BINF 630: Bioinformatics Methods

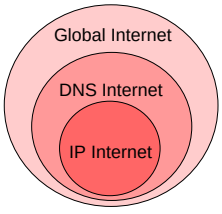
Iosif Vaisman

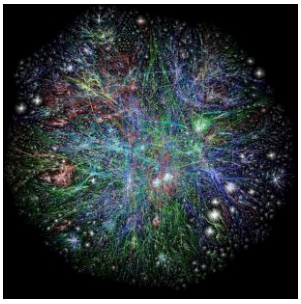
Email: ivaisman@gmu.edu

What is Internet?

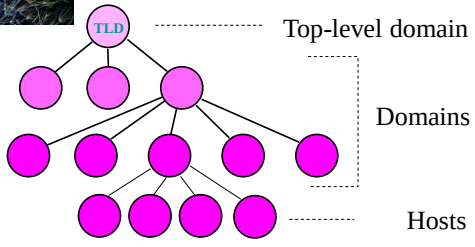
S: (n) **internet**, net, cyberspace (a computer network consisting of a worldwide network of computer networks that use the TCP/IP network protocols to facilitate data transmission and exchange)

WordNet: An Electronic Lexical Database (MIT Press)





Domain Name System

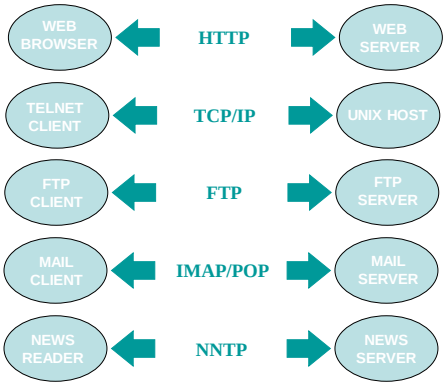


TCP/IP model

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 -- databse 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)

Organzational models of the electronic publishing:

- **Centralized (similar to the conventional informaton chain)**

- **Distributed (with no intermediaries)**

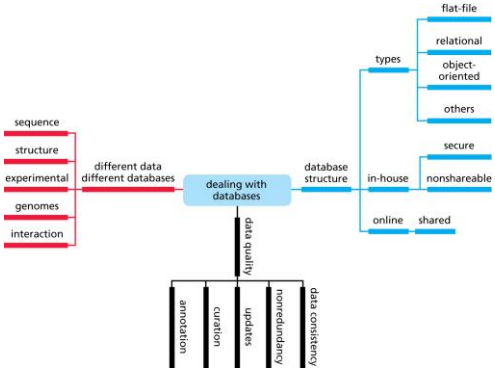
Hypertext Functionality in Scientific Literature

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



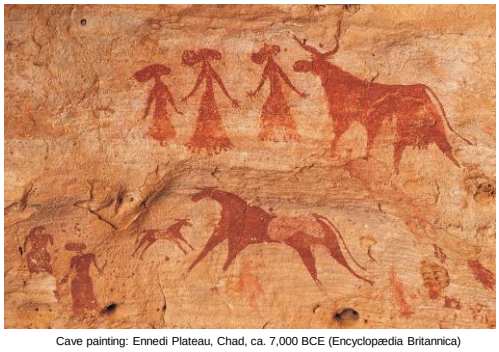
Data management and utilization

An early example of biological data depository:

Cave painting: Lascaux Grotto, near Montignac, France., ca. 15,000 BCE (Ralph Morse, Getty Images)

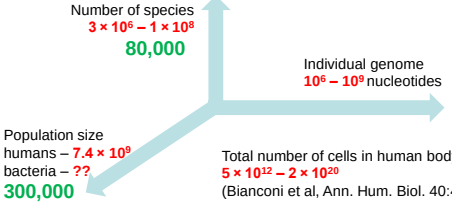
Data management and utilization

An early example of biological data depository:



Cave painting: Ennedi Plateau, Chad, ca. 7,000 BCE (Encyclopædia Britannica)

Genomic data

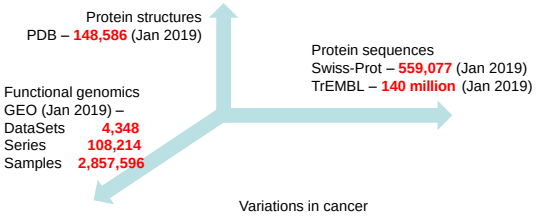


Total number of cells in human body
$5 \times 10^{12} – 2 \times 10^{20}$
(Bianconi et al, Ann. Hum. Biol. 40:463, 2013)

Total gut microbiota
$3 \times 10^{13} – 4 \times 10^{14}$
(Suau et al, Appl Environ Microbiol. 65:4799, 1999)

Recent estimate
Human: 3×10^{13}
Bacterial: 4×10^{13}
(Sender et al, Cell 164:337, 2016)

Genomic data

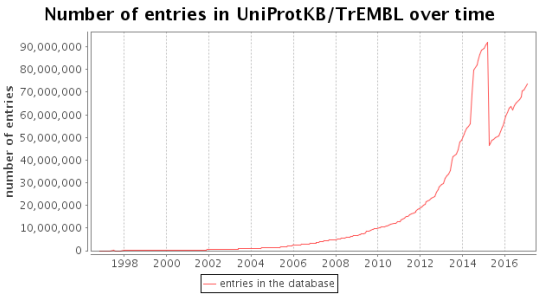


Variations in cancer

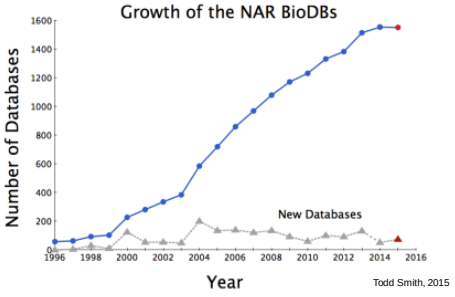
COSMIC (Nov 2018) –	
Samples	1,403,267
Coding Mutations	5,992,260
Whole Genomes	35,490
Copy Number Variants	1,179,545
Gene Expression Variants	9,147,833

Molecular Databases

Nucleic acid sequences:	GenBank	(211 million, Dec 2018)
	WGS	(773 million, Dec 2018)
Protein sequences:	UniProtKB:	
	Swiss-Prot	(550 thousand, Jan 2019)
	TrEMBL	(140 million, Jan 2019)
Protein structures:	PDB	(148 thousand, Jan 2019)



NAR Molecular Databases



NAR 2016 database issue: **15** categories, **41** subcategories, **1685** databases

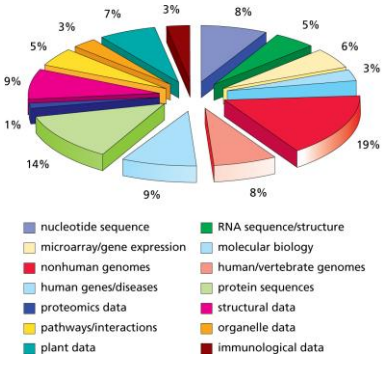
NAR Molecular Databases

NAR Database Summary Paper Category List

Nucleotide Sequence Databases
- International Nucleotide Sequence Database Collaboration
- Coding and non-coding DNA
- Gene structure, introns and exons, splice sites
- Transcriptional regulator sites and transcription factors

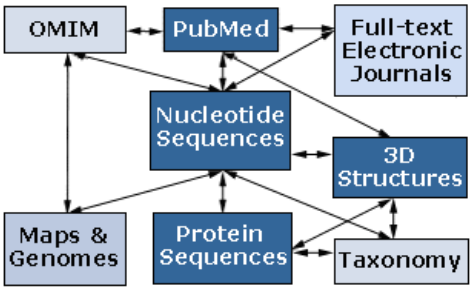
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

Molecular Databases



NCBI integrated search and retrieval system

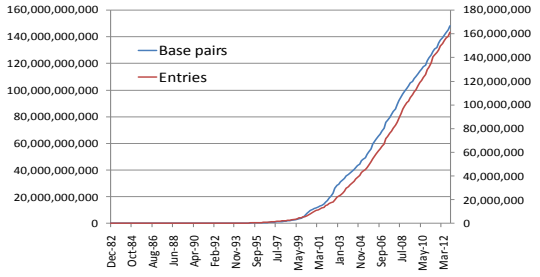
Entrez



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

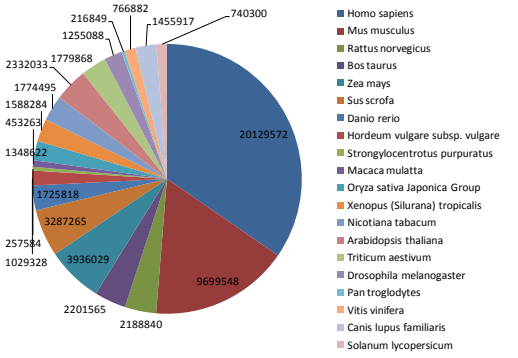
Growth of GenBank



December 1982
680,338 bp
606 seq

December 2012
148,390,863,904 bp
161,140,325 seq

GenBank Selected Per-Organism Statistics



NCBI Databases

Table 1. The Entrez databases (as of 3 September 2013)

Database	Records	Section within this article	Data source
NCBI Web Site	21 929	Introduction	N
PubMed	23 052 796	Literature	C
PMC	2 836 592	Literature	D, C
NLM Catalog	1 485 089	Literature	C, N
MeSH	243 770	Literature	N
Books	222 232	Literature	C, N
Taxonomy[a]	1 153 795	Taxonomy	C, N
Nucleotide[a]	101 599 766	DNA and RNA	D (GenBank), C, N
EST[a]	74 911 096	DNA and RNA	D (GenBank)
GSS[a]	36 959 049	DNA and RNA	D (GenBank)
BioSample	2 100 817	DNA and RNA	N
SRA[a]	475 684	DNA and RNA	D
PopSet[a]	183 110	DNA and RNA	D (GenBank)
Protein[a]	94 102 424	Proteins	C, N
Protein Clusters[a]	382 691	Proteins	N
GEO Profiles[a]	91 392 791	Genes and expression	D
Probe	31 367 498	Genes and expression	D
Gene[a]	14 167 800	Genes and expression	C, N
UniGene[a]	6 467 085	Genes and expression	N
GEO Datasets[a]	1 044 344	Genes and expression	N
Biosystems[a]	522 277	Genes and expression	C
Homologene[a]	133 548	Genes and expression	N
Clone[a]	33 135 797	Genomes	D, N
UniSTS[a]	545 913	Genomes	D (dbSTS)
BioProject[a]	98 358	Genomes	D
Assembly	17 707	Genomes	C, N
Genome[a]	10 929	Genomes	C, N

NCBI Databases (cont.)

MedGenEpigenomics[a]	10 811	Genomes	D
SNP[a]	300 258 943	Genetics and medicine	D (dbSNP), N
dbVar[a]	3 584 019	Genetics and medicine	D
MedGen[a]	169 433	Genetics and medicine	C, N
dbGaP	154 971	Genetics and medicine	D
ClinVar[a]	49 040	Genetics and medicine	D, N
PubMed Health	41 262	Genetics and medicine	C
GTR[a]	29 212	Genetics and medicine	D
OMIA	2844	Genetics and medicine	C
PubChem Substance[a]	119 813 846	Chemicals and bioassays	D
PubChem Compound[a]	47 757 896	Chemicals and bioassays	N
PubChem Bioassay[a]	717 429	Chemicals and bioassays	D
Structure[a]	92 993	Domains and structures	C, N
CDD[a]	48 034	Domains and structures	C, N

[a]Indicates that the data in this resource are available by FTP.
D, direct submission; C, collaboration/agreement; N, internal NCBI/NLM curation.

doi:10.1093/nar/gkt1146

Derivative databases

Systems	Year	Major features
Ranking search results		
PeMed	2010	Featuring multi-level relevance feedback for ranking
Querrle	2009	Allowing searches with concept categories
MedlineRanker	2009	Finding relevant documents through classification
MiSearch	2009	Using implicit feedback for improving ranking
Hakia	2008	Powered by Hakia's proprietary semantic search technology
SemanticMEDLINE	2008	Powered by cognition's proprietary search technology
MiScanner	2008	Finding relevant documents through classification
eTBLAST	2007	Finding documents similar to input text
PubFocus	2006	Sorting by impact factor and citation volume
Twease	2005	Query expansion with relevance ranking technique
Clustering results into topics		
Anne O'Tate	2008	Clustering by important words, topics, journals, authors, etc.
McSyBi	2007	Clustering by MeSH or UMLS concepts
GoPubMed	2005	Clustering by MeSH or GO terms
ClusterMed	2004	Clustering by MeSH, title/abstract, author, affiliation, or date
XplorMed	2001	Clustering by extracted keywords from abstracts
Extracting and displaying semantics and relations		
MedEvi	2008	Providing textual evidence of semantic relations in output
EBIMed	2007	Displaying proteins, GO annotations, drugs and species
CiteXplore	2006	EBI's tool for integrating biomedical literature and data
MEDIE	2006	Extracting text fragments matching queried semantics
PubNet	2005	Visualizing literature-derived network of bio-entities
Improving search interface and retrieval experience		
iPubMed	2010	Allow fuzzy search and approximate match
PubGet	2007	Retrieving results in PDFs
BabelMeSH	2006	Multi-language search interface
HubMed	2006	Export data in multiple format; visualization; etc
askMEDLINE	2005	Converting questions into formulated search as PICO
SLIM	2005	Slider interface for PubMed searches
PICO	2004	Search with patient, intervention, comparison, outcome
PubCrawler	1999	Alerting users with new articles based on saved searches

Lu, 2011, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3025693/pdf/baq036.pdf

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 (10), 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            707..1774
                     /gene="luxA"
     CDS             707..1774
                     /gene="luxA"
                     /codon_start=1
                     /product="luciferase alpha subunit"
                     /db_xref="PID:g155175"
                     /transl_table=11
                     /translation="MKFGNFLLTYQPPELSQTEVMKRLVNLGKASEGCGFDTVWLLEH
                     HFTEFGLLGNPYVAAAHLLGATETLNVGTAAIVLPTAHPVRQAEDVNLLDQMSKGRFR
                     FGICRGLYDKDFRVFGTDMDNSRALMDCWYDLMKEGFNEGYIAADNEHIKFPKIQLNP
                     SAYTQGGAPVYVVAESASTTEWAAERGLPMILSWIINTHEKKAQLDLYNEVATEHGYD
                     VTKIDHCLSYITSVDHDSNRAKDICRNFLGHWYDSYVNATKIFDDSDQTKGYDFNKGQ
                     WRDFVLKGHKDTNRRIDYSYEINPVGTPEECIAIIQQDIDATGIDNICCGFEANGSEE
                     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 aaaacattgg cttgattgcg gcaagtctct cagctcgtgt cgcctatgaa
      121 gttatctctg atctggagct gtcttttctg attactgcgg ttggtgtggt gaacttgcgt
      181 gacacactag aaaaagcgct tggttttgat tacctcagtt tgcctatcga tgagctacca
          ....
```

Example of a Pubmed entry

```
PMID- 15887224
OWN - NLM
STAT- MEDLINE
DA  - 20050718
DCOM- 20060522
IS  - 1097-0134 (Electronic)
IS  - 0887-3585 (Linking)
DP  - 2005 Aug 15
TI  - New method for protein secondary structure assignment based on a simple
      topological descriptor.
PG  - 513-24
AB  - A simple, five-element descriptor, derived from the Delaunay tessellation of a
      protein structure in a single point per residue representation, can be assigned
      to each residue in the protein. The descriptor characterizes main-chain topology
      and connectivity in the neighborhood of the residue and does not explicitly
      depend on putative hydrogen bonds or any geometric parameter, including bond
      length, angles, and areas. Rules based on this descriptor can be used for
      accurate, robust, and computationally efficient secondary structure assignment
      that correlates well with the existing methods.
AD  - Laboratory for Structural Bioinformatics, School of Computational Sciences,
      George Mason University, Manassas, Virginia 20110, USA.
FAU - Taylor, Todd
AU  - Taylor T
FAU - Rivera, Margarita
AU  - Rivera M
FAU - Wilson, Glenda
AU  - Wilson G
FAU - Vaisman, Iosif I
AU  - Vaisman II
LA  - eng
PT  - Journal Article
PT  - Research Support, U.S. Gov't, Non-P.H.S.
PL  - United States
```

Example of a Pubmed entry

```
TA  - Proteins
JT  - Proteins
JID - 8700181
RN  - 0 (Proteins)
SB  - IM
MH  - Algorithms
MH  - Amino Acid Sequence
MH  - Biophysical Phenomena
MH  - Biophysics
MH  - Computational Biology/*methods
MH  - Computer Simulation
MH  - Hydrogen Bonding
MH  - Models, Chemical
MH  - Models, Molecular
MH  - Molecular Sequence Data
MH  - Protein Conformation
MH  - Protein Folding
MH  - Protein Structure, Secondary
MH  - Protein Structure, Tertiary
MH  - Proteins/chemistry
MH  - Proteomics/*methods
MH  - Sensitivity and Specificity
MH  - Software
MH  - Structural Homology, Protein
EDAT- 2005/05/12 09:00
MHDA- 2006/05/23 09:00
CRDT- 2005/05/12 09:00
```

ID Converters

Features of mapping services	Gene/Clone ID converter	ID mapping by UniProt	MatchMiner	DAVID gene ID conversion tool
Interface	Web-based GUI form	Web-based GUI form	Web-based GUI form, command line	Web-based GUI form
Output format	Html, text, spreadsheet	Html, text	Html, text, spreadsheet	Html, text, spreadsheet
Organisms	Human, mouse, rat	Human, mouse, rat and many other species	Human, mouse	Human, about another 90,000 species
Input/output gene	HUGO gene names, Entrez gene Ids, Ensembl gene Ids, UniGene cluster Ids, RefSeq RNAs (Additional output: CCDS)[a]	Entrez Gene, HGNC, Ensembl, UniGene, TIGR (JCVI)	Gene Symbol HUGO/Alias, Name, UniGene Cluster Id, Entrez Gene Id, RefSeq RNA	Entrez gene Id, Ensembl gene/transcript Id, RefSeq mRNA accession, UniGene Id
Input/output protein	RefSeq peptides, SwissProt names (Additional output: IPI, PDB)*	UniProtKB, RefSeq, GenPept, IPI, PDB	RefSeq protein	PIR accession, PIR Id, PIR NREF Id, RefSeq Protein accession, Uniprot Id/accession, UniRef Id

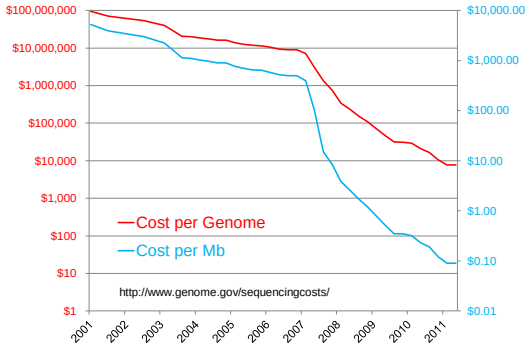
Chavan et al., 2011

ID Converters

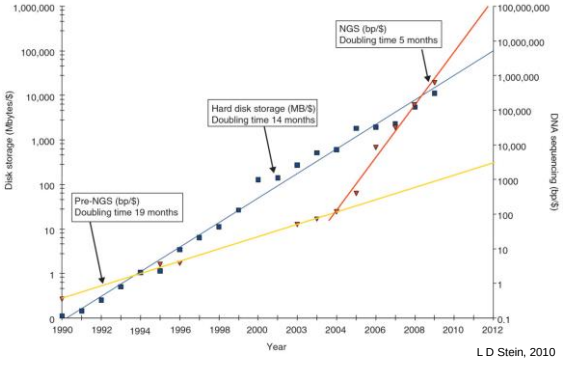
Mapping services	Link
Gene/Clone ID converter	http://idconverter.bioinfo.cnio.es/
ID mapping by UniProt	http://www.uniprot.org/?tab=mapping
MatchMiner	http://discover.nci.nih.gov/matchminer/
DAVID gene ID conversion tool	http://david.abcc.ncifcrf.gov/conversion.jsp
g:Convert	http://biit.cs.ut.ee/gprofiler/gconvert.cgi
CRONOS	http://mips.helmholtz-muenchen.de/genre/proj/cronos/
bioDBnet:db2db	http://biodbnet.abcc.ncifcrf.gov/db/db2db.php

Chavan et al., 2011

Genome sequencing costs



Sequencing and storage cost



L D Stein, 2010