



Databases

Methods Course Bioinformatics
WS 2007/08
Hoff/Crass/Haubrock/Michael



The aim of this part of the methods course is to introduce you to the existence and usage of different databases associated with biological research.

Contents

1 Information Retrieval	1
1.1 Central Providers of Biological/Medical Databases	1
1.2 Literature Databases	2
1.3 Taxonomy	3
2 Sequence Retrieval	3
2.1 Accession Numbers	3
2.2 Data Formats	3
2.3 GenBank	6
2.4 Ensembl	6
2.5 UniProt/SwissProt	7
3 Specific Databases	7

1 Information Retrieval

Finding important information on your personal field of research today is not limited to books and article collected in your local library. The internet is a valuable resource for research related information. In this part of the methods course, we discuss important sequence and literature databases as well as organism/subject related databases.


1.1 Central Providers of Biological/Medical Databases

Two major entry points provide access to a virtual research community organized around databases, software, and data formats:

- **National Center for Biotechnology Information (NCBI, Bethesda, USA)**
Starting in 1988, the National Library of Medicine and the National Institutes of Health in the United States decided to build a national resource for molecular biology information called NCBI. This organization develops databases, computer programs for genomic data, guides research in computational biology, and runs a central server for published biomedical information. The duties and responsibilities of the NCBI are to build up a better understanding of molecular processes affecting human health and diseases.

- **European Bioinformatics Institute (EBI, Hinxton, UK)**

The European Bioinformatics Institute (EBI) is a non-profit academic organization that is part of the European Molecular Biology Laboratory (EMBL). The EBI is a center for research in bioinformatics, also providing tools, database, and defining data formats for genomic data.

 **Exercise 1: EBI, NCBI**

1. Open a web browser and try to find the web pages of the two central institutes that are mentioned above.
2. What are the topics of the two institutes? Which type of information is available?

1.2 Literature Databases

Literature databases are very important when you are looking for information about a research topic without consulting all the different relevant journal webpages separately. In the following, two different literature databases are described.

- **PubMed**

PubMed is the most widely used free literature search tool for biologists. It is integrated in the database-retrieval system at the NCBI and it is essentially a web interface to the Medline database which indexes more than 11 million journal abstracts. PubMed provides links to more than 1100 journals which are available on the web.

The interface allows the user to specify a search term and a search field (e.g. title, text word, journal or author). Queries retrieve abstracts for most journals.

Search operators in PubMed:

- Boolean operators AND, OR, NOT
- wildcard function *. For example `schizo* AND 8q*` tries to find publications which present evidence in schizophrenia associated on chromosome 8q21. The wildcard function is useful when several words indicate articles as important literature for your research field (`schizo*` → `schizoaffective`, `schizophrenia`, `schizophrenic`).
- *search tags* or *filters* limit the search, e.g.
 - * `Search Term Mody 4` (no filter) - results: more than 13000 papers
 - * `Search Term Mody 4 [ti]` (restriction of your search just on the title of all available papers in PubMed) - result: 1 paper

Further filters and operators are listed on the help pages of PubMed!

- **ISI Web of Knowledge**

ISI Web of Knowledge is a widely used but not free literature database for many scientific research fields. ISI enables you to search for publications in many different databases, also including PubMed. Regrettably, you cannot access ISI without subscription but university of Göttingen has a subscription, which means you can use it from any computer inside the university network. A detailed search mask allows you to narrow search results.

 **Exercise 2: PubMed and ISI Web of Knowledge**

1. Start a browser and use the URL `www.pubmed.org`.
2. Which gene is involved in the `Mody4` disease? Read the abstract. (Hint: You will not find this information in the `Mody4 [ti]` paper! Check spelling of your query for spaces!)
3. Search for `ethics of cloning`. Check `Details` tag in the PubMed search engine to see how the terms are mapped (rightmost panel of the pubmed query interface).

4. Find out what happens when you enter the following search terms:
 - B Crohn
 - B Crohn[au]
 - Crohn B[au]
5. Find the webpage of ISI Web of Knowledge.
6. Use both literature search interfaces to find the original paper of the first cloning on a mammalian species by nuclear transfer from a cultured cell line. Which species was cloned and how many people were involved? If you cannot find anything... the first cloned mammal was a sheep.
7. Kary Banks Mullis received a Nobel Prize in chemistry in 1993, for his invention of the Polymerase Chain Reaction (PCR). Find the paper where he described the PCR method the first time by using as many web-resources as you wish. Hint: The first paper was *not* in Nature and it did *not* contain *Polymerase Chain Reaction* in its title!

1.3 Taxonomy

The NCBI taxonomy website includes a taxonomy browser for the major divisions of living organisms including archaea, bacteria, eukaryota, and viruses. This database includes taxonomy information and additional features like the genetic codes, molecular data, extinct organisms and recent changes in the classification schemes.

Exercise 3: NCBI Taxonomy Browser

1. Browse the taxonomy database. You can find this database at the NCBI server.
2. Look for the *Elephas Maximus* (Asiatic elephant). What kind of information does the taxonomy browser offer?
3. Which type of genomic sequences are stored for this species. Try the *Entrez* record *nucleotide* for this species entry.
4. The taxonomy database is a useful database to retrieve all available sequence data for an organism. How many genes could you retrieve for the human (*homo sapiens*) and the rat (*rattus norvegicus*) species? How many genes are available for these two species on the X chromosome.

2 Sequence Retrieval

In this section, data formats and general databases for nucleotide and protein sequence retrieval are introduced.

2.1 Accession Numbers

DNA and protein sequence records are tagged with accession numbers. They consist of a string about four to ten numbers and/or alphabetic characters that are associated with a molecular sequence record. Often, you find accession numbers at the beginning of a data entry. The accession number of a genome/gene/protein is given in most publications and it allows you to find this exact piece of sequence information.

2.2 Data Formats

Different file formats are used for file exchange in bioinformatics. More or less every database enables the user to get sequence data in some format. This data format can be used as an input for sequence analysis software.

GenBank Format GenBank is the NIH genetic sequence database format. A GenBank entry includes a concise description of sequence, scientific name and taxonomy of the source organism, and a table of features that identifies coding regions and other sites of biological significance, such as transcription units, sites of mutations, and repeats.

```

LOCUS       X14897                4145 bp    mRNA    linear    ROD 18-APR-2005
DEFINITION  Mouse fosB mRNA.
ACCESSION  X14897
VERSION    X14897.1  GI:50991
KEYWORDS   fos cellular oncogene; fosB oncogene; oncogene.
SOURCE     Mus musculus (house mouse)
  ORGANISM Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE  1 (bases 1 to 4145)
AUTHORS    Zerial,M., Toschi,L., Ryseck,R.P., Schuermann,M., Muller,R. and
            Bravo,R.
TITLE      The product of a novel growth factor activated gene, fos B,
            interacts with JUN proteins enhancing their DNA binding activity
JOURNAL    EMBO J. 8 (3), 805-813 (1989)
PUBMED    2498083
COMMENT    clone=AC113-1; cell line=NIH3T3.
FEATURES   Location/Qualifiers
     source          1..4145
                   /organism="Mus musculus"
                   /mol_type="mRNA"
                   /db_xref="taxon:10090"
     CDS             1202..2218
                   /note="unnamed protein product; fosB protein (AA 1-338)"
                   /codon_start=1
                   /protein_id="CAA33026.1"
                   /db_xref="GI:50992"
                   /translation="MFQAFP GDYDSGSRCS SSPSAESQYLSSVDSFGSPPTAAASQEC
                   AGLGEMPGSFVPTVTAITTSQDLQWL VQPTLISSMAQSQQPLASQPPAVDPYDMPGT
                   KPGCKIPYEEGPGPLAEVRDLPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNL
                   LLAL"
ORIGIN
1 ataaattctt attttgacac tcaccaaagt agtcacctgg aaaaccgct tttgtgaca
61 aagtacagaa ggcttggtca catttaaatc actgagaact agagagaaat actatcgcaa
121 actgtaatag acattacatc cataaaagt tccccagtc ttattgtaat attgcacagt
181 gcaattgcta catggcaaac tagtgtagca tagaagtcaa agcaaaaaca aaccaagaa
241 aggagccaca agagtaaac tgttcaacag ttaatagttc aaactaagcc attgaaatca
301 tcattgggat cgttaaaatg aatcttctca caccttgca gttgatgatt aacttttaca
361 gaacacaagc caagtttaaa atcagcagta gagatattaa aatgaaaagg tttgctaata
421 gagtaacatt aaataccctg aaggaaaaaa aacctaaata tcaaaataac tgattaaat
...
//

```

The sequence information in GenBank is organized into fields, each with an identifier, shown as the first text on each line.

FASTA Format This format contains a one line header followed by lines of sequence data. Sequences in FASTA formatted files are preceded by a line starting with a > symbol. The first word on this line is the name of the sequence. The rest of the line is a description of the sequence. The remaining lines in that format should contain sequence information.

```

>gi|50991|emb|X14897.1| Mouse fosB mRNA
ATAAATCTTATTTTGACACTCACAAAATAGTCACCTGGAACCCGCTTTTGTGACAAAGTACAGAA
GGCTTGGTACATTTAAATCACTGAGAAGTAGAGAGAAATACTATCGCAAACGTAAATAGACATTACATC
CATAAAAGTTTCCCGAGTCCTTATTGTAATATTGCACAGTGCAATTGCTACATGGCAAAC TAGGTAGCA
TAGAAGTCAAAGCAAAAACAACCAAGAAAGGAGCCACAAGAGTAAAACCTGTTCAACAGTTAATAGTTC
AAACTAAGCCATTGAATCTATCATTGGGATCGTTAAAATGAATCTTCTACACCTTGCAAGTATGATT

```

The FASTA format is one of the most frequently used formats in bioinformatics. It is easy to handle and contains additional information within the sequence.

EMBL Format The European Molecular Biology Laboratory (EMBL) maintains DNA and protein sequence databases. The format of their database entries are shown in the listing below. Similar to GenBank entries, a large amount of information is given for each sequence. The EMBL-format uses a two letter format to type single data

fields in one EMBL-entry. Every EMBL entry finishes with a sequence block, starting with the sequence shortcut SQ and finishing with the // symbol, which marks the end of the entry.

```

ID   X14897; SV 1; linear; mRNA; STD; MUS; 4145 BP.
XX
AC   X14897;
XX
DT   23-NOV-1989 (Rel. 21, Created)
DT   18-APR-2005 (Rel. 83, Last updated, Version 3)
XX
DE   Mouse fosB mRNA
XX
KW   fos cellular oncogene; fosB oncogene; oncogene.
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC   Muridae; Murinae; Mus.
XX
RN   [1]
RP   1-4145
RX   PUBMED; 2498083.
RA   Zerial M., Toschi L., Ryseck R.P., Schuermann M., Mueller R., Bravo R.;
RT   "The product of a novel growth factor activated gene, fos B, interacts with
RT   JUN proteins enhancing their DNA binding activity";
RL   EMBO J. 8(3):805-813(1989).
XX
DR   ASTD; TRAN00000152487.
DR   TRANSFAC; T00291; T00291.
XX
CC   clone=AC113-1; cell line=NIH3T3;
XX
FH   Key          Location/Qualifiers
FH
FT   source          1..4145
FT                   /organism="Mus musculus"
FT                   /mol_type="mRNA"
FT                   /db_xref="taxon:10090"
FT   CDS             1202..2218
FT                   /note="fosB protein (AA 1-338) "
FT                   /db_xref="GOA:P13346"
FT                   /protein_id="CAA33026.1"
FT                   /translation="MFQAFPGDYDSGSRCS SSPSAESQYLSSVDSFGSPPTAASQECA
FT                   GLGEMPGSFVPTVTAITTSQDLQWLQVPTLSSMAQSQGQPLASQPPAVDPYDMPGTSY
FT                   THSEVQVLGDPFPVSPSYTSSVLTCEVSAFAGAQR TSGSEQPSDPLNSP SLLAL"
XX
SQ   Sequence 4145 BP; 960 A; 1186 C; 1007 G; 991 T; 1 other;
      ataaattcctt attttgacac tcacccaaat agtcacctgg aaaaccgcgt ttttgtgaca      60
      aagtacagaa ggcttggtca catttaaadc actgagaact agagagaaat actatcgcaa      120
      actgtaatat acattacatc cataaaagt tccccagtc ttattgtaat attgcacagt      180
      gcaattgcta catggcaaac tagtgtagca tagaagtcaa agcaaaaaca aaccaaaagaa      240
      aggagccaca agagtaaaac tgttcaacag ttaatatgtc aaactaagcc attgaatcta      300
      tcattgggat cgttaaaatg aatcttccta caccttcgag tgtatgattt aacttttaca      360
      gaacacaagc caagtttaaa atcagcagta gagatattaa aatgaaaagg tttgctaata      420
      ...
      //

```

SwissProt Sequence Format The SwissProt sequence format is very similar to the EMBL format but it contains more information about physical and biochemical properties of the protein.

Exercise 4: Data Formats

1. Browse to the NCBI taxonomy database (<http://www.ncbi.nlm.nih.gov/Taxonomy>) and retrieve all available nucleotide sequences of the Asiatic elephant (*Elephas maximus*) in FASTA and GenBank format. Please use the Entrez records table to retrieve all genes (upper right corner, section gene).
2. How many sequences are available? Which part of the elephant genome has been sequenced?
3. Focus on the *cytochrome b* gene. Use the *links* reference to answer the following questions:
 - How many publications are stored in PubMed linked to this gene?
 - How long is the mitochondrial genome of the elephant?

- How many protein entries are available for this gene?
 - Store the entries in FASTA and GenBank format.
4. Browse to the following website: <http://www.ebi.ac.uk/cgi-bin/readseq.cgi>
 5. Use the downloaded GenBank format and convert it to EMBL- and FASTA format. Store the sequences.

2.3 GenBank

DDBJ, EBI and GenBank are equivalent in their sequence information content. They update their content every night. We will focus on GenBank. GenBank is a database storing DNA and protein sequences. In addition, GenBank entries contain bibliographic and biological annotation. The number of bases in GenBank doubles approximately every 14 months.

Entrez is the backbone of the NCBI database infrastructure. It is an integrated database retrieval system that enables the user to retrieve and browse datasets from the NCBI databases through a single gateway system.

Exercise 5: GenBank Sequence Retrieval

1. Browse to the NCBI homepage. Select the nucleotide database for the search interface (upper left popup box).
2. Retrieve the GenBank entry with accession number X76930.
3. Read the annotation and get the coding sequence, only.
4. What's happening when you use the CDS html interface?
5. Cut out the coding sequence and convert it to the corresponding amino acid sequence. Use a text editor to build up the coding sequence. To translate the sequence, you can try the following web tool:
 - <http://www.expasy.org/tools/dna.html>
6. Compare the results of your amino acid sequences with the annotated sequence in the GenBank entry. Consider the concept of reading frames.

2.4 Ensembl

Ensembl is a joint project between EMBL and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation of selected eukaryotic genomes. There are different entry points to access the database. It is possible to search for genes, chromosomal locations, SNPs, etc.

Exercise 6: Using Ensembl database

1. Browse to <http://www.ensembl.org>
2. Use the existing web interface to retrieve organisms which have a gene entry for the glucokinase gene. How many species which are stored in Ensembl database system contain an entry for this gene?
3. Select the human Ensembl gene entry for glucokinase. (You want to have the gene OTTHUMG00000022903...)
 - Which chromosomal location is linked to the human glucokinase?
 - How many exons can be observed for the proteins?
 - What is the biochemical function of the coded protein?
 - Which type of information can you find in the protein feature plot?
 - Use the genomic location link to get an overview for the retrieved gene and the corresponding genomic organization.
 - How many ESTs (expressed sequence tags) can be observed for that gene?
 - How many orthologous gene(s) are given for mouse and rat?

2.5 UniProt/SwissProt

UniProt is a central database of protein sequences and functions created by joining the information contained in SwissProt, TrEMBL, and PIR. SwissProt is a resource for protein sequences produced in collaboration between the University of Geneva and the EMBL Data Library (curated & high level annotation). TrEMBL is an unannotated supplement SwissProt. It consists of entries in SwissProt-like format derived from the translation of all coding sequences in EMBL, except CDS which are already included in Swiss-Prot. The protein information resource (PIR) is located at Georgetown University and joined the UniProt consortium in 2002. It also contains functional annotations of protein sequences.

Exercise 7: Retrieval of SwissProt entries

1. Use the SRS system to retrieve the protein sequence for the L-lactate dehydrogenase A chain (LDHA) for human and chimp.
2. Store the human and chimp entries in the local file system.
3. Cut out the first 30 aminoacids of each sequence. How similar are the two subsequences? How many aminoacids are different?

3 Specific Databases

Specific databases have been created for most model organisms/topics that are studied by molecular biologists. Some examples:

OMIM at NCBI, <http://www.arabidopsis.org>, <http://www.nematode.net>, <http://elegans.swmed.edu>, <http://rgd.mcw.edu>, <http://www.gdb.org>, <http://www.xenbase.org>, <http://zf.nichd.nih.gov/pubzf>, <http://www.fruitfly.org>

Exercise 8: Information/topics at specific databases

1. What is the main topic/organism of each of the above listed specific databases?
2. Which main features does each website offer? Are there any advantage of using the specific database in comparison to GenBank & relatives?
3. Can you find more databases for the same topics/model organisms?