Biological databases - overview
Aims

• To understand general types of data that are collected to support biomedical sciences
• To know the basic infrastructure of bioinformatics
• To know the sites and responsibilities of major service providers
• To explore the contents of databases using Web interfaces
What is a (biological) database?

- set of structured **entries**
- composed of **fields**
- flat files!
- needs an **unique identifier**
- **cross-references**
- **keywords**
- controlled vocabularies
- “semistructured content”
- open source!
Biological database types

- **archival**
  - primary repository
  - submitter owns the data
    - submitter is the only one who can make changes
    - tertiary annotations!

- **interpreting**
  - secondary
  - combine primary data, add to it
  - computed, curated
Database classes

- Nucleic acid sequences
- Amino acid sequences
- Proteomics resources
- Structure databases
- Metabolite databases
- Pathways
- Ontologies
- Textual databases (Literature,...)
What you will be doing

- "browsing" approach
  - fine for learning about things
- accessing directly
  - using programs to manipulate data
  - Web services
    - Taverna
    - BioMoby
  - API (Application Program Interface)
  - direct access to relational databases
Sequence Databases

Nucleotide Databases:

EMBL-bank: European Molecular Biology Laboratory, EBI
GenBank: National Center for Biotechnology Information (NCBI)
DDBJ: DNA Data Bank of Japan

Current Release: 49,498,755 entries

Submission obligatory before publication
Unannotated -> Preliminary -> Unreviewed -> Standard

Two or more identifies (ID, ACC, GI), Accession numbers are unique to each entry. One alphabetical character is followed by five digits, or two alphabetical characters are followed by six digits.
Database Growth

- Now > 50 million sequences in EMBL
- Exponential growth
  - fuelled by genome projects
- Most bioinformatics databases mirror this pattern of growth
EMBL entry for a sequence fragment implicated in Human Breast Cancer

ID       AY144588 standard; DNA; HUM; 68 BP.

AC       AY144588;

SV       AY144588.1

DT       23-SEP-2002 (Rel. 73, Created)

DT       23-SEP-2002 (Rel. 73, Last updated, Version 1)

DE       Homo sapiens truncated breast and ovarian cancer susceptibility protein

DE       (BRCA1) gene, partial cds.

KW       .

OS       Homo sapiens (human)

OC       Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;

OC       Eutheria; Primates; Catarrhini; Hominidae; Homo.
<table>
<thead>
<tr>
<th>RN</th>
<th>[1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP</td>
<td>1-68</td>
</tr>
<tr>
<td>RA</td>
<td>Rajkumar T., Soumittra N., Nirmala Nancy K., Shanta V.;</td>
</tr>
<tr>
<td>RT</td>
<td>&quot;Novel 5bp deletion in BRCA1 gene in South Indian family&quot;;</td>
</tr>
<tr>
<td>RL</td>
<td>Unpublished.</td>
</tr>
<tr>
<td>RN</td>
<td>[2]</td>
</tr>
<tr>
<td>RP</td>
<td>1-68</td>
</tr>
<tr>
<td>RA</td>
<td>Rajkumar T., Soumittra N., Nirmala Nancy K., Shanta V.;</td>
</tr>
<tr>
<td>RT</td>
<td>;</td>
</tr>
<tr>
<td>RL</td>
<td>Submitted (27-AUG-2002) to the EMBL/GenBank/DDBJ databases.</td>
</tr>
<tr>
<td>RL</td>
<td>Molecular Oncology, Cancer Institute (WIA), Canal Bank Road, Adyar, RL Chennai, TN 600020, India</td>
</tr>
</tbody>
</table>
FH Key Location/Qualifiers

FH

FT source 1..68

FT /country="India: South India"

FT /db_xref="taxon:9606"

FT /note="identical sequence found in daughter with breast cancer"

FT /sex="female"

FT /organism="Homo sapiens"

FT /isolation_source="mother with breast cancer"

FT /dev_stage="adult"

FT /mRNA 68

FT /gene="BRCA1"

FT /product="truncated breast and ovarian cancer susceptibility protein"
FT CDS <1..68
FT /codon_start=3
FT /note="contains premature stop codon due to frameshift caused by deletion"
FT /product="truncated breast and ovarian cancer susceptibility protein"
FT /protein_id="AAN10167.1"
FT /translation="EAASGCESETSVSEDCSGLSE"
FT exon 1..68
FT /number=12
FT /gene="BRCA1"
FT misc_feature 61..62
FT /note="site of deletion"
FT /gene="BRCA1"
SQ Sequence 68 BP; 19 A; 12 C; 23 G; 14 T; 0 other;
  gtgaagcagc atctgggtgt gagagtgaaa caagcgcttc tgaagactgc tcagggctat 60
cagagtga
  //
Sequence Databases

**Nucleotide Databases:**

**RefSeq:** Reference Sequence

Current Release: 93,285 entries

**NC_123456**

Complete Prokaryote Genome

Complete Eukaryote Chromosome

**NG_123456**

*Homo sapiens* Genomic Region

Not part of the sequence database collaboration!

**NM_123456**

mRNA of several organisms, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*

Those accession numbers beginning with X indicate model entries produced as a result of the Genome Annotation process.
Protein Databases

Uniprot/SwissProt:

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

Release 7.1 : 208,005 entries

UniProt/PIR:

Protein Information Resource

http://pir.georgetown.edu/

Entry names are often the name of the gene followed by the species. Accession numbers are of the following format:

[O,P,Q] [0-9] [A-Z, 0-9] [A-Z, 0-9] [A-Z, 0-9] [0-9],

e.g. P26367 (PAX6_HUMAN)
Sequence Databases

Protein Databases:

UniProt/TrEMBL:
Translated EMBL

Current Release: 1,748,002 entries

SpTrEMBL & RemTrEMBL

- Acts as a supplement to SwissProt and contains translated EMBL sequences with automatic annotation. TrEMBL entries are manually annotated before being entered into SwissProt.
- Remaining TrEMBL contains entries that will never be incorporated into SwissProt. These include: immunoglobulins; T-cell receptors; small fragments; synthetic sequences; CDS not coding for real proteins; patent application sequences.
- SwissProt TrEMBL contains entries which will eventually be integrated into the SwissProt database. SwissProt accession numbers have been assigned.
Swiss-Prot entry I

<table>
<thead>
<tr>
<th>ID</th>
<th>BPT2_BOVIN</th>
<th>STANDARD; PRT; 100 AA.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>P04815;</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>13-AUG-1987 (Rel. 05, Created)</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>01-MAR-1989 (Rel. 10, Last sequence update)</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>13-SEP-2005 (Rel. 48, Last annotation update)</td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>Spleen trypsin inhibitor I precursor (SI-I) [Contains: Spleen trypsin inhibitor II (SI-II); Spleen trypsin inhibitor III (SI-III)].</td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Bos taurus (Bovine).</td>
<td></td>
</tr>
<tr>
<td>OC</td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos.</td>
<td></td>
</tr>
<tr>
<td>OX</td>
<td>NCBI_TaxID=9913;</td>
<td></td>
</tr>
</tbody>
</table>
Swiss-Prot entry II

RN [1]  
RP NUCLEOTIDE SEQUENCE.  
RX MEDLINE=87283904; PubMed=2441071;  
RA Creighton T.E., Charles I.G.;  
RT "Sequences of the genes and polypeptide precursors for two bovine protease inhibitors.";  
RN [2]  
RP NUCLEOTIDE SEQUENCE.  
RX MEDLINE=88295740; PubMed=2456884;  
RA Creighton T.E., Charles I.G.;  
RT "Biosynthesis, processing, and evolution of bovine pancreatic trypsin inhibitor.";  
RN [3]  
RP NUCLEOTIDE SEQUENCE OF 34-97.  
RX MEDLINE=86158754; PubMed=2420326;  
RA Kingston I.B., Anderson S.;  
RT "Sequences encoding two trypsin inhibitors occur in strikingly similar genomic environments.";  
Swiss-Prot entry III

-**- SUBCELLULAR LOCATION: Secreted.

-**- SIMILARITY: Contains 1 BPTI/Kunitz inhibitor domain.

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 as long as its content is in no way modified and this statement is not removed.

EMBL; M20935; AAA51418.1; -; Genomic_DNA.
EMBL; M20931; AAA51418.1; JOINED; Genomic_DNA.
EMBL; M20933; AAA51418.1; JOINED; Genomic_DNA.
EMBL; X05275; CAA28887.1; -; mRNA.
EMBL; X06685; CAA29881.1; -; Genomic_DNA.
EMBL; X03366; CAA27064.1; ALT_SEQ; Genomic_DNA.
EMBL; X03366; CAA27065.1; -; Genomic_DNA.
PIR; S00274; TIBOSP.
HSSP; F00974; 1BPI.
SMR; P04815; 36-93.
InterPro; IPR002223; Prot_Inh_Kunz-m.
Pfam; PF00014; Kunitz_BPTI; 1.
PRINTS; PR00759; BASICPTASE.
ProDom; PD000222; Prot_Inh_Kunz-m; 1.
SMART; SM00131; KU; 1.
PROSITE; PS00280; BPTI_KUNITZ_1; 1.
PROSITE; PS50279; BPTI_KUNITZ_2; 1.
Swiss-Prot entry IV

KW Direct protein sequencing; Protease inhibitor;
KW Serine protease inhibitor; Signal.
FT SIGNAL 1 21 Potential.
FT PROPEP 22 33
FT CHAIN 34 99 Spleen trypsin inhibitor I.
FT CHAIN 36 97 Spleen trypsin inhibitor III.
FT CHAIN 36 93 Spleen trypsin inhibitor II.
FT PROPEP 100 100
FT DOMAIN 40 90 BPTI/Kunitz inhibitor.
FT SITE 50 51 Reactive bond (By similarity).
FT DISULFID 40 90 By similarity.
FT DISULFID 49 73 By similarity.
FT DISULFID 65 86 By similarity.
SQ SEQUENCE 100 AA; 10843 MW; 39069734B8ACF4E3 CRC64;
MKMSRLCLSI ALLVLLGTLa ASTPGCDTsn QAALKRPDFC LePPYTGPCK AKMIRYFYNA
KAGFCETFvY GGCKAKSNbF RSAEDQMRTC GGAIGPENL
//
Sequence Databases

Protein Databases:

RefSeqP: Reference Sequence Proteins

Current Release: 402,006 entries

Accession numbers for all proteins are of the format: NP_123456

RefSeqP provides a protein reference standard for the central dogma. It is used, as is RefSeq, to provide a foundation for the functional annotation of the human genome.
Sequence Version History

- Sequences change over time. What if I need a specific version?

- EMBL Sequence Version Archive
  - [http://www.ebi.ac.uk/cgi-bin/sva/sva.pl](http://www.ebi.ac.uk/cgi-bin/sva/sva.pl)

- GenBank Revision History

- SWISS-PROT
  - see the release number on the entry
Nucleotide Sequence Related DBs

- Rebase
- The Alternative Splicing Database (ASD)
- Protein Kinase Resource
- HIV protease Database
- IMGT, ImMunoGeneTics database
- KABAT, database of protein sequences of biological interest
Proteomics

- ENZYME
  - EC number (Enzyme commission)
- InterPro
  - umbrella db for PROSITE, Pfam, PRINTS, SMART, ProDom
- NRL_3D
  - sequences from Protein Data Bank
- ExPASy
  - Expert Protein Analysis System
Integrated Databases

These contain overview information garnered from a variety of different databases, and then offer links to further information.

GeneCards: http://bioinformatics.weizmann.ac.il/cards
http://www.nbn.ac.za/genecards/

An extremely thorough overview of a particular gene, with links to various other integrated and clinical databases.

Interpro: http://www.ebi.ac.uk/interpro

Integration of individual protein resources PRINTS; PROSITE; SMART; ProDom; Pfam; TIGRfam into one database. A search will scan entries of each and output results.
Genome Browsers

• Ensembl
  – http://ensembl.nbn.ac.za/
  – http://www.ensembl.org/

• UCSC
  – http://genome.ucsc.edu/

• NCBI

• model organism databases...
Integrated Databases

Ensembl: http://www.ensembl.org

A joint project by EBI and Sanger to annotate all the information currently known about the human genome in one larger database.

Browsing Contig Display

In addition to sequence displays a map of DNA fragments is shown giving the location of genes.

Each display is a magnified view of the red window in the display above.

- Landmark map markers
- Genes positions are shown under the map
- Use these buttons to move and resize your view
- Use these menus to reconfigure your view and access advanced features.
Structural Databases

*Tertiary protein structure prediction is possibly the Holy Grail of bioinformatics.*

Worldwide Protein Data Bank
http://www.wwpdb.org/

PDB: Protein DataBank, New Jersey, USA
Research Collaboratory for Structural Bioinformatics
http://www.rcsb.org/

Protein Data Bank Japan
MSD: EBI Macromolecular Structure Database
http://www.ebi.ac.uk/msd/index.html

Management and distribution of data on macromolecular structures in close collaboration with the PDB.

This houses a collection of 3D coordinates of each atom in a protein, allowing the structure to be displayed by viewing software. Protein structures are submitted by individual researchers and have been determined by x-ray diffraction, or NMR.

Currently 36,615 entries
Comparing Database Sizes

- DNA sequences far outnumber protein
- Human annotation is slow
- Solving protein structures is slower!
Structural Databases

**SCOP:** Structural Classification of Proteins

http://scop.mrc-lmb.cam.ac.uk/scop/

*Current Release: 686 folds; 1073 Superfamilies; 1827 Families representing 15,979 PDB entries*

**CATH:** Classification, Architecture, Topology, Homology

http://www.biochem.ucl.ac.uk/bsm/cath_new

*Current Release: 36,480 Domains*

Metabolite I

- Metabolite mass spectral database.
  - http://metlin.scripps.edu/

- The Human Metabolite Database
  - electronic repository for identification of small molecule metabolites.
  - http://hmdb.ca/
Metabolites II

• The Human Natural Products Database
  – formulas, exact masses and descriptions of endogenous metabolites.
  – http://massspec.scripps.edu/research/hnp/search.html

• The Golm Metabolome Database
  – mass spectra libraries, metabolite profiling experiments and others
  – http://csbldb.mpimp-golm.mpg.de/csbldb/gmd/gmd.html
Metabolomics III

• The Spectral Database for Organic Compounds SDBS
  – spectra of organic compounds (NMR, MS, IR).
  – http://www.aist.go.jp/RIODB/SDBS/cgi-bin/direct_f

• Chemical Entities of Biological Interest (ChEBI)
  – http://www.ebi.ac.uk/chebi/
Pathways

• Kyoto Encyclopedia of Genes and Genomes
  – pathway + genes + ligand + brite(ontology)

• Reactome
  – Curated database of biological processes in humans.
  – http://www.reactome.org/
  – evidence tracking!
BioModels

• The Systems Biology Markup Language (SBML)
  – http://www.ebi.ac.uk/biomodels/

• Members:
  – European Bioinformatics Institute (UK)
  – the SBML Team, including
    • Keck Graduate Institute (USA)
    • The Systems Biology Institute (Japan)
    • Stellenbosch University (South Africa)
Ontologies

• The National Center for Biomedical Ontology
  – http://bioontology.org/

• Open Biomedical Ontologies
  – http://obo.sourceforge.net/
  – Gene Ontology
    • http://geneontology.org/
  – Sequence Ontology
    • http://song.sourceforge.net/
  – eVOC
    • http://www.evoontology.org/
Bibliographic Databases

Used for searching for reference articles


Currently holds over 15 million MEDLINE entries.

Also, remember http://www.plos.org/

Online tools:
- connotea http://www.connotea.org/
- del.icio.us http://del.icio.us/
Clinical Databases

Human Gene Mutation Database, Cardiff, UK:
http://www.hgmd.org

Registers published mutations in the human genome and the diseases they cause.

dbSNP, Bethesda, USA:

HapMap
http://www.hapmap.org/

Online Mendelian Inheritance of Man, OMIM:
More

• 2can
  – Bioinformatics Educational Resource
  – http://www.ebi.ac.uk/2can/databases/

• Annual database issue of Nucleic Acid Research journal:
  – the first issue of the year
  – http://nar.oxfordjournals.org/content/vol34/suppl_1/

• http://www.bork.embl-heidelberg.de/Hotlist/
Searching the databases with a “search engine”:

The Sequence Retrieval System (SRS) from LION Bioscience AG is a very common search tool

The NCBI in the USA has its own search engine called Entrez.

Entrez is a retrieval system for searching several linked databases.

It provides access to:

- PubMed: The biomedical literature (PubMed)
- Nucleotide sequence database (Genbank)
- Protein sequence database
- Structure: three-dimensional macromolecular structures
- Genome: complete genome assemblies
- PopSet: Population study data sets
- Taxonomy: organisms in GenBank
- OMIM: Online Mendelian Inheritance in Man
Computers are THICK!

Database entries often presented as **flatfiles**

Each piece of information is on a separate line, distinguished by a code. Computers index this code, so they can search for the relevant entry.
Sequence Databases

Searching for a sequence:

Text Search: Use text with a boolean operator

BRCA1 & BRCA2 – searches for BRCA1 AND BRCA2

BRCA1 | BRCA2 – searches for one gene OR the other

BRCA1 ! BRCA2 – searches for BRCA1 BUT NOT BRCA2
To search for the BRCA1 gene in Homo sapiens in the EMBL database:

BRCA1 [DE] & Human [OC]