data node on NCI's caBIG

P PEPTIDE SEARCH ?

DATABASE: UniProtKB

Use single letter amino acid code

T TEXT SEARCH ?

DATABASE: iProClass

Protein sequence databanks III.

- UniProt: Universal Protein Resource
  - http://www.uniprot.org
  - Fusion of *EBI/SIB Swiss-Prot* + *TrEMBL* and *PIR* → UniProt Consortium (2002)
  - Three database layers:
    - **UniProtKB**: *UniProt Knowledgebase*: contains the sequences and the annotations
      - 2 parts:


annotated by hand: like Swiss-Prot (2004)


annotated by computers: like TrEMBL

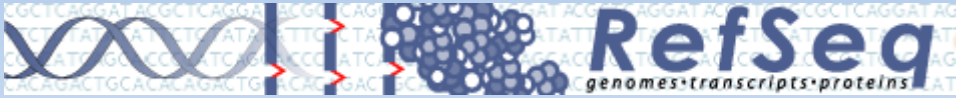
    - **UniRef**: *UniProt Non-redundant Reference*, non redundant clastered sequences → accelerates the similarity searches (BLAST)
    - **UniParc**: *UniProt Archive*, sequence version archive

UniProt record (part)

```
ID   RBL_WHEAT               Reviewed;         477 AA.
AC   P11383; Q7YKX2;
DT   01-JUL-1989, integrated into UniProtKB/Swiss-Prot.
DT   01-AUG-1990, sequence version 2.
DT   16-JUN-2009, entry version 81.
DE   RecName: Full=Ribulose bisphosphate carboxylase large chain;
DE            Short=RuBisCO large subunit;
DE            EC=4.1.1.39;
DE   Flags: Precursor;
GN   Name=rbcL;
OS   Triticum aestivum (Wheat).
OG   Plastid; Chloroplast.
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BEP clade;
OC   Pooideae; Triticeae; Triticum.
OX   NCBI_TaxID=4565;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RA   Terachi T., Ogihara Y., Tsunewaki K.;
RT   "The molecular basis of genetic diversity among cytoplasms of Triticum
RT   and Aegilops. VI. Complete nucleotide sequences of the rbcL genes
RT   encoding H- and L-type rubisco large subunits in common Wheat and Ae.
RT   crassa 4x.";
RL   Jpn. J. Genet. 62:375-387(1987).
RN   [2]
...
```

Non redundant databases

- NCBI RefSeq



  - comprehensive, integrated, well annotated
  - genomic DNA, cDNA, protein

- NCBI UniGene

  - ordered display of transcriptome

3-D protein structure databases

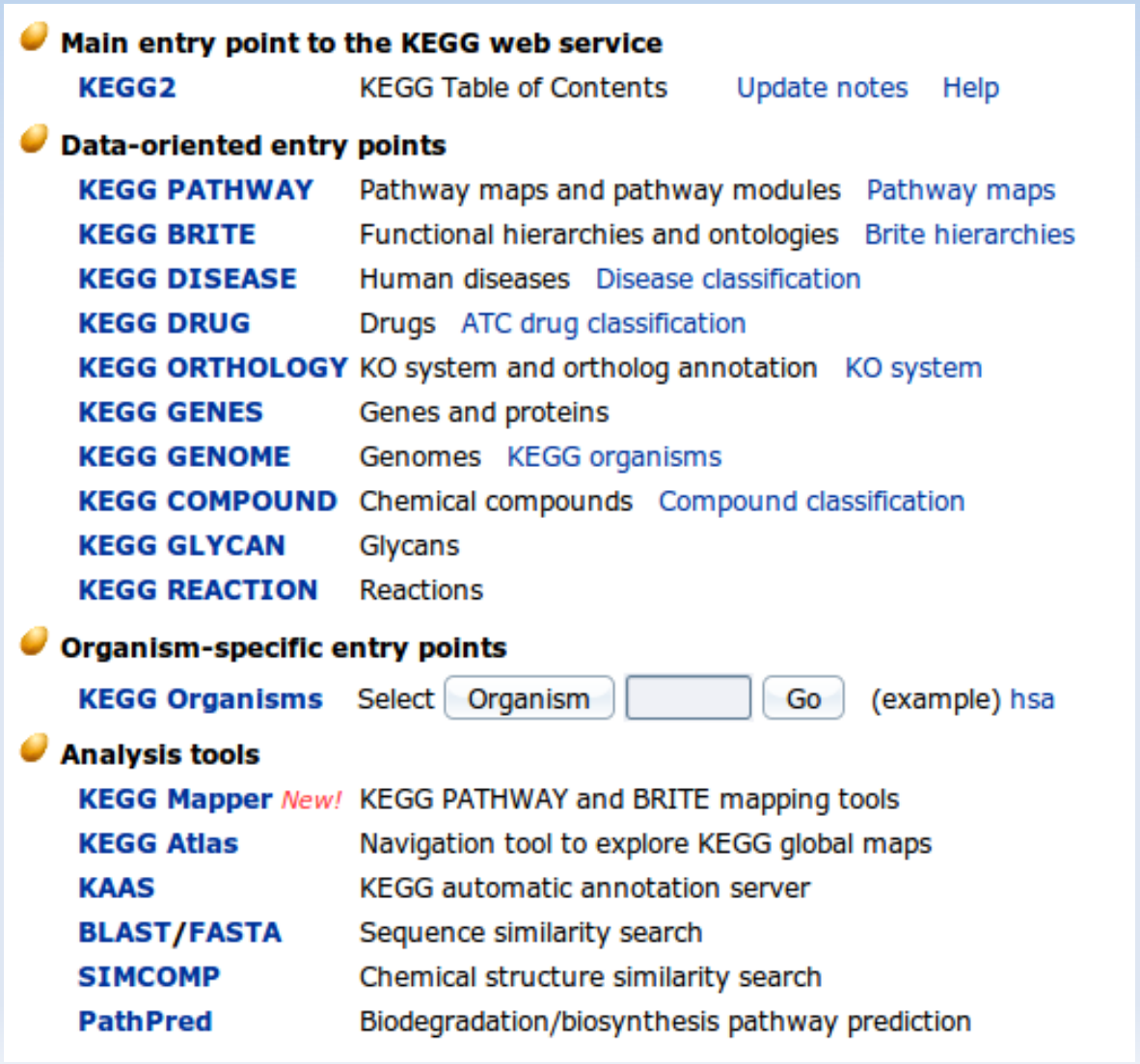
- PDB (Protein Data Bank)
  - Research Collaboratory for Structural Bioinformatics, USA
  - http://www.rcsb.org/pdb/
  - experimentally specified structures (X-ray diffraction, NMR, MRI)
- NCBI Structure:
  - MMDB: Molecular Modeling Database
  - http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
- EBI-MSD (~PDB)
  - http://www.ebi.ac.uk/pdb/
- SCOP
  - hierarchical classification of 3-D structures
  - http://scop.mrc-lmb.cam.ac.uk/scop/
- CATH
  - classified protein domain structures
  - http://www.cathdb.info/

Gene ontology databank I.: GO

- The Gene Ontology Consortium
  - http://www.geneontology.org/
- Aim of standardizing the representation of gene and gene product attributes across species and databases.
- The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data
- 3 kind of approaches:
  - Molecular function (e.g. *chatalitic activity*)
  - Biological function (e.g. *pyrimidin metabolism*)
  - Cellular component (e.g. *endoplazmatic retiaulum*)
- Connecting projects e.g:
  - *FlyBase*
  - *Mouse Genome Database (MGD) and Gene Expression Database (GXD)*
  - *GOA: Gene Ontology Annotation @ EBI*
  - *Saccharomyces Genome Database (SGD)*
  - *The J. Craig Venter Institute (JCVI)*
  - *WormBase*
  - *etc...*

Gene ontology databank II: KEGG

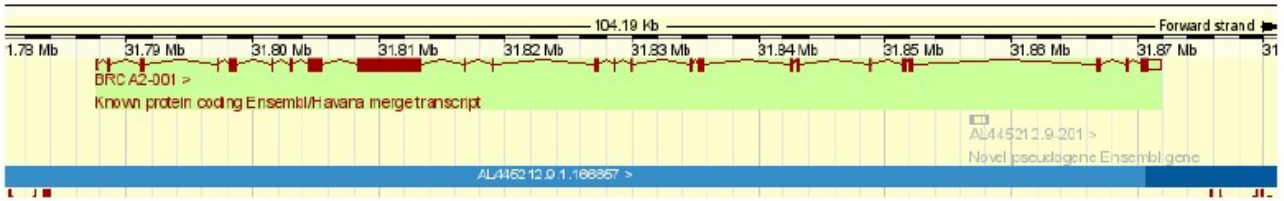
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- http://www.genome.jp/kegg/

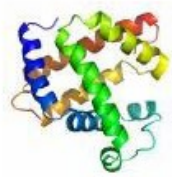


Genome browsers

- Ensembl:
  - Vertebrates and Eucaryotes
  - http://www.ensembl.org/index.html
  - By EBI and Sanger Institute
- NCBI Map Viewer:
  - http://www.ncbi.nlm.nih.gov/mapview/
- UCSC Genome Browser:
  - http://genome.ucsc.edu

Ensembl Genome Browser





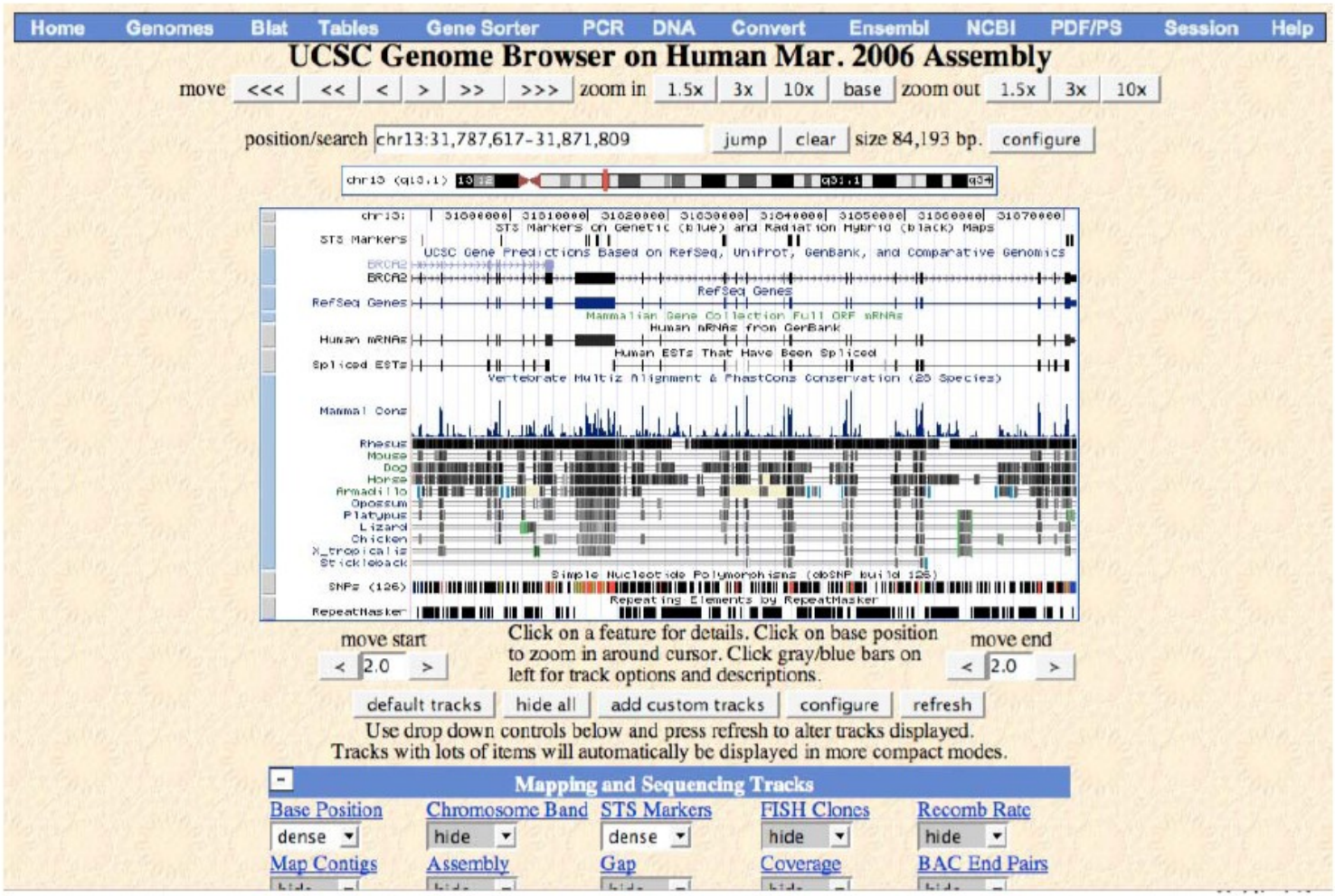
+

→ **Ensembl Genes**

Protein/ mRNA Sequence Assembly

NCBI Map Viewer

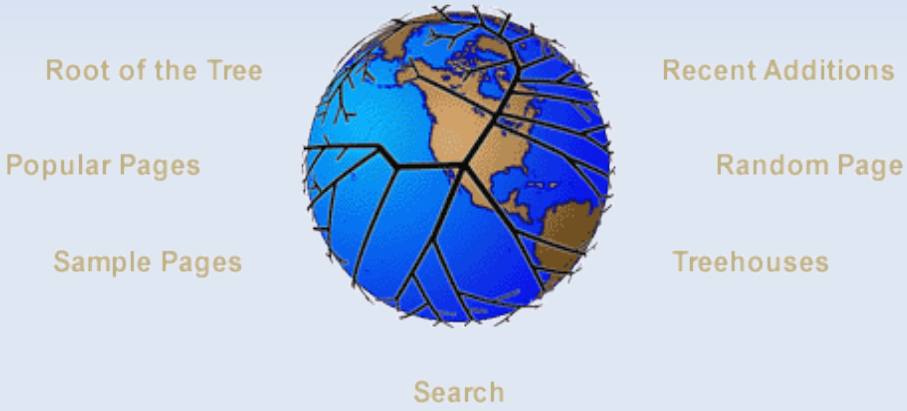
UCSC Genome Browser

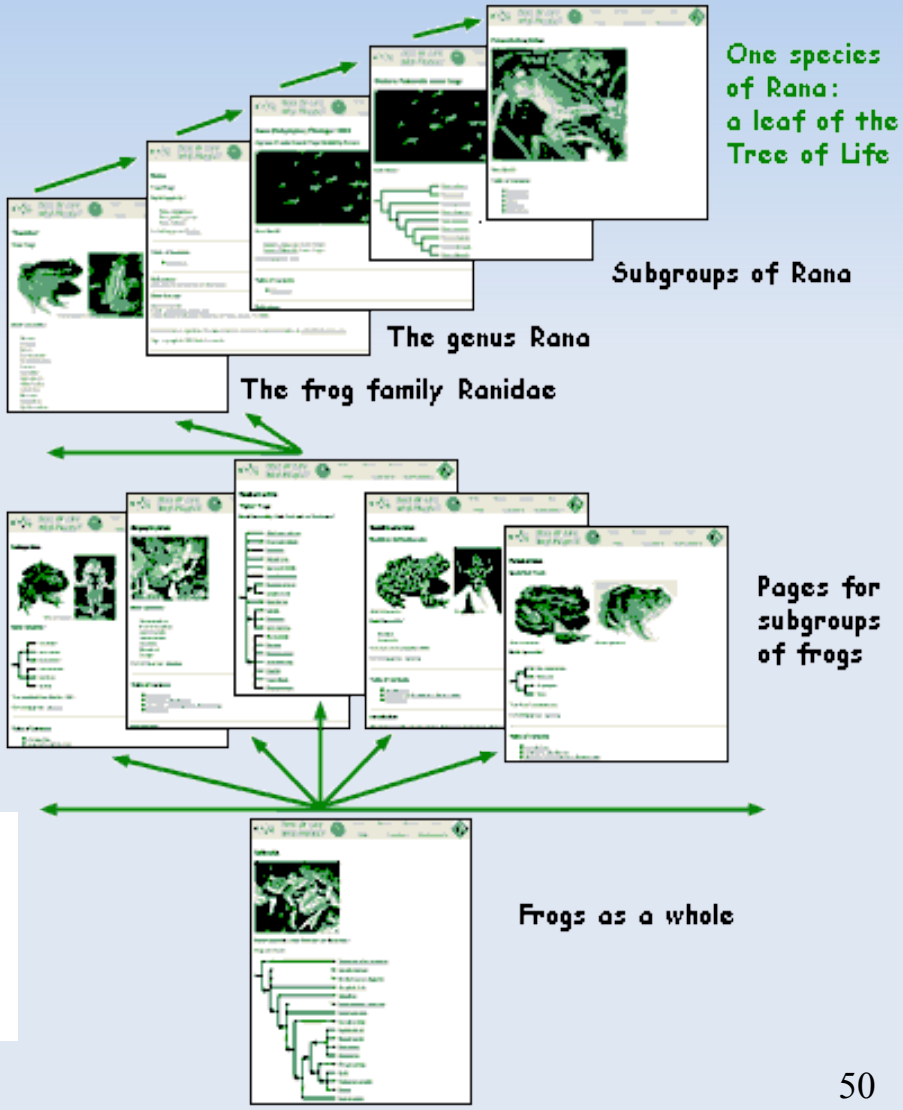


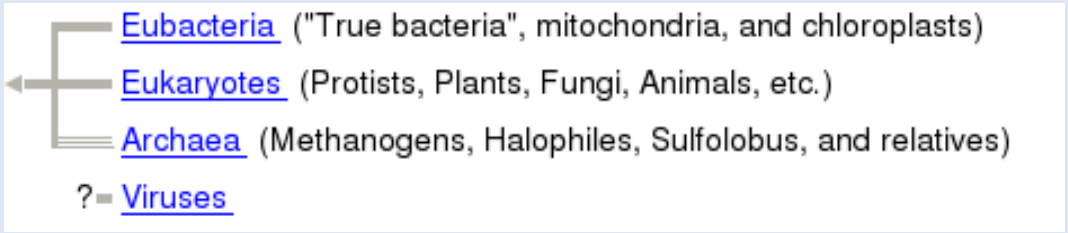
"Well, look who has stock in Genomes-R-Us."

Evolutionary databases I: Tree of Life

- A tree wich will contain all the species
- http://tolweb.org/tree/







TREE OF LIFE web project

Explore the Tree of Life

Browse the Site

Root of the Tree
Popular Pages
Sample Pages
Recent Additions
Random Page
Treehouses
Images, Movies,...

Search
advanced

Learn about ...

Crassigyrinus scoticus
(an extinct vertebrate)

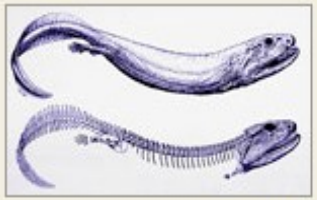


image info

Crassigyrinus scoticus is an aquatic stem-tetrapod from the Late Mississippian and Early Pennsylvanian (Visean and basal Namurian) of Scotland....

read more

previously featured pages

News

New article about the Tree of Life Web Project in **Zootaxa**...

read more

The Tree of Life Web Project (ToL) is a collaborative effort of biologists from around the world. On more than 9000 World Wide Web pages, the project provides information about the diversity of organisms on Earth, their evolutionary history (phylogeny), and characteristics.

Each page contains information about a particular group of organisms (e.g., echinoderms, tyrannosaurs, phlox flowers, cephalopods, club fungi, or the salamanderfish of Western Australia). ToL pages are linked one to another hierarchically, in the form of the evolutionary tree of life. Starting with the root of all Life on Earth and moving out along diverging branches to individual species, the structure of the ToL project thus illustrates the genetic connections between all living things.

read more about the Tree of Life Web Project...

Evolutionary databases II

- Treebase
  - Database of phylogenetic trees
  - http://www.treebase.org/
- Encyclopedia of Life (EOL)
  - http://www.eol.org/

Hylobates lar (Linnaeus, 1771)

Lar gibbon

Species recognized by The Integrated Taxonomic Information System, T Orrell (custodian) in Catalogue of Life

IUCN RED LIST STATUS: ENDANGERED (EN)

SWITCH TO COMMON NAMES

IMAGES | MAPS | COMMENTS

IMAGES

COPYRIGHT: Some rights reserved (cc) BY-NC-SA

SUPPLIER: Flickr flickr

AUTHOR: Robertsphotos1

Photographed at the Philadelphia Zoo - America's First Zoo

CLASSIFICATION : TEXT | GRAPHIC |

- Animalia +
  - Chordata +
    - Mammalia +
      - Primates +
        - Hylobatidae +
          - Hylobates +
            - *Hylobates lar* (Linnaeus, 1771)
- Archaea +
- Bacteria +
- Chromista +
- Fungi +
- Plantae +
- Protozoa +
- Viruses +

Authoritative INFORMATION All

TABLE OF CONTENTS

BEHAVIOR

SOURCE AND ADDITIONAL INFORMATION

AUTHOR Andrea Smith

COPYRIGHT ©1995-2008, The Regents of the University of Michigan and its licensors. Some rights reserved (cc) BY-NC-SA

SUPPLIER Animal Diversity Web ADW

SOURCE URL View original data object

These gibbons form small groups consisting of one mated pair and their offspring. Mated pairs tend to stay together in the same territory for their entire life-span, and they continue to have new young as mature offspring leave the group. There is some evidence of "divorces," where the male or female leaves his or her mate for no obvious reason and mates with another individual.

All gibbons are known to defend their territories from conspecifics using calls. These calls are usually very loud, and typically are duets, with both males and females calling.

CONTRIBUTE

This page has 1 curators.

Last curated: 18 Aug 2009

- Latest Changes
- Submit an image
- Submit text
- More information on how to help

EXPLORE

Pogonomyrmex bigbendensis Francke & Merickel, 1982

Veronica alpina L.
Alpine speedwell

Endothenia heinrichi McDunnough 1929

Useful links

- 2010 NAR Database Summary Papers: http://www3.oup.co.uk/nar/database/cap/
- EMBL, GenBank, DDJB Feature table definitions: http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html
- EBI: Introduction to Molecular Biology Databases: http://www.ebi.ac.uk/panda/Publications/mbd1.html
- The NCBI Handbook: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook
- NCBI Education: http://www.ncbi.nlm.nih.gov/Education/

Thank you for your attention!