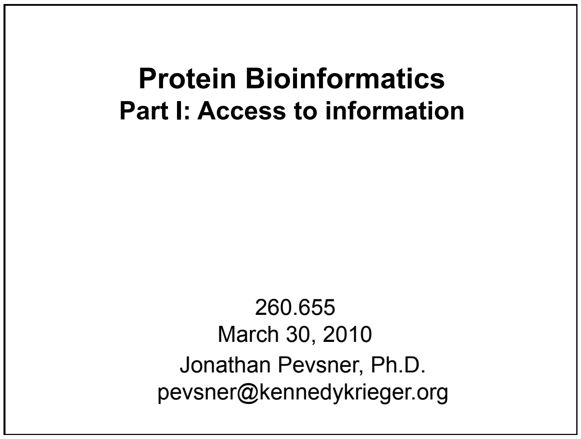
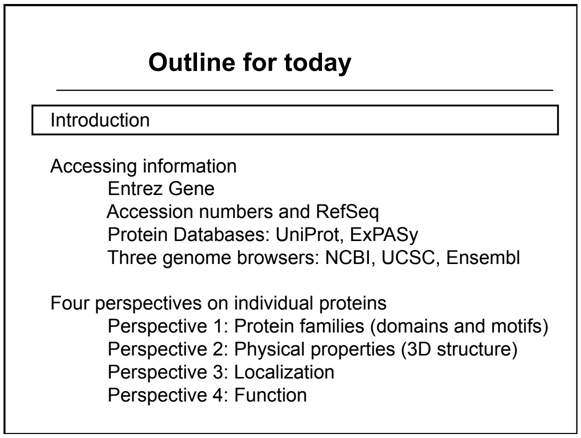


Protein Bioinformatics
Part I: Access to information

260.655
March 30, 2010

Jonathan Pevsner, Ph.D.
pevsner@kennedykrieger.org

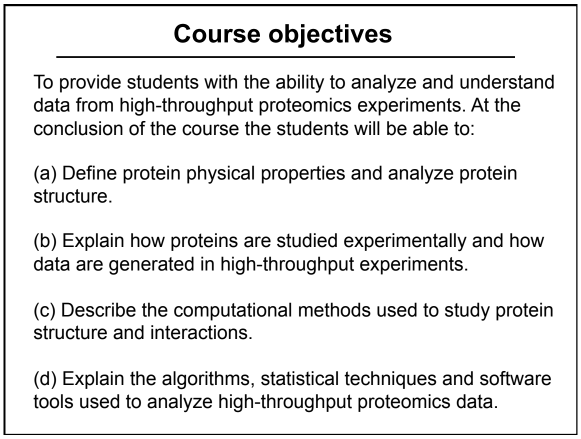


Outline for today

Introduction

Accessing information
- Entrez Gene
- Accession numbers and RefSeq
- Protein Databases: UniProt, ExPASy
- Three genome browsers: NCBI, UCSC, Ensembl

Four perspectives on individual proteins
- Perspective 1: Protein families (domains and motifs)
- Perspective 2: Physical properties (3D structure)
- Perspective 3: Localization
- Perspective 4: Function



Course objectives

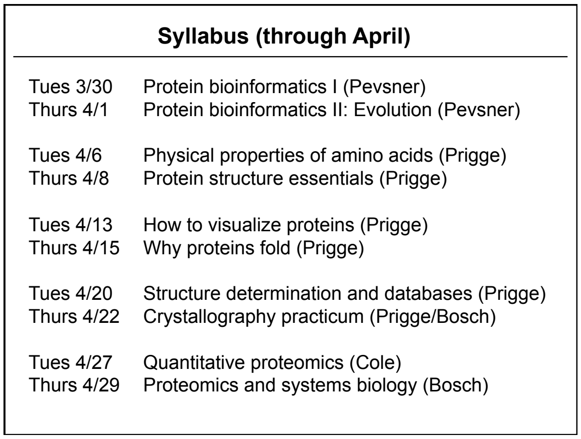
To provide students with the ability to analyze and understand data from high-throughput proteomics experiments. At the conclusion of the course the students will be able to:

(a) Define protein physical properties and analyze protein structure.

(b) Explain how proteins are studied experimentally and how data are generated in high-throughput experiments.

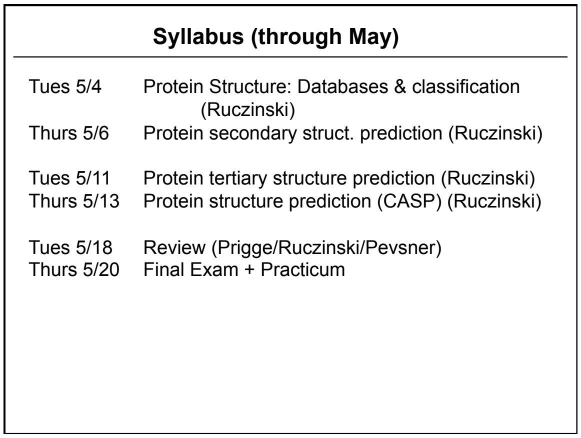
(c) Describe the computational methods used to study protein structure and interactions.

(d) Explain the algorithms, statistical techniques and software tools used to analyze high-throughput proteomics data.



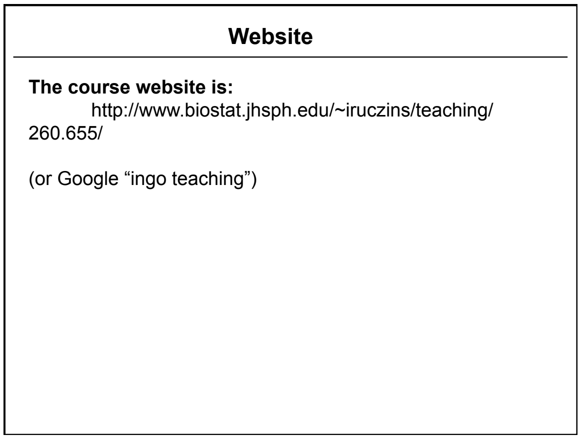
Syllabus (through April)

Tues 3/30	Protein bioinformatics I (Pevsner)
Thurs 4/1	Protein bioinformatics II: Evolution (Pevsner)
Tues 4/6	Physical properties of amino acids (Prigge)
Thurs 4/8	Protein structure essentials (Prigge)
Tues 4/13	How to visualize proteins (Prigge)
Thurs 4/15	Why proteins fold (Prigge)
Tues 4/20	Structure determination and databases (Prigge)
Thurs 4/22	Crystallography practicum (Prigge/Bosch)
Tues 4/27	Quantitative proteomics (Cole)
Thurs 4/29	Proteomics and systems biology (Bosch)



Syllabus (through May)

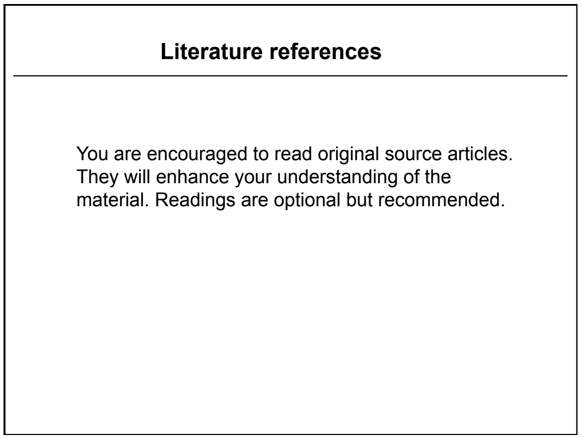
Tues 5/4	Protein Structure: Databases & classification (Ruczinski)
Thurs 5/6	Protein secondary struct. prediction (Ruczinski)
Tues 5/11	Protein tertiary structure prediction (Ruczinski)
Thurs 5/13	Protein structure prediction (CASP) (Ruczinski)
Tues 5/18	Review (Prigge/Ruczinski/Pevsner)
Thurs 5/20	Final Exam + Practicum



Website

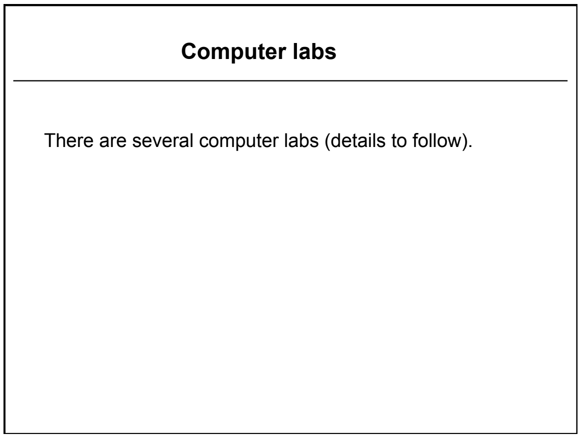
The course website is:
http://www.biostat.jhsph.edu/~iruczins/teaching/260.655/

(or Google “ingo teaching”)



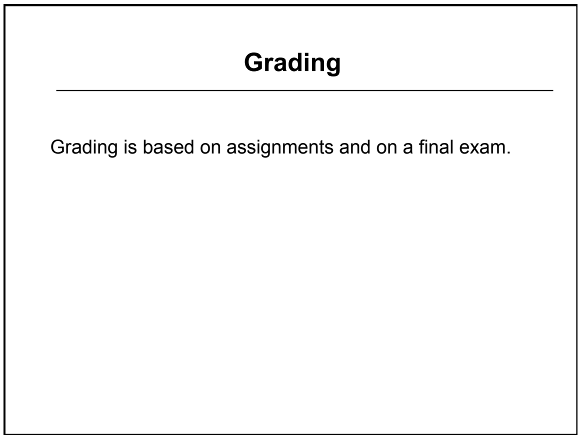
Literature references

You are encouraged to read original source articles. They will enhance your understanding of the material. Readings are optional but recommended.



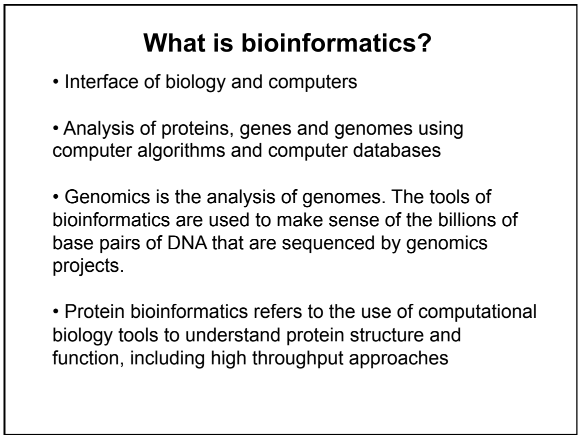
Computer labs

There are several computer labs (details to follow).



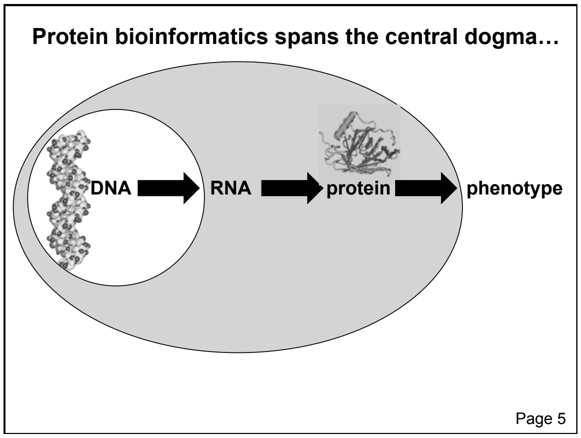
Grading

Grading is based on assignments and on a final exam.



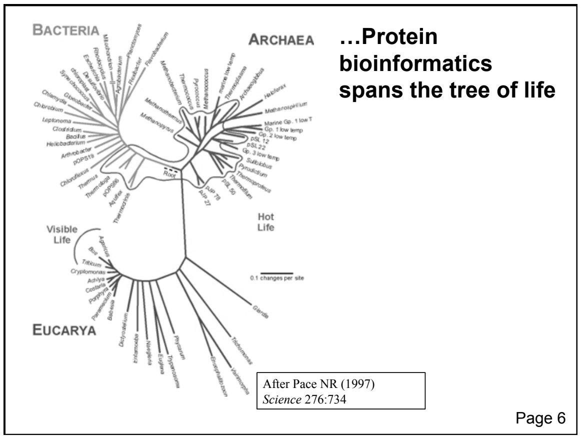
What is bioinformatics?

- Interface of biology and computers
- Analysis of proteins, genes and genomes using computer algorithms and computer databases
- Genomics is the analysis of genomes. The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.
- Protein bioinformatics refers to the use of computational biology tools to understand protein structure and function, including high throughput approaches



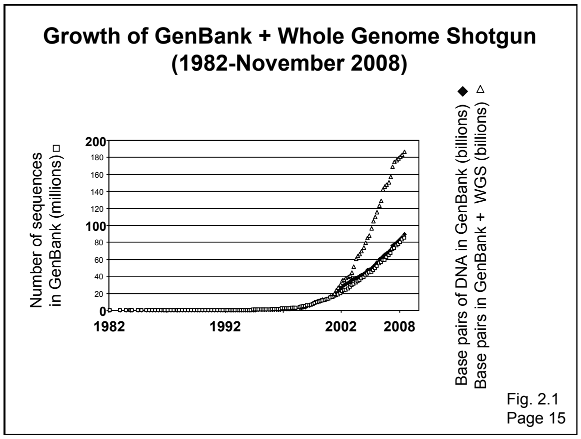
Protein bioinformatics spans the central dogma…

DNA → RNA → protein → phenotype



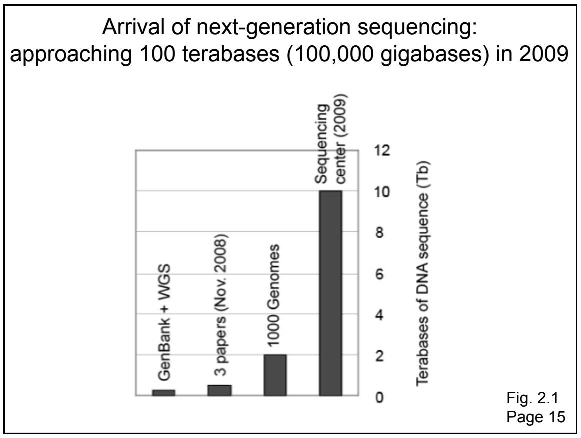
…Protein bioinformatics spans the tree of life

After Pace NR (1997) *Science* 276:734



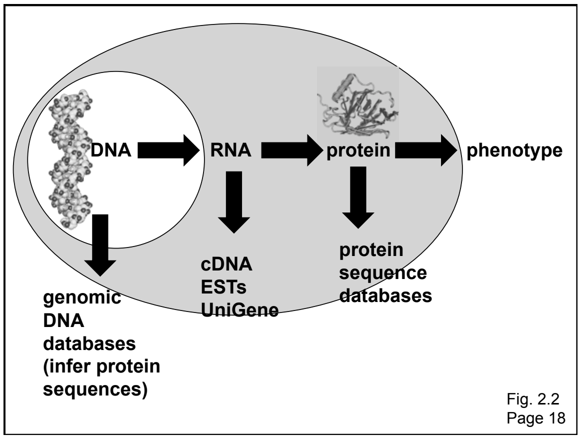
Growth of GenBank + Whole Genome Shotgun (1982-November 2008)

Fig. 2.1
Page 15



Arrival of next-generation sequencing: approaching 100 terabases (100,000 gigabases) in 2009

Fig. 2.1
Page 15



DNA

RNA

protein

phenotype

genomic DNA databases (infer protein sequences)

cDNA
ESTs
UniGene

protein sequence databases

Fig. 2.2
Page 18

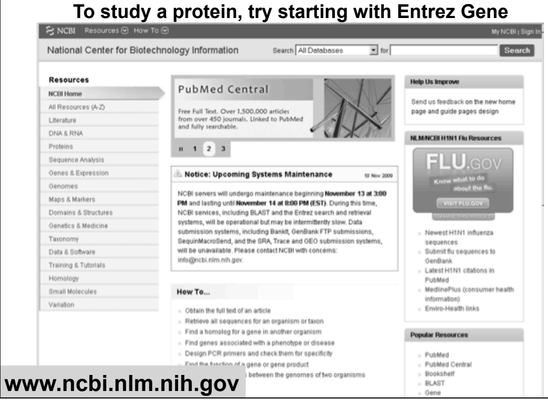
Outline for today

Introduction

Accessing information
- Entrez Gene
- Accession numbers and RefSeq
- Protein Databases: UniProt, ExPASy
- Three genome browsers: NCBI, UCSC, Ensembl

Four perspectives on individual proteins
- Perspective 1: Protein families (domains and motifs)
- Perspective 2: Physical properties (3D structure)
- Perspective 3: Localization
- Perspective 4: Function

New NCBI homepage (November 2009): To study a protein, try starting with Entrez Gene



Free Full Text. Over 1,500,000 articles from over 450 journals. Linked to PubMed and fully searchable.

Notice: Upcoming Systems Maintenance — 10 Nov 2009

NCBI servers will undergo maintenance beginning **November 13 at 3:00 PM** and lasting until **November 14 at 8:00 PM (EST)**. During this time, NCBI services, including BLAST and the Entrez search and retrieval systems, will be operational but may be intermittently slow. Data submission systems, including BankIt, GenBank FTP submissions, SequinMacroSend, and the SRA, Trace and GEO submission systems, will be unavailable. Please contact NCBI with concerns: info@ncbi.nlm.nih.gov.

How To...
- Obtain the full text of an article
- Retrieve all sequences for an organism or taxon
- Find a homolog for a gene in another organism
- Find genes associated with a phenotype or disease
- Design PCR primers and check them for specificity
- Find the function of a gene or gene product
- between the genomes of two organisms

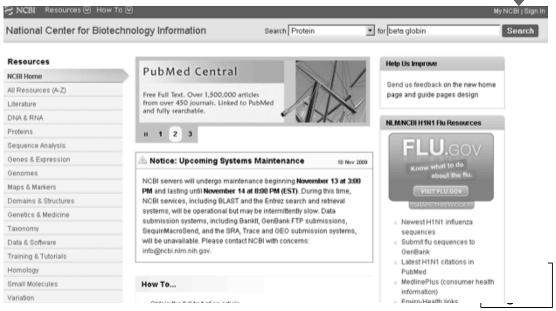
Send us feedback on the new home page and guide pages design

- Newest H1N1 influenza sequences
- Submit flu sequences to GenBank
- Latest H1N1 citations in PubMed
- MedlinePlus (consumer health information)
- Enviro-Health links

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene

www.ncbi.nlm.nih.gov

From the NCBI home page, type "beta globin" and hit "Search"



Free Full Text. Over 1,500,000 articles from over 450 journals. Linked to PubMed and fully searchable.

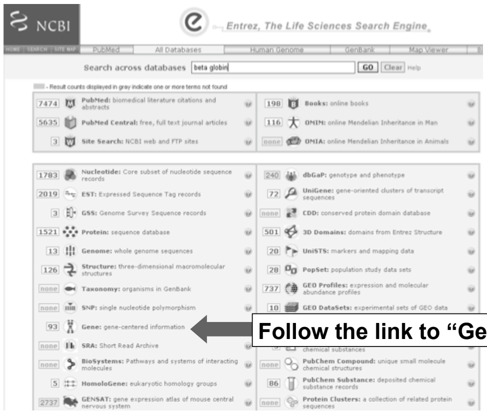
Notice: Upcoming Systems Maintenance — 10 Nov 2009

NCBI servers will undergo maintenance beginning **November 13 at 3:00 PM** and lasting until **November 14 at 8:00 PM (EST)**. During this time, NCBI services, including BLAST and the Entrez search and retrieval systems, will be operational but may be intermittently slow. Data submission systems, including BankIt, GenBank FTP submissions, SequinMacroSend, and the SRA, Trace and GEO submission systems, will be unavailable. Please contact NCBI with concerns: info@ncbi.nlm.nih.gov.

How To...

Send us feedback on the new home page and guide pages design

- Newest H1N1 influenza sequences
- Submit flu sequences to GenBank
- Latest H1N1 citations in PubMed
- MedlinePlus (consumer health information)
- Enviro-Health links



Search across databases beta globin GO Clear Help

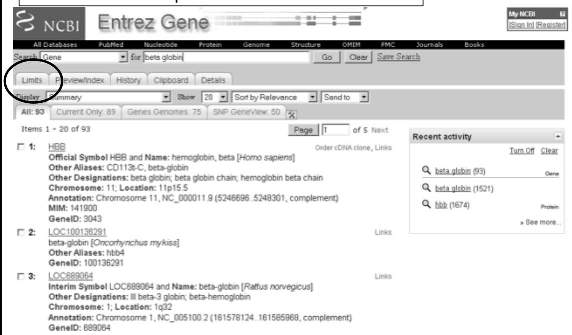
- Result counts displayed in gray indicate one or more terms not found

7474	**PubMed:** biomedical literature citations and abstracts	198	**Books:** online books
5635	**PubMed Central:** free, full text journal articles	116	**OMIM:** online Mendelian Inheritance in Man
3	**Site Search:** NCBI web and FTP sites	none	**OMIA:** online Mendelian Inheritance in Animals
1783	**Nucleotide:** Core subset of nucleotide sequence records	240	**dbGaP:** genotype and phenotype
2019	**EST:** Expressed Sequence Tag records	72	**UniGene:** gene-oriented clusters of transcript sequences
3	**GSS:** Genome Survey Sequence records	none	**CDD:** conserved protein domain database
1521	**Protein:** sequence database	501	**3D Domains:** domains from Entrez Structure
13	**Genome:** whole genome sequences	20	**UniSTS:** markers and mapping data
126	**Structure:** three-dimensional macromolecular structures	28	**PopSet:** population study data sets
none	**Taxonomy:** organisms in GenBank	737	**GEO Profiles:** expression and molecular abundance profiles
none	**SNP:** single nucleotide polymorphism	10	**GEO DataSets:** experimental sets of GEO data
93	**Gene:** gene-centered information		
none	**SRA:** Short Read Archive		chemical substances
none	**BioSystems:** Pathways and systems of interacting molecules	none	**PubChem Compound:** unique small molecule chemical structures
5	**HomoloGene:** eukaryotic homology groups	86	**PubChem Substance:** deposited chemical substance records
2737	**GENSAT:** gene expression atlas of mouse central nervous system	none	**Protein Clusters:** a collection of related protein sequences

Follow the link to "Gene"

Fig. 2.5
Page 28

Entrez Gene is in the header
Note the "Official Symbol" HBB for beta globin
Note the "limits" option



Search Gene for beta globin

Display Summary Show 20 Sort by Relevance

All: 93 Current Only: 89 Genes Genomes: 75 SNP GeneView: 50

Items 1 - 20 of 93 Page 1 of 5

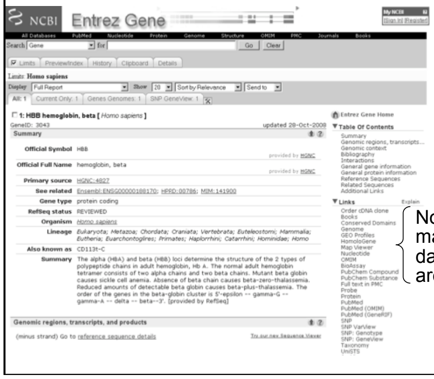
□ 1: HBB Order cDNA clone, Links
Official Symbol HBB and **Name:** hemoglobin, beta [*Homo sapiens*]
Other Aliases: CD113t-C, beta-globin
Other Designations: beta globin; beta globin chain; hemoglobin beta chain
Chromosome: 11; **Location:** 11p15.5
Annotation: Chromosome 11, NC_000011.9 (5246696..5248301, complement)
MIM: 141900
GeneID: 3043

□ 2: LOC100136291 Links
beta-globin [*Oncorhynchus mykiss*]
Other Aliases: hbb4
GeneID: 100136291

□ 3: LOC689064 Links
Interim Symbol LOC689064 and **Name:** beta-globin [*Rattus norvegicus*]
Other Designations: ll beta-3 globin; beta-hemoglobin
Chromosome: 1; **Location:** 1q32
Annotation: Chromosome 1, NC_005100.2 (161578124..161585968, complement)
GeneID: 689064

beta globin (93) Gene
beta globin (1521)
hbb (1674) Protein

Entrez Gene (top of page)



Search Gene for

☑ Limits

Limits: Homo sapiens

Display Full Report Show 20 Sort by Relevance

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

□ 1: HBB hemoglobin, beta [*Homo sapiens*]

GeneID: 3043 updated 28-Oct-2008

Summary

Official Symbol	HBB provided by HGNC
Official Full Name	hemoglobin, beta provided by HGNC
Primary source	HGNC:4827
See related	Ensembl:ENSG00000188170; HPRD:00786; MIM:141900
Gene type	protein coding
RefSeq status	REVIEWED
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	CD113t-C
Summary	The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'. [provided by RefSeq]

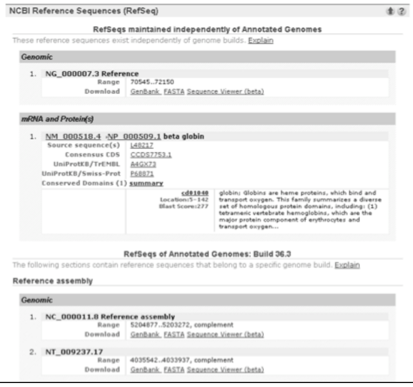
Genomic regions, transcripts, and products

(minus strand) Go to reference sequence details Try our new Sequence Viewer

Note that links to many other HBB database entries are available

Page 30

Entrez Gene (bottom of page): RefSeq accession numbers



NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

Genomic

1. **NG_000007.3 Reference**
   Range 70545..72150
   Download GenBank FASTA Sequence Viewer (beta)

mRNA and Protein(s)

1. NM_000518.4 NP_000509.1 **beta globin**
   Source sequence(s) L48217
   Consensus CDS CCDS7753.1
   UniProtKB/TrEMBL A4GX73
   UniProtKB/Swiss-Prot P68871
   Conserved Domains (1) **summary**
   cd01040 Location:5-142 Blast Score:277 globin: Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen...

RefSeqs of Annotated Genomes: Build 36.3

The following sections contain reference sequences that belong to a specific genome build. Explain

Reference assembly

Genomic

1. **NC_000011.8 Reference assembly**
   Range 5204877..5203272, complement
   Download GenBank FASTA Sequence Viewer (beta)
2. **NT_009237.17**
   Range 4035542..4033937, complement
   Download GenBank FASTA Sequence Viewer (beta)

Outline for today

Introduction

Accessing information
- Entrez Gene
- Accession numbers and RefSeq
- Protein Databases: UniProt, ExPASy
- Three genome browsers: NCBI, UCSC, Ensembl

Four perspectives on individual proteins
- Perspective 1: Protein families (domains and motifs)
- Perspective 2: Physical properties (3D structure)
- Perspective 3: Localization
- Perspective 4: Function

Access to sequences: Entrez Gene at NCBI

Entrez Gene is a great starting point: it collects key information on each gene/protein from major databases. It covers all major organisms.

RefSeq provides a curated, optimal accession number for each DNA (NM_000518 for beta globin DNA corresponding to mRNA) or protein (NP_000509)

Page 29

Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences.
You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.

Page 26

What is an accession number?

An accession number is label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	**DNA**
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	**RNA**
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	**protein**
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

Page 27

NCBI's important RefSeq project: best representative sequences

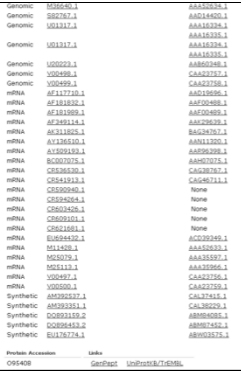
RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon "reference" version of a sequence.

RefSeq identifiers include the following formats:

Complete genome	NC_######
Complete chromosome	NC_######
Genomic contig	NT_######
mRNA (DNA format)	NM_###### e.g. NM_006744
Protein	NP_###### e.g. NP_006735

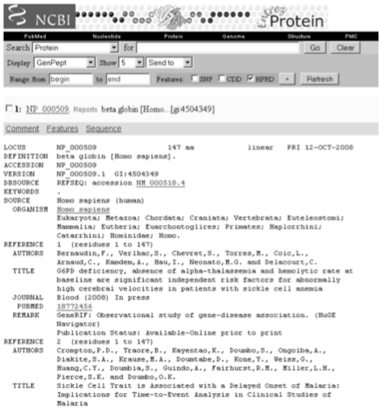
Page 27

Entrez Gene (bottom of page): non-RefSeq accessions
(it’s unclear what these are, highlighting usefulness of RefSeq)



Genomic	M36640.1	AAA52634.1
Genomic	S82767.1	AAD14420.1
Genomic	U01317.1	AAA16334.1 AAA16335.1
Genomic	U01317.1	AAA16334.1 AAA16335.1
Genomic	U20223.1	AAB60348.1
Genomic	V00498.1	CAA23757.1
Genomic	V00499.1	CAA23758.1
mRNA	AF117710.1	AAD19696.1
mRNA	AF181832.1	AAF00488.1
mRNA	AF181989.1	AAF00489.1
mRNA	AF349114.1	AAK29639.1
mRNA	AK311825.1	BAG34767.1
mRNA	AY136510.1	AAN11320.1
mRNA	AY509193.1	AAR96398.1
mRNA	BC007075.1	AAH07075.1
mRNA	CR536530.1	CAG38767.1
mRNA	CR541913.1	CAG46711.1
mRNA	CR590940.1	None
mRNA	CR594264.1	None
mRNA	CR603426.1	None
mRNA	CR609101.1	None
mRNA	CR621681.1	None
mRNA	EU694432.1	ACD39349.1
mRNA	M11428.1	AAA52633.1
mRNA	M25079.1	AAA35597.1
mRNA	M25113.1	AAA35966.1
mRNA	V00497.1	CAA23756.1
mRNA	V00500.1	CAA23759.1
Synthetic	AM392537.1	CAL37415.1
Synthetic	AM393351.1	CAL38229.1
Synthetic	DQ893159.2	ABM84085.1
Synthetic	DQ896453.2	ABM87452.1
Synthetic	EU176774.1	ABW03575.1

Protein Accession	Links
O95408	GenPept UniProtKB/TrEMBL



Search Protein for

Display GenPept Show 5 Send to

Range: from begin to end Features: □ SNP □ CDD ☑ HPRD

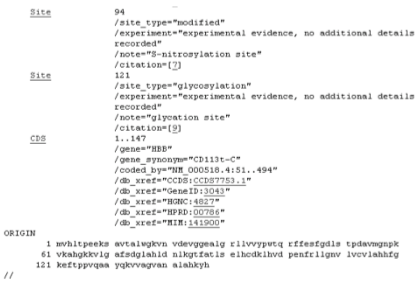
□ 1: NP_000509. Reports beta globin [Homo...[gi:4504349]

```
LOCUS       NP_000509               147 aa            linear   PRI 12-OCT-2008
DEFINITION  beta globin [Homo sapiens].
ACCESSION   NP_000509
VERSION     NP_000509.1  GI:4504349
DBSOURCE    REFSEQ: accession NM_000518.4
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 147)
  AUTHORS   Bernaudin,F., Verlhac,S., Chevret,S., Torres,M., Coic,L.,
            Arnaud,C., Kamdem,A., Hau,I., Neonato,M.G. and Delacourt,C.
  TITLE     G6PD deficiency, absence of alpha-thalassemia and hemolytic rate at
            baseline are significant independent risk factors for abnormally
            high cerebral velocities in patients with sickle cell anemia
  JOURNAL   Blood (2008) In press
   PUBMED   18772456
  REMARK    GeneRIF: Observational study of gene-disease association. (HuGE
            Navigator)
            Publication Status: Available-Online prior to print
REFERENCE   2  (residues 1 to 147)
  AUTHORS   Crompton,P.D., Traore,B., Kayentao,K., Doumbo,S., Ongoiba,A.,
            Diakite,S.A., Krause,M.A., Doumtabe,D., Kone,Y., Weiss,G.,
            Huang,C.Y., Doumbia,S., Guindo,A., Fairhurst,R.M., Miller,L.H.,
            Pierce,S.K. and Doumbo,O.K.
  TITLE     Sickle Cell Trait is Associated with a Delayed Onset of Malaria:
            Implications for Time-to-Event Analysis in Clinical Studies of
            Malaria
```

Entrez Protein:
accession,
organism,
literature…

Fig. 2.8
Page 31

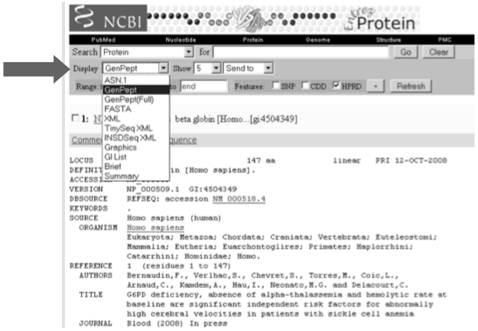
Entrez Protein:
…features of a protein, and its sequence
in the one-letter amino acid code



```
     Site            94
                     /site_type="modified"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="S-nitrosylation site"
                     /citation=[7]
     Site            121
                     /site_type="glycosylation"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="glycation site"
                     /citation=[9]
     CDS             1..147
                     /gene="HBB"
                     /gene_synonym="CD113t-C"
                     /coded_by="NM_000518.4:51..494"
                     /db_xref="CCDS:CCDS7753.1"
                     /db_xref="GeneID:3043"
                     /db_xref="HGNC:4827"
                     /db_xref="HPRD:00786"
                     /db_xref="MIM:141900"
ORIGIN
        1 mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
       61 vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg
      121 keftppvqaa yqkvvagvan alahkyh
//
```

Fig. 2.8
Page 31

Entrez Protein:
You can change the display (as shown)…

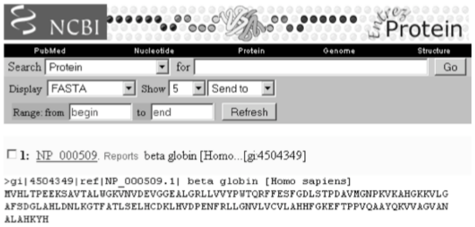


□ 1: N beta globin [Homo...[gi:4504349]

```
LOCUS                        147 aa          linear   PRI 12-OCT-2008
DEFINIT           in [Homo sapiens].
ACCESSI
VERSION     NP_000509.1  GI:4504349
DBSOURCE    REFSEQ: accession NM_000518.4
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 147)
  AUTHORS   Bernaudin,F., Verlhac,S., Chevret,S., Torres,M., Coic,L.,
            Arnaud,C., Kamdem,A., Hau,I., Neonato,M.G. and Delacourt,C.
  TITLE     G6PD deficiency, absence of alpha-thalassemia and hemolytic rate at
            baseline are significant independent risk factors for abnormally
            high cerebral velocities in patients with sickle cell anemia
  JOURNAL   Blood (2008) In press
```

Page 31

FASTA format:
versatile, compact with one header line
followed by a string of nucleotides or amino acids
in the single letter code



□ 1: NP_000509. Reports beta globin [Homo...[gi:4504349]

```
>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Fig. 2.9
Page 32

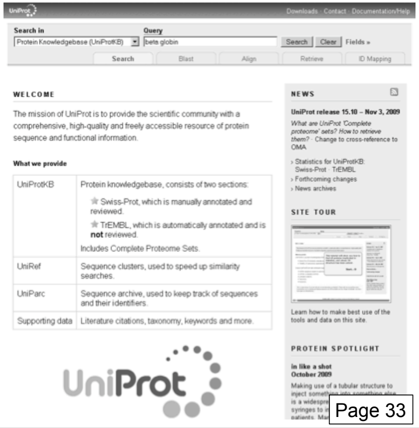
Outline for today

Introduction

Accessing information
- Entrez Gene
- Accession numbers and RefSeq
- Protein Databases: UniProt, ExPASy
- Three genome browsers: NCBI, UCSC, Ensembl

Four perspectives on individual proteins
- Perspective 1: Protein families (domains and motifs)
- Perspective 2: Physical properties (3D structure)
- Perspective 3: Localization
- Perspective 4: Function

UniProt:
a centralized
protein
database
(uniprot.org)



WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ★ TrEMBL, which is automatically annotated and is **not** reviewed. Includes Complete Proteome Sets.
UniRef	Sequence clusters, used to speed up similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords and more.

NEWS

UniProt release 15.10 – Nov 3, 2009

What are UniProt 'Complete proteome' sets? How to retrieve them? · Change to cross-reference to OMA

- Statistics for UniProtKB: Swiss-Prot · TrEMBL
- Forthcoming changes
- News archives

SITE TOUR

Learn how to make best use of the tools and data on this site.

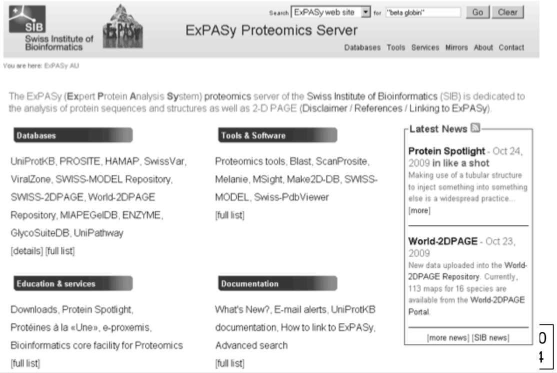
PROTEIN SPOTLIGHT

in like a shot
October 2009

Making use of a tubular structure to inject something into something else is a widesprea syringes to i patients. Ma

Page 33

ExPASy: vast proteomics resources (www.expasy.ch)



You are here: ExPASy AU

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) **proteomics** server of the **Swiss Institute of Bioinformatics** (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (**Disclaimer** / **References** / **Linking to ExPASy**).

Databases

UniProtKB, PROSITE, HAMAP, SwissVar, ViralZone, SWISS-MODEL Repository, SWISS-2DPAGE, World-2DPAGE Repository, MIAPEGelDB, ENZYME, GlycoSuiteDB, UniPathway
[details] [full list]

Tools & Software

Proteomics tools, Blast, ScanProsite, Melanie, MSight, Make2D-DB, SWISS-MODEL, Swiss-PdbViewer
[full list]

Education & services

Downloads, Protein Spotlight, Protéines à la «Une», e-proxemis, Bioinformatics core facility for Proteomics
[full list]

Documentation

What's New?, E-mail alerts, UniProtKB documentation, How to link to ExPASy, Advanced search
[full list]

Latest News

Protein Spotlight - Oct 24, 2009 **in like a shot**
Making use of a tubular structure to inject something into something else is a widespread practice...
[more]

World-2DPAGE - Oct 23, 2009
New data uploaded into the **World-2DPAGE Repository**. Currently, 113 maps for 16 species are available from the **World-2DPAGE Portal**.

[more news] [SIB news]

0
4

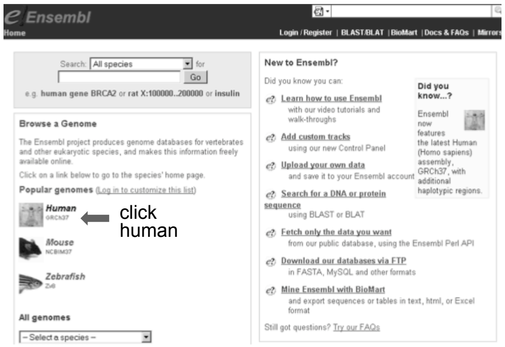
Outline for today

Introduction

Accessing information
- Entrez Gene
- Accession numbers and RefSeq
- Protein Databases: UniProt, ExPASy
- Three genome browsers: NCBI, UCSC, Ensembl

Four perspectives on individual proteins
- Perspective 1: Protein families (domains and motifs)
- Perspective 2: Physical properties (3D structure)
- Perspective 3: Localization
- Perspective 4: Function

Ensembl genome browser (www.ensembl.org)


Ensembl
Home
Login / Register | BLAST/BLAT | BioMart | Docs & FAQs | Mirrors
Go

Search: All species for

e.g. **human gene BRCA2** or **rat X:100000..200000** or **insulin**

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

Human GRCh37

Mouse NCBIM37

Zebrafish Zv8

click human

All genomes

-- Select a species --

New to Ensembl?

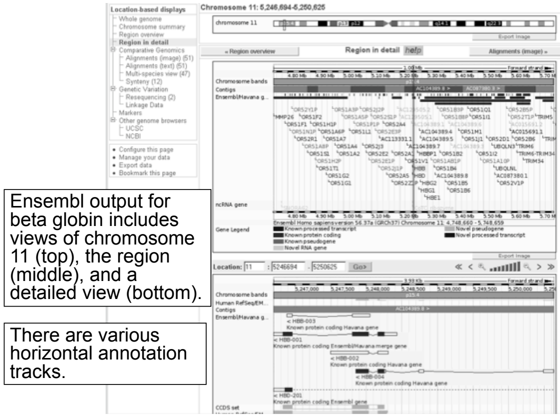
Did you know you can:

- Learn how to use Ensembl with our video tutorials and walk-throughs
- Add custom tracks using our new Control Panel
- Upload your own data and save it to your Ensembl account
- Search for a DNA or protein sequence using BLAST or BLAT
- Fetch only the data you want from our public database, using the Ensembl Perl API
- Download our databases via FTP in FASTA, MySQL and other formats
- Mine Ensembl with BioMart and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs

Did you know...?

Ensembl now features the latest Human (Homo sapiens) assembly, GRCh37, with additional haplotypic regions.



Chromosome 11: 5,246,694-5,250,625

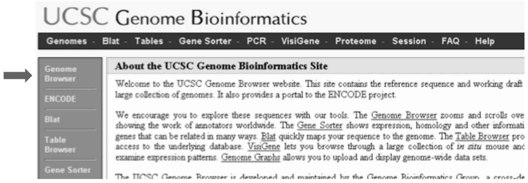
Ensembl Homo sapiens version 56.37a (GRCh37) Chromosome 11: 4,748,660 - 5,748,659

Location: 11 : 5246694 - 5250625

Ensembl output for beta globin includes views of chromosome 11 (top), the region (middle), and a detailed view (bottom).

There are various horizontal annotation tracks.

[1] Visit http://genome.ucsc.edu/, click Genome Browser


UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Proteome - Session - FAQ - Help
Genome Browser
ENCODE
Blat
Table Browser
Gene Sorter

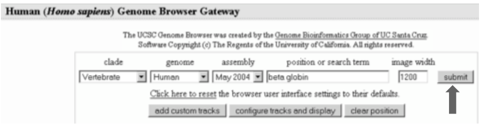
About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft large collection of genomes. It also provides a portal to the ENCODE project.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls ove showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other informati genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser pro access to the underlying database. VisiGene lets you browse through a large collection of *in situ* mouse anc examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-de

[2] Choose organisms, enter query (beta globin), hit submit



Human (*Homo sapiens*) Genome Browser Gateway

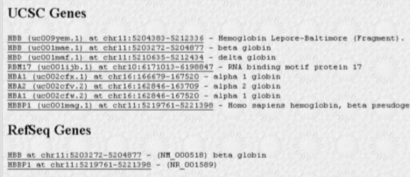
The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term	image width
Vertebrate	Human	May 2004	beta globin	1200

Click here to reset the browser user interface settings to their defaults.

Page 36

[3] Choose the RefSeq beta globin gene



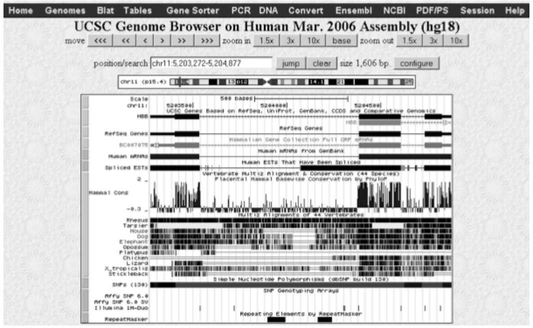
UCSC Genes

HBB (uc009yem.1) at chr11:5204383-5212336 - Hemoglobin Lepore-Baltimore (Fragment).
HBB (uc001mae.1) at chr11:5203272-5204877 - beta globin
HBD (uc001maf.1) at chr11:5210635-5212434 - delta globin
RBM17 (uc001ijb.1) at chr10:6171013-6198847 - RNA binding motif protein 17
HBA1 (uc002cfx.1) at chr16:166679-167520 - alpha 1 globin
HBA2 (uc002cfv.2) at chr16:162846-163709 - alpha 2 globin
HBA1 (uc002cfw.2) at chr16:162846-167520 - alpha 1 globin
HBBP1 (uc001mag.1) at chr11:5219761-5221398 - Homo sapiens hemoglobin, beta pseudoge

RefSeq Genes

HBB at chr11:5203272-5204877 - (NM_000518) beta globin
HBBP1 at chr11:5219761-5221398 - (NR_001589)

[4] The UCSC Genome Browser is an essential resource
--choose which tracks to display
--add custom tracks
--the Table Browser is complementary



UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)

position/search chr11:5,203,272-5,204,877 size 1,606 bp.

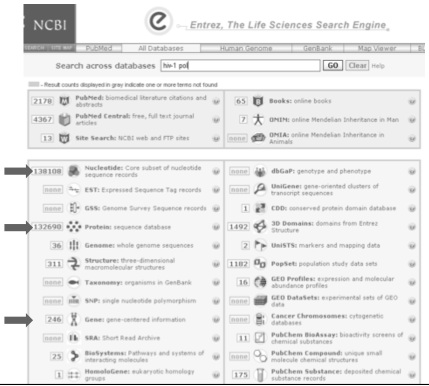
Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump clear configure
chr11 (p15.4)
Scale
chr11:
500 bases
5203500
5204000
5204500
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
HBB
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC007075
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 44 Vertebrates
Rhesus
Tarsier
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 130)
SNPs (130)
SNP Genotyping Arrays
Affy SNP 6.0
Affy SNP 6.0 SV
Illumina 1M-Duo
Repeating Elements by RepeatMasker
RepeatMasker

Example of how to access sequence data: HIV-1 *pol*

There are many possible approaches. Begin at the main page of NCBI, and type an Entrez query: hiv-1 pol

Page 36

Searching for HIV-1 *pol*: >130,000 nucleotide, protein hits



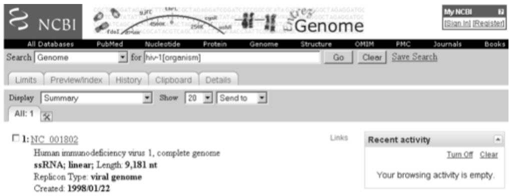
Search across databases: hiv-1 pol

- Result counts displayed in gray indicate one or more terms not found

Count	Database	Count	Database
2178	PubMed: biomedical literature citations and abstracts	65	Books: online books
4367	PubMed Central: free, full text journal articles	7	OMIM: online Mendelian Inheritance in Man
13	Site Search: NCBI web and FTP sites	none	OMIA: online Mendelian Inheritance in Animals
138108	Nucleotide: Core subset of nucleotide sequence records	none	dbGaP: genotype and phenotype
none	EST: Expressed Sequence Tag records	none	UniGene: gene-oriented clusters of transcript sequences
none	GSS: Genome Survey Sequence records	1	CDD: conserved protein domain database
132690	Protein: sequence database	1492	3D Domains: domains from Entrez Structure
36	Genome: whole genome sequences	2	UniSTS: markers and mapping data
311	Structure: three-dimensional macromolecular structures	1182	PopSet: population study data sets
none	Taxonomy: organisms in GenBank	16	GEO Profiles: expression and molecular abundance profiles
none	SNP: single nucleotide polymorphism	none	GEO DataSets: experimental sets of GEO data
246	Gene: gene-centered information	none	Cancer Chromosomes: cytogenetic databases
none	SRA: Short Read Archive	11	PubChem BioAssay: bioactivity screens of chemical substances
25	BioSystems: Pathways and systems of interacting molecules	none	PubChem Compound: unique small molecule chemical structures
1	HomoloGene: eukaryotic homology groups	175	PubChem Substance: deposited chemical substance records

11/09

Searching for HIV-1 *pol*: using the command hiv-1[organism] limits the output to just one entry



Search: Genome for hiv-1[organism]

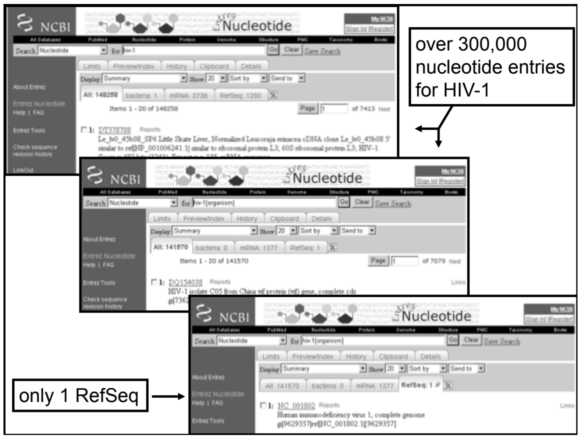
Display: Summary Show: 20 Send to

All: 1

□ 1: NC_001802
Human immunodeficiency virus 1, complete genome
ssRNA; linear; Length: **9,181 nt**
Replicon Type: **viral genome**
Created: **1998/01/22**

Your browsing activity is empty.

Try Taxonomy Browser to easily limit your query to your favorite organism(s). *Example*:
NCBI home → Taxonomy → Taxonomy browser → human → protein to find a human protein



Search: Nucleotide for hiv-1

Display: Summary Show: 20 Sort by Send to

All: 148258 | bacteria: 1 | mRNA: 3738 | RefSeq: 1250

Items 1 - 20 of 148258 Page 1 of 7413

□ 1: DT378788
Le_lv0_45b08_SP6 Little Skate Liver, Normalized Leucoraja erinacea cDNA clone Le_lv0_45b08 5' similar to ref|NP_001006241.1| similar to ribosomal protein L3; 60S ribosomal protein L3; HIV-1.

Search: Nucleotide for hiv-1[organism]

Display: Summary Show: 20 Sort by Send to

All: 141570 | bacteria: 0 | mRNA: 1377 | RefSeq: 1

Items 1 - 20 of 141570 Page 1 of 7079

□ 1: DQ154038
HIV-1 isolate C05 from China vif protein (vif) gene, complete cds

Search: Nucleotide for hiv-1[organism]

Display: Summary Show: 20 Sort by Send to

All: 141570 | bacteria: 0 | mRNA: 1377 | RefSeq: 1

□ 1: NC_001802
Human immunodeficiency virus 1, complete genome
gi|9629357|ref|NC_001802.1|[9629357]

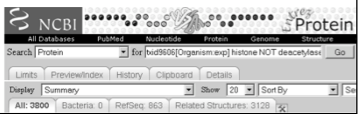
over 300,000 nucleotide entries for HIV-1

only 1 RefSeq

Example of how to access sequence data: histone

query for “histone”	# results
protein records	85,000
RefSeq entries	32,000
RefSeq (limit to human)	1129
NOT deacetylase	863

At this point, select a reasonable candidate (e.g. histone 2, H4) and follow its link to Entrez Gene. There, you can confirm you have the right protein.



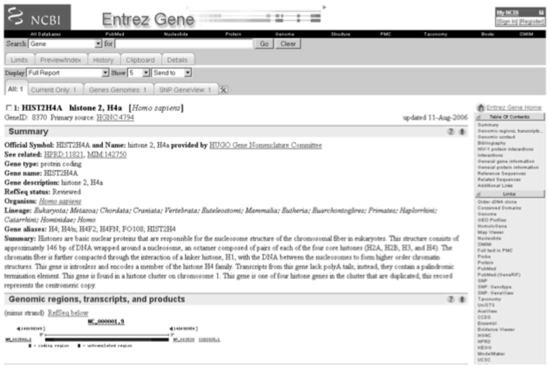
Search Protein for txid9606[Organism:exp] histone NOT deacetylase

Display Summary Show 20 Sort By

All: 3800 Bacteria: 0 RefSeq: 863 Related Structures: 3128

11-09

Entrez Gene result for a histone



Search Gene for

Display Full Report Show 5 Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

□ 1: **HIST2H4A histone 2, H4a** [*Homo sapiens*]

GeneID: 8370 Primary source: HGNC:4794 updated 11-Aug-2006

Summary

Official Symbol: HIST2H4A **and Name:** histone 2, H4a **provided by** HUGO Gene Nomenclature Committee
See related: HPRD:11821, MIM:142750
Gene type: protein coding
Gene name: HIST2H4A
Gene description: histone 2, H4a
RefSeq status: Reviewed
Organism: *Homo sapiens*
Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Gene aliases: H4; H4/n; H4F2; H4FN; FO108; HIST2H4
Summary: Histones are basic nuclear proteins that are responsible for the nucleosome structure of the chromosomal fiber in eukaryotes. This structure consists of approximately 146 bp of DNA wrapped around a nucleosome, an octamer composed of pairs of each of the four core histones (H2A, H2B, H3, and H4). The chromatin fiber is further compacted through the interaction of a linker histone, H1, with the DNA between the nucleosomes to form higher order chromatin structures. This gene is intronless and encodes a member of the histone H4 family. Transcripts from this gene lack polyA tails; instead, they contain a palindromic termination element. This gene is found in a histone cluster on chromosome 1. This gene is one of four histone genes in the cluster that are duplicated; this record represents the centromeric copy.

Genomic regions, transcripts, and products

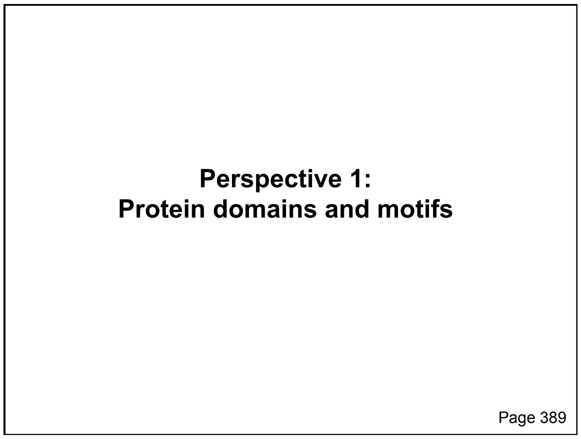
(minus strand) RefSeq below

Outline for today

Introduction

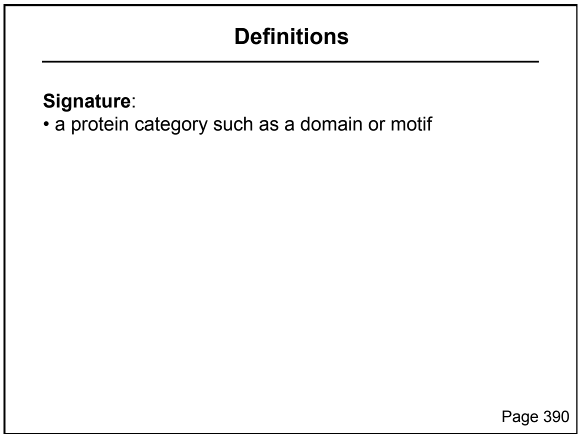
Accessing information
- Entrez Gene
- Accession numbers and RefSeq
- Protein Databases: UniProt, ExPASy
- Three genome browsers: NCBI, UCSC, Ensembl

Four perspectives on individual proteins
- Perspective 1: Protein families (domains and motifs)
- Perspective 2: Physical properties (3D structure)
- Perspective 3: Localization
- Perspective 4: Function



Perspective 1:
Protein domains and motifs

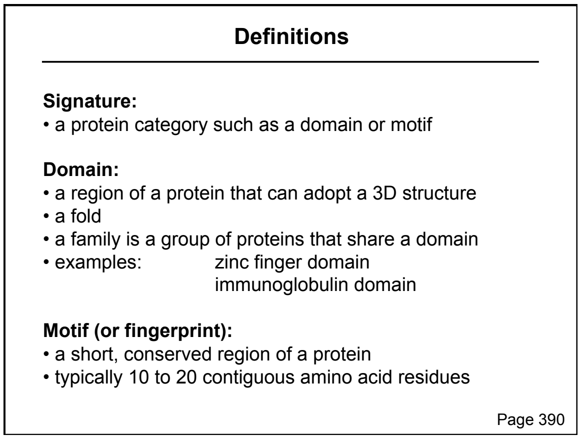
Page 389



Definitions

Signature:
- a protein category such as a domain or motif

Page 390



Definitions

Signature:
- a protein category such as a domain or motif

Domain:
- a region of a protein that can adopt a 3D structure
- a fold
- a family is a group of proteins that share a domain
- examples: zinc finger domain
  immunoglobulin domain

Motif (or fingerprint):
- a short, conserved region of a protein
- typically 10 to 20 contiguous amino acid residues

Page 390

15 most common domains (human)

Zn finger, C2H2 type	1093 proteins
Immunoglobulin	1032
EGF-like	471
Zn-finger, RING	458
Homeobox	417
Pleckstrin-like	405
RNA-binding region RNP-1	400
SH3	394
Calcium-binding EF-hand	392
Fibronectin, type III	300
PDZ/DHR/GLGF	280
Small GTP-binding protein	261
BTB/POZ	236
bHLH	226
Cadherin	226

Source: Integr8 at EBI website

Page 391

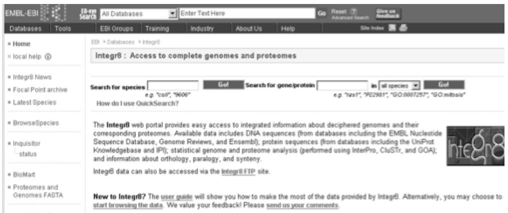
15 most common domains (various species)

The European Bioinformatics Institute (EBI) offers many key proteomics resources at the Integr8 site:

http://www.ebi.ac.uk/proteome/

Page 391

1. Go to the Integr8 site: http://www.ebi.ac.uk/proteome/
2. Browse species; choose *Homo sapiens*.
3. Click “Proteome analysis”
4. Obtain a variety of statistics, such as common repeats, domains, average protein length



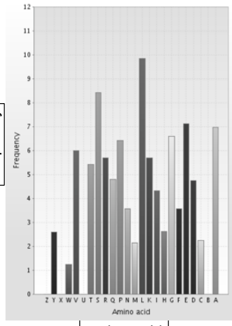
Integr8 : Access to complete genomes and proteomes

The **Integr8** web portal provides easy access to integrated information about deciphered genomes and their corresponding proteomes. Available data includes DNA sequences (from databases including the EMBL Nucleotide Sequence Database, Genome Reviews, and Ensembl); protein sequences (from databases including the UniProt Knowledgebase and IPI); statistical genome and proteome analysis (performed using InterPro, CluSTr, and GOA); and information about orthology, paralogy, and synteny.

Integr8 data can also be accessed via the Integr8 FTP site.

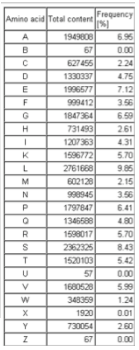
New to Integr8? The user guide will show you how to make the most of the data provided by Integr8. Alternatively, you may choose to start browsing the data. We value your feedback! Please send us your comments.

Amino acid composition



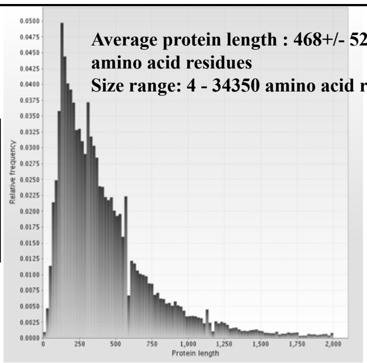
frequency

amino acid



Amino acid	Total content	Frequency [%]
A	1949808	6.95
B	67	0.00
C	627455	2.24
D	1330337	4.75
E	1996577	7.12
F	999412	3.56
G	1847364	6.59
H	731493	2.61
I	1207363	4.31
K	1596772	5.70
L	2761668	9.85
M	602128	2.15
N	998945	3.56
P	1797847	6.41
Q	1346588	4.80
R	1598017	5.70
S	2362325	8.43
T	1520103	5.42
U	57	0.00
V	1680528	5.99
W	348359	1.24
X	1920	0.01
Y	730054	2.60
Z	67	0.00

Source: Integr8 at EBI website (updated 7/09)



Average protein length : 468+/- 522 amino acid residues
Size range: 4 - 34350 amino acid residues

Relative frequency

protein length

Source: Integr8 at EBI website (updated 7/09)

Definition of a domain

According to InterPro at EBI (http://www.ebi.ac.uk/interpro/):

A domain is an independent structural unit, found alone or in conjunction with other domains or repeats. Domains are evolutionarily related.

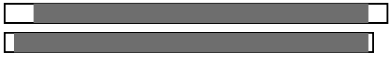
According to SMART (http://smart.embl-heidelberg.de):

A domain is a conserved structural entity with distinctive secondary structure content and a hydrophobic core. Homologous domains with common functions usually show sequence similarities.

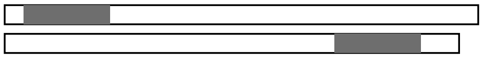
Page 390

Varieties of protein domains

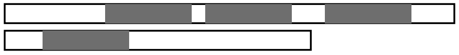
Extending along the length of a protein



Occupying a subset of a protein sequence

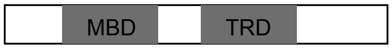


Occurring one or more times



Page 393

Example of a protein with domains: Methyl CpG binding protein 2 (MeCP2)

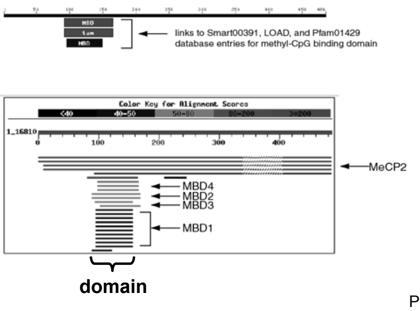


The protein includes a methylated DNA binding domain (MBD) and a transcriptional repression domain (TRD). MeCP2 is a transcriptional repressor.

Mutations in the gene encoding MeCP2 cause Rett Syndrome, a neurological disorder affecting girls primarily.

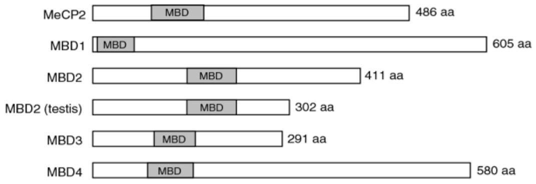
Page 393

Result of an MeCP2 blastp search: A methyl-binding domain shared by several proteins


MBD
MBD
links to Smart00391, LOAD, and Pfam01429 database entries for methyl-CpG binding domain
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_16810
0
100
200
300
400
MeCP2
MBD4
MBD2
MBD3
MBD1
domain

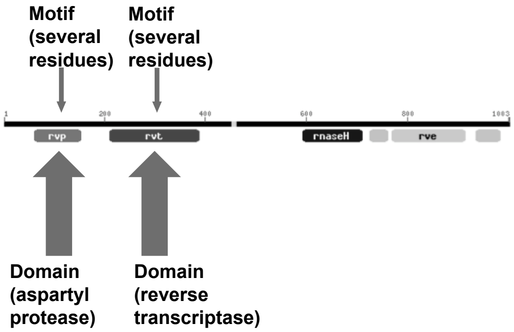
Page 393

Are proteins that share only a domain homologous?



Page 393

Proteins can have both domains and motifs (patterns)

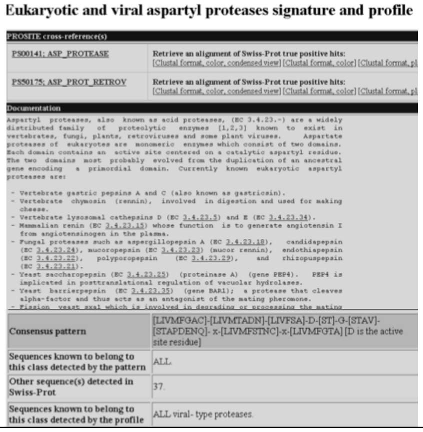


Motif (several residues)

Motif (several residues)

Domain (aspartyl protease)

Domain (reverse transcriptase)



Eukaryotic and viral aspartyl proteases signature and profile

PROSITE cross-reference(s)

PS00141; ASP_PROTEASE	**Retrieve an alignment of Swiss-Prot true positive hits:** [Clustal format, color, condensed view] [Clustal format, color] [Clustal format, pl
PS50175; ASP_PROT_RETROV	**Retrieve an alignment of Swiss-Prot true positive hits:** [Clustal format, color, condensed view] [Clustal format, color] [Clustal format, pl

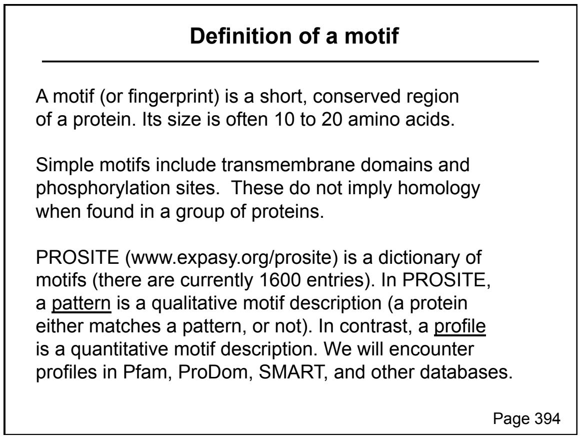
Documentation

```
Aspartyl  proteases, also  known as acid proteases, (EC 3.4.23.-) are a widely
distributed family   of   proteolytic   enzymes  [1,2,3]  known  to  exist  in
vertebrates, fungi, plants, retroviruses and some plant viruses.     Aspartate
proteases of  eukaryotes are  monomeric  enzymes which consist of two domains.
Each domain contains an  active site centered on a catalytic aspartyl residue.
The two  domains  most  probably  evolved from the duplication of an ancestral
gene encoding   a  primordial  domain.  Currently  known  eukaryotic  aspartyl
proteases are:

 - Vertebrate gastric pepsins A and C (also known as gastricsin).
 - Vertebrate  chymosin  (rennin),  involved  in digestion and used for making
   cheese.
 - Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34).
 - Mammalian renin (EC 3.4.23.15) whose function  is to generate angiotensin I
   from angiotensinogen in the plasma.
 - Fungal proteases such as aspergillopepsin A (EC 3.4.23.18),   candidapepsin
   (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin),  endothiapepsin
   (EC 3.4.23.22),   polyporopepsin    (EC 3.4.23.29),    and   rhizopuspepsin
   (EC 3.4.23.21).
 - Yeast saccharopepsin (EC 3.4.23.25)  (proteinase A)  (gene PEP4).   PEP4 is
   implicated in posttranslational regulation of vacuolar hydrolases.
 - Yeast  barrierpepsin  (EC 3.4.23.35)  (gene BAR1);  a protease that cleaves
   alpha-factor and thus acts as an antagonist of the mating pheromone.
 - Fission  yeast sxa1 which is involved in degrading or processing the mating
```

Consensus pattern	[LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]- x-[LIVMFSTNC]-x-[LIVMFGTA] [D is the active site residue]
Sequences known to belong to this class detected by the pattern	ALL.
Other sequence(s) detected in Swiss-Prot	37.
Sequences known to belong to this class detected by the profile	ALL viral- type proteases.

Page 396



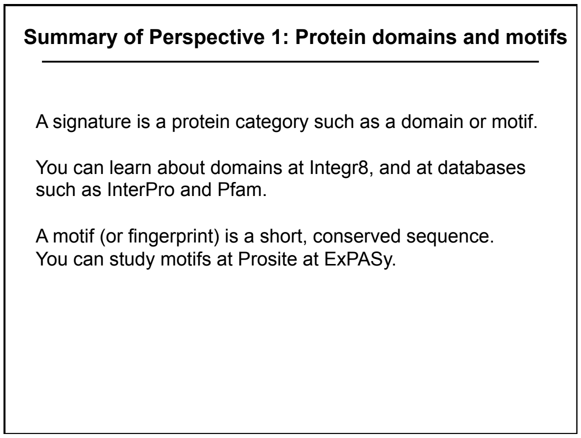
Definition of a motif

A motif (or fingerprint) is a short, conserved region of a protein. Its size is often 10 to 20 amino acids.

Simple motifs include transmembrane domains and phosphorylation sites. These do not imply homology when found in a group of proteins.

PROSITE (www.expasy.org/prosite) is a dictionary of motifs (there are currently 1600 entries). In PROSITE, a pattern is a qualitative motif description (a protein either matches a pattern, or not). In contrast, a profile is a quantitative motif description. We will encounter profiles in Pfam, ProDom, SMART, and other databases.

Page 394

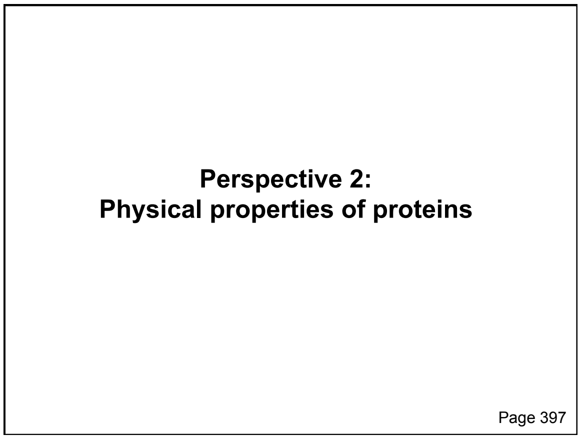


Summary of Perspective 1: Protein domains and motifs

A signature is a protein category such as a domain or motif.

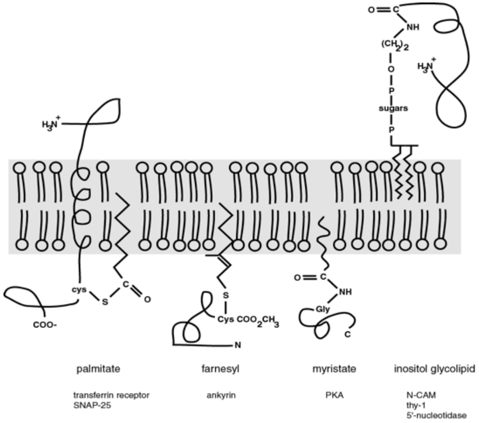
You can learn about domains at Integr8, and at databases such as InterPro and Pfam.

A motif (or fingerprint) is a short, conserved sequence. You can study motifs at Prosite at ExPASy.



Perspective 2: Physical properties of proteins

Page 397



Page 398

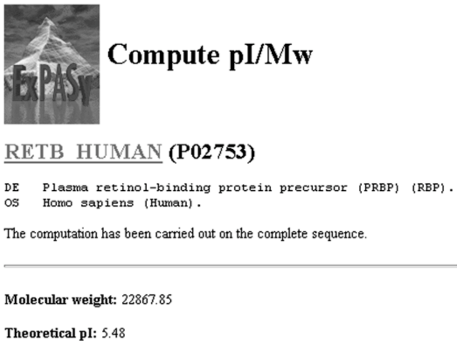
Physical properties of proteins

Many websites are available for the analysis of individual proteins. ExPASy and ISREC are two excellent resources.

The accuracy of these programs is variable. Predictions based on primary amino acid sequence (such as molecular weight prediction) are likely to be more trustworthy. For many other properties (such as posttranslational modification of proteins by specific sugars), experimental evidence may be required rather than prediction algorithms.

Page 399

Access a variety of protein analysis programs from the top right of the ExPASy home page



Compute pI/Mw

RETB_HUMAN (P02753)

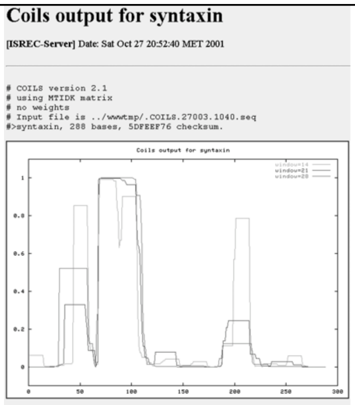
```
DE   Plasma retinol-binding protein precursor (PRBP) (RBP).
OS   Homo sapiens (Human).
```

The computation has been carried out on the complete sequence.

Molecular weight: 22867.85

Theoretical pI: 5.48

Page 399



Coils output for syntaxin

[ISREC-Server] Date: Sat Oct 27 20:52:40 MET 2001

```
# COILS version 2.1
# using MTIDK matrix
# no weights
# Input file is ../wwwtmp/.COILS.27003.1040.seq
#>syntaxin, 288 bases, 5DFEEF76 checksum.
```

Page 400

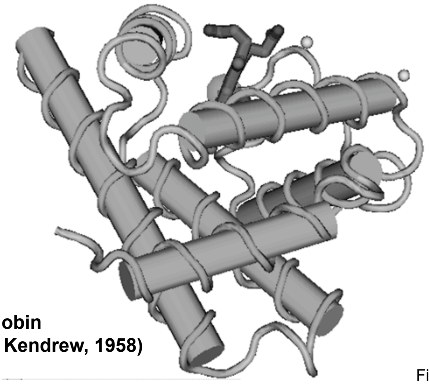
Protein secondary structure

Protein secondary structure is determined by the amino acid side chains.

Myoglobin is an example of a protein having many α-helices. These are formed by amino acid stretches 4-40 residues in length.

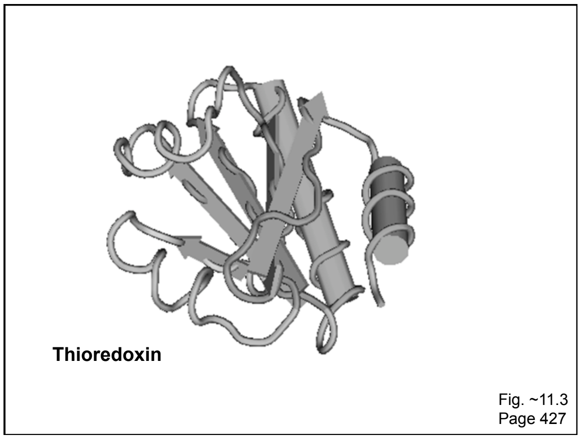
Thioredoxin from *E. coli* is an example of a protein with many β sheets, formed from β strands composed of 5-10 residues. They are arranged in parallel or antiparallel orientations.

Page 425



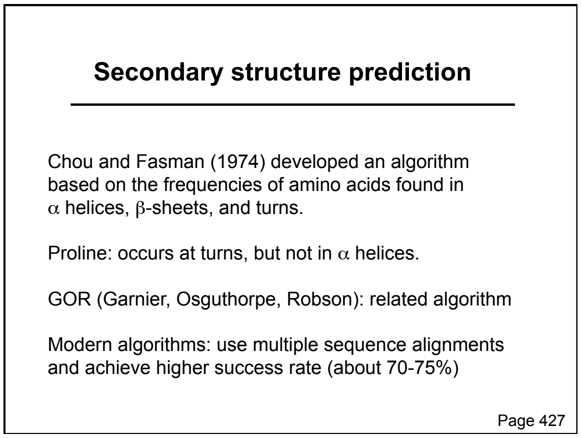
Myoglobin
(John Kendrew, 1958)

Fig. 11.3
Page 427



Thioredoxin

Fig. ~11.3
Page 427



Secondary structure prediction

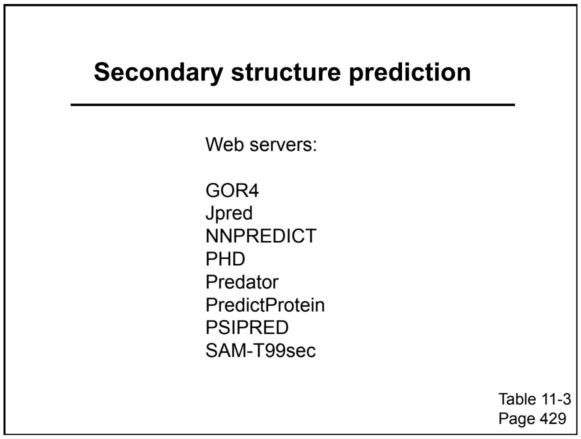
Chou and Fasman (1974) developed an algorithm based on the frequencies of amino acids found in α helices, β-sheets, and turns.

Proline: occurs at turns, but not in α helices.

GOR (Garnier, Osguthorpe, Robson): related algorithm

Modern algorithms: use multiple sequence alignments and achieve higher success rate (about 70-75%)

Page 427

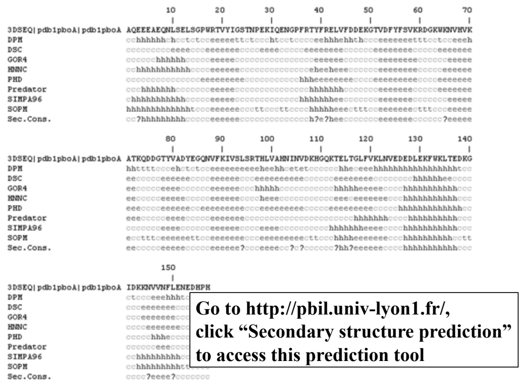


Secondary structure prediction

Web servers:

GOR4
Jpred
NNPREDICT
PHD
Predator
PredictProtein
PSIPRED
SAM-T99sec

Table 11-3
Page 429



```
                             10        20        30        40        50        60        70
                             |         |         |         |         |         |         |
3DSEQ|pdb1pboA|pdb1pboA  AQEEEAEQNLSELSGPWRTVYIGSTNPEKIQENGPFRTYFRELVFDDEKGTVDFYFSVKRDGKWKNVHVK
DPM                      cchhhhhhchcctctcceeeeettctccchtcctcccceehhhehhthccceeeeeeetttcctcceeeh
DSC                      cccccccccceeeecccceeeeeeecccceeeeccccceeeeeeeeccccccceeeeeeeecccccceeeeeeee
GOR4                     cccccchhhhhhccccceeeeeccccccccccccccceeeeeeccccccceeeeeeeeccccccceeeee
HNNC                     cchhhhhhhhhhhcccceeeeeecccccccccccccceheeheeeccccccceeeeeeeecccccceeeeee
PHD                      cccccccccccccccceeeeeeeeeccccccccccchhheeeeeeeeecccceeeeeeeeecccceeeeeeee
Predator                 ccchhhhhhhccccccceeeeeccccccccccccccchhhhhhhcccccccceeeeeeeccccceeeeeeee
SIMPA96                  chhhhhhhhhhhccccceeeeecccccccccccccchhhhhhheeccccccceeeeeeecccccccceeeee
SOPM                     hhhhhhhhhhhhhhtccceeeeeeccctttccctttcccchhhhheeectttcceeeeeeectttccceeeee
Sec.Cons.                cc?hhhhhhhhhccccceeeeeeccccccccccccccch?e?heeccccccceeeeeeecccccc?eeeee

                             80        90        100       110       120       130       140
                             |         |         |         |         |         |         |
3DSEQ|pdb1pboA|pdb1pboA  ATKQDDGTYVADYEGQNVFKIVSLSRTHLVAHNINVDKHGQKTELTGLFVKLNVEDEDLEKFWKLTEDKG
DPM                      hhttttccceehctctcceeeeeeeeeeeheehhcetetcccccchhhceeehhhhhhhhhhhhhhhhhhhtcc
DSC                      eecccccceeeeeeccccceeeeeeccceeeeeeeeeeccccccceeeeeeeeeccccccccchhhhhheecccc
GOR4                     eeccccceeeeccccccceeeeeecccchhhhhccccccccccchhhhheeeeeecccchhhhhhhhhhhccc
HNNC                     eeccccceeeeeccccceeeeeeecccheeeccccccccccccceeeceeeeeeeccchhhhhhhhhhhccc
PHD                      eeeccccceeeeeecceeeeeeeeeccceeeeeeeeeccccceeeeeeeeeeccccchhhhhhhhhhhhhcc
Predator                 eeeccccceeeecccccceeeeeeeccccccccccccccccccccccchhhhhhhccccchhhhhhhhhhccc
SIMPA96                  ccccccceeeecccccceeeeeeecccceeeecccccccccccchhhhhhheeeecccchhhhhhhhhhhccc
SOPM                     eccttttceeeeeettcceeeeeeeccceeeeeeeeeecttccchhhheeeeeeeccchhhhhhhhhhcctt
Sec.Cons.                eeccccceeeeeccccceeeeee?ccceeeeecc?c?cccccc?hh?eeeeeeccccchhhhhhhhhhhccc

                            150
                             |
3DSEQ|pdb1pboA|pdb1pboA  IDKKNVVNFLENEDHPH
DPM                      ctccceeehhhtc
DSC                      ccceeeeeeeccc
GOR4                     ccceeeeeeeccc
HNNC                     cccceeeeeeecc
PHD                      ccccchhhecccc
Predator                 cccccceeecccc
SIMPA96                  cchhhhhhhhhcc
SOPM                     cchhhhhhhhhtt
Sec.Cons.                cccc?eeee?ccccccc
```

Go to http://pbil.univ-lyon1.fr/, click "Secondary structure prediction" to access this prediction tool

Tertiary protein structure: protein folding

Main approaches:

[1] Experimental determination
(X-ray crystallography, NMR)

[2] Prediction

► Comparative modeling (based on homology)

► Threading

► *Ab initio* (de novo) prediction

Page 430

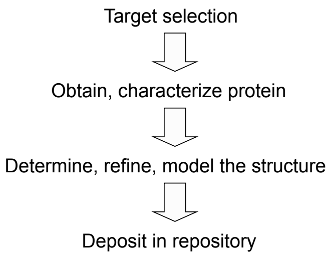
Experimental approaches to protein structure

[1] X-ray crystallography
-- Used to determine 80% of structures
-- Requires high protein concentration
-- Requires crystals
-- Able to trace amino acid side chains
-- Earliest structure solved was myoglobin

[2] NMR
-- Magnetic field applied to proteins in solution
-- Largest structures: 350 amino acids (40 kD)
-- Does not require crystallization

Page 430

Steps in obtaining a protein structure



Target selection

Obtain, characterize protein

Determine, refine, model the structure

Deposit in repository

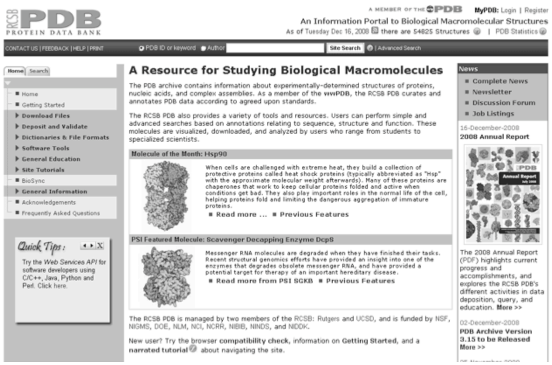
Fig 11.5
page 431

The Protein Data Bank (PDB)

- PDB is the principal repository for protein structures
- Established in 1971
- Accessed at http://www.rcsb.org/pdb or simply http://www.pdb.org
- Currently contains 64,000 structure entities

Updated 3/26/10

Page 434



An Information Portal to Biological Macromolecular Structures

As of Tuesday Dec 16, 2008 there are 54825 Structures | PDB Statistics

Try the *Web Services API* for software developers using C/C++, Java, Python and Perl. Click here.

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Molecule of the Month: Hsp90

When cells are challenged with extreme heat, they build a collection of protective proteins called heat shock proteins (typically abbreviated as "Hsp" with the approximate molecular weight afterwards). Many of these proteins are chaperones that work to keep cellular proteins folded and active when conditions get bad. They also play important roles in the normal life of the cell, helping proteins fold and limiting the dangerous aggregation of immature proteins.

■ Read more ... ■ Previous Features

PSI Featured Molecule: Scavenger Decapping Enzyme DcpS

Messenger RNA molecules are degraded when they have finished their tasks. Recent structural genomics efforts have provided an insight into one of the enzymes that degrades obsolete messenger RNA, and have provided a potential target for therapy of an important hereditary disease.

■ Read more from PSI SGKB ■ Previous Features

The RCSB PDB is managed by two members of the RCSB: Rutgers and UCSD, and is funded by NSF, NIGMS, DOE, NLM, NCI, NCRR, NIBIB, NINDS, and NIDDK.

New user? Try the browser **compatibility check**, information on **Getting Started**, and a **narrated tutorial** about navigating the site.

16-December-2008
2008 Annual Report

The 2008 Annual Report (PDF) highlights current progress and accomplishments, and explores the RCSB PDB's different activities in data deposition, query, and education. **More >>**

02-December-2008
PDB Archive Version 3.15 to be Released
More >>

Fig. 11.7
Page 435

PDB content growth (www.pdb.org)

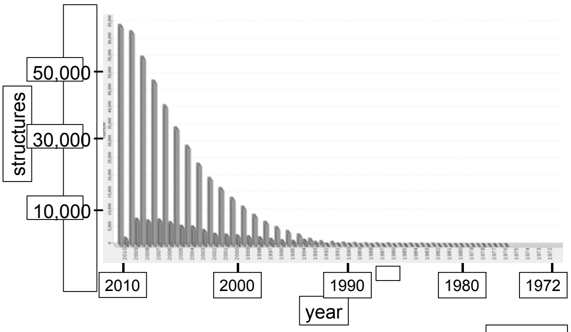


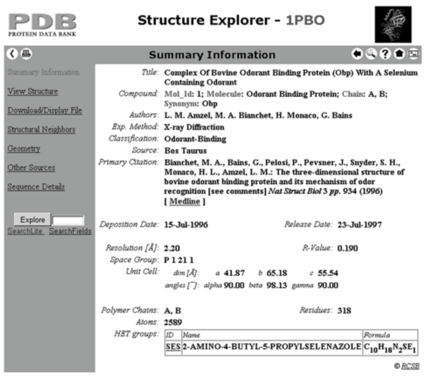
Fig. 9.6
Page 281

Updated 3/26/10

PDB holdings (12/08)

50,621	proteins, peptides
2,225	protein/nucl. complexes
1,946	nucleic acids
33	other; carbohydrates
54,825	total

Table 11-4
Page 435



PDB PROTEIN DATA BANK

Structure Explorer - 1PBO

Summary Information

Title: Complex Of Bovine Odorant Binding Protein (Obp) With A Selenium Containing Odorant

Compound: Mol_Id: 1; Molecule: Odorant Binding Protein; Chain: A, B; Synonym: Obp

Authors: L. M. Amzel, M. A. Bianchet, H. Monaco, G. Bains

Exp. Method: X-ray Diffraction

Classification: Odorant-Binding

Source: Bos Taurus

Primary Citation: Bianchet, M. A., Bains, G., Pelosi, P., Pevsner, J., Snyder, S. H., Monaco, H. L., Amzel, L. M.: The three-dimensional structure of bovine odorant binding protein and its mechanism of odor recognition [see comments] *Nat Struct Biol* 3 *pp.* 934 (1996) [Medline]

Deposition Date: 15-Jul-1996 *Release Date:* 23-Jul-1997

Resolution [Å]: 2.20 *R-Value:* 0.190

Space Group: P 1 21 1

Unit Cell: *dim [Å]:* *a* 41.87 *b* 65.18 *c* 55.54
angles [°]: *alpha* 90.00 *beta* 98.13 *gamma* 90.00

Polymer Chains: A, B *Residues:* 318

Atoms: 2589

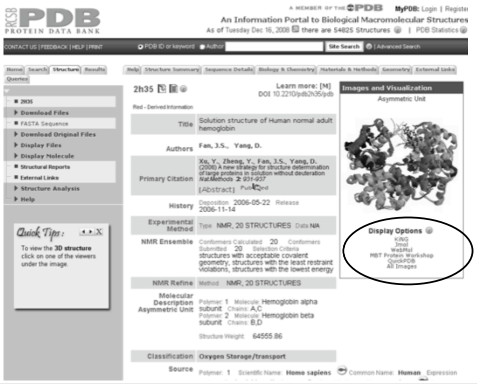
HET groups:

ID	*Name*	*Formula*
SES	2-AMINO-4-BUTYL-5-PROPYLSELENAZOLE	$C_{10}H_{18}N_2SE_1$

© RCSB

Fig. ~11.10
Page 436

Viewing hemoglobin (accession 2H35) at PDB



An Information Portal to Biological Macromolecular Structures

As of Tuesday Dec 16, 2008 there are 54825 Structures | PDB Statistics

Quick Tips: To view the **3D structure** click on one of the viewers under the image.

2h35

Learn more: [M]
DOI 10.2210/pdb2h35/pdb

Red - Derived Information

Title	Solution structure of Human normal adult hemoglobin
Authors	Fan, J.S., Yang, D.
Primary Citation	Xu, Y., Zheng, Y., Fan, J.S., Yang, D. (2006) A new strategy for structure determination of large proteins in solution without deuteration Nat.Methods 3: 931-937 [Abstract] PubMed
History	Deposition 2006-05-22 Release 2006-11-14
Experimental Method	Type NMR, 20 STRUCTURES Data N/A
NMR Ensemble	Conformers Calculated 20 Conformers Submitted 20 Selection Criteria structures with acceptable covalent geometry, structures with the least restraint violations, structures with the lowest energy
NMR Refine	Method NMR, 20 STRUCTURES
Molecular Description Asymmetric Unit	Polymer: 1 Molecule: Hemoglobin alpha subunit Chains: A,C Polymer: 2 Molecule: Hemoglobin beta subunit Chains: B,D Structure Weight: 64555.86
Classification	Oxygen Storage/transport
Source	Polymer: 1 Scientific Name: Homo sapiens Common Name: Human Expression

Viewing structures at PDB: WebMol

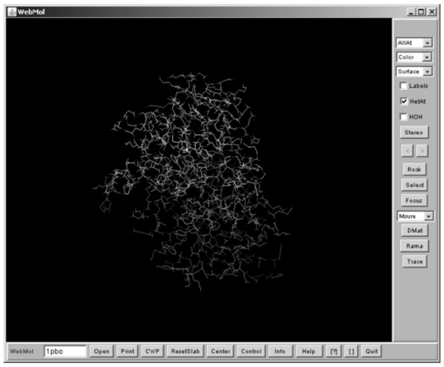
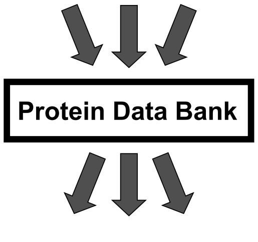


Fig. 11.11
Page 437



gateways to access PDB files

Swiss-Prot, NCBI, EMBL

Protein Data Bank

CATH, Dali, SCOP, FSSP

databases that interpret PDB files

The CATH Hierarchy

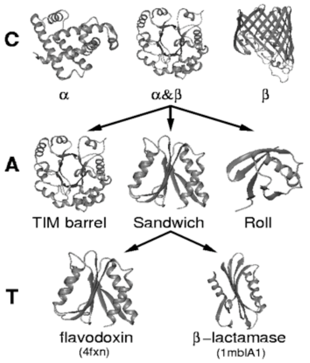


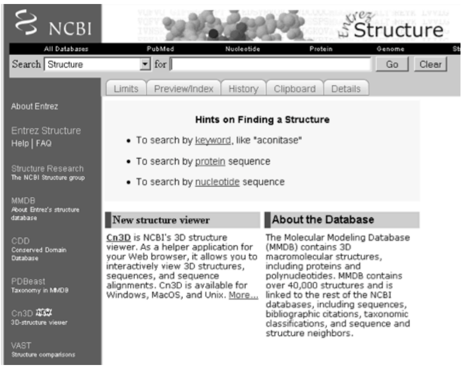
Fig. 11.18
Page 444

Access to PDB through NCBI

You can access PDB data at the NCBI several ways.

- Go to the Structure site, from the NCBI homepage
- Use Entrez
- Perform a BLAST search, restricting the output to the PDB database

Page 437



Hints on Finding a Structure

- To search by keyword, like "aconitase"
- To search by protein sequence
- To search by nucleotide sequence

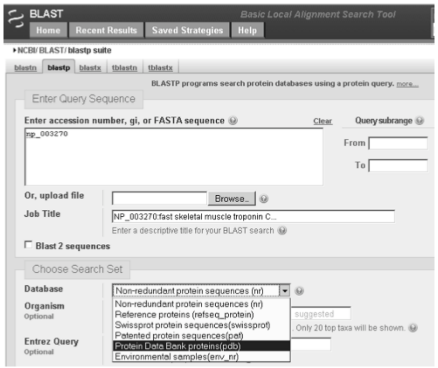
New structure viewer

Cn3D is NCBI's 3D structure viewer. As a helper application for your Web browser, it allows you to interactively view 3D structures, sequences, and sequence alignments. Cn3D is available for Windows, MacOS, and Unix. More...

About the Database

The Molecular Modeling Database (MMDB) contains 3D macromolecular structures, including proteins and polynucleotides. MMDB contains over 40,000 structures and is linked to the rest of the NCBI databases, including sequences, bibliographic citations, taxonomic classifications, and sequence and structure neighbors.

Fig. ~11.12
Page 438



BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

▸NCBI/ BLAST/ blastp suite

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence: np_003270 — Clear

Query subrange: From [ ] To [ ]

Or, upload file: [ ] Browse..

Job Title: NP_003270:fast skeletal muscle troponin C...

Enter a descriptive title for your BLAST search

□ Blast 2 sequences

Choose Search Set

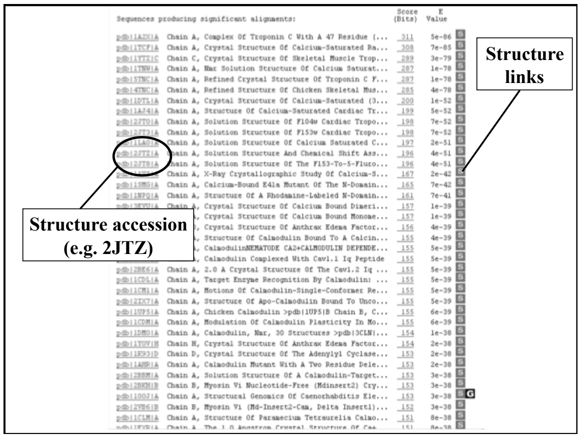
Database: Non-redundant protein sequences (nr)

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Swissprot protein sequences(swissprot)
- Patented protein sequences(pat)
- Protein Data Bank proteins(pdb)
- Environmental samples(env_nr)

Organism (Optional): suggested — Only 20 top taxa will be shown.

Entrez Query (Optional)

Do a blastp search;
set the database to pdb (Protein Data Bank)

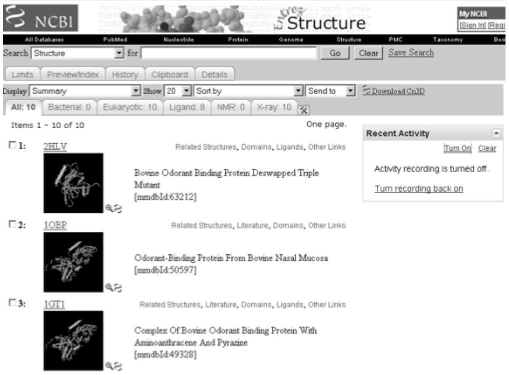


```
                                                                   Score     E
Sequences producing significant alignments:                        (Bits)  Value

pdb|1A2X|A  Chain A, Complex Of Troponin C With A 47 Residue (...   311    5e-86  S
pdb|1TCF|A  Chain A, Crystal Structure Of Calcium-Saturated Ra...   308    7e-85  S
pdb|1YTZ|C  Chain C, Crystal Structure Of Skeletal Muscle Trop...   289    3e-79  S
pdb|1TNW|A  Chain A, Nmr Solution Structure Of Calcium Saturat...   287    1e-78  S
pdb|5TNC|A  Chain A, Refined Crystal Structure Of Troponin C F...   287    1e-78  S
pdb|4TNC|A  Chain A, Refined Structure Of Chicken Skeletal Mus...   285    4e-78  S
pdb|1DTL|A  Chain A, Crystal Structure Of Calcium-Saturated (3...   200    1e-52  S
pdb|1AJ4|A  Chain A, Structure Of Calcium-Saturated Cardiac Tr...   199    5e-52  S
pdb|2JT0|A  Chain A, Solution Structure Of F104w Cardiac Tropo...   198    7e-52  S
pdb|2JT3|A  Chain A, Solution Structure Of F153w Cardiac Tropo...   198    7e-52  S
pdb|1LA0|A  Chain A, Solution Structure Of Calcium Saturated C...   197    2e-51  S
pdb|2JTZ|A  Chain A, Solution Structure And Chemical Shift Ass...   196    4e-51  S
pdb|2JT8|A  Chain A, Solution Structure Of The F153-To-5-Fluro...   196    4e-51  S
[illegible] Chain A, X-Ray Crystallographic Study Of Calcium-S...   167    2e-42  S
pdb|1SMG|A  Chain A, Calcium-Bound E41a Mutant Of The N-Domain...   165    7e-42  S
pdb|1NPQ|A  Chain A, Structure Of A Rhodamine-Labeled N-Domain...   161    7e-41  S
pdb|3EVU|A  Chain A, Crystal Structure Of Calcium Bound Dimeri...   157    1e-39  S
            A, Crystal Structure Of Calcium Bound Monome...         157    1e-39  S
            , Crystal Structure Of Anthrax Edema Factor...          156    4e-39  S
            A, Structure Of Calmodulin Bound To A Calcin...         155    4e-39  S
            A, CalmodulinNEMATODE CA2+CALMODULIN DEPENDE...         155    5e-39  S
            A, Calmodulin Complexed With Cav1.1 Iq Peptide          155    5e-39  S
pdb|2BE6|A  Chain A, 2.0 A Crystal Structure Of The Cav1.2 Iq ...   155    5e-39  S
pdb|1CDL|A  Chain A, Target Enzyme Recognition By Calmodulin: ...   155    5e-39  S
pdb|1CM1|A  Chain A, Motions Of Calmodulin-Single-Conformer Re...   155    5e-39  S
pdb|2IX7|A  Chain A, Structure Of Apo-Calmodulin Bound To Unco...   155    5e-39  S
pdb|1UP5|A  Chain A, Chicken Calmodulin >pdb|1UP5|B Chain B, C...   155    6e-39  S
pdb|1CDM|A  Chain A, Modulation Of Calmodulin Plasticity In Mo...   155    6e-39  S
pdb|1DM0|A  Chain A, Calmodulin, Nmr, 30 Structures >pdb|3CLN|...   154    1e-38  S
pdb|1Y0V|H  Chain H, Crystal Structure Of Anthrax Edema Factor...   154    2e-38  S
pdb|1K93|D  Chain D, Crystal Structure Of The Adenylyl Cyclase...   153    2e-38  S
pdb|1AHR|A  Chain A, Calmodulin Mutant With A Two Residue Dele...   153    2e-38  S
pdb|2BBM|A  Chain A, Solution Structure Of A Calmodulin-Target...   153    3e-38  S
pdb|2BKH|B  Chain B, Myosin Vi Nucleotide-Free (Mdinsert2) Cry...   153    3e-38  S
pdb|1OOJ|A  Chain A, Structural Genomics Of Caenorhabditis Ele...   153    3e-38  S G
pdb|2VB6|B  Chain B, Myosin Vi (Md-Insert2-Cam, Delta Insert1)...   152    3e-38  S
pdb|1CLM|A  Chain A, Structure Of Paramecium Tetraurelia Calmo...   151    8e-38  S
pdb|1EXR|A  Chain A, The 1.0 Angstrom Crystal Structure Of Ca+...   151    8e-38  S
```

Structure links

Structure accession (e.g. 2JTZ)



NCBI Entrez Structure — My NCBI [Sign In] [Regi

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Boo

Search: Structure for [ ] Go | Clear | Save Search

Limits | Preview/Index | History | Clipboard | Details

Display: Summary | Show: 20 | Sort by | Send to | Download Cn3D

All: 10 | Bacterial: 0 | Eukaryotic: 10 | Ligand: 8 | NMR: 0 | X-ray: 10

Items 1 - 10 of 10 — One page.

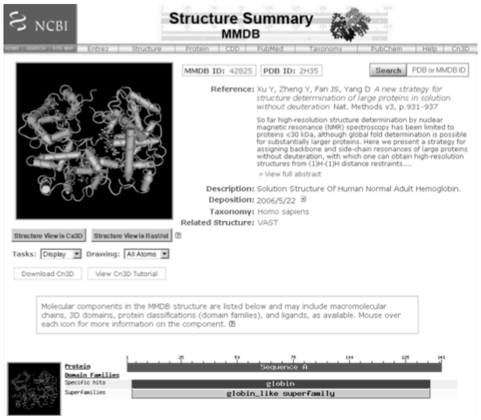
□ 1: 2HLV — Related Structures, Domains, Ligands, Other Links
Bovine Odorant Binding Protein Deswapped Triple Mutant
[mmdbId:63212]

□ 2: 1OBP — Related Structures, Literature, Domains, Other Links
Odorant-Binding Protein From Bovine Nasal Mucosa
[mmdbId:50597]

□ 3: 1GT1 — Related Structures, Literature, Domains, Ligands, Other Links
Complex Of Bovine Odorant Binding Protein With Aminoanthracene And Pyrazine
[mmdbId:49328]

Recent Activity — Turn On | Clear
Activity recording is turned off.
Turn recording back on

Fig. 9.14
Page 289



Structure Summary
MMDB

MMDB ID: 42825 **PDB ID:** 2H35 Search PDB or MMDB ID

Reference: Xu Y, Zheng Y, Fan JS, Yang D *A new strategy for structure determination of large proteins in solution without deuteration* Nat. Methods v3, p.931-937

So far high-resolution structure determination by nuclear magnetic resonance (NMR) spectroscopy has been limited to proteins <30 kDa, although global fold determination is possible for substantially larger proteins. Here we present a strategy for assigning backbone and side-chain resonances of large proteins without deuteration, with which one can obtain high-resolution structures from (1)H-(1)H distance restraints....

» View full abstract

Description: Solution Structure Of Human Normal Adult Hemoglobin.
Deposition: 2006/5/22
Taxonomy: Homo sapiens
Related Structure: VAST

Structure View in Cn3D | Structure View in RasMol

Tasks: Display Drawing: All Atoms

Download Cn3D | View Cn3D Tutorial

Molecular components in the MMDB structure are listed below and may include macromolecular chains, 3D domains, protein classifications (domain families), and ligands, as available. Mouse over each icon for more information on the component.

Access to PDB structures through NCBI

Molecular Modeling DataBase (MMDB)

Cn3D ("see in 3D" or three dimensions):
structure visualization software

Vector Alignment Search Tool (VAST):
view multiple structures

Page 291

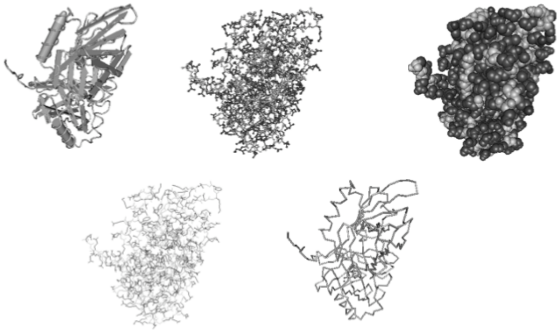


Fig. 9.16
Page 291

Introduction to Perspectives 3 and 4: Gene Ontology (GO) Consortium

Page 237

The Gene Ontology Consortium

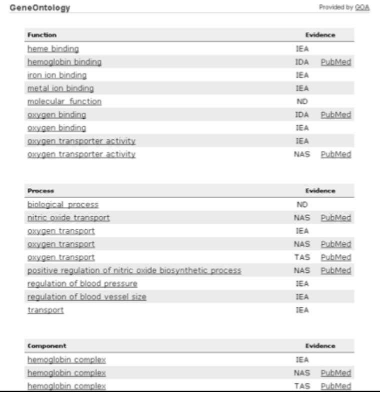
An ontology is a description of concepts. The GO Consortium compiles a dynamic, controlled vocabulary of terms related to gene products.

There are three organizing principles:
Molecular function
Biological process
Cellular compartment

You can visit GO at http://www.geneontology.org. There is no centralized GO database. Instead, curators of organism-specific databases assign GO terms to gene products for each organism.

Page 237

GO terms are assigned to Entrez Gene entries



GeneOntology

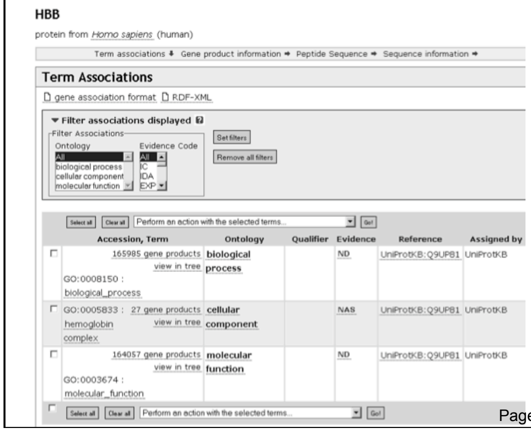
Provided by GOA

Function	Evidence	
heme binding	IEA	
hemoglobin binding	IDA	PubMed
iron ion binding	IEA	
metal ion binding	IEA	
molecular_function	ND	
oxygen binding	IDA	PubMed
oxygen binding	IEA	
oxygen transporter activity	IEA	
oxygen transporter activity	NAS	PubMed

Process	Evidence	
biological_process	ND	
nitric oxide transport	NAS	PubMed
oxygen transport	IEA	
oxygen transport	NAS	PubMed
oxygen transport	TAS	PubMed
positive regulation of nitric oxide biosynthetic process	NAS	PubMed
regulation of blood pressure	IEA	
regulation of blood vessel size	IEA	
transport	IEA	

Component	Evidence	
hemoglobin complex	IEA	
hemoglobin complex	NAS	PubMed
hemoglobin complex	TAS	PubMed

Page 241



HBB

protein from *Homo sapiens* (human)

Term associations ⬇ Gene product information ➡ Peptide Sequence ➡ Sequence information ➡

Term Associations

	Accession, Term	Ontology	Qualifier	Evidence	Reference	Assigned by
□	165985 gene products view in tree GO:0008150 : biological_process	biological process		ND	UniProtKB:Q9UP81	UniProtKB
□	GO:0005833 : 27 gene products hemoglobin view in tree complex	cellular component		NAS	UniProtKB:Q9UP81	UniProtKB
□	164057 gene products view in tree GO:0003674 : molecular_function	molecular function		ND	UniProtKB:Q9UP81	UniProtKB

Page 241

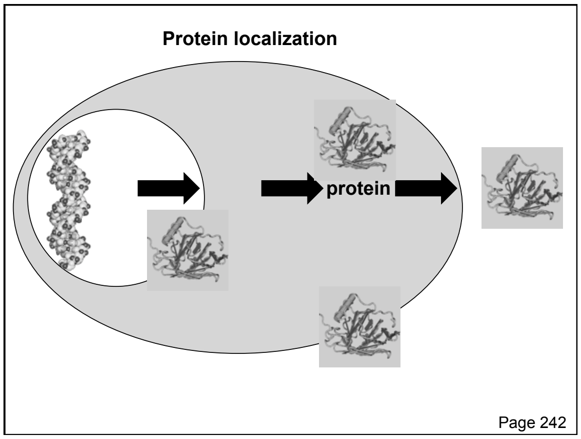
The Gene Ontology Consortium: Evidence Codes

IC	Inferred by curator
IDA	Inferred from direct assay
IEA	Inferred from electronic annotation
IEP	Inferred from expression pattern
IGI	Inferred from genetic interaction
IMP	Inferred from mutant phenotype
IPI	Inferred from physical interaction
ISS	Inferred from sequence or structural similarity
NAS	Non-traceable author statement
ND	No biological data
TAS	Traceable author statement

Page 240

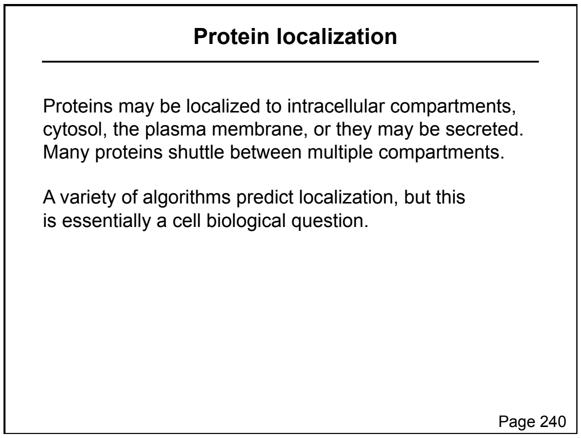
Perspective 3:
Protein localization

Page 242



Protein localization

Page 242

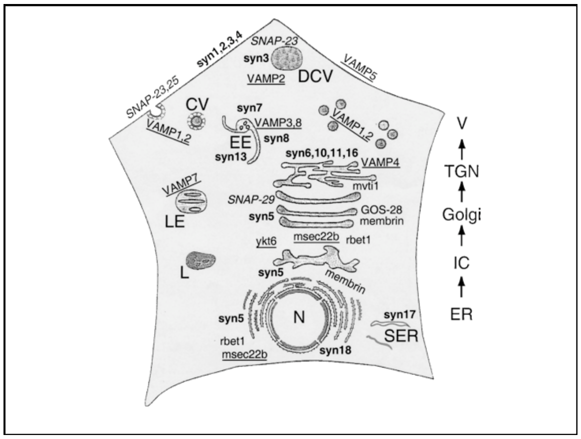


Protein localization

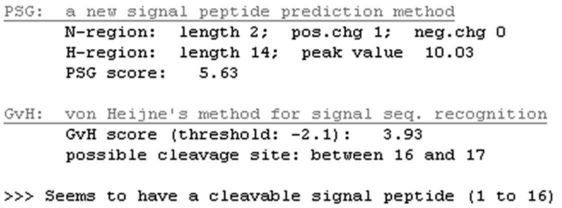
Proteins may be localized to intracellular compartments, cytosol, the plasma membrane, or they may be secreted. Many proteins shuttle between multiple compartments.

A variety of algorithms predict localization, but this is essentially a cell biological question.

Page 240



Results of Subprograms

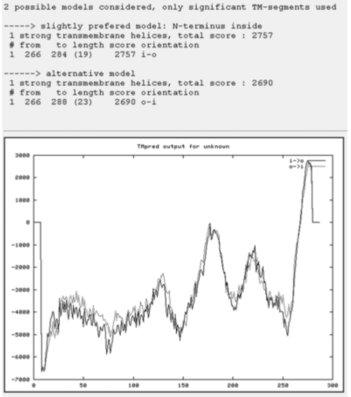


```
PSG:  a new signal peptide prediction method
      N-region:  length 2;  pos.chg 1;  neg.chg 0
      H-region:  length 14;  peak value  10.03
      PSG score:   5.63

GvH:  von Heijne's method for signal seq. recognition
      GvH score (threshold: -2.1):   3.93
      possible cleavage site: between 16 and 17

>>> Seems to have a cleavable signal peptide (1 to 16)
```

Page 242



```
2 possible models considered, only significant TM-segments used

-----> slightly prefered model: N-terminus inside
 1 strong transmembrane helices, total score : 2757
 # from   to length score orientation
 1  266  284 (19)    2757 i-o

------> alternative model
 1 strong transmembrane helices, total score : 2690
 # from   to length score orientation
 1  266  288 (23)    2690 o-i
```

Page 244