Biological Databases

Unit-2

Paper- Bioinformatics (DSE-1)

B.Sc. (H) Microbiology V Sem

Biological databases

- A biological database is a collection of data that is structured, searchable, updated periodically and cross referenced.
- The data is stores, maintained, annotated, curated and stored for public/research use.
- Data collected and organized in a specific but useful way

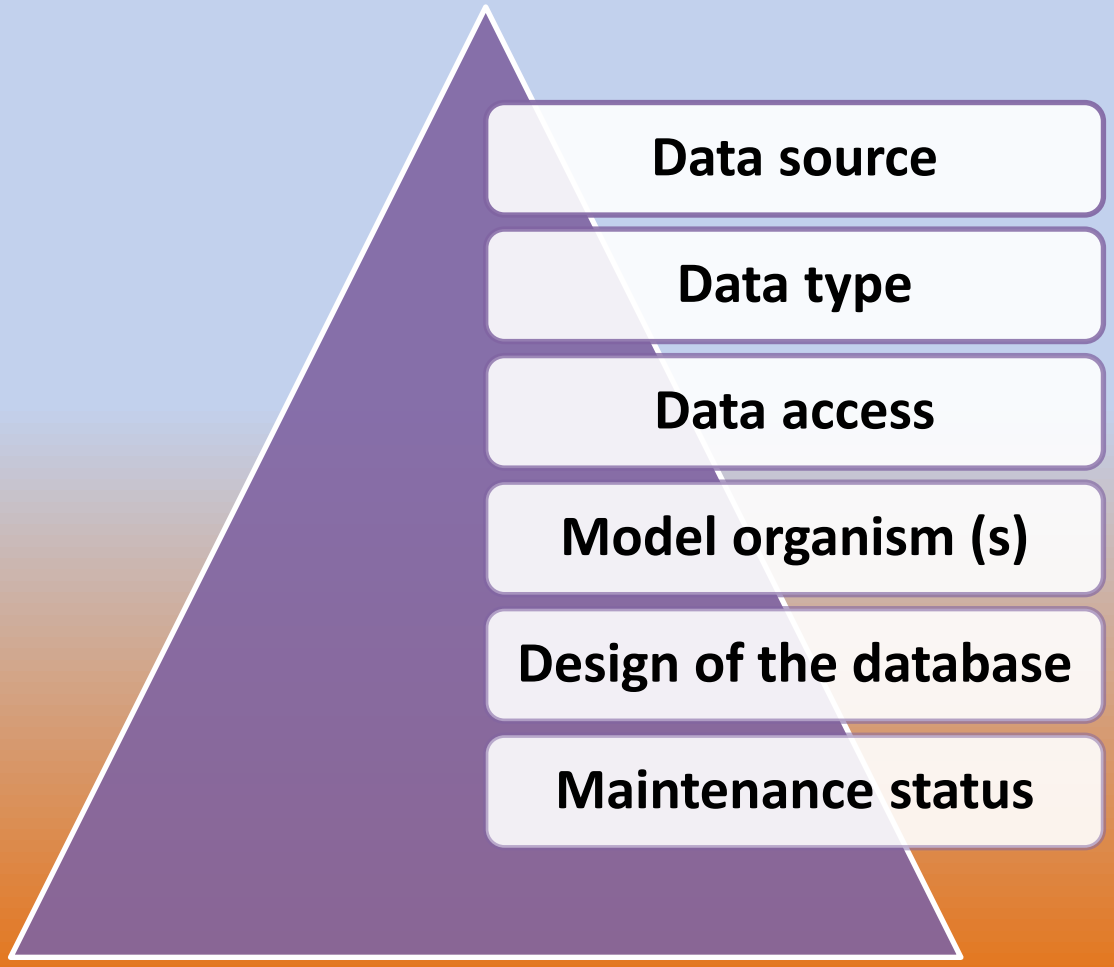
Biological databases: why?

- Need for storing and communicating large datasets has grown
- Distributed resources (experimental platforms & bioinformatics platform different)
- Make biological data available to scientists.
- To make biological data available in computer-readable form with ease of access.

Databases features

- Large volume data
- Data heterogeneity
- Easy to add new data
- Dynamic
- Uncertainty
- Data curation
- Large scale data intergartion
- Data sharing

Classifications of databases in Bioinformatics



Classifications of databases in Bioinformatics

- On the basis of -Type of data
  - nucleotide sequences
  - protein sequences
  - Genomes database
  - proteins sequence patterns or motifs
  - macromolecular 3D structure
  - gene expression data/transcriptome
  - metabolic pathways
  - Literature database

Different classifications of databases….

- On the basis of data and analysis -Primary or derived databases
  - Primary databases: experimental results directly into database
  - Secondary databases: results of analysis of primary databases
  - Composite database (Aggregate of many databases)
    - Links to other data items
    - Combination of data
    - Consolidation of data

Different classifications of databases….

- On the basis of -Availability
  - Publicly available, no restrictions
  - Available, but with copyright
  - Accessible, but not downloadable
  - Academic, but not freely available
  - Proprietary, commercial; possibly free for academics

Some of the Bioinformatics Databases

GenBank	www.ncbi.nlm.nih.gov	nucleotide sequences
Ensembl others)	www.ensembl.org	human/mouse genome (and
PubMed	www.ncbi.nlm.nih.gov	literature references
NR	www.ncbi.nlm.nih.gov	protein sequences
SWISS-PROT	www.expasy.ch	protein sequences
InterPro	www.ebi.ac.uk	protein domains
OMIM	www.ncbi.nlm.nih.gov	genetic diseases
Enzymes	www.chem.qmul.ac.uk	enzymes
PDB	www.rcsb.org/pdb/	protein structures
KEGG	www.genome.ad.jp	metabolic pathways

DNA Sequence databases

- Main repositories:
  - GenBank (US) (http://www.ncbi.nlm.nih.gov/Genbank/index.html)
  - EMBL (Europe) (http://www.ebi.ac.uk/embl/)
  - DDBJ (Japan) (http://www.ddbj.nig.ac.jp/)

– Primary databases

Sequence Databases

- Annotated sequence databases
  - SWISS-PROT, GenBank etc…
  - Usage: identifying function, retrieving information
- Low-annotation sequence databases
  - EST databases, high-throughput genome sequences
  - Usage: discovery of new genes

Genome Databases

Focus on one organism or group of organisms:

- Colibase (*E. coli* and related species) - OBSOLETE http://colibase.bham.ac.uk/
- GDB (human) http://www.gdb.org/ - OBSOLETE
- Flybase (Drosophila) http://flybase.bio.indiana.edu/
- WormBase (*C. elegans*) http://wormbase.org
- AtDB (Arabidopsis) http://www.arabidopsis.org
- TAIR
- RiceDB
- SGD (*S. cerevisiae*) http://genome-www.stanford.edu/Saccharomyces/

General Protein Databases

- SWISS-PROT -Manually curated
- high-quality annotations, less data
- GenPept/TREMBL
  - Translated coding sequences from GenBank/EMBL
  - Few annotations, more up to date
- Uniprot KB -The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation
- **All 3 now are now under UniProt (http://www.uniprot.org)**
- PIR - Phylogenetic-based annotations -

Protein domain databases

- Pfam (http://www.sanger.ac.uk/Software/Pfam/)
  - Collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families
- SMART **(a Simple Modular Architecture Research Tool)**
  - Identification and annotation of genetically mobile domains and the analysis of domain architectures
  - (http://smart.embl-heidelberg.de/help/smart_about.shtml
- CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)
  - Combines SMART and Pfam databases
  - Easier and quicker search

Sequence & Structure Databases

- PDB (Protein Databank)
  - Stores 3-dimensional atomic coordinates for biological molecules including protein and nucleic acids
  - Data obtained by X-ray crystallography, NMR, or computer modelling
  - http://www.rcsb.org/pdb/
- MMDB (Molecular Modelling database)
  - Over 28,000 3D macromolecular structures, including proteins and polynucleotides
  - (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure)
- SCOP (Structural Classification of Proteins)

Expression Databases

- **RNA expression**
  - Results of microarray experiments measuring the change in specific mRNA content under certain conditions
  - Array Express (EBI) and GEO (NCBI)
  - Not user friendly
- **Proteome databases**
  - 2D gel electrophoresis images representing the protein content of a cell or tissue under specific conditions
  - SWISS 2D PAGE at http://us.expasy.org/ch2d/

KEGG: Kyoto Encyclopedia of Genes and Genomes

- KEGG:http://www.genome.jp/kegg/
  - Also look at MetaCyc, another metabolic pathways database http://metacyc.org/
- Primarily used for metabolic reaction pathways, which are manually curated from published materials
  - http://www.genome.jp/kegg/pathway.html

KEGG PATHWAY DATABSE:
http://www.genome.jp/kegg/pathway.html

- Breakdown into major categories:
  - metabolism (the most important one),
  - genetic information processing (including protein folding and sorting),
  - environmental information processing (including membrane transport and intracellular signaling),
  - cellular processes & plus some others
- Broken down into subcategories, e.g. carbohydrate metabolism, and then into individual pathways, e.g. glycolysis/gluconeogenesis (http://www.genome.jp/kegg/pathway/map/map00010.html)

Other Database Types

- Literature — MEDLINE (http://ncbi.nlm.nih.gov/PubMed/), HighWire (http://www.highwire.org)
- Variation — dbSNP(http://ncbi.nlm.nih.gov/SNP/), HGBase (http://hgbase/interactiva/de)
- Metabolic pathways

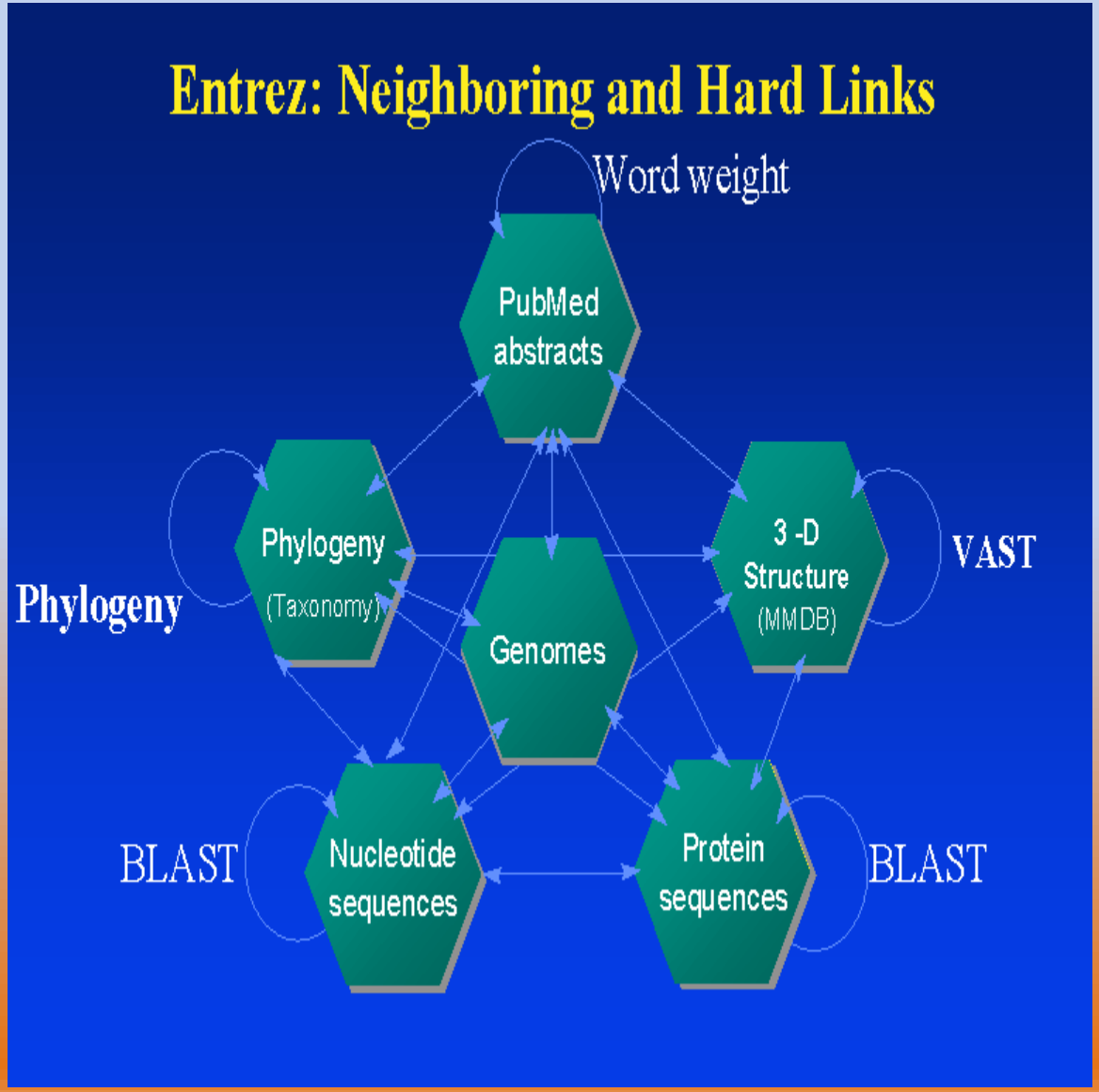
KEGG (http://kegg.genome.ad.jp/kegg/)
WIT(http://wit.mcs/anl.gov/WIT2)

- Organisms and nomenclature — Taxonomies(e.g.:http://ncbi.nlm.nih.gov/Taxonomy/)

Mendel(http://mbclserver.rutgers.edu/CPGN)

- DATABASE: the Journal of Biological Databases and Curation

Data type	Explanation	Example
Bibliographic DB	Contains article and research papers of different journals	MEDLINE, Pubmed
Genome DB	Contains whole genome sequences of viruses, eukaryotes or prokaryotes	Genome Information Broker (GIB), Entrez genome of NCBI
Sequence DB	Contains protein and nucleotide sequence	DDBJ, EMBL, SWISS prot
Structure DB	Contains 3D structure of proteins and nucleic acids	Nucleotide Database (NDB), Protein Data Bank (PDB)
Metabolic DB	Contains data about various biological pathways	Kyoto Encyclopedia of genes and genomes (KEGG)
Enzyme DB	Contains data about structure, function and pathways of various enzymes	ExPasy, REBASE
Disease DB	Disease related information	OMIM
Chemical DB	Data about biological activity of several small molecules	PubChem
Microarray DB	Mainly Data obtained from microarray experiments	Gene Expression Omnibus (GEO), Human Gene Expression Index (HUGE)



http://www.ncbi.nlm.nih.gov/gquery/

TABLE 2.3. Search Field Qualifiers for GenBank

Qualifier	Field Name	Definition
[ACCN]	Accession	Contains the unique accession number of the sequence or record, assigned to the nucleotide, protein, structure, or genome record.
[ALL]	All fields	Contains all terms from all searchable database fields in the database.
[AUTH]	Author name	Contains all authors from all references in the database records.
[ECNO]	EC/RN number	Number assigned by the Enzyme Commission or Chemical Abstract Service to designate a particular enzyme or chemical, respectively.
[FKEY]	Feature key	Contains the biological features assigned or annotated to the nucleotide sequences. Not available for the protein or structure databases.
[GENE]	Gene name	Contains the standard and common names of genes found in the database records.
[JOUR]	Journal name	Contains the name of the journal in which the data were published.
[KYWD]	Keyword	Contains special index terms from the controlled vocabularies associated with the GenBank, EMBL, DDBJ, SWISS-Prot, PIR, PRF, or PDB databases.
[MDAT]	Modification date	Contains the date that the most recent modification to that record is indexed in Entrez, in the format YYYY/MM/DD.
[MOLWT]	Molecular weight	Molecular weight of a protein, in daltons (Da), calculated by the method described in the Searching by Molecular Weight section of the Entrez help document.
[ORGN]	Organism	Contains the scientific and common names for the organisms associated with protein and nucleotide sequences.
[PROP]	Properties	Contains properties of the nucleotide or protein sequence. For example, the nucleotide database's properties index includes molecule types, publication status, molecule locations, and GenBank divisions.
[PROT]	Protein name	Contains the standard names of proteins found in database records.
[PDAT]	Publication date	Contains the date that records are released into Entrez, in the format YYYY/MM/DD.
[SQID]	SeqID	Contains the special string identifier for a given sequence.
[SLEN]	Sequence length	Contains the total length of the sequence.
[WORD]	Text word	Contains all of the "free text" associated with a record.
[TITL]	Title word	Includes only those words found in the definition line of a record.

Note: Some of these qualifiers are interchangeable with PubMed qualifiers.
Source: www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html.