Module 1: Sequence databases
Nucleic acid and amino acid sequences

**DNA** = chain of nucleotides

4 nucleotides: **Adenine** – **Cytosine** – **Guanine** – **Thymine**

DNA can be written as a succession of letters (A,C,G,T)

By convention reported from 5' to 3' end

Has capacity to represent information

**Protein** = chain of amino acids

20 amino acids


By convention reported from the N- to the C-terminal end
mRNA: idem as DNA but using A, C, G, U
However, in records of sequence databases they write mRNA sequences using A, C, G, T!
Databases are used in biology to cope with data flood

Sequencing becomes cheaper and cheaper

https://www.genome.gov/sequencingcosts/

Cost per Genome

Moore's Law

NIH National Human Genome Research Institute
genome.gov/sequencingcosts
Databases are used in biology to cope with data flood

More and more data -> you need databases to handle these amounts of data

Databases store data in a structured way -> easy to retrieve data

Biological databases contain biological data

Genome databases:
- Ensembl
- UCSC

Protein databases:
- UniProt
- Protein

Sequence databases:
- Genbank
- RefSeq
- Unigene

Expression databases:
- GEO
- ArrayExpress

Structure databases:
- PDB
- SCOP

Interaction databases:
- Intact
- BIND
- DIP

Annotation databases:
- Gene
- Gene Ontology

Pathway databases:
- KEGG
- Reactome
Many biological databases exist

Nucleic Acids Research: annual database issue

list of biological databases:

http://www.oxfordjournals.org/our_journals/nar/database/a/
Sequence databases contain sequences

- **Genbank**
- **EMBL-bank**
- **DDBJ**
- **RefSeq**
- **Unigene**
- **Protein**
- **UniProt**

**Sequence databases**

- **all sequences**
- **clean non-redundant sequences**
- **mRNA**
- **protein**

**Organizations**

- **NCBI**
- **EBI**
- **NGI**
Many smaller more specialized sequence databases exist.

Genome sequence databases:
- Ensembl
- UCSC

Organism specific sequence databases:
- SGD
- TAIR
- FlyBase

Databases containing a specific type of sequences:
- Repbase: repeats
- MiRBase: miRNAs
- EPD: promoters
- UniVec: vectors
Databases containing all sequences

Sequence databases

all sequences

Genbank

Used to be collection of all available nucleotide and protein sequences

But NGS has turned this intention upside down

Free to use

EMBL-bank

High redundancy: contain many sequences for the same genomic region

No curation: nobody checks if submitted data is accurate and complete

DDBJ
NGS generates so many sequences that they are stored in separate databases.

Trace Archive:
- raw data from NGS experiments:
  - base calls + quality scores

SRA:
- sequences from NGS experiments

Daily exchange of data among the 3 primary sequence dbs

All three contain the same data

How do they differ?
organization of data
tools and database links
A sequence record contains two categories of information

1. **Info** about sequence, gene, components of the sequence, submitters and literature (annotations, upper part)

2. Actual **sequence** (data, at the bottom)

A sequence record is 'annotated' when biological info is added

Sequence annotations define components of the sequence (e.g. CDS, exons...) and are linked to a position in the sequence

Sequence annotations are also called 'features'

Annotation and feature names are indicated at the start of the lines in the record = keywords
Annotations are divided into different sections

Record divided into sections / subsections preceded by keywords

Homo sapiens PCCX1 mRNA for protein containing CXXC domain 1, complete cds

GenBank: AB031069.1
FASTA  Graphics

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>AB031069</th>
<th>2487 bp mRNA linear PRI 27-MAY-2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Homo sapiens PCCX1 mRNA for protein containing CXXC domain 1, complete cds.</td>
<td></td>
</tr>
<tr>
<td>ACCESSION</td>
<td>AB031069</td>
<td></td>
</tr>
<tr>
<td>VERSION</td>
<td>AB031069.1 GI:8100074</td>
<td></td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>Homo sapiens (human)</td>
<td></td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Homo sapiens</td>
<td></td>
</tr>
<tr>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontognires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFERENCE</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Fujino,T., Hasegawa,M., Shibata,S., Kishimoto,T., Imai,S. and Takano,T.</td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>PCCX1, a novel DNA-binding protein with PHD finger and CXXC domain, is regulated by proteolysis</td>
<td></td>
</tr>
<tr>
<td>PUBMED</td>
<td>10799292</td>
<td></td>
</tr>
<tr>
<td>REFERENCE</td>
<td>2 (bases 1 to 2487)</td>
<td></td>
</tr>
</tbody>
</table>

Sections / fields

Keywords

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Location/Qualifiers</td>
</tr>
<tr>
<td>gene</td>
<td>1..2487</td>
</tr>
<tr>
<td>CDS</td>
<td>229..2199</td>
</tr>
</tbody>
</table>

**Example annotations:**
- regulatory regions
- coding sequence
- exons and introns
- protein translation

**db_xref = cross references**
- links to records of other databases that contain info for this sequence
  Format `dbname:identifier`

**The actual sequence:**
```
1 agatggcgcg gctgagggggt cctgggggct ctaggcgcgc cactcactggt ttggcacat
```
What about the different IDs: GI numbers

**GI (GenInfo Identifier) numbers** - GI: 8100074

Series of digits consecutively assigned to each sequence record processed

No logical representation or implied meaning: just a serial number

Unique identifier for each record

Used independently by GenBank and ENA: neither uses others' numbering

What about the different IDs: accession numbers

**Accession Numbers** – AB031069

Represent and uniquely identify each sequence

Do not change, even if the info in a record is changed at the author's request

2 ≠ versions of same sequence have the same accession number

different GI numbers

Consistent across all databases at NCBI, ENA and DDBJ
ENA record of this mRNA

ID      AB031069; SV 1; linear; mRNA; STD; HUM; 2487 BP.
XX
AC      AB031069;
XX
DT      29-MAY-2000 (Rel. 63, Created)
DT      07-OCT-2008 (Rel. 97, Last updated, Version 2)
XX
DE      Homo sapiens PCCX1 mRNA for protein containing CXXC domain 1, complete cds.
XX
KW
XX
OS      Homo sapiens (human)
OC      Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC      Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC      Homo.
XX
RN      [1]
RP      1-2487
RA      Fujino T., Hasegawa M., Shibata S., Kishimoto T., Imai S., Takano T.;
RT      ;
RL      Submitted (15-AUG-1999) to the INSDC.
RL      Tadahiro Fujino, Keio University School of Medicine, Department of Microbiology; Shinanomachi 35, Shinjuku-ku, Tokyo 160-8582, Japan
RL      (E-mail:fujino@microb.med.keio.ac.jp, Tel:+81-3-3353-1211(ex.62692),
RL      Fax:+81-3-3360-1508)
XX
RN      [2]
RX      DOI: 10.1006/bbrc.2000.2614.
What about the different IDs: version numbers

Version numbers – AB031069.1

Represent and uniquely identify each correction of a sequence

Incremented every time sequence is changed e.g. corrected by author

Uses accession.version format

Implemented by GenBank/ENA/DDBJ

If change is made to sequence => it receives new GI number

Correct sequence and resubmit – Accession ? Version ? GI nr ?

Which is the best choice for tools that can use IDs as input ?

The use of accession numbers versus gene names

Genes have names / symbols e.g. BRCA2, PRKAR1A, DRP1, MAX2, BAI1...
Most genes have multiple symbols e.g. BDKRB1 aka BKR1, B1BKR, bradyb1

   HTR1D aka RDC4, HT1DA, 5-HT1D

One is chosen as the approved symbol but not everyone sticks to this
Many symbols change over time   e.g. HTR1D used to be HTRL

Tool to check symbols (human): http://www.genenames.org/cgi-bin/symbol_checker

<table>
<thead>
<tr>
<th>Input</th>
<th>Match type</th>
<th>Approved symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAI1</td>
<td>Previous symbol</td>
<td>ADGRB1</td>
</tr>
</tbody>
</table>

!! Excel messes up gene symbols and Riken identifiers (seq of mouse genome)

http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-80
e.g. DEC1  is transformed into 1 December

   2310009E13 is converted to floating point number: 2.31E+13
RefSeq is a clean high quality subset of Genbank

Collection of selected nucleotide and protein sequence records

No redundancy:
contains only one sequence for each genomic region

Curation:
people@NCBI select/create most complete and accurate data from Genbank
but large-scale sequencing/gene prediction turned this intention upside-down

Validates RefSeq records are curated

RefSeq chapter in the NCBI handbook:
Model RefSeq sequences are not curated
Main differences between Genbank and RefSeq

- **Primary sequence database Genbank**
  - **Only submitter** can make corrections
  - **No filtering**: all sequences
  - **Redundancy** (= sequence, ≠ records)

- **Cleaned / Derived database RefSeq**
  - **Provider** can derive/correct data
  - **Filtering on organisms or quality**
  - **No redundancy**

The sequence records look the same
Data retrieval from NCBI can be done in various ways


On home page select database, leave search term box empty and click Search


On home page select All Databases, leave search term box empty and click Search

Searching nucleotide sequences on the NCBI website

**Step 1:** Select database you want to search in

**Step 2:** Provide terms to define what you search for

You don’t see Genbank in this list!
DNA and RNA sequences from Genbank and Refseq are combined into one large sequence repository called ‘Nucleotide’
DNA/RNA sequences are divided into 3 databases

**CoreNucleotide = Nucleotide**
DNA / mRNA / genome sequences but no ESTs and GSSs: highest quality
Data from Genbank / RefSeq / other databases

**dbEST = EST**
EST sequences from Genbank: variable quality

**dbGSS = GSS**
Genomic Survey Sequences from Genbank: variable quality
Genomic counterpart of EST: short sequences resulting from sequencing genome

You can search each of them separately
Which nucleotide database do you choose to search in?

 Doesn’t matter much:

 searching any of the 3 databases will provide links to results in the others

 Unless you know you want to find a specific EST or GSS sequence

 Searching Nucleotide with a text query will produce the results in Nucleotide

 But you can follow links to results in EST and GSS

 Nucleotide database search results:

 See CXXC1 (PCCX1) CXXC finger protein 1 in the Gene database
 pccx1 reference sequences: Transcript (4) Protein (4)

 Items: 14

 Found 13253337 nucleotide sequences. Nucleotide (14) EST (10464181) GSS (2789142)
Use Send to download selected records from the list of results

Coding Sequences: sequences of features annotated as CDS
Gene Features: sequences of features annotated as gene

Destination
File: Download
Clipboard: Copy (you have to paste it in a file on your pc)
Collections: only if you have a NCBI account
Analysis tool: BLAST / Primer-BLAST

Format
FASTA, Accession list:
standards for bioinformatics tools (see next slides)

GFF3: standard for visualization & mapping RNASeq reads

ASN.1, XML, INSDSeq XML, TinySeq XML:
for exchange between databases not for users
FASTA is a standard format for representing sequences

Ideal editor: WordPad / NotePad / TextEdit

Save as .txt

Use single letter codes

Description line:  >sequenceID (+ description)

↑ no space here

not too long (< 80 chars)
Accession list generates a list of accession numbers

AB031069.1
NC_000018.10
NT_010966.15
NC_018929.2
BC014940.2
BC015733.1
BC029922.1
NM_001101654.1
NM_014593.3
NG_029505.1
NW_004929410.1
XM_011525941.1
XM_011525940.1
AB031230.1

Most sequence analysis tools accept accession number lists as input
(instead of complete sequences)
GFF3 is a standard format for sequence annotations

<table>
<thead>
<tr>
<th>Seq ID</th>
<th>source</th>
<th>type</th>
<th>start</th>
<th>end</th>
<th>score</th>
<th>strand</th>
<th>phase</th>
<th>info</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>region</td>
<td>1</td>
<td>2487</td>
<td>.</td>
<td>+</td>
<td></td>
<td>ID=id0;</td>
</tr>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>gene</td>
<td>1</td>
<td>2487</td>
<td>+</td>
<td>.</td>
<td>ID=gene0; Name=PCCX1;</td>
<td></td>
</tr>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>CDS</td>
<td>229</td>
<td>2199</td>
<td>+</td>
<td>0</td>
<td>ID=cds0; Parent=gene0,</td>
<td></td>
</tr>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>sequence_alteration</td>
<td>863</td>
<td>863</td>
<td>.</td>
<td>+</td>
<td></td>
<td>ID=id1</td>
</tr>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>sequence_alteration</td>
<td>1132</td>
<td>1132</td>
<td>.</td>
<td>+</td>
<td></td>
<td>ID=id2</td>
</tr>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>sequence_alteration</td>
<td>2003</td>
<td>2003</td>
<td>.</td>
<td>-</td>
<td></td>
<td>ID=id3</td>
</tr>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>sequence_alteration</td>
<td>2222</td>
<td>2222</td>
<td>.</td>
<td>+</td>
<td></td>
<td>ID=id4</td>
</tr>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>sequence_alteration</td>
<td>2446</td>
<td>2446</td>
<td>.</td>
<td>+</td>
<td></td>
<td>ID=id5</td>
</tr>
<tr>
<td>AB031069.1</td>
<td>DDBJ</td>
<td>sequence_alteration</td>
<td>2447</td>
<td>2447</td>
<td>.</td>
<td>+</td>
<td></td>
<td>ID=id6</td>
</tr>
</tbody>
</table>

Score: only for predicted features – no score = .

Phase: reading frame, only for CDS - no phase = .
You can see your search query in the search details box

Search Homo sapiens in the Organism field or anywhere in the records

Search pccx1 anywhere in the records
Words between [] correspond to keywords that marks fields in records

**Search details**

```
("Homo sapiens" OR human) AND pccx1
```

[Organism] means that the preceding search term: “Homo sapiens” is searched only in the Organism field of the Genbank records

So returns only human records (for which Organism = Homo sapiens)

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>AB031069</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Homo sapiens PCCX1 mRNA for protein containing CXXC domain 1, complete cds.</td>
</tr>
<tr>
<td>ACCESSION</td>
<td>AB031069</td>
</tr>
<tr>
<td>VERSION</td>
<td>AB031069.1 GI:8100074</td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>.</td>
</tr>
<tr>
<td>SOURCE</td>
<td>Homo sapiens (human)</td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Homo sapiens</td>
</tr>
</tbody>
</table>

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
The query generates a massive amount of hits in EST and GSS

Why so many hits here?

See CXXC1 (PCCX1) CXXC finger protein 1 in the Gene database
pccx1 reference sequences Transcript (4) Protein (4)

Items: 14

Found 13253337 nucleotide sequences. Nucleotide (14)

1. Homo sapiens PCCX1 mRNA for protein containing CXXC domain 1. complete cds
   2,487 bp linear mRNA
   Accession: AB031069.1 GI: 8100074
   GenBank FASTA Graphics

2. Homo sapiens chromosome 18, GRCh38.p2 Primary Assembly
   80,373,285 bp linear DNA
   Accession: NC_000018.10 GI: 568815580
   GenBank FASTA Graphics
The query used on the EST and GSS database is not correct

**PCCX1 is not found in any record of these databases**

⇒ **PCCX1 is dropped as a search term**

⇒ **a search is done using only Homo sapiens as a search term**

All records that mention “Homo sapiens” are retrieved

---

**Search details**

"Homo sapiens"[Organism] OR homo sapiens[All Fields]

---

**Solved by using Boolean operators**
Boolean operators are used in searches

**AND**
Narrow search
Retrieve records containing each term
e.g. mouse AND pseudogene
Retrieve records that contain both words

**NOT**
Narrow search
Retrieve records that do not contain the term
e.g. pseudogene NOT mouse
retrieve records that contain pseudogene but not mouse

**OR**
Broaden search
Retrieve records containing any of the terms
e.g. mouse OR rat
 retrieve records that contain mouse add records that contain rat

Boolean operators avoid problem of unfound terms

No hits on EST and GSS
Boolean operators are processed from left to right

Search records that contain “BRCA2”

From this list remove the records that contain “human”

Then add records that contain “mouse”

Why so many hits?
Use parentheses to change the priority of the operators

Search records that contain “BRCA2”

From this list remove the records that contain either “human” or “mouse”
Use wild cards to represent variable characters

1. **LP05734.5prime LP Drosophila melanogaster larval-early pupal pOT2 Drosophila melanogaster cDNA clone LP05734 5 similar to J01098: D melanogaster locus 67B: heat shock protein hsp22 gene., mRNA sequence**
   - 552 bp linear mRNA
   - Accession: At257387.1 GI: 3864912
   - EST  GenBank  FASTA

   - 521 bp linear mRNA
   - Accession: AA263486.1 GI: 1899680
   - EST  GenBank  FASTA

   - 582 bp linear mRNA
Use quotes to search for expressions

Search records that contain both “insulin” and “dependent”
Use quotes to search for expressions

Search records that contains the full expression “insulin dependent”
If your search still returns too many records you can

Use an **Advanced Search** to refine your search

**Filter** the search results

Advanced searches allow to specify the field to search in

Term to search for: "insulin dependent" AND homo sapiens[Organism]

Fields to search in:
- All Fields
- Accession
- Author
- BioProject
- Component Accession
- Gene
- Gene Name
- Issue
- Journal
- Keyword
- Modification Date
- Organism
- Page Number
- Primary Accession
- Primary Organism
- Properties
- Protein Name
- Publication Date

Search history: overview of past searches:

<table>
<thead>
<tr>
<th>Query</th>
<th>Items found</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>homo sapiens AND BRCA2</td>
<td>395</td>
<td>06:39:49</td>
</tr>
</tbody>
</table>

operator
Filters allow to select certain categories of results.
Filters stay active until you remove them!!

Show additional filters

Clear all

Species
Animals
More ...

Molecule types
genomic DNA/RNA
mRNA
More ...

Source databases
GenBank
RefSeq
More ...

Remove filter by clicking clear or by ticking it off
NCBI search tutorials

Entrez Help:
http://www.ncbi.nlm.nih.gov/books/NBK3837/

Entrez Sequences Help:
http://www.ncbi.nlm.nih.gov/books/NBK44864/
http://www.ncbi.nlm.nih.gov/books/NBK44863/

Entrez Nucleotide FAQ:
http://www.ncbi.nlm.nih.gov/books/NBK49541/

Search field descriptions by database:
http://www.ncbi.nlm.nih.gov/books/NBK49540/

The tutorials are sometimes a bit outdated!
NCBI search tutorials

Dates and Other Ranges

Certain fields can accept ranges of values. Common examples are Publication Date, Modification Date, Accession, Molecular Weight, and Sequence Length. In these cases the low and high numbers of the range are entered with a colon ":" as the range operator between them followed by the field:

110:500[Sequence Length]
2009/3/1:2009/9/30[Publication Date]
You can change the information shown in sequence records.

If you use Send you will download the updated information.
You can display a part of the sequence

If you use Send you will download the partial sequence
Refseq: data retrieval

Same format as GenBank records

Searching ‘Nucleotide’ searches both Genbank and RefSeq but

- Users can filter the RefSeq records

- RefSeq accession numbers always contain an underscore

  Genbank nucleotide accession numbers never contain an underscore
After a nucleotide search you can filter the RefSeq records

- **Homo sapiens breast cancer 2, early onset (BRCA2), mRNA**
  - Accession: NM_000059.3  GI: 119395733
  - GenBank  FASTA  Graphics  Related Sequences

- **Homo sapiens BRCA2 and CDKN1A interacting protein (BCCIP), transcript variant C, mRNA**
  - Accession: NM_078469.2  GI: 169790846
  - GenBank  FASTA  Graphics  Related Sequences
<table>
<thead>
<tr>
<th>Accession No.</th>
<th>Molecule</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>X999999</td>
<td>Nucleotide</td>
<td>1 Alpha plus 5 Numeric: Oldest GenBank originated nucleotide records with original accession numbering scheme.</td>
</tr>
<tr>
<td>XX999999</td>
<td>Nucleotide</td>
<td>2 Alpha plus 6 Numeric: GenBank originated nucleotide records entered since change in February 1999.</td>
</tr>
<tr>
<td>XXX999999</td>
<td>Protein</td>
<td>3 Alpha plus 5 Numeric: GenBank originated protein translations of nucleotide records.</td>
</tr>
<tr>
<td>A_B</td>
<td>Protein</td>
<td>Up to 5 alpha plus underscore plus up to 5 alpha: Not strictly an accession, but rather a SwissProt originated protein &quot;entry name&quot;, where A is mnemonic code for protein name and B is mnemonic code for species. NCBI Entrez will return results for this ID.</td>
</tr>
<tr>
<td>Ex: PSBP_CUCSA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A_B</td>
<td>Protein</td>
<td>Up to 12 upper case alphanumeric with underscore separator: Not strictly an accession, but rather a UniProt/TrEMBL originated protein &quot;entry name&quot;, where A is the accession number and B is mnemonic code for species. NOTE: NCBI Entrez will not return results for this format. You must remove the '_B' portion (underscore and species name) and use only the accession.</td>
</tr>
<tr>
<td>Ex: O95417_HUMAN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXXXXX</td>
<td>Protein</td>
<td>Up to 5 alpha: A shortened version of above originating with one of the protein databases SwissProt/UniPort/pir. Usually also listed as the locus.</td>
</tr>
<tr>
<td>Accession No.</td>
<td>Molecule</td>
<td>Method*</td>
</tr>
<tr>
<td>---------------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>AC_123456</td>
<td>Genomic</td>
<td>Mixed</td>
</tr>
<tr>
<td>AP_123456</td>
<td>Protein</td>
<td>Mixed</td>
</tr>
<tr>
<td>NC_123456</td>
<td>Genomic</td>
<td>Mixed</td>
</tr>
<tr>
<td>NG_123456</td>
<td>Genomic</td>
<td>Mixed</td>
</tr>
<tr>
<td>NM_123456</td>
<td>mRNA</td>
<td>Mixed</td>
</tr>
<tr>
<td>NM_123456789</td>
<td>mRNA</td>
<td>Mixed</td>
</tr>
</tbody>
</table>
NCBI Protein database summarizes protein info

Collection of protein sequences

Translations from annotated coding regions in GenBank, RefSeq and TPA
Protein sequence records from UniProt, PIR, PRF and PDB

Quality ?
Redundancy ?

High quality protein databases will be discussed later
Access Protein database via the NCBI website

Search in exactly the same way as in the Nucleotide database
# Searches on all NCBI databases using GQuery

Integrated, text-based search and retrieval system used for NCBI databases.

NCBI databases are all connected: allows cross database searches.

GQuery allows you to search all NCBI databases simultaneously.

## List of databases that are searched

<table>
<thead>
<tr>
<th>Literature</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>expressed sequence tag sequences</td>
</tr>
<tr>
<td>MeSH</td>
<td>collected information about gene loci</td>
</tr>
<tr>
<td>NLM Catalog</td>
<td>functional genomics studies</td>
</tr>
<tr>
<td>PubMed</td>
<td>gene expression and molecular abundance profiles</td>
</tr>
<tr>
<td>PubMed Central</td>
<td>homologous gene sets for selected organisms</td>
</tr>
<tr>
<td>Health</td>
<td>sequence sets from phylogenetic and population studies</td>
</tr>
<tr>
<td>ClinVar</td>
<td>clusters of expressed transcripts</td>
</tr>
<tr>
<td>dbGaP</td>
<td></td>
</tr>
<tr>
<td>GTR</td>
<td></td>
</tr>
<tr>
<td>MedGen</td>
<td></td>
</tr>
<tr>
<td>OMIM</td>
<td></td>
</tr>
<tr>
<td>PubMed Health</td>
<td></td>
</tr>
<tr>
<td>Genomes</td>
<td></td>
</tr>
<tr>
<td>Assembly</td>
<td>conserved protein domains</td>
</tr>
<tr>
<td></td>
<td>protein sequences</td>
</tr>
<tr>
<td></td>
<td>sequence similarity-based protein clusters</td>
</tr>
<tr>
<td></td>
<td>experimentally-determined biomolecular structures</td>
</tr>
</tbody>
</table>

### Literature

- **Books**: books and reports
- **MeSH**: ontology used for PubMed indexing
- **NLM Catalog**: books, journals and more in the NLM Collections
- **PubMed**: scientific & medical abstracts/citations
- **PubMed Central**: full-text journal articles

### Health

- **ClinVar**: human variations of clinical significance
- **dbGaP**: genotype/phenotype interaction studies
- **GTR**: genetic testing registry
- **MedGen**: medical genetics literature and links
- **OMIM**: online mendelian inheritance in man
- **PubMed Health**: clinical effectiveness, disease and drug reports

### Genes

- **EST**: gene expression and molecular abundance profiles
- **Gene**: homologous gene sets for selected organisms
- **GEO DataSets**: sequence sets from phylogenetic and population studies
- **GEO Profiles**: clusters of expressed transcripts

### Proteins

- **Conserved Domains**: conserved protein domains
- **Protein**: protein sequences
- **Protein Clusters**: sequence similarity-based protein clusters
- **Structure**: experimentally-determined biomolecular structures

### Genomes

- **Assembly**: genomic assembly information
Searches on GQuery return a summary page containing the number of records returned by each database.

### Search NCBI databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>0</td>
</tr>
<tr>
<td>MeSH</td>
<td>0</td>
</tr>
<tr>
<td>NLM Catalog</td>
<td>0</td>
</tr>
<tr>
<td>PubMed</td>
<td>1</td>
</tr>
<tr>
<td>PubMed Central</td>
<td>1</td>
</tr>
<tr>
<td>ClinVar</td>
<td>0</td>
</tr>
<tr>
<td>dbGaP</td>
<td>0</td>
</tr>
<tr>
<td>GTR</td>
<td>0</td>
</tr>
<tr>
<td>MedGen</td>
<td>0</td>
</tr>
<tr>
<td>OMIM</td>
<td>0</td>
</tr>
<tr>
<td>PubMed Health</td>
<td>0</td>
</tr>
<tr>
<td>Assembly</td>
<td>0</td>
</tr>
<tr>
<td>BioProject</td>
<td>0</td>
</tr>
<tr>
<td>BioSample</td>
<td>0</td>
</tr>
<tr>
<td>Clone</td>
<td>0</td>
</tr>
<tr>
<td>dbVar</td>
<td>2</td>
</tr>
<tr>
<td>Epigenomics</td>
<td>0</td>
</tr>
<tr>
<td>Genome</td>
<td>0</td>
</tr>
<tr>
<td>EST</td>
<td>0</td>
</tr>
<tr>
<td>Gene</td>
<td>2</td>
</tr>
<tr>
<td>GEO DataSets</td>
<td>0</td>
</tr>
<tr>
<td>GEO Profiles</td>
<td>2,007</td>
</tr>
<tr>
<td>HomoloGene</td>
<td>1</td>
</tr>
<tr>
<td>PopSet</td>
<td>0</td>
</tr>
<tr>
<td>UniGene</td>
<td>1</td>
</tr>
<tr>
<td>Conserved Domains</td>
<td>0</td>
</tr>
<tr>
<td>Protein</td>
<td>9</td>
</tr>
<tr>
<td>Protein Clusters</td>
<td>0</td>
</tr>
<tr>
<td>Structure</td>
<td>6</td>
</tr>
<tr>
<td>BioSystems</td>
<td>0</td>
</tr>
<tr>
<td>PubChem BioAssay</td>
<td>4</td>
</tr>
<tr>
<td>PubChem Compound</td>
<td>0</td>
</tr>
<tr>
<td>PubChem Substance</td>
<td>0</td>
</tr>
</tbody>
</table>

**About 2,350 search results for "pccx1 AND homo sapiens"**
Using the correct syntax for GQueries is crucial

**AND, OR, NOT**  Boolean operators to combine search terms

**()**  Parentheses to determine the order of the search terms

**[*]**  Wild cards to represent 1/more variable characters

**"**  Quotes to search for expressions

**[Field name]**  Square brackets to limit the search to a certain field in the records
You can find fields via Search details

Query Translation:

```
```

**Sequence length range**

**mRNA sequences only**

**Release date range**
Ensembl is a eukaryotic genome sequence database

Only completely sequenced and assembled genomes of eukaryotic model organisms

Genomes come from sequencing centers

Ensembl does the annotation = location of features on the genome

Video tutorial: https://www.ensembl.org/Multi/Help/Movie?db=core;id=188
Gene annotation is done by aligning mRNAs to genome

Gene based on all transcripts with (partly) overlapping CDS

If CDS annotation of Ensembl and HAVANA agrees -> golden transcripts
If CDS annotation of Ensembl/HAVANA/UCSC/NCBI agrees -> CCDS
Gene is the key concept in the Ensembl database

Gene encodes transcripts encode proteins

ENSG... ENST... ENSP...

Gene page Transcript pages on a Transcript page you can ask for protein information

Genes are located on genome and have a sequence

Location page Sequences can vary between individuals

Variation page
### Flags in transcript table represent evidence for splice variant

<table>
<thead>
<tr>
<th>Name</th>
<th>Transcript ID</th>
<th>bp</th>
<th>Protein</th>
<th>Biotype</th>
<th>CCDS</th>
<th>RefSeq</th>
<th>Flags</th>
</tr>
</thead>
<tbody>
<tr>
<td>CXXC1-001</td>
<td>ENST00000285106</td>
<td>2936</td>
<td>656 aa</td>
<td>Protein coding</td>
<td>CCDS11945</td>
<td>NM_014593</td>
<td>TSL1, GENCODE basic, APPRIS CI</td>
</tr>
<tr>
<td>CXXC1-002</td>
<td>ENST00000412036</td>
<td>2280</td>
<td>660 aa</td>
<td>Protein coding</td>
<td>CCDS45866</td>
<td>NM_001101654</td>
<td>TSL1, GENCODE basic, APPRIS CI2</td>
</tr>
<tr>
<td>CXXC1-004</td>
<td>ENST00000589940</td>
<td>2242</td>
<td>613 aa</td>
<td>Protein coding</td>
<td>-</td>
<td>-</td>
<td>TSL1, GENCODE basic</td>
</tr>
<tr>
<td>CXXC1-018</td>
<td>ENST00000589548</td>
<td>849</td>
<td>253 aa</td>
<td>Protein coding</td>
<td>-</td>
<td>-</td>
<td>TSL3, CDS 3' incomplete</td>
</tr>
<tr>
<td>CXXC1-017</td>
<td>ENST00000591474</td>
<td>845</td>
<td>250 aa</td>
<td>Protein coding</td>
<td>-</td>
<td>-</td>
<td>TSL3, CDS 3' incomplete</td>
</tr>
<tr>
<td>CXXC1-019</td>
<td>ENST00000588837</td>
<td>838</td>
<td>231 aa</td>
<td>Protein coding</td>
<td>-</td>
<td>-</td>
<td>TSL5, CDS 3' incomplete</td>
</tr>
<tr>
<td>CXXC1-008</td>
<td>ENST00000587396</td>
<td>476</td>
<td>No protein</td>
<td>Processed transcript</td>
<td>-</td>
<td>-</td>
<td>TSL2</td>
</tr>
<tr>
<td>CXXC1-006</td>
<td>ENST00000590001</td>
<td>2562</td>
<td>No protein</td>
<td>Retained intron</td>
<td>-</td>
<td>-</td>
<td>TSL2</td>
</tr>
</tbody>
</table>

**TSL1**: all splice junctions supported by at least 1 high quality mRNA

**TSL2**: transcript annotation supported by multiple ESTs / 1 low quality mRNA

**TSL3**: transcript annotation supported by 1 EST

**TSL4**: transcript annotation supported by 1 low quality EST

**TSL5**: transcript annotation not supported by Ensembl (but by UCSC)

*Transcript support levels: [http://www.ensembl.org/Help/Glossary?id=492](http://www.ensembl.org/Help/Glossary?id=492)*
Changing the appearance of graphical displays

You can change how the sequence is displayed
Works for any kind of graph on any page in Ensembl
Navigation in the most detailed display of the Location tab

- **Change location**

- **Zoom and scroll region**

- **Only region of gene itself is shown**

- **All transcripts of gene of interest**
Navigation in the most detailed display

Clicking a feature opens a box containing more info

Clicking outside features allows to adjust zoom/center level
Export sequences

Choose output format

Select the sequence you want to download
Species in Ensembl

http://www.ensembl.org

No plants nor fungi nor bacteria!! These are stored on http://ensemblgenomes.org

Genomes that are not yet fully annotated are available at http://pre.ensembl.org/index.html
Ensemblgenomes has a similar interface
Gene pages in Ensemblgenomes are similar to these in Ensembl.

Gene: PDF1.2 AT5G44420

Description: plant defensin 1.2
Source: TAIR LOCUS AT5G44420

Location: Chromosome 5: 17,907,069-17,907,598 reverse strand.

Transcripts:
This gene has 1 transcript (splice variant)

<table>
<thead>
<tr>
<th>Name</th>
<th>Transcript ID</th>
<th>Length (bp)</th>
<th>Protein ID</th>
<th>Length (aa)</th>
<th>Biotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF1.2-201</td>
<td>AT5G44420.1</td>
<td>438</td>
<td>AT5G44420.1</td>
<td>80</td>
<td>Protein coding</td>
</tr>
</tbody>
</table>

Summary

Name: PDF1.2 (TAIR Gene Name)

Synonyms: LCR77, PDF1.2A

Gene type: Protein coding

Prediction Method:
Gene annotation by TAIR through a process of automatic and manual curation.

Annotation is adopted from TAIR

Not matched with annotation from other sources: no CCDS, no golden transcripts.
Help on Ensembl and EnsemblGenomes

Ensembl

Documentation: http://www.ensembl.org/info/

Video tutorials: http://www.ensembl.org/info/website/tutorials/

FAQ: http://www.ensembl.org/Help/Faq

Publications: http://www.ensembl.org/info/about/publications.html

Blog: http://www.ensembl.info/

Ensembl genomes

EnsemblPlants Documentation: http://plants.ensembl.org/info/

EnsemblBacteria Video tutorials: http://www.ensembl.org/Multi/Help/Movie?db=core;id=210

FAQ: http://ensemblgenomes.org/info/faqs

Publications: http://ensemblgenomes.org/info/publications

Questions: helpdesk@ensembl.org
Alternative plant sequence databases

Phytozome: http://www.phytozome.net/

PlantGDB: http://www.plantgdb.org/

PLAZA: http://bioinformatics.psb.ugent.be/plaza/
   compares plant genomes to improve annotation
Specialized sequence databases:

Uniprot contains a high quality protein sequence database

http://www.uniprot.org

Created by EBI + Swiss IB + Georgetown University

Focus on biological function of proteins derived from literature

Only contains protein sequences, no nucleotide sequences

Divided into two separate sections:

Swiss-Prot: manually annotated and curated records with added info

e.g. binding sites for drugs...

TrEMBL: computationally derived records that await annotation and curation
NCBIs protein sequence databases

**Genbank protein**: translations of all CDS annotated in Genbank
so all redundancy / errors of Genbank nucleotide -> Genbank protein corresponds to UniProt TrEMBL

**Refseq protein**: each transcript in RefSeq is linked to a protein
no redundancy / less errors than Genbank protein
only for major organisms for which sufficient info is available
not as heavily curated as Swiss-Prot

So quality of information in Swiss-Prot is much higher than in Genbank protein
But Genbank protein contains a lot more sequences than Swiss-Prot

First search Swiss-Prot
No results => search Genbank protein / TrEMBL

Checking the quality of a UniProt record

<table>
<thead>
<tr>
<th>Entry</th>
<th>Entry name</th>
<th>Protein names</th>
<th>Gene names</th>
<th>Organism</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q6GZR1</td>
<td>064R_FRG3G</td>
<td>Caspase recruitment domain-</td>
<td>FV3-064R</td>
<td>Frog virus 3 (isolate Goorha) (FV-3)</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>containi...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A0A009FA48</td>
<td>A0A009FA48_ACIBA</td>
<td>Caspase domain protein</td>
<td>J521_3085</td>
<td>Acinetobacter baumannii 1035119</td>
<td>497</td>
</tr>
</tbody>
</table>

**Swiss Prot: curated**

**TrEMBL: not curated**

Video tutorial on UniProt: [https://www.youtube.com/watch?v=9IYl4QDVPa0](https://www.youtube.com/watch?v=9IYl4QDVPa0)

Video tutorial on UniProtKB: [https://www.youtube.com/watch?v=ado1r8IDm3U](https://www.youtube.com/watch?v=ado1r8IDm3U)

UniProt text search help: [http://www.uniprot.org/help/text-search](http://www.uniprot.org/help/text-search)

Boolean operators in UniProt

If you type multiple keywords AND is implied

Other Boolean operators are used like in NCBI databases
Parentheses to specify priority of Boolean operators

OR mouse is simply ignored!

With parentheses the query is handled correctly
Advanced searches on UniProt are similar to NCBI’s
Square brackets and wild cards do not work in Uniprot

Advanced search

Search using field in between square brackets

Translation:
- Square brackets and wild cards do not work in Uniprot.
- Advanced search: Search for caspase [DE] in UniProtKB, showing 1 to 25 of 3,963 results.
- Search using field in between square brackets: Search for caspase [DE] in UniProtKB, showing 1 to 25 of 1,605 results.
- HSP2* search returns 0 results in UniProtKB.
Quotes to indicate expressions: similar as in NCBI databases
Filters to limit the number of search results

<table>
<thead>
<tr>
<th>Entry</th>
<th>Entry name</th>
<th>Protein names</th>
<th>Gene names</th>
<th>Organism</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q58NA2</td>
<td>Q58NA2_NE0V1</td>
<td>Insulin-dependent glucose transport...</td>
<td></td>
<td>Neovison vison (American mink) (Mustela vison)</td>
<td>100</td>
</tr>
<tr>
<td>Q9TQE4</td>
<td>Q9TQE4_HUMAN</td>
<td>MHC class II HLA-DQ-beta-1</td>
<td>HLA-DQB1</td>
<td>Homo sapiens (Human)</td>
<td>79</td>
</tr>
<tr>
<td>P70169</td>
<td>DOC2B_MOUSE</td>
<td>Double C2-like domain-containing pr...</td>
<td>Doc2b</td>
<td>Mus musculus (Mouse)</td>
<td>412</td>
</tr>
<tr>
<td>Q9BZD2</td>
<td>S29A3_HUMAN</td>
<td>Equilibrative nucleoside transporter...</td>
<td>SLC29A3, ENT3, UNQ717/PRO1380</td>
<td>Homo sapiens (Human)</td>
<td>475</td>
</tr>
<tr>
<td>P14672</td>
<td>GTR4_HUMAN</td>
<td>Solute carrier family 2, facilitate...</td>
<td>SLC2A4, GLUT4</td>
<td>Homo sapiens (Human)</td>
<td>509</td>
</tr>
<tr>
<td>Q6EEV6</td>
<td>SUMO4_HUMAN</td>
<td>Small ubiquitin-related modifier 4</td>
<td>SUMO4, SMT3H4</td>
<td>Homo sapiens (Human)</td>
<td>95</td>
</tr>
<tr>
<td>P19357</td>
<td>GTR4_RAT</td>
<td>Solute carrier family 2, facilitate...</td>
<td>Slc2a4, Glut-4, Glut4</td>
<td>Rattus norvegicus</td>
<td>509</td>
</tr>
</tbody>
</table>
Uniprot records contain Evidence: how sure is the actual existence of this protein?

P70169 - DOC2B_MOUSE

Protein: Double C2-like domain-containing protein beta
Gene: Doc2b
Organism: Mus musculus (Mouse)
Status: Reviewed - Experimental evidence at protein level

In UniProtKB there are 5 types of evidence for the existence of a protein:

- 1. Experimental evidence at protein level
- 2. Experimental evidence at transcript level
- 3. Protein inferred from homology
- 4. Protein predicted
- 5. Protein uncertain

Video tutorial on UniProt records: https://www.youtube.com/watch?v=lHeqHpNaSis
You can download sequences from UniProt

Select the sequences you wish to download
No selection will download all sequences

Expand basket to see its content
Sequences can be downloaded in various formats

**Canonical:**
One of multiple isoforms (splice variants) is selected
- The most prevalent
- The most similar to orthologs
- Allows clearest description of domains, isoforms...
- If no info is available => the longest sequence

**Canonical & isoforms:** all variants are included

Canonical sequences and isoforms: [http://www.uniprot.org/help/canonical_and_isoforms](http://www.uniprot.org/help/canonical_and_isoforms)
When multiple isoforms exist, one is chosen as canonical.

 Isoform 1 (identifier: P07305-1) [UniParc]  
This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

 Isoform 2 (identifier: P07305-2) [UniParc]  
The sequence of this isoform differs from the canonical sequence as follows: 3-19: Missing.
All locations in a record refer to the canonical sequence.
The Retrieve/ID mapping tool

Video tutorial on Retrieve/ID mapping: https://www.youtube.com/watch?v=kLdgjqWoMZc
Uniprot provides additional protein databases

**Proteome** = full set of proteins expressed by completely sequenced organisms
- attempt to organize UniProt by species (~Ensembl)
- contain Swiss-Prot and TrEMBL sequences
- search by organism name

**UniRef**: attempt to remove redundancy from UniProt
- All UniProt sequences are pairwise compared
- Sequences are merged based on % identity

**UniRef100**: identical sequences are merged into a single UniRef entry

**UniRef90**: sequences that are at least 90% identical are merged

---

Wiki:
Exercises on UniProt

Video tutorial on Proteomes: [https://www.youtube.com/watch?v=ZLt3ug0mZ7A](https://www.youtube.com/watch?v=ZLt3ug0mZ7A)

Proteome [http://www.uniprot.org/help/proteome](http://www.uniprot.org/help/proteome)

UniRef [http://www.uniprot.org/uniref/](http://www.uniprot.org/uniref/)
Summary three main sequence databases

Genbank
EMBL-bank / ENA
DDBJ

They are filled with sequences by scientists / consortia / patent offices
They store sequences in their own way but they do exchange data every day
Every sequence gets unique identifiers, some identifiers are shared
A sequence record contains two categories of information:
  annotation: general + sequence features
  the actual sequence
These databases contain a lot of *redundancy + errors*!
Summary derived sequence databases

**RefSeq** is a **clean non-redundant** subset of Genbank

**Ensembl** annotates fully sequenced + assembled genomes

  easy to distinguish reliable annotations from predictions: CCDS

  lots of easily accessible external information via left menu

  visualization on location page

**UniProt** contains a **high quality** protein database

Many more sequence databases exist ...
Summary searching databases

Make your searches as specific as possible using

- Boolean operators, parentheses
- Quotes, wild cards
- Advanced searches in fields
- Filters