Databases in bioinformatics

NCBI resources & GenBank

Till Helge Helwig

Proseminar “Genome Bioinformatics“
(Dr. Kay Nieselt)

22.05.2007

Outline

1. Introductional thoughts
2. Entrez
3. DNA / Genes
4. Genome
5. Proteins
6. Phenotypes
7. Summary

Introductional thoughts

- many research groups all over the world
- huge amount of data generated daily
- interpretation impossible without comparing to other data
- collecting all data in a central repository is essential

Introductional thoughts

- many research groups all over the world
- huge amount of data generated daily
- interpretation impossible without comparing to other data
- collecting all data in a central repository is essential

- **1988: creation of the NCBI**

What is the NCBI?

- **N**ational **C**enter for **B**iotechnology **I**nformation
- project of the **N**ational **L**ibrary of **M**edicine (**NLM**)
- thus belongs to the **N**ational **I**nstitute of **H**ealth (**NIH**)
- primarily a project to build information systems for molecular biology
- today a huge collection of databases and tools is available on the website **http://www.ncbi.nlm.nih.gov**

How to access a huge database?

- many different databases
- huge amounts of data
- lots of very different queries
- results readable and usable for everybody

How to access a huge database?

- many different databases
- huge amounts of data
- lots of very different queries
- results readable and usable for everybody

- **Entrez - The Life Sciences Search Engine**

Entrez

The main component of the NCBI resources

- integrated database retrieval system
- links all NCBI resources together
- provides access to more than 91 million DNA and protein sequences
- the whole NCBI website is searchable, too
- accessible via webbrowser or using the "**E-Utilities**"

Entrez (2)

Collecting information from everywhere

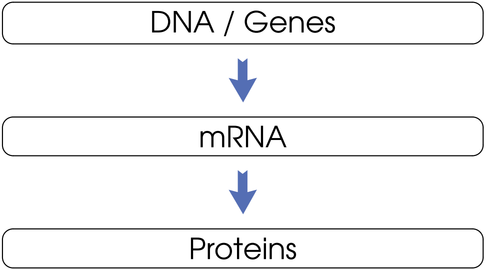
- search results are extended with links to biomedical literature using **PubMed** and **PubMed Central**
  - 16.5 million citations
  - 750 000 fulltext-articles, including whole articles from well known science journals
- **LinkOut** provides links to external projects and resources
- most query results also contain related data from other NCBI databases and direct links to tools

Entrez (3)

Customization

- using “**My NCBI**” you can customize the behaviour of **Entrez**
- query results can be shown directly, sent by email or provided as RSS feed
- the output is possible in many different formats including:
  - FastA format
  - XML documents
  - GenBank flat file
  - ...

Central dogma of biology



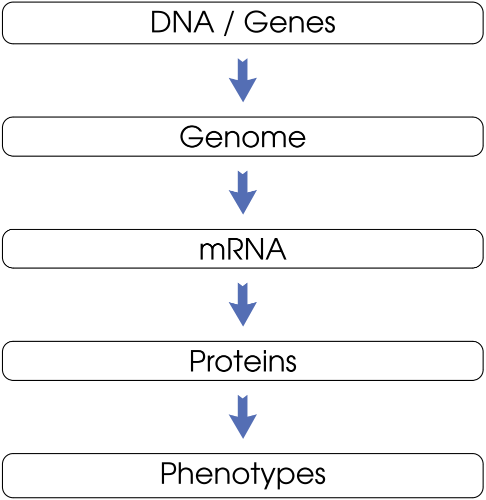
DNA / Genes

mRNA

Proteins

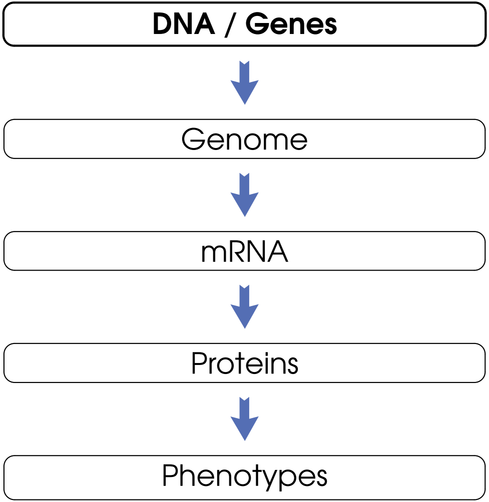
Central dogma of biology

Extended



Central dogma of biology

Extended



DNA / Genes

Genome

mRNA

Proteins

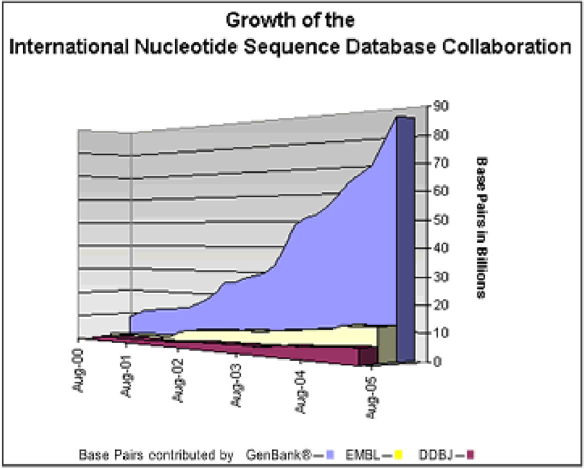
Phenotypes

GenBank

General information

- fastest growing and therefore largest public database of nucleotide sequences
- currently about 61 million sequences from more than 240 000 organism recorded
- collaboration with **DDBJ** and **EMBL**
- each entry contains:
  - unique accession id, e.g. AC202656, which is shared in DDBJ and EMBL
  - description & scientific name
  - biologically relevant sections (mutation sites, ...)
  - taxonomy information & literature references

GenBank (2)



Source: http://www.ncbi.nlm.nih.gov/Genbank/index.html

GenBank (3)

Organization of the records

- entries are stored in divisions
- first divisions were build from taxonomic relations: Bacteria (**BAC**), Viruses (**VRL**), Primates (**PRI**) and Rodents (**ROD**)
- later other divisions were added describing the used sequencing strategy:
  Expressed sequence tag (**EST**), Genome survey (**GSS**), High throughput genomic (**HTG**), ...
- today there are 18 divisions

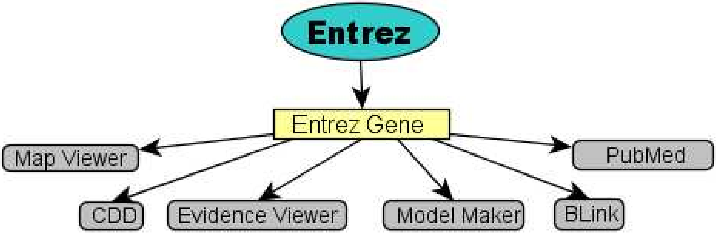
Gene databases

Entrez Gene

- gene-specific information
- focus is set on completely sequenced genomes
- actively researched genomes are included as well
- database is build by information accumulated by the NCBI staff and international collaborations
- all entries are linked to other resources

Gene databases (2)

Entrez Gene - The infrastructure



Gene databases (3)

UniGene & ProtEST

- **UniGene**
  - contains clusters of sequences from **GenBank** describing one unique gene
  - attempts to handle the redundancy of the **GenBank** entries for selected organisms
  - 87 000 clusters for human sequences in the release from 2006

Gene databases (3)

UniGene & ProtEST

- **UniGene**
  - contains clusters of sequences from **GenBank** describing one unique gene
  - attempts to handle the redundancy of the **GenBank** entries for selected organisms
  - 87 000 clusters for human sequences in the release from 2006
- **ProtEST**
  - provides pre-computed protein alignments and translations for **UniGene** entries

Gene databases (4)

HomoloGene & dbMHC

- **HomoloGene**
  - collection of homologs among the genes of 18 completely sequenced eukaryotic genomes
  - finding homologs is difficult and a typical problem of bioinformatics
  - records are linked to information from **OMIM** and **COGs**

Gene databases (4)

HomoloGene & dbMHC

- **HomoloGene**
  - collection of homologs among the genes of 18 completely sequenced eukaryotic genomes
  - finding homologs is difficult and a typical problem of bioinformatics
  - records are linked to information from **OMIM** and **COGs**
- **dbMHC**
  - **MHC** is of high interest to researchers, because it encodes the **HLA**
  - information on variations in the relevant genes are collected all over the world
  - integrated databases for several other medical relevant topics

Gene databases (5)

dbSNP & RefSeq

- **dbSNP**
  - records about single nucleotide polymorphisms (**SNP**)
  - 12 million entries about the human genome, 22 milion about other organisms
  - **SNP**s are displayed graphically

Gene databases (5)

dbSNP & RefSeq

- **dbSNP**
  - records about single nucleotide polymorphisms (**SNP**)
  - 12 million entries about the human genome, 22 milion about other organisms
  - **SNP**s are displayed graphically
- **RefSeq**
  - information about DNA, RNA and proteins of major research organisms
  - attempt to build non-redundant data sets from the huge amount of information in the other databases
  - currently about 4.1 million sequences represent about 3700 organisms

Gene tools

ORF, Splign & Spidey

- **Open reading frame finder** (**ORF**)
  - takes sequence or **GenBank** accession number
  - searches all possible reading frames of a specified length

Gene tools

ORF, Splign & Spidey

- **Open reading frame finder** (**ORF**)
  - takes sequence or **GenBank** accession number
  - searches all possible reading frames of a specified length
- **Splign**
  - aligns cDNA with genomic DNA or spliced sequences (Needleman-Wunsch + BLAST)

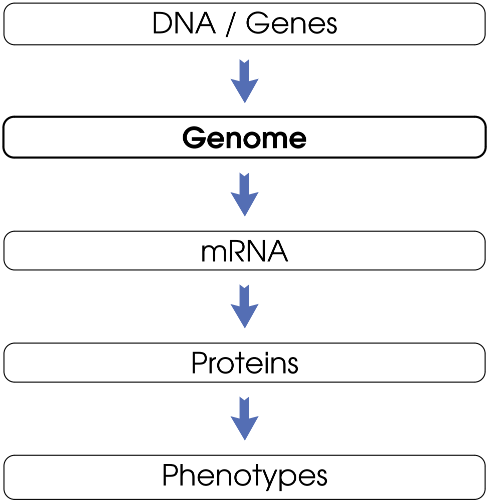
Gene tools

ORF, Splign & Spidey

- **Open reading frame finder** (**ORF**)
    - takes sequence or **GenBank** accession number
    - searches all possible reading frames of a specified length
- **Splign**
    - aligns cDNA with genomic DNA or spliced sequences (Needleman-Wunsch + BLAST)
- **Spidey**
    - creates alignments of eukaryotic mRNA with single genomic sequences using a splice-site model

Central dogma of biology

Extended



DNA / Genes

Genome

mRNA

Proteins

Phenotypes

Genome databases

Entrez Genome

- 370 complete microbial genomic sequences
- 2 500 viral sequences
- 1 050 reference sequences for eukaryotic organelles
- 20 genomes from higher organisms
- results are linked to resources for graphical views of sequences
- relevant **COGs** are included in the results as well as pre-calculated neighbours for microbial genomes

Genome databases (2)

Trace Archives & Genome Project

- **The Trace Archives**
  - 1.3 billion traces (raw sequence data from sequencing projects) stored currently
  - 860 organisms represented

Genome databases (2)

Trace Archives & Genome Project

- **The Trace Archives**
    - 1.3 billion traces (raw sequence data from sequencing projects) stored currently
    - 860 organisms represented
- **Genome Project**
    - records about status, progress and results of sequencing projects
    - keeps track even of projects that have not yet produced results
    - many biological facts are added to each description

Genome tools

Map Viewer

- **Map Viewer**
  - maps for every zoom factor are available from single genes to complete chromosomes
  - genetic and physical markers can be shown along the sequence
  - taxonomic list shows organisms for which maps are available
  - accessed by many other resources to display query results (e.g. **Entrez Gene**)

Special resources

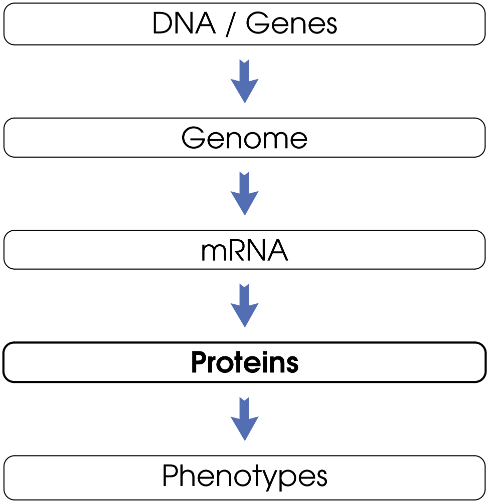
- resources for medical purposes
    - databases containing information on specific topics like influenza or cancer
    - opportunity for quick check on pathogenicity of an organism

Special resources

- resources for medical purposes
  - databases containing information on specific topics like influenza or cancer
  - opportunity for quick check on pathogenicity of an organism
- resources for organization of databases
  - **Clusters of Orthologous Groups** (**COGs**) combine information from completely sequenced prokaryotic protein sequences
  - similar database available for eukaryotic proteins (**KOGs**)

Central dogma of biology

Extended



DNA / Genes

Genome

mRNA

Proteins

Phenotypes

Protein resources

- **BLAST Link** (**BLink**)
  - produces pre-calculated alignments of protein sequences

Protein resources

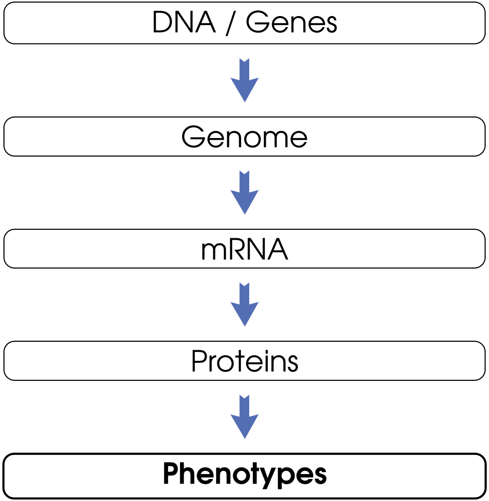
- **BLAST Link (BLink)**
    - produces pre-calculated alignments of protein sequences
- **Open Mass-Spectometry Search Algorithm**
    - **OMSSA** helps to identify spectra taken from tandem mass spectometry
    - similiar to BLAST as it calculates an "Expect-Value"

Protein resources

- **BLAST Link** (**BLink**)
  - produces pre-calculated alignments of protein sequences
- **Open Mass-Spectometry Search Algorithm**
  - **OMSSA** helps to identify spectra taken from tandem mass spectometry
  - similiar to BLAST as it calculates an "Expect-Value"
- **Molecular Modeling Database (MMDB)**
  - The molecular modeling database is built from entries of the Protein Database (**PDB**)
  - contains information of experimentally found 3D structures in biomolecules

Central dogma of biology

Extended



DNA / Genes

Genome

mRNA

Proteins

Phenotypes

Phenotype databases

- **Online Mendelian Inheritance in Man** (**OMIM**)
    - list of genes and genetic disorders
    - connection to disease phenotypes
    - 17 000 entries

Phenotype databases

- **Online Mendelian Inheritance in Man** (**OMIM**)
  - list of genes and genetic disorders
  - connection to disease phenotypes
  - 17 000 entries
- **Online Mendelian Inheritance in Animals** (**OMIA**)
  - same as **OMIM** but with information taken from animals

Phenotype databases

- **Online Mendelian Inheritance in Man** (**OMIM**)
  - list of genes and genetic disorders
  - connection to disease phenotypes
  - 17 000 entries
- **Online Mendelian Inheritance in Animals** (**OMIA**)
  - same as **OMIM** but with information taken from animals
- **GENSAT**
  - atlas of gene expressions in the central nervous system of the mouse

Phenotype databases (2)

- **Gene Expression Omnibus** (**GEO**)
  - repository for high-throughput data
  - results of microarray experiments, serial analysis of gene expressions (**SAGE**) experiments, mass spectometry peptide profiling, ...
  - about 3 billion measurements are recorded

Summary

- databases, tools and other resources for every imaginable kind of information
- effective search through all these with **Entrez**
- heavy linkage between the resources
- visit the homepage and take a look around yourselves: **http://www.ncbi.nlm.nih.gov**

Summary

- databases, tools and other resources for every imaginable kind of information
- effective search through all these with **Entrez**
- heavy linkage between the resources
- visit the homepage and take a look around yourselves: **http://www.ncbi.nlm.nih.gov**

Thank you for listening.