Bioinformatics for biomedicine

Course, 1 credit point

Per Kraulis

Goals for this course

- Introduction to bioinformatics
- Databases
  - Contents, background
- Analysis methods
  - Algorithms, properties, background
- Basic search and analysis
  - Applied to some problems of interest

Practical details

http://biomedicum.ut.ee/~kraulis

Time: Tuesday 14-16

Room: 1024, Biomedicum

Email: per.kraulis@ut.ee

Phone: 737 4052

Course design

1. What is bioinformatics? Basic databases and tools
2. Sequence searches: BLAST, Fasta
3. Multiple alignments, phylogenetic trees
4. Protein domains and 3D structure
5. Seminar: Sequence analysis of a favourite gene
6. Gene expression data, methods of analysis
7. Gene and protein annotation, Gene Ontology, pathways
8. Seminar: Further analysis of a favourite gene

Bioinformatics?

- “Information technology applied to the management and analysis of biological data”

  Attwood & Parry-Smith 1999

- “Collection, archiving, organization and interpretation of biological data”

  Thornton 2003

Databases (DB)

- Sets of data
- Stored on computer

- Explicit data model
  - What is in the DB? What is not?
  - Well-defined data structure

Analysis methods

- Searches
  - Keywords or free text
  - Similarity
- Comparison
  - Sequence-sequence alignment
  - Structures
- Features
  - Identification of (sites, domains)
  - Prediction of (secondary structure)

Important web sites

- EBI
  - www.ebi.ac.uk/
  - European Bioinformatics Institute
  - Databases: EMBL, UniProt, Ensembl,…
- NCBI
  - www.ncbi.nlm.nih.gov/
  - National Center for Biotechnology Information
  - Databases: PubMed, Entrez, OMIM,…

Historical background, 1

- “Atlas of Protein Sequence and Structure”, Margaret Dayhoff et al 1965
  - Printed book with all published sequences
  - New editions into the 1970s
  - Basis for Protein Information Resource (PIR), pir.georgetown.edu/
  - Since 2003 part of UniProt, www.uniprot.org/

Historical background, 2

- SwissProt
  - Amos Bairoch, University of Geneva
  - Swiss Institute of Bioinformatics, http://www.isb-sib.ch/
  - Data from literature, carefully curated
  - Started in 1986
  - Since 2003 part of UniProt http://www.uniprot.org/

DB properties

- Quality
  - Error rates, types of errors
  - Update policy
- Comprehensiveness
  - Data sources
- Redundancy
  - Multiple entries for same biological item?

Consider when choosing a DB

- Central data type
- Data entry and quality
- Primary or derived data
- Maintainer status
- Availability

Central data types

- Nucleotide sequences
  - EMBL, GenBank
- Protein sequences
  - UniProt (PIR, Swiss-Prot)
- Genes, genomes
  - Ensembl, EntrezGene
- 3D structure
  - PDB (RSCB)

Data entry and quality

- Method of data entry
  - Scientists deposit data
  - Curators enter data (from literature)
- Quality control
  - Consistency, redundancy, conflicts
  - Are checks applied?
- Update policy
  - Regularity
  - Are errors removed?

Primary or derived data?

- Primary data
  - Experimental data, more or less
  - Sequences, 3D structure, expression data
- Derived data
  - Obtained by analysis methods
  - Domains, secondary structure
- Aggregated data
  - Unified from several data sources

DB vs. Interface

- Confusion: Interface is not same as DB!
- Interface is the method of access
- Database (DB) is the data itself

- Same DB accessed by different interfaces (UniProt from ExPASy or EBI)
- One interface may be used to access different databases (SRS)

Maintainer status

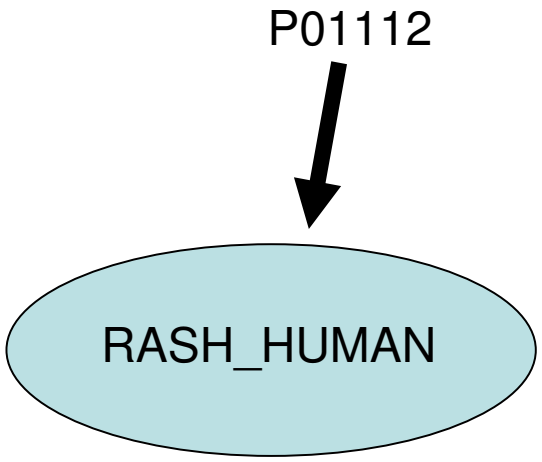
- Large, academic, public institute
  - EBI, NCBI
- Quasi-academic institute
  - TIGR, SIB
- Research group or scientist
- Company

Availability

- Publicly available, no restrictions
  - EMBL, GenBank
- Available, but with copyright
  - May not be re-used in other DB
  - UniProt
- Commercial
  - Copyright
  - May be accessible to academics at no charge

DB identifiers

- Identify a DB item uniquely
- A primary key for an item
  - Unique
  - Permanent
- “Accession code”
  - E.g. P01112
  - Use this for UniProt, EMBL, etc
- “Entry name”
  - E.g. RASH_HUMAN
  - Warning: may change!



Accession codes and updates

- DB items may be merged or split
  - Two sequence entries merged, e.g. they were actually the same protein
  - A sequence entry split, e.g. actually from two different genes
- Primary accession code
  - The new, recommended code
- Secondary accession code
  - The old; kept for trackability
- Version numbers in some DBs

Nucleotide sequence DBs

- Primary
  - EMBL, www.ebi.ac.uk/embl
  - GenBank, www.ncbi.nlm.nih.gov/GenBank
- Collaboration and synchronization
- Data submitted directly from sequencing projects, scientists
- Large, with subdivisions
- Redundant, fragments, rather messy…
- http://www.ebi.ac.uk/cgi-bin/expasyfetch?AF493

- Publicly available, no restrictions

Genome DBs, 1

- Nucleotides, but complete genome
- Usually high quality, careful annotation
- Ensembl
  - www.ensembl.org
  - Eukaryotes
  - Automatic annotation, links to other DBs
  - http://www.ensembl.org/Homo_sapiens/geneview?gen
- Vega
  - http://vega.sanger.ac.uk/
  - Manually curated, built on top of Ensembl

Genome DBs, 2

- UniGene
  - www.ncbi.nlm.nih.gov/UniGene
  - “An organized view of the transcriptome”
- EntrezGene (LocusLink)
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
- Species-specific databases
  - SGD, Saccharomyces cerevisiae, www.yeastgenome.org
  - Berkeley Drosophila Genome Project, www.fruitfly.org

Protein sequence DBs

- UniProt
  - www.uniprot.org
  - UniProtKB = UniProt Knowledge Base
  - UniProt = Swiss-Prot + PIR + TrEMBL
- Swiss-Prot
  - Credible sequences
  - Manual expert annotation
- TrEMBL
  - Translations of EMBL nucleotide sequences
  - Automatic, basic annotation
  - Eventually integrated into Swiss-Prot

UniProt

- ExPASY interface (Swiss Inst Bioinfo)
  - http://www.expasy.org/uniprot/P01112
- EBI interface
  - http://www.ebi.uniprot.org/uniprot-srv/uniProtV
- Same data, different look; your choice

Protein sequence domains

- Domains, motifs, families
  - Patterns of similar residues in a section of a protein sequence
  - Often a functional and structural unit
  - The presence of a domain: hints on function
- Pfam
  - Protein sequence domains
  - www.sanger.ac.uk/Software/Pfam/
  - Hidden Markov Models, software HMMER

Macromolecular 3D structure

- Protein Data Bank, PDB
  - Oldest computer-based bio-DB (1971)
  - http://www.rcsb.org/pdb/
  - 3D structures of proteins, oligonucleotides
  - X-ray crystallography and NMR
- SCOP, Structural Classification
  - http://scop.mrc-lmb.cam.ac.uk/scop/
  - Hierarchical scheme of classification
  - Similar 3D structures in families

Others, 1

- OMIM
  - Online Mendelian Inheritance in Man
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?
  - Human genes and genetic disorders
  - http://www.ncbi.nlm.nih.gov/entrez/dispomim.c

Others, 2

- GeneCards
  - Aggregate database; human
  - Links from gene to other DBs
  - http://www.genecards.org/
- GeneLynx
  - Aggregate database; human, rat, mouse
  - Links from gene to other DBs
  - http://www.genelynx.org/