"In the 21st century, the database is the marketplace."
-- Stan Rapp

"To a database person, every nail looks like a thumb. Or something like that."
-- Jamie Zawinski

Biology Collects Lots of Data

- Trillions bases of DNA sequence
- Hundreds of thousands of species
- Millions of articles in scientific journals
- Genetic information:
  - gene names (thousands)
  - phenotype of mutants (infinite?)
  - location of genes/mutations on chromosmes
  - linkage (distances between genes)

Biological Databases:

- Data Domains
- Types of Databases - By Scope
- Types of Databases - By Level of Curation
  - »GenBank
  - »RefSeq

NCBI

Acknowledgement: The presentation includes adaptations from NCBI's Introduction to Molecular Biology Information Resources Modules

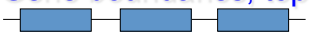
Data Domains

- **Types of data generated by molecular biology research:**
  - nucleotide sequences (DNA and mRNA)
  - protein sequences
  - 3-D protein structures
  - complete genomes and maps
- **Also now have:**
  - gene expression
  - genetic variation (polymorphisms)

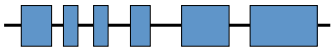
Curated Biological Data

DNA, nucleotide sequences

Gene boundaries, topology

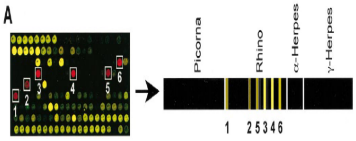


Gene structure

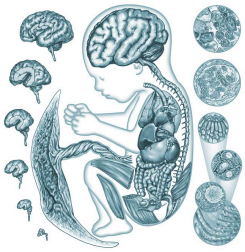


Introns, exons, ORFs, splicing

Expression data



Cell types



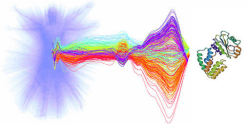
Curated Biological Data

Proteins, residue sequences

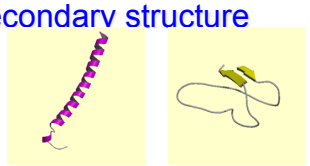
Extended sequence information

MCTUYTCUYFSTYRCCTYFSCD

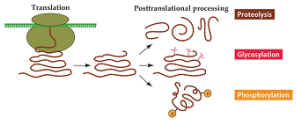
Mass spectometry (metabolomics, proteomics)



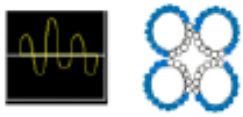
Secondary structure



Post-Translational protein Modification (PTM)

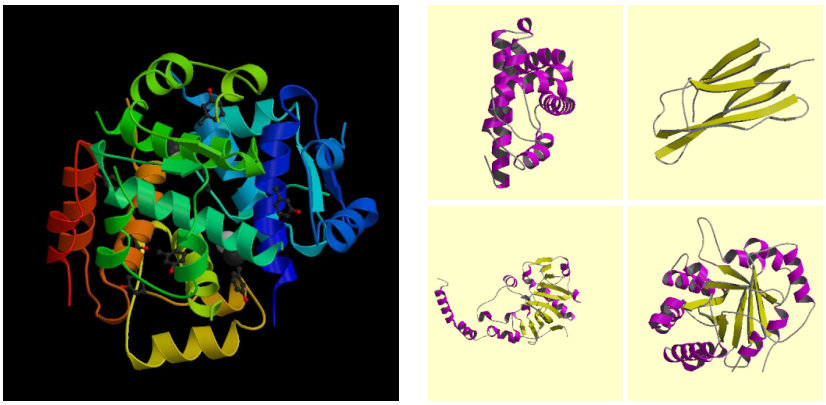


Hydrophobicity, motif data



Protein-protein interaction

Curated Biological data
3D Structures, folds



WHAT is a database?

- A collection of data that needs to be:
  - Structured
  - Searchable
  - Updated (periodically)
  - Cross referenced

- Challenge:
  - To change “meaningless” data into useful information that can be accessed and analysed the best way possible.

For example:

HOW would YOU organize all biological sequences so that the biological information is optimally accessible?

A Spreadsheet can be a Database

- columns are **Fields**
- Rows are **Records**
- Can search for a term within just one field
- Or combine searches across several fields

SNP ID	SNPSeq ID	Gene	+primer	-primer	Hap A	Hap B	Hap C
D1Mit160_1	10.MMHAP67FLD1.seq	lymphocyte antigen 84	AAGGTAAAAGGCAATCAGCACAGCC	TCAACCTGGAGTCAGAGGCT	C	—	A
M-05554_1	12.MMHAP31FLD3.seq	procollagen, type III, alpha	TGCGCAGAAGCTGAAGTCTA	TTTTGAGGTGTTAATGGTTCT	C	—	A
M-05554_2	X60184	complement component factor i	ACTTCCAGCCCTGGCTCT	ATATGCCACCAAGAAGCA	A	C	—
M-09947_3	AF067835	caspase 8	TCACAGAGGGAAACATGAAG	CTCCACATTGAACCAAAGCA	G	C	T
M-11415_1	U02023	insulin-like growth factor binding protein	GGGAAAAGCCTGAAAGAAGC	AGCTGAAACCGGACATCAAT	T	G	—
D1Mit284_3	J05234	nucleolin	TGTTGGAACCGACTTCTTCA	AAGAGTCAAAGAATTTATGGAATGA	G	T	T

Structured Data

- Repository of information
- managed and accessed differently
- Flat-file (text)
- Relational (key)
- “talk” to each other

(A)

NAME	TELEPHONE	ADDRESS
S. Claus	0203 450	The North Pole, Lapland
M. Mouse	0202 453	Disneyworld, Florida
A. Moonman	0104 459	Craterland, The Moon

(B) GenBank Flat-File Format

```
LOCUS       SCU49845     5028 bp    DNA
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and
            Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina;
            Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
```

protab1

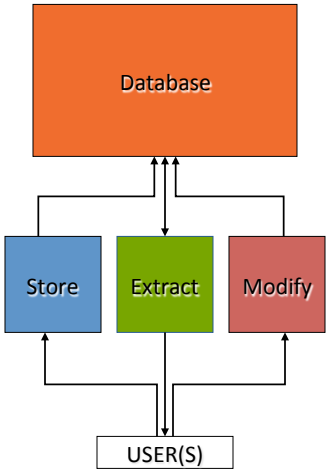
Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

protab2

Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS…
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT…
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP…
.....	

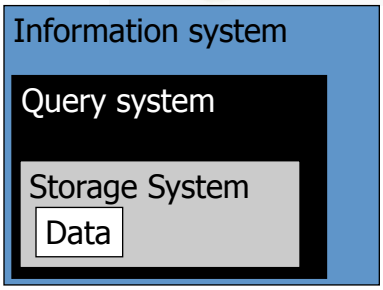
DBMS

- Internal organization
  - Controls speed and flexibility
- A unity of programs that
  - Store
  - Extract
  - Modify



DBMS organisation types

- Flat file databases (flat DBMS)
  - Simple, restrictive, table
- Hierarchical databases (hierarchical DBMS)
  - Simple, restrictive, tables
- Relational databases (RDBMS)
  - Complex,versatile, tables
- Object-oriented databases (ODBMS)
  - Complex, versatile, objects



Relational databases

- Data is stored in multiple related tables
- Data relationships across tables can be either many-to-one or many-to-many
- A few rules allow the database to be viewed in many ways

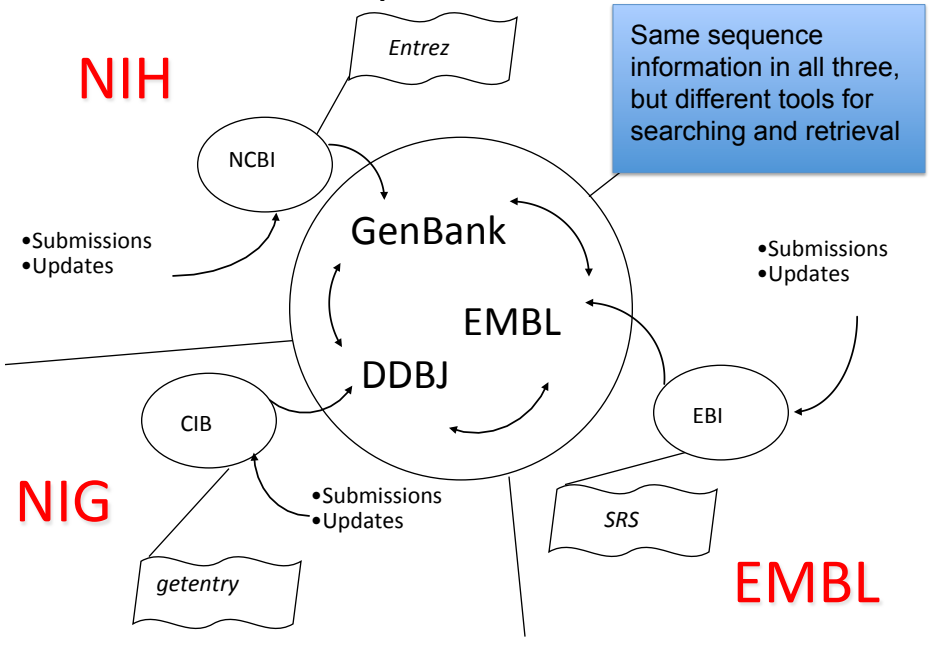
Three reasons to care …

- Database proliferation
  - hundreds to thousands at the moment
- More and more scientific discoveries result from inter-database analysis and mining
- Rising complexity of required data-combinations
  - E.g. translational medicine: “from bench to bedside” (genomic data vs. clinical data)

Types of Databases - By Scope

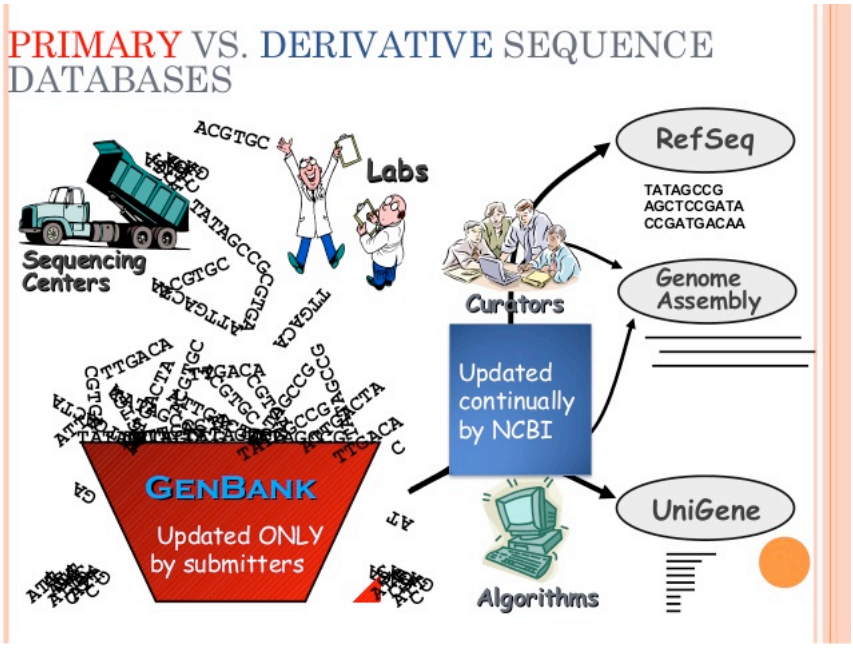
- **Comprehensive**
  - **Contain data from many organisms and many different types of sequences. Examples:**
  - Nucleotide
    - GenBank (overview)
    - EMBL: European Molecular Biology Laboratory
    - DDBJ: DNA Data Bank of Japan
      (The three databases above comprise the International Nucleotide Sequence Database Collaboration and currently include sequence data from >120,000 species.)
  - Protein, such as Swiss-Prot
  - Protein Structure, such as PDB: Protein Data Bank
  - Genomes and Maps, such as Entrez Genomes
- **Specialized**
  - **Contain data from individual organisms, specific categories/functions of sequences, or data generated by specific sequencing technologies.**

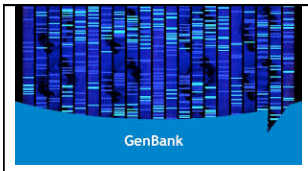
Public Sequence Databases



Types of Databases - By Level of Curation

- **Archival data**
  - repository of information
  - redundant; might have many sequence records for the same gene, each from a different lab
  - submitters maintain editorial control over their records: what goes in is what comes out
  - no controlled vocabulary
  - variation in annotation of biological features
- **Curated data**
  - non-redundant; one record for each gene, or each splice variant
  - each record is intended to present an encapsulation of the current understanding of a gene or protein, similar to a review article
  - records contain value-added information that have been added by an expert(s)



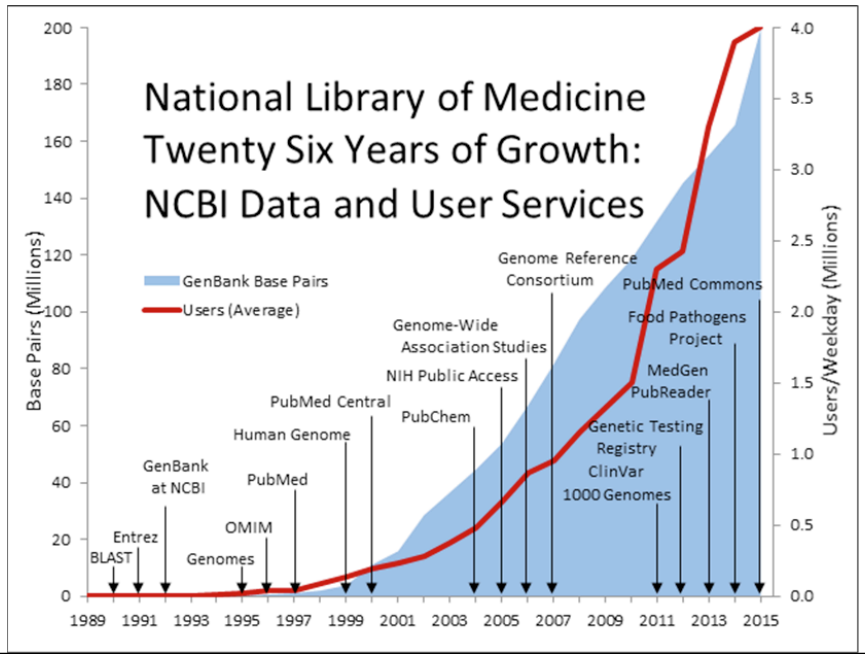


GenBank

- Contains **all** DNA and protein sequences described in the scientific literature or collected in publicly funded research
- Flatfile: Composed entirely of text

- Each submitted sequence is a record
- Had fields for Organism, Date, Author, etc.
- Unique identifier for each sequence
  - Locus and Accession #

Growth of Genbank



http://www.ncbi.nlm.nih.gov/Genbank

- Once upon a time, **GenBank** mailed out sequences on CD-ROM disks a few times per year.
- At least doubles in size every 18 months
- There are approximately 253,630,708,098 bases, from 207,040,555 reported sequences in the traditional GenBank divisions as of January 2018.

Distribution of sequence databases

- Books, articles 1968 -> 1985
- Computer tapes 1982 ->1992
- Floppy disks 1984 -> 1990
- CD-ROM 1989 -> ?
- FTP 1989 -> ?
- On-line services 1982 -> 1994
- WWW 1993 -> ?
- DVD 2001 -> ?
- Mailing hard drives 2009 -> ?

GenBank Flat File (GBFF)

Header
- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

Fields

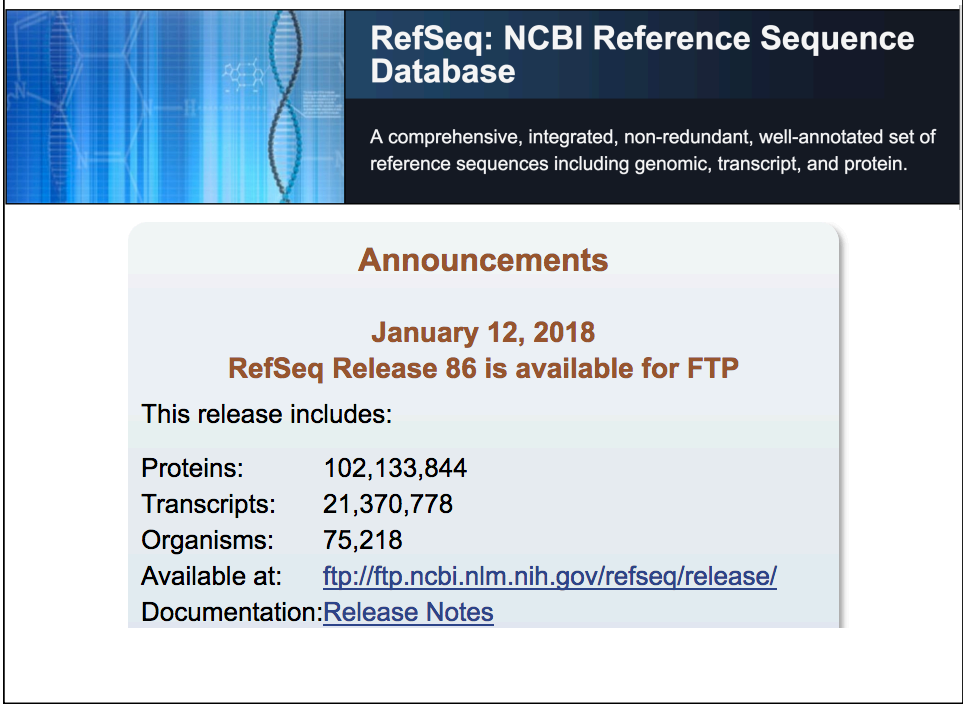
```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
NID         g1293613
VERSION     U49845.1  GI:1293613
KEYWORDS
SOURCE      baker's yeast.
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
            Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  MEDLINE   95176709
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  MEDLINE   96194260
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="PID:g1293614"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
```

Accession Numbers!!

- Databases are designed to be searched by accession numbers (and locus IDs)
- These are **guaranteed** to be non-redundant, accurate, and not to change.
- Searching by gene names and keywords is doomed to frustration and probable failure
- Neither scientists nor computers can be trusted to accurately and consistently annotate database entries!!

Last thoughts on Genbank ...

- Often only use FASTA files (eg for BLAST)
- GBFF are simply human readable versions of these records
- GBFF have become a vehicle for a lot more information than they where meant to do
- Keep in mind that GenBank is DNA centric and is a poor vehicle for protein and mRNA expression/ interaction information



A few words about RefSeq

- Many sequences in GenBank correspond to the same gene
  - genomic clones, full length mRNA, various kinds of ESTs, submitted by different investigators
- RefSeq is the "Reference Sequence" for a gene - as determined by GenBank curators
  - best guess given the current evidence, can change
  - usually based on the longest mRNA
  - usually has both 5' and 3' UTR
- A representative GenBank record is used as the source for a RefSeq record

A few more words about RefSeq

- Each record is intended to present an encapsulation of the current understanding of a gene or protein, similar to a review article
- RefSeq database includes genomic DNA, mRNA, and protein sequences. E.g. Organized by the central dogma.
- Not necessarily totally reliable
  - A lot is not yet known… eg, alternative splicing

RefSeq and Accessions

- Genomic DNA
  - NC_123456 - complete genome, complete chromosome, complete plasmid
  - NG_123456 - genomic region
  - NT_123456 - genomic contig
- mRNA - NM_123456
- Protein - NP_123456
- Gene and protein models from genome annotation projects:
  - XM_123456 - mRNA
  - XR_123456 - RNA (non-coding transcripts)
  - XP_123456 - protein

RefSeq Status Codes

- Level of manual curation
- Examples
  - Provisional
    - has not yet been subject to individual review and is thought to be well supported and to represent a valid transcript and protein
  - Reviewed
    - has been the reviewed by NCBI staff or by a collaborator
  - Predicted
    - is predicted and has not been subject to individual review
  - Genome Annotation
    - identifies RefSeq records provided by the NCBI Genome Annotation process

Many Datasets at NCBI

The NCBI hosts a huge interconnected database system that, in addition to DNA and protein, includes:

- Journal Articles (PubMed)
- Genetic Diseases (OMIM)
- Polymorphisms (dbSNP)
- Cytogenetics (CGH/SKY/FISH & CGAP)
- Gene Expression (GEO)
- Taxonomy
- Chemistry (PubChem)

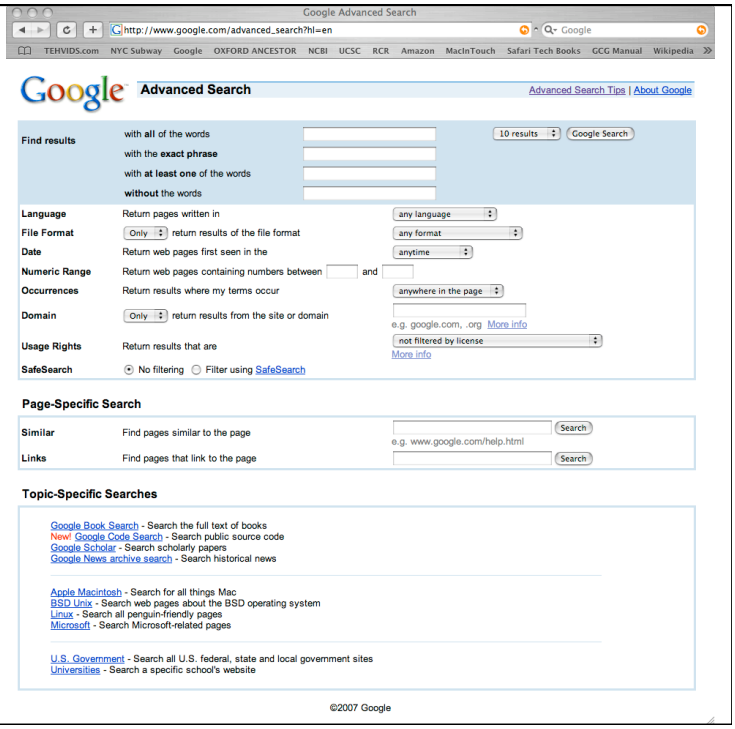
Accessing database information

- A request for data from a database is called a *query*
- *Queries* can be of three forms:
  - Choose from a list of parameters
  - Query by example (QBE)
  - Query language

Web Query

- Most databases have a web-based query tool
- It may be simple…

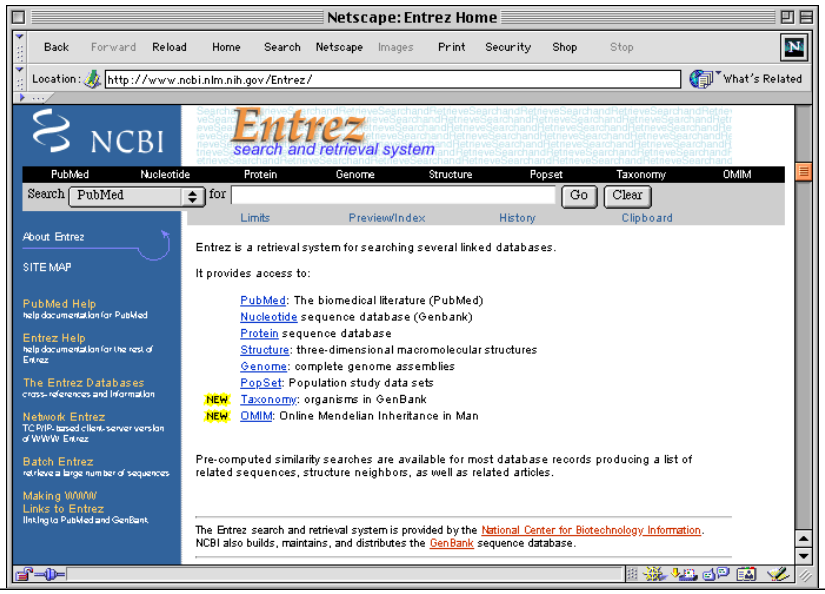
… or complex



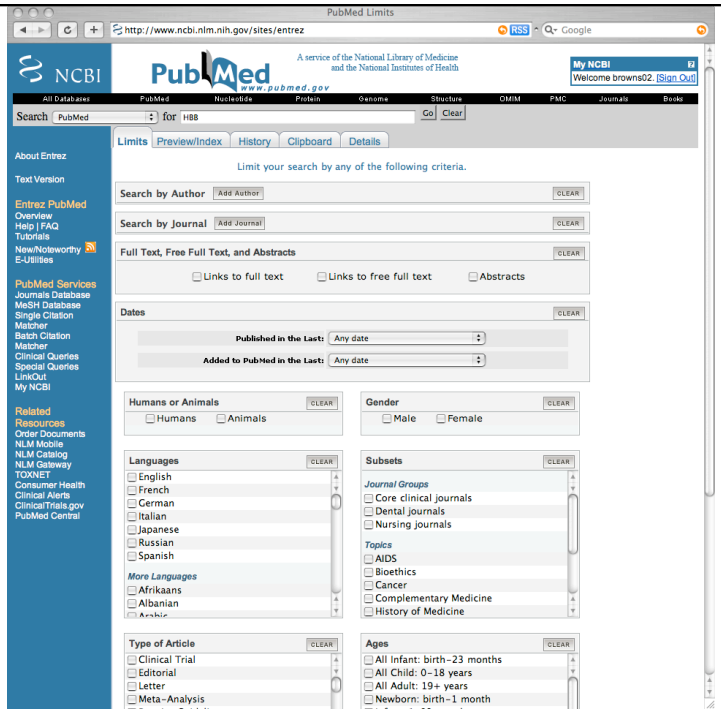
Query Languages

- The standard Relational Database Query Language
  - SQL (Structured Query Language) originally called SEQUEL (Structured English QUEry Language)
  - Developed by IBM in 1974; introduced commercially in 1979 by Oracle Corp.
  - Standard interactive and programming language for getting information from and updating a database.
  - RDMS (SQL), ODBMS (Java, C++, OQL etc)

ENTREZ is the GenBank web query tool



Advanced query interface:



Database Searching

A database can only be searched in ways that it was designed to be searched

Boolean: "AND" and "OR" searches

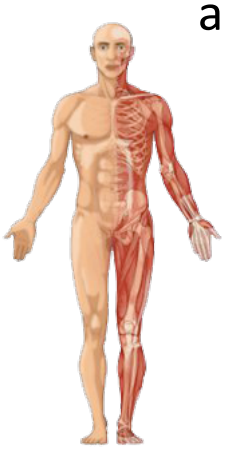
Bad to search for "human hemoglobin" in a '*Description*' field

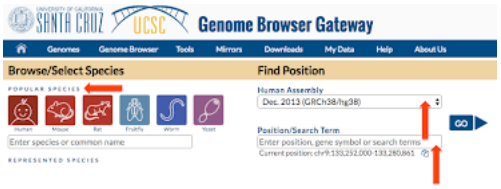
Much better to search for "homo sapiens in '*Organism*' AND "HBB" in *'gene name'*

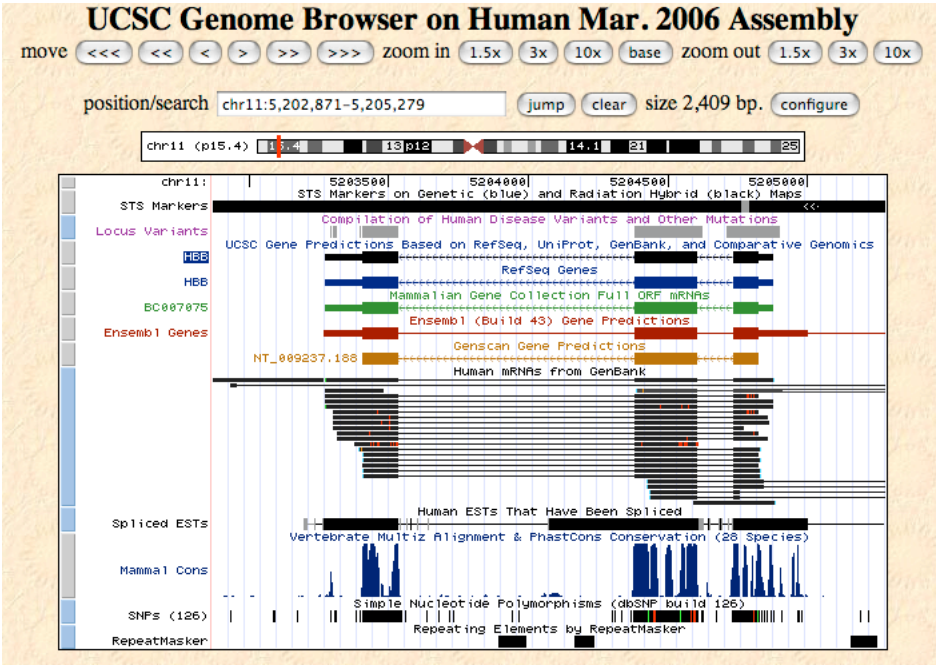
Strategies

- Use accession numbers whenever possible
- Start with broad keywords and narrow the search using more specific terms
- Try variants of spelling, numbers, etc.
- Search all relevant databases
- **Be persistent!!**

In addition to NCBI, there are specialized data resources by the big annotation consortia.



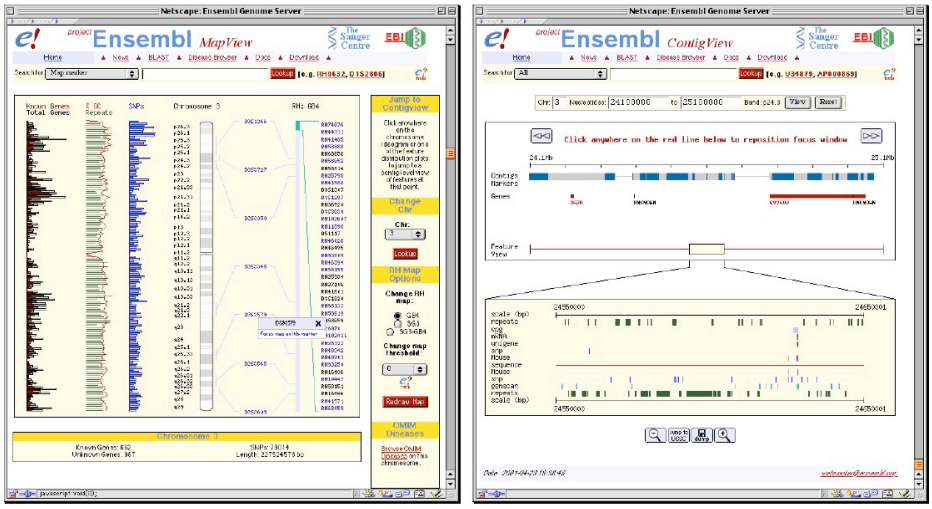




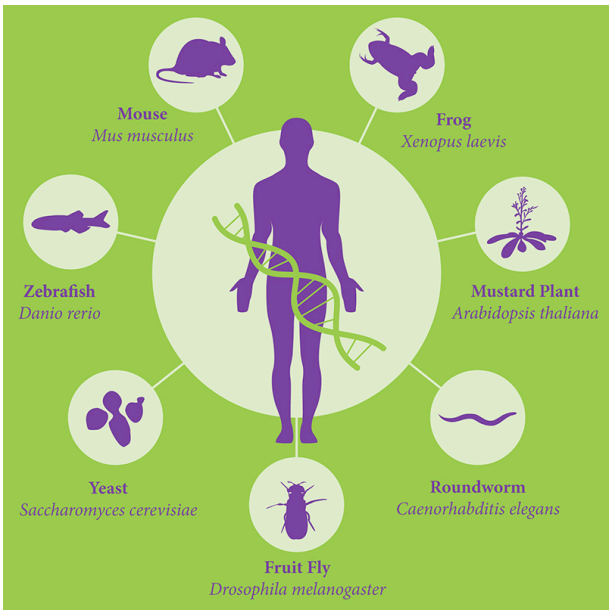
Lots of additional data can be added as optional "tracks"

- anything that can be mapped to locations on the genome

Ensembl at EBI/EMBL

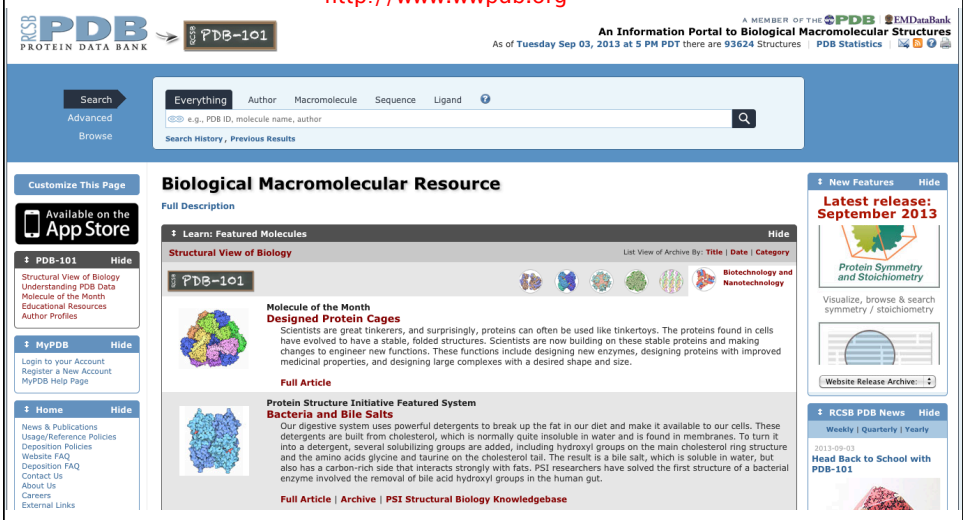


In addition to human data there are many model organisms.



RCSB PDB
PROTEIN DATA BANK

http://www.wwpdb.org



Specialized Databases

- More than 1000 different specialized databases
- Generally accessible through the web
  (useful link: www.expasy.ch/alinks.html)

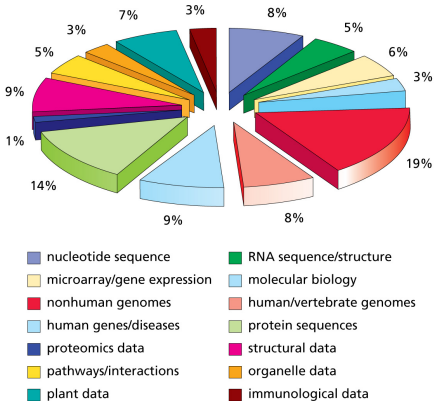
- Variable size: <100Kb to >10Gb
  - DNA: > 10 Gb
  - Protein: 1 Gb
  - 3D structure: 5 Gb
  - Other: smaller

- Update frequency: daily to annually

NAR Database Issue

- Online collection of biological databases:

http://www.oxfordjournals.org/nar/database/c/



Golden Rules

- Use published databases and methods
  - Supported, maintained, trusted by community
- ***Document what you have done !!!***
  - Sequence identification numbers
  - Server, database, program VERSION
  - Program parameters
- Assess reliability of results
- Check quality of ALL data before you use it. (Garbage in, garbage out)

Bio-databases: A short word on problems

- Even today we face some key limitations
  - There is no standard format
    - Every database or program has its own format
  - There is no standard nomenclature
    - Every database has its own names
  - Data is not fully optimized
    - Some datasets have missing information without indications of it
  - Data errors
    - Data is sometimes of poor quality, erroneous, misspelled
    - Error propagation resulting from computer annotation