Protein Bioinformatics
Part I: Access to information

260.655
April 6, 2006
Jonathan Pevsner, Ph.D.
pevsner@kennedykrieger.org

Outline

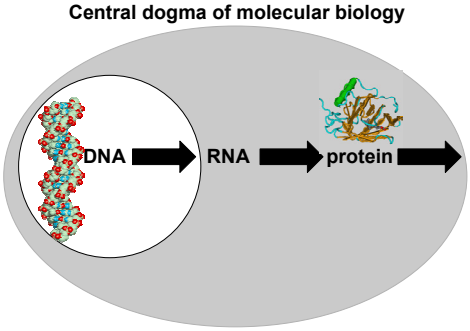
[1] Proteins at NCBI
- RefSeq accession numbers
- Cn3D to visualize structures

[2] The Protein Data Bank (PDB)

[3] UniProt

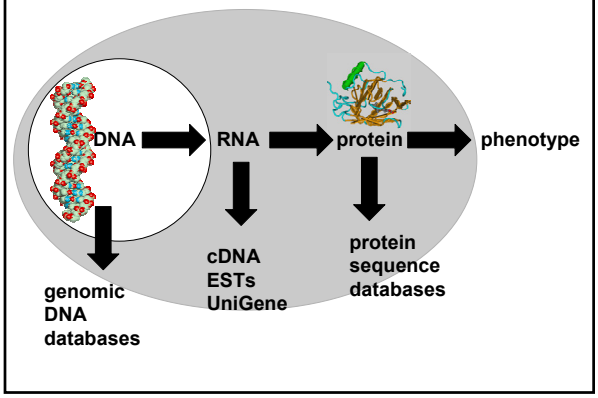
[4] ExPASy (Expert Protein Analysis System)
- DeepView, the Swiss-Pdb Viewer.

Central dogma of molecular biology

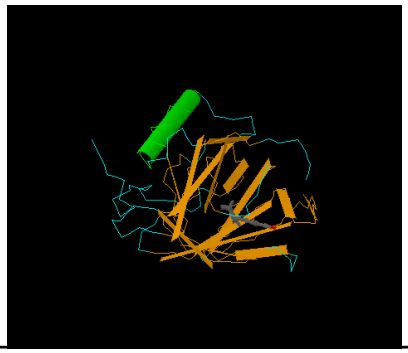


genome → transcriptome → proteome

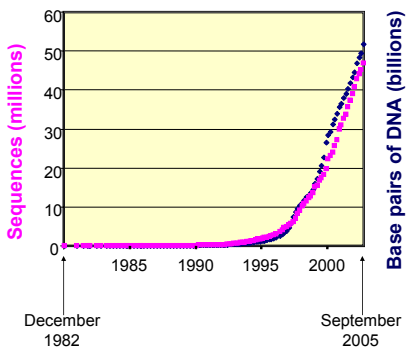
Central dogma of bioinformatics and genomics



[1] NCBI



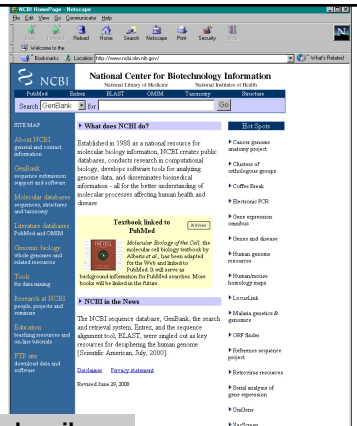
Growth of GenBank



There are three major public DNA databases

EMBL ⟷ **GenBank** ⟷ **DDBJ**

Housed at EBI	Housed at NCBI	Housed in Japan
European Bioinformatics Institute	National Center for Biotechnology Information	



www.ncbi.nlm.nih.gov

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.

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 NT_030059 Rs7079946	GenBank genomic DNA sequence Genomic contig dbSNP (single nucleotide polymorphism)	**DNA**
N91759.1 NM_006744	An expressed sequence tag (1 of 328) RefSeq DNA sequence (from a transcript)	**RNA**
NP_007635 AAC02945 Q28369 1KT7	RefSeq protein GenBank protein SwissProt protein Protein Data Bank structure record	**protein**

Accessing protein sequences via Entrez

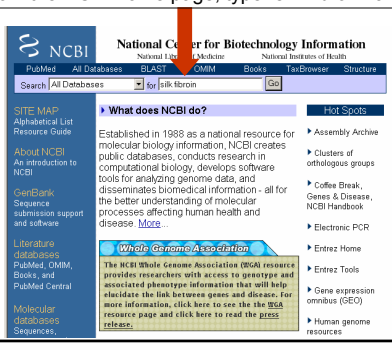
Entrez Gene with RefSeq

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_006744) or protein (NP_007635)

Example #1. Sean mentioned silk fibroin. How do you find its sequence?

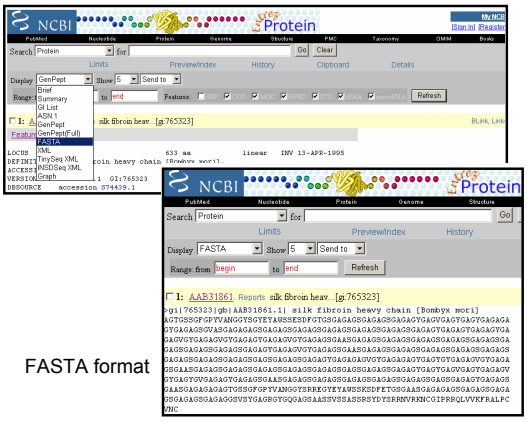
From the NCBI home page, type “silk fibroin” and hit “Go”




NCBI
Entrez, The Life Sciences Search Engine
PubMed
All Databases
Human Genome
GenBank
Map Viewer
Search across databases
silk fibroin
Go
Clear
Help
PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Site Search: NCBI web and FTP sites
Books: online books
OMIM: online Mendelian Inheritance in Man
OMIA: Online Mendelian Inheritance in Animals
Nucleotide: sequence database (GenBank)
Protein: sequence database
Genome: whole genome sequences
Structure: three-dimensional macromolecular structures
Taxonomy: organisms in GenBank
SNP: single nucleotide polymorphism
Gene: gene-centered information
HomoloGene: eukaryotic homology groups
PubChem Compound: unique small molecule chemical structures
PubChem Substance: deposited chemical substance records
Genome Project: genome project information
UniGene: gene-oriented clusters of transcript sequences
CDD: conserved protein domain database
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
PopSet: population study data sets
GEO Profiles: expression and molecular abundance profiles
GEO DataSets: experimental sets of GEO data
Cancer Chromosomes: cytogenetic databases
PubChem BioAssay: bioactivity screens of chemical substances
GENSAT: gene expression atlas of mouse central nervous system
Probe: sequence-specific reagents


NCBI
Protein
My NCBI
[Sign In] [Register]
Search Protein for silk fibroin
Go
Clear
Save Search
Limits
Preview/Index
History
Clipboard
Details
Display Summary
Show 20
Sort by Relevance
Send to
All: 110
Bacteria: 0
RefSeq: 3
Related Structures: 67
Items 1 - 20 of 110
Page 1 of 6 Next
1: AAL83649 Reports BLink, Links
silk fibroin [Bombyx mori]
gi|19221230|gb|AAL83649.1|[19221230]
2: AAA27839 Reports Links
silk fibroin
gi|155999|gb|AAA27839.1|[155999]
3: AAC32606 Reports BLink, Links
fibroin [Antheraea pernyi]
gi|3450883|gb|AAC32606.1|[3450883]
4: P46804 Reports BLink, Links
Spidroin 2 (Dragline silk fibroin 2)
gi|1174415|pir||P46804|SFD2_NEPCL[1174415]
5: P19837 Reports BLink, Links
Spidroin 1 (Dragline silk fibroin 1)
gi|1174414|pir||P19837|SFD1_NEPCL[1174414]
6: P46802 Reports BLink, Links
Spidroin 1 (Dragline silk fibroin 1)
gi|1174413|pir||P46802|SFD1_ARABI[1174413]
7: AAB31861 Reports BLink, Links
silk fibroin heavy chain [Bombyx mori]
gi|765323|gb|AAB31861.1|[bbm:354280|bbs|157676[765323]

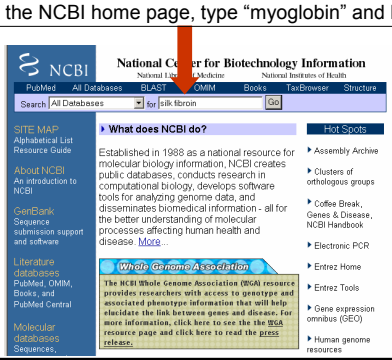

NCBI
Protein
Search Protein for
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Display GenPept Show 5 Send to
Range from begin to end Features Refresh
1: AAB31861 Reports silk fibroin heav. [gi:765323] BLink, Link
Features Sequence
LOCUS AAB31861 633 aa linear INV 13-APR-1995
DEFINITION silk fibroin heavy chain [Bombyx mori].
ACCESSION AAB31861
VERSION AAB31861.1 GI:765323
RESOURCE accession S74439.1
KEYWORDS .
SOURCE Bombyx mori (domestic silkworm)
ORGANISM Bombyx mori
Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Bombycoidea; Bombycidae; Bombyx.
REFERENCE 1 (residues 1 to 633)
AUTHORS Hita,K., Ichimura,S. and James,T.C.
TITLE Highly repetitive structure and its organization of the silk fibroin gene
JOURNAL J. Mol. Evol. 38 (6), 583-592 (1994)
PUBMED 7916056
REMARK GenBank staff at the National Library of Medicine created this entry [NCBI gibbsq 157676] from the original journal article.
COMMENT Method: conceptual translation supplied by author.
FEATURES Location/Qualifiers
source 1..633
/organism="Bombyx mori"
/db_xref="taxon:7091"
Protein 1..633
/product="silk fibroin heavy chain"
CDS 1..633
/gene="silk fibroin heavy chain"
/coded_by="S74439.1:2..1903"
ORIGIN



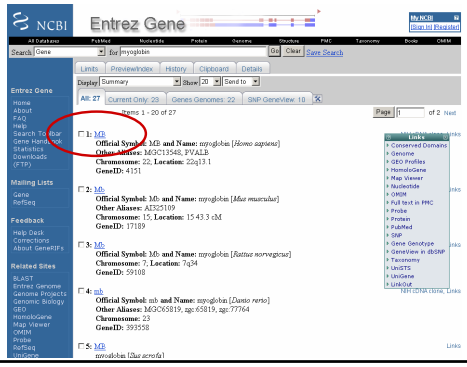
FASTA format

Example #2. Find the sequence of myoglobin

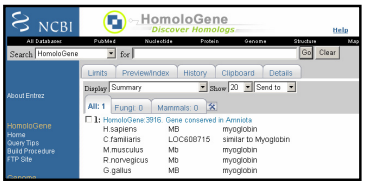
From the NCBI home page, type “myoglobin” and hit “Go”

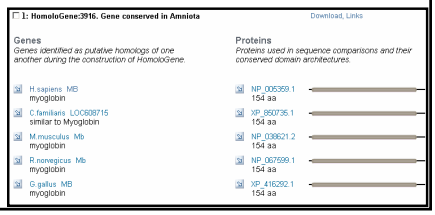


1. Entrez Gene entries offer a wealth of information, including links to RefSeq entries and to the Human Protein Reference Database (HPRD; Akhilesh)



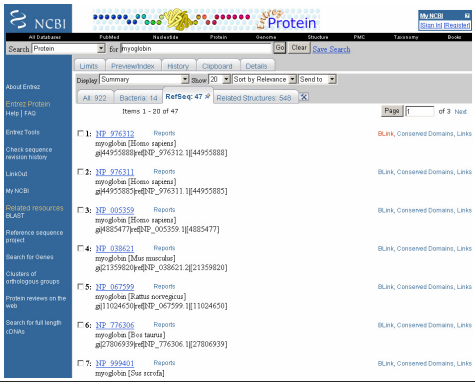
2. HomoloGene provides access to RefSeq identifiers of a protein family, and offers links (domains, pairwise alignments)



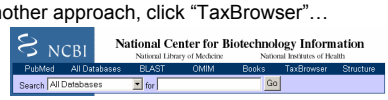


3. Entrez Protein shows 922 myoglobins (too many), including 47 RefSeq (still a lot).

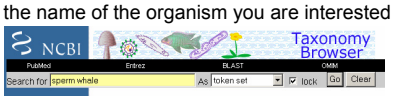
You can try scrolling through the RefSeq list, or apply “Limits”



As another approach, click “TaxBrowser”…



Enter the name of the organism you are interested in…



Follow a link of interest…

sperm whale

- Kogia simus [genbank common name: dwarf sperm whale]
- Physeter catodon [genbank common name: sperm whale]
- Kogia breviceps [genbank common name: pygmy sperm whale]
- Dwarf sperm whale gammaherpesvirus

Now click protein…

Physeter catodon

Taxonomy ID: 9755
Genbank common name: **sperm whale**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
synonym: **Physeter macrocephalus**

Lineage (full):
cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Laurasiatheria; Cetartiodactyla; Cetacea; Odontoceti; Physeteridae; Physeter

Entrez records	
Database name	Direct links
Nucleotide	130
Protein	234
Structure	160
Genome Sequences	1
Popset	11
3D Domains	152
UniSTS	10
PubMed Central	20
Gene	12
Taxonomy	1

You now can view all sperm whale proteins…

…and restrict the output to sperm whale myoglobin.

4. Structure provides links to myoglobin structures

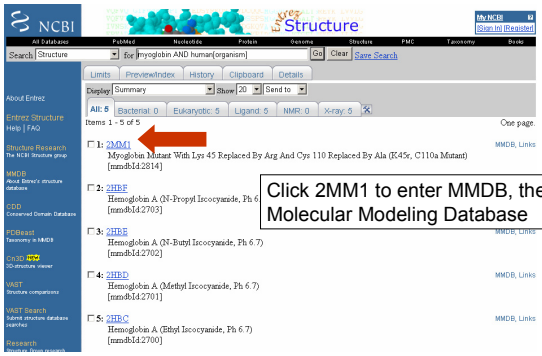
Access to PDB through NCBI

Molecular Modeling DataBase (MMDB)

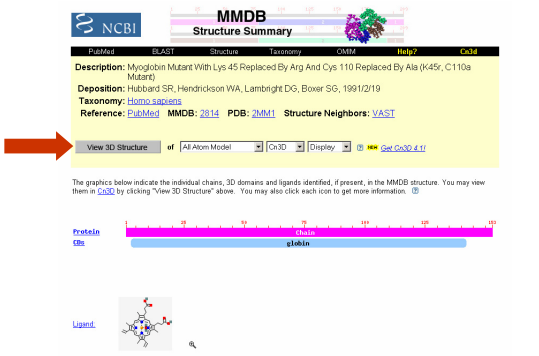
Cn3D (“see in 3D” or three dimensions): structure visualization software

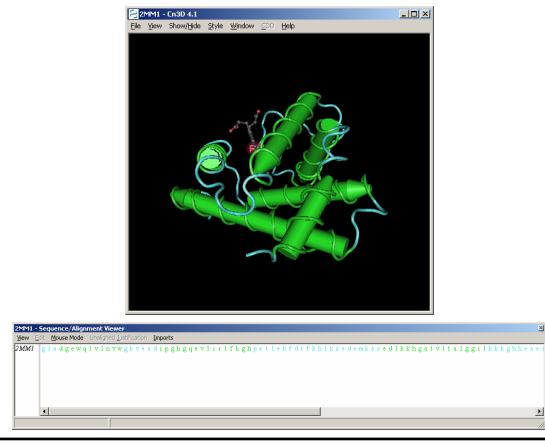
Vector Alignment Search Tool (VAST): view multiple structures

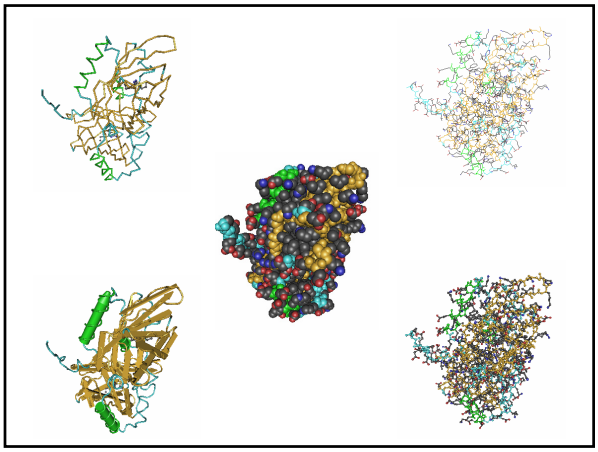
You can limit the output to particular species, e.g. with the command human[organism]



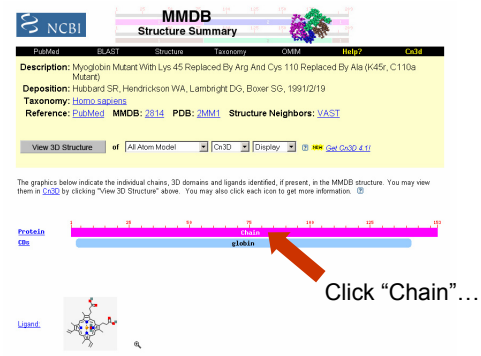
Click 2MM1 to enter MMDB, the Molecular Modeling Database

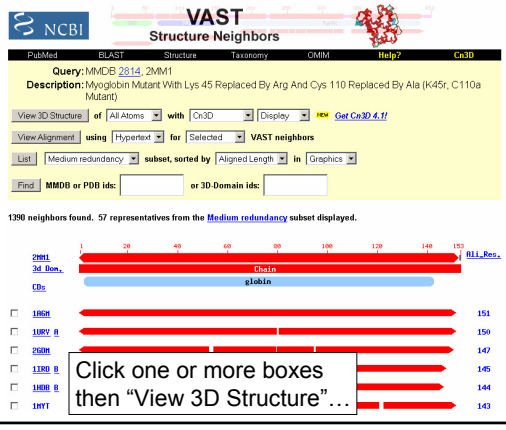




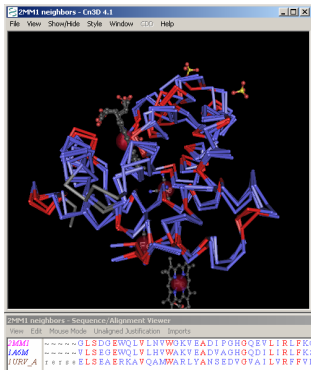


Overlay two or more structures with VAST at NCBI

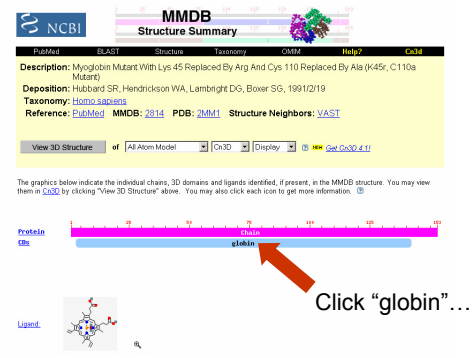




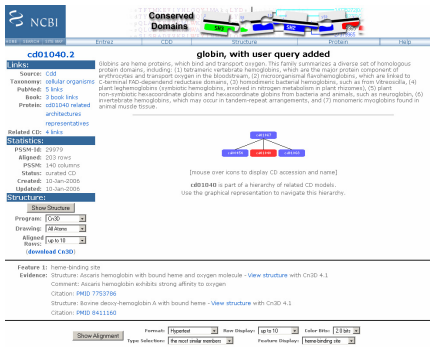
Overlay two or more structures with VAST at NCBI



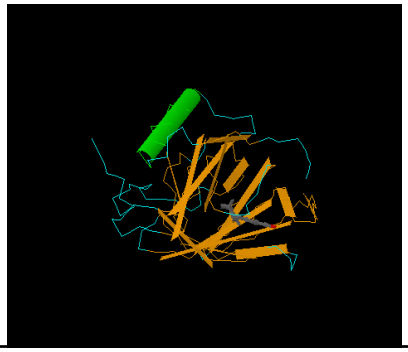
Overlay two or more structures with VAST at NCBI



Access the Conserved Domain database at NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)



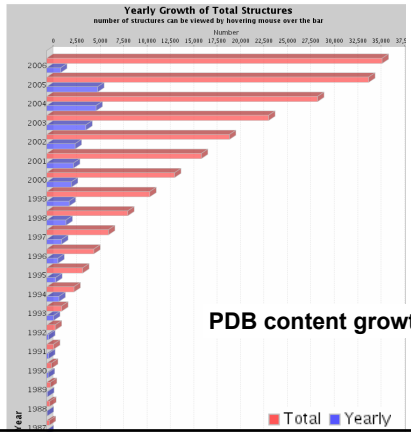
[2] PDB



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 over 35,000 structure entities

Updated 3/06



PDB content growth

PDB holdings (September, 2005)

29,876	proteins, peptides
1,338	protein/nucl. complexes
1,500	nucleic acids
13	carbohydrates
32,727	total

Search for keyword DNAC
yields mouse zinc finger binding proteins

gateways to access PDB files

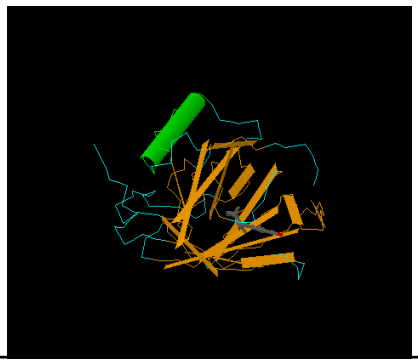
Swiss-Prot, NCBI, EMBL

Protein Data Bank

CATH, Dali, SCOP, FSSP

databases that interpret PDB files

[3] UniProt



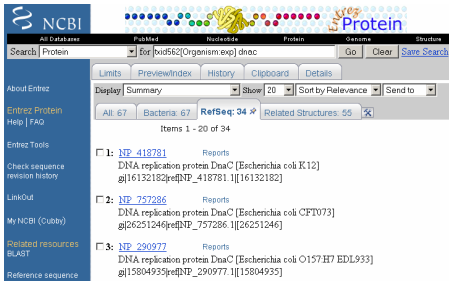
UniProt (Universal Protein Resource) at www.uniprot.org

UniProt combines information in Swiss-Prot, TrEMBL, and PIR. UniProt is comprised of three components

- The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information.
- The **UniProt Reference Clusters (UniRef)** databases combine closely related sequences into a single record.
- The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.

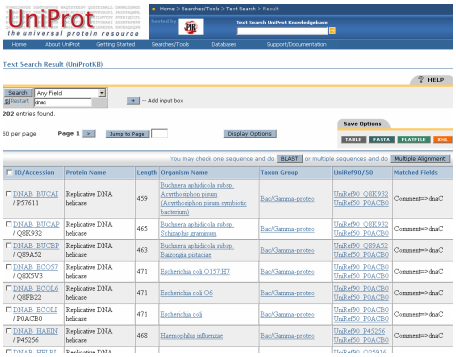
www.uniprot.org

Example: search for *E. coli* DnaC at NCBI



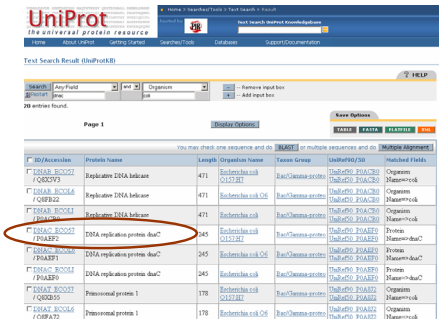
Approach: NCBI→ TaxBrowser → *E. coli* → proteins → dnac → RefSeq → three entries shown here.

Example: search for *E. coli* DnaC at UniProt



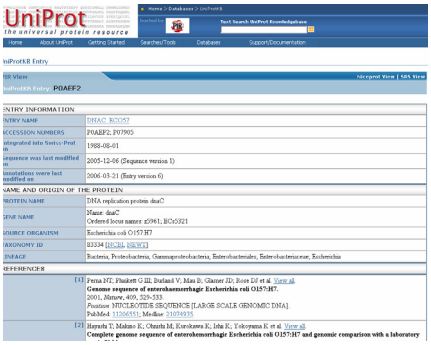
202 entries found

Example: search for *E. coli* DnaC at UniProt



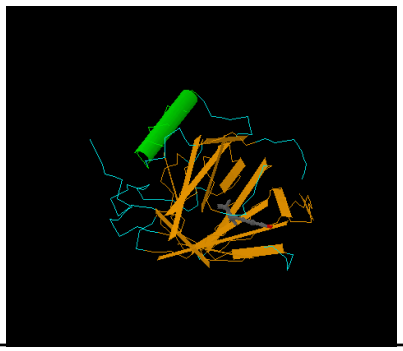
Add input box (organism, coli) and 20 entries found

Example: search for *E. coli* DnaC at UniProt



The UniProt entry links to Entrez Gene (but not RefSeq)

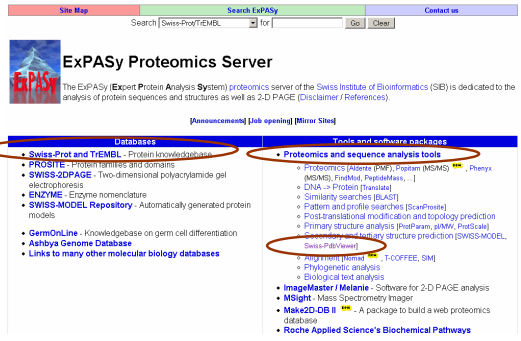
[4] ExPASy / DeepView



ExPASy to access protein and DNA sequences

ExPASy sequence retrieval system
(ExPASy = Expert Protein Analysis System)

Visit http://www.expasy.ch/



ExPASy to access protein and DNA sequences

When you search the ExPASy database, you are now querying the UniProt Knowledgebase.

► UniProtKB/Swiss-Prot; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

UniProtKB/Swiss-Prot Release 49.3 of 21-Mar-2006: 212,425 entries (More statistics)

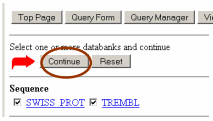
ExPASy to access protein and DNA sequences

When you search the ExPASy database, you are now querying the UniProt Knowledgebase.

► UniProtKB/TrEMBL; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

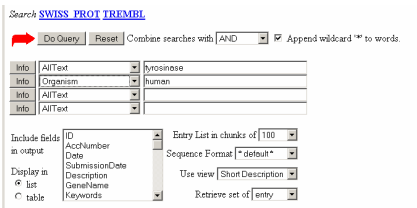
UniProtKB/TrEMBL Release 32.3 of 21-Mar-2006: 2,666,963 entries

Example: find human tyrosinase at ExPASy



From the ExPASy home, click Swiss-Prot → SRS → Start → Continue

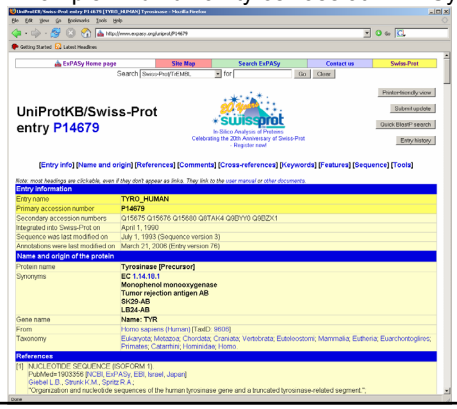
Example: find human tyrosinase at ExPASy



Example: find human tyrosinase at ExPASy

Note the lack of RefSeq and multiple accession numbers
Choose the right protein by inspection (P14679)

Example: find human tyrosinase at ExPASy



Example: find human tyrosinase at ExPASy

Hum. Mutat. 23:630-631(2004).

Comments

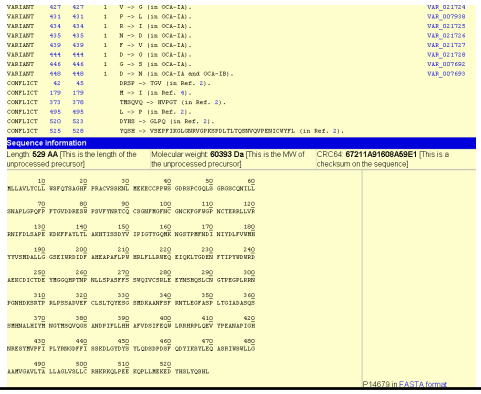
- ***FUNCTION***: This is a copper-containing oxidase that functions in the formation of pigments such as melanins and other polyphenolic compounds. Catalyzes the rate-limiting conversions of tyrosine to DOPA, DOPA to DOPA-quinone and possibly 5,6-dihydroxyindole to indole-5,6 quinone.
- ***CATALYTIC ACTIVITY***: L-tyrosine + L-dopa + O_2 = L-dopa + dopaquinone + H_2O.
- ***COFACTOR***: Binds 2 copper ions per subunit.
- ***SUBCELLULAR LOCATION***: Type I membrane protein. Melanosomal.
- ***ALTERNATIVE PRODUCTS***:
  - Display all isoform sequences in FASTA format
  - **Alternative splicing [2 named forms]**

Name 1
Isoform ID:P14679-1
This is the isoform sequence displayed in this entry.

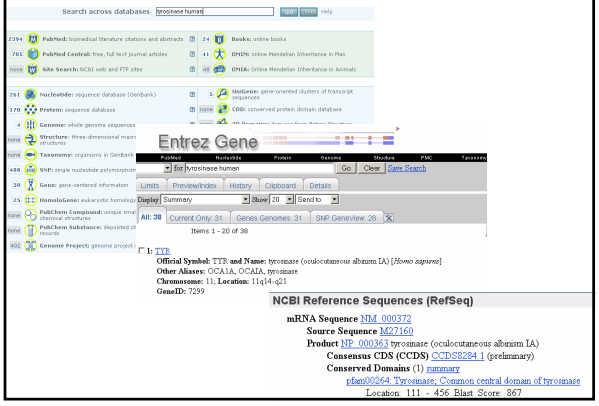
Name 2
Isoform ID:P14679-2
Note: No experimental confirmation available.
Features which should be applied to build the isoform sequence: VSP_006701, VSP_006702

- ***INDUCTION***: Increased expression after UV-B radiation.
- ***POLYMORPHISM***: Compound heterozygosity for the R402Q polymorphism and a mutant allele of TYR is a common cause of autosomal recessive ocular albinism. The R402Q polymorphism is also found in Waardenburg syndrome type II with ocular albinism (WS2-OA) in association with a deletion in the MITF gene.
- ***DISEASE***: Defects in TYR are the cause of oculocutaneous albinism type IA (OCA-IA) [MIM:203100]. OCA-I, also known as tyrosinase negative oculocutaneous albinism, is an autosomal recessive disorder characterized by absence of pigment in hair, skin and eyes. OCA-I is divided into 2 types: type IA, characterized by complete lack of tyrosinase activity due to production of an inactive enzyme, and type IB characterized by reduced activity of tyrosinase. OCA-IA patients presents with the life-long absence of melanin pigment after birth and manifest increased sensitivity to ultraviolet radiation and to predisposition to skin cancer.
- ***DISEASE***: Defects in TYR are the cause of oculocutaneous albinism type IB (OCA-IB) [MIM:606952]; also known as albinism yellow mutant type. OCA-IB patients have white hair at birth that rapidly turns yellow or blond. They manifest the development of minimal-to-moderate amounts of cutaneous and ocular pigment.
- ***DISEASE***: Defects in TYR are the cause of oculocutaneous albinism type I temperature-sensitive (OCA-ITS) [MIM:606952]. OCA-ITS patients have white axillary and scalp hair and pigmented arm and leg hair.
- ***SIMILARITY***: Belongs to the tyrosinase family.
- ***DATABASE***: NAME=Albinism database (ADB); NOTE=OCA-I mutations; WWW="http://www.cbc.umn.edu/tad/".
- ***DATABASE***: NAME=Mutations of the TYR gene; NOTE=Retina International's Scientific Newsletter,

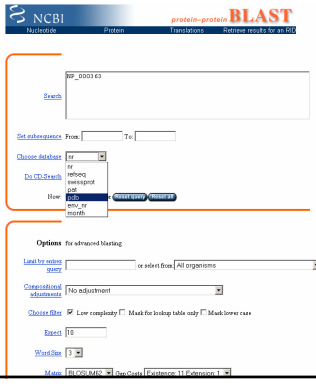
Example: find human tyrosinase at ExPASy



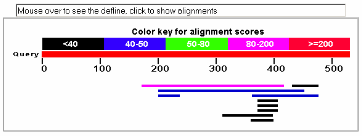
Example: find human tyrosinase at NCBI



Human tyrosinase: blastp against pdb to find known structures



Human tyrosinase: blastp against pdb



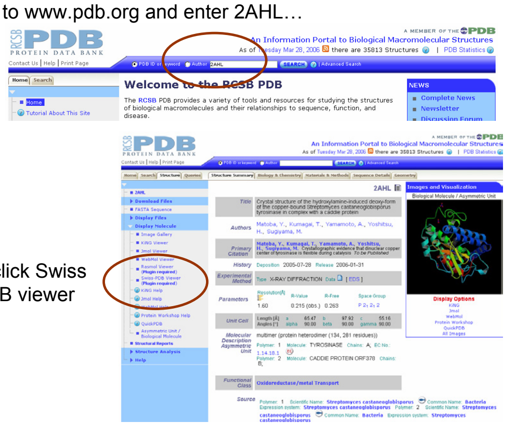
Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value
gi|88192060|pdb|2AHL|A Chain A, Crystal Structure Of The Hydr...	103	9e-23
gi|15988084|pdb|1JS8|B Chain B, Structure Of A Functional Uni...	43.9	8e-05
gi|6137618|pdb|1BT3|A Chain A, Catechol Oxidase From Ipomoea ...	41.6	4e-04
gi|31615506|pdb|1LNL|C Chain C, Structure Of Deoxygenated Hem...	41.6	4e-04
gi|1421539|pdb|1NOL| Oxygenated Hemocyanin (Subunit Type Ii)...	32.7	0.19
gi|2392430|pdb|1LL1| Hydroxo Bridge Met Form Hemocyanin From Li	32.7	0.19
gi|999996|pdb|1OXY| Hemocyanin Subunit Ii Complexed With Oxygen	32.7	0.19
gi|49259128|pdb|1SR9|B Chain B, Crystal Structure Of Leua Fro...	29.6	1.6
gi|494086|pdb|1HCY| Arthropodan Hemocyanin (Deoxygenated) Re...	27.7	6.1
gi|55670764|pdb|1X72|A Chain A, Crystal Structure Of A Putati...	27.3	8.0

Human tyrosinase: best blastp pdb match to structure 2AHL

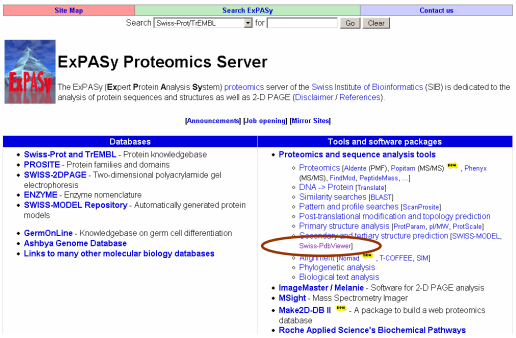
```
>gi|88192060|pdb|2AHL|A Chain A, Crystal Structure Of The Hydroxylamine-Induced Deoxy-Form
Of The Copper-Bound Streptomyces Castaneoglobisporus Tyrosinase
In Complex With A Caddie Protein
 gi|88192058|pdb|2AHK|A Chain A, Crystal Structure Of The Met-Form Of The Copper-Bound
Streptomyces Castaneoglobisporus Tyrosinase In Complex With
A Caddie Protein Obtained By Soking In Cupric Sulfate For
6 Months
 gi|88191780|pdb|1WXC|A Chain A, Crystal Structure Of The Copper-Free Streptomyces Castaneoglobisporus
Tyrosinase Complexed With A Caddie Protein
 gi|88191778|pdb|1WX5|C Chain C, Crystal Structure Of The Copper-Free Streptomyces Castaneoglobisporus
Tyrosinase Complexed With A Caddie Protein
In The Monoclinic Crystal
 gi|88191776|pdb|1WX5|A Chain A, Crystal Structure Of The Copper-Free Streptomyces Castaneoglobisporus
Tyrosinase Complexed With A Caddie Protein
In The Monoclinic Crystal
 gi|88191774|pdb|1WX4|A Chain A, Crystal Structure Of The Oxy-Form Of The Copper-Bound
Streptomyces Castaneoglobisporus Tyrosinase Complexed With
A Caddie Protein Prepared By The Addition Of Dithiothreitol
 gi|88191772|pdb|1WX3|A Chain A, Crystal Structure Of The Met-Form Of The Copper-Bound
Streptomyces Castaneoglobisporus Tyrosinase Complexed With
A Caddie Protein
 gi|88191770|pdb|1WX2|A Chain A, Crystal Structure Of The Oxy-Form Of The Copper-Bound
Streptomyces Castaneoglobisporus Tyrosinase Complexed With
A Caddie Protein Prepared By The Addition Of Hydrogenperoxide
Length=281

 Score =  103 bits (257),  Expect = 9e-23
 Identities = 78/253 (30%), Positives = 115/253 (45%), Gaps = 49/253 (19%)

Query  173  YDLFVWMHYYVSMDALLGGSEIWRDIDFAHEAPAFLPWHRLFLLRWEQEIQKLTGDENFT  232
            YD FV  H    M     G           H +P+FLPWHR FLL +EQ +Q +   D + T
Sbjct  31   YDEFVRTHNEFIMSDTDSGERT------GHRSPSFLPWHRRFLLDFEQALQSV--DSSVT  82

Query  233  IPYWDWRDAE--KCDICTDEYMGGQHPTNPNLLSPASFFSS---WQI---VCSRLEEYNS  284
            +PYWDW       +  +    +++GG    +      F +S    W I    V SR
Sbjct  83   LPYWDWSADRTVRASLWAPDFLGGTGRSTDGRVMDGPFAASTGNWPINVRVDSRTY----  138

Query  285  HQSLCNGTPEGPLRRNPGNHDKSRTPRLPSSADVEFCLSLTQYESGSMDKAANFSFRNTL  344
                         LRR+ G          LP+ A+VE  L+++ Y+     + A+    FRN L
Sbjct  139  ------------LRRSLGGS----VAELPTRAEVESVLAISAYDLPPYNSASE-GFRNHL  181

Query  345  EGFASPLTGIADASQSSMHNALHIYMNGTMSQVQGSANDPIFLLHHAFVDSIFEQWLRRH  404
            EG+               ++HN +H+++ G M+     S NDP+F LHHA+VD ++ +W RRH
Sbjct  182  EGWRGV----------NLHNRVHVWVGGQMA-TGVSPNDPVFWLHHAYVDKLWAEWQRRH  230

Query  405  RPLQEVYPEANAP  417
             P    P   P
Sbjct  231  -PDSAYVPTGGTP  242
```

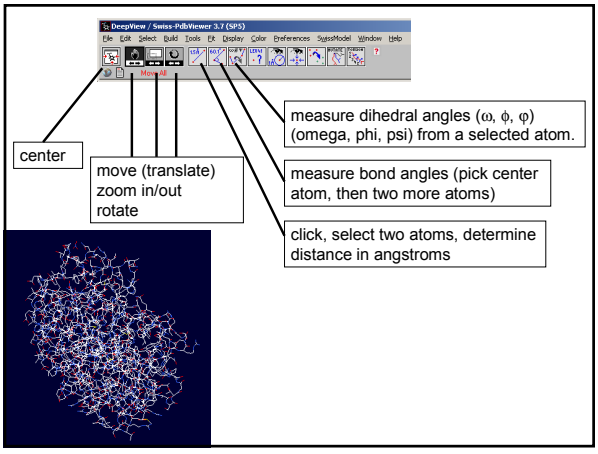
Go to www.pdb.org and enter 2AHL…

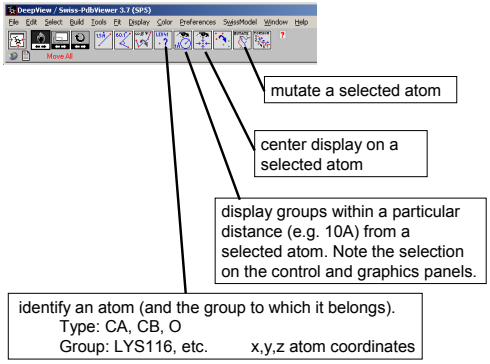


…click Swiss PDB viewer

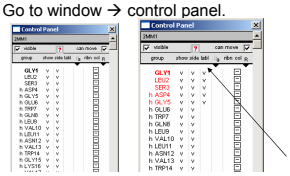
View protein structures with DeepView



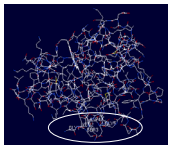




► Go to window → control panel.



►Shift/click to select the first five amino acid residues of myoglobin. They should appear red.
► Click “labl” (i.e. label)(see arrow, above right). Those five residues now have a “v”.
►Inspect the display panel; those five residues are labeled.



Download and practice using DeepView!
Try using myoglobin.

The ExPASy download site includes a helpful web-based tutorial
http://www.usm.maine.edu/~rhodes/SPVTut/