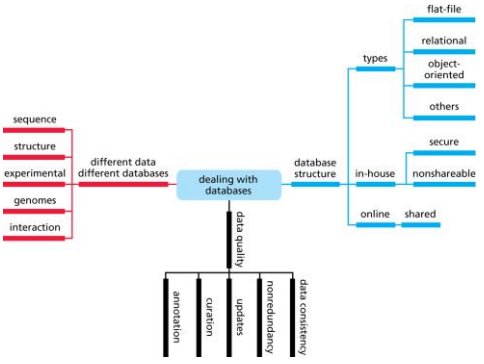
BINF 630: Bioinformatics Methods

Iosif Vaisman

Email: ivaisman@gmu.edu

Databases in Bioinformatics

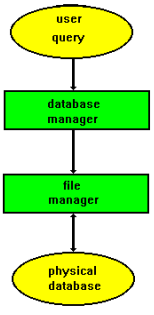


Database

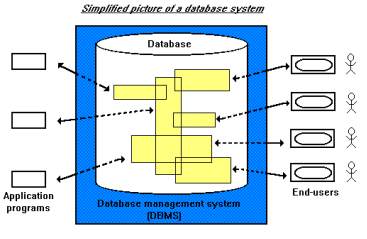
Database

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

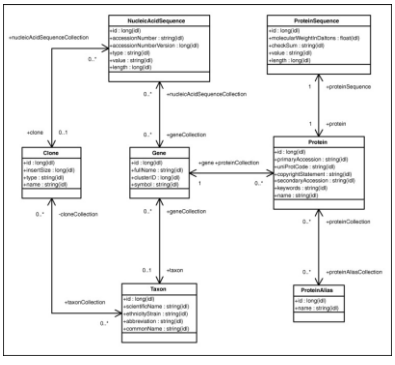
protab1

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	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS…
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT…
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP…
.....	

Object-oriented Database Model

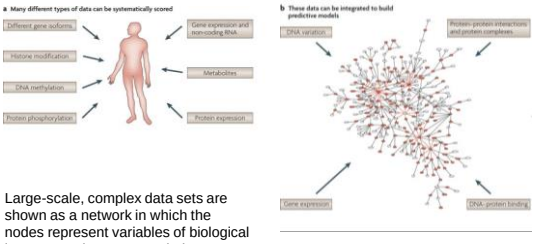


G A Komatsoulis et al., 2008

Database administration

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

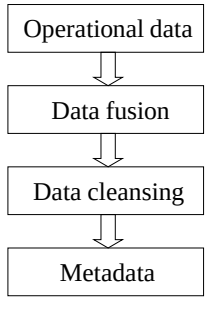
Data integration



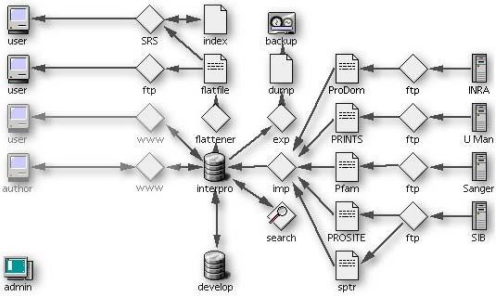
Large-scale, complex data sets are shown as a network in which the nodes represent variables of biological interest, such as DNA variation, RNA variation, protein levels, protein states, metabolite levels and disease-associated traits, and the edges between these nodes represent causal relationships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

Schadt et al., 2010

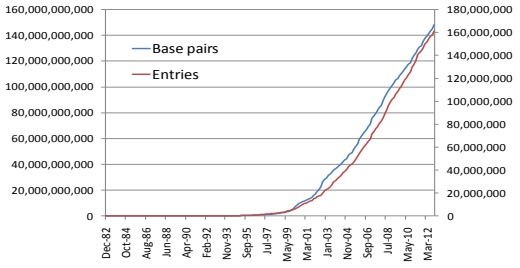
Data Warehouse



InterPro Database



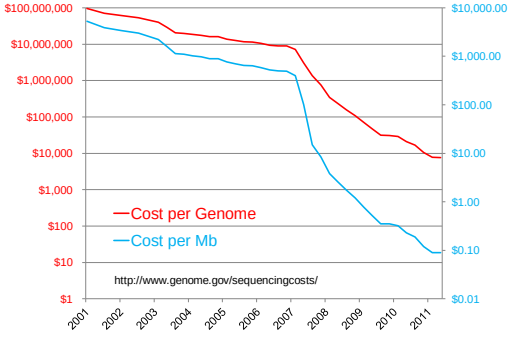
Growth of GenBank



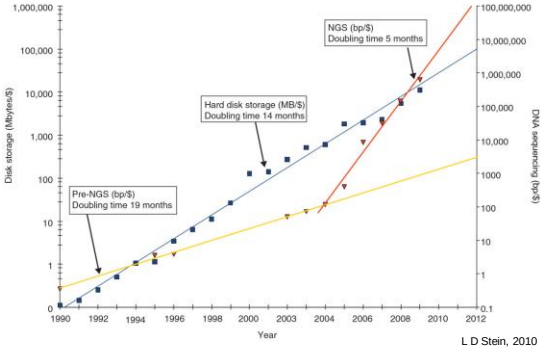
December 1982
680,338 bp
606 seq

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

Genome sequencing costs



Sequencing and storage cost



L D Stein, 2010

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
```