Prediction of structure and function of proteins

University of Thessaly

Introduction to protein databases

Protein database development

- Protein databases (DBs) are the second largest biological DBs (after DNA DBs)
- Important because proteins exhibit large variability in their structure and function
- An important focus of modern bioinformatics is the analysis of protein sequences and functional data related to them, that are constantly being produced through wet-lab experiments

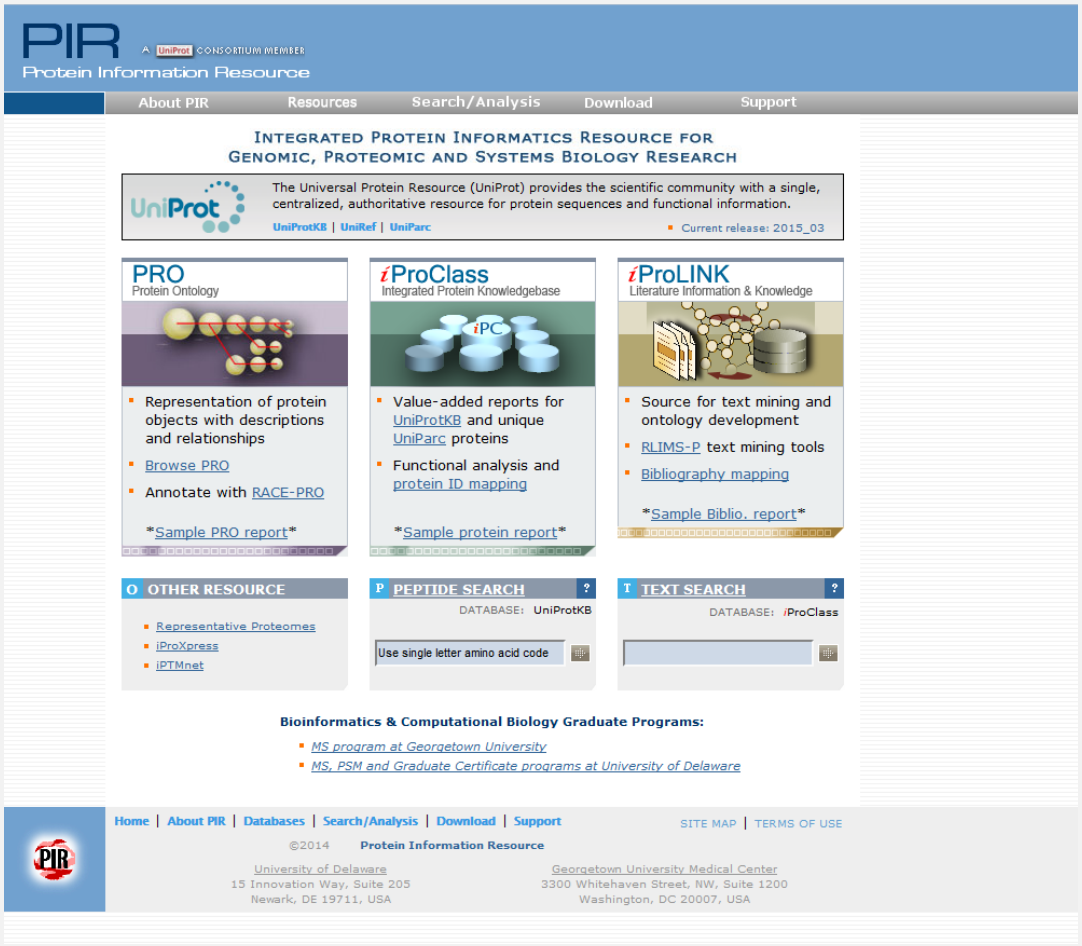
- Atlas of protein sequence and structure (1966) → the first protein sequence DB (before the advent of bioinformatics). Today known as Protein Information Resource (PIR)
- Protein data bank (PDB, 1971) → DB for structural data (1971), still remains the most widely used DB regarding macromolecular structures
- United Protein Databases (UniProt, 2003) → the largest protein sequence and protein function DB, created by the unificcation of SWISS-PROT, TrEMBL andPIR

PIR
(Protein Information Resourse)

http://pir.georgetown.edu

PIR was developed in 1984 by National Biomedical Research Foundation (NBRF) in USA in order to provide researchers with information regarding protein sequences.

Between 1965-1978, NBRF had the first complete collection of macromolecular sequences (Atlas of Protein Sequence and Structure).



PIR entry: IPPG

```
>P1;IPPG
insulin precursor - pig
C;Species: Sus scrofa domestica (domestic pig)
…
C;Accession: A01583; A94572; S16492; A60835; B60835
…
C;Keywords: hormone; pancreas
F;1-30/Domain: insulin chain B #status experimental
F;1-30,64-84/Product: insulin #status experimental
F;33-63/Domain: connecting peptide #status experimental
F;64-84/Domain: insulin chain A #status experimental
F;7-70,19-83,69-74/Disulfide bonds: #status experimental
>P1;IPPG

FVNQHLCGSH LVEALYLVCG ERGFFYTPKA RREAENPQAG AVELGGGLGG LQALALEGPP
QKRGIVEQCC TSICSLYQLE NYCN*
```

SwissProt

http://www.expacy.ch/sprot/

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

EBI and SIB created SwissProt and TrEMBL DBs. SwissProt was the main project of Amos Bairoch Msc and PhD studies back in 1990s at SIB and then was further developed by Rolf Apweiler at EBI.

Swissprot is different than other protein DBs, because:

- Manual annotation – high quality data, including function, classification, post-translational modifications
- Minimal redundancy
- Cross-references to many other DBs
- Detailed manual

Swiss Institute of Bioinformatics

Institut Suisse de Bioinformatique
Schweizerisches Institut für Bioinformatik

Home | The Institute | Info | Groups | Projects | Teaching

Finding People

--- News ---

Swiss Consortium to Manage GISAID Database. In order to contribute to the worldwide efforts against the spread of avian flu the Global Initiative on Sharing Avian Influenza Data (GISAID) has entered into an agreement with the Swiss Institute of Bioinformatics to lead a consortium that will develop a database on influenza viruses ...
> Read more.

Protein Spotlight Update: Heavy Metal. Our grandmothers used to make jam in huge copper pans. The same copper pans that you would see hanging over the stove, with that distinctive green patina lining the inside. The same green patina that children instinctively knew was poisonous ...
> Read more.

Brand new PhD School in

Home

The SIB is an academic not-for-profit foundation established on March 30, 1998 whose mission is to promote research, the development of databanks and computer technologies, teaching and service activities in the field of bioinformatics, in Switzerland with international collaborations.

servers:
- ExPASy proteomics server
- Swiss node of EMBnet

databases:
- Ashbya Genome Database
- Cancer Immunome Database
- Eukaryotic Promoter Database (EPD)
- GermOnline
- MyHits
- PROSITE
- Swiss-Prot and TrEMBL
- SWISS-2DPAGE
- SWISS-MODEL Repository

software tools:
- ESTScan
- GoCluster
- ImageMaster / Melanie
- iMolTalk
- Make2D-DB II
- MSight
- SIBsim4
- SWISS-MODEL
- Swiss-PdbViewer
- tromer

partners:
- Bioalps (Lake Geneva Biocluster)
- Biozentrum, Basel
- University Hospital Center of Vaud (CHUV)
- Swiss Federal Institute of Technology Lausanne (EPFL)
- Swiss Federal Institute of Technology Zürich (ETHZ)
- GeneBio
- Les Hôpitaux Universitaires de Genève (HUG)
- Ludwig Institute for Cancer Research (LICR)
- Swiss Institute for Experimental Cancer Research (ISREC)
- University of Geneva
- University of Lausanne

external links:
- B3 Biotech Center

ExPASy
Bioinformatics Resource Portal

Find resources | search | help

Visual Guidance
Categories
- proteomics
  - protein sequences and identification
  - mass spectrometry and 2-DE data
  - protein characterisation and function
  - families, patterns and profiles
  - post-translational modifications
  - protein structure
  - protein-protein interaction
  - similarity search/alignment
- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A..Z
Links/Documentation

SIB resources
External resources - *(No support from the ExPASy Team)*

Databases

- UniProtKB • functional information on proteins • [more]
- UniProtKB/Swiss-Prot • protein sequence database • [more]
- STRING • protein-protein interactions • [more]
- SWISS-MODEL Repository • protein structure homology models • [more]
- PROSITE • protein domains and families • [more]
- ViralZone • portal to viral UniProtKB entries • [more]
- neXtProt • human proteins • [more]

- EMBnet services • bioinformatics tools, databases and courses • [more]
- ENZYME • enzyme nomenclature • [more]
- GPSDB • gene and protein synonyms • [more]
- HAMAP • UniProtKB family classification and annotation • [more]
- MetaNetX • Metabolic Network Repository & Analysis • [more]
- MIAPEGelDB • MIAPE document edition • [more]
- MyHits • protein domains database and tools • [more]
- PANDITplus • protein families and domains resources • [more]
- PaxDb • protein abundance database • [more]
- Prolune • Popular science articles (in French) • [more]
- Protein Model Portal • structural information for a protein • [more]
- Protein Spotlight • Informally written reviews on proteins • [more]
- SugarBind • pathogen sugar-binding • [more]
- SWISS-2DPAGE • proteins on 2-D and SDS PAGE maps • [more]
- SwissBiolsostere • biolsosteres for small molecules • [more]
- SwissSidechain • non-natural amino-acid sidechains • [more]

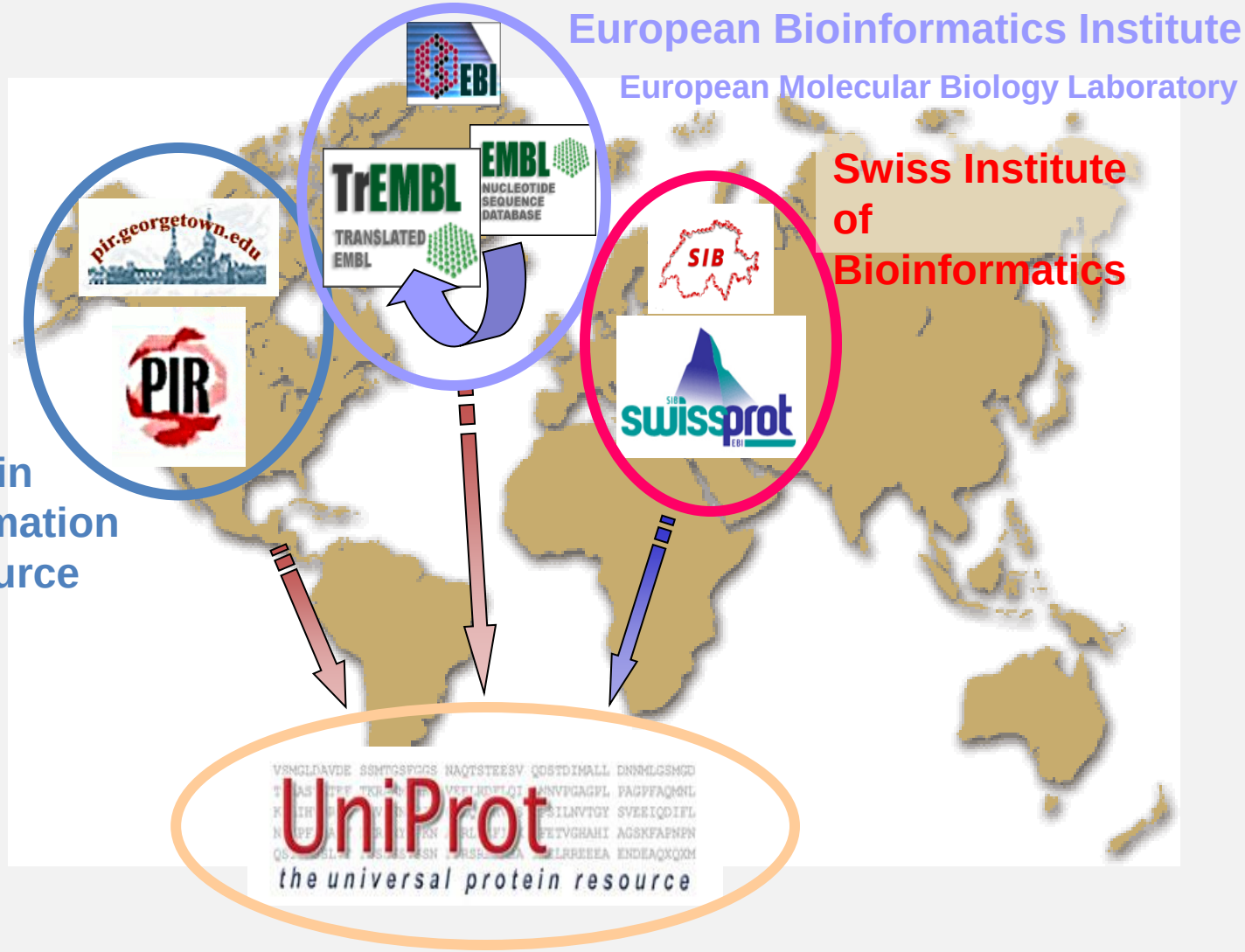
Tools

- SWISS-MODEL Works
- SwissDock • protein lig
- 2ZIP • Prediction of leuc
- 3of5 • find user-defined
- AACompIdent • protein
- AACompSim • amino a
- Agadir • Prediction of th
- ALF • simulation of gen
- Alignment tools • Four t
- AllAll • protein sequenc
- APSSP • Advanced Pro
- Ascalaph • Molecular m
- big-PI • predict GPI mod
- Biochemical Pathways
- BLAST • sequence sim
- BLAST (UniProt) • BLA
- BLAST - NCBI • Biologi
- BLAST - PBIL • BLAST
- Blast2Fasta • Blast to F
- boxshade • MSA pretty
- CFSSP • Protein secon
- ChloroP • chloroplast tra
- Click2Drug • Directory o

Swissprot text entry

```
ID   INS_PIG        STANDARD;      PRT;   108 AA.
AC   P01315; Q9TSJ5;
…
DE   INSULIN PRECURSOR.
GN   INS.
OS   Sus scrofa (Pig).
…
CC   -!- FUNCTION: INSULIN DECREASES BLOOD GLUCOSE CONCENTRATION. IT
CC       INCREASES CELL PERMEABILITY TO MONOSACCHARIDES, AMINO ACIDS AND
…
DR   EMBL; AF064555; AAC77920.1; ALT_INIT. [EMBL / GenBank / DDBJ]
…
KW   Insulin family; Hormone; Glucose metabolism; Signal; 3D-structure.
FT   SIGNAL        1     24
FT   CHAIN        25     54       INSULIN B CHAIN.
…
SQ   SEQUENCE   108 AA;  11671 MW;  CB4491B429858EBE CRC64;
     MALWTRLLPL LALLALWAPA PAQAFVNQHL CGSHLVEALY LVCGERGFFY TPKARREAEN
     PQAGAVELGG GLGGLQALAL EGPPQKRGIV EQCCTSICSL YQLENYCN
//
```

The UniProt consortium



UniProt

http://www.uniprot.org

- ✓ UniProt (Uniprot Knowledge Base) is a collaborative project between 3 institutes, namely the European Bioinformatics Institute (EBI -UK), the Swiss Institute of Bioinformatics (SIB-CH), and the Protein Information Resource (PIR - USA).
- ✓ SIB contributed a well-annotated protein sequence DB
- ✓ EBI contributed TrEMBL, an automatic, not annotated translated nucleotide DB
- ✓ PIR contributed their own protein sequence DB, as well as a group of protein families (PSD)
- ✓ Uniprot contains 3 sub-DBs:
  - ✓ UniProtKB (Swiss-Prot + TrEMBL)
  - ✓ UniRef
  - ✓ UniParc
- ✓ Uniprot is updated monthly and has 3 main FTP servers, one in each institute
- ✓ Offers lots of functionalities, e.g. text or BLAST searches, as well as Multiple Sequence Alignment (MSA) tool (ClustalO) and retrieve/mapping of protein IDs to respective entries

TrEMBL

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

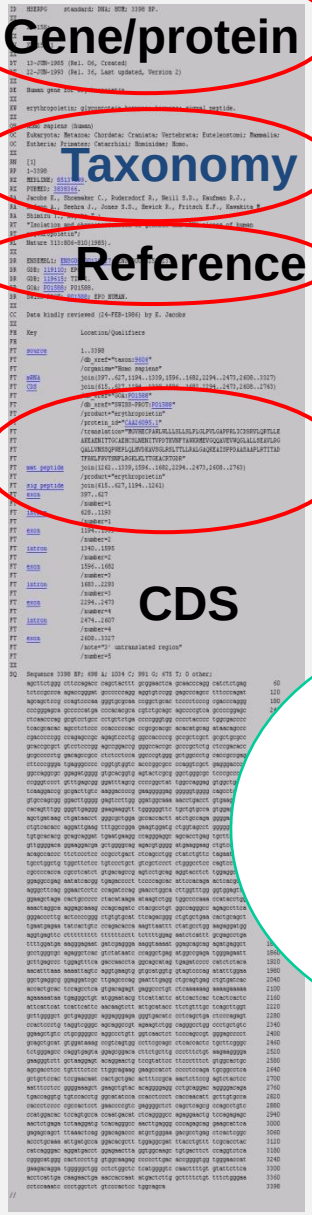
- TrEMBL is comprised of two parts: SP-TrEMBL which contains entries that will be included in Swiss-Prot and REM-TrEMBL which contains entries that will not be part of Swiss-Prot. It can have very short protein sequences (from 8 amino-acids long) or sequences that are under patent.

- Contrary to Swiss-Prot, TrEMBL is based on automated annotation instead of manual curation

- TrEMBL does not translate DNA sequences, nor does it use gene finding software. It provides only the coding sequence (CDS) that is recommended by the researchers that deposit it in genomic databases (EMBL/Genbank/DDBJ)

- The CDS and the respective protein sequence can have been experimentally verified or derived from prediction methods. This is not clear in a TrEMBL entry.

- TrEMBL does not validate any sequence. The quality of the data is solely dependent on the researcher that submits it.



EMBL

TrEMBL

Swiss-Prot

Automated extraction of protein sequence – CDS, gene name and literature references. **Automated annotation**

Manual curation of the sequence and other biologically-related data (literature, other DBs); addition of cross-references, alternative splicing

The Universal Protein resource components

Release 2018_11 (Dec 2018)

UniProtKB	UniRef100 UniRef 90 UniRef 50	UniProt Archives
UniProtKB/TrEMBL Computer annotated protein sequences 137,213,158 entries 832,433 species **UniProtKB/Swiss-Prot** Manually annotated protein sequences 558,898 entries ~9,463 species produced by SIB and EBI	• One **UniRef100** entry = **All identical sequences** (including fragments).≈ 170M entries • One **UniRef90** entry = Sequences that have at least **90% or more identity** ≈ 85M entries • One **UniRef50** entry = Sequences that are at least **50% or more identity** ≈ 32M entries Independent of species. produced by PIR	~244,490,125 entries **Archived raw protein sequences,** found in publicly accessible databases: Swiss-Prot, TrEMBL, PIR, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, Patent Offices etc produced by EBI

Use with extreme caution: Contains pseudogenes, incorrect CDS predictions, etc…

Uniprot home page

UniProt

UniProtKB | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

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.

UniProtKB
UniProt Knowledgebase

Swiss-Prot (551,987)
Manually annotated and reviewed.

TrEMBL (66,905,753)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data
Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2016_08
Butterfly fashion: all they need is cortex | Cross-references to CDD | Change of the cross-references to VectorBase and WormBase | Pepti...

UniProt release 2016_07
(Bacterial) immigration under control

News archive

Getting started

Text search
Our basic text search allows you to search all the resources available

BLAST
Find regions of similarity between your sequences

Sequence alignments
Align two or more protein sequences using the Clustal Omega program

Retrieve/ID mapping
This tool merges the "Retrieve" and "ID Mapping" tools

UniProt data

Download latest release
Get the UniProt data

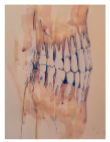
Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us
The UniProt Consortium

Submit your data
Submit your sequences and annotation updates

SPARQL
Query UniProt data using a SQL like graph query language

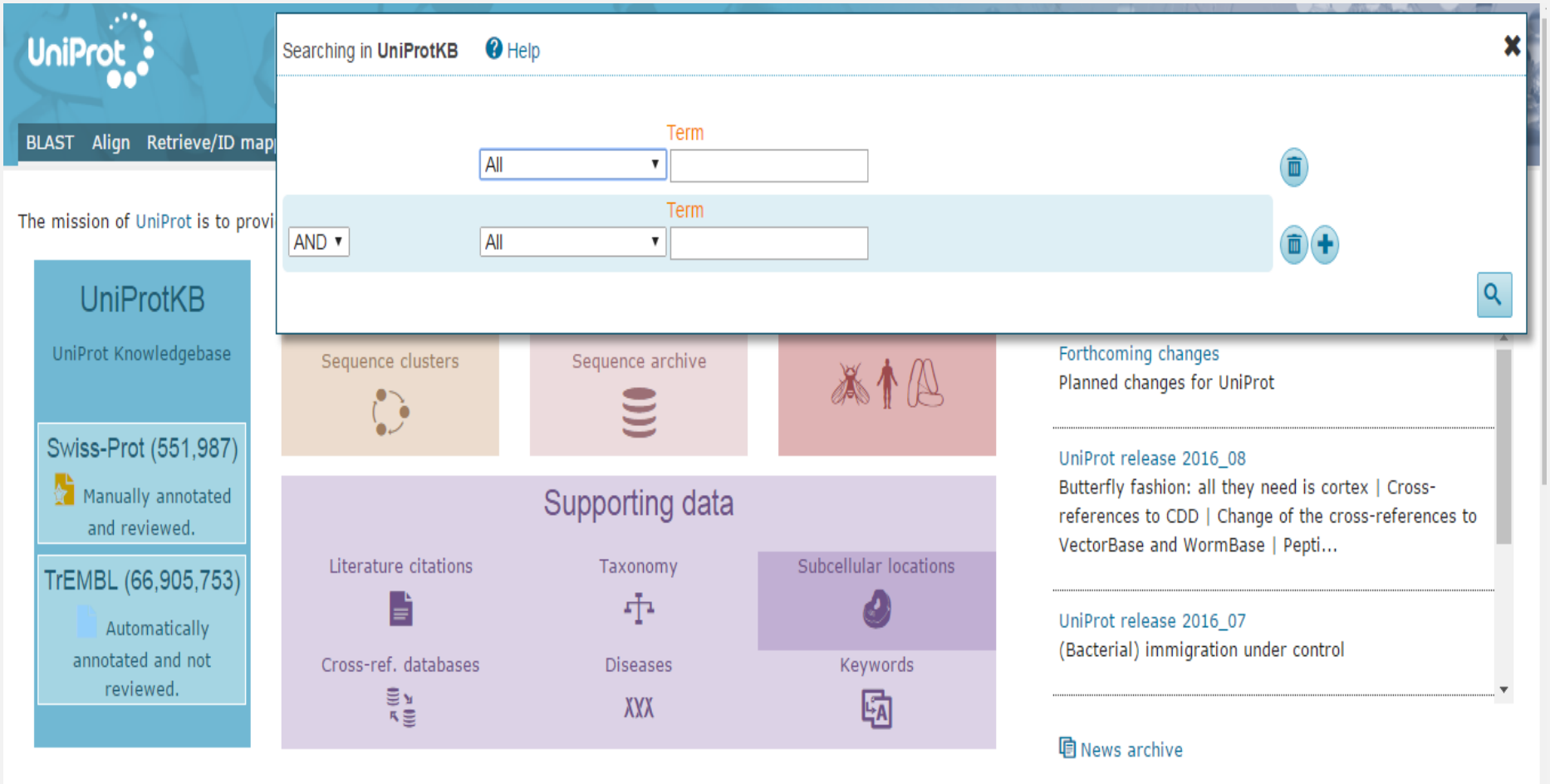
Protein spotlight



A Loosening Of Habits
August 2016

We are not alone. From the day we are born, we carry with us hordes of microorganisms which, if all is going well, live off us while giving something in return. This micro-universe which is an integral part of our physiology has been called the microbiome. In the recent years, scientists have demonstrated the importance a microbiome has on our overall health; how it can influence our well-being as it can be at the heart of a disease. Take our mouths for instance...

Uniprot search



Uniprot search results

UniProtKB | database:(type:pfam id:PF00001) AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]" | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search | Help Contact

UniProtKB results

About UniProtKB | Basket

Filter by

Reviewed (285) Swiss-Prot

Popular organisms

Human (285)

Proteomes

UP000005640 (285)

View by

Taxonomy

Keywords

Gene Ontology

Enzyme class

Pathway

UniRef

Your results in sequence

BLAST | Align | Download | Add to basket | Columns

1 to 25 of 285 | Show 25

Download selected (285)
Download all (285)
Format: Text
Compressed Uncompressed
Preview first 10
Go

Entry	Entry			Gene names	Organism	Length
			lected. (or select only the results on this page)			
P61073	CXCR		r type 4 (CXC-R4) (CXCR-4) (FB22) (Fusin) derived seven transmembrane domain saccharide-associated protein 3) (LAP-3) (NPYRL) (Stromal cell-derived factor 1 (CD antigen CD184)	CXCR4	Homo sapiens (Human)	352
Q99527	GPER1_HUMAN		G-protein coupled estrogen receptor 1 (Chemoattractant receptor-like 2) (Flow-induced endothelial G-protein coupled receptor 1) (FEG-1) (G protein-coupled estrogen receptor 1) (G-protein coupled receptor 30) (GPCR-Br) (IL8-related receptor DRY12) (Lymphocyte-derived G-protein coupled receptor) (LYGPR) (Membrane estrogen receptor) (mER)	GPER1 CEPR, CMKRL2, DRY12, GPER, GPR30	Homo sapiens (Human)	375
P51681	CCR5_HUMAN		C-C chemokine receptor type 5 (C-C CKR-5) (CC-CKR-5) (CCR-5) (CCR5) (CHEMR13) (HIV-1 fusion coreceptor) (CD antigen CD195)	CCR5 CMKBR5	Homo sapiens (Human)	352
P55085	PAR2_HUMAN		Proteinase-activated receptor 2 (PAR-2) (Coagulation factor II receptor-like 1) (G-protein coupled receptor 11) (Thrombin receptor-like 1) [Cleaved into: Proteinase-activated receptor 2, alternate cleaved 1; Proteinase-activated receptor 2, alternate cleaved 2]	F2RL1 GPR11, PAR2	Homo sapiens (Human)	397

Non-redundant databases

- Definition: Repeated entries
- Cause: Identical or overlapping sequences originating from the same or different author(s)
- **No** redundancy in Swiss-Prot
- **How**? When different genes in the same species code for the same protein sequences, they are merged under the same entry in UniProtKB/Swiss-Prot and all gene names appear in one field (http://www.uniprot.org/uniprot/P68431).
- Non-redundancy inUniProtKB/Swiss-Prot means that identical sequences are presented in one entry. However, if the identical sequences derive from different species, then they are multiple entries, one per species.

Redundancy

- Contain only sequences / search can be done using a sequence
- Created by combining more than one DBs, e.g:
  - NR Nucleic (genbank+EMBL+DDBJ+PDB DNA)
  - NR Protein (SWISS-PROT+TrEMBL+GenPept+PDB protein)

Protein existence evidence

- Since most protein sequences are derived from translation of nucleotide sequences (i.e. predictions), the PE line in a Uniprot entry informs us regarding the existence of the protein

- 'Protein existence evidence' has 5 confidence levels:
    1. Evidence at protein level
    2. Evidence at transcript level
    3. Inferred from homology
    4. Predicted
    5. Protein uncertain - Unassigned (used mostly in TrEMBL)

Annotation errors

- C. Hardley, EMBO reports, 4(9), 2003.

  "Sequences are rarely deposited in a "mature" state; as with all scientific research, DNA and protein annotation is a continual process of learning, revision and corrections." ….

  "Sequencing error rates: ~1 base in 10,000" ….

  "Making people aware of errors is good and great; making people aware that they're responsible also for correcting errors is even greater"

- Fixing sequence errors is a key point in the effort for providing the scientific community with reliable entries

- The manually annotated entries consist of information derived from literature, specialised DBs, expert researchers/curators, idea exchange and brainstorming

- Clear distinction from data/information obtained by computational analyses

A manually annotated entry

P63284 - CLPB_ECOLI

Protein: **Chaperone protein ClpB**

Gene: **clpB**

Organism: *Escherichia coli (strain K12)*

Status: Reviewed - Annotation score: ⓪⓪⓪⓪⓪ - Experimental evidence at protein level

Display: None

- Function
- Names & Taxonomy
- Subcellular location
- Pathology & Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequences (2)
- Cross-references
- Publications
- Entry information
- Miscellaneous
- Similar proteins

BLAST | Align | Format | Add to basket | History

Function

Part of a stress-induced multi-chaperone system, it is involved in the recovery of the cell from heat-induced damage, in cooper
protein aggregates. Protein binding stimulates the ATPase activity; ATP hydrolysis unfolds the denatured protein aggregates, w
ClpB-bound aggregates, contributing to the solubilization and refolding of denatured protein aggregates by DnaK. 3 Publications

Regions

Feature key	Position(s)	Length	Description
Nucleotide binding	206 - 213	8	ATP 1
Nucleotide binding	605 - 612	8	ATP 2

GO - Molecular function

- ATP binding Source: UniProtKB-KW
- identical protein binding

GO - Biological process

- protein processing Source: InterPro
- response to heat Source
- response to unfolded protein Source: EcoCyc

Complete GO annotation...

Keywords - Molecular function

Chaperone

Keywords - Biological process

Important fields in a Uniprot entry

- All protein and respective gene names
- Biological origin of the protein, with links to taxonomic DBs
- Literature references
- Summary of what is known regarding the protein, e.g. function, alternative splicing, post-translational modifications, tissue expression, 3D structure etc
- Multiple cross-references to other DBs
- Selected keywords
- Description of important sequence features of the protein, e.g. signal peptide, transmembrane segments, PTMs, sequence variations

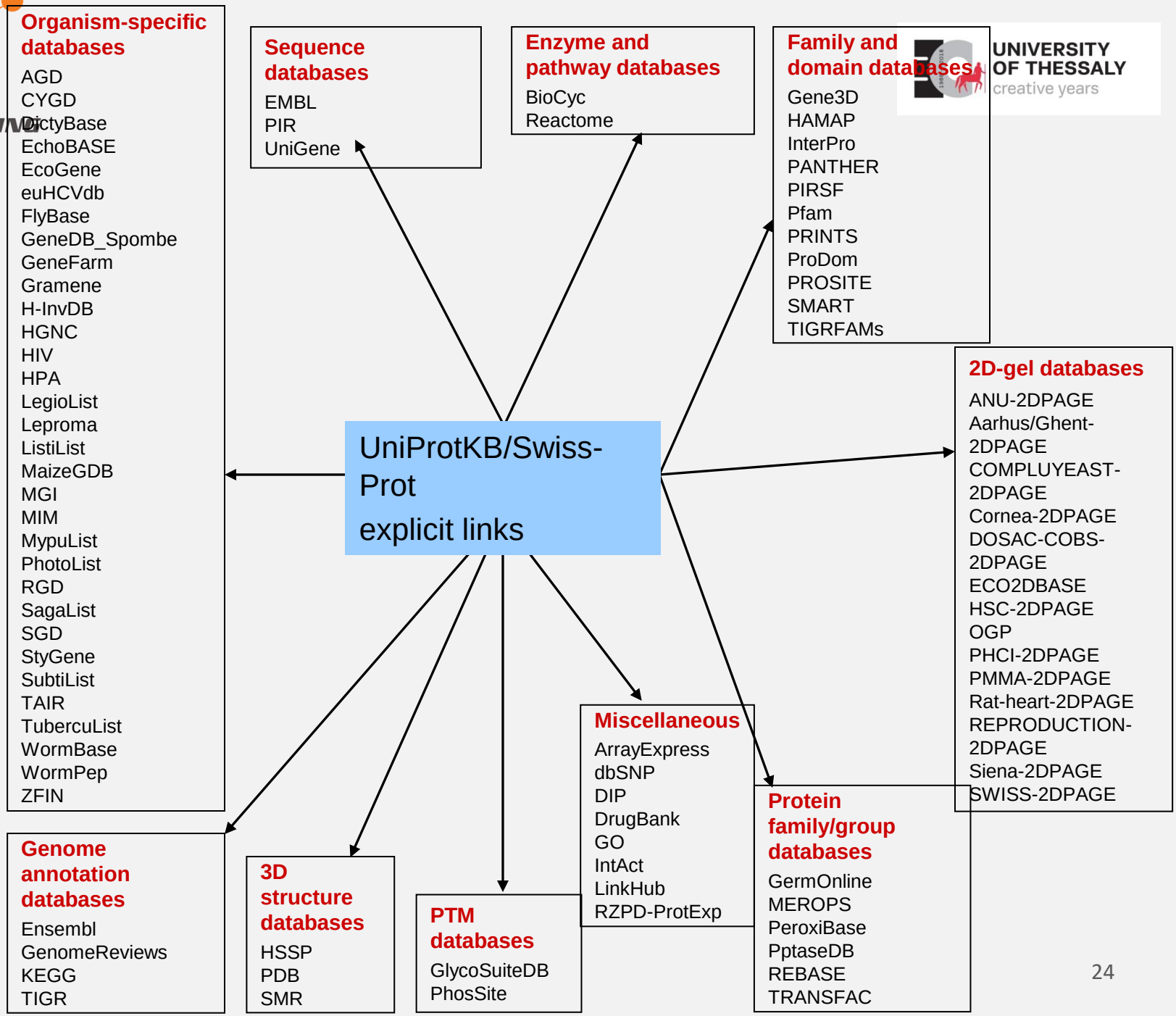
Cross-references in Uniprot

http://www.uniprot.org/docs/dbxref

- Swiss-Prot was the first DB that contained cross-references to other DBs

- A Uniprot Accession Number (AC) can be used by various other DBs (e.g. BLOCKS domain db) as an identifier for their entries, i.e. facilitating direct link to Uniprot, without being explicitly referenced in Uniprot (implicit cross-references)

- Currently (Dec 2018), there are >170 cross-referenced DBs in Uniprot
  - DNA (EMBL/GenBank/DDBJ)
  - 3D-structure (PDB)
  - Family and domain (InterPro, HAMAP, PROSITE, Pfam, etc.)
  - genomic (OMIM, MGI, FlyBase, SGD, SubtiList, etc.)
  - specialized DBs (e.g.GlycoSuiteDB, PhosSite, MEROPS)
  - literature (PubMed)


UniProtKB/Swiss-Prot
explicit links
Organism-specific databases
AGD
CYGD
DictyBase
EchoBASE
EcoGene
euHCVdb
FlyBase
GeneDB_Spombe
GeneFarm
Gramene
H-InvDB
HGNC
HIV
HPA
LegioList
Leproma
ListiList
MaizeGDB
MGI
MIM
MypuList
PhotoList
RGD
SagaList
SGD
StyGene
SubtiList
TAIR
TubercuList
WormBase
WormPep
ZFIN
Sequence databases
EMBL
PIR
UniGene
Enzyme and pathway databases
BioCyc
Reactome
Family and domain databases
Gene3D
HAMAP
InterPro
PANTHER
PIRSF
Pfam
PRINTS
ProDom
PROSITE
SMART
TIGRFAMs
2D-gel databases
ANU-2DPAGE
Aarhus/Ghent-2DPAGE
COMPLUYEAST-2DPAGE
Cornea-2DPAGE
DOSAC-COBS-2DPAGE
ECO2DBASE
HSC-2DPAGE
OGP
PHCI-2DPAGE
PMMA-2DPAGE
Rat-heart-2DPAGE
REPRODUCTION-2DPAGE
Siena-2DPAGE
SWISS-2DPAGE
Miscellaneous
ArrayExpress
dbSNP
DIP
DrugBank
GO
IntAct
LinkHub
RZPD-ProtExp
Protein family/group databases
GermOnline
MEROPS
PeroxiBase
PptaseDB
REBASE
TRANSFAC
Genome annotation databases
Ensembl
GenomeReviews
KEGG
TIGR
3D structure databases
HSSP
PDB
SMR
PTM databases
GlycoSuiteDB
PhosSite

UniProt

Keywords | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

Keywords results

UniProtKB entries are tagged with keywords that can be used to retrieve particular subsets of entries. There are 10 categories of keywords:

- Biological process
- Cellular component
- Coding sequence diversity
- Developmental stage
- Disease
- Domain
- Ligand
- Molecular function
- Post-translational modification
- Technical term

Information about the usage of this controlled vocabulary in UniProtKB entries can be found in the user manual.

Help | Tutorials and videos | Downloads

Keywords

Molecular function	Hydrolase, Protease, Serine protease
Biological process	Blood coagulation, Fibrinolysis, Hemostasis

UniProt

Keywords | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

Keyword - Blood coagulation (KW-0094)

Map to

- UniProtKB (513)
- Reviewed (343) Swiss-Prot
- Unreviewed (170) TrEMBL

Format

Keywords navigation	› Hemostasis › Fibrinolysis › Hemophilia › Thrombophilia › von Willebrand disease
Definition	Protein involved in blood clotting, a complex enzymatic cascade, in which the activated form of one factor catalyzes the activation of the next factor. Both, the extrinsic clotting pathway, induced by a damaged surface, and the intrinsic pathway, induced by a trauma, converge in a final common pathway to form cross-linked fibrin clots.
Category	› Biological process
GO	› blood coagulation [GO:0007596]
Graphical	KW-0094 Blood coagulation → KW-0356 Hemostasis → KW-9999 Biological process

UniProt

UniProtKB | keyword:"Blood coagulation [KW-0094]" | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB results

About UniProtKB | Basket

Filter by

- Reviewed (343) Swiss-Prot
- Unreviewed (170) TrEMBL

Popular organisms

- Human (50)
- Mouse (49)
- Bovine (34)
- Rat (33)
- Zebrafish (3)
- Other organisms

Go

View by

- Results table
- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

UniRef

Your results in sequence clusters with identity of: 100%, 90% or 50%

Demo

Help video

BLAST | Align | Download | Add to basket | Columns

1 to 25 of 513 | Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
P23605	ACH2_LONAC	**Achelase-2**		Lonomia achelous (Giant silkworm moth) (Saturnid moth)	214
P01009	A1AT_HUMAN	**Alpha-1-antitrypsin**	**SERPINA1** AAT, PI, PRO0684, PRO2209	Homo sapiens (Human)	418
P23604	ACH1_LONAC	**Achelase-1**		Lonomia achelous (Giant silkworm moth) (Saturnid moth)	213
P08758	ANXA5_HUMAN	**Annexin A5**	**ANXA5** ANX5, ENX2, PP4	Homo sapiens (Human)	320
Q4FZU6	ANXA8_RAT	**Annexin A8**	**Anxa8**	Rattus norvegicus (Rat)	327
P32262	ANT3_SHEEP	**Antithrombin-III**	**SERPINC1** AT3	Ovis aries (Sheep)	465
P15358	ANTA_HAEOF	**Antistasin**		Haementeria officinalis (Mexican leech)	136
P81287	ANXA5_BOVIN	**Annexin A5**	**ANXA5** ANX5	Bos taurus (Bovine)	321
P41361	ANT3_BOVIN	**Antithrombin-III**	**SERPINC1** AT3	Bos taurus (Bovine)	465
P81050	ANT3_MESAU	**Antithrombin-III**	**SERPINC1** AT3	Mesocricetus auratus (Golden hamster)	25
Q5R5A3	ANT3_PONAB	**Antithrombin-III**	**SERPINC1** AT3	Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii)	464
O97529	ANXA8_RABIT	**Annexin A8**	**ANXA8**	Oryctolagus cuniculus (Rabbit)	327
P48036	ANXA5_MOUSE	**Annexin A5**	**Anxa5** Anx5	Mus musculus (Mouse)	319
A5A6L7	ANXA8_PANTR	**Annexin A8**	**ANXA8**	Pan troglodytes (Chimpanzee)	327
P38977	ANTA_HYDVU	**Antistasin**		Hydra vulgaris (Hydra) (Hydra attenuata)	220
P14668	ANXA5_RAT	**Annexin A5**	**Anxa5** Anx5	Rattus norvegicus (Rat)	319
P01008	ANT3_HUMAN	**Antithrombin-III**	**SERPINC1** AT3, PRO0309	Homo sapiens (Human)	464
Q95L54	ANXA8_BOVIN	**Annexin A8**	**ANXA8**	Bos taurus (Bovine)	327
P13928	ANXA8_HUMAN	**Annexin A8**	**ANXA8** ANX8	Homo sapiens (Human)	327
Q5R1W0	ANXA5_PANTR	**Annexin A5**	**ANXA5**	Pan troglodytes (Chimpanzee)	320
Q03352	ANT3_CHICK	**Antithrombin-III**	**SERPINC1** AT3	Gallus gallus (Chicken)	105
P32261	ANT3_MOUSE	**Antithrombin-III**	**Serpinc1** At3	Mus musculus (Mouse)	465
Q4R4H7	ANXA5_MACFA	**Annexin A5**	**ANXA5** QnpA-14191	Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey)	320
O35640	ANXA8_MOUSE	**Annexin A8**	**Anxa8** Anx8	Mus musculus (Mouse)	327
Q76936	ATS13_MOUSE	**A disintegrin and metalloproteinase...**	**Adamts13** Gm710	Mus musculus (Mouse)	1,426

1 to 25 of 513 | Show 25

Use of keywords in UniprotKB/Swiss-Prot

Link to Gene Ontology DB for further analysis and information retrieval

Reference proteomes

- Are created for model organisms, for which is a demand for extensive and comprehensive information (> 16,000 as of Dec. 2018)
- They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny.
- Example of model organisms: *E.coli*, *B.subtilis*, human, mouse, fruitfly, *C.elegans*, yeast, *S.pombe*, *A.thaliana*.

https://www.uniprot.org/proteomes/

Downloads and updates

ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase/complete/

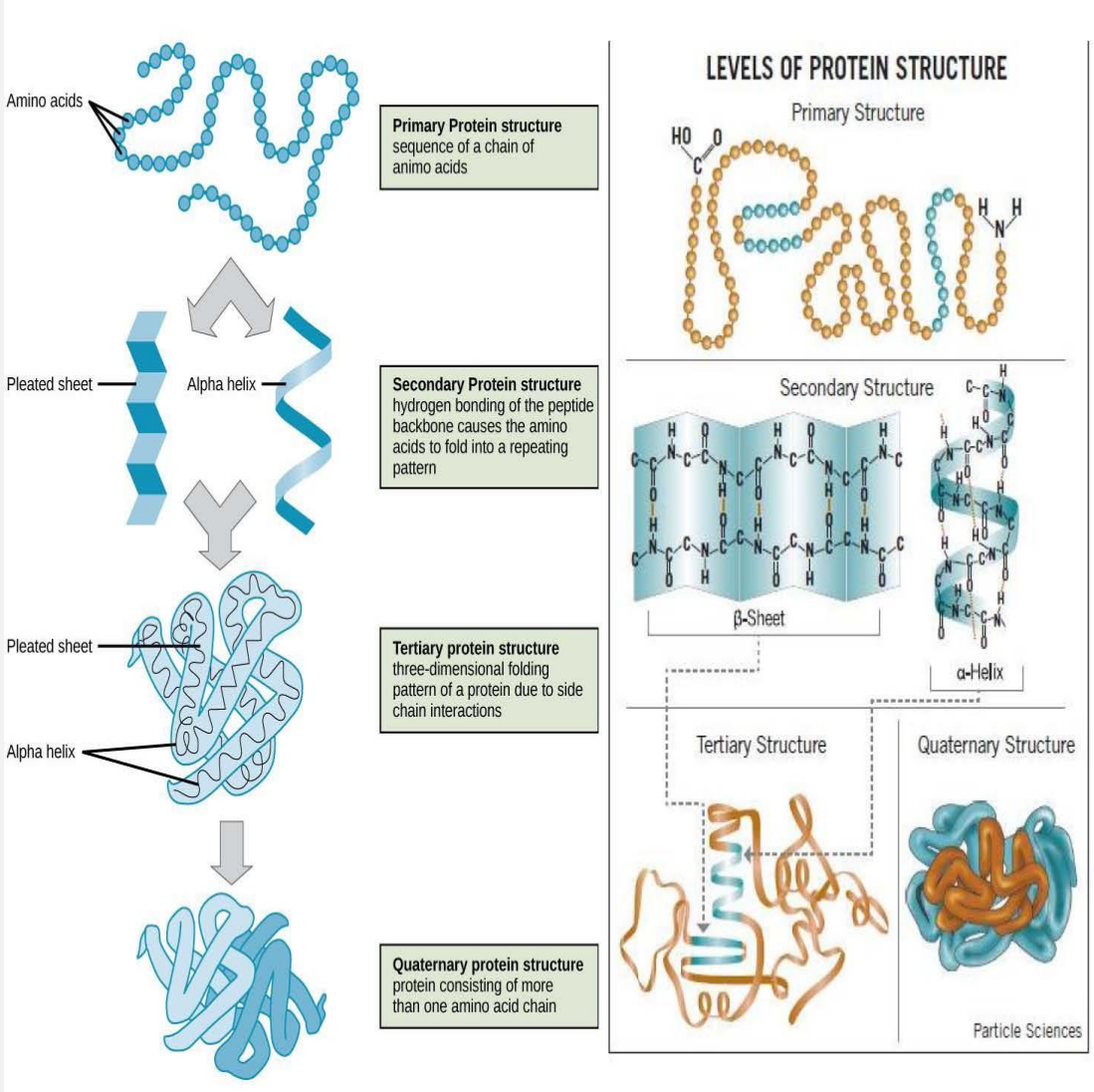
- New release every month
- Various formats (flat file, XML file, FASTA file)
- **Always** cite the Accession number, not the entry name (ID)
- Information included in the entries can be altered by Uniprot curators if they deem necessary (not possible in genomic databases, where only the submitting authors are responsible for the information
- User manual is frequently updated

Other important protein-related databases

Structural DBs

What can we learn from structures?

- Secondary structure
- Function
- Similarity, evolutionary relationships
- Shape, size
- Folding
- Structural motifs
- Distances, angles
- Surface interactions
- Effect of mutations
- Transmembrane segments



Worldwide Protein Databank (wwPDB)

- It was established in 1971 at Brookhaven National Laboratories (BNL) in USA and it was moved to the Research Collaboratory for Structural Bioinformatics (RCSB) in 1998.
- Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.
- The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

The PDB database

CRYSTALLOGRAPHY

Protein Data Bank

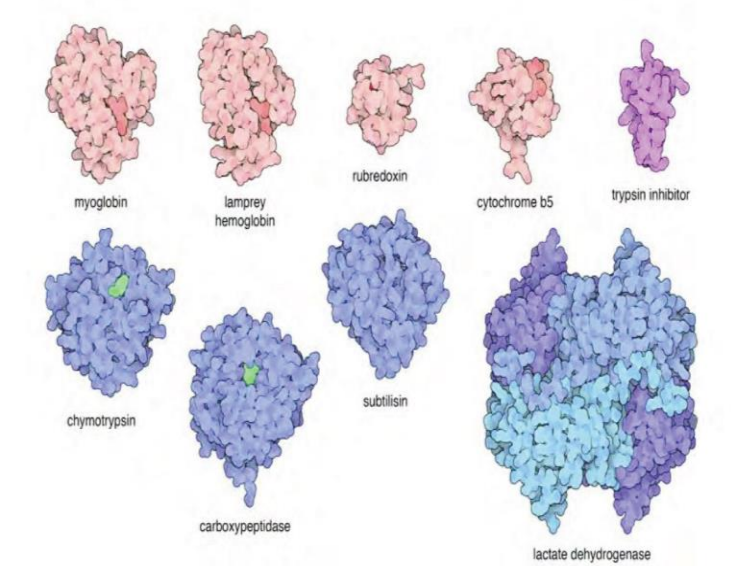
A repository system for protein crystallographic data will be operated jointly by the Crystallographic Data Centre, Cambridge, and the Brookhaven National Laboratory. The system will be responsible for storing atomic coordinates, structure factors and electron density maps and will make these data available on request. Distribution will be on magnetic tape in machine-readable form whenever possible. There will be no charge for the service other than handling costs. Files will be updated as new material is received. The total holding will be announced annually in the organic bibliographic volumes of the reference series "Molecular Structures and Dimensions" published for the Crystallographic Data Centre and

Announcing the Protein Data Bank

Nature New Biology
Vol. 233 October 20 1971

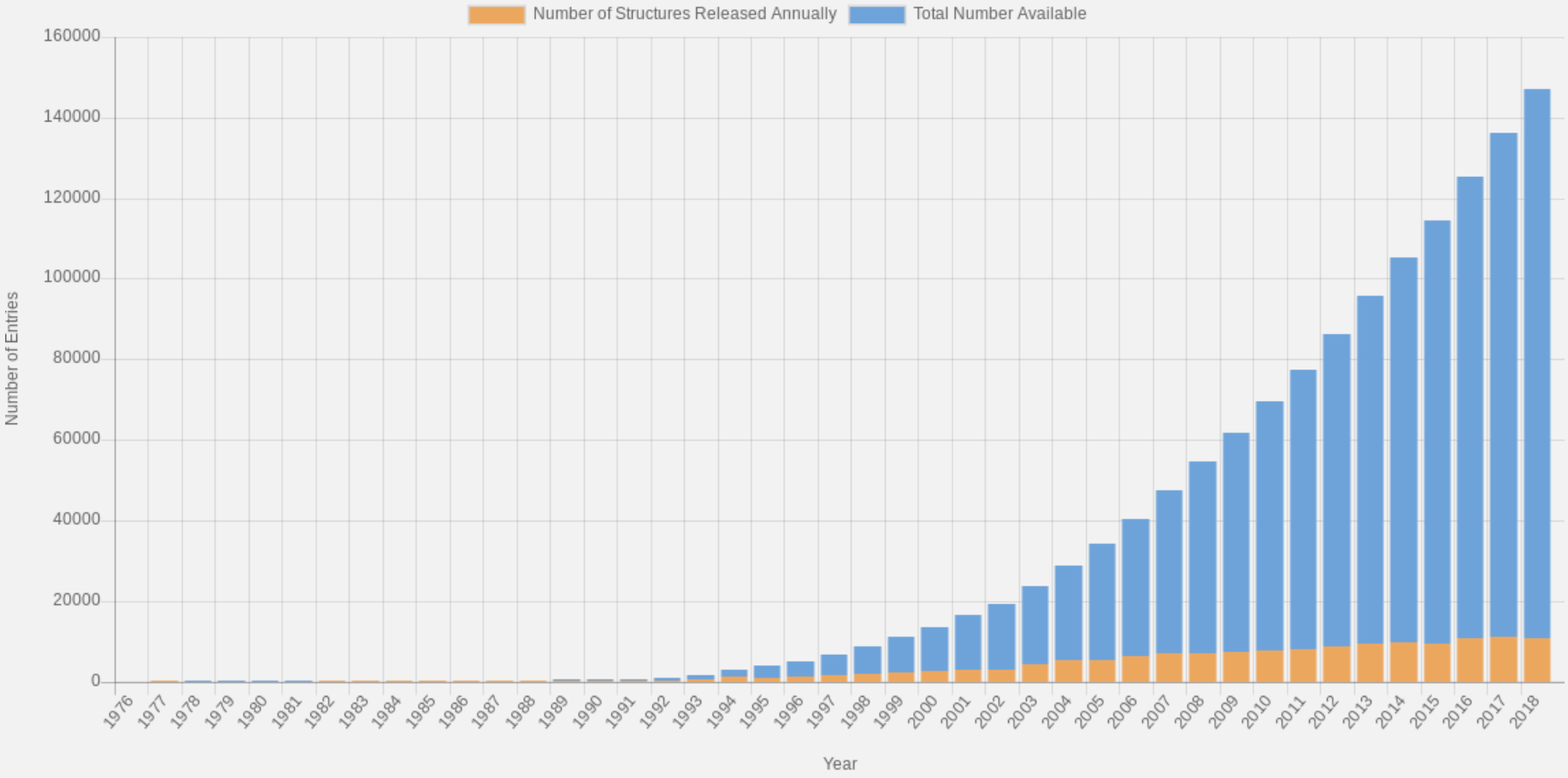
https://www.rcsb.org/

The Protein Data Bank in 1973

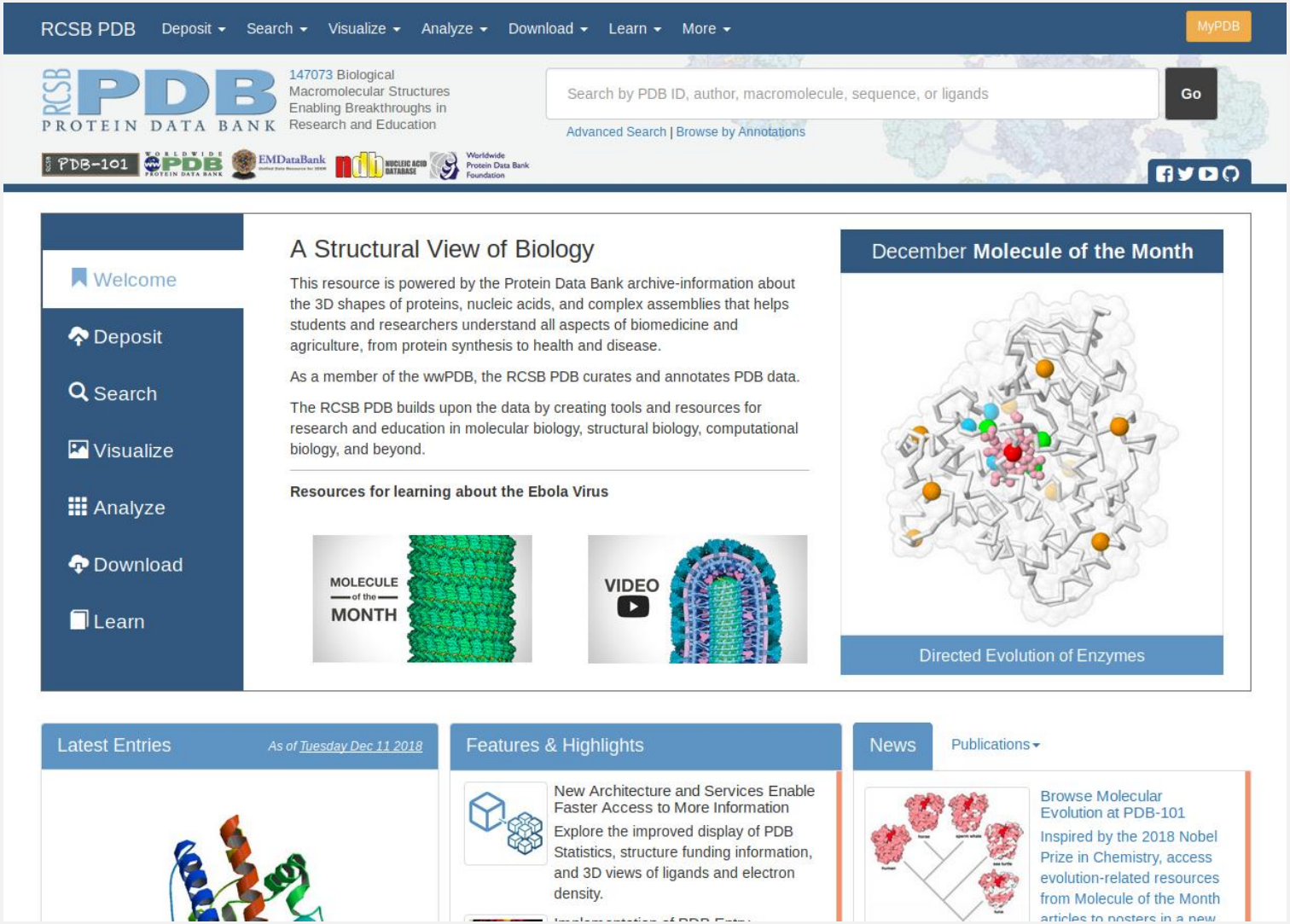


9 structures

Growth of PDB



PDB homepage



PDB browse

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | MyPDB

RCSB PDB PROTEIN DATA BANK — 147073 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands — Go

Advanced Search | Browse by Annotations

PDB-101 | PDB | EMDataBank | NDB Nucleic Acid Database | Worldwide Protein Data Bank Foundation

147,073 Structures | 57838 Citations | 26818 Ligands

Refine Search | Save Search to MyPDB

Search Parameter:

Holdings : All Structures

Refinements

ORGANISM

- Homo sapiens (42247)
- Escherichia coli (9245)
- Mus musculus (6265)
- Saccharomyces cerevisiae (4117)
- synthetic construct (3658)
- Rattus norvegicus (2976)
- Bos taurus (2838)
- Other (76541)

UNIPROT MOLECULE NAME

- Uncharacterized protein (4035)
- Gag-Pol polyprotein (1111)
- Genome polyprotein (1078)
- Beta-2-microglobulin (1048)
- Lysozyme C (1012)
- Carbonic anhydrase 2 (766)
- Endolysin (723)
- Refine Query

TAXONOMY

Currently showing 1 - 25 of 147073 Page: 1 of 5883 ← Previous Next → Displaying 25 Results

View: Detailed | Reports: Select a Report | Sort: ↓ Release Date: Newest to Oldest | Download Files

5OD1 — Download File | View File

Structure of the engineered metalloesterase MID1sc10 complexed with a phoshponate transition state analogue

Mittl, P.R.E., Studer, S.

PubMed ID is not available.

3D View

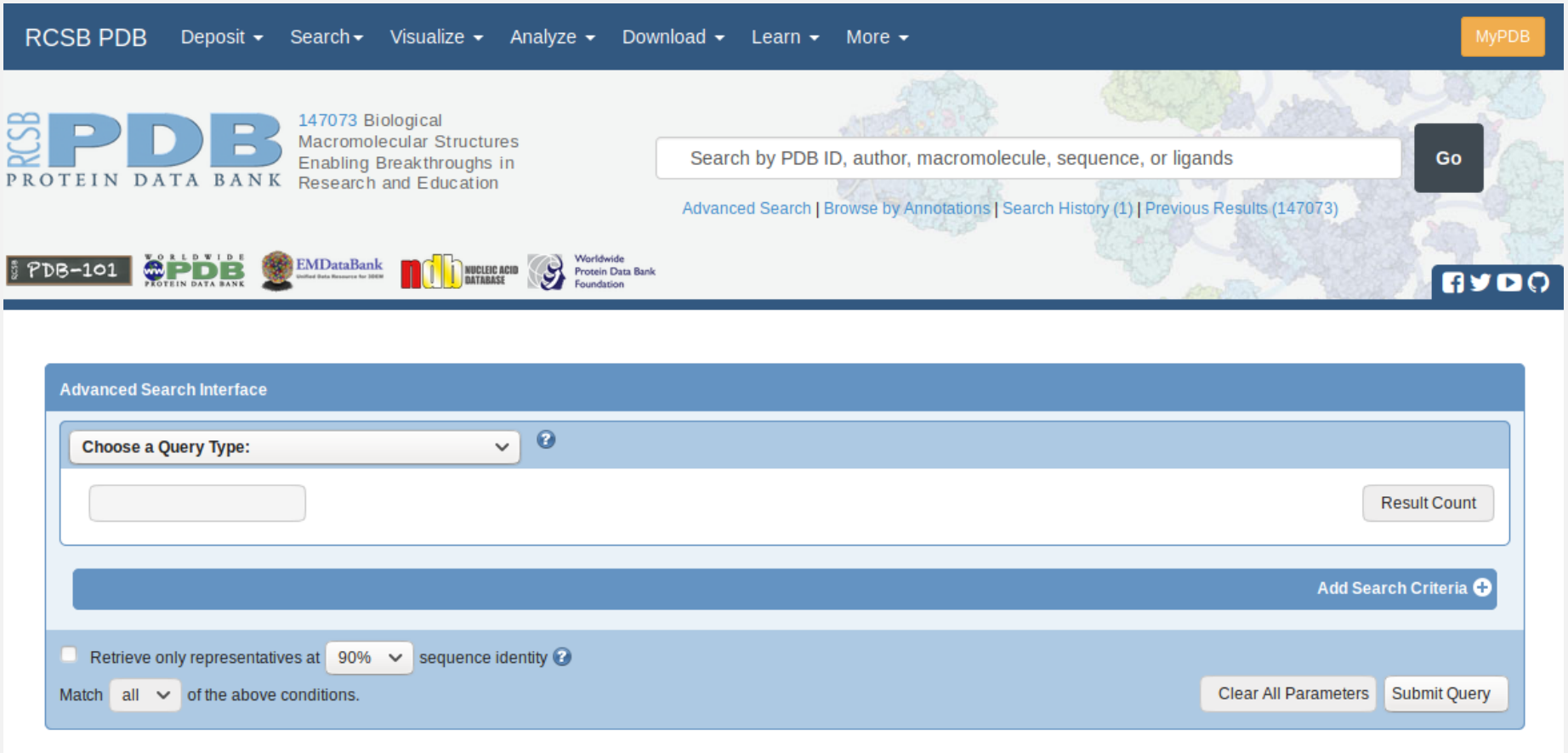
Released: 12/12/2018
Method: X-ray Diffraction
Resolution: 1.34 Å
Residue Count: 97
Macromolecule: --
Unique Ligands: 9RQ, GOL, ZN

5OD9 — Download File | View File

Structure of the engineered metalloesterase MID1sc9

Studer, S., Mittl, P.R.E., Hilvert, D.

PDB search


RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
147073 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations | Search History (1) | Previous Results (147073)
PDB-101
EMDataBank
Worldwide Protein Data Bank Foundation
Advanced Search Interface
Choose a Query Type:
Result Count
Add Search Criteria
Retrieve only representatives at 90% sequence identity
Match all of the above conditions.
Clear All Parameters
Submit Query

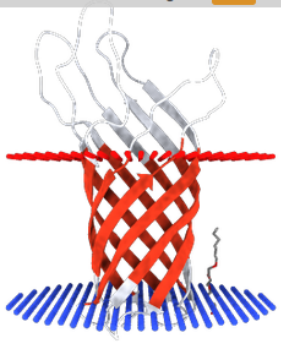
PDB entry

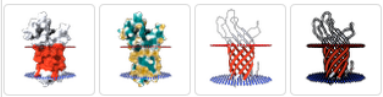
RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | MyPDB

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Transmembrane View

transmembrane regions OPM





3D View: Structure | Electron Density | Ligand Interaction

Standalone Viewers
Protein Workshop | Ligand Explorer

Macromolecule Content

- Total Structure Weight: 19199.34
- Atom Count: 1351
- Residue Count: 172
- Unique protein chains: 1

Display Files | Download Files

1BXW

OUTER MEMBRANE PROTEIN A (OMPA) TRANSMEMBRANE DOMAIN

DOI: 10.2210/pdb1BXW/pdb

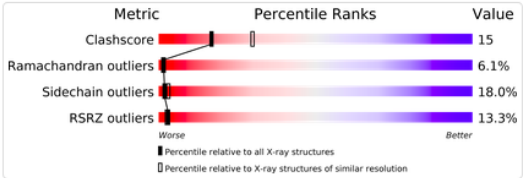
Classification: MEMBRANE PROTEIN
Organism(s): Escherichia coli (strain K12)
Expression System: Escherichia coli BL21(DE3)
Mutation(s): 3

Deposited: 1998-10-03 **Released:** 1998-10-14
Deposition Author(s): Schulz, G.E., Pautsch, A.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.5 Å
R-Value Free: 0.235
R-Value Work: 0.189

wwPDB Validation | 3D Report | Full Report



This is version 1.4 of the entry. See complete history.

Literature | Download Primary Citation

Structure of the outer membrane protein A transmembrane domain.

Pautsch, A., Schulz, G.E.
(1998) Nat.Struct.Mol.Biol. **5**: 1013-1017

PubMed: 9808047 Search on PubMed
DOI: 10.1038/2983
Also Cited By: 1QJP

PDB information

- Coordinates of atoms that make up the structure
- Literature references
- Details regarding structure determination (e.g. experimental procedures)
- Flat file with defined format
- Every structure, before being published, is checked for errors using a computer software. Subsequently, it obtains a unique code and is deposited in the database

PDB flat entry

Selected Protein Data Bank Record Types	
Record Type	
ATOM	atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in standard residues (amino acids and nucleic acids).
HETATM	atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in nonstandard residues. Nonstandard residues include inhibitors, cofactors, ions, and solvent. The only functional difference from ATOM records is that HETATM residues are by default not connected to other residues. Note that water residues should be in HETATM records.
TER	indicates the end of a chain of residues. For example, a hemoglobin molecule consists of four subunit chains which are not connected. TER indicates the end of a chain and prevents the display of a connection to the next chain.
SSBOND	defines disulfide bond linkages between cysteine residues.
HELIX	indicates the location and type (right-handed alpha, *etc.*) of helices. One record per helix.
SHEET	indicates the location, sense (anti-parallel, *etc.*) and registration with respect to the previous strand in the sheet (if any) of each strand in the model. One record per strand.

Protein Data Bank Format

Record Type	Columns	Data	Justifi-cation	Data Type
ATOM	1-4	"ATOM"	left	character
	7-11	Atom serial number	right	integer
	13-16	Atom name	left*	character
	17	Alternate location indicator		character
	18-20	Residue name	right	character
	22	Chain identifier		character
	23-26	Residue sequence number	right	integer
	27	Code for insertions of residues		character
	31-38	X orthogonal Angstrom coordinate	right	floating
	39-46	Y orthogonal Angstrom coordinate	right	floating
	47-54	Z orthogonal Angstrom coordinate	right	floating
	55-60	Occupancy	right	floating
	61-66	Temperature factor	right	floating
	73-76	Segment identifier (optional)	left	character
	77-78	Element symbol	right	character
	79-80	Charge (optional)		character
HETATM	1-6	"HETATM"		
	7-80	same as ATOM records		
TER	1-3	"TER"		character
	7-11	Serial number	right	integer
	18-20	Residue name	right	character
	22	Chain identifier		character

Blast

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | Sign in to NCBI

BLAST®

Home | Recent Results | Saved Strategies | Help

COVID-19 is an emerging, rapidly evolving situation.

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

A new version IgBLAST (1.17) is here.

We've added a new field "V frame shift" to the IgBLAST output to indicate if there is an internal frame shift in the normal V gene translation frame.

Thu, 14 Jan 2021 12:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

Protein Blast

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | Sign in to NCBI

BLAST® » blastp suite

Home | Recent Results | Saved Strategies | Help

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

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

Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

New columns added to the Description Table

Click 'Select Columns' or 'Manage Columns'.

New

Or, upload file

Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested

exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection

Algorithm

Quick BLASTP (Accelerated protein-protein BLAST)

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Protein Blast

BLAST® » blastp suite

Home | Recent Results | Saved Strategies | Help

Align Sequences Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

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

Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

New columns added to the Description Table

Click 'Select Columns' or 'Manage Columns'.

New

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Subject subrange

From

To

Or, upload file: Choose File No file chosen

Program Selection

Algorithm

◉ blastp (protein-protein BLAST)

Choose a BLAST algorithm

Clustal

Clustal Omega

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback | Share

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

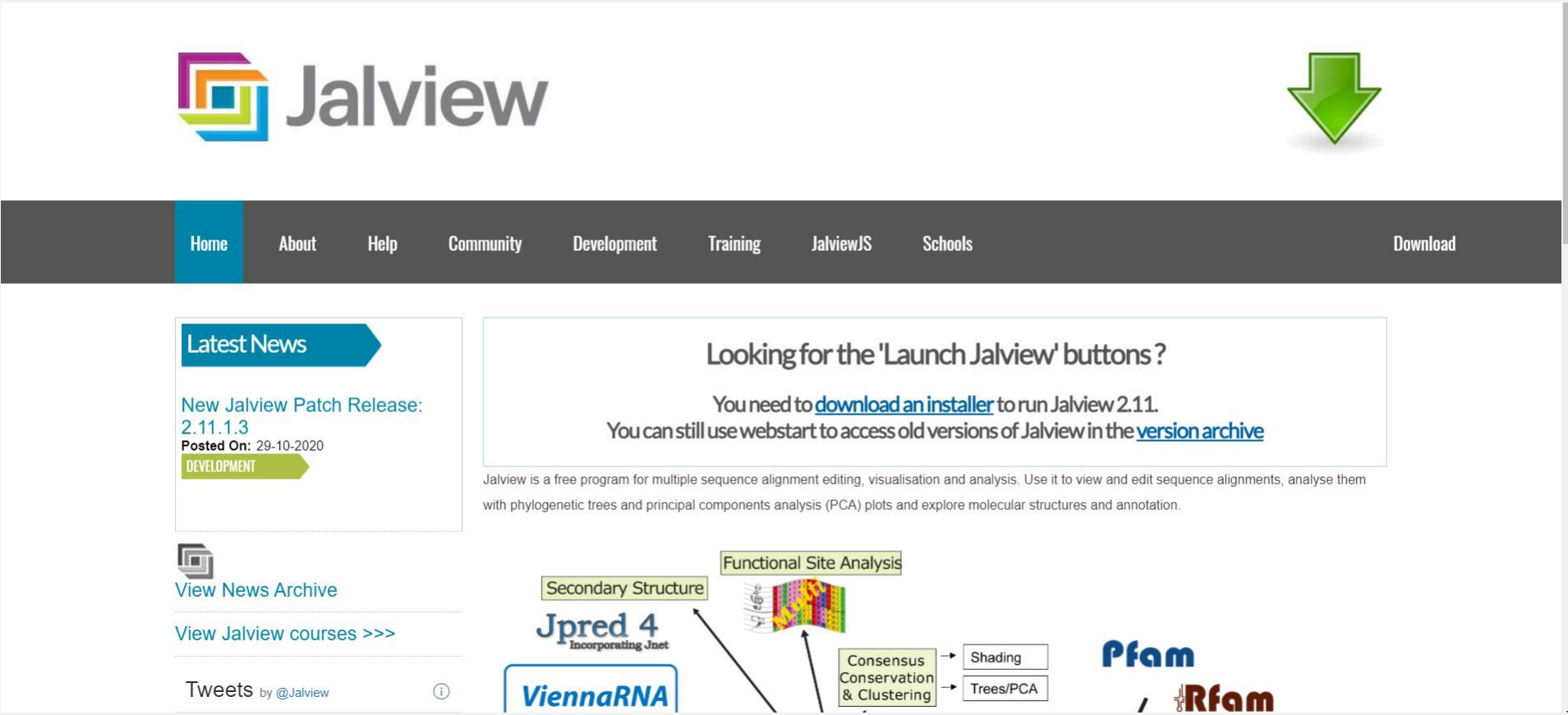
PROTEIN

sequences in any supported format:

Or, upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs

Jalview


Jalview
Home
About
Help
Community
Development
Training
JalviewJS
Schools
Download
Latest News
New Jalview Patch Release: 2.11.1.3
Posted On: 29-10-2020
DEVELOPMENT
View News Archive
View Jalview courses >>>
Tweets by @Jalview
Looking for the 'Launch Jalview' buttons?
You need to download an installer to run Jalview 2.11.
You can still use webstart to access old versions of Jalview in the version archive
Jalview is a free program for multiple sequence alignment editing, visualisation and analysis. Use it to view and edit sequence alignments, analyse them with phylogenetic trees and principal components analysis (PCA) plots and explore molecular structures and annotation.
Functional Site Analysis
Secondary Structure
Jpred 4
Incorporating Jnet
ViennaRNA
Consensus Conservation & Clustering
Shading
Trees/PCA
Pfam
Rfam