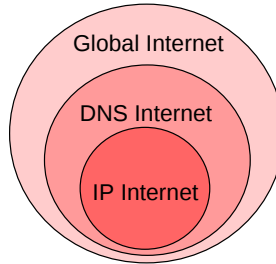


BINF 630: Introduction to Bioinformatics

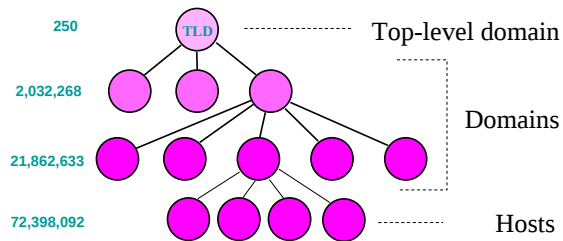
Iosif Vaisman

Email: ivaisman@gmu.edu

What is Internet



Domain Name System

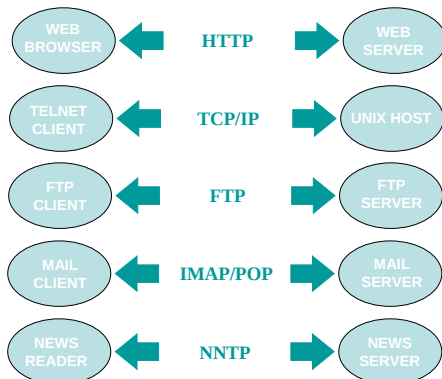


Data from ISC
(January 2000)

Client - Server Model



Client - Server Model



Uniform Resource Locator (URL)

protocol://host.domain[:port]/path/filename

ftp - an anonymous FTP server (<ftp://ftp.pdb.pdb.gov>)
http - a World Wide Web server (<http://mmlin4.pha.unc.edu/~cmb96>)
telnet - a telnet session (<telnet://nun.oit.unc.edu>)

Network applications in science

- **Virtual Laboratory**
- **Virtual Library**
- **Virtual Conference**
- **Virtual Classroom**

Network collaboration

Real-time data sharing -- exchange of information between remote participants in the project

Resources sharing -- remote access to the instruments and computers

Resources integration -- simultaneous use of remote instruments and computers

Bioinformatics servers

Remote data access -- database search, cross-links between the databases

Remote computing -- use of server's processing capabilities (sequence alignment, structure prediction, homology modeling)

Infospace navigation -- pointers to the available resources

Bioinformatics servers

Real-time

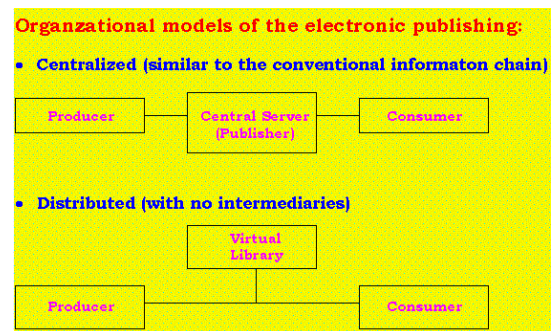
Asynchronous

Digital information cycle

Creation and capture
Storage and management
Rights management
Search and access
Distribution

Electronic publishing

Quality (peer review, retrospective evaluation)
Reliability (stability of serves, control over alterations, proper archiving and mirroring)



Hypertext Functionality in Scientific Literature

Active references
Forwarding references
Dynamic publishing

Ethical, Legal, and Economical Issues of Electronic Publishing

Intellectual property rights
Ownership of information
Information as a commodity

Molecular Databases

Nucleic acid sequences: GenBank, DNA Data Bank of Japan, EMBL Nucleotide Sequence Database

Nucleic acid structures: NDB - Nucleic Acid Database

Protein sequences: PIR - Protein Information Resource, SWISS-PROT

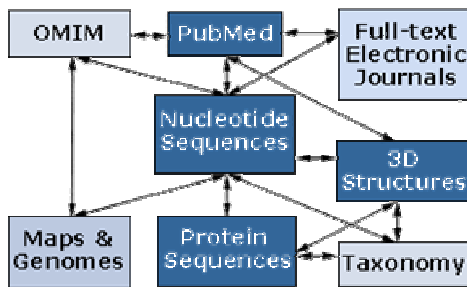
Protein structures: PDB - Protein Data Bank , NRL_3D

Physical properties: Biological Macromolecule Crystallization Database, BioMagResBank

Molecular images: SWISS-3DIMAGE, Molecules R US

NCBI integrated search and retrieval system

Entrez



NCBI Databases

- **nr** - All non-redundant GenBank CDS translations+PDB+SwissProt+PIR
- **month** - All new or revised GenBank CDS released in the last 30 days
- **swissprot** - the last major release of the SWISS-PROT protein sequence database (no updates)
- **yeast** - Yeast (*Saccharomyces cerevisiae*) protein sequences.
- **E. coli** - E. coli genomic CDS translations
- **pdb** - Sequences derived from the 3-dimensional structure Brookhaven Protein Data Bank
- **kabat** - Kabat's database of sequences of immunological interest

Database

database	a collection of related structured information about entities
file	a collection of records
record	a set of fields
field	a single characteristic of an entity
character	a symbol used in data field

Example of a Genbank entry

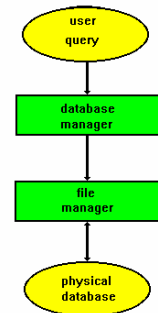
```
LOCUS      VIBHALUXA   3141 bp    DNA             BCT           15-FEB-1996
DEFINITION V.harveyi luciferase alpha and beta subunit (luxA and luxB) genes,
            complete cds.
ACCESSION  M10961 M13494
NID        g155174
KEYWORDS   luciferase.
SOURCE     Vibrio harveyi DNA.
ORGANISM   Vibrio harveyi
            Eubacteria; Proteobacteria; gamma subdivision; Vibrionaceae;
            Vibrio.
REFERENCE  1 (bases 1 to 1838)
AUTHORS   Cohn,D.H., Mileham,A.J., Simon,M.I., Nealson,K.H., Rausch,S.K.,
            Bonam,D. and Baldwin,T.O.
TITLE      Nucleotide sequence of the luxA gene of Vibrio harveyi and the
            complete amino acid sequence of the alpha subunit of bacterial
            luciferase
JOURNAL    J. Biol. Chem. 260 (18), 6139-6146 (1985)
MEDLINE    85207595
REFERENCE  2 (bases 1745 to 3141)
AUTHORS   Johnston,T.C., Thompson,R.B. and Baldwin,T.O.
TITLE      Nucleotide sequence of the luxB gene of Vibrio harveyi and the
            complete amino acid sequence of the beta subunit of bacterial
            luciferase
JOURNAL    J. Biol. Chem. 261 (11), 4805-4811 (1986)
MEDLINE    86168191
```

Example of a Genbank entry

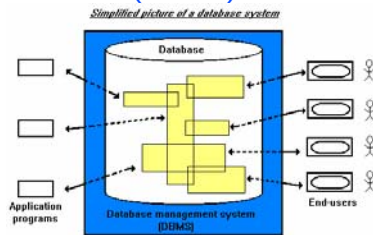
```

FEATURES             Location/Qualifiers
     gene             797..1774
                     /gene="luxA"
     CDS              797..1774
                     /gene="luxA"
                     /codon_start=1
                     /product="luciferase alpha subunit"
                     /db_xref="PID:g155175"
                     /transl_table=11
                     /translation="MKFGNFLLTYQPPELSQTEVMKRLVNLGKASEGCGFDTVMLEH
                     HFTEFGLLGNPYVAAHLLGATETLVNVTAAIVLPTAHPVRQAEDVNLDDQMSKGRFR
                     FGICRGLYDKDFRVFGTDMDNSRALMDWYDLMKEGFNEGYIAADNEHIKFPKIQLNP
                     SAYTGGAPVYVAESASTTEMAAERGLPMLSWIINTHEKKAQLDLYNEVATENGVD
                     VTKIDHCLSYITSVDHDSNRKADICRNFLGHMYDSYVNAIKIFDSDSQTKGYDFNKGQ
                     WRDFVLKGKDTNRRIDYSYEINPVGTPEECIAIQQIDATGIDNICCGFEANGSEE
                     EIIASMKLFQSDVMPYLKEKQ"
BASE COUNT      883 a   665 c   741 g   852 t
ORIGIN          1 bp upstream of EcoRI site.
               1 gaattcacca tgacgacggg caaaaatagt ttgtgcactg tttatcactg gctgcagacc
               61 aagggcacac aaacacatgg ctgatgctg gcaagctct cagtcgtgt cgccatgaa
               121 gtatctctg atctggact gtcctttctg attactgcg ttggtgtgt gaacttcgt
               181 gacacactag aaaaagcgt ttgtttgat tacctcagt ttgcctatga tgactacca
               ....
    
```

Database Organization



Database Management System (DBMS)



Four major components of DBMS:

Data * Hardware * Software * Users

Data Model

- A named logical unit (record type, data item)
- Relationships among logical units

Relationships among logical units

- one to one
- one to many
- many to one

Database administration

- Redundancy eliminated
- Inconsistency avoided
- Data shared
- Standards enforced
- Security applied
- Integrity maintained
- Requirements balanced