Bioinformatik
Stefanie Hartmann
Wintersemester 2019 / 2020, Universität Potsdam

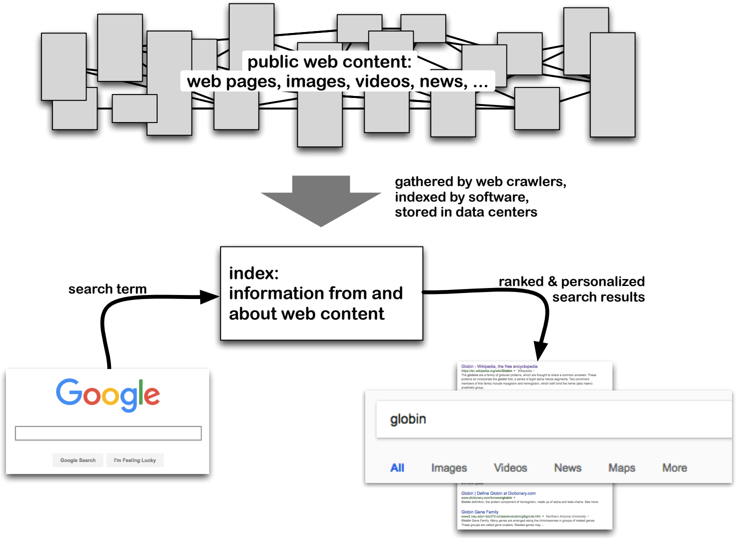
Online resources (I)
Oct 25, 2019

- get a Computer Pool account as soon as possible
  - https://www.chem.uni-potsdam.de/groups/pools/Studierende/studierende.html

- 10h-10:45h, four pools (64 computers) 12 69
- 11h-11:45h, three pools (48 computers) 54

- bitte bei PULS ummelden / abmelden!

Google search



public web content:
web pages, images, videos, news, ...

gathered by web crawlers,
indexed by software,
stored in data centers

search term

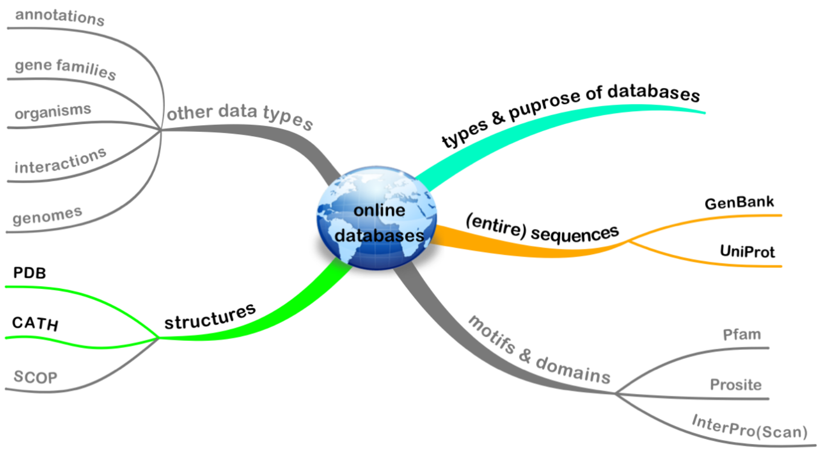
index:
information from and
about web content

ranked & personalized
search results

Literature databases

Google Scholar	PubMed	Web of Science
search engine for filtered web content	human-curated database	human-curated database
all subject areas	focus on biomedical literature	interdisciplinary
journals, conference proceedings, books, reports, ...	all available languages	selected journals, English language
no tags	articles are tagged (“review”, “mouse”)	articles are tagged (“review”)
full text searches	searches based on abstract and tags	searches based on abstract and tags

Today's topics



online databases

types & puprose of databases

(entire) sequences

GenBank

UniProt

motifs & domains

Pfam

Prosite

InterPro(Scan)

structures

PDB

CATH

SCOP

other data types

annotations

gene families

organisms

interactions

genomes

Biology: virtual spaces

networks

- internet, local networks, shared computers
- research, collaboration, interaction, exchange

access data on other computers

- files
- programs
- data in databases
  (sequences, literature, taxonomy, structure, pathways, and much more!)

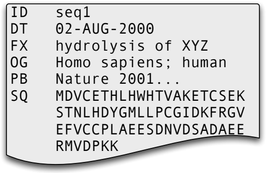
Biological databases (first generation)

purpose

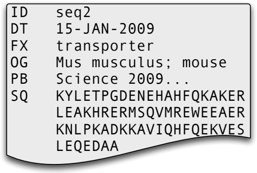
- centralize biological sequences
- make sequences available for computer analysis
- retrieve (mostly) one sequence at a time

format

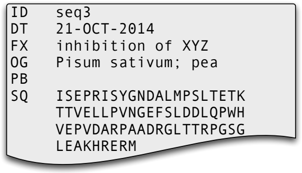
- "flat file": separate plain-text files
- human readable, computer readable



```
ID   seq1
DT   02-AUG-2000
FX   hydrolysis of XYZ
OG   Homo sapiens; human
PB   Nature 2001...
SQ   MDVCETHLHWHTVAKETCSEK
     STNLHDYGMLLPCGIDKFRGV
     EFVCCPLAEESDNVDSADAEE
     RMVDPKK
```



```
ID   seq2
DT   15-JAN-2009
FX   transporter
OG   Mus musculus; mouse
PB   Science 2009...
SQ   KYLETPGDENEHAHFQKAKER
     LEAKHRERMSQVMREWEEAER
     KNLPKADKKAVIQHFQEKVES
     LEQEDAA
```

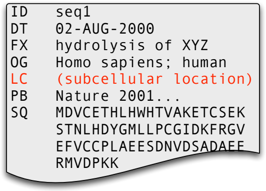


```
ID   seq3
DT   21-OCT-2014
FX   inhibition of XYZ
OG   Pisum sativum; pea
PB
SQ   ISEPRISYGNDALMPSLTETK
     TTVELLPVNGEFSLDDLQPWH
     VEPVDARPAADRGLTTRPGSG
     LEAKHRERM
```

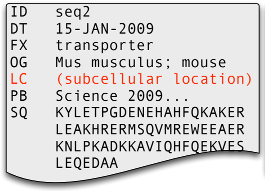
Biological databases (first generation)

problems

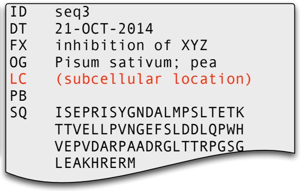
- updating types of information
- other data difficult to include
  (splice variants, regulatory regions, etc)
- entering and updating sequence data
- show/find complex relationships between entries



```
ID   seq1
DT   02-AUG-2000
FX   hydrolysis of XYZ
OG   Homo sapiens; human
LC   (subcellular location)
PB   Nature 2001...
SQ   MDVCETHLHWHTVAKETCSEK
     STNLHDYGMLLPCGIDKFRGV
     EFVCCPLAEESDNVDSADAEE
     RMVDPKK
```



```
ID   seq2
DT   15-JAN-2009
FX   transporter
OG   Mus musculus; mouse
LC   (subcellular location)
PB   Science 2009...
SQ   KYLETPGDENEHAHFQKAKER
     LEAKHRERMSQVMREWEEAER
     KNLPKADKKAVIQHFQEKVES
     LEQEDAA
```



```
ID   seq3
DT   21-OCT-2014
FX   inhibition of XYZ
OG   Pisum sativum; pea
LC   (subcellular location)
PB
SQ   ISEPRISYGNDALMPSLTETK
     TTVELLPVNGEFSLDDLQPWH
     VEPVDARPAADRGLTTRPGSG
     LEAKHRERM
```

Biological databases (second generation)

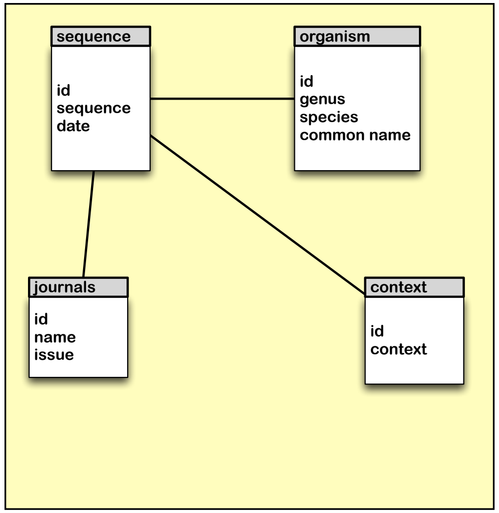
purpose

- centralize biological sequences
- make sequences available for computer analysis
- allow retrieving information across multiple entries

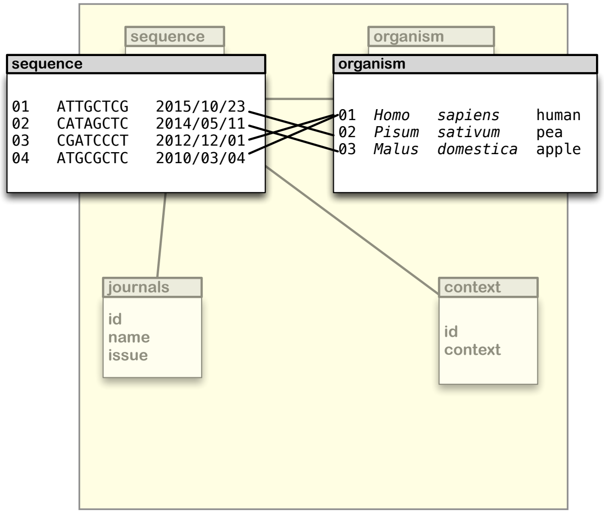
format

- computer readable
- can be made human readable
- relational databases (linked tables)

Biological databases (second generation)



Biological databases (second generation)



sequence

01	ATTGCTCG	2015/10/23
02	CATAGCTC	2014/05/11
03	CGATCCCT	2012/12/01
04	ATGCGCTC	2010/03/04

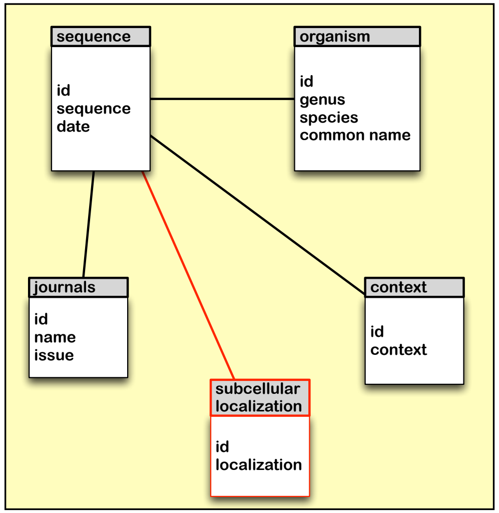
organism

01	*Homo*	*sapiens*	human
02	*Pisum*	*sativum*	pea
03	*Malus*	*domestica*	apple

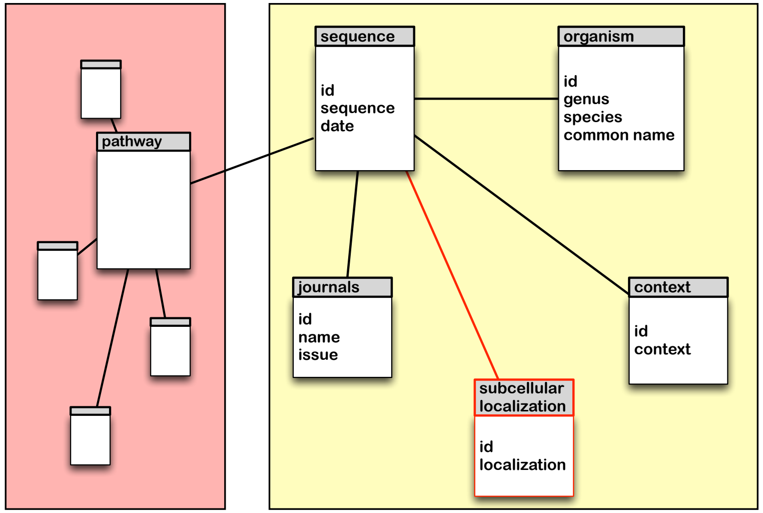
Biological databases (second generation)

challenges:

- organizing & linking independent databases



Biological databases (third generation)



pathway

sequence
id
sequence
date

organism
id
genus
species
common name

journals
id
name
issue

context
id
context

subcellular localization
id
localization

Using online databases

- Online databases have the relevant records
  - there are hundreds of online databases
  - there are often different online databases that have the same / similar information
  - how to find the relevant records / information is not always obvious
- You need to understand
  - the data they have, and how they are organized
  - relationships between types of data
  - links within and between databases

Types of databases

Characterization of databases by

- type of data
- supported activities
- organism(s)
- technical design
- availability (publicly available, commercial)
- primary vs. derived (secondary) data, both

Types of databases

Characterization of databases by

- type of data
- supported activities
- organism(s)
- technical design
- availability (publicly available, commercial)
- primary vs. derived (secondary) data, both
  - primary: DNA sequence, protein structure
  - secondary: protein sequence, protein structure

Data!

primary database

primary / raw data

```
ATGATCTTCATTTTGACCGTAAACTTCCGTTGGAG
ATATTTAATCTTACTGATTTGCAAGTCTTTAATGT
TGCTGGAAATCAGTTGTCCGGTGAAATACCCGGAG
AGGTTCCTCGGAGTTTGCGATACTTTGATCTTTCG
TCAAACTTGTTTACAGGAGATATTCCGAGGTACTT
GTCGGATTTGTCTCAGTTGCTTTTTATTAATCTTT
CTTACAATCGTTTTTCCGGCGAAATTCCGGCGAGT
ATTGGTCGGCTTCAGCAGCTTCAGTACCTTTGGCT
CGCGTATAATGACTTGGTTGGAACTTTGCCTTCGG
CAATTGCTAACTGTTCGTCGTTGGTTCATTGAGTG
CTTAAGGAAATGCTATTCGAGGTGTTATTCCAGCG
GCAATTGCTGCTTTACCAAAGCTTCAGGTGATATC
TTTATCACATACTACTCTGTCTGGTTCTTTGCCAG
CTTCATTGTTCTGCAATGTTTCGATTTATCCTCCT
TCCCTTGGGATTGTTCAGTTGGGTTTCAATGGGTT
```

metadata

- organism: species
- sampling date & location
- DNA extraction protocol
- sequencing protocol
- processing of sequence data
- name of investigator
- ...

secondary database

secondary / derived data

- information about / link to primary data
- positions of start, stop, intron, exon, ...
- prediction of protein sequence
- prediction of functional domains
- prediction of subcellular location
- similarity with other sequences
- evolutionary relationship with other sequences
- ...

databases:

- primary, secondary, mixed

curation by curators (or computers) of secondary databases

- collect, annotate, validate, consolidate, monitor data quality, completeness, and consistency

analysis by researchers

- analyze to gain knowledge, integrate, find patterns, ...

DNA sequencing data

raw data

base calling

DNA data & quality info

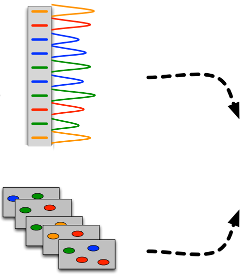
further processing

Sanger:

trace chromatograms

NGS data:

series of image files



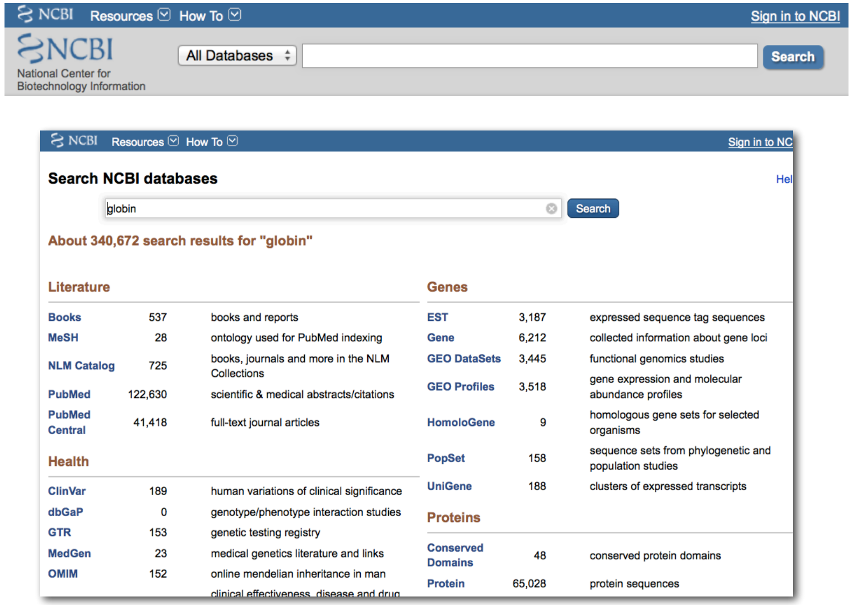
GCTTAGATNNNNTTACTTG

Primary nucleotide databases

	NCBI	EMBL-EBI	DDBJ DNA Data Bank of Japan
DB	GenBank	EMBL	DDBJ
maintained by	NCBI	EBI	NIG
access & search	Entrez	SRS	getentry
URL	www.ncbi.nlm.nih.gov	www.ebi.ac.uk	www.ddbj.nig.ac.jp

National Center for Biotechnology Information

- created in 1988 to develop information systems for molecular biology
- databases
  - nucleic acid database GenBank
  - many other databases: literature, taxonomy, DNA & RNA, proteins, genomes, etc
  - Entrez: search & retrieval system for the databases at NCBI
- data analysis tools
  - BLAST sequence comparison



Search NCBI databases

globin

About 340,672 search results for "globin"

Literature

Books	537	books and reports
MeSH	28	ontology used for PubMed indexing
NLM Catalog	725	books, journals and more in the NLM Collections
PubMed	122,630	scientific & medical abstracts/citations
PubMed Central	41,418	full-text journal articles

Health

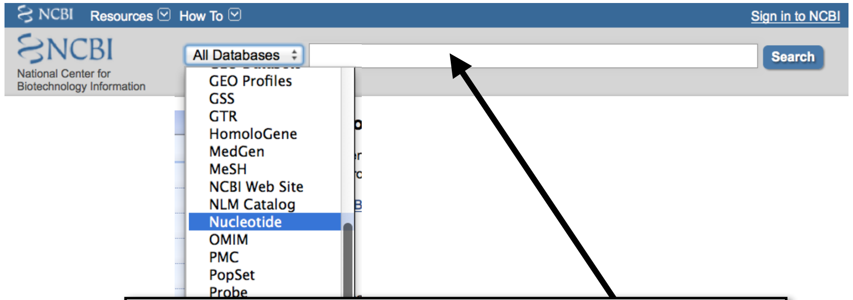
ClinVar	189	human variations of clinical significance
dbGaP	0	genotype/phenotype interaction studies
GTR	153	genetic testing registry
MedGen	23	medical genetics literature and links
OMIM	152	online mendelian inheritance in man
		clinical effectiveness, disease and drug

Genes

EST	3,187	expressed sequence tag sequences
Gene	6,212	collected information about gene loci
GEO DataSets	3,445	functional genomics studies
GEO Profiles	3,518	gene expression and molecular abundance profiles
HomoloGene	9	homologous gene sets for selected organisms
PopSet	158	sequence sets from phylogenetic and population studies
UniGene	188	clusters of expressed transcripts

Proteins

Conserved Domains	48	conserved protein domains
Protein	65,028	protein sequences



- horse
- horse[Organism]
- horse[Organism] AND 110:500[Sequence Length] AND 2009[Publication Date]

GenBank sequences

individual research projects

- result of studying a specific biological process
- sequences are full-length and well annotated

large-scale projects

- genome or transcriptome projects
- often fragments, low quality, no functional annotation, no experimental verification

metagenome projects

- sequence data from environmental samples
- often fragments, low quality, source organism unknown

GenBank statistics (v233): top 10

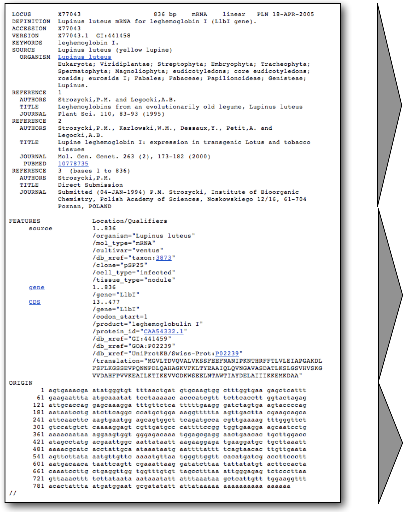
Entries	Bases	Species
26,917,743	20,390,017,939	*Homo sapiens*
1,937,298	17,186,497,195	*Triticum aestivum*
10,018,042	10,443,110,196	*Mus musculus*
22,978	9,981,129,079	*Triticum turgidum subsp. durum*
1,347,029	8,071,264,876	*Hordeum vulgare subsp. vulgare*
2,200,465	6,530,442,551	*Rattus norvegicus*
2,234,258	5,433,577,654	*Bos taurus*
4,211,701	5,250,234,927	*Zea mays*
213,865,349	366,733,917,629	TOTAL

v233, Aug 15, 2019

Divisions

BCT	bacterial sequences	Entrez nucleotide
INV	invertebrate sequences	Entrez nucleotide
MAM	other mammalian sequences	Entrez nucleotide
PHG	bacteriophage sequences	Entrez nucleotide
PLN	plant, fungal, and algal sequences	Entrez nucleotide
PRI	primate sequences	Entrez nucleotide
ROD	rodent sequences	Entrez nucleotide
SYN	synthetic sequences	Entrez nucleotide
UNA	unannotated sequences	Entrez nucleotide
VRL	viral sequences	Entrez nucleotide
VRT	other vertebrate sequences	Entrez nucleotide
ENV	Environmental sampling sequences	Entrez nucleotide
EST	expressed sequence tags	Entrez EST
GSS	genome survey sequences	Entrez GSS
HTC	high throughput cDNA sequences	Entrez nucleotide
HTG	high throughput genomic sequences	Entrez nucleotide
STS	sequence tagged sites	Entrez nucleotide
TSA	transcriptome shotgun sequences	Entrez nucleotide
PAT	patent sequences	Entrez nucleotide
WGS	whole genome shotgun sequences	Entrez nucleotide

A sequence in GenBank format



```
LOCUS       X77043                   836 bp    mRNA    linear   PLN 18-APR-2005
DEFINITION  Lupinus luteus mRNA for leghemoglobin I (LlbI gene).
ACCESSION   X77043
VERSION     X77043.1  GI:441458
KEYWORDS    leghemoglobin I.
SOURCE      Lupinus luteus (yellow lupine)
  ORGANISM  Lupinus luteus
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; eurosids I; Fabales; Fabaceae; Papilionoideae; Genisteae;
            Lupinus.
REFERENCE   1
  AUTHORS   Strozycki,P.M. and Legocki,A.B.
  TITLE     Leghemoglobins from an evolutionarily old legume, Lupinus luteus
  JOURNAL   Plant Sci. 110, 83-93 (1995)
REFERENCE   2
  AUTHORS   Strozycki,P.M., Karlowski,W.M., Dessaux,Y., Petit,A. and
            Legocki,A.B.
  TITLE     Lupine leghemoglobin I: expression in transgenic Lotus and tobacco
            tissues
  JOURNAL   Mol. Gen. Genet. 263 (2), 173-182 (2000)
   PUBMED   10778735
REFERENCE   3  (bases 1 to 836)
  AUTHORS   Strozycki,P.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-JAN-1994) P.M. Strozycki, Institute of Bioorganic
            Chemistry, Polish Academy of Sciences, Noskowskiego 12/16, 61-704
            Poznan, POLAND

FEATURES             Location/Qualifiers
     source          1..836
                     /organism="Lupinus luteus"
                     /mol_type="mRNA"
                     /cultivar="ventus"
                     /db_xref="taxon:3873"
                     /clone="pSP25"
                     /cell_type="infected"
                     /tissue_type="nodule"
     gene            1..836
                     /gene="LlbI"
     CDS             13..477
                     /gene="LlbI"
                     /codon_start=1
                     /product="leghemoglobulin I"
                     /protein_id="CAA54332.1"
                     /db_xref="GI:441459"
                     /db_xref="GOA:P02239"
                     /db_xref="UniProtKB/Swiss-Prot:P02239"
                     /translation="MGVLTDVQVALVKSSFEEFNANIPKNTHRFFTLVLEIAPGAKDL
                     FSFLKGSSEVPQNNPDLQAHAGKVFKLTYEAAIQLQVNGAVASDATLKSLGSVHVSKG
                     VVDAHFPVVKEAILKTIKEVVGDKWSEELNTAWTIAYDELAIIIKKEMKDAA"
ORIGIN
        1 agtgaaacga atatgggtgt tttaactgat gtgcaagtgg ctttggtgaa gagctcattt
       61 gaagaattta atgcaaatat tcctaaaaac acccatcgtt tcttcacctt ggtactagag
      121 attgcaccag gagcaaagga tttgttctca tttttgaagg gatctagtga agtaccccag
      181 aataatcctg atcttcaggc ccatgctgga aaggttttta agttgactta cgaagcagca
      241 attcaacttc aagtgaatgg agcagtggct tcagatgcca cgttgaaaag tttgggttct
      301 gtccatgtct caaaaggagt cgttgatgcc cattttccgg tggtgaagga agcaatcctg
      361 aaaacaataa aggaagtggt gggagacaaa tggagcgagg aactgaacac tgcttggacc
      421 atagcctatg acgaattggc aattataatt aagaaggaga tgaaggatgc tgcttaaatt
      481 aaaacgcatc acctattgca ataaataatg aattttattt tcagtaacac ttgttgaata
      541 agttcttata aatgttgttc aaaatgttaa tgggttggtt cacatgatcg accttccctt
      601 aatgacaaca taattcagtt cgaaattaag gatatcttaa tattatatgt acttccacta
      661 caaatccttg ctgaggttgg tggtttgtgt tagcctttaa attgggagag tctcccttaa
      721 gttaaacttt tcttataata aataaatatt atttaaataa gctcattgtt tggaaggttt
      781 acactattta atgatggaat gcgatatatt attataaaaa aaaaaaaaaa aaaaaa
//
```

Header

- Accession
- Taxonomy
- Citation

Features

(AA sequence)

DNA sequence

GenBank: Header

```
LOCUS       X77043                   836 bp    mRNA    linear   PLN 18-APR-2005
DEFINITION  Lupinus luteus mRNA for leghemoglobin I (LlbI gene).
ACCESSION   X77043
VERSION     X77043.1
KEYWORDS    leghemoglobin I.
SOURCE      Lupinus luteus (yellow lupine)
  ORGANISM  Lupinus luteus
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; eurosids I; Fabales; Fabaceae; Papilionoideae; Genisteae;
            Lupinus.
REFERENCE   1
  AUTHORS   Strozycki,P.M. and Legocki,A.B.
  TITLE     Leghemoglobins from an evolutionarily old legume, Lupinus luteus
  JOURNAL   Plant Sci. 110, 83-93 (1995)
```

GenBank: Header

```
LOCUS       X77043                   836 bp    mRNA    linear   PLN 18-APR-2005
DEFINITION  Lupinus luteus mRNA for leghemoglobin I (LlbI gene).
ACCESSION   X77043
VERSION     X77043.1
KEYWORDS    leghemoglobin I.
SOURCE      Lupinus luteus (yellow lupine)
  ORGANISM  Lupinus luteus
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; eurosids I; Fabales; Fabaceae; Papilionoideae; Genisteae;
            Lupinus.
REFERENCE   1
  AUTHORS   Strozycki,P.M. and Legocki,A.B.
  TITLE     Leghemoglobins from an evolutionarily old legume, Lupinus luteus
  JOURNAL   Plant Sci. 110, 83-93 (1995)
```

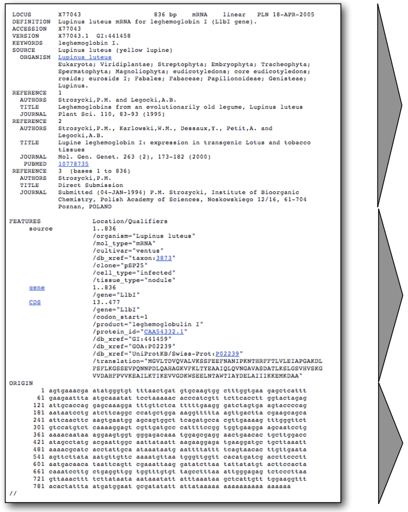
```
KEYWORDS    RefSeq.
COMMENT     VALIDATED REFSEQ: This record has undergone validation or
            preliminary review. The reference sequence was derived from
            AL844581.7, BC079883.1, BC082596.1 and AL845523.6.
```

GenBank: Features

http://www.insdc.org/documents/feature_table.html

```
FEATURES             Location/Qualifiers
     source          1..836
                     /organism="Lupinus luteus"
                     /mol_type="mRNA"
                     /cultivar="ventus"
                     /db_xref="taxon:3873"
                     /clone="pSP25"
                     /cell_type="infected"
                     /tissue_type="nodule"
     gene            1..836
                     /gene="LlbI"
     CDS             13..477
                     /gene="LlbI"
                     /codon_start=1
                     /product="leghemoglobulin I"
                →    /protein_id="CAA54332.1"
                     /db_xref="GI:441459"
                     /db_xref="GOA:P02239"
                     /db_xref="UniProtKB/Swiss-Prot:P02239"
                     /translation="MGVLTDVQVALVKSSFEEFNANIPKNTHRFFTLVLEIAPGAKDL
                     FSFLKGSSEVPQNNPDLQAHAGKVFKLTYEAAIQLQVNGAVASDATLKSLGSVHVSKG
                     VVDAHFPVVKEAILKTIKEVVGDKWSEELNTAWTIAYDELAIIIKKEMKDAA"
```

A sequence in GenBank format



```
LOCUS       X77043                   836 bp    mRNA    linear   PLN 18-APR-2005
DEFINITION  Lupinus luteus mRNA for leghemoglobin I (LlbI gene).
ACCESSION   X77043
VERSION     X77043.1  GI:441458
KEYWORDS    leghemoglobin I.
SOURCE      Lupinus luteus (yellow lupine)
  ORGANISM  Lupinus luteus
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; eurosids I; Fabales; Fabaceae; Papilionoideae; Genisteae;
            Lupinus.
REFERENCE   1
  AUTHORS   Strozycki,P.M. and Legocki,A.B.
  TITLE     Leghemoglobins from an evolutionarily old legume, Lupinus luteus
  JOURNAL   Plant Sci. 110, 83-93 (1995)
REFERENCE   2
  AUTHORS   Strozycki,P.M., Karlowski,W.M., Dessaux,Y., Petit,A. and
            Legocki,A.B.
  TITLE     Lupine leghemoglobin I: expression in transgenic Lotus and tobacco
            tissues
  JOURNAL   Mol. Gen. Genet. 263 (2), 173-182 (2000)
   PUBMED   10778735
REFERENCE   3  (bases 1 to 836)
  AUTHORS   Strozycki,P.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-JAN-1994) P.M. Strozycki, Institute of Bioorganic
            Chemistry, Polish Academy of Sciences, Noskowskiego 12/16, 61-704
            Poznan, POLAND

FEATURES             Location/Qualifiers
     source          1..836
                     /organism="Lupinus luteus"
                     /mol_type="mRNA"
                     /cultivar="ventus"
                     /db_xref="taxon:3873"
                     /clone="pSP25"
                     /cell_type="infected"
                     /tissue_type="nodule"
     gene            1..836
                     /gene="LlbI"
     CDS             13..477
                     /gene="LlbI"
                     /codon_start=1
                     /product="leghemoglobulin I"
                     /protein_id="CAA54332.1"
                     /db_xref="GI:441459"
                     /db_xref="GOA:P02239"
                     /db_xref="UniProtKB/Swiss-Prot:P02239"
                     /translation="MGVLTDVQVALVKSSFEEFNANIPKNTHRFFTLVLEIAPGAKDL
                     FSFLKGSSEVPQNNPDLQAHAGKVFKLTYEAAIQLQVNGAVASDATLKSLGSVHVSKG
                     VVDAHFPVVKEAILKTIKEVVGDKWSEELNTAWTIAYDELAIIIKKEMKDAA"
ORIGIN
        1 agtgaaacga atatgggtgt tttaactgat gtgcaagtgg ctttggtgaa gagctcattt
       61 gaagaattta atgcaaatat tcctaaaaac acccatcgtt tcttcacctt ggtactagag
      121 attgcaccag gagcaaagga tttgttctca tttttgaagg gatctagtga agtaccccag
      181 aataatcctg atcttcaggc ccatgctgga aaggttttta agttgactta cgaagcagca
      241 attcaacttc aagtgaatgg agcagtggct tcagatgcca cgttgaaaag tttgggttct
      301 gtccatgtct caaaaggagt cgttgatgcc cattttccgg tggtgaagga agcaatcctg
      361 aaaacaataa aggaagtggt gggagacaaa tggagcgagg aactgaacac tgcttggacc
      421 atagcctatg acgaattggc aattataatt aagaaggaga tgaaggatgc tgcttaaatt
      481 aaaacgcatc acctattgca ataaataatg aattttattt tcagtaacac ttgttgaata
      541 agttcttata aatgttgttc aaaatgttaa tgggttggtt cacatgatcg accttccctt
      601 aatgacaaca taattcagtt cgaaattaag gatatcttaa tattatatgt acttccacta
      661 caaatccttg ctgaggttgg tggtttgtgt tagcctttaa attgggagag tctcccttaa
      721 gttaaacttt tcttataata aataaatatt atttaaataa gctcattgtt tggaaggttt
      781 acactattta atgatggaat gcgatatatt attataaaaa aaaaaaaaaa aaaaaa
//
```

Header

- Accession
- Taxonomy
- Citation

Features

(AA sequence)

DNA sequence

The same sequence in FASTA format

```
>gi|441458|emb|X77043.1| Lupinus luteus LlbI gene
AGTGAAACGAATATGGGTGTTTTAACTGATGTGCAAGTGGCTTTGGTGAAGAGCTCATTTGAAGAATTTAAT
GCAAATATTCCTAAAAACACCCATCGTTTCTTCACCTTGGTACTAGAGATTGCACCAGGAGCAAAGGATTTG
TTCTCATTTTTGAAGGGATCTAGTGAAGTACCCCAGAATAATCCTGATCTTCAGGCCCATGCTGGAAAGGTT
TTTAAGTTGACTTACGAAGCAGCAATTCAACTTCAAGTGAATGGAGCAGTGGCTTCAGATGCCACGTTGAAA
AGTTTGGGTTCTGTCCATGTCTCAAAAGGAGTCGTTGATGCCCATTTTCCGGTGGTGAAGGAAGCAATCCTG
AAAACAATAAAGGAAGTGGTGGGAGACAAATGGAGCGAGGAACTGAACACTGCTTGGACCATAGCCTATGAC
GAATTGGCAATTATAATTAAGAAGGAGATGAAGGATGCTGCTTAAATTAAAACGCATCACCTATTGCAATAA
ATAATGAATTTTATTTTCAGTAACACTTGTTGAATAAGTTCTTATAAATGTTGTTCAAAATGTTAATGGGTT
GGTTCACATGATCGACCTTCCCTTAATGACAACATAATTCAGTTCGAAATTAAGGATATCTTAATATTATAT
GTACTTCCACTACAAATCCTTGCTGAGGTTGGTGGTTTGTGTTAGCCTTTAAATTGGGAGAGTCTCCCTTAA
GTTAAACTTTTCTTATAATAAATAAATATTATTTAAATAAGCTCATTGTTTGGAAGGTTTACACTATTTAAT
GATGGAATGCGATATATTATTATAAAAAAAAAAAAAAAAAAAAA
```

Primary structure database: PDB

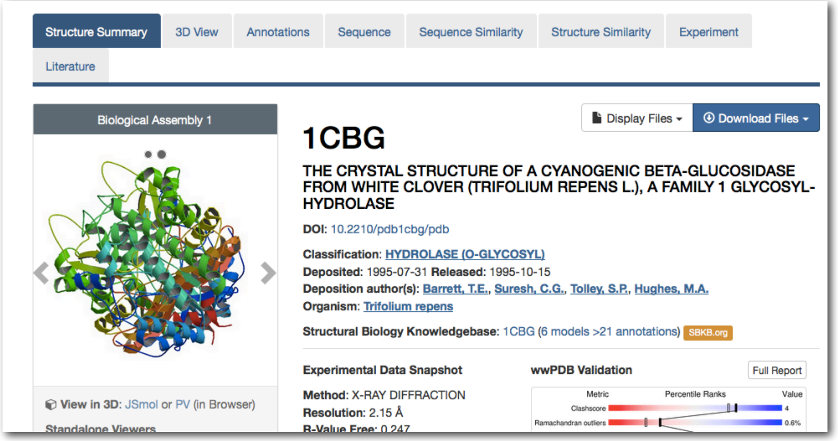
- single worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids
- managed by the Research Collaboratory for Structural Bioinformatics (RCSB)
- http://www.rcsb.org/pdb/

Experimental Method	
X-RAY	159,973
NMR	12,804
ELECTRON MICR.	3,914
…	
Total	157,145

Organism (top 5)	
Homo sapiens	41,327
Escherichia coli	6,859
Mus musculus	4,999
S. cerevisiae	3,567
Rattus norvegicus	3,040

Polymer Type	
Protein	145,695
Mixed	8,021
D/RNA	3,399

A PDB structure summary



1CBG

THE CRYSTAL STRUCTURE OF A CYANOGENIC BETA-GLUCOSIDASE FROM WHITE CLOVER (TRIFOLIUM REPENS L.), A FAMILY 1 GLYCOSYL-HYDROLASE

DOI: 10.2210/pdb1cbg/pdb

Classification: HYDROLASE (O-GLYCOSYL)
Deposited: 1995-07-31 **Released:** 1995-10-15
Deposition author(s): Barrett, T.E., Suresh, C.G., Tolley, S.P., Hughes, M.A.
Organism: Trifolium repens

Structural Biology Knowledgebase: 1CBG (6 models >21 annotations)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.15 Å
R-Value Free: 0.247

wwPDB Validation

View in 3D: JSmol or PV (in Browser)

Standalone Viewers

A PDB entry

```
HEADER    HYDROLASE (O-GLYCOSYL)                  31-JUL-95   1CBG
TITLE     THE CRYSTAL STRUCTURE OF A CYANOGENIC BETA-GLUCOSIDASE FROM WHITE
TITLE    2 CLOVER (TRIFOLIUM REPENS L.), A FAMILY 1 GLYCOSYL-HYDROLASE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CYANOGENIC BETA-GLUCOSIDASE;
COMPND   3 CHAIN: A;
COMPND   4 EC: 3.2.1.21
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: TRIFOLIUM REPENS;
SOURCE   3 ORGANISM_COMMON: WHITE CLOVER;
SOURCE   4 ORGANISM_TAXID: 3899;
SOURCE   5 VARIANT: L;
SOURCE   6 ORGAN: LEAVES;
SOURCE   7 TISSUE: LEAVES
KEYWDS    CYANOGENIC BETA-GLUCOSIDASE, HYDROLASE (O-GLYCOSYL)
EXPDTA    X-RAY DIFFRACTION
AUTHOR    T.E.BARRETT,C.G.SURESH,S.P.TOLLEY,M.A.HUGHES
...
ATOM      1  N   PHE A   1      60.319  44.445  68.521  1.00 38.85           N
ATOM      2  CA  PHE A   1      60.228  43.024  68.138  1.00 38.70           C
ATOM      3  C   PHE A   1      61.491  42.643  67.355  1.00 39.17           C
ATOM      4  O   PHE A   1      61.998  43.412  66.522  1.00 39.33           O
ATOM      5  CB  PHE A   1      58.975  42.736  67.325  1.00 37.25           C
ATOM      6  CG  PHE A   1      58.451  41.343  67.209  1.00 35.96           C
ATOM      7  CD1 PHE A   1      59.146  40.349  66.527  1.00 35.81           C
ATOM      8  CD2 PHE A   1      57.216  41.020  67.769  1.00 35.36           C
ATOM      9  CE1 PHE A   1      58.661  39.042  66.419  1.00 35.30           C
ATOM     10  CE2 PHE A   1      56.702  39.738  67.682  1.00 35.22           C
ATOM     11  CZ  PHE A   1      57.423  38.748  67.010  1.00 35.20           C
ATOM     12  N   LYS A   2      61.942  41.450  67.667  1.00 39.48           N
ATOM     13  CA  LYS A   2      63.096  40.795  67.022  1.00 39.72           C
ATOM     14  C   LYS A   2      62.768  39.294  67.181  1.00 39.56           C
ATOM     15  O   LYS A   2      62.464  38.798  68.278  1.00 39.57           O
ATOM     16  CB  LYS A   2      64.467  41.188  67.512  1.00 40.44           C
```

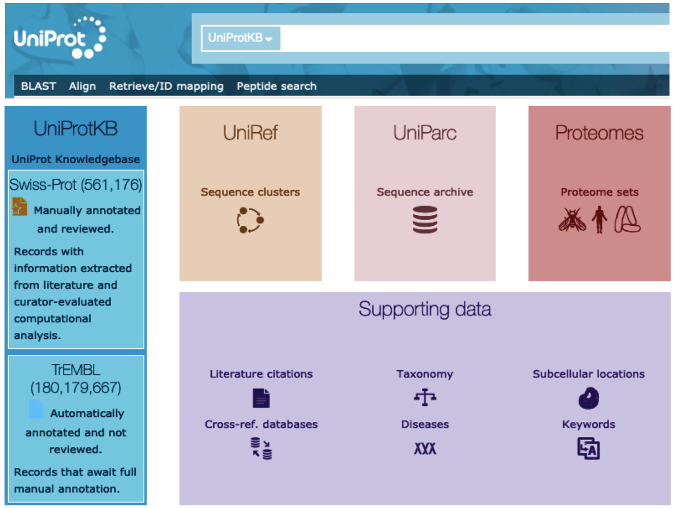
Secondary databases

- ✶ annotated structures
- ✶ annotated sequences
- ✶ motifs and domains
- gene families / sets of orthologous genes
- controlled vocabularies
- genome databases
- organism-specific databases
- ✶ pathways
- specialized databases
- …

UniProt (Universal Protein Resource)

- is a central repository of protein sequence and their annotation
- is a collaboration between
  - European Bioinformatics Institute (TrEMBL)
  - Swiss Institute of Bioinformatics (Swiss-Prot)
  - Georgetown University (PIR)
- users can
  - search the data using
    - text searches
    - BLAST similarity searches
  - download the data
- data is based on data in primary databases

UniProt (Universal Protein Resource)



UniProtKB

UniProt Knowledgebase

Swiss-Prot (561,176)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (180,179,667)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

Sequence clusters

UniParc

Sequence archive

Proteomes

Proteome sets

Supporting data

Literature citations

Taxonomy

Subcellular locations

Cross-ref. databases

Diseases

Keywords

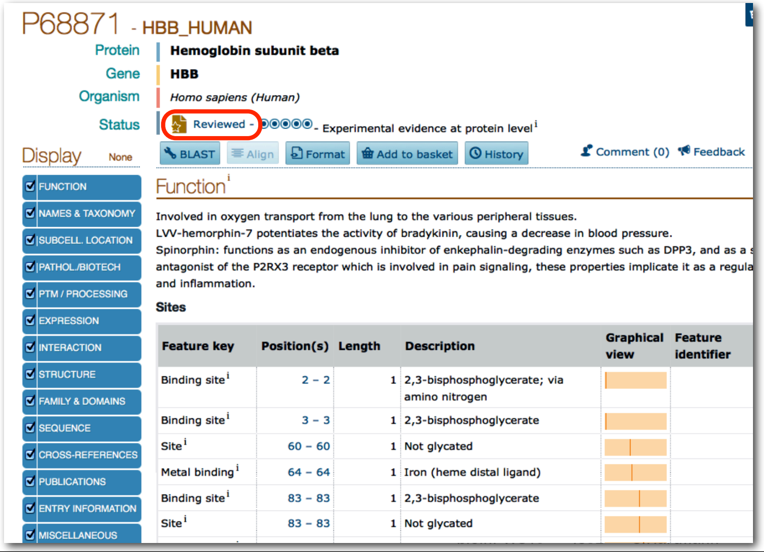
UniProt data: UniProtKB

Swiss-Prot (sprot): manually curated

- information extracted from the literature
- curator-evaluated computational analysis
- information:
  - function, catalytic activity
  - subcellular location
  - structure, posttranslational modification
  - splice variants
  - cross-references to primary & secondary dbs

TrEMBL (trembl): computer-annotated

UniProt format



P68871 - HBB_HUMAN

Protein	**Hemoglobin subunit beta**
Gene	**HBB**
Organism	*Homo sapiens (Human)*
Status	Reviewed - ●●●●● - Experimental evidence at protein level[i]

Function[i]

Involved in oxygen transport from the lung to the various peripheral tissues.
LVV-hemorphin-7 potentiates the activity of bradykinin, causing a decrease in blood pressure.
Spinorphin: functions as an endogenous inhibitor of enkephalin-degrading enzymes such as DPP3, and as a s
antagonist of the P2RX3 receptor which is involved in pain signaling, these properties implicate it as a regula
and inflammation.

Sites

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Binding site[i]	2 – 2	1	2,3-bisphosphoglycerate; via amino nitrogen		
Binding site[i]	3 – 3	1	2,3-bisphosphoglycerate		
Site[i]	60 – 60	1	Not glycated		
Metal binding[i]	64 – 64	1	Iron (heme distal ligand)		
Binding site[i]	83 – 83	1	2,3-bisphosphoglycerate		
Site[i]	83 – 83	1	Not glycated		

UniProt format

```
ID   LGB1_LUPLU              Reviewed;         154 AA.
AC   P02239;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   23-JAN-2007, sequence version 3.
DT   08-APR-2008, entry version 66.
DE   Leghemoglobin-1 (Leghemoglobin I).
OS   Lupinus luteus (European yellow lupin).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
OC   rosids; eurosids I; Fabales; Fabaceae; Papilionoideae; Genisteae;
OC   Lupinus.
OX   NCBI_TaxID=3873;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RX   MEDLINE=87316940; PubMed=3628011; DOI=10.1093/nar/15.16.6742;
RA   Konieczny A.;
RT   "Nucleotide sequence of lupin leghemoglobin I cDNA.";
RL   Nucleic Acids Res. 15:6742-6742(1987).
RN   [2]
RP   NUCLEOTIDE SEQUENCE.
RC   STRAIN=cv. Ventus; TISSUE=Root nodule;
RA   Strozycki P.S.P.;
RL   Submitted (JAN-1994) to the EMBL/GenBank/DDBJ databases.
RN   [3]
RP   NUCLEOTIDE SEQUENCE.
RC   STRAIN=cv. Ventus;
RA   Strozycki P.M., Karlowski W.M., Legocki A.B.;
RT   "Yellow lupine gene coding for leghemoglobin I.";
```

UniProt format

```
RL   (er) Plant Gene Register PGR98-017.
RN   [4]
RP   PROTEIN SEQUENCE OF 2-154.
RC   TISSUE=Root nodule;
RA   Egorov T.A., Feigina M.Y., Kazakov V.K., Shakhparonov M.I.,
RA   Mimaleva S.I., Ovchinnikov Y.A.;
RT   "The complete amino acid sequence of the leghemoglobin I from yellow
RT   lupin root nodules.";
RL   Bioorg. Khim. 2:125-128(1976).
CC   -!- FUNCTION: Provides oxygen to the bacteroids. This role is
CC       essential for symbiotic nitrogen fixation.
CC   -!- SUBUNIT: Monomer.
CC   -!- TISSUE SPECIFICITY: Root nodules.
CC   -!- SIMILARITY: Belongs to the plant globin family.
CC   -----------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   -----------------------------------------------------------------------
DR   EMBL; Y00401; CAA68462.1; -; mRNA.
DR   EMBL; X77043; CAA54332.1; -; mRNA.
DR   EMBL; U50083; AAC04853.1; -; Genomic_DNA.
DR   PIR; A26808; GPYL.
DR   HSSP; P02240; 2GDM.
DR   SMR; P02239; 2-154.
DR   InterPro; IPR012292; Globin.
DR   InterPro; IPR000971; Globin_subset.
DR   InterPro; IPR001032; Leghaemoglobin.
DR   Gene3D; G3DSA:1.10.490.10; Globin_related; 1.
```

UniProt format

```
DR   Pfam; PF00042; Globin; 1.
DR   PRINTS; PR00188; PLANTGLOBIN.
DR   PROSITE; PS01033; GLOBIN; 1.
DR   PROSITE; PS00208; PLANT_GLOBIN; 1.
PE   1: Evidence at protein level;
KW   Direct protein sequencing; Heme; Iron; Metal-binding;
KW   Nitrogen fixation; Oxygen transport; Transport.
FT   INIT_MET      1      1       Removed.
FT   CHAIN         2    154       Leghemoglobin-1.
FT                                /FTId=PRO_0000192984.
FT   METAL        64     64       Iron (heme distal ligand) (By
FT                                similarity).
FT   METAL        98     98       Iron (heme proximal ligand) (By
FT                                similarity).
FT   CONFLICT     80     80       Q -> E (in Ref. 4; AA sequence).
FT   CONFLICT    121    121       E -> G (in Ref. 1; CAA68462).
SQ   SEQUENCE   154 AA;  16753 MW;  58101C830CB21F14 CRC64;
     MGVLTDVQVA LVKSSFEEFN ANIPKNTHRF FTLVLEIAPG AKDLFSFLKG SSEVPQNNPD
     LQAHAGKVFK LTYEAAIQLQ VNGAVASDAT LKSLGSVHVS KGVVDAHFPV VKEAILKTIK
     EVVGDKWSEE LNTAWTIAYD ELAIIIKKEM KDAA
//
```

http://web.expasy.org/docs/userman.html

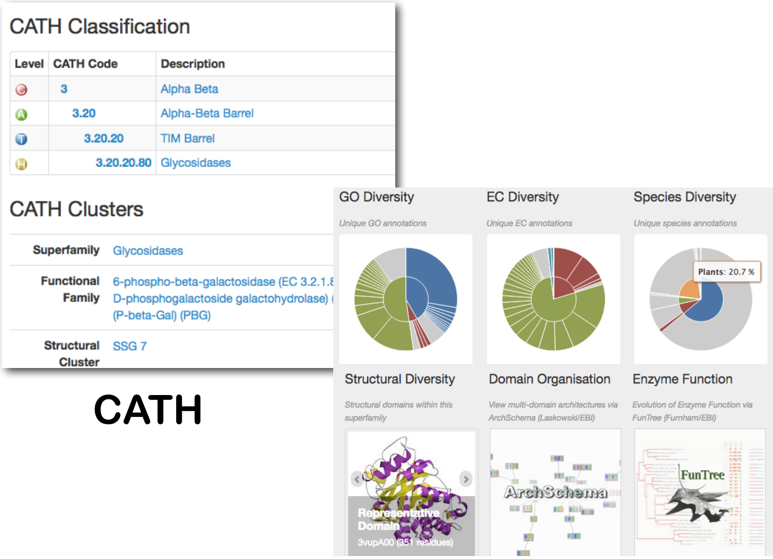
SCOPe, CATH, and others

- manual & automated curation of PDB entries
- description of the structural and evolutionary relationships between known structures
- hierarchical classification of structures
    - unit for analysis / classification: domain
        - structural unit
        - evolutionary building block
        - often multiple domains per protein
    - comparison of domains (sequences, structures)
    - grouping of similar domains
- different definitions & methods: different results!

example: CATH

- protein structures from PDB: split up into domains

C	class	architectures with similar content of secondary structure (e.g., mainly alpha or beta, mixed, etc)
A	architecture	topologies that share a roughly similar spatial arrangement of secondary structures
T	topology	homologous superfamilies that share the same fold; no clear evidence for evolutionary relationship
H	homologous superfamily	domains that share a clear common ancestor



CATH Classification

Level	CATH Code	Description
C	3	Alpha Beta
A	3.20	Alpha-Beta Barrel
T	3.20.20	TIM Barrel
H	3.20.20.80	Glycosidases

CATH Clusters

Superfamily	Glycosidases
Functional Family	6-phospho-beta-galactosidase (EC 3.2.1.8 D-phosphogalactoside galactohydrolase) ((P-beta-Gal) (PBG)
Structural Cluster	SSG 7

GO Diversity

Unique GO annotations

EC Diversity

Unique EC annotations

Species Diversity

Unique species annotations

Structural Diversity

Structural domains within this superfamily

Domain Organisation

View multi-domain architectures via ArchSchema (Laskowski/EBI)

Enzyme Function

Evolution of Enzyme Function via FunTree (Furnham/EBI)

CATH

Today’s exercise

- online databases: sequences, motifs/domains
  - NCBI: GenBank
  - UniProt
  - optional: PDB, CATH
- take notes as you complete the exercise
  - information, available cross-links
  - relevant information about the site
  - your experience with a site
  - results, answers to questions

Key terms and concepts

- flat-file vs. relational databases
- computer-readable database entries
- primary/raw vs. annotated/curated data
  - content, examples
- GenBank, UniProt (sprot, trembl)
- GenBank: IDs, accession number
- PDB, CATH
- genbank/uniprot, fasta format for sequence data