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

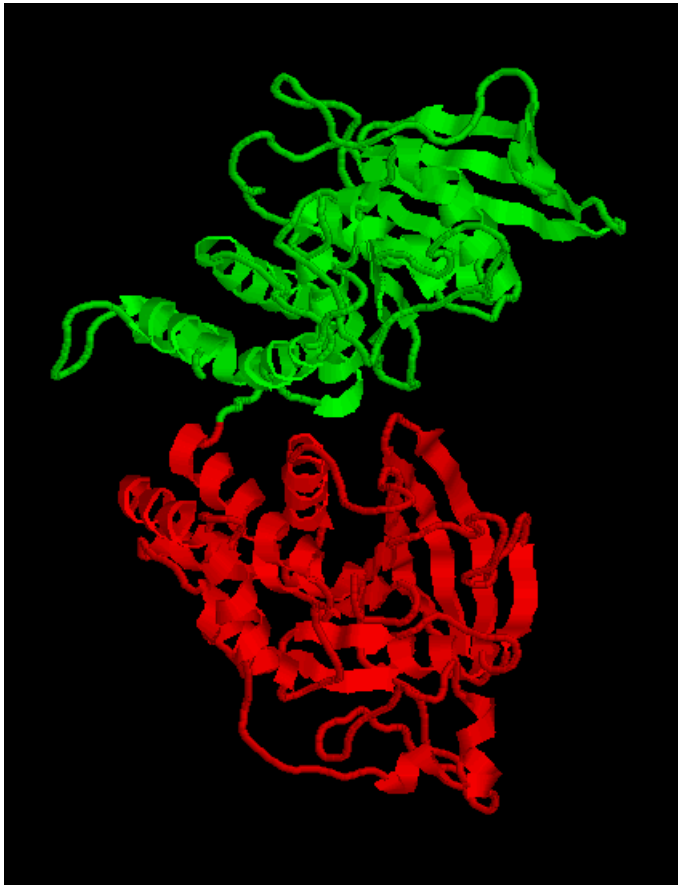
Proteins normally are multidomain (average 300 aa)

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)

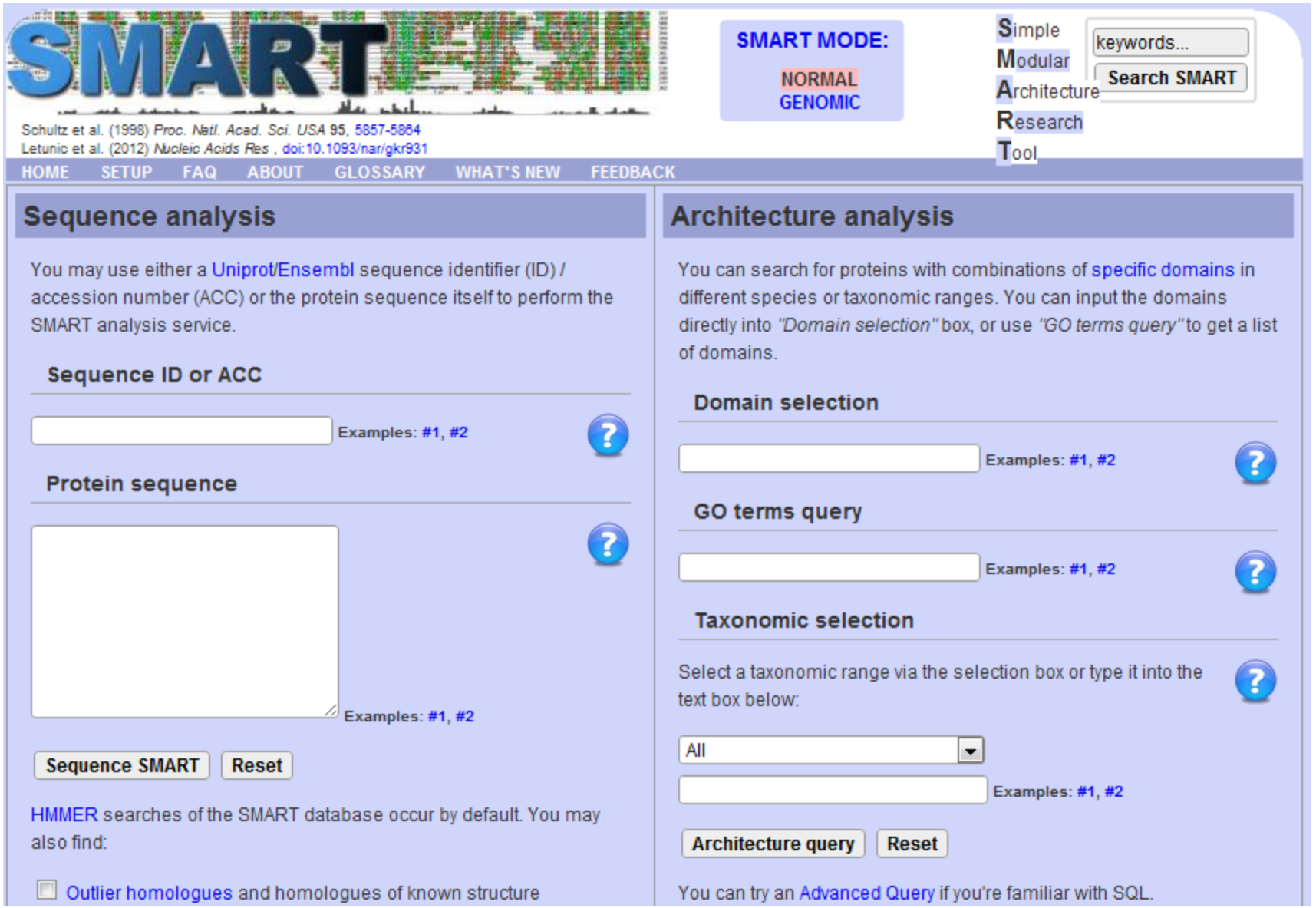
Generate profile from instances of domain
Search for remote homologs (HMMer)
Include them in profile
Iterate until convergence

Schultz et al (1998) *PNAS*

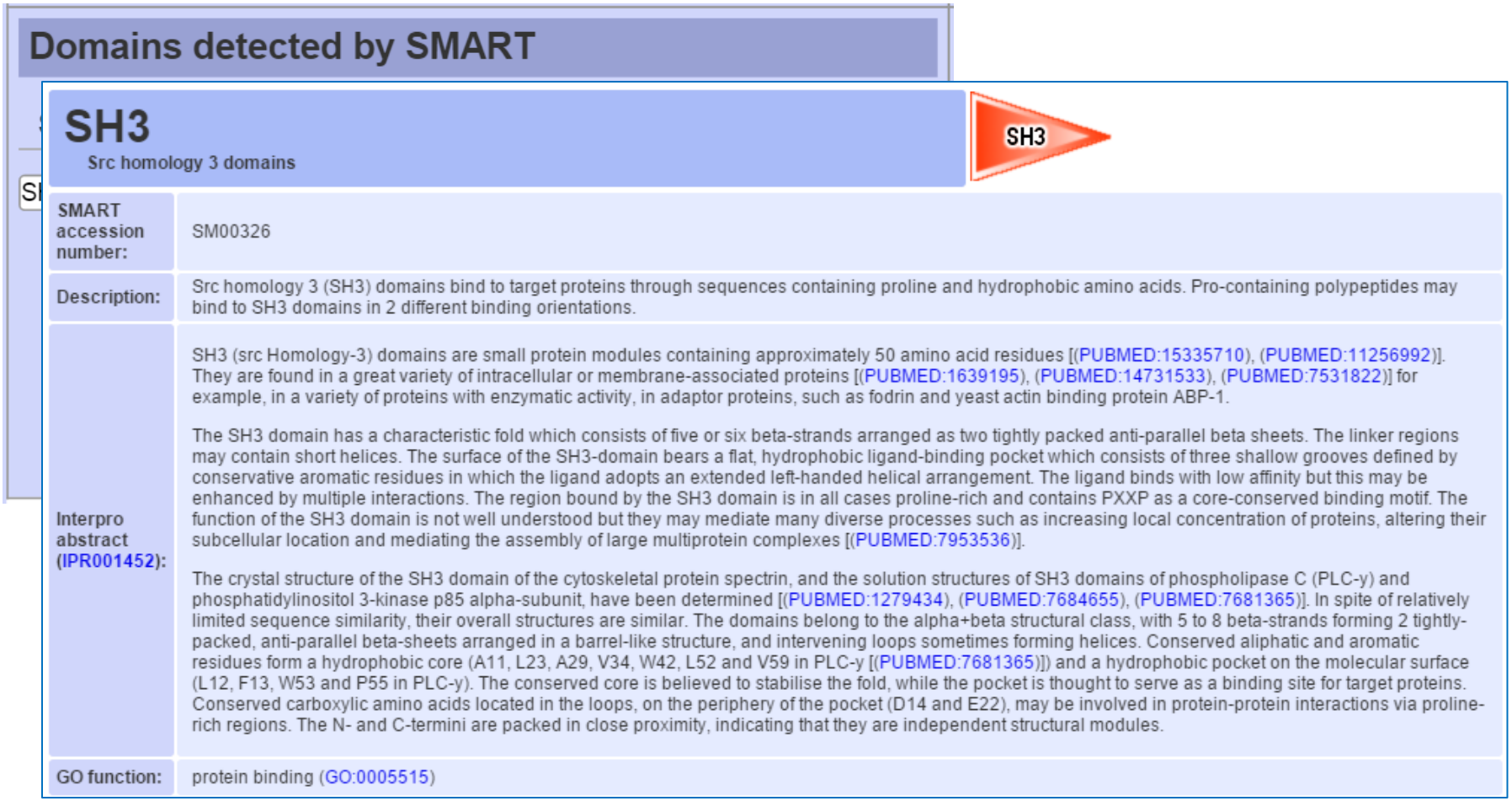
…

Letunic et al (2014) *Nucleic Acids Research*

Domain databases



Domain databases
SMART



Domains detected by SMART

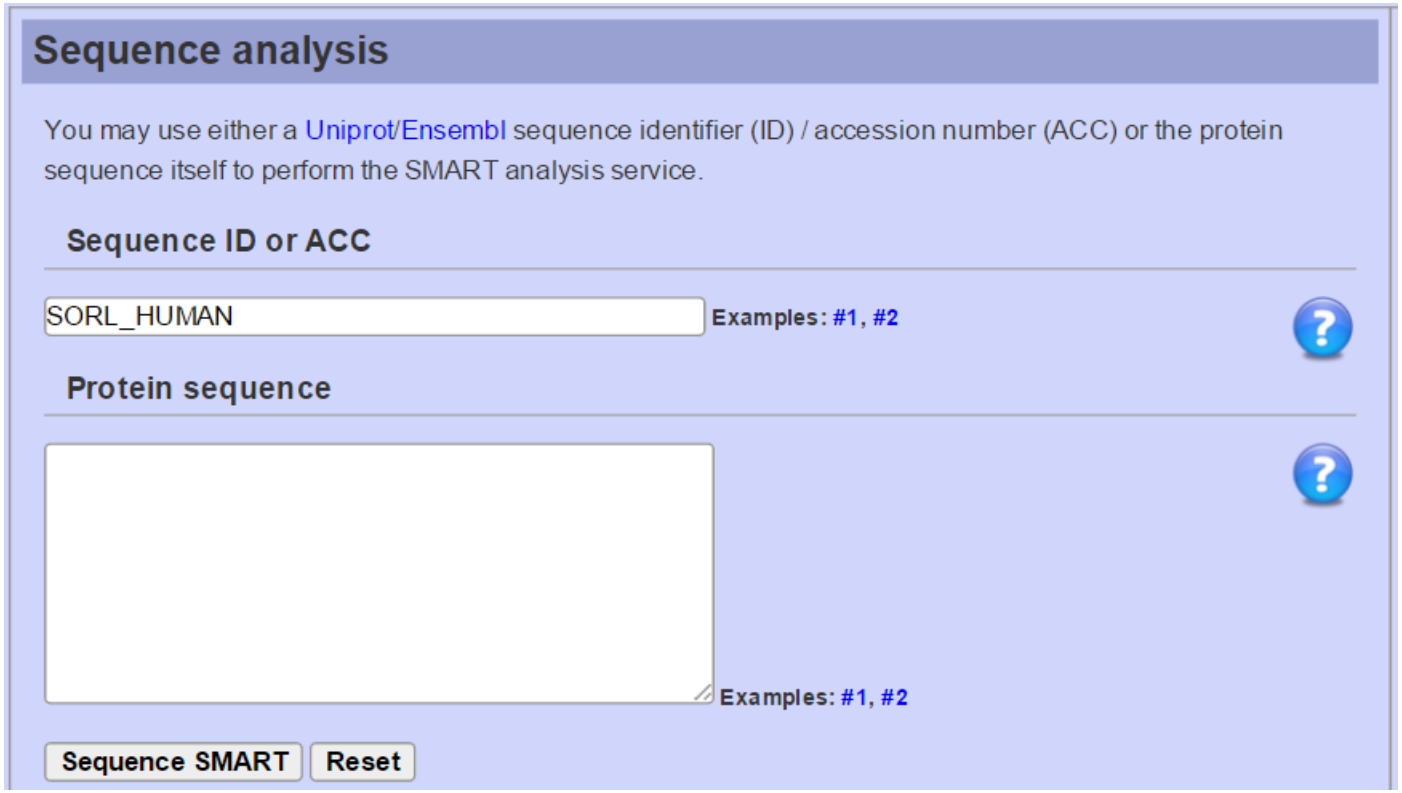
SH3

Src homology 3 domains

SH3

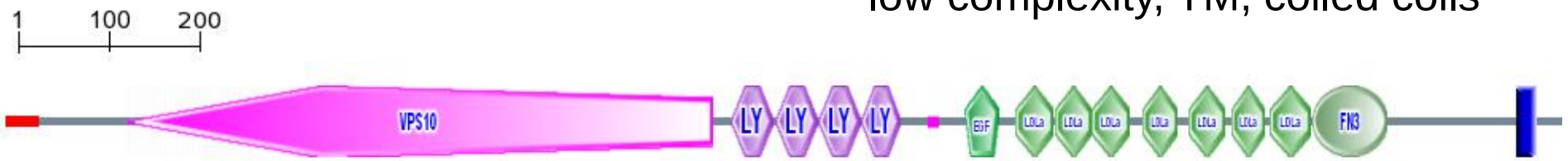
SMART accession number:	SM00326
Description:	Src homology 3 (SH3) domains bind to target proteins through sequences containing proline and hydrophobic amino acids. Pro-containing polypeptides may bind to SH3 domains in 2 different binding orientations.
Interpro abstract (IPR001452):	SH3 (src Homology-3) 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



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
SMART

The following proteins have the same domain composition as your query protein.

You can [display the domain architecture] of [A L L (17)] or selected (below) proteins.

If you want only single domain sequences in the fasta file, type domain name here: [ ]

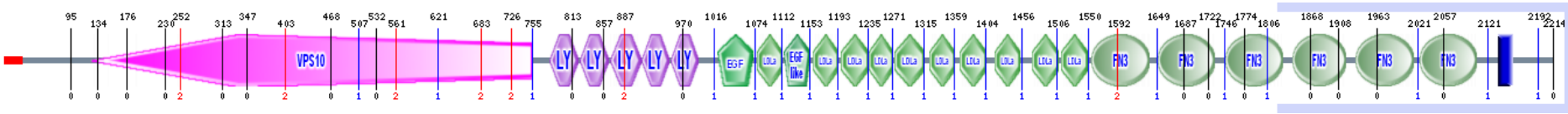
Taxonomic tree of query results.

- Eukaryota (17)
  - Metazoa (17)
    - Arthropoda (5)

Protein	UPI000013D0B1 (source)
Description	Sortilin-related receptor precursor (Sorting protein-related receptor containing LDLR class A repeats) (SorLA) (SorLA-1) (Low-density lipoprotein receptor relative with 11 ligand-binding repeats) (LDLR relative with 11 ligand-binding repeats) (LR11).
Species	*Homo sapiens*
Domain architecture invented in	Eutheria
Representative of protein cluster	CLUST_UPI000013D0B1

1 100 200

Due to overlapping domains, there are 4 representations of the protein



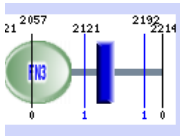
Domain databases
SMART

Protein
Descrip
Species
Domain
invente
Repres
cluster
1
Due to o

EGF

Epidermal growth factor-like domain.

SMART accession number:	SM00181
Description:	
Interpro abstract (IPR006210):	Epidermal growth factors and transforming growth factors belong to a general class of proteins that share a repeat pattern involving a number of conserved Cys residues. Growth factors are involved in cell recognition and division. The repeating pattern, especially of cysteines (the so-called EGF repeat), is thought to be important to the 3D structure of the proteins, and hence its recognition by receptors and other molecules. The type 1 EGF signature includes six conserved cysteines believed to be involved in disulphide bond formation. The EGF motif is found frequently in nature, particularly in extracellular proteins.
Family alignment:	View Alignment consensus sequence or Family alignment in CHROMA format



There are 43703 EGF domains in 14525 proteins in SMART's nrdb database.

Click on the following links for more information.

Evolution (species in which this domain is found)

Structure (3D structures containing this domain)

3D Structures of EGF domains in PDB

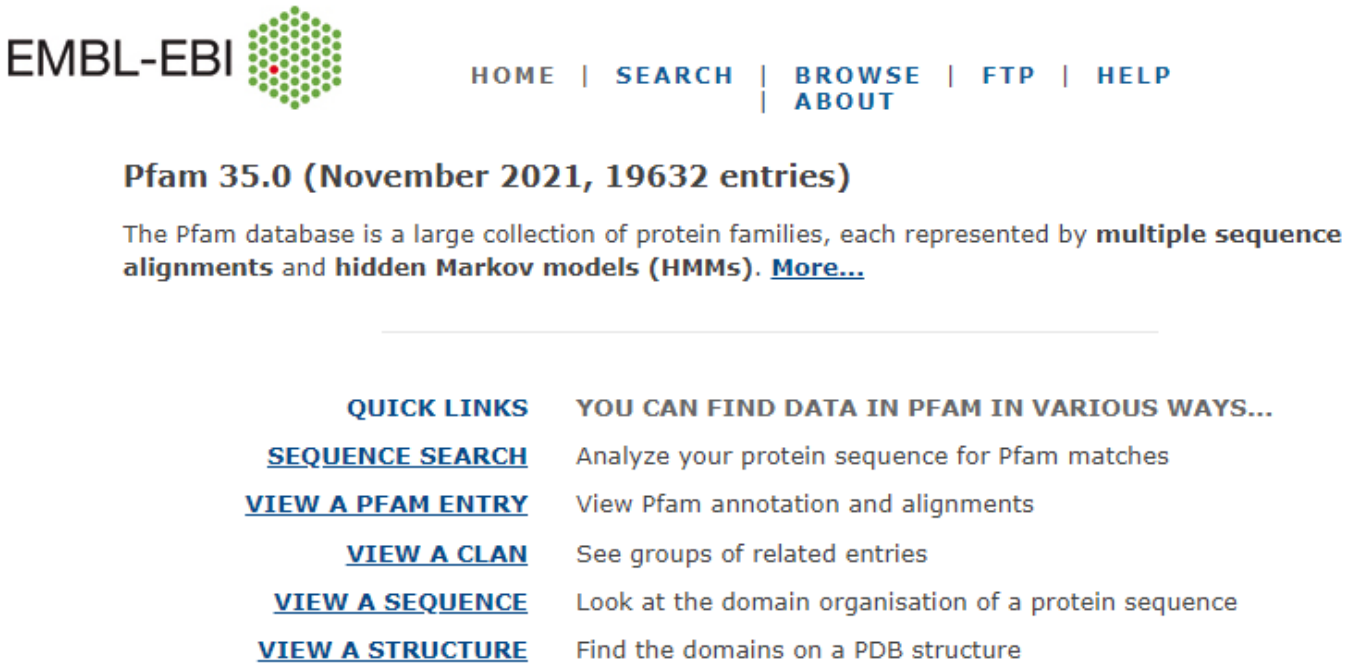
1a3p, 1adx, 1cqe, 1cqe, 1cvu, 1cvu, 1cvw, 1cx2, 1cx2, 1cx2, 1cx2, 1ddx, 1ddx, 1ddx, 1ddx, 1diy, 1dqb, 1dx5, 1dx5, 1dx5, 1dx5, 1ebv, 1egf, 1epg, 1eph, 1epi, 1epj, 1eqg, 1eqg, 1eqh, 1eqh, 1esl, 1fe2, 1fjs, 1fsb, 1g1q, 1g1q, 1g1q, 1g1q, 1g1r, 1g1r, 1g1r, 1g1r, 1g1s, 1g1s, 1g1t, 1gk5, 1gl4, 1hae, 1haf, 1hcg, 1hre, 1hrf, 1ht5, 1ht5, 1ht8, 1ht8, 1igx, 1igz, 1ijq, 1ijq, 1iox, 1ip0, 1ivo, 1ivo, 1j9c, 1jbu, 1jl9, 1jl9, 1k36, 1k37, 1kig, 1kli, 1klj, 1kye, 1mox, 1mox, 1mq5, 1mq6, 1nql, 1p9j, 1pge, 1pge, 1pgf, 1pgf, 1pgg, 1pgg, 1prh, 1prh, 1pth, 1pth, 1pxx, 1pxx, 1pxx, 1pxx, 1q4g, 1q4g, 1qfk, 1rfn, 1tpg, 1u67, 1v3x, 1w7x, 1w8b, 1xdt, 1xfe, 1ygc, 1yo8, 1yuf, 1yug, 1z1y, 1z1y, 1z27, 1z3g, 1z3g, 1z6e, 1zaq, 2adx, 2ayl, 2ayl, 2bmg, 2bok, 2bq6, 2bq7, 2bqw, 2bz6, 2d1j, 2ddu, 2e26, 2fzz, 2g00, 2gd4, 2gd4, 2gy5, 2gy7, 2i9a, 2i9a, 2i9a, 2i9a, 2i9b, 2i9b, 2i9b, 2i9b, 2oye, 2oyu, 2p16, 2p3f, 2p3t, 2p3u, 2p93, 2p94, 2p95, 2pe4, 2pr3, 2puq, 2q1j, 2ra0, 2tgf, 3egf, 3pgh, 3pgh, 3pgh, 3pgh, 3tgf, 4cox, 4cox, 4cox, 4cox, 4tgf, 5cox, 5cox, 5cox, 5cox, 6cox, 6cox

Domain databases

PFAM (until Jan 2023)

Erik Sonnhammer/Ewan Birney/Alex Bateman

http://pfam.xfam.org/



Sonnhammer et al (1997) *Proteins*

...

Mistry et al (2021) *Nucleic Acids Research*

Domain databases
PFAM

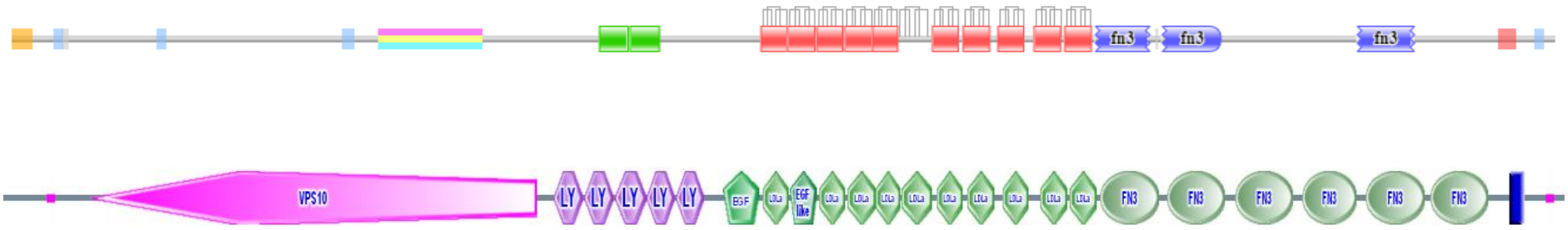
This is the summary of UniProt entry SORL_HUMAN (Q92673).

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.

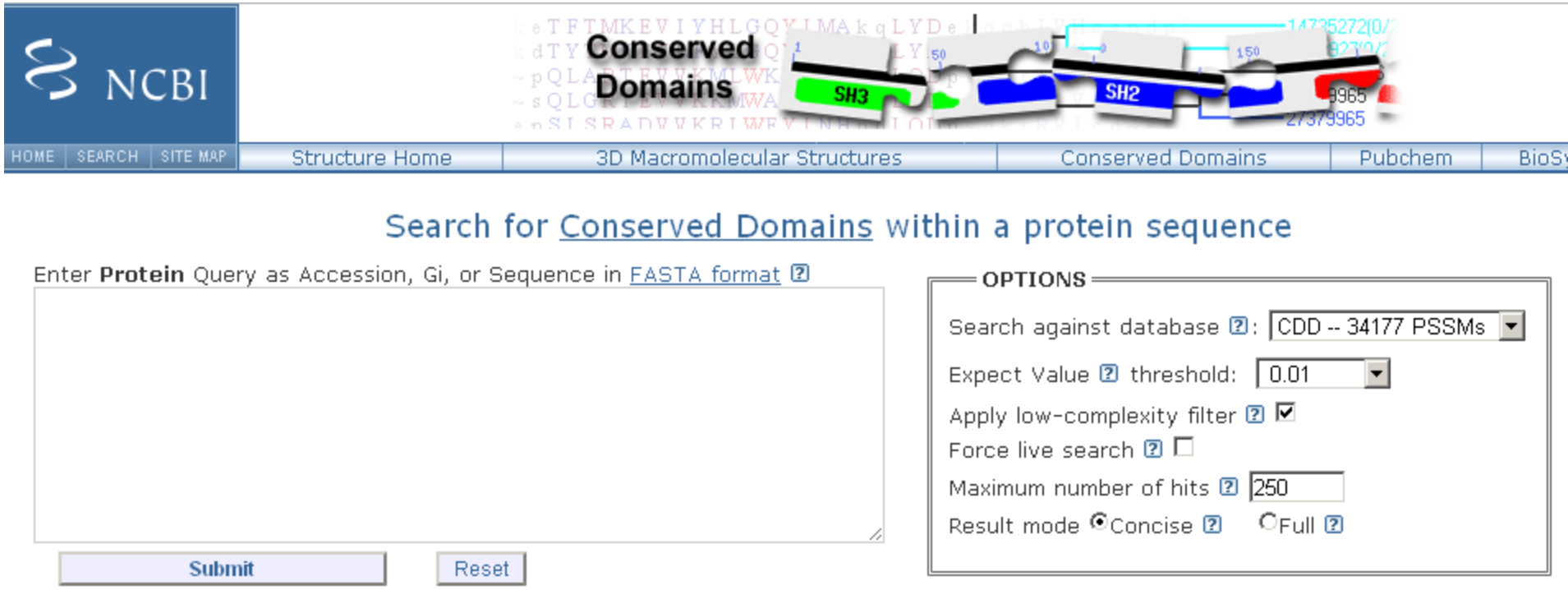
Pfam domains

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...**



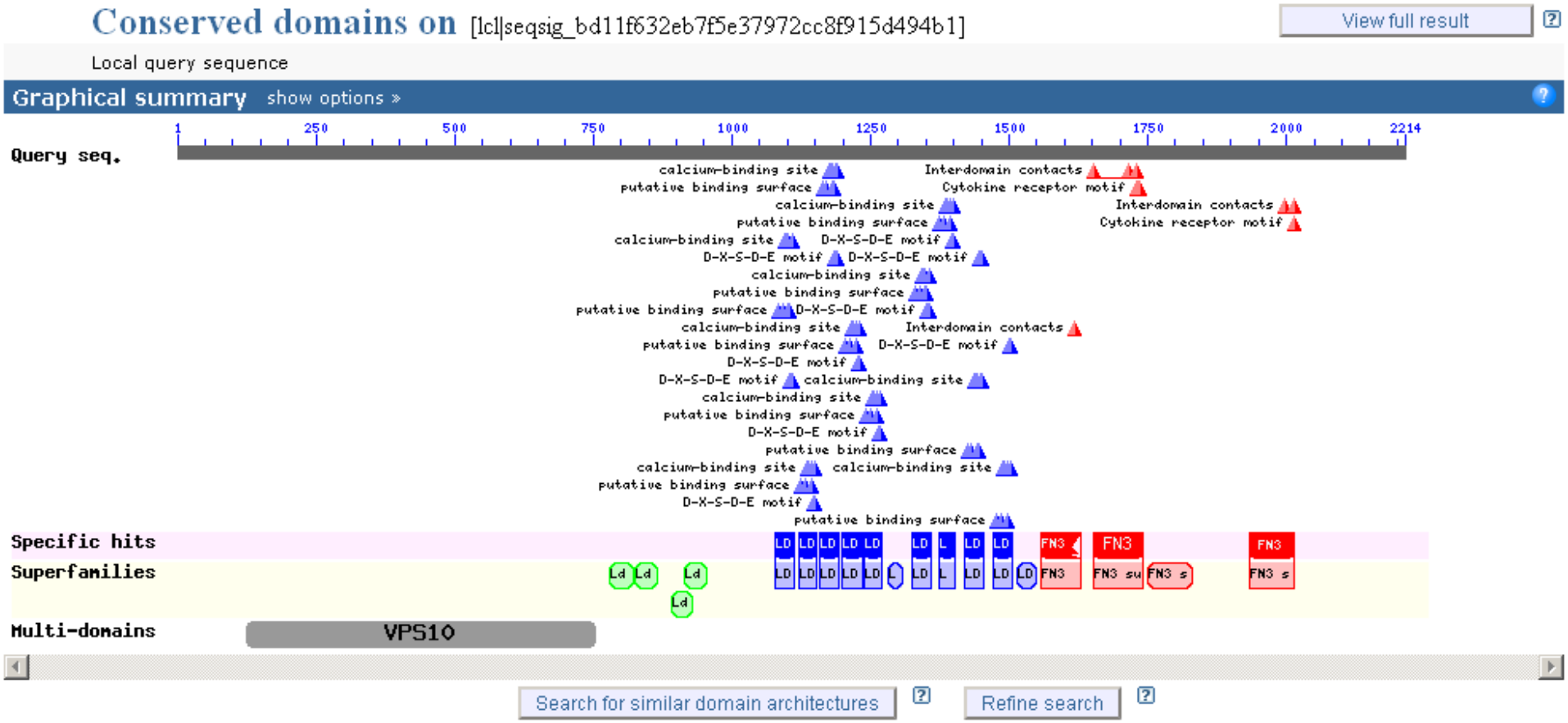
Domain databases
CDD

Stephen Bryant http://www.ncbi.nlm.nih.gov/cdd



Wang et al (2022) *Nucleic Acids Res*

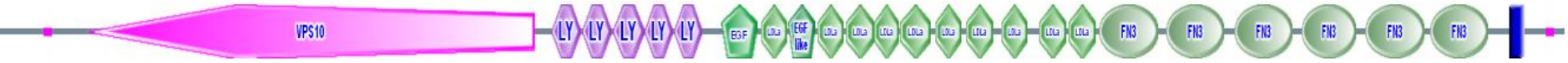
Domain databases
CDD



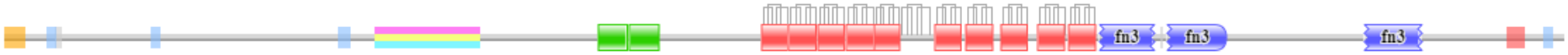
Domain databases

SORLA/SORL1 from *Homo sapiens*

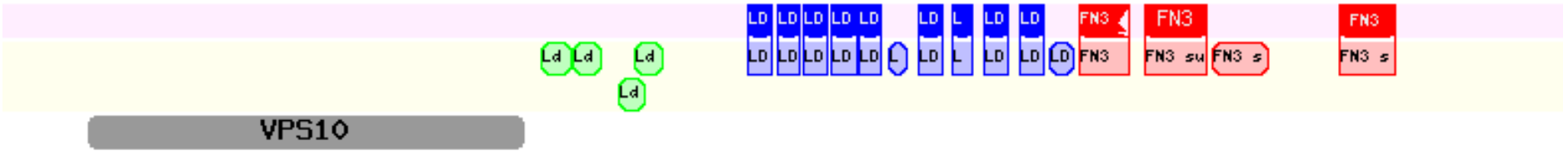
SMART



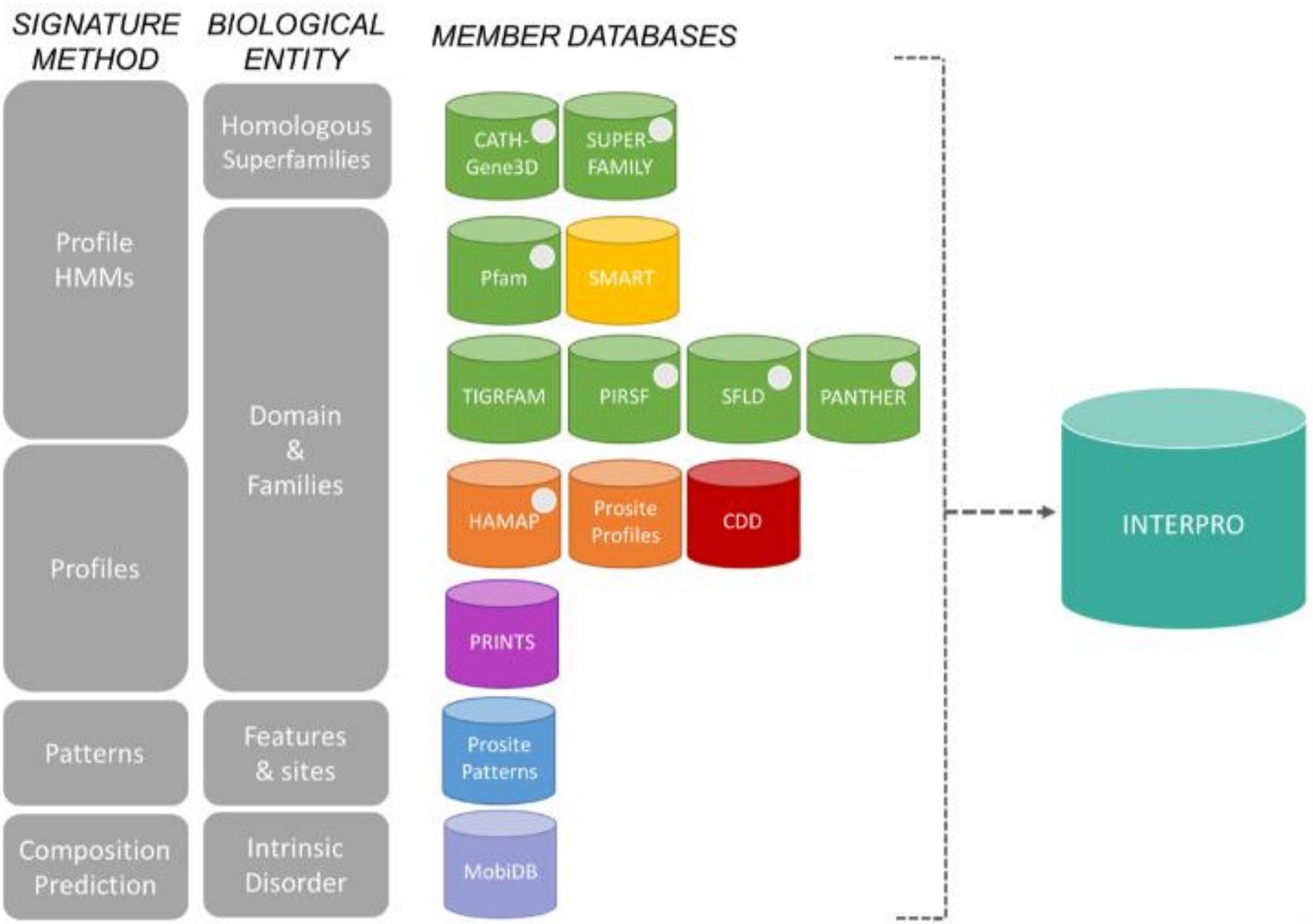
PFAM



CDD



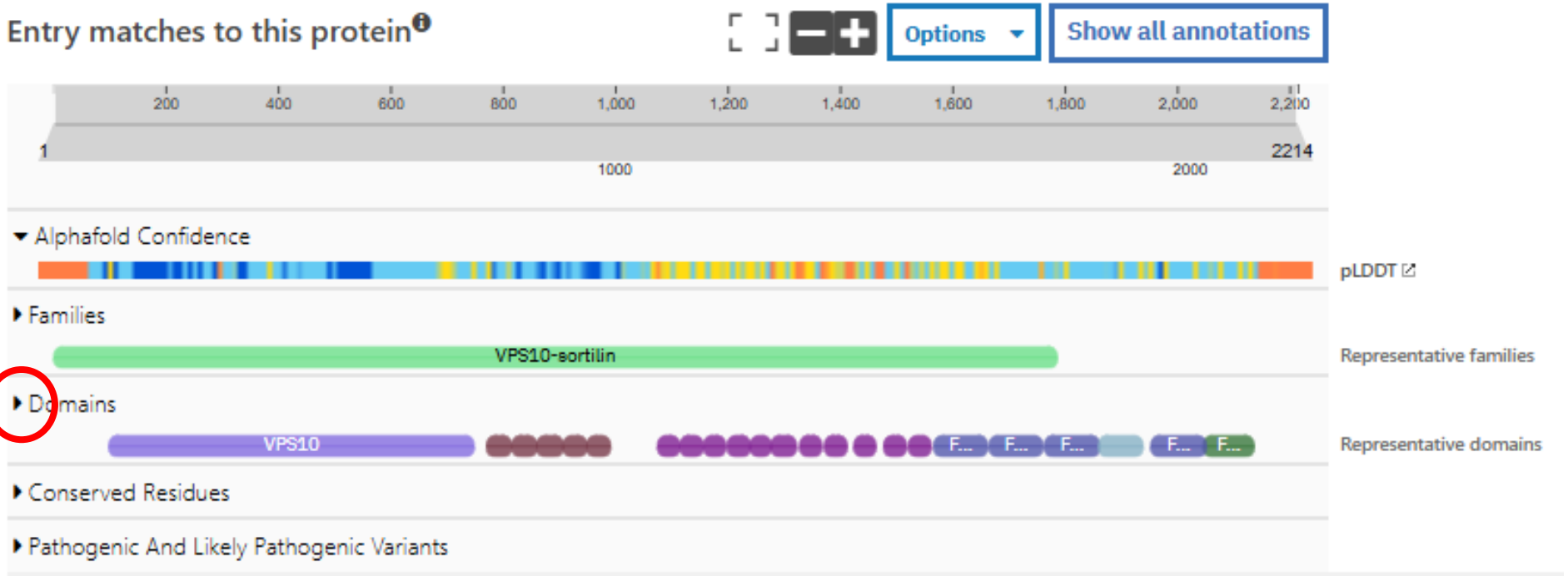
InterPro

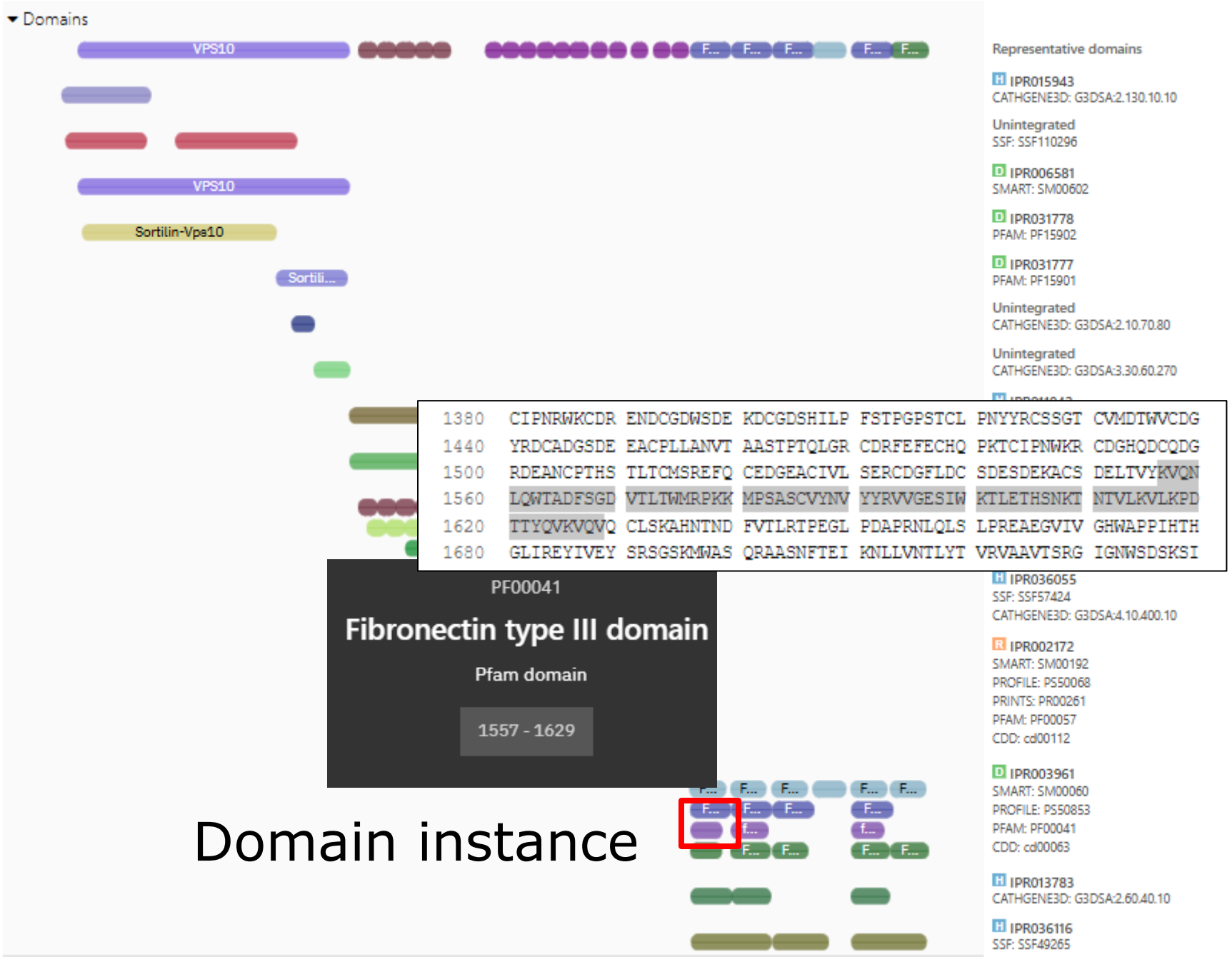

SIGNATURE METHOD
BIOLOGICAL ENTITY
MEMBER DATABASES
Profile HMMs
Profiles
Patterns
Composition Prediction
Homologous Superfamilies
Domain & Families
Features & sites
Intrinsic Disorder
CATH-Gene3D
SUPER-FAMILY
Pfam
SMART
TIGRFAM
PIRSF
SFLD
PANTHER
HAMAP
Prosite Profiles
CDD
PRINTS
Prosite Patterns
MobiDB
INTERPRO

InterPro

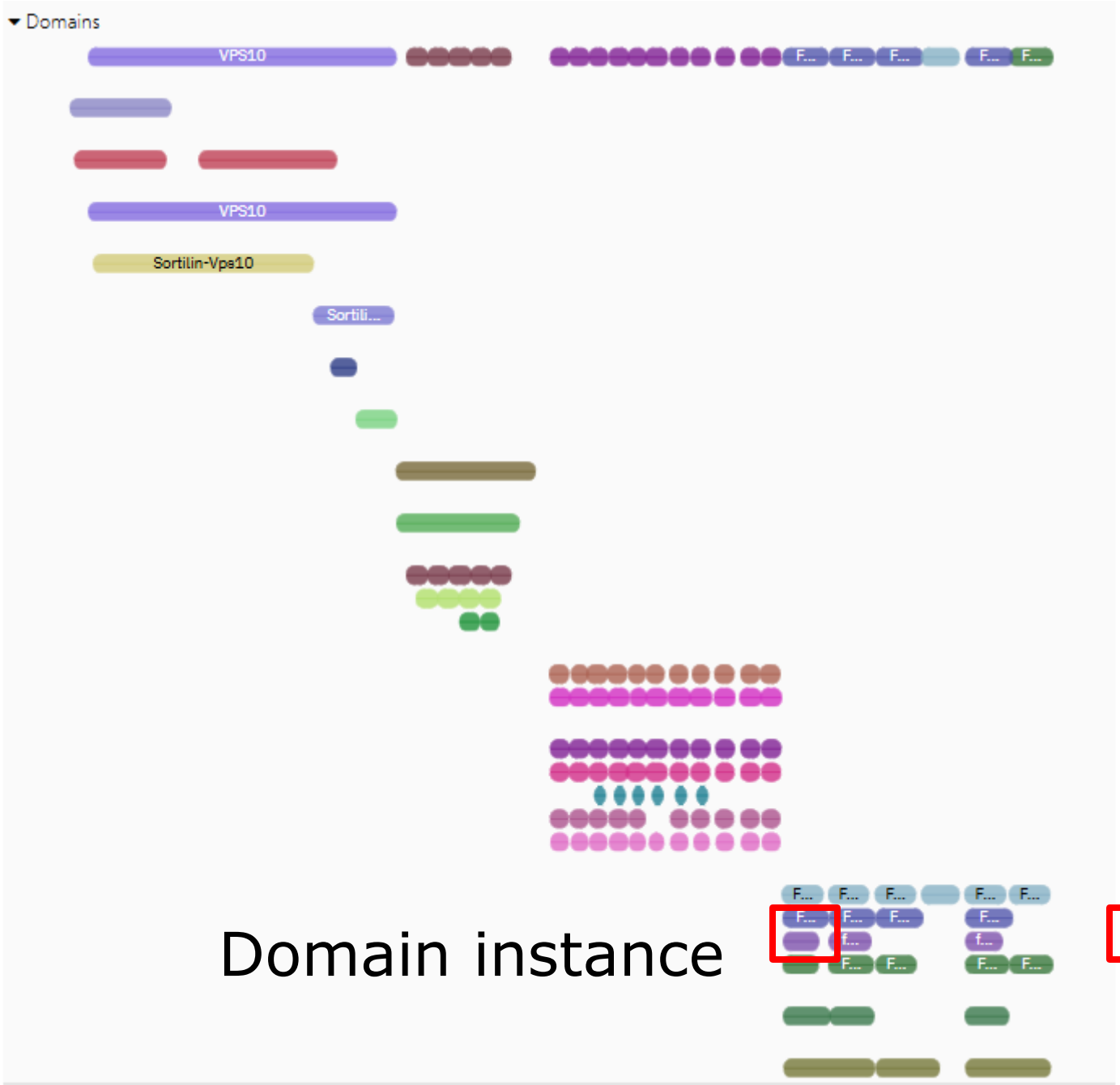
SORLA/SORL1 from *Homo sapiens*

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





Domain instance


Domains
VPS10
VPS10
Sortilin-Vps10
Sortili...
Representative domains
IPR015943
CATHGENE3D: G3DSA:2.130.10.10
Unintegrated
SSF: SSF110296
IPR006581
SMART: SM00602
IPR031778
PFAM: PF15902
IPR031777
PFAM: PF15901
Unintegrated
CATHGENE3D: G3DSA:2.10.70.80
Unintegrated
CATHGENE3D: G3DSA:3.30.60.270
IPR011042
CATHGENE3D: G3DSA:2.120.10.30
Unintegrated
SSF: SSF63825
IPR000033
SMART: SM00135
PROFILE: PS51120
PFAM: PF00058
IPR036055
SSF: SSF57424
CATHGENE3D: G3DSA:4.10.400.10
IPR002172
PFAM: PF00057
CDD: cd00112
IPR003961
SMART: SM00060
PROFILE: PS50853
PFAM: PF00041
IPR013783
CATHGENE3D: G3DSA:2.60.40.10
IPR036116
SSF: SSF49265
Domain instance
Database domain entry

InterPro

Pfam PF00041 Fibronectin type III domain

Pfam entry

Member database	Pfam
Pfam type	domain
Short name	*fn3*
Clan	E-set
Author	Sonnhammer ELL;0000-0002-9015-5588
Sequence Ontology	0000417

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].

Provide feedback

Integrated to

IPR003961

Representative structure

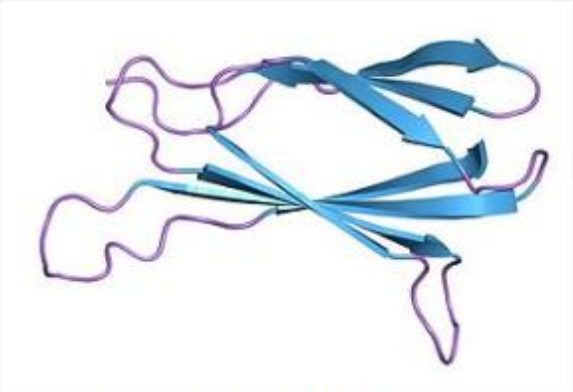
1ten: STRUCTURE OF A FIBRONECTIN TYPE III DOMAIN FROM TENASCIN PHASED BY MAD ANALYSIS OF THE SELENOMETHIONYL PROTEIN

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



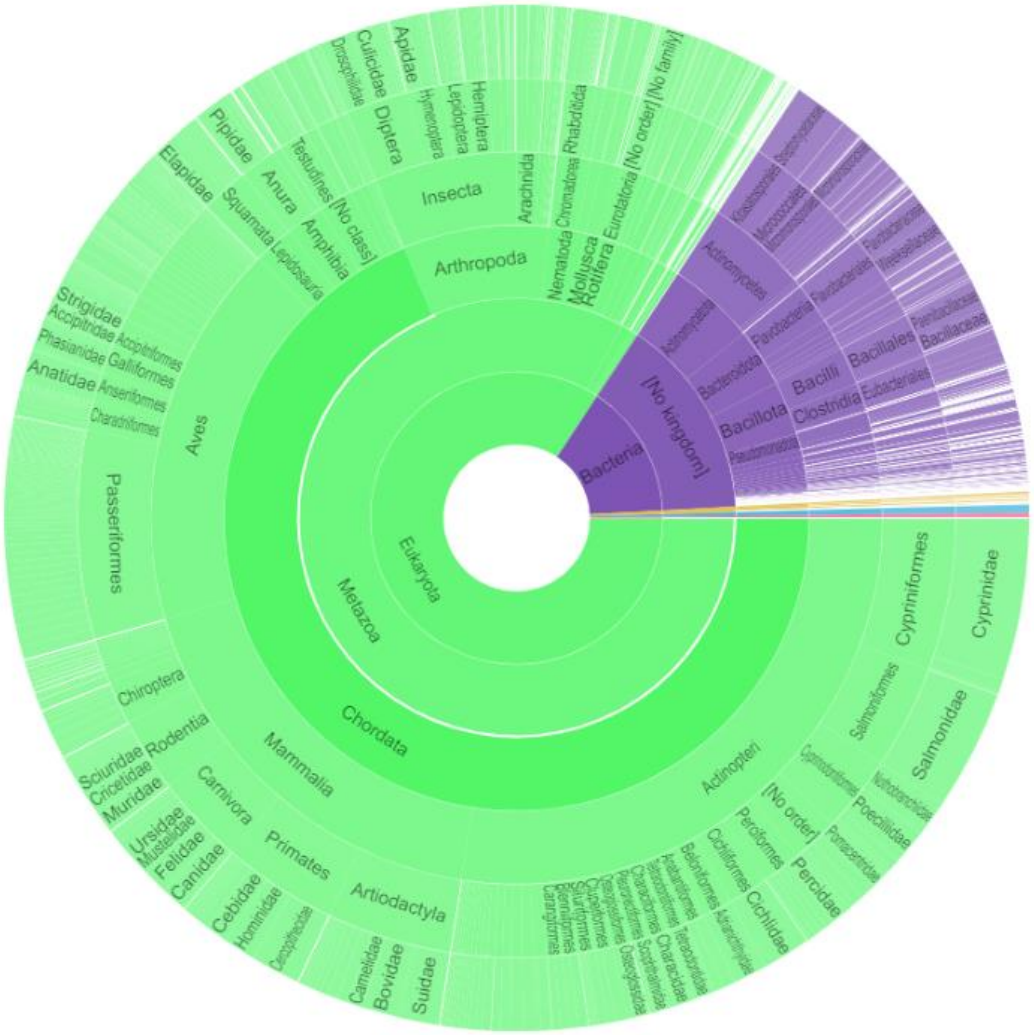
The tenth type III domain of fibronectin

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

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

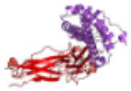
- Overview
- Proteins 266k
- Domain Architectures 21k
- Taxonomy 23k
- Proteomes 5k
- **Structures** 463
- Signature
- AlphaFold 125k
- Alignment
- Curation

This entry matches these structures:

1 - 20 of **450** structures

Search

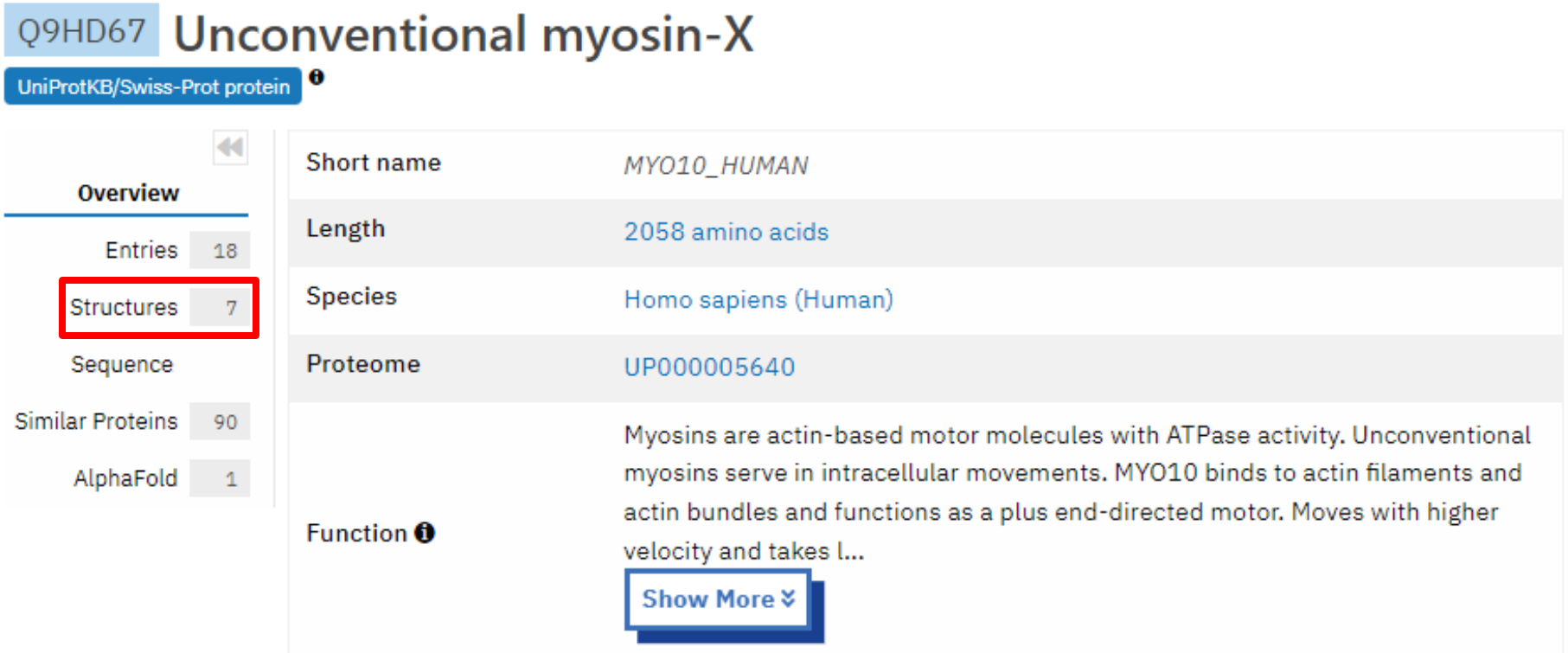
Export

ACCESSION	NAME	SOURCE DATABASE	STRUCTURE	MATCHES
1a22	HUMAN GROWTH HORMONE BOUND TO SINGLE RECEPTOR	PDB		B 100 200
1axi	STRUCTURAL PLASTICITY AT THE HGH:HGHBP INTERFACE	PDB		B 100 200
1bj8	THIRD N-TERMINAL DOMAIN OF GP130, NMR, MINIMIZED AVERAGE STRUCTURE	PDB		A 50 100
1bpv	TITIN MODULE A71 FROM HUMAN CARDIAC MUSCLE, NMR, 50 STRUCTURES	PDB		A 50 100

Exercise 1

Find structures in the PDB for human myosin X

Search InterPro by text using UniProt identifier Q9HD67
https://ebi.ac.uk/interpro/protein/reviewed/Q9HD67/



Exercise 1

Find structures in the PDB for human myosin X

- Which domains of myosin X are covered by the solved structures?

- Is there a part of the protein for which there are no know structures? Does it have predicted domains?

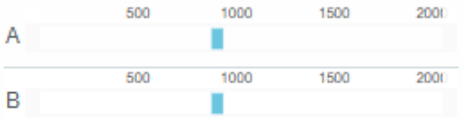
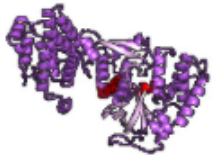
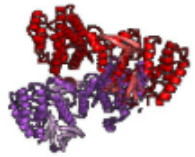
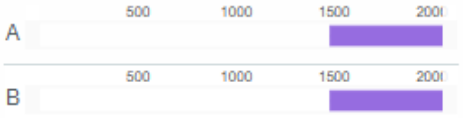
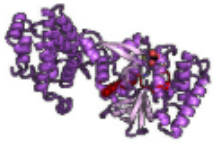
Exercise 2

Compare domain predictions to structure

1 - 7 of 7 structures

Search

H Download

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		

Exercise 2

Compare domain predictions to structure

- Open the structure of the 4th hit (3PZD) in Chimera

Now colour the fragments corresponding to the representative domains MyTH4 (in pink), B41 (in blue) and FERM (in purple).

How do the domain annotations fit the structure?

- Chain B in this structure is a small peptide.

Which domain is interacting with this peptide?