

BINF 630: Bioinformatics Methods

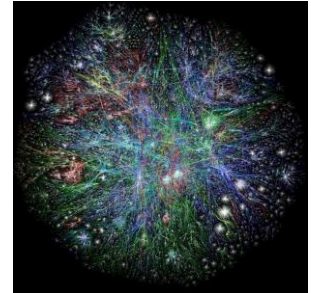
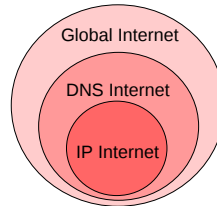
Iosif Vaisman

Email: ivaisman@gmu.edu

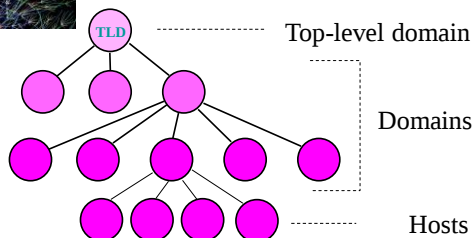
What is Internet?

S: (n) **internet**, net, cyberspace (a computer network consisting of a worldwide network of computer networks that use the TCP/IP network protocols to facilitate data transmission and exchange)

WordNet: An Electronic Lexical Database (MIT Press)



Domain Name System



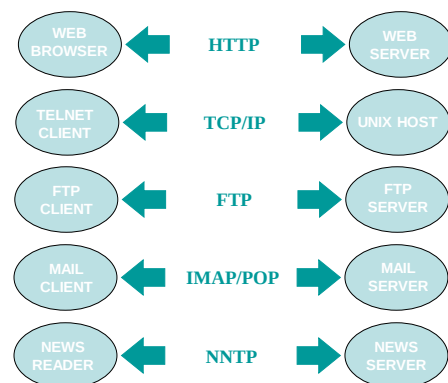
TCP/IP model

Layers	Protocols
1 - Physical Layer This layer defines the network hardware and device drivers.	Ethernet, ADSL, 802.11
2 - Network Layer This layer is used for basic communication, addressing and routing.	IPv6
3 - Transport Handles communication among programs on a network.	TCP, UDP
4 - Application End-user applications reside at this layer.	HTTP, FTP, IMAP

Client - Server Model



Client - Server Model



Uniform Resource Locator (URL)

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

`http://binf.gmu.edu/vaisman/binf630`

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

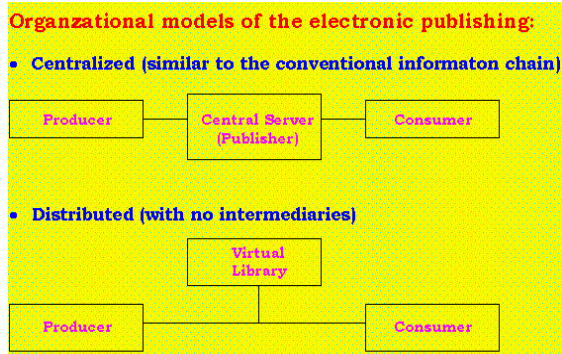
Asynchronous

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)



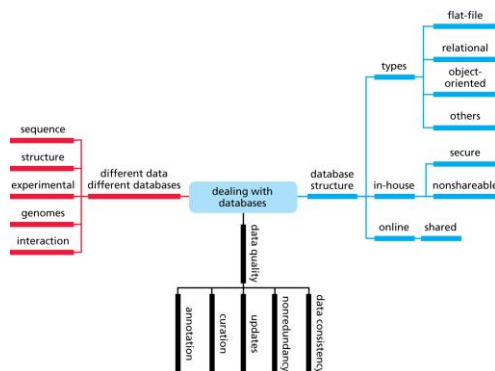
Hypertext Functionality in Scientific Literature

Active references
Forwarding references
Dynamic publishing

Ethical, Legal, and Economical Issues of Electronic Publishing

Intellectual property rights
Ownership of information
Information as a commodity

Databases in Bioinformatics

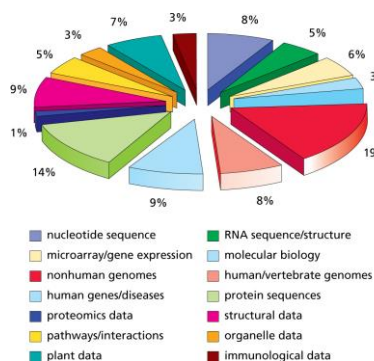


Molecular Databases

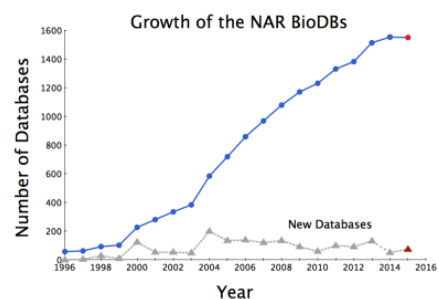
Nucleic acid sequences:	GenBank	(189 million, Dec 2015)
	WGS	(317 million, Dec 2015)
Protein sequences:	UniProtKB:	
	Swiss-Prot	(550 thousand, Jan 2016)
	TrEMBL	(59 million, Jan 2016)
Protein structures:	PDB	(115 thousand, Jan 2016)

NAR annual database issue

NAR Molecular Databases



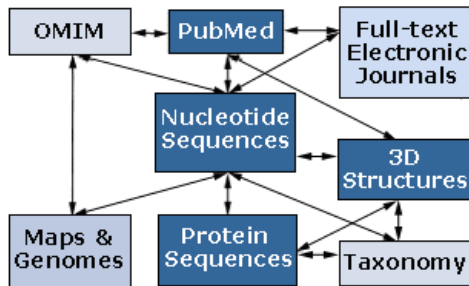
NAR Molecular Databases



Todd Smith, 2015

NCBI integrated search and retrieval system

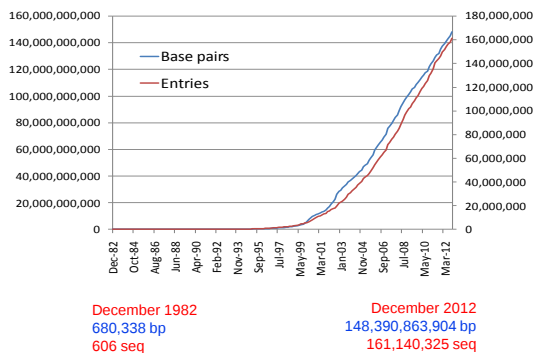
Entrez



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

Growth of GenBank



GenBank Selected Per-Organism Statistics

