Biological Databases

- Sequence Databases
- Genome Databases
- Structure Databases

Sequence Databases

- The sequence databases are the oldest type of biological databases, and also the most widely used

Sequence Databases

- Nucleotide: ATGC
- Protein: MERITSAPLG

The nucleotide sequence repositories

- There are three main repositories for nucleotide sequences: EMBL, GenBank, and DDBJ.
- All of these should in theory contain "all" known public DNA or RNA sequences
- These repositories have a collaboration so that any data submitted to one of databases will be redistributed to the others.

- The three databases are the only databases that can issue sequence accession numbers.
- Accession numbers are unique identifiers which permanently identify sequences in the databases.
- These accession numbers are required by many biological journals before manuscripts are accepted.

EST databases

- Expressed sequence tags (ESTs) are short sequences from expressed mRNAs.
- The basic idea is to get a handle on the parts of the genome that is expressed as mRNA (often called the transcriptome).
- ESTs are generated by end-sequencing clones from cDNA libraries from different sources.

EST cluster databases

- UniGene
- UniGene is a database at NCBI that contains clusters (UniGene clusters) of sequences that represent unique genes. These cluster are made automatically by partitioning GenBank sequences into a non-redundant set of gene-oriented clusters.

Ideal minimal content of a « sequence » db

Sequences !!
Accession number (AC)
References
Taxonomic data
ANNOTATION/CURATION
Keywords
Cross-references
Documentation

Sequence database: example

- …a SWISS-PROT entry, in fasta format:

- >sp|P01588|EPO_HUMAN ERYTHROPOIETIN PRECURSOR - Homo sapiens(Human).
- MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRV LERYLLEAKEAE
- NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQA VEVWQGLALLSEA
- VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLR ALGAQKEAISPPD
- AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEAC RTGDR

SWISS-PROT knowledgebase

- Created by Amos Bairoch in 1986
- Collaboration between the SIB (CH) and EBI (UK)
- Annotated (manually), non-redundant, cross-referenced, documented protein sequence database.
- ~122 ’000 sequences from more than 7’700 different species; 192 ’000 references (publications); 958 ’000 cross-references (databases); ~400 Mb of annotations.
- Weekly releases; available from more than 50 servers across the world, the main source being ExPASy

SWISS-PROT: species

- 7’700 different species
- 20 species represent about 42% of all sequences in the database
- 5’000 species are only represented by one to three sequences. In most cases, these are sequences which were obtained in the context of a phylogenetic study

Some protein motif databases

- Prosite - Regular expression built from SWISS-PROT
- ➢ PRINTS - aligned motif consensus built from OWL
- • (http://bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html)
- ➢ BLOCKS - PRINTS-like generated from PROSITE families
- • (http://www.blocks.fhcrc.org/)
- ➢ IDENTIFY - Fuzzy regular expressions derived from PROSITE
- ➢ pfam - Hidden Markov Model built from SWISS-PROT
- • (http://www.sanger.ac.uk/Software/Pfam)
- ➢ Profiles - Weight Matrix profiles built from SWISS-PROT
- ➢ Interpro - All of the above (almost)
- • (http://www.ebi.ac.uk/InterPro)

Genomic Databases

- Genome databases differ from sequence databases in that the data contained in them are much more diverse.
- The idea behind a genome database is to organize all information on an organism (or as much as possible).
- In many cases they stem out of the necessity for a centralized resource for a particular genome project. But of course they are also important resources for the research community.

Genomic Databases

- Ensembl
- Genome Browser
- NCBI

Structure Databases

- PDB
- SCOP

PDB

- The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) (1) in 1971 as an archive for biological macromolecular crystal structures.
- The three dimensional structures in PDB are primarily derived from experimental data obtained by X-ray crystallography and NMR .

SCOP

- The SCOP database groups different protein structures according to their evolutionary relationship.The evolutionary relationship of all known protein structures have been determined by manual inspection and automated methods.
- The goal of SCOP is to provide detail information about close relatives of proteins and protein and to provide an evolutionary based protein classification resource.