



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Protein domains

Miguel Andrade

Faculty of Biology,

Institute of Organismic Molecular Evolution,

Johannes Gutenberg University

Mainz, Germany

andrade@uni-mainz.de

Introduction

Protein domains are structural units (average 160 aa) that share:

Function

Folding

Evolution

Proteins normally are multidomain (average 300 aa)

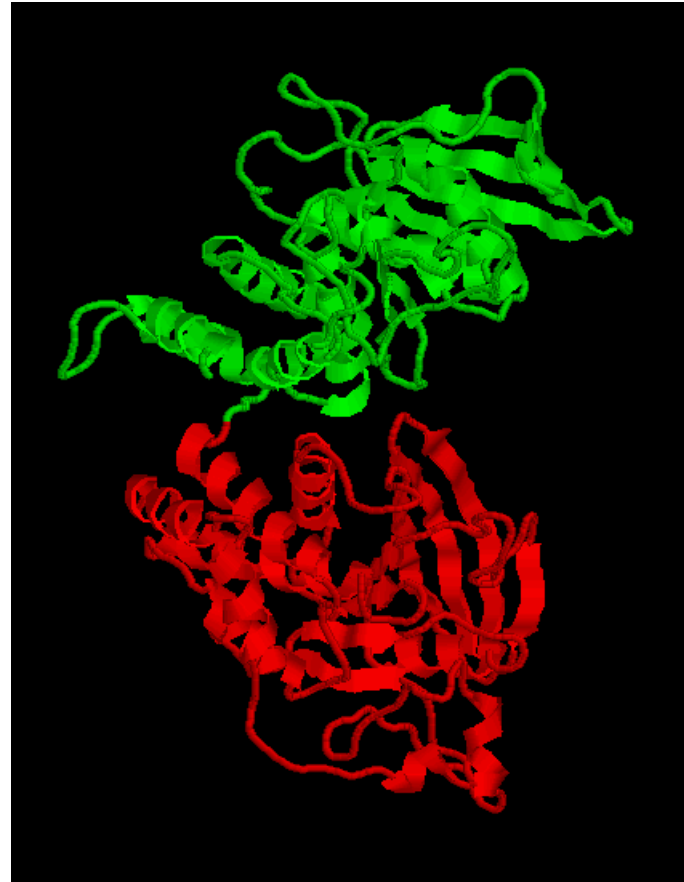


Introduction

Protein domains are structural units (average 160 aa) that share:

Function
Folding
Evolution

Proteins normally are multidomain (average 300 aa)



Domains

Why to search for domains:

Protein structural determination methods such as X-ray crystallography and NMR have size limitations that limit their use.

Multiple sequence alignment at the domain level can result in the detection of homologous sequences that are more difficult to detect using a complete chain sequence.

Methods used to gain an insight into the structure and function of a protein work best at the domain level.

Domain databases

SMART

Peer Bork

<http://smart.embl.de/>

Manual definition of domain (bibliography)

Generate profile from instances of domain

Search for remote homologs (HMMer)

Include them in profile


Iterate until convergence

Schultz et al (1998) *PNAS*

...

Letunic et al (2014) *Nucleic Acids Research*

Domain databases



Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* 95, 5857-5864
Letunic et al. (2012) *Nucleic Acids Res* , doi:10.1093/nar/gkr931

HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK


SMART MODE:
NORMAL
GENOMIC

Simple
Modular
Architecture
Research
Tool


Sequence analysis

You may use either a [Uniprot/Ensembl](#) sequence identifier (ID) / accession number (ACC) or the protein sequence itself to perform the SMART analysis service.

Sequence ID or ACC

Examples: #1, #2 

Protein sequence

Examples: #1, #2 


HMMER searches of the SMART database occur by default. You may also find:

[Outlier homologues](#) and homologues of known structure


Architecture analysis

You can search for proteins with combinations of [specific domains](#) in different species or taxonomic ranges. You can input the domains directly into "Domain selection" box, or use "GO terms query" to get a list of domains.


Domain selection

Examples: #1, #2 

GO terms query

Examples: #1, #2 

Taxonomic selection

Select a taxonomic range via the selection box or type it into the text box below: 

All
Examples: #1, #2

You can try an [Advanced Query](#) if you're familiar with SQL.

Domain databases

SMART

Domains detected by SMART

SH3

Src homology 3 domains



SMART
accession
number:

SM00326

Description:

Src homology 3 (SH3) domains bind to target proteins through sequences containing proline and hydrophobic amino acids. Pro-containing polypeptides may bind to SH3 domains in 2 different binding orientations.

Interpro
abstract
([IPR001452](#)):

SH3 (src Homology-3) domains are small protein modules containing approximately 50 amino acid residues [([PUBMED:15335710](#)), ([PUBMED:11256992](#))]. They are found in a great variety of intracellular or membrane-associated proteins [([PUBMED:1639195](#)), ([PUBMED:14731533](#)), ([PUBMED:7531822](#))] for example, in a variety of proteins with enzymatic activity, in adaptor proteins, such as fodrin and yeast actin binding protein ABP-1.

The SH3 domain has a characteristic fold which consists of five or six beta-strands arranged as two tightly packed anti-parallel beta sheets. The linker regions may contain short helices. The surface of the SH3-domain bears a flat, hydrophobic ligand-binding pocket which consists of three shallow grooves defined by conservative aromatic residues in which the ligand adopts an extended left-handed helical arrangement. The ligand binds with low affinity but this may be enhanced by multiple interactions. The region bound by the SH3 domain is in all cases proline-rich and contains PXXP as a core-conserved binding motif. The function of the SH3 domain is not well understood but they may mediate many diverse processes such as increasing local concentration of proteins, altering their subcellular location and mediating the assembly of large multiprotein complexes [([PUBMED:7953536](#))].

The crystal structure of the SH3 domain of the cytoskeletal protein spectrin, and the solution structures of SH3 domains of phospholipase C (PLC-y) and phosphatidylinositol 3-kinase p85 alpha-subunit, have been determined [([PUBMED:1279434](#)), ([PUBMED:7684655](#)), ([PUBMED:7681365](#))]. In spite of relatively limited sequence similarity, their overall structures are similar. The domains belong to the alpha+beta structural class, with 5 to 8 beta-strands forming 2 tightly-packed, anti-parallel beta-sheets arranged in a barrel-like structure, and intervening loops sometimes forming helices. Conserved aliphatic and aromatic residues form a hydrophobic core (A11, L23, A29, V34, W42, L52 and V59 in PLC-y [([PUBMED:7681365](#))] and a hydrophobic pocket on the molecular surface (L12, F13, W53 and P55 in PLC-y). The conserved core is believed to stabilise the fold, while the pocket is thought to serve as a binding site for target proteins. Conserved carboxylic amino acids located in the loops, on the periphery of the pocket (D14 and E22), may be involved in protein-protein interactions via proline-rich regions. The N- and C-termini are packed in close proximity, indicating that they are independent structural modules.

GO function:

protein binding ([GO:0005515](#))

Domain databases

SMART

Sequence analysis

You may use either a [Uniprot/Ensembl](#) sequence identifier (ID) / accession number (ACC) or the protein sequence itself to perform the SMART analysis service.

Sequence ID or ACC

Examples: #1, #2



Protein sequence

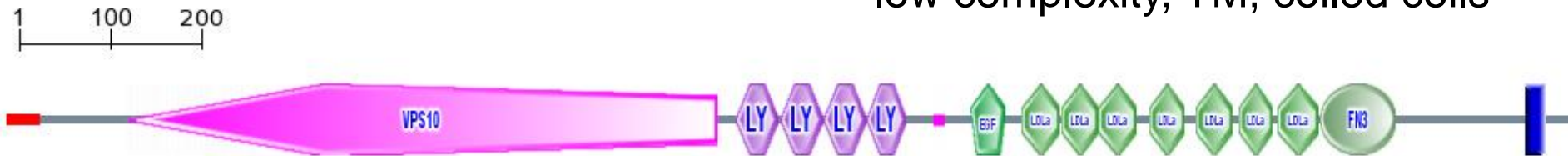
Examples: #1, #2



Domain databases

SMART

Extra features:
Signal-peptide,
low complexity, TM, coiled coils



Confidently predicted domains, repeats, motifs and features:

Name	Begin	End	E-value
signal peptide	1	36	-
VPS10	125	741	0.00e+00
LY	761	806	2.88e+00
LY	807	851	3.94e-04
LY	852	896	5.31e-10
LY	897	939	1.76e-15
low complexity	968	979	-
EGF	1006	1042	1.87e+01
LDLa	1059	1098	2.69e-10
LDLa	1100	1138	1.62e-13
EGF_like	1138	1177	5.24e+01
LDLa	1139	1178	1.46e-11
LDLa	1193	1230	2.07e-11
LDLa	1240	1278	2.91e-06
LDLa	1286	1321	3.21e-08
LDLa	1326	1369	1.27e-06
FN3	1370	1448	1.36e-03
transmembrane	1584	1606	-

Additional information

[Display](#) other IDs, orthology and alternative splicing data for this sequence.

Domain architecture analysis

This domain architecture was probably invented with the emergence of [Hydra viridis](#).

[Display](#) all proteins with similar domain [organisation](#).

[Display](#) all proteins with similar domain [composition](#).

Domain databases

PFAM (until Jan 2023)

Erik Sonnhammer/Ewan Birney/Alex Bateman

<http://pfam.xfam.org/>



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#)
| [ABOUT](#)

Pfam 35.0 (November 2021, 19632 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW A PFAM ENTRY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

Analyze your protein sequence for Pfam matches

View Pfam annotation and alignments

See groups of related entries

Look at the domain organisation of a protein sequence

Find the domains on a PDB structure

Sonnhammer et al (1997) *Proteins*

...

Mistry et al (2021) *Nucleic Acids Research*

Domain databases

PFAM

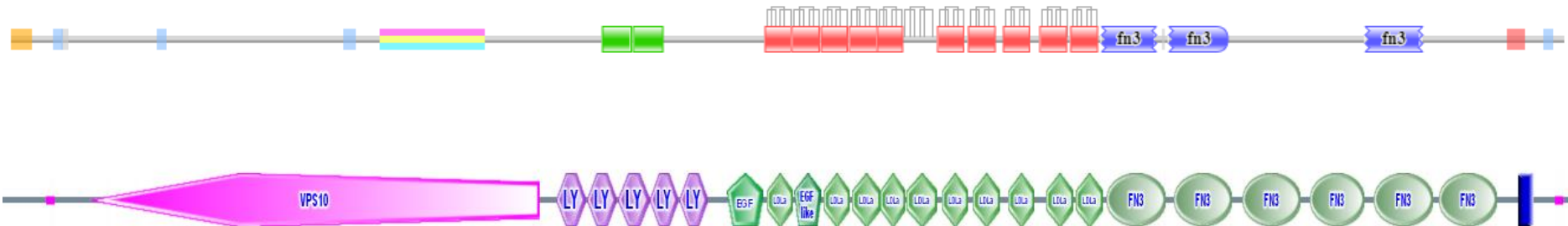
This is the summary of UniProt entry [SORL_HUMAN](#) (Q92673).

Description:	Sortilin-related receptor
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	2214 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the next Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains. [More...](#)

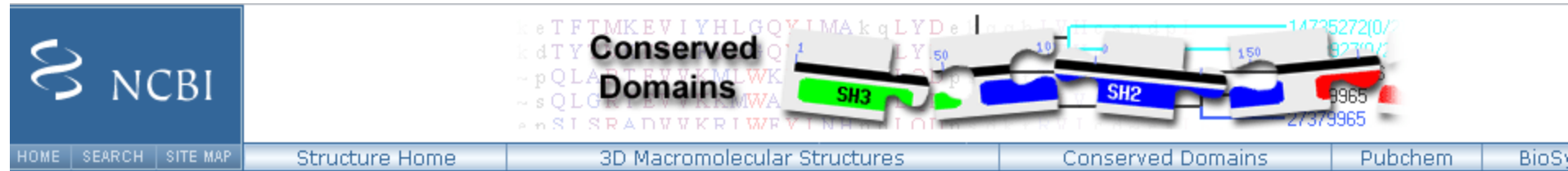


Domain databases

CDD

Stephen Bryant

<http://www.ncbi.nlm.nih.gov/cdd>



NCBI

HOME SEARCH SITE MAP

Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioS

Search for Conserved Domains within a protein sequence

Enter **Protein** Query as Accession, Gi, or Sequence in [FASTA format](#) [?](#)

Submit

Reset

OPTIONS

Search against database [?](#): CDD -- 34177 PSSMs

Expect Value [?](#) threshold: 0.01

Apply low-complexity filter [?](#)

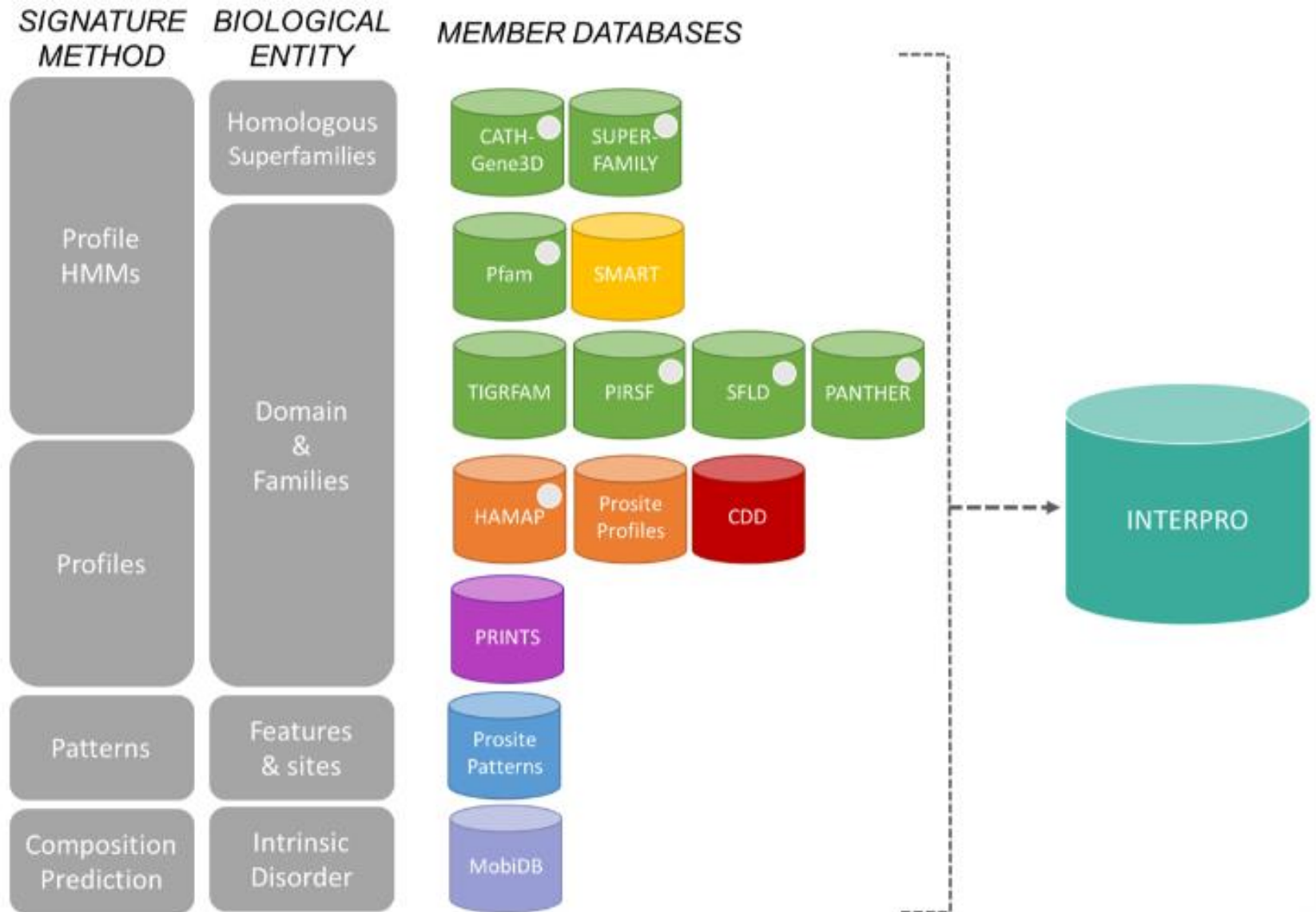
Force live search [?](#)

Maximum number of hits [?](#) 250

Result mode Concise [?](#) Full [?](#)

Marchler-Bauer et al (2015) *Nucleic Acids Res*

InterPro

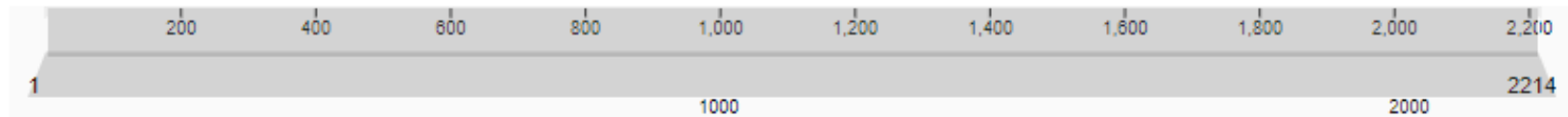


InterPro

SORLA/SORL1 from *Homo sapiens*

<https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/>

Entry matches to this proteinⁱ



▼ AlphaFold Confidence

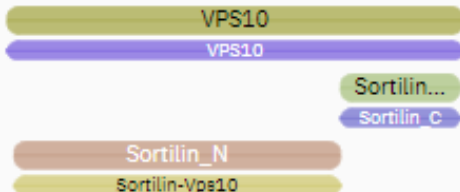


pLDDT

▼ Representative Domains



▼ Domain



D IPR003961
SM00060
P550853
PF00041
cd00063

D IPR006581
SM00602

D IPR031777
PF15901

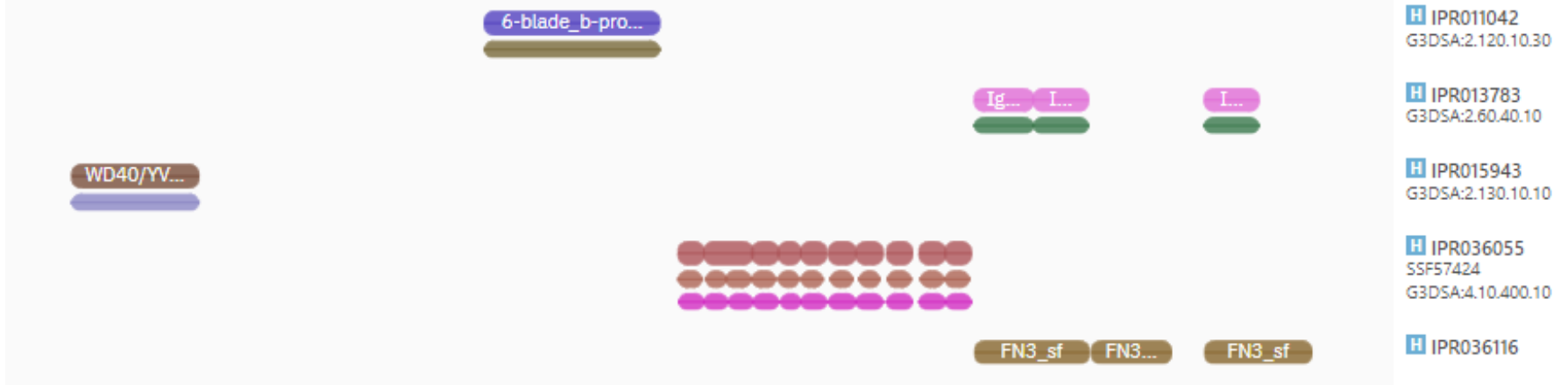
D IPR031778
PF15902

InterPro

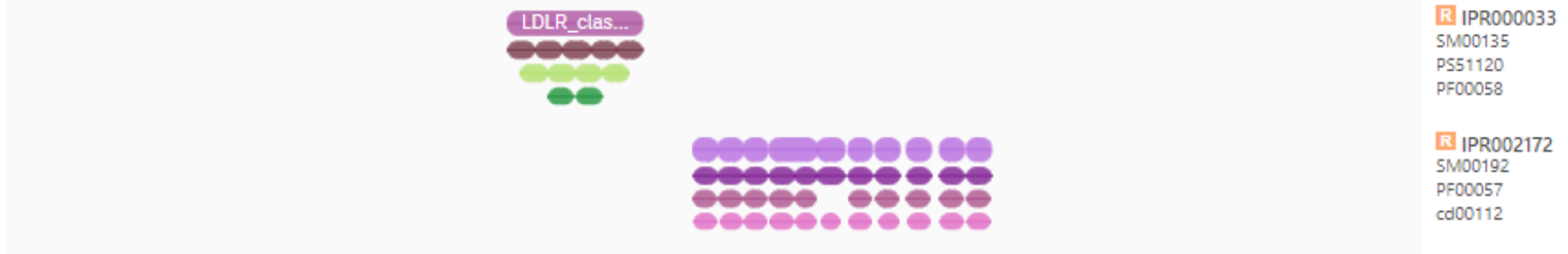
SORLA/SORL1 from *Homo sapiens*

<https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/>

▼ Homologous Superfamily



▼ Repeat

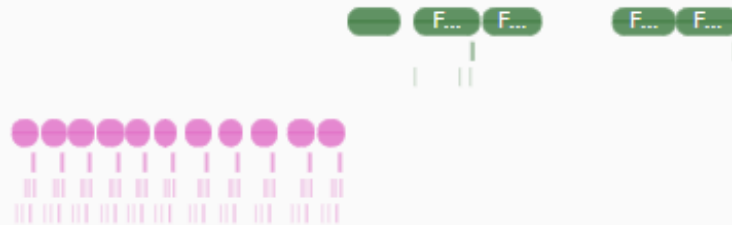


InterPro

SORLA/SORL1 from *Homo sapiens*

<https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/>

▼ Residues



cd00063
Cytokine receptor motif
Interdomain contacts

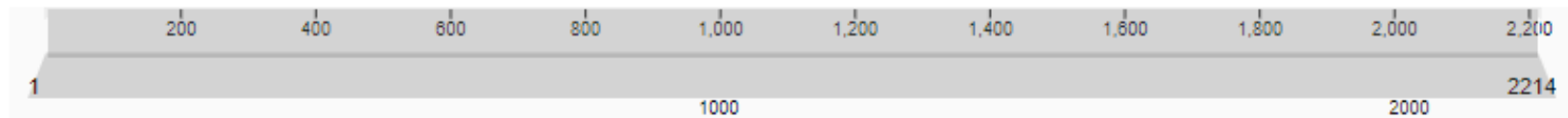
cd00112
D-X-S-D-E motif
Calcium-binding site
Putative binding surface

InterPro

SORLA/SORL1 from *Homo sapiens*

<https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/>

Entry matches to this proteinⁱ

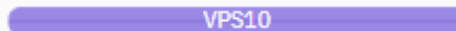


AlphaFold Confidence



pLDDT

Representative Domains



VPS10

Fibronectin type III domain

Pfam domain

1557 - 1629



Domain

1380	CIPNRWKCDR	ENDCGDWSDE	KDCGDHILP	FSTPGPSTCL	PNYYRCSSGT	CVMDTWVCDG	
1440	YRDCADGSDE	EACPLLAVT	AASTPTQLGR	CDRFEFECHQ	PKTCIPNWKR	CDGHQDCQDG	
1500	RDEANCPHS	TLTCMSREFQ	CEDEACIVL	SERCDGFLDC	SDESDEKACS	DELTVYKVQN	
1560	LQWTADFSGD	VTLTWMPKK	MPSASCYVNV	YYRVVGESIW	KTLETHSNKI	NTVLKVLKPD	
Sortilin	1620	TTYQVKVQVQ	CLSKAHNTND	FVTLRTP EGL	PDAPRNQLS	LPREAEGVIV	GHWAPPIHTH
Sortilin-V	1680	GLIREYIVEY	SRSGSKMWAS	QRAASNFT EI	KNLLVNTLYT	VRVAAVTSRG	IGNWSDSKSI

D IPR003961
SM00060
P550853
PF00041
cd00063

D IPR006581
SM00602

D IPR031777
PF15901

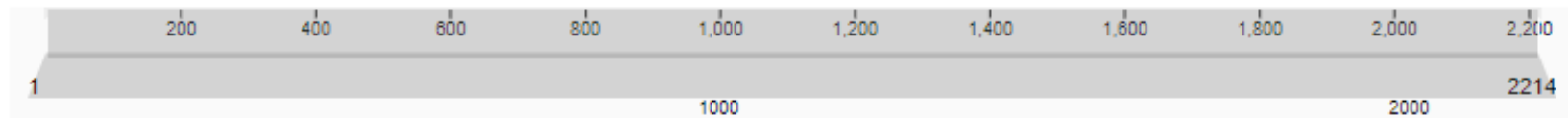
D IPR031778
PF15902

InterPro

SORLA/SORL1 from *Homo sapiens*

<https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/>

Entry matches to this proteinⁱ



▼ AlphaFold Confidence



▼ Representative Domains



▼ Domain



- IPR003961 SM00060
- PSS0853 PF00041
- IPR006581 SM00602
- IPR031777 PF15901
- IPR031778 PF15902

InterPro

[Home](#) / [Browse](#) / [By Entry](#) / [Pfam](#) / [PF00041](#) / [Overview](#)

Pfam

PF00041

Fibronectin type III domain

Pfam entry ⓘ

[Add your annotation](#) ▼

Integrated to

[> IPR003961](#)

Overview

Proteins 260k

Domain Architectures 20k

Taxonomy 22k

Proteomes 5k

Structures 324

Signature

AlphaFold 123k

Alignment

Curation

Member database [Pfam](#) ⓘ

Pfam type domain

Short name *fn3*

Set [E-set](#)

Description ⓘ Imported from [IPR003961](#)

Fibronectin is a dimeric glycoprotein composed of disulfide-linked subunits with a molecular weight of 220-250kDa each. It is involved in cell adhesion, cell morphology, thrombosis, cell migration, and embryonic differentiation. Fibronectin is a modular protein composed of homologous repeats of three prototypical types of domains known as types I, II, and III ^[4].

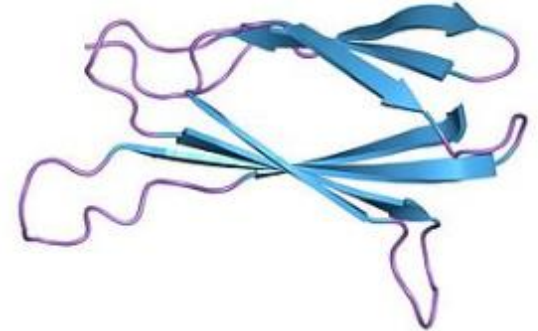
Fibronectin type-III (FN3) repeats are both the largest and the most common of the fibronectin subdomains. Domains homologous to FN3 repeats have been found in various animal protein families including other extracellular-matrix molecules, cell-surface receptors, enzymes, and muscle proteins ^[2]. Structures of individual FN3 domains have revealed a conserved β -sandwich fold with one β -sheet containing four strands and the other sheet containing three strands (see for example [1TEN](#)) ^[1]. This fold is topologically very similar to that of Ig-like domains, with a notable difference being the lack of a conserved disulfide bond in FN3 domains. Distinctive hydrophobic core packing and the lack of detectable sequence homology between immunoglobulin and FN3 domains suggest, however, that these domains are not evolutionarily related ^[1].

InterPro

Fibronectin type III domain [Wikipedia](#)

The **Fibronectin type III domain** is an evolutionarily conserved protein domain that is widely found in animal proteins. The fibronectin protein in which this domain was first identified contains 16 copies of this domain. The domain is about 100 amino acids long and possesses a beta sandwich structure. Of the three fibronectin-type domains, type III is the only one without disulfide bonding present. Fibronectin domains are found in a wide variety of extracellular proteins. They are widely distributed in animal species, but also found sporadically in yeast, plant and bacterial proteins.

Fibronectin type III domain



The tenth type III domain of fibronectin

Identifiers

Symbol	fn3
Pfam	PF00041
Pfam_clan	CL0159
InterPro	IPR003961
SMART	FN3
PROSITE	PDOC00214

InterPro

Domain Architectures 20k

Taxonomy 22k

Proteomes 5k

Structures 324

Signature

AlphaFold 123k

Alignment

Curation

i The number of species for this sunburst is 13055. The depth of the visualisation has been limited. You can modify this with the controller in the right side. however, please note this might affect the performance in your browser.



Legends

- bacteria
- viruses
- archaea
- eukaryota
- Other

Weight Segments by

Number of sequences

Font Size

14

Sunburst Depth

6 rings

2 8

Selected Taxon

Name

Chordata

Number of sequences

178358

Number of species

1738

Lineage

root; Eukaryota; Metazoa; Chordata;

InterPro

Pfam

PF00041

Fibronectin type III domain

Pfam entry ⓘ



This entry matches these structures:

Overview

Proteins 266k

Domain Architectures 21k

Taxonomy 23k

Proteomes 5k

Structures 463

Signature

AlphaFold 125k

Alignment

Curation

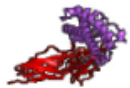
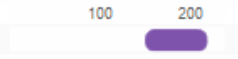
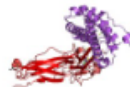


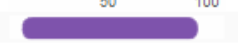

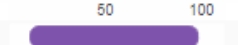
1 - 20 of 450 structures



Search

Export



ACCESSION	NAME	SOURCE DATABASE	STRUCTURE	MATCHES
1a22	HUMAN GROWTH HORMONE BOUND TO SINGLE RECEPTOR	PDB		B 
1axi	STRUCTURAL PLASTICITY AT THE HGH:HGHBP INTERFACE	PDB		B 
1bj8	THIRD N-TERMINAL DOMAIN OF GP130, NMR, MINIMIZED AVERAGE STRUCTURE	PDB		A 
1bpv	TITIN MODULE A71 FROM HUMAN CARDIAC MUSCLE, NMR, 50 STRUCTURES	PDB		A 

Exercise 1

Find structures in the PDB for human myosin X

The corresponding UniProt page is

<https://www.ebi.ac.uk/interpro/protein/reviewed/Q9HD67/>

Q9HD67 Unconventional myosin-X

UniProtKB/Swiss-Prot protein ⓘ

Overview

Entries 18

Structures 7

Sequence

Similar Proteins 90

AlphaFold 1

Short name *MYO10_HUMAN*

Length 2058 amino acids

Species *Homo sapiens* (Human)

Proteome UP000005640

Function ⓘ

Myosins are actin-based motor molecules with ATPase activity. Unconventional myosins serve in intracellular movements. MYO10 binds to actin filaments and actin bundles and functions as a plus end-directed motor. Moves with higher velocity and takes l...

Show More ▾

Exercise 1

Find structures in the PDB for human myosin X

Tip: The details of the structures are at the bottom of the page. You have to slide down.


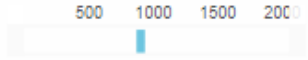

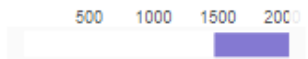







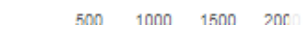
- Which domains of myosin X are covered by the solved structures?
- Is there a part of the protein for which there are no known structures? Does it have predicted domains?

Exercise 2

Analyse domain predictions

Slide down to see the details of the structures.

This protein matches these structures:

ACCESSION	NAME	SOURCE DATABASE	STRUCTURE	MATCHES
2lw9	NMR solution structure of Myo10 anti-CC	PDB		
3au4	Structure of the human myosin-X MyTH4-FERM cassette bound to its specific cargo, DCC	PDB		
3au5	Structure of the human myosin-X MyTH4-FERM cassette	PDB		
3pzd	Structure of the myosin X MyTH4-FERM/DCC complex	PDB		
5i0h	Crystal structure of myosin X motor domain in pre-powerstroke state	PDB		
	Crystal structure of myosin X motor domain with DCC motif in			

Exercise 2

Analyse domain predictions

- Examine the structure of 3pzd

How do the domain predictions fit the structure?

- Chain B in this structure is a small peptide.

Which domain in Myosin X is interacting with this peptide?

Exercise 3

AlphaFold prediction

- There is a predicted structure

Q9HD67 Unconventional myosin-X

UniProtKB/Swiss-Prot protein ⓘ

Overview	⏪
Entries	18
Structures	7
Sequence	
Similar Proteins	90
AlphaFold	1

Short name *MYO10_HUMAN*

Length 2058 amino acids

Species *Homo sapiens* (Human)

Proteome UP000005640

Function ⓘ
Myosins are actin-based motor molecules with ATPase activity. Unconventional myosins serve in intracellular movements. MYO10 binds to actin filaments and actin bundles and functions as a plus end-directed motor. Moves with higher velocity and takes l...

Show More ▾

Exercise 3

AlphaFold prediction

- There is a predicted structure
- Download the PDB file and load it in Chimera
- Select the central region without PDB information (Select/Atom specifier), 934-1485, inverse the selection, and delete everything else (Actions/Atoms/Delete).

Describe the structure predicted for this region and how this could affect structure determination.

- Examine the PH domains. How many domains do you see? Is there anything particular about them?

Exercise 3

AlphaFold prediction

