

“In the 21<sup>st</sup> 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 chromosomes
  - linkage (distances between genes)

## Biological Databases:



- Data Domains
  - Types of Databases - By Scope
  - Types of Databases - By Level of Curation



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

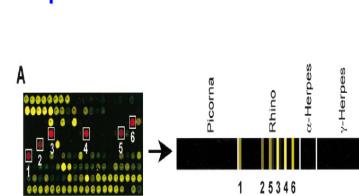


Gene structure

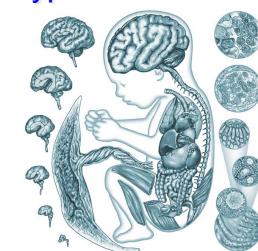


Introns, exons, ORFs, splicing

Expression data



Cell types



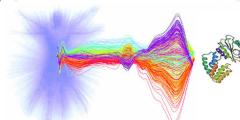
## Curated Biological Data

Proteins, residue sequences

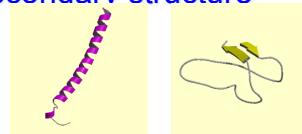
Extended sequence information

MCTUYTCUYFSTYRCCTYFSCD

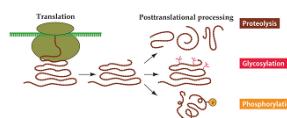
Mass spectrometry  
(metabolomics, proteomics)



Secondary structure



Post-Translational protein Modification (PTM)

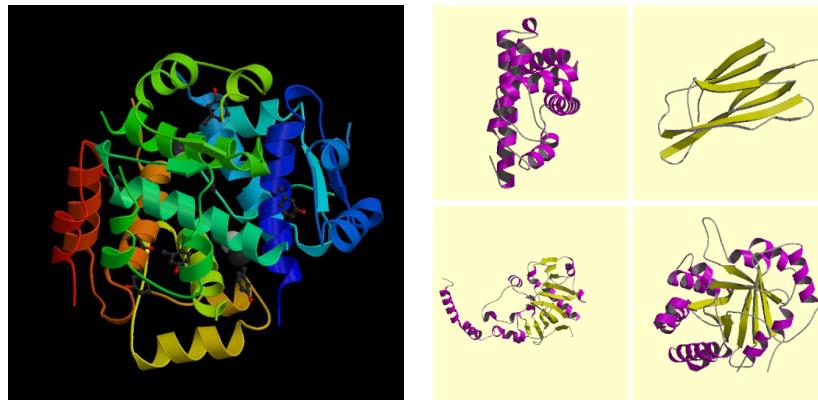


Hydrophobicity, motif data



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.MMHAP6 7FLD1.seq	lymphocyte antigen 84	AAGGTAAA GGCAATCAG CACAGCC	TCAACCTGG AGTCAGAGG CT	C	—	A
M-05554_1	12.MMHAP3 1FLD3.seq	procollagen, type III, alpha	TGCCAGAA GCTGAAGTC TA	TTTGAGGT GTTAATGGTT CT	C	—	A
M-05554_2	X60184	complement component factor I	ACTTCAGC CCTGGCTCT	ATATGCCACC AAGAAAGCA	A	C	—
M-09947_3	AF067835	caspase 8	TCACAGAGG GAACATGA AG	CTCCACATG AACCAAAGC A	G	C	T
M-11415_1	U02023	insulin-like growth factor binding protein	GGGAAAAGC CTGAAAGAA GC	AGCTGAAAC CGGACATCA AT	T	G	—
D1Mit284_3	J05234	nucleolin	TGTTGAAAC CGACTTCTTC A	AAGAGTCAA AGAATTATG GAATGA	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	0020 450	The North Pole, Lapland	
M. Mouse	0020 453	Disneyworld, Florida	
A. Moonman	0104 459	Craterland, The Moon	

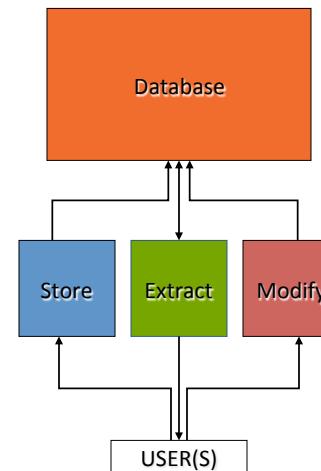
(B)	GenBank Flat-File Format						
LOCUS	SC049845	5028 bp	DNA				
DEFINITION	Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Ax1 genes (Ax1α2) and Rev7p (REV7) genes, complete cds.						
ACCESSION	U199845.1	GI:1293613					
VERSION	U199845.1	GI:1293613					
KEYWORDS							
SOURCE	Saccharomyces cerevisiae (baker's yeast)						
ORGANISM	Saccharomyces cerevisiae Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetidae; 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	MDRTTHGFDLKLLSPRTVNQWLMALFFGHS...
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT...
P1003	SRTHEEEGKLMQWPPLPYIALFTEPPYP...
....	

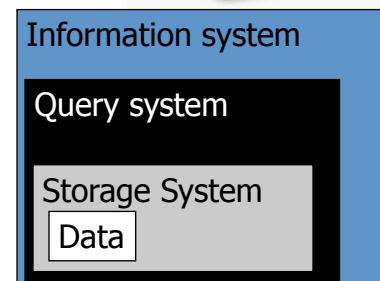
## DBMS

- 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)

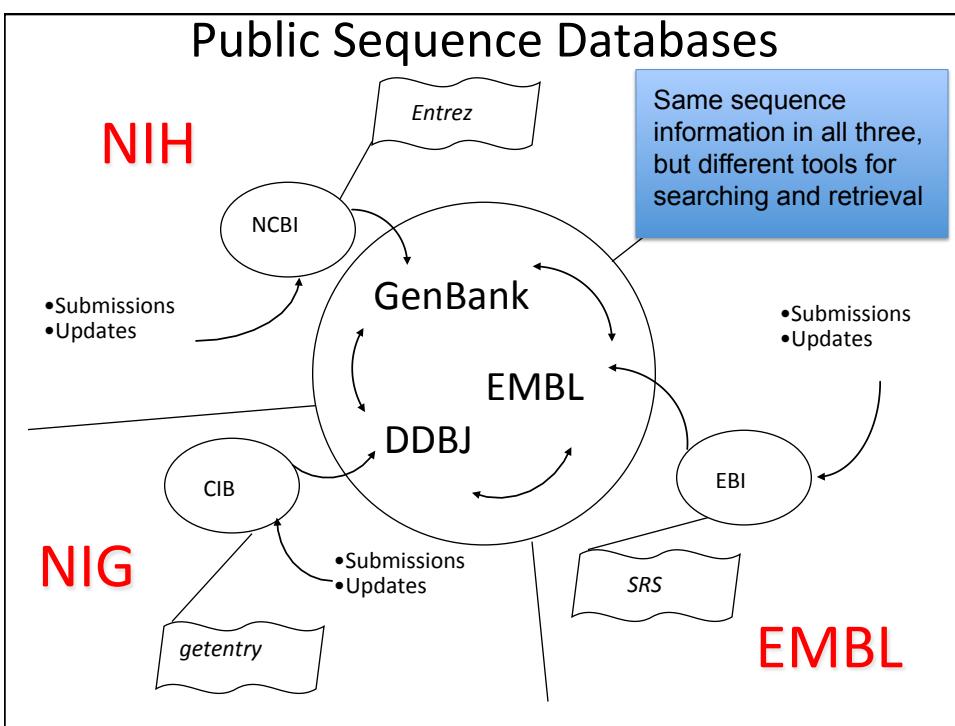
## Types of Databases - By Scope

- **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.



## Types of Databases

### - By Level of Curation

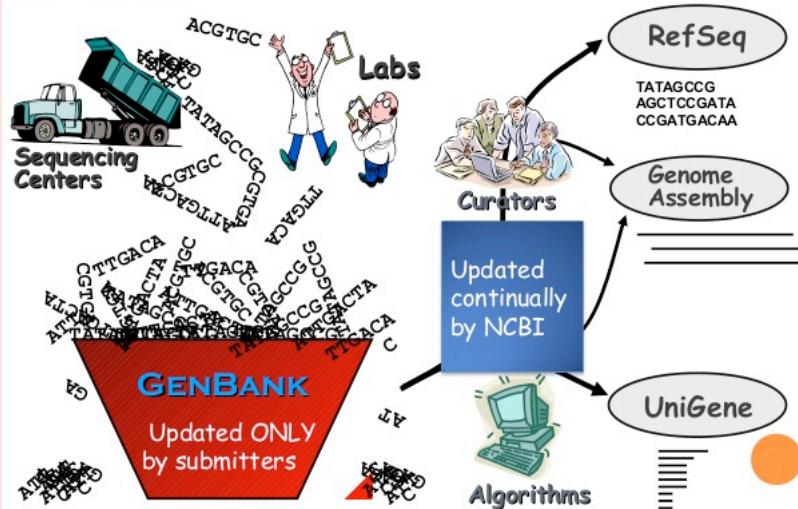
- **Archival data**

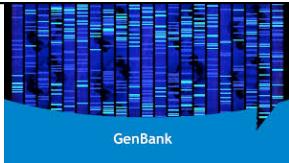
- repository of information
- redundant; might have many sequence records for the same gene, each from a different lab
- submitters maintain editorial control over their records: what goes in is what comes out
- no controlled vocabulary
- variation in annotation of biological features

- **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)

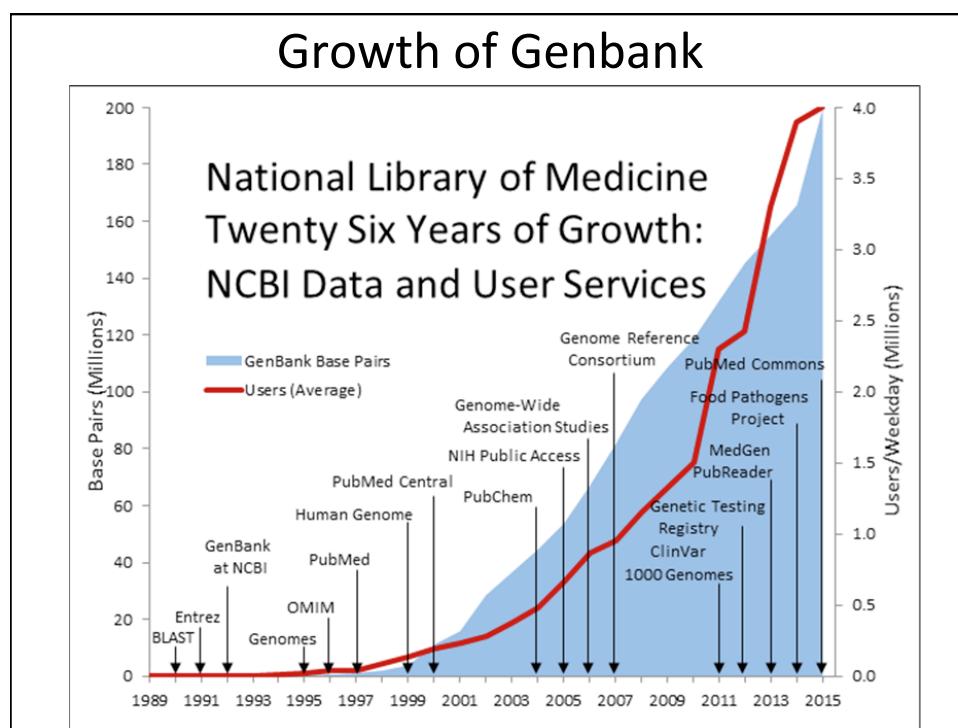
### PRIMARY VS. DERIVATIVE SEQUENCE DATABASES





## 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 #



<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)

- Title
  - Taxonomy
  - Citation

## Header

## Features (AA seq)

# DNA Sequence

## Fields

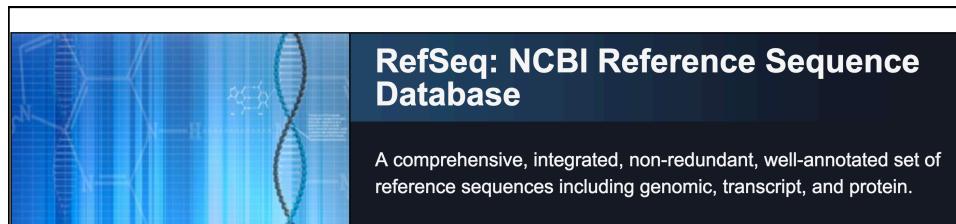
<b>LOCUS DEFINITION</b>	SCU49845      5028 bp      DNA	<b>PLN</b>	<b>21-JUN-1993</b>
	Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.		
<b>ACCESSION</b>	U49845		
<b>MD5</b>	91293613		
<b>VERSION</b>	U49845.1	<a href="#">GT</a> :1293613	
<b>KEYWORDS</b>	baker's yeast		
<b>SOURCE</b>	Saccharomyces cerevisiae		
<b>ORGANISM</b>	Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.		
<b>REFERENCE</b>	1. (bases 1 to 5028)		
<b>AUTHORS</b>	Torpey, L.E., Gibbs, P.E., Nelson, J. and Lawrence, C.W.		
<b>TITLE</b>	Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in <i>Saccharomyces cerevisiae</i>		
<b>JOURNAL</b>	Yeast, 10 (11), 1503-1509 (1994)		
<b>MEDLINE</b>	95176709		
<b>REFERENCE</b>	2. (bases 1 to 5028)		
<b>AUTHORS</b>	Roemer, T., Madden, K., Chang, J. and Snyder, M.		
<b>TITLE</b>	Selection of axial growth sites in yeast requires Axl2p, a novel plasma membrane glycoprotein		
<b>JOURNAL</b>	Genes Dev, 10 (7), 777-793 (1996)		
<b>MEDLINE</b>	96194260		
<b>REFERENCE</b>	3. (bases 1 to 5028)		
<b>AUTHORS</b>	Roemer, T.		
<b>TITLE</b>	<a href="#">Direct Submission</a>		
<b>JOURNAL</b>	Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New Haven, CT, USA		
<b>FEATURES</b>	Location/Qualifiers		
<b>source</b>	1..5028		
	/organism="Saccharomyces cerevisiae"		
<b>CDS</b>	<a href="#">/db_xref="taxon:4932"</a>		
	/chromosome="IX"		
	/map="S"		
	<1..206		
	/codon_start=3		
	/product="TCP1-beta"		
	<a href="#">/protein_id="AAAH98665.1"</a>		
	/db_xref="PDB:g1293614"		
	/db_xref="GT:1293614"		
	<a href="#">/translation="SSYYNGIESTSGLDLNNGTIAIMRQLGIVESYKLKRAYVSSASEA</a>		
<b>gene</b>	AEVLLLRDVHILTRARPTENRQHM"		
<b>CDS</b>	687..3158		
	/gene="AXL2"		
	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 were meant to do
- Keep in mind that GenBank is DNA centric and is a poor vehicle for protein and mRNA expression/interaction information



**RefSeq: NCBI Reference Sequence Database**

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

### Announcements

**January 12, 2018**  
**RefSeq Release 86 is available for FTP**

This release includes:

Proteins: 102,133,844  
Transcripts: 21,370,778  
Organisms: 75,218  
Available at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>  
Documentation: [Release Notes](#)

## 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 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

The screenshot shows the Google Advanced Search interface. On the left, there is a sidebar with the text "... or complex". The main area contains several search filters:

- Find results:** with all of the words, with the exact phrase, with at least one of the words, without the words.
- Language:** Return pages written in [any language dropdown].
- File Format:** Only [dropdown] return results of the file format [any format dropdown].
- Date:** Return web pages first seen in the [anytime dropdown].
- Numeric Range:** Return web pages containing numbers between [ ] and [ ].
- Occurrences:** Return results where my terms occur [anywhere in the page dropdown].
- Domain:** Only [dropdown] return results from the site or domain [e.g. google.com, .org More info dropdown].
- Usage Rights:** Return results that are [not filtered by license More info dropdown].
- SafeSearch:** No filtering [radio button], Filter using SafeSearch [radio button].

**Page-Specific Search:**

- Similar:** Find pages similar to the page [e.g. www.google.com/help.html Search button].
- Links:** Find pages that link to the page [Search button].

**Topic-Specific Searches:**

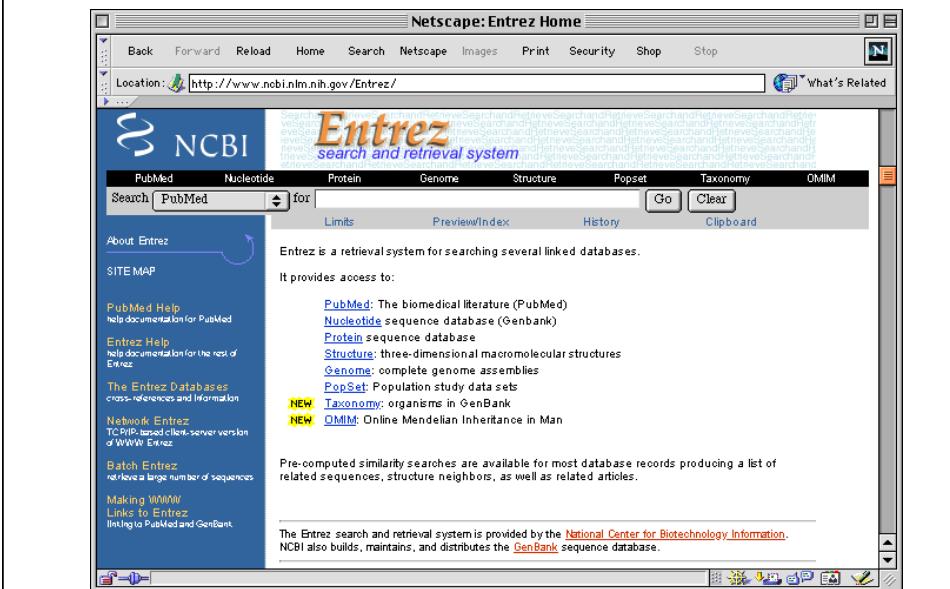
- [Google Book Search](#) - Search the full text of books
- [New Google Code Search](#) - Search public source code
- [Google Scholar](#) - Search scholarly papers
- [Google News archive search](#) - Search historical news
- [Apple Macintosh](#) - Search for all things Mac
- [BSD](#) - Search for all things BSD
- [Linux](#) - Search for all penguin-friendly pages
- [Microsoft](#) - Search Microsoft-related pages
- [U.S. Government](#) - Search all U.S. federal, state and local government sites
- [Universities](#) - Search a specific school's website

©2007 Google

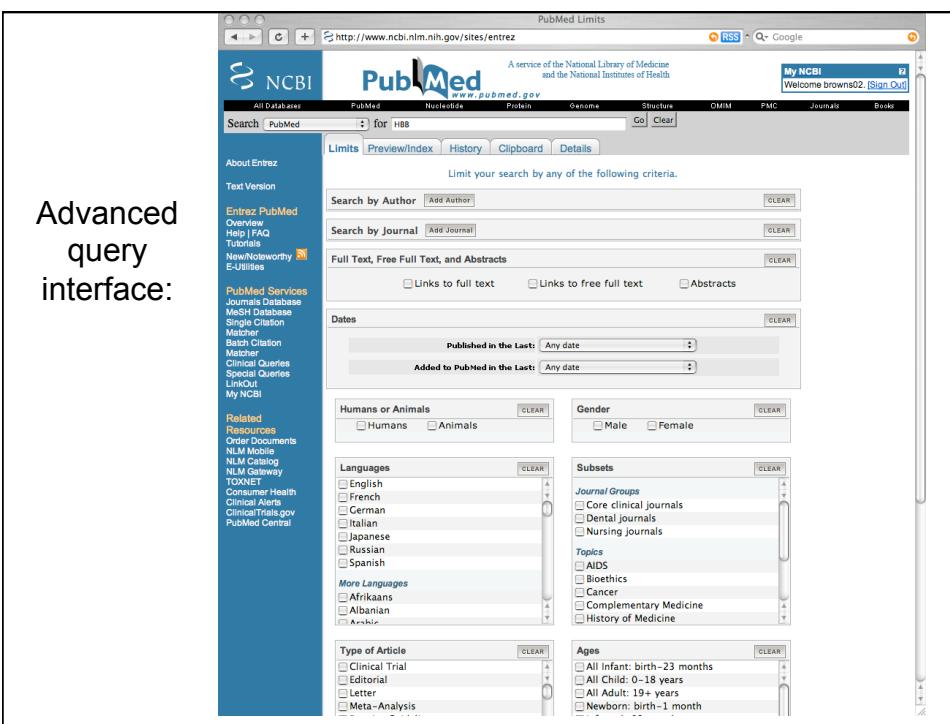
## Query Languages

- 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

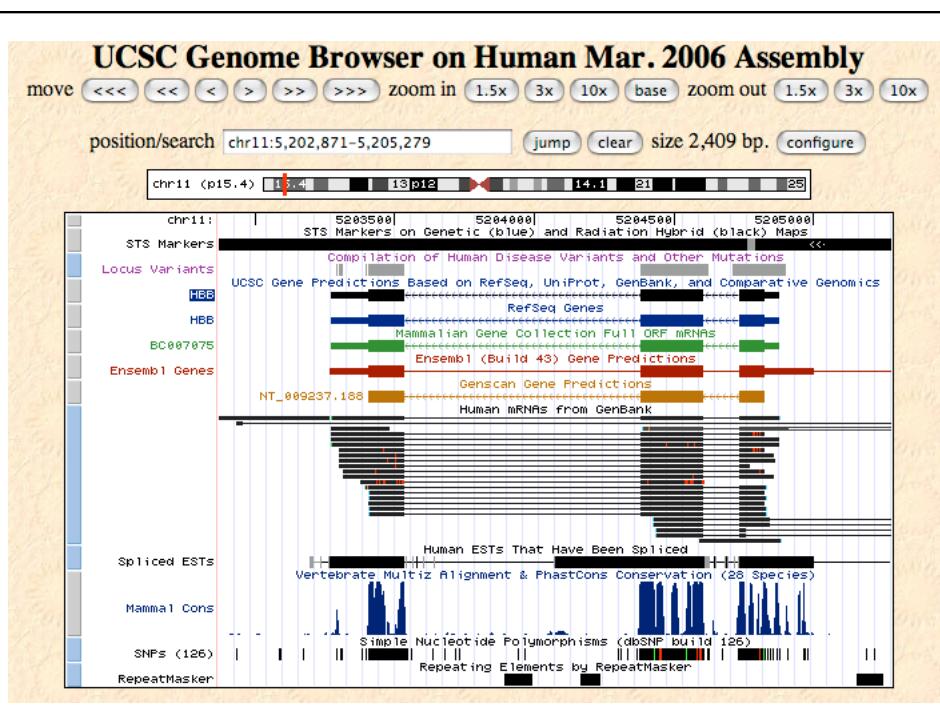
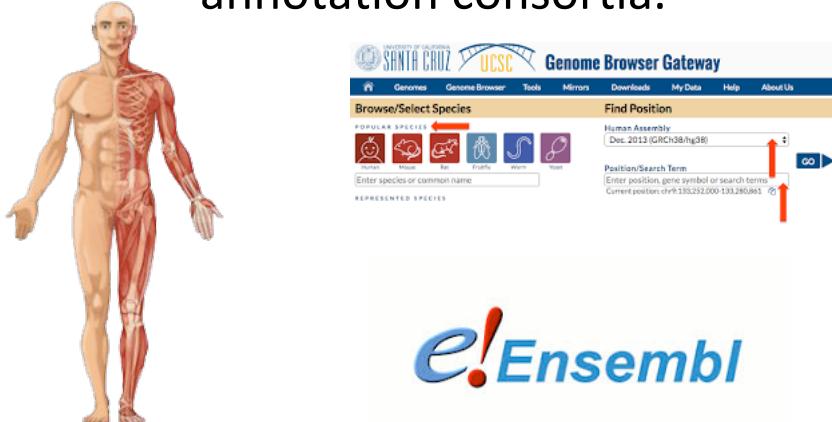
Bad to search for "human hemoglobin" in a '*Description*' field

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

The screenshot shows a complex web-based genome browser interface with a sidebar on the left containing a list of track categories and their sub-options. The categories include:

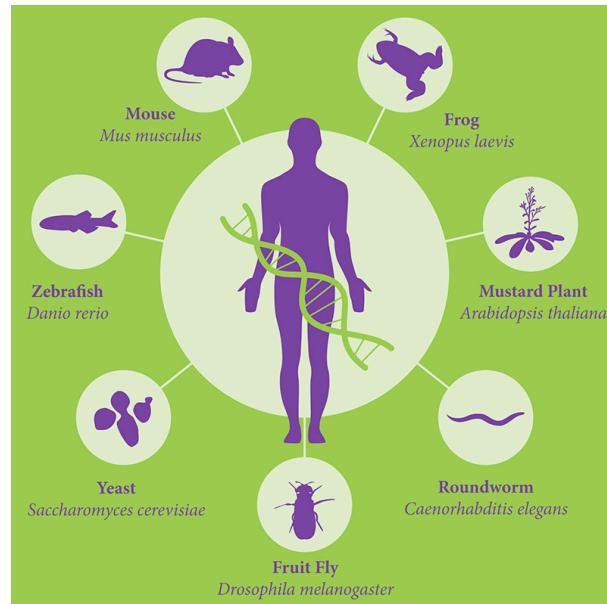
- Mapping and Sequencing Tracks:** Base Position, Map Contigs, Fosmid End Pairs, Chromosome Band, STS Markers, FISH Clones, Recomb Rate.
- Phenotype and Disease Associations:**
- Genes and Gene Prediction Tracks:** UCSC Genes, Other RefSeq, N-SCAN, Augustus, Old Known Genes, MGC Genes, SGP Genes, Superfamily, Alt Events, ORFeome Clones, Ensembl Genes, Genscan Genes, CCDS, RefSeq Genes, AceView Genes, Exoniphy, sno/miRNA.
- mRNA and EST Tracks:** Human mRNAs, Spliced ESTs, H-Inv, UniGene, Human ESTs, Poly(A), Other mRNAs, CGAP SAGE.
- Expression and Regulation:** Affy All Exon, Bertone Yale TAR, CpG Islands, ORegAnno, Affy HuEx 1.0, Affy U133, FirstEF, TX Reg Potential, Allen Brain, GNF Atlas 2, GNF Ratio, Affy U133Plus2, Eponine TSS, TFBS Conserved.
- Comparative Genomics:** Conservation, Medaka Net, Tetraodon Chain, Most Conserved, Stickleback Chain, Tetrapod Net, 17-Way Cons, Stickleback Net, Tetrapod Ecores, 17-Way Most Cons, Fugu Chain, Zebrafish chain, Medaka Chain, Fugu Net, Zebrafish Net.
- Variation and Repeats:** SNPs (126), Exapeted Repeats, SNP Arrays, RepeatMasker, HapMap SNPs, Simple Repeats, Segmental Dups, Microsatellite, Structural Var, Self Chain.

## Ensembl at EBI/EMBL

The Ensembl interface consists of two main windows:

- MapView (Left):** This window displays a genomic track for chromosome 5. It includes a track selector on the left, a main visualization area with a grey bar representing the genome, and a detailed track information panel on the right. The panel shows various tracks like 'Total Genes', 'IC Repeats', 'SNPs', and 'Diseases'. A specific SNP (rs2052727) is highlighted with a blue box, and a zoomed-in view of its details is shown in a separate window.
- ConfigView (Right):** This window provides a more detailed view of a genomic region. It features a zoomed-in track for chromosome 5 between coordinates 24100000 and 25100000. The visualization includes tracks for 'Genes', 'Repeats', 'SNPs', and 'Diseases'. A red line marks the current focus window, and a tooltip indicates 'Click anywhere on the red line below to reposition focus window'.

In addition to human data there are many model organisms.



**RCSB PDB PROTEIN DATA BANK**

<http://www.wwpdb.org>

A MEMBER OF THE An Information Portal to Biological Macromolecular Structures As of Tuesday Sep 03, 2013 at 5 PM PDT there are 93624 Structures PDB Statistics

Search Advanced Browse

Everything Author Macromolecule Sequence Ligand ?  
e.g., PDB ID, molecule name, author

Search History, Previous Results

Customize This Page Available on the App Store

PDB-101 Hide Structural View of Biology Understanding PDB Data Molecule of the Month Educational Resources Author Profiles

MyPDB Hide Login to your Account Register a New Account MyPDB Help Page

Home Hide News & Publications Usage/Reference Policies Deposit Policies Website FAQ Deposition FAQ Contact Us About Us Careers External Links

**Biological Macromolecular Resource**

**Full Description**

**Learn: Featured Molecules**

**Structural View of Biology**

**PDB-101**

**Designed Protein Cages**

Scientists are great tinkerers, and surprisingly, proteins can often be used like tinkertoys. The proteins found in cells have evolved to have a stable, folded structures. Scientists are now building on these stable proteins and making changes to engineer new functions. These functions include designing new enzymes, designing proteins with improved medicinal properties, and designing large complexes with a desired shape and size.

**Full Article**

**Protein Structure Initiative Featured System**

**Bacteria and Bile Salts**

Our digestive system uses powerful detergents to break up the fat in our diet and make it available to our cells. These detergents are built from bile acids, which are the same kind of molecules that are found in membranes. To clean it into a detergent, several solubilizing groups are added, including hydroxyl groups on the main cholesterol ring structure and the amino acids glycine and taurine on the cholesterol tail. The result is a bile salt, which is soluble in water, but also has a carbon-rich side that interacts strongly with fats. PSI researchers have solved the first structure of a bacterial enzyme involved the removal of bile acid hydroxyl groups in the human gut.

**Full Article | Archive | PSI Structural Biology Knowledgebase**

**New Features Hide**

**Latest release: September 2013**

**Protein Symmetry and Stoichiometry**

Visualize, browse & search symmetry / stoichiometry

**Website Release Archive**

**RCSB PDB News Hide**

Weekly | Quarterly | Yearly

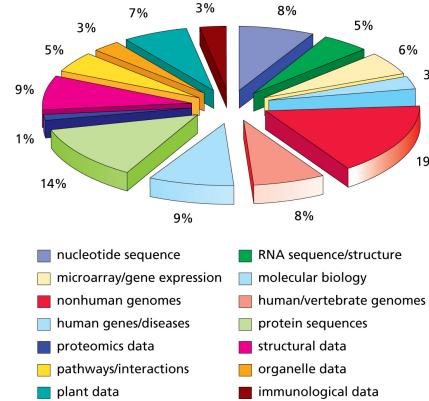
**Head Back to School with PDB-101**

## Specialized Databases

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