

BINF 630: Bioinformatics Methods

Iosif Vaisman

Email: ivaisman@gmu.edu

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

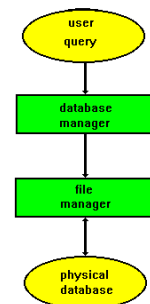
Database



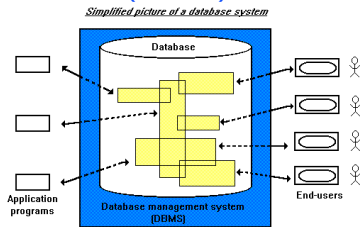
Database

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

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

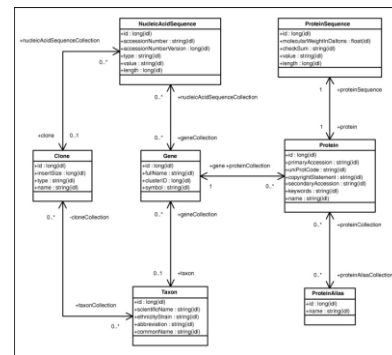
- one to one
- one to many
- many to one

Relational Database Model

protab1			
Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

protab2	
Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLSPRTVNQWLMLALFFGHS...
P1002	MDKTSHGFEIKLLTPKKLQQLMIAIYFGHT...
P1003	SRTHEEKGKLMQWPPRPPLYIALFTPEPPYP...
.....	

Object-oriented Database Model

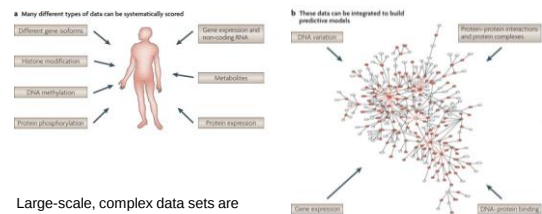


G A Komatsoulis et al., 2008

Database administration

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

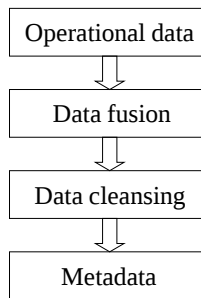
Data integration



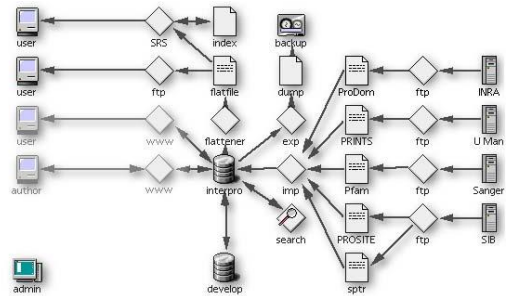
Large-scale, complex data sets are shown as a network in which the nodes represent variables of biological interest, such as DNA variation, RNA variation, protein levels, protein states, metabolite levels and disease-associated traits, and the edges between these nodes represent causal relationships between the variables.

Schadt et al., 2010

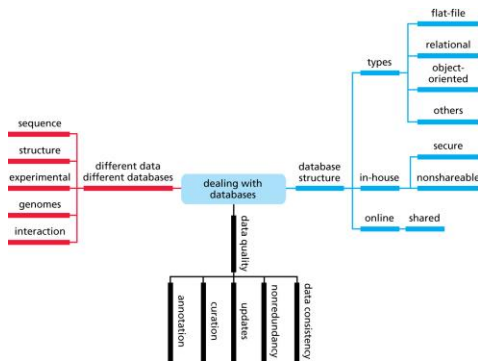
Data Warehouse



InterPro Database



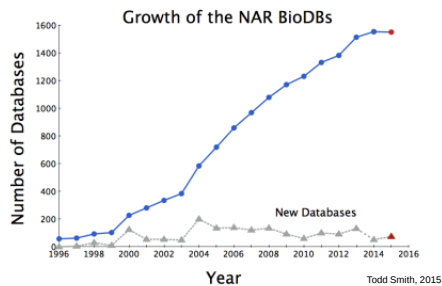
Databases in Bioinformatics



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 Molecular Databases



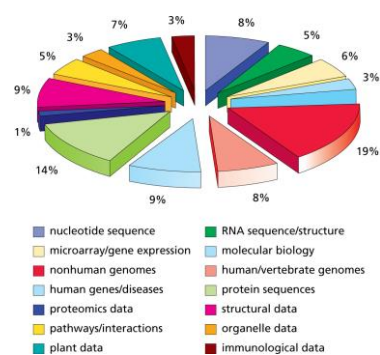
NAR 2016 database issue: 15 categories, 41 subcategories, 1685 databases

NAR Molecular Databases

NAR Database Summary Paper Category List

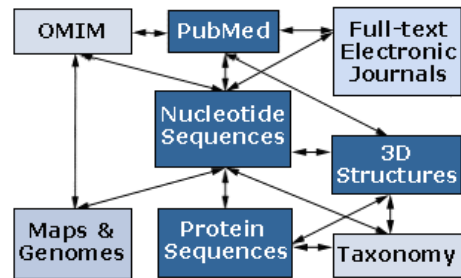
- Nucleotide Sequence Databases
 - International Nucleotide Sequence Database Collaboration
 - Coding and non-coding DNA
 - Gene structure, introns and exons, splice sites
 - Transcriptional regulator sites and transcription factors
- RNA sequence databases
- Protein sequence databases
- Structure Databases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

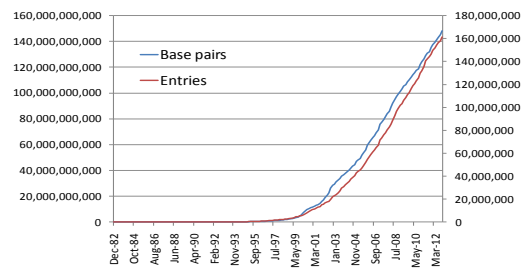
Entrez



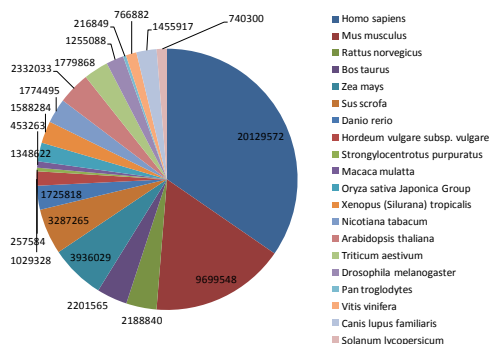
NCBI Databases

- 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



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	28 366 592	Literature	D, C
NLM Catalog	1 485 089	Literature	C, N
MeSH	243 770	Literature	N
Books	222 232	Literature	C, N
Taxonomy ^a	1153 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	21 000 817	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	1044 344	Genes and expression	C, 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.)

MadGenEpiGenomics*	10811	Genomes	D
SNP*	300 258 943	Genetics and medicine	D (dbSNP), N
dbVar*	3584 019	Genetics and medicine	D
MedGen*	169 433	Genetics and medicine	C, N
dbGaP	154 971	Genetics and medicine	D
ClinVar*	49 040	Genetics and medicine	D, N
PubMed Health	41 262	Genetics and medicine	C
GTR*	29 212	Genetics and medicine	D
OMIA	2844	Genetics and medicine	C
PubChem Substance*	119 813 846	Chemicals and bioassays	D
PubChem Compound*	47 757 896	Chemicals and bioassays	N
PubChem Bioassay*	717 429	Chemicals and bioassays	D
Structure*	92 593	Domains and structures	C, N
CDD*	48 404	Domains and structures	C, N

*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		
Refined	2010	Featuring multi-level relevance feedback for ranking
Quartile	2009	Allowing searches with concept categories
MedlineRanker	2009	Finding relevant documents through classification
MiscRank	2009	Using implicit feedback for improving ranking
Halia	2008	Powered by Halia's proprietary semantic search technology
SemanticMEDLINE	2008	Powered by cognihon's proprietary search technology
MiscRanker	2008	Finding relevant documents through classification
eBLAST	2007	Finding documents similar to input text
PubFocus	2006	Sorting by impact factor and citation volume
Treasure	2005	Query expansion with relevance ranking technique
Clustering results into topics		
Anne O'Tate	2008	Clustering by important words, topics, journals, authors, etc.
MediBy	2007	Clustering by MeSH or UMLS concepts
GIRMed	2005	Clustering by MeSH or GO terms
ClusterMed	2004	Clustering by MeSH, title/abstract, author, affiliation, or date
igRankMed	2001	Clustering by extracted keywords from abstracts
Extracting and displaying semantics and relations		
MedEvi	2008	Providing textual evidence of semantic relations in output
EBMed	2007	Displaying proteins, GO annotations, drugs and species
ChiroCore	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		
PubMed	2010	Allow fuzzy search and approximate match
PubGet	2007	Retrieving results in PDFs
bioRxivSH	2006	Multi-language search interface
HubMed	2006	Export data in multiple format; visualization; etc
ashMEDLINE	2005	Converting questions into formulated search as PICO
SLIM	2005	Sliding 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        gl55174
KEYWORDS   luciferase.
SOURCE     Vibrio harveyi DNA.
ORGANISM   Vibrio harveyi
            Eubacteria; Proteobacteria; gamma subdivision; Vibrionaceae;
            Vibrio.
REFERENCE  1 (bases 1 to 1938)
AUTHORS   Cohn,D.H., Mileham,A.J., Simon,M.I., Nealson,K.H., Rausch,S.K.,
            Bonan,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:gl55175"
                     /transl_table=11
                     /translation="MKGFGNFLTYPPELSQTEVMKRLVNLGHASRGCGFDTWLLHH
HTEPGGLGNFYVAARHLGATETLNVGTAAIVLPFAHPVRQAEVDNLLDQMSKGRFR
FGICRGLYDKDFRVFGTMDMNSRALDCWYDLKSGFNGEYIADNEHIFFKIQLNP
SAYTQGGAPVYVAESASTTEWAERGLPMLSWIINTEHKAQLDLNEVATEHGVD
VTKIDHCLSYITSVDHDSNRKADICNFFLGRWYDSYNATKIFDSDSQTKGYDFNKGQ
WRDFVLKHKOTNRIRIDSYEINPVGTEPCETAIQQDDIDATGIDNICGFEANGSEE
EIIASMLFQSDVMFPLKEKQ"
BASE COUNT      883 a      665 c      741 g      852 t
ORIGIN            1 bp upstream of EcoRI site.
1   gaattcaccac  tgacgacggcg  caaaaatatg  ttgtgcactg  ttatcatcgt  gctgcagacc
61  aaaggccacac  aaaaacattg  ctgtgatggc  gcaagttctt  cagctcgtgt  cgcctatgaa
121  gttattctctg  atctggagct  gttctttctg  attactcgcg  ttggtgtggt  gaacttcgct
181  gacacactagc  aaaaagcgct  tggttttgat  tacctcagtt  tgcctatcga  tgagctacca
.....
```

Example of a Pubmed entry

```
PMID- 15887224
OMW - 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 E
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 - IN
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
EADT- 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)*	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

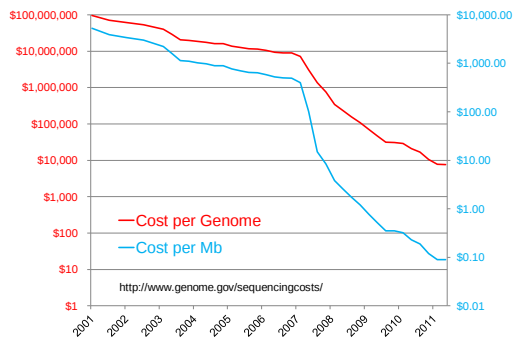
Chavan et al., 2011

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://bit.cs.ut.ee/gprofiler/gconvert.cgi
CRONOS	http://mips.helmholtz-muenchen.de/genel/proj/cronos/
bioDBnet:db2db	http://biobdnet.abcc.ncifcrf.gov/db/db2db.php

Chavan et al., 2011

Genome sequencing costs



Sequencing and storage cost

