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

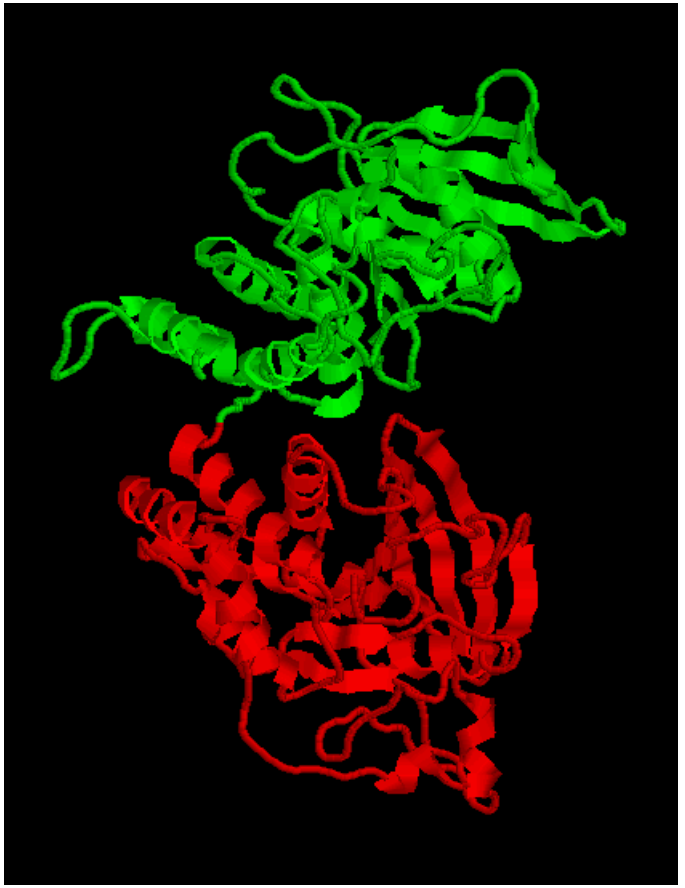
Proteins normally are multidomain (average 300 aa)

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)

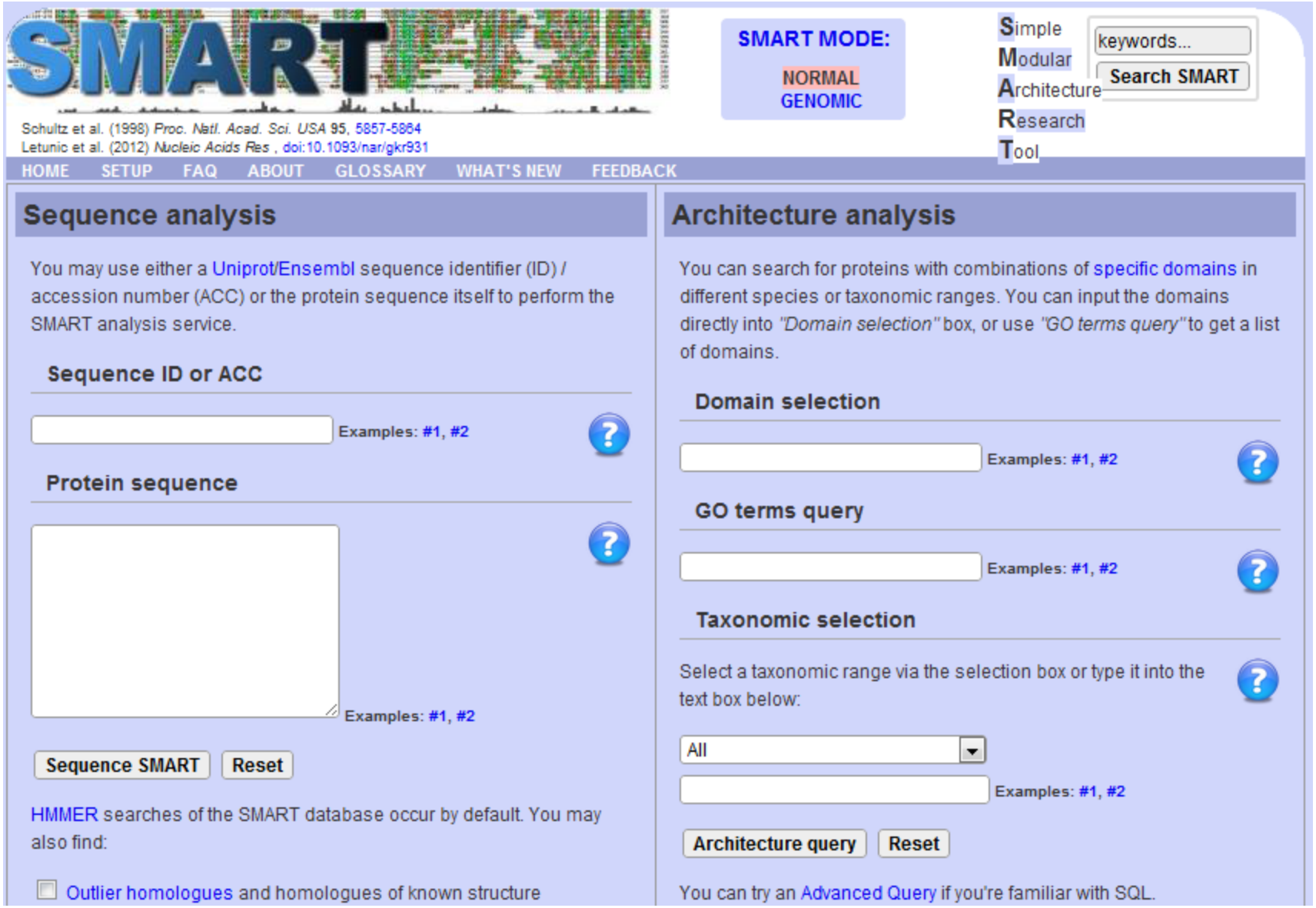
Generate profile from instances of domain
Search for remote homologs (HMMer)
Include them in profile
Iterate until convergence

Schultz et al (1998) *PNAS*

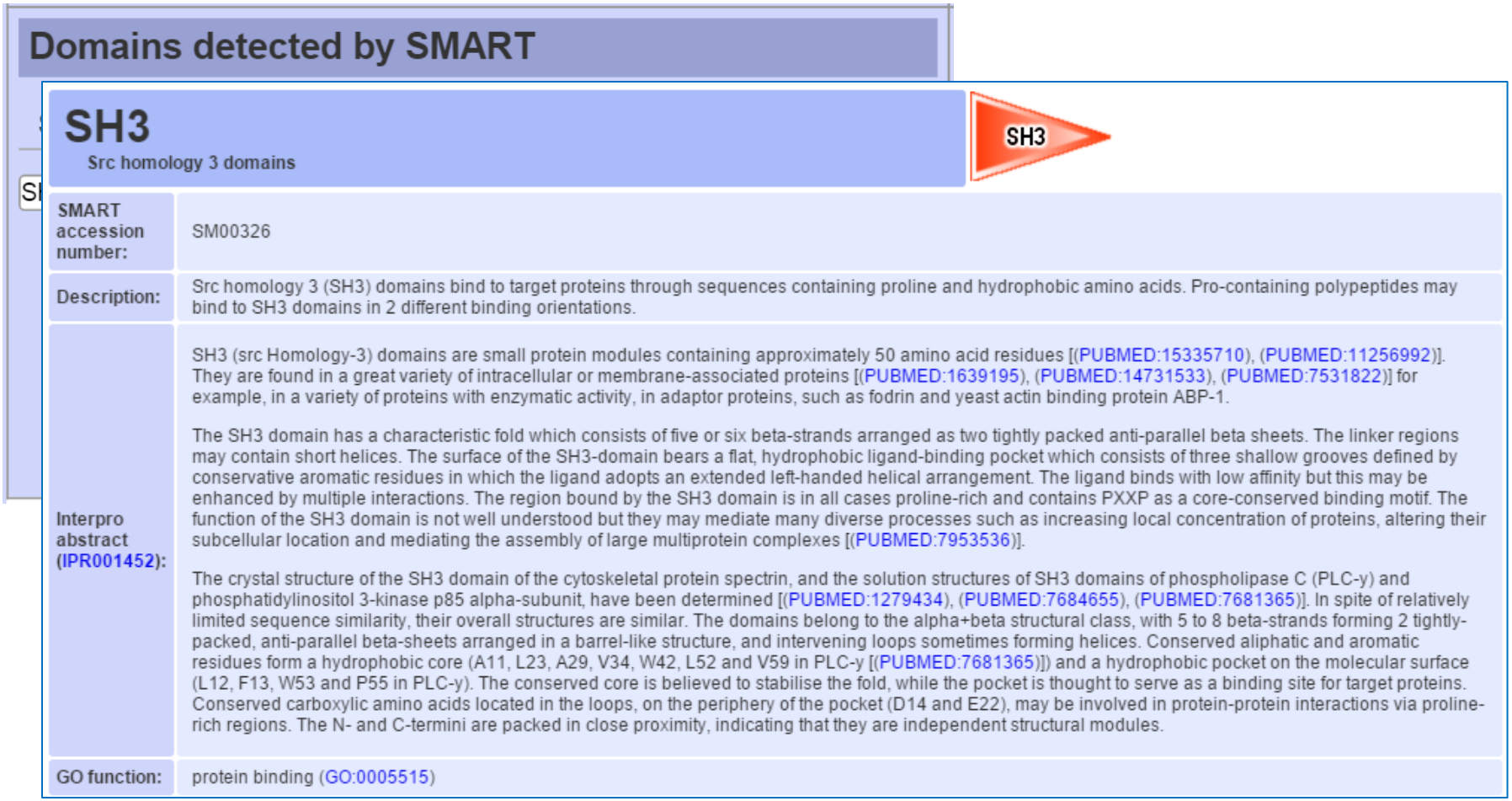
…

Letunic et al (2014) *Nucleic Acids Research*

Domain databases



Domain databases
SMART



Domains detected by SMART

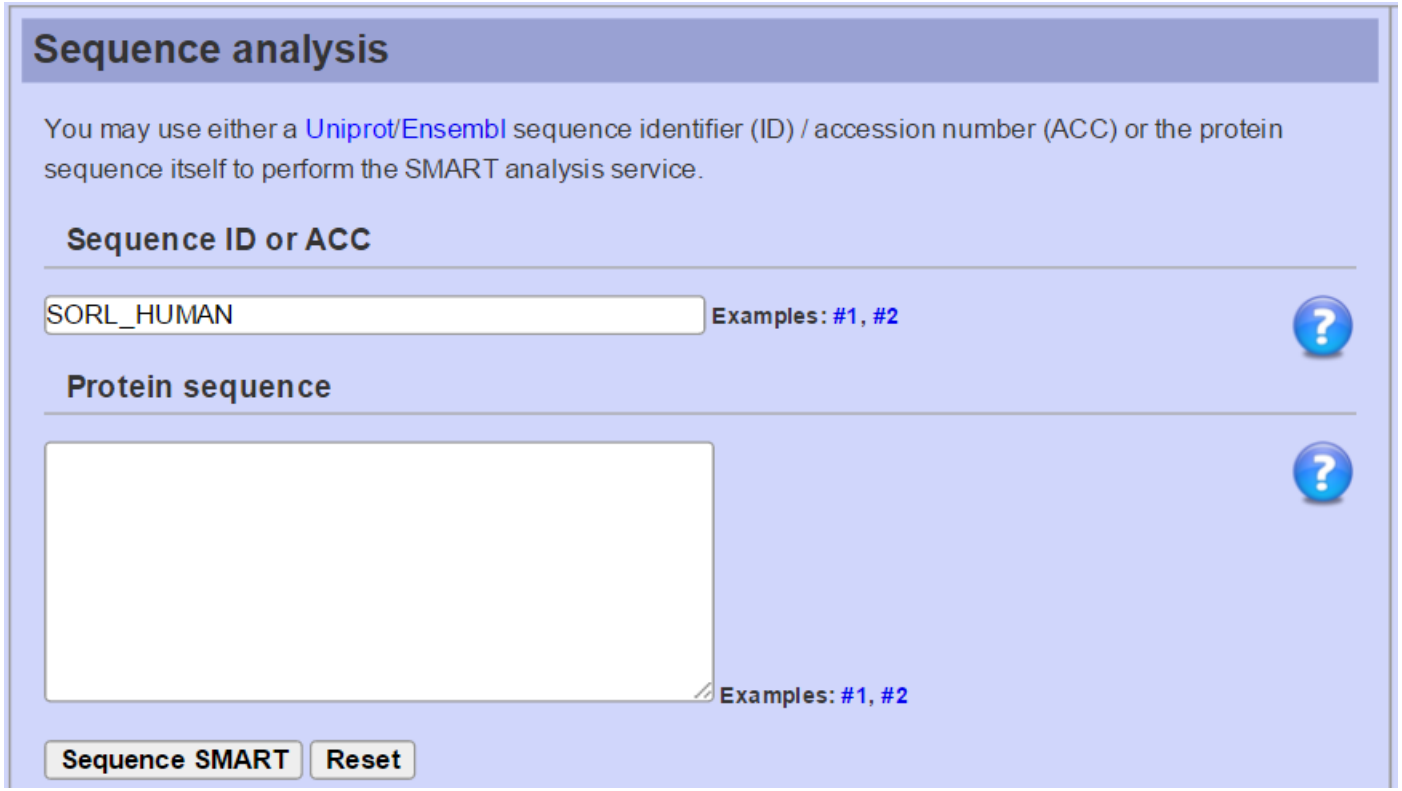
SH3

Src homology 3 domains

SH3

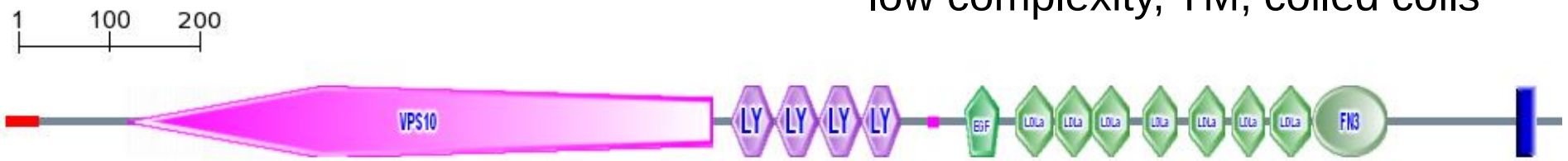
SMART accession number:	SM00326
Description:	Src homology 3 (SH3) domains bind to target proteins through sequences containing proline and hydrophobic amino acids. Pro-containing polypeptides may bind to SH3 domains in 2 different binding orientations.
Interpro abstract (IPR001452):	SH3 (src Homology-3)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

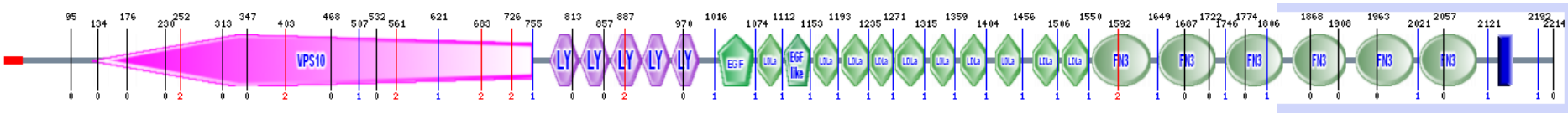
Taxonomic tree of query results.

- 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



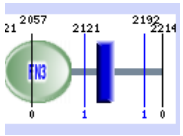
Domain databases
SMART

Protein
Descrip
Species
Domain inventec
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)

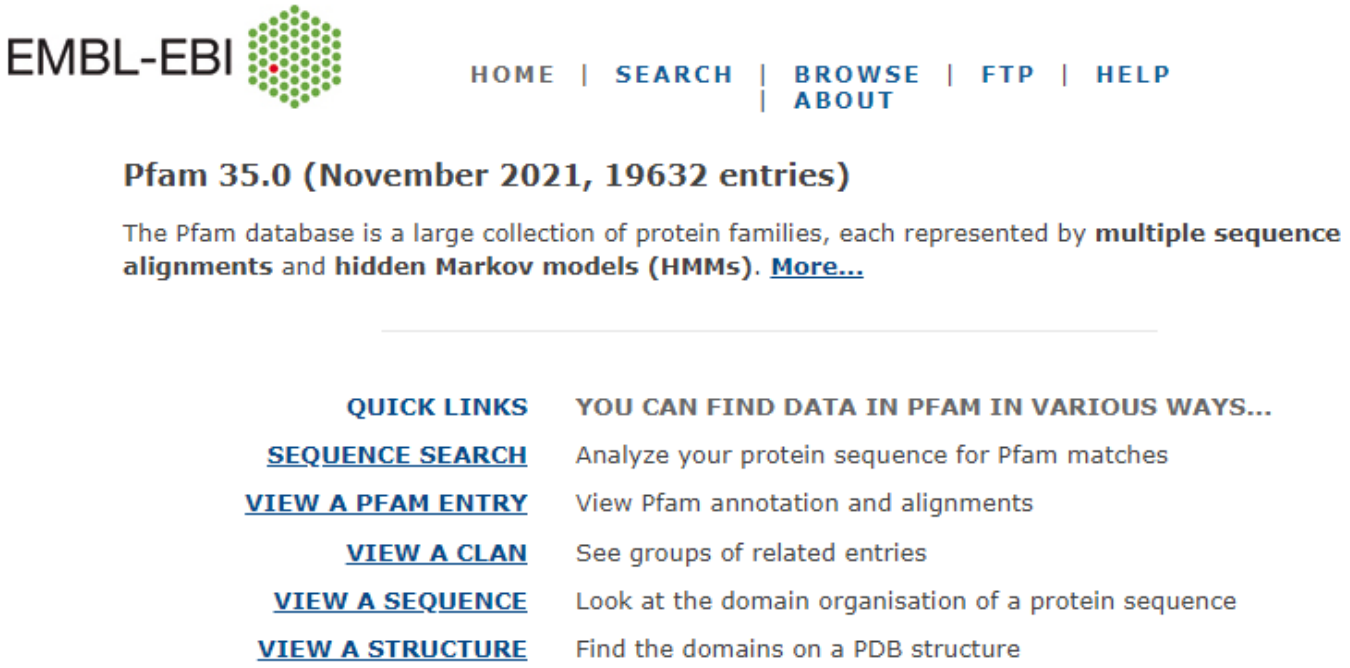
3D Structures of EGF domains in PDB

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

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

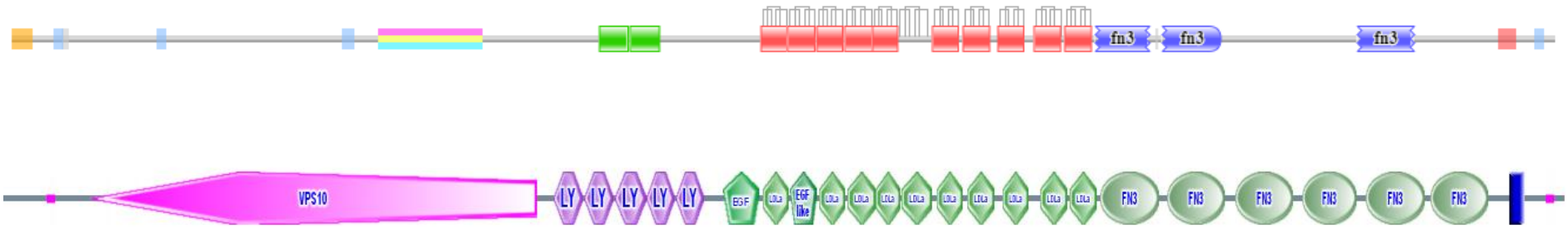
This is the summary of UniProt entry SORL_HUMAN (Q92673).

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.

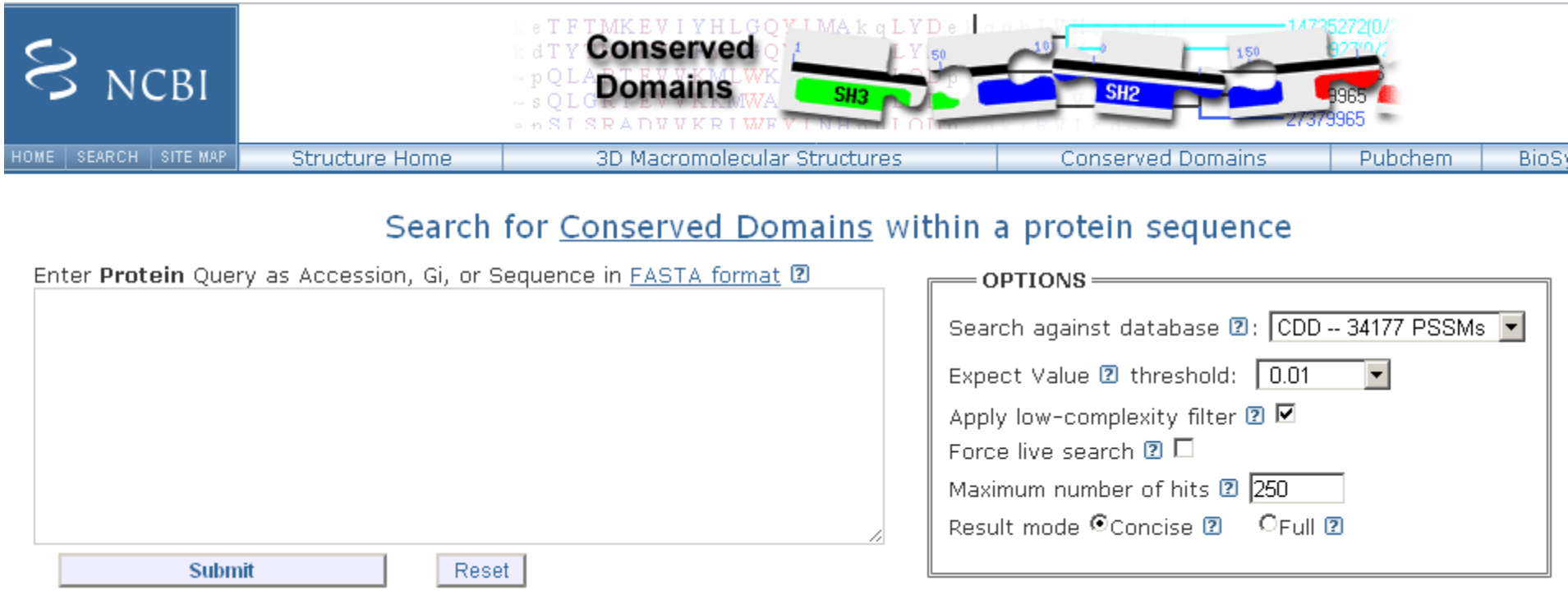
Pfam domains

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...**



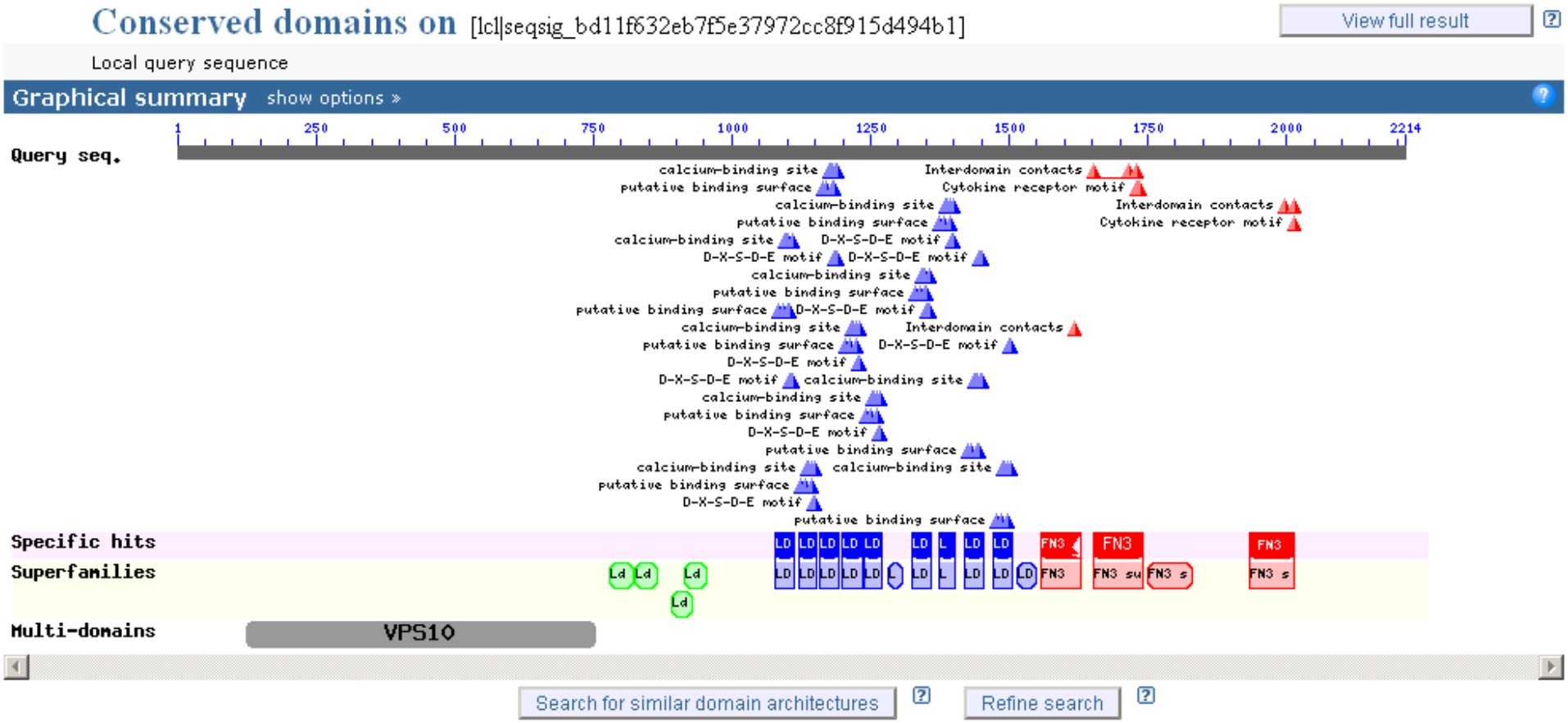
Domain databases
CDD

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



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

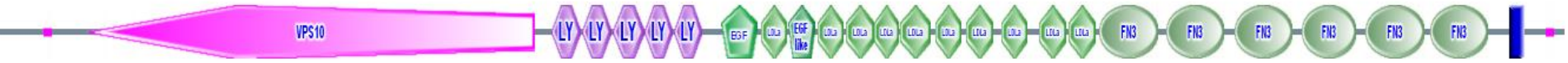
Domain databases
CDD



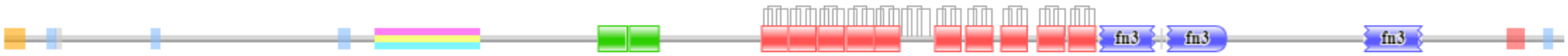
Domain databases

SORLA/SORL1 from *Homo sapiens*

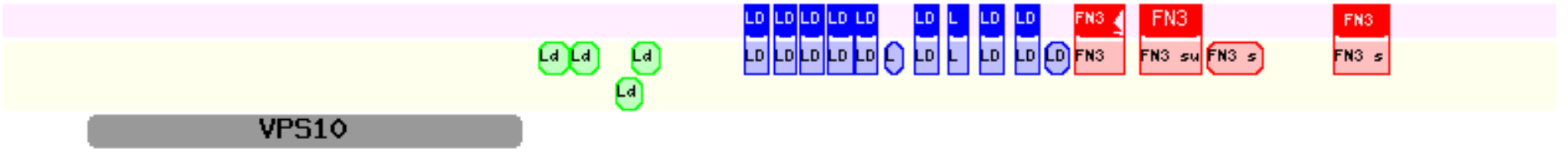
SMART



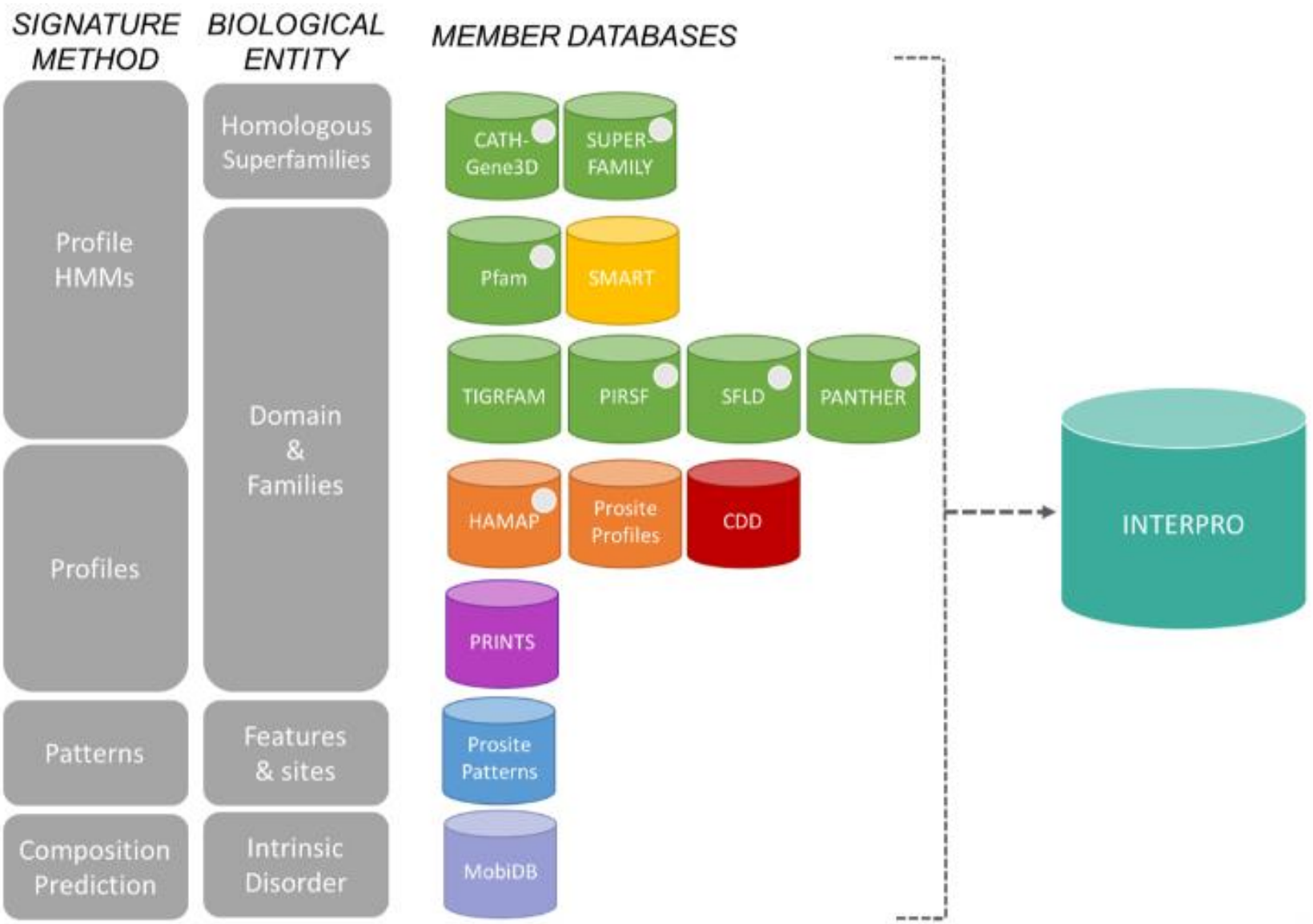
PFAM



CDD



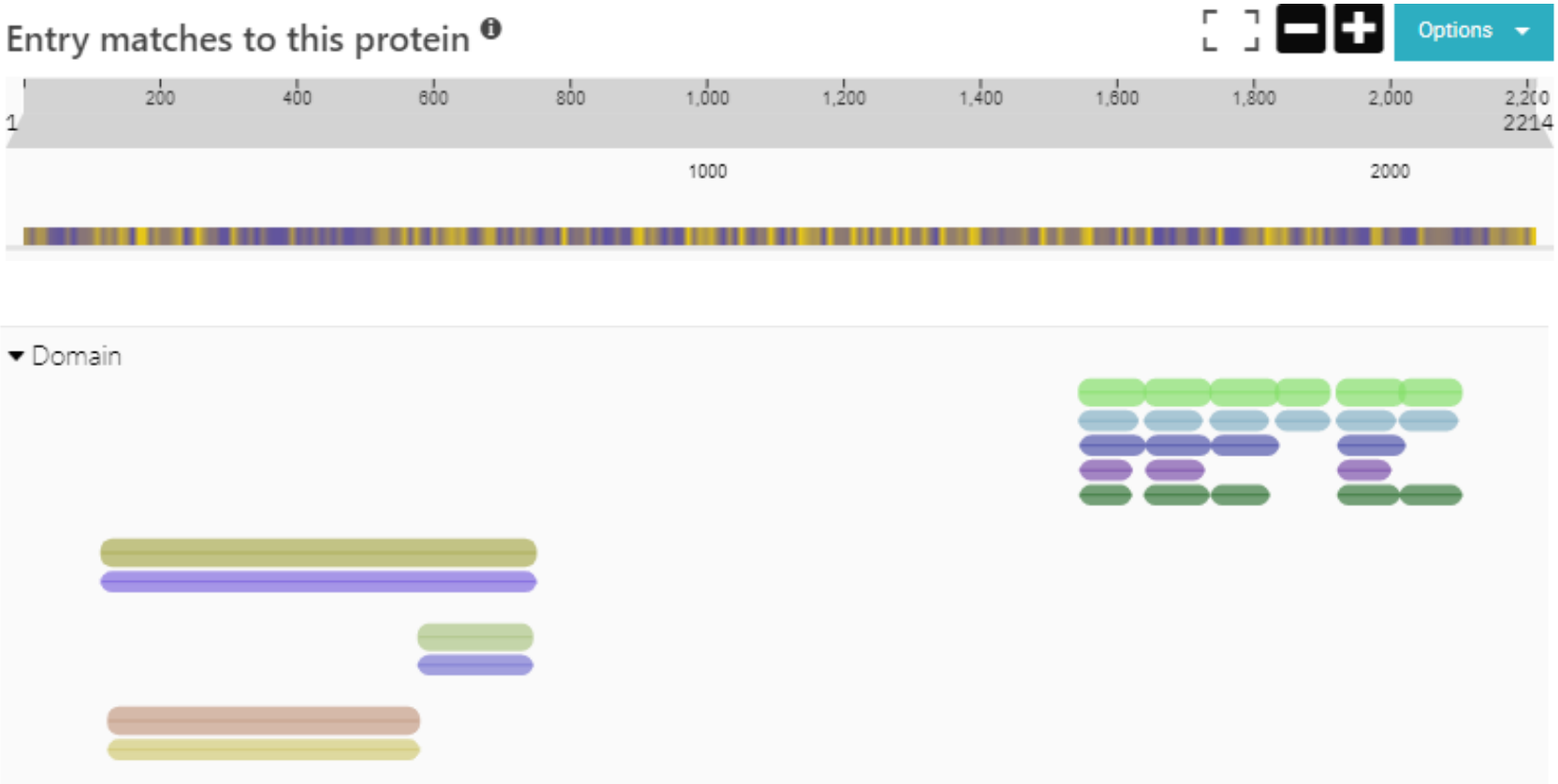
InterPro


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*

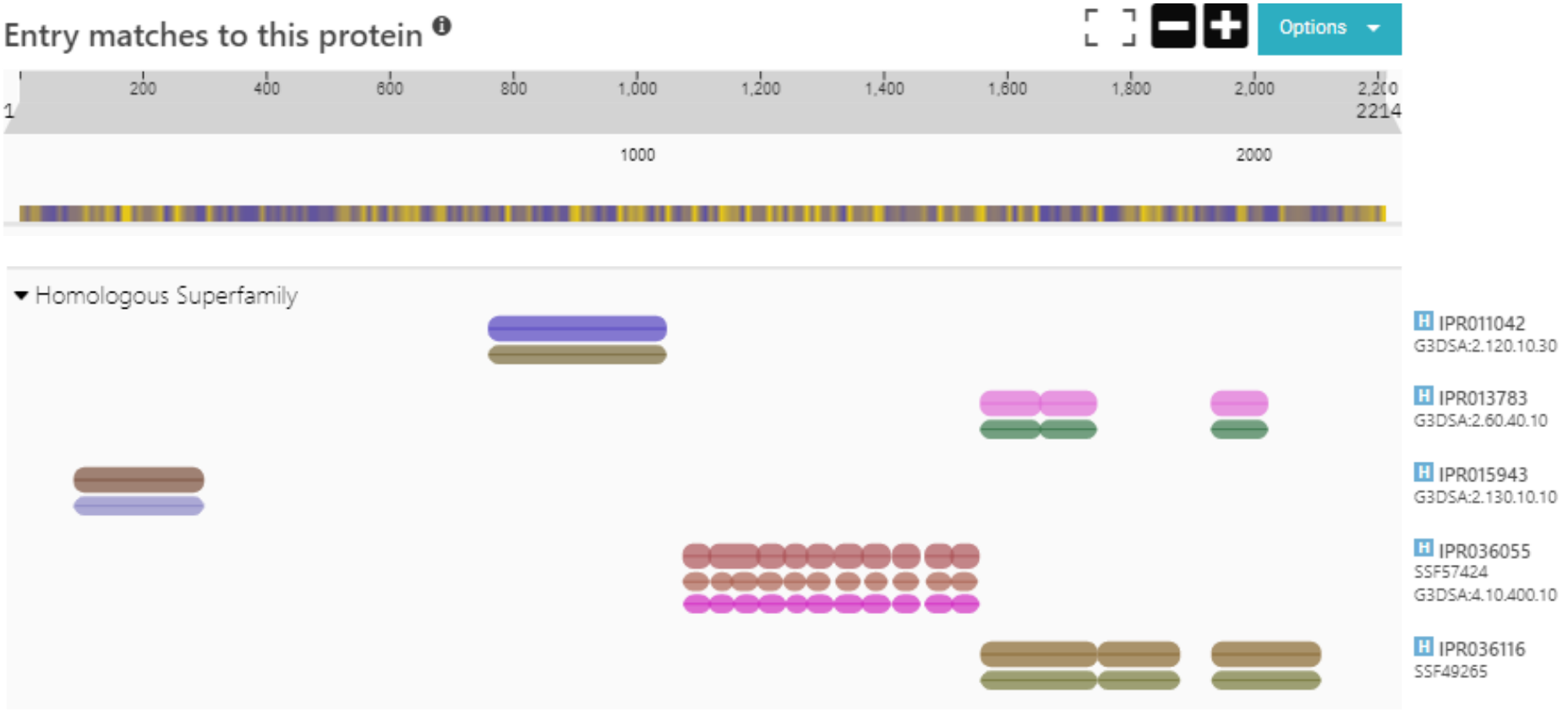
https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/



InterPro

SORLA/SORL1 from *Homo sapiens*

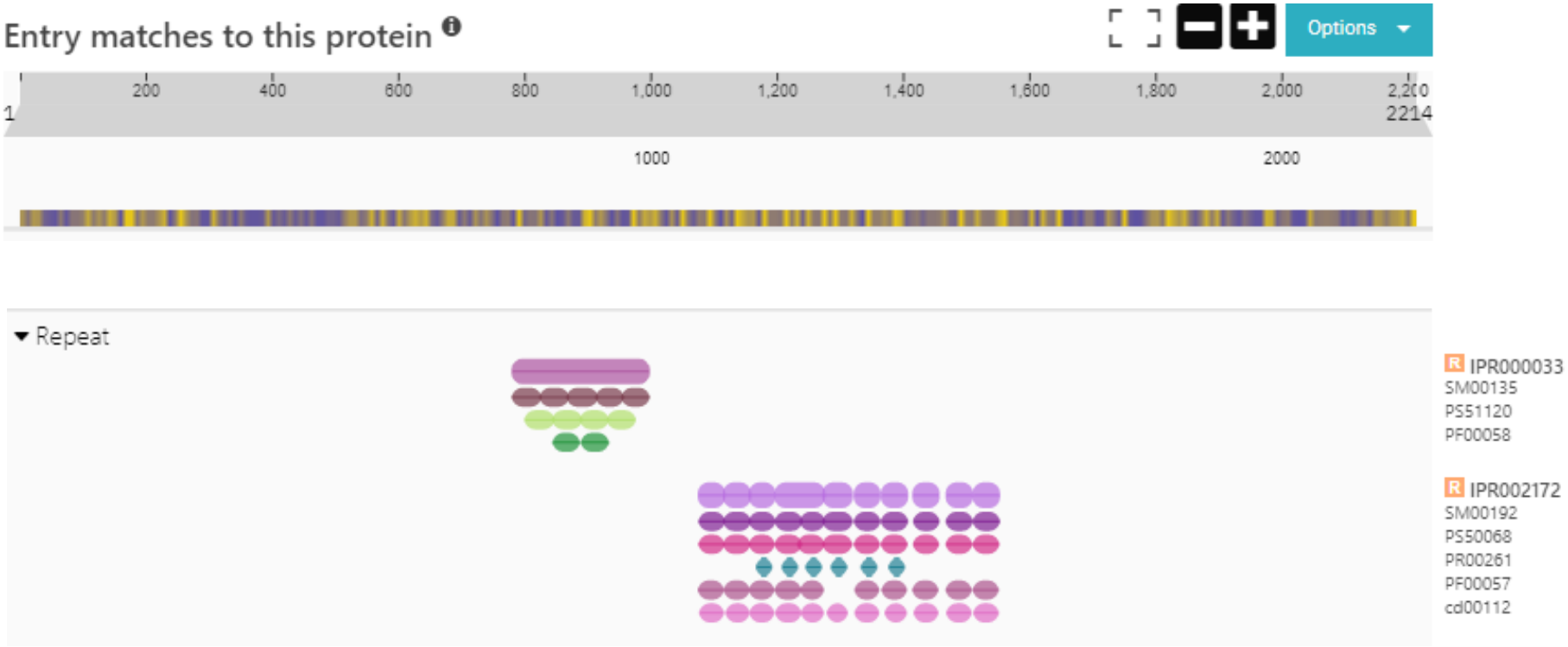
https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/



InterPro

SORLA/SORL1 from *Homo sapiens*

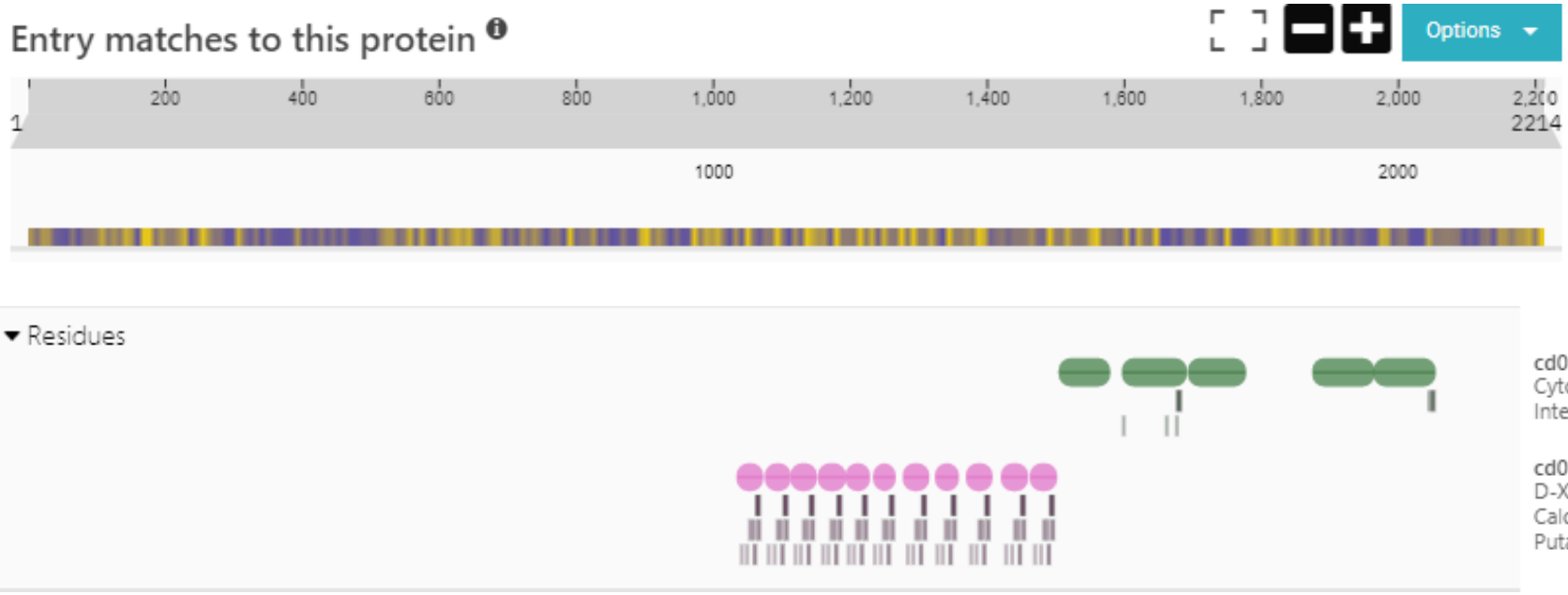
https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/



InterPro

SORLA/SORL1 from *Homo sapiens*

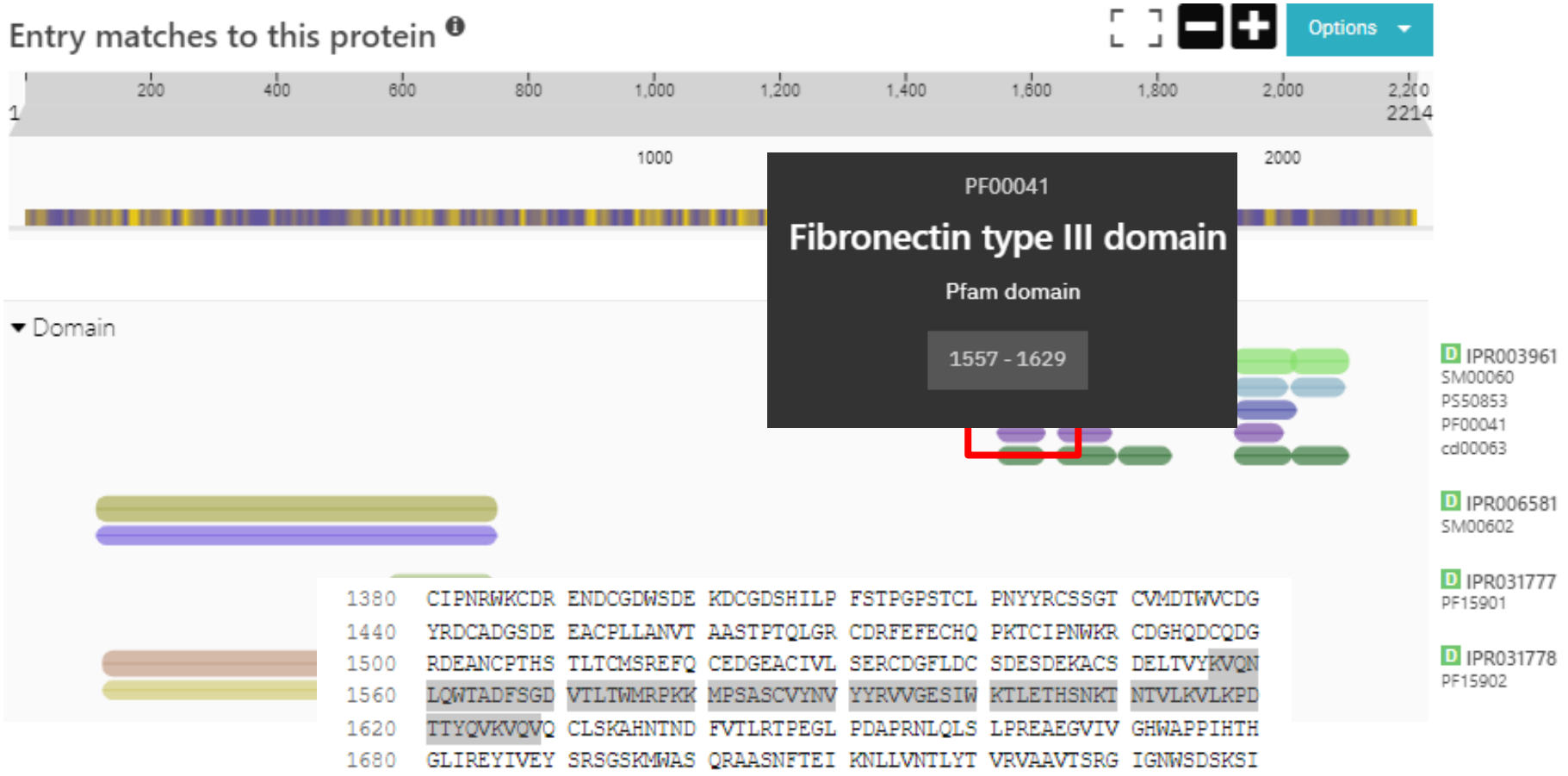
https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/



InterPro

SORLA/SORL1 from *Homo sapiens*

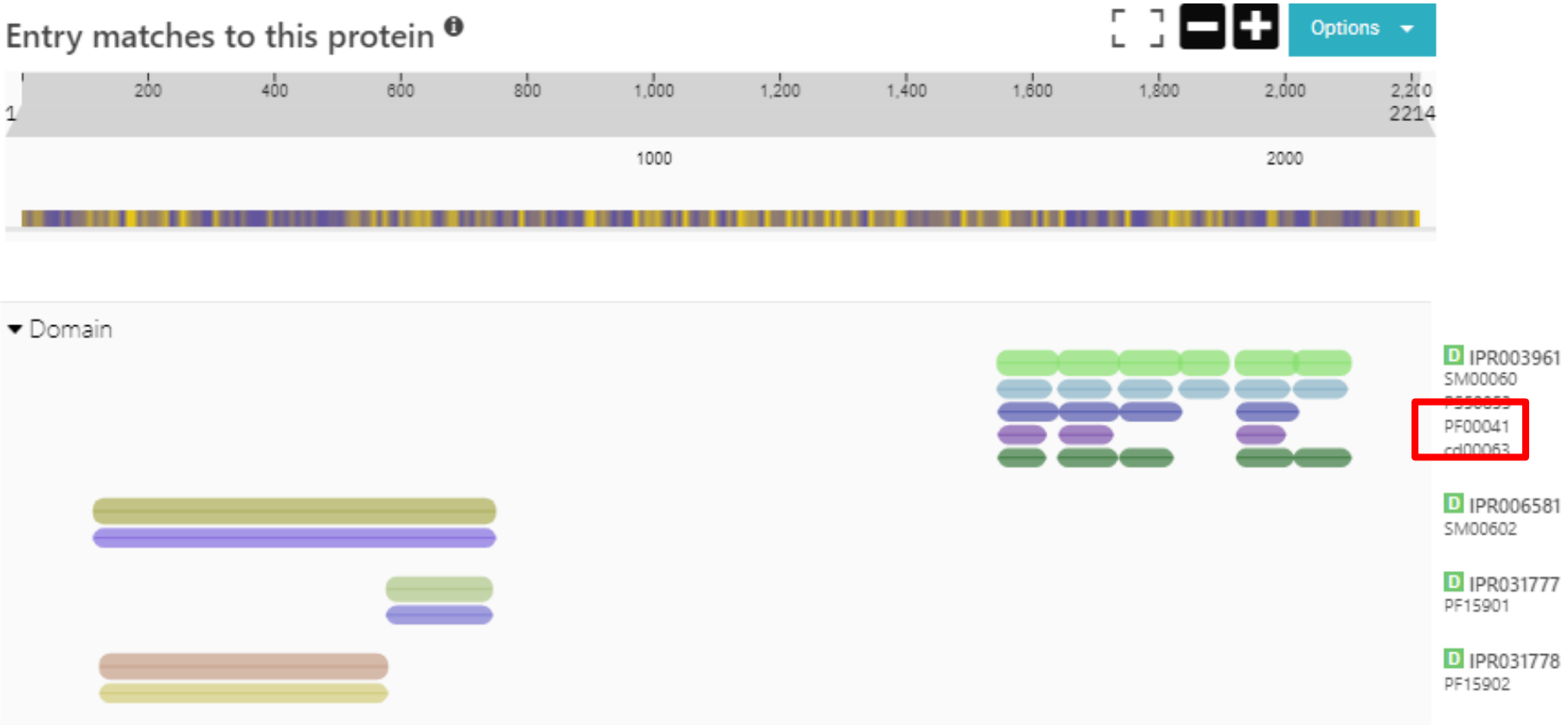
https://www.ebi.ac.uk/interpro/protein/reviewed/Q92673/



InterPro

SORLA/SORL1 from *Homo sapiens*

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



InterPro

/ Browse / By Entry / Pfam / PF00041 / Overview

Pfam Fibronectin type III domain PF00041

Pfam entry

- Overview
- Proteins 251k
- Domain Architectures 18k
- Taxonomy 21k
- Proteomes 5k
- Structures 303
- Signature
- AlphaFold 126k
- Alignment
- Curation

Member database	Pfam
Pfam type	Domain
Short name	*fn3*
Set	cl0159

Add your annotation

Integrated to

> IPR003961

References

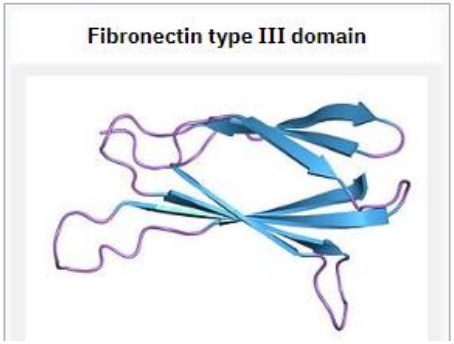
[1]. Primary structure of human fibronectin: differential splicing may generate at least 10 polypeptides from a single gene. Kornblihtt AR, Umezawa K, Vibe-Pedersen K, Baralle FE. *EMBO J.* 4, 1755-9, (1985). View article PMID: 2992939

[2]. Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. Little E, Bork P, Doolittle RF. *J. Mol. Evol.* 39, 631-43, (1994). View article PMID: 7528812

[3]. Structural design and molecular evolution of a cytokine receptor superfamily. Bazan JF. *Proc. Natl. Acad. Sci. U.S.A.* 87, 6934-8, (1990). View article PMID: 2169613

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.



InterPro

- Domain Architectures 18k
- **Taxonomy** 21k
- Proteomes 5k
- Structures 303
- Signature
- AlphaFold 126k
- Alignment
- Curation

The number of species for this sunburst is 12594. 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

175913

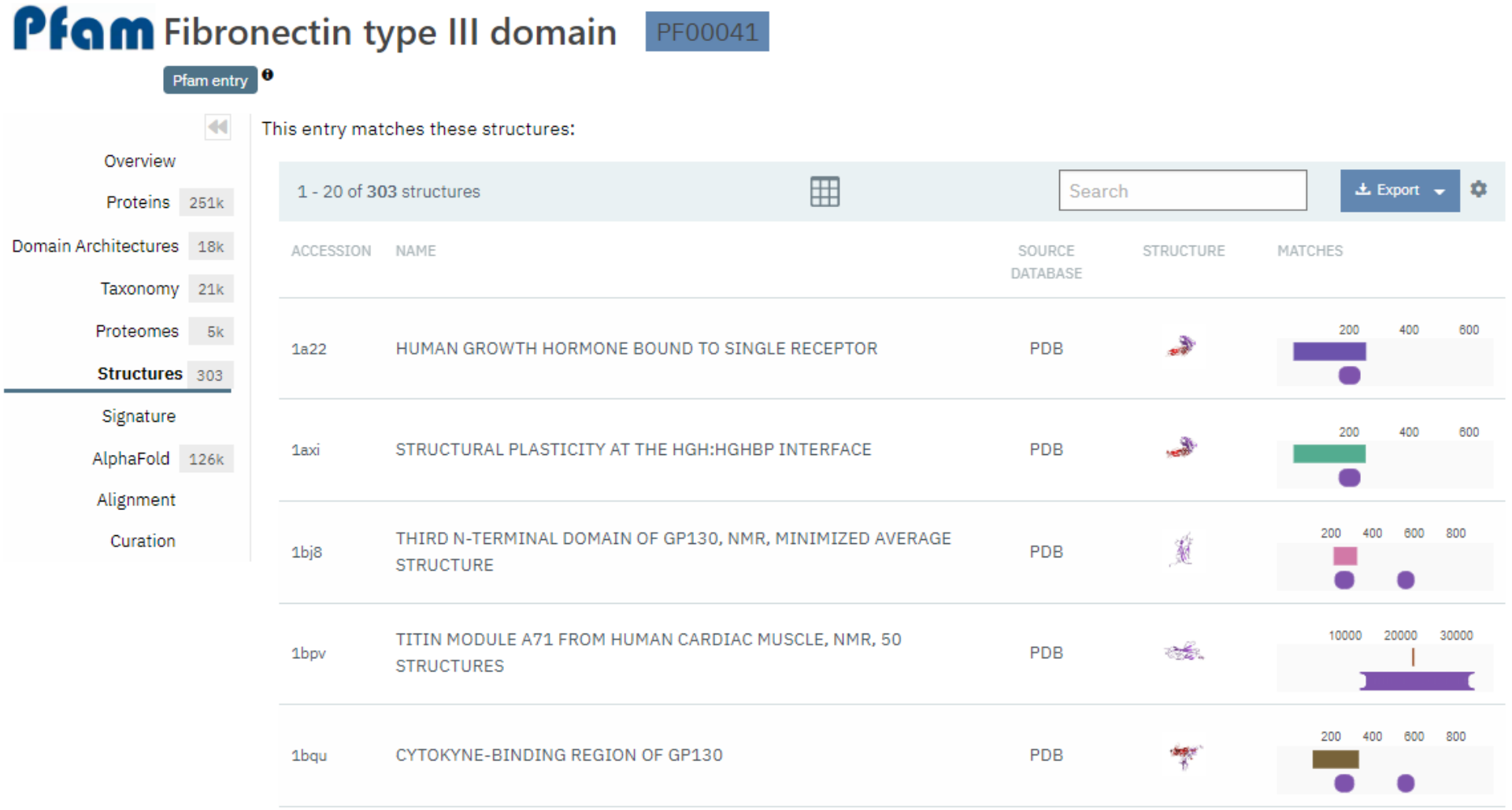
Number of species

1737

Lineage

root; *Eukaryota*; *Metazoa*; *Chordata*;

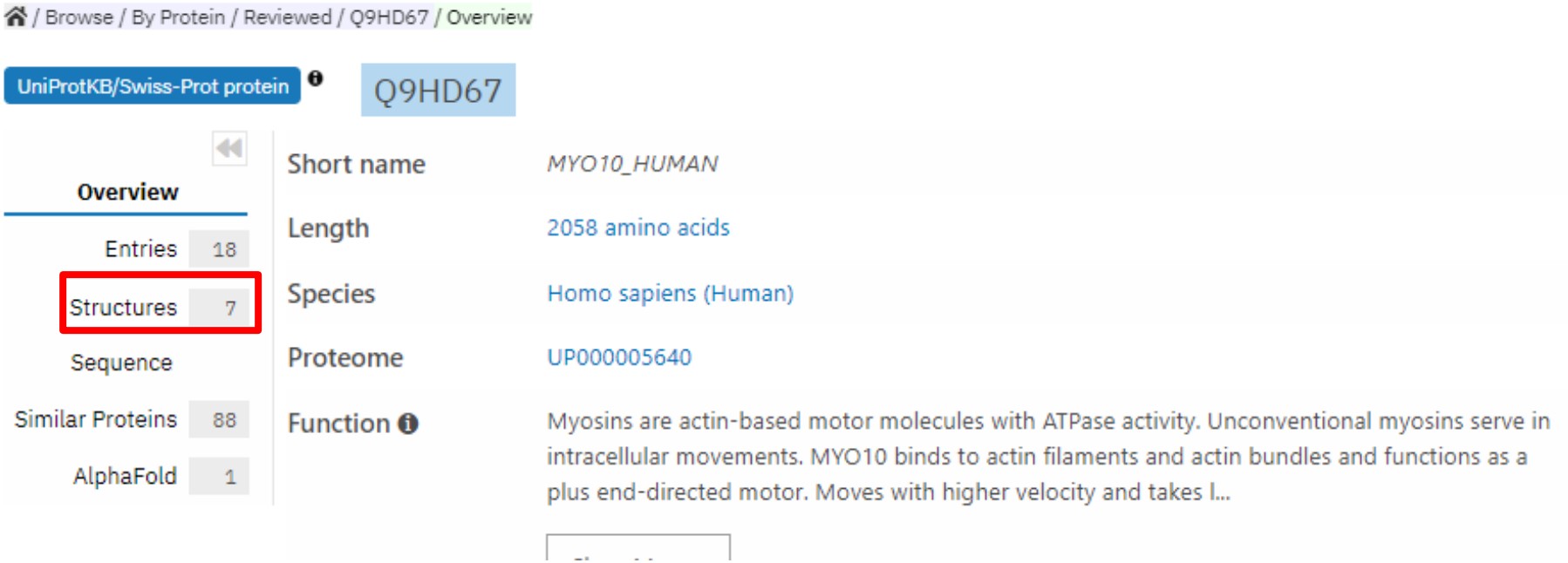
InterPro


Pfam
Fibronectin type III domain
PF00041
Pfam entry
This entry matches these structures:
Overview
Proteins 251k
Domain Architectures 18k
Taxonomy 21k
Proteomes 5k
Structures 303
Signature
AlphaFold 126k
Alignment
Curation
1 - 20 of 303 structures
Search
Export
ACCESSION
NAME
SOURCE DATABASE
STRUCTURE
MATCHES
1a22
HUMAN GROWTH HORMONE BOUND TO SINGLE RECEPTOR
PDB
200
400
600
1axi
STRUCTURAL PLASTICITY AT THE HGH:HGHBP INTERFACE
PDB
200
400
600
1bj8
THIRD N-TERMINAL DOMAIN OF GP130, NMR, MINIMIZED AVERAGE STRUCTURE
PDB
200
400
600
800
1bpv
TITIN MODULE A71 FROM HUMAN CARDIAC MUSCLE, NMR, 50 STRUCTURES
PDB
10000
20000
30000
1bqu
CYTOKYNE-BINDING REGION OF GP130
PDB
200
400
600
800

Exercise 1

Find structures in PDB for human myosin X

Find this sequence in InterPro.
https://www.ebi.ac.uk/interpro/protein/reviewed/Q9HD67/



Exercise 1

Find structures in 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

Analyse domain predictions

Slide down to see the details of the structures.

This protein matches these structures:

1 - 7 of 7 structures

Search

Export

ACCESSION	NAME	SOURCE DATABASE	STRUCTURE	MATCHES
2lw9	NMR solution structure of Myo10 anti-CC	PDB		500 1000 1500 2000
3au4	Structure of the human myosin-X MyTH4-FERM cassette bound to its specific cargo, DCC	PDB		500 1000 1500 2000
3au5	Structure of the human myosin-X MyTH4-FERM cassette	PDB		500 1000 1500 2000
3pzd	Structure of the myosin X MyTH4-FERM/DCC complex	PDB		500 1000 1500 2000
5i0h	Crystal structure of myosin X motor domain in pre-powerstroke state	PDB		500 1000 1500 2000

Exercise 2

Analyse domain predictions

- Examine the structure of 3pzd

How do the domain predictions fit the structure?

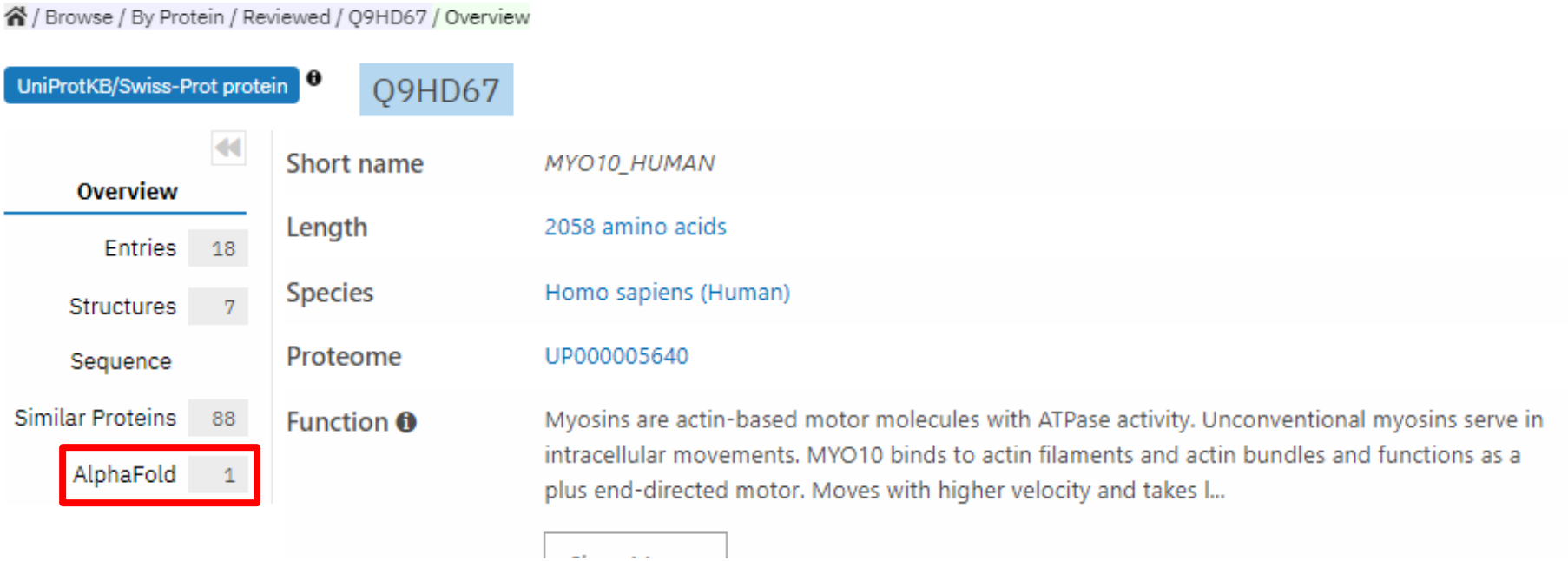
- Chain B in this structure is a small peptide.

Which domain in Myosin X is interacting with this peptide?

Exercise 3

AlphaFold prediction

- There is a predicted structure



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), 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