JOHANNES GUTENBERG UNIVERSITÄT MAINZ

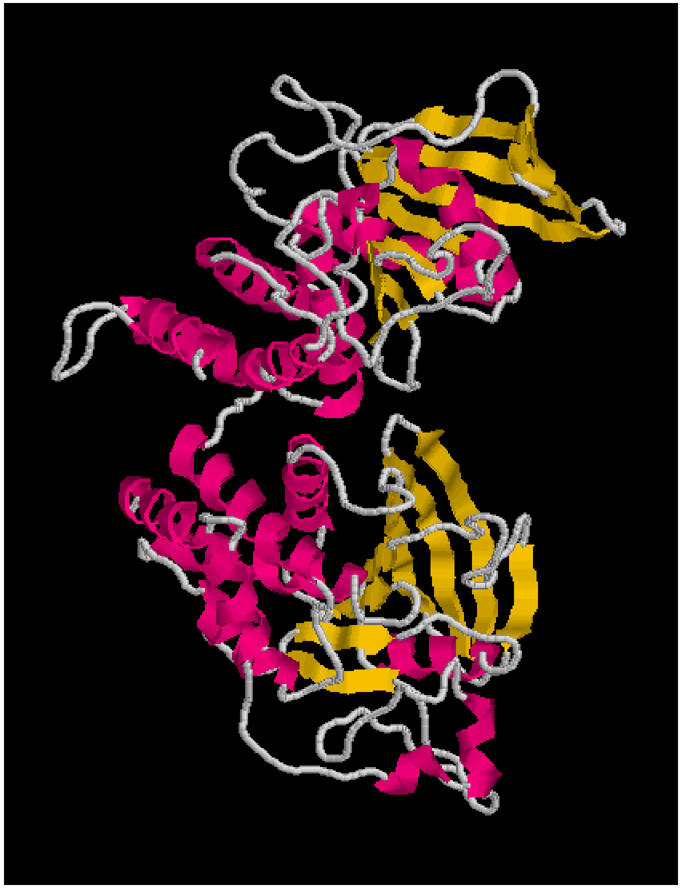
Protein domains

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

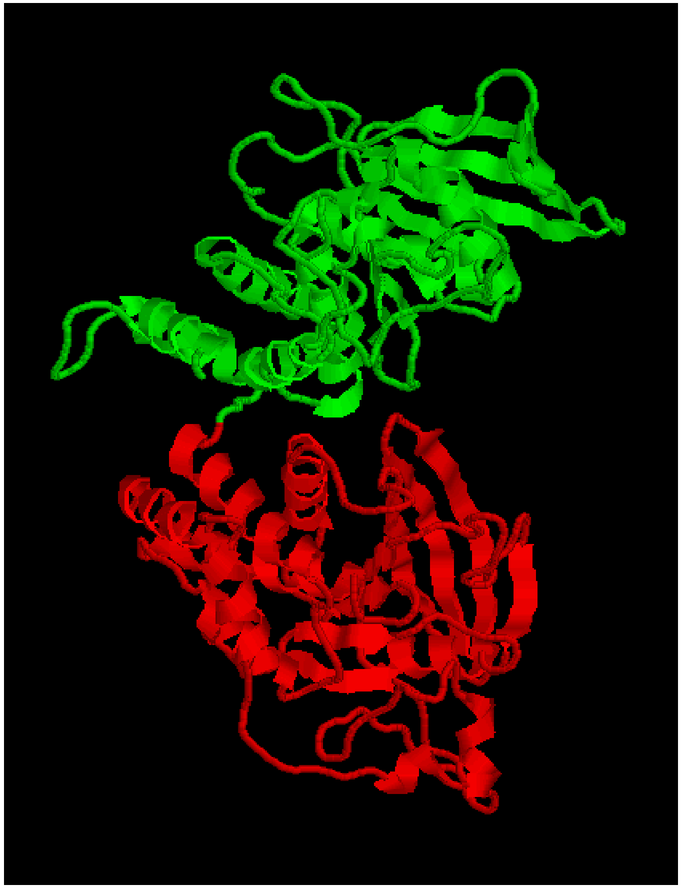


Proteins normally are multidomain (average 300 aa)

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.

Experiments used to gain insight into the function of a protein might work better at the domain level.

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.

Domain databases SMART

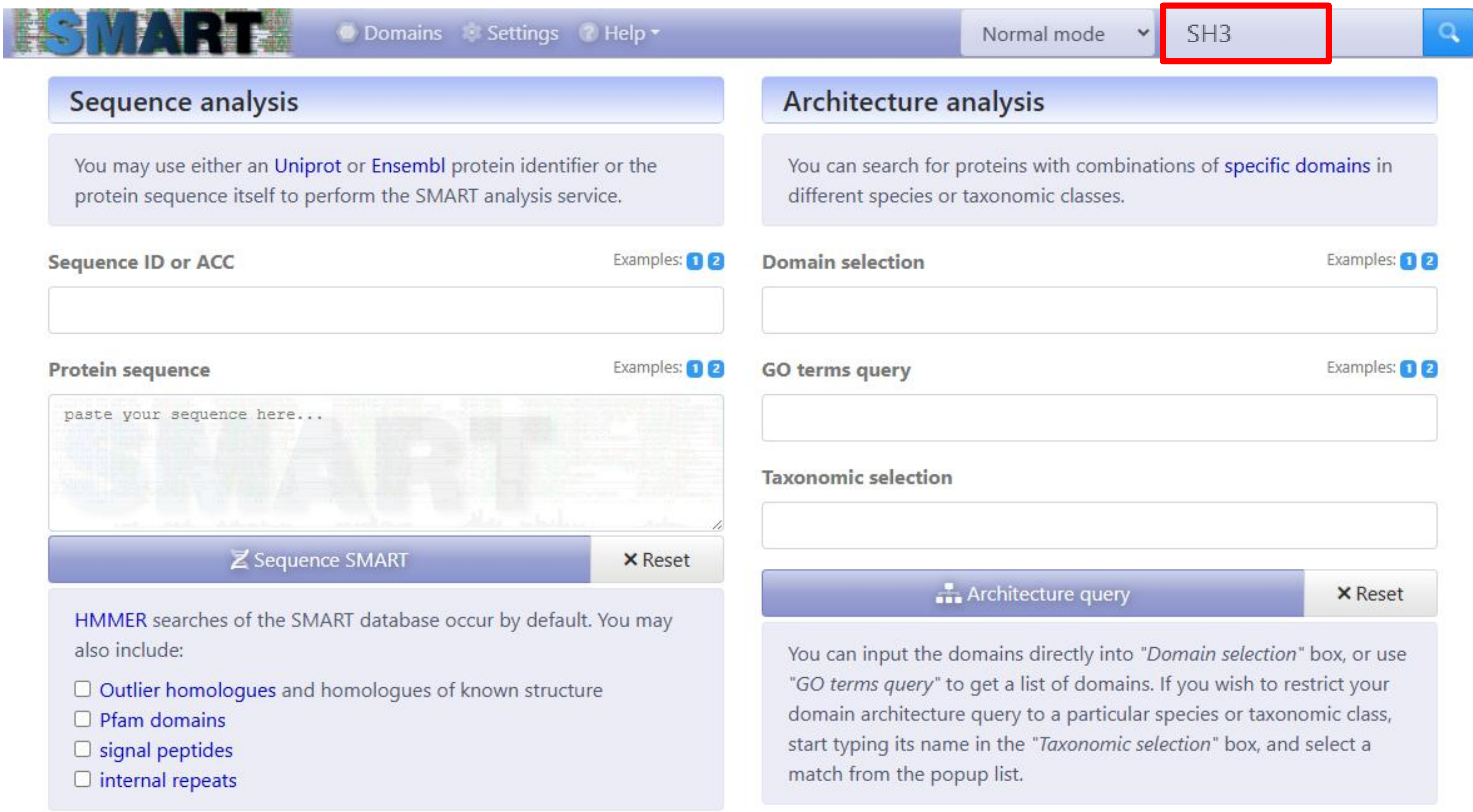
Peer Bork http://smart.embl.de/

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 (2020) *Nucleic Acids Research*

Domain databases



Sequence analysis

You may use either an Uniprot or Ensembl protein identifier or the protein sequence itself to perform the SMART analysis service.

Sequence ID or ACC Examples: 1 2

Protein sequence Examples: 1 2

paste your sequence here...

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

- ☐ Outlier homologues and homologues of known structure
- ☐ Pfam domains
- ☐ signal peptides
- ☐ internal repeats

Architecture analysis

You can search for proteins with combinations of specific domains in different species or taxonomic classes.

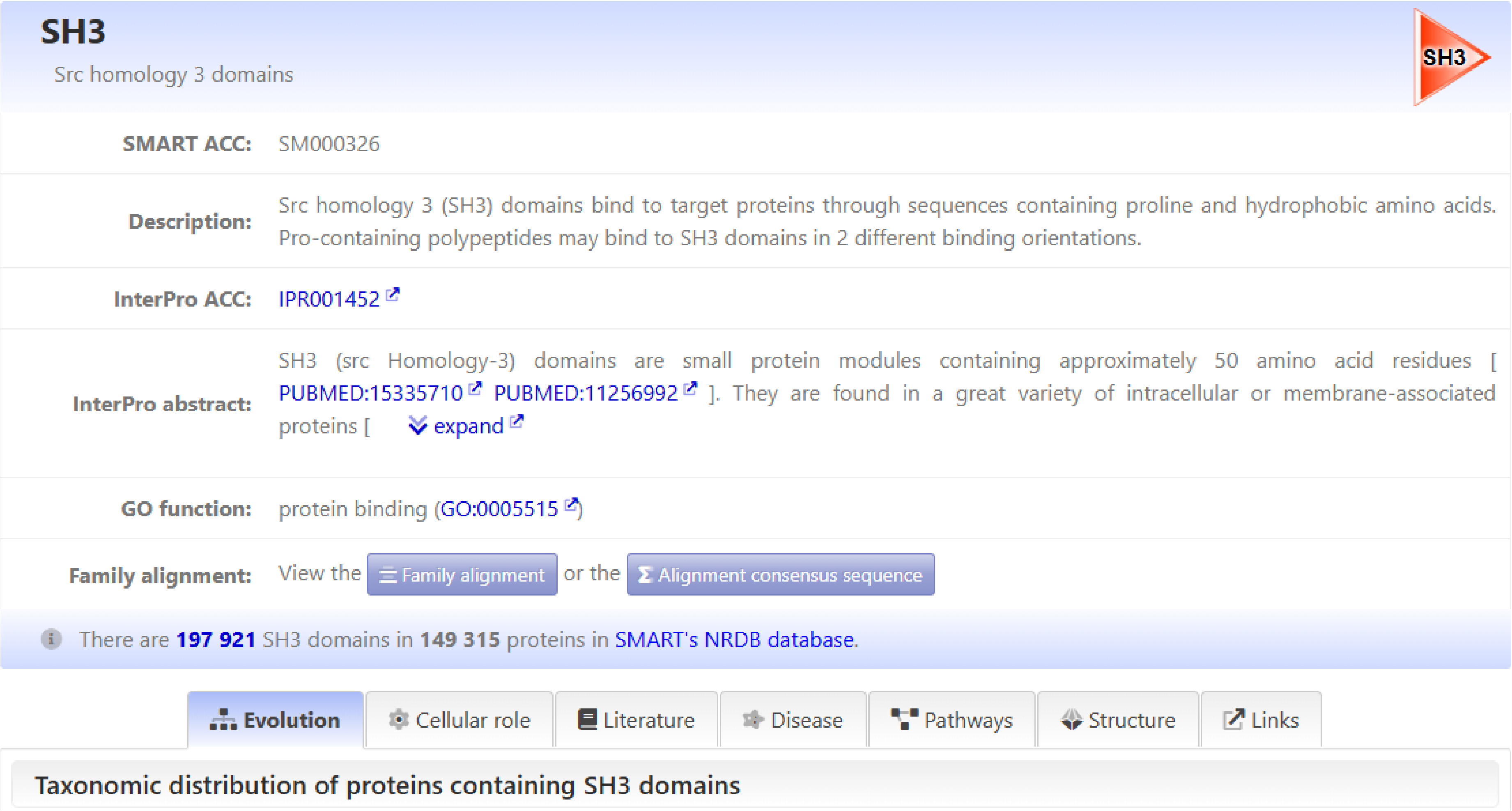
Domain selection Examples: 1 2

GO terms query Examples: 1 2

Taxonomic selection

You can input the domains directly into "*Domain selection*" box, or use "*GO terms query*" to get a list of domains. If you wish to restrict your domain architecture query to a particular species or taxonomic class, start typing its name in the "*Taxonomic selection*" box, and select a match from the popup list.

Domain databases
SMART



SH3

Src homology 3 domains

SMART ACC: SM000326

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 ACC: IPR001452

InterPro abstract: 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 [expand

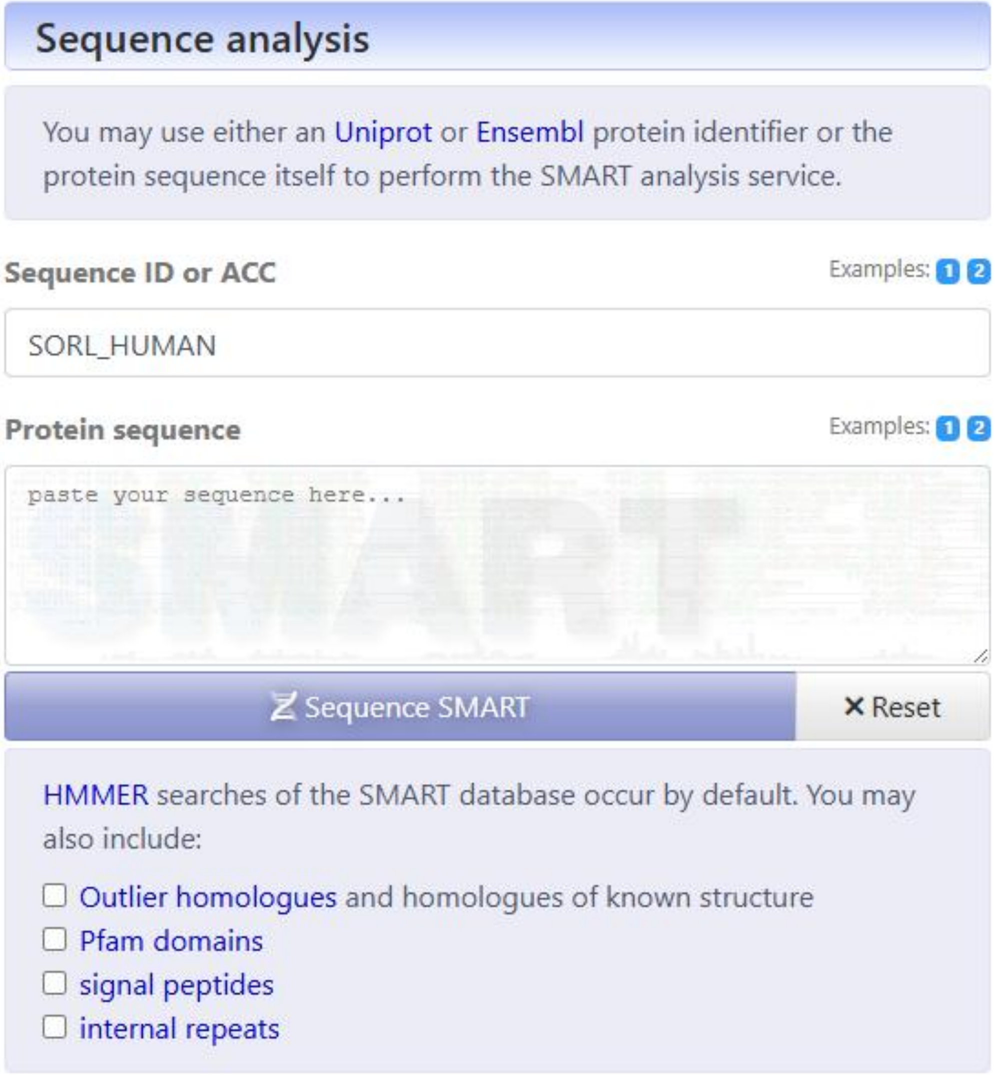
GO function: protein binding (GO:0005515)

Family alignment: View the Family alignment or the Alignment consensus sequence

There are **197 921** SH3 domains in **149 315** proteins in SMART's NRDB database.

Taxonomic distribution of proteins containing SH3 domains

Domain databases
SMART



Sequence analysis

You may use either an Uniprot or Ensembl protein identifier or the protein sequence itself to perform the SMART analysis service.

Sequence ID or ACC: SORL_HUMAN

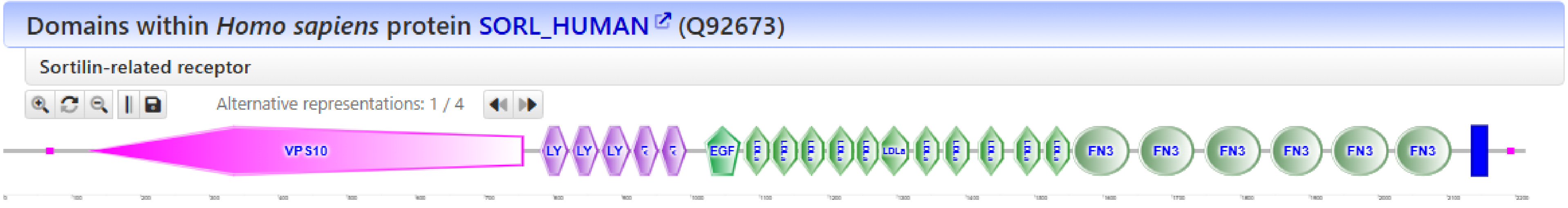
Protein sequence: paste your sequence here...

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

- □ Outlier homologues and homologues of known structure
- □ Pfam domains
- □ signal peptides
- □ internal repeats

Domain databases
SMART

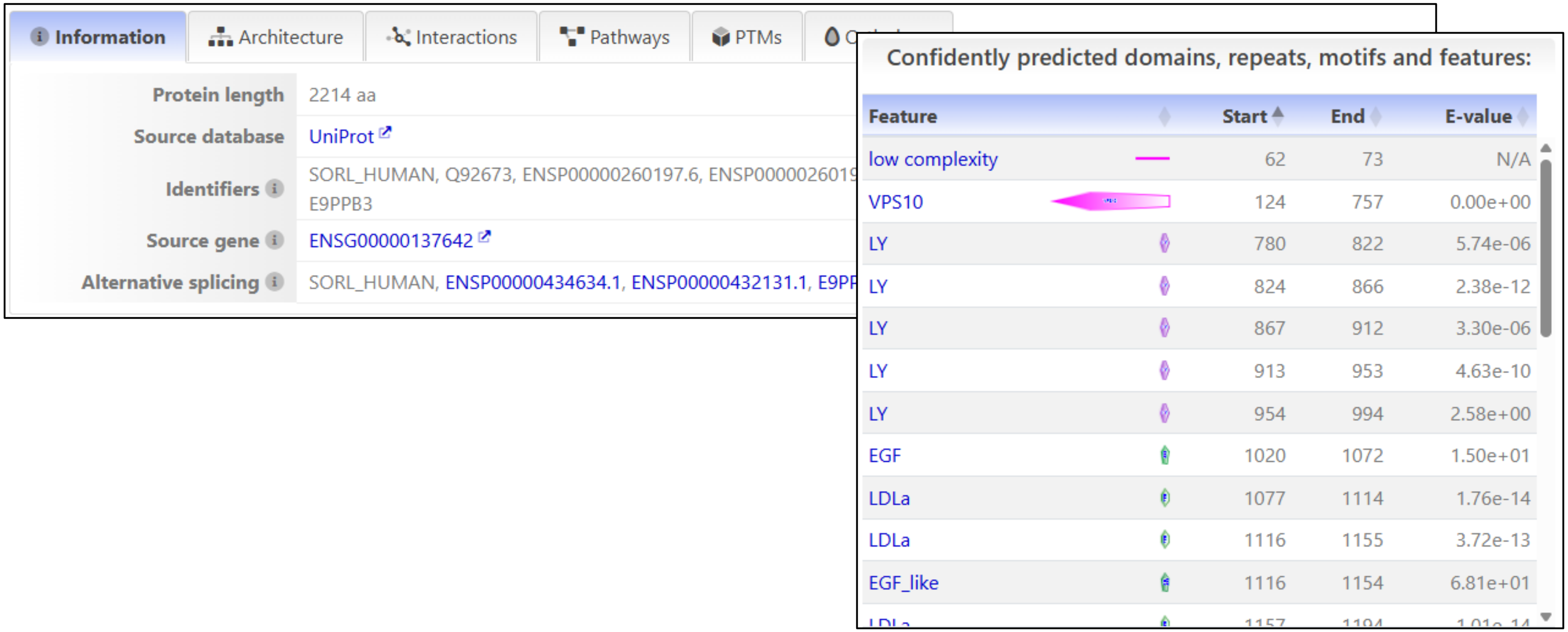
Extra features:
low complexity, TM, coiled coils



Domains within *Homo sapiens* protein SORL_HUMAN (Q92673)

Sortilin-related receptor

Alternative representations: 1 / 4



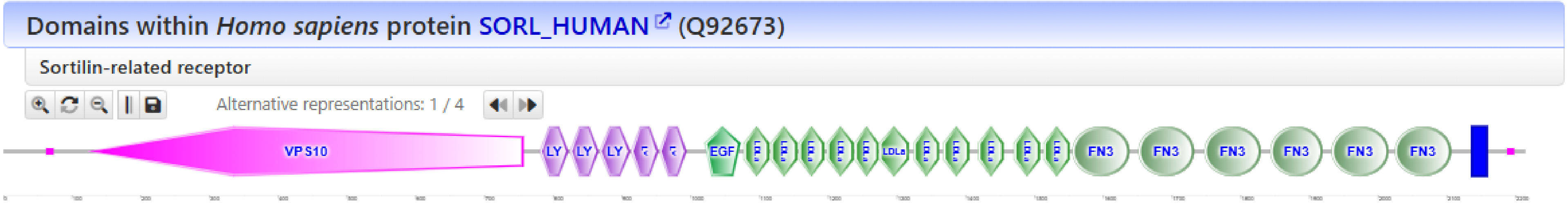
Protein length	2214 aa
Source database	UniProt
Identifiers	SORL_HUMAN, Q92673, ENSP00000260197.6, ENSP0000026019 E9PPB3
Source gene	ENSG00000137642
Alternative splicing	SORL_HUMAN, ENSP00000434634.1, ENSP00000432131.1, E9PP

Confidently predicted domains, repeats, motifs and features:

Feature	Start	End	E-value
low complexity	62	73	N/A
VPS10	124	757	0.00e+00
LY	780	822	5.74e-06
LY	824	866	2.38e-12
LY	867	912	3.30e-06
LY	913	953	4.63e-10
LY	954	994	2.58e+00
EGF	1020	1072	1.50e+01
LDLa	1077	1114	1.76e-14
LDLa	1116	1155	3.72e-13
EGF_like	1116	1154	6.81e+01
LDLa	1157	1194	1.01e-14

Domain databases
SMART

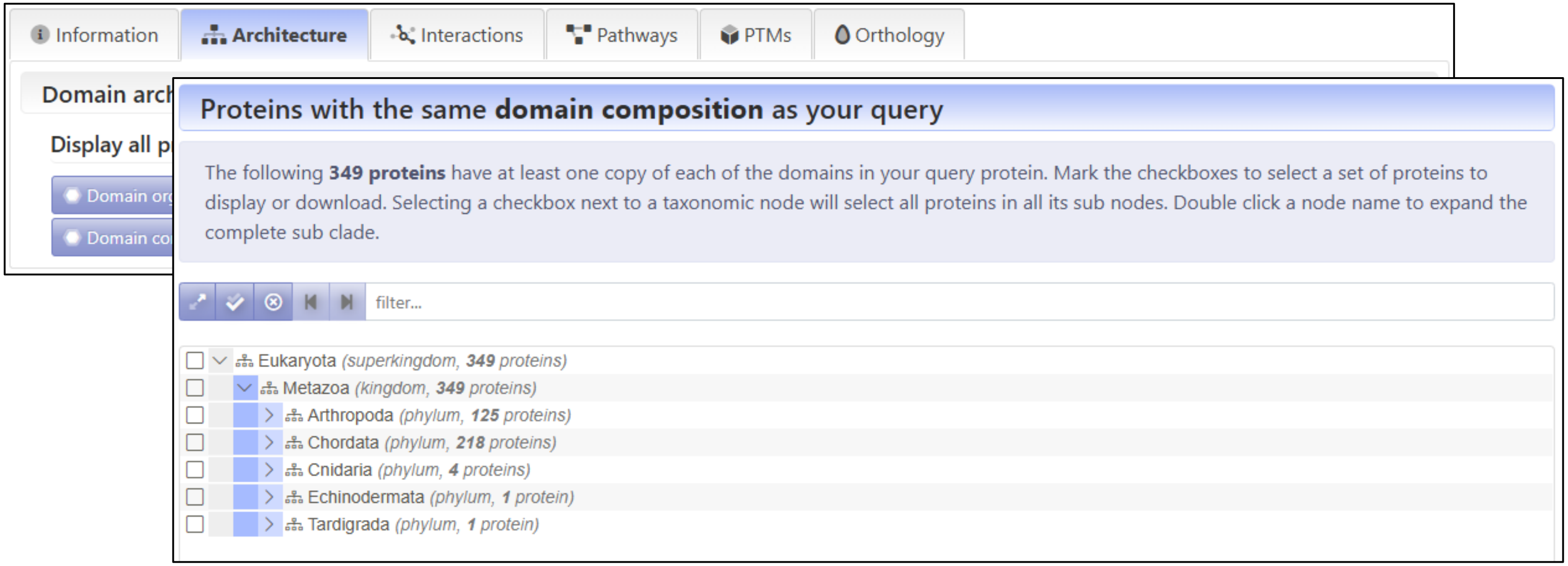
Extra features:
low complexity, TM, coiled coils



Domains within *Homo sapiens* protein SORL_HUMAN (Q92673)

Sortilin-related receptor

Alternative representations: 1 / 4



Proteins with the same **domain composition** as your query

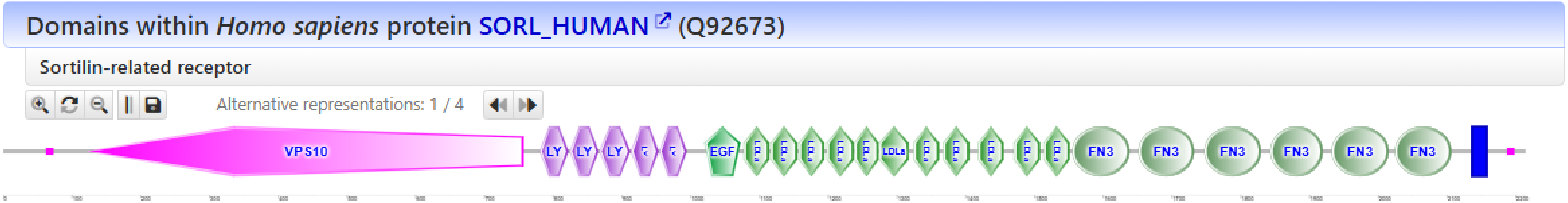
The following **349 proteins** have at least one copy of each of the domains in your query protein. Mark the checkboxes to select a set of proteins to display or download. Selecting a checkbox next to a taxonomic node will select all proteins in all its sub nodes. Double click a node name to expand the complete sub clade.

filter...

- □ Eukaryota *(superkingdom, **349** proteins)*
  - □ Metazoa *(kingdom, **349** proteins)*
    - □ Arthropoda *(phylum, **125** proteins)*
    - □ Chordata *(phylum, **218** proteins)*
    - □ Cnidaria *(phylum, **4** proteins)*
    - □ Echinodermata *(phylum, **1** protein)*
    - □ Tardigrada *(phylum, **1** protein)*

Domain databases
SMART

Extra features:
low complexity, TM, coiled coils



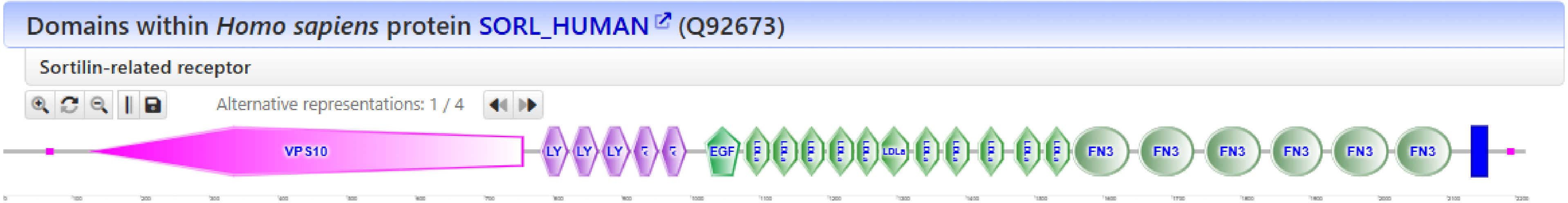
Domains within *Homo sapiens* protein SORL_HUMAN (Q92673)

Sortilin-related receptor

Alternative representations: 1 / 4

Domain databases SMART

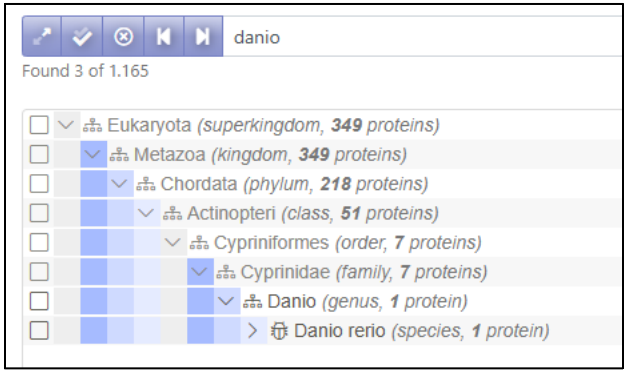
Extra features:
low complexity, TM, coiled coils



Domains within *Homo sapiens* protein SORL_HUMAN (Q92673)

Sortilin-related receptor

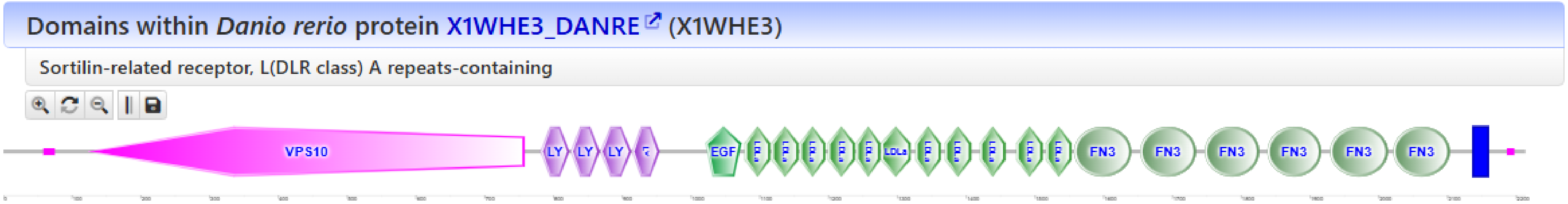
Alternative representations: 1 / 4



danio

Found 3 of 1.165

- □ Eukaryota *(superkingdom, **349** proteins)*
- □ Metazoa *(kingdom, **349** proteins)*
- □ Chordata *(phylum, **218** proteins)*
- □ Actinopteri *(class, **51** proteins)*
- □ Cypriniformes *(order, **7** proteins)*
- □ Cyprinidae *(family, **7** proteins)*
- □ Danio *(genus, **1** protein)*
- □ Danio rerio *(species, **1** protein)*

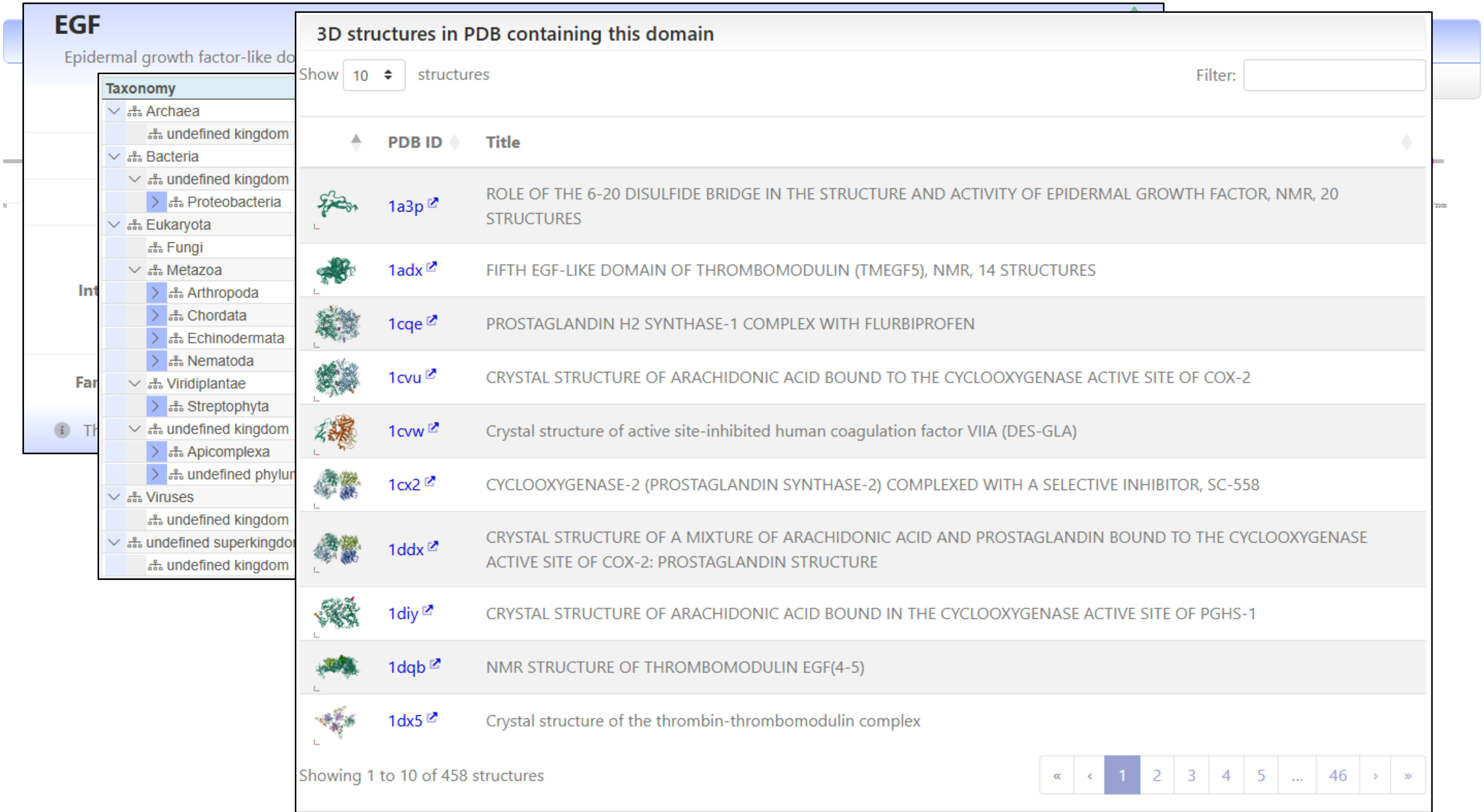


Domains within *Danio rerio* protein X1WHE3_DANRE (X1WHE3)

Sortilin-related receptor, L(DLR class) A repeats-containing

Domain databases
SMART

Extra features:
low complexity, TM, coiled coils



EGF

Epidermal growth factor-like do

Int

Far

Th

Taxonomy

- Archaea
  - undefined kingdom
- Bacteria
  - undefined kingdom
    - Proteobacteria
- Eukaryota
  - Fungi
  - Metazoa
    - Arthropoda
    - Chordata
    - Echinodermata
    - Nematoda
  - Viridiplantae
    - Streptophyta
  - undefined kingdom
    - Apicomplexa
    - undefined phylum
- Viruses
  - undefined kingdom
- undefined superkingdom
  - undefined kingdom

3D structures in PDB containing this domain

Show 10 structures

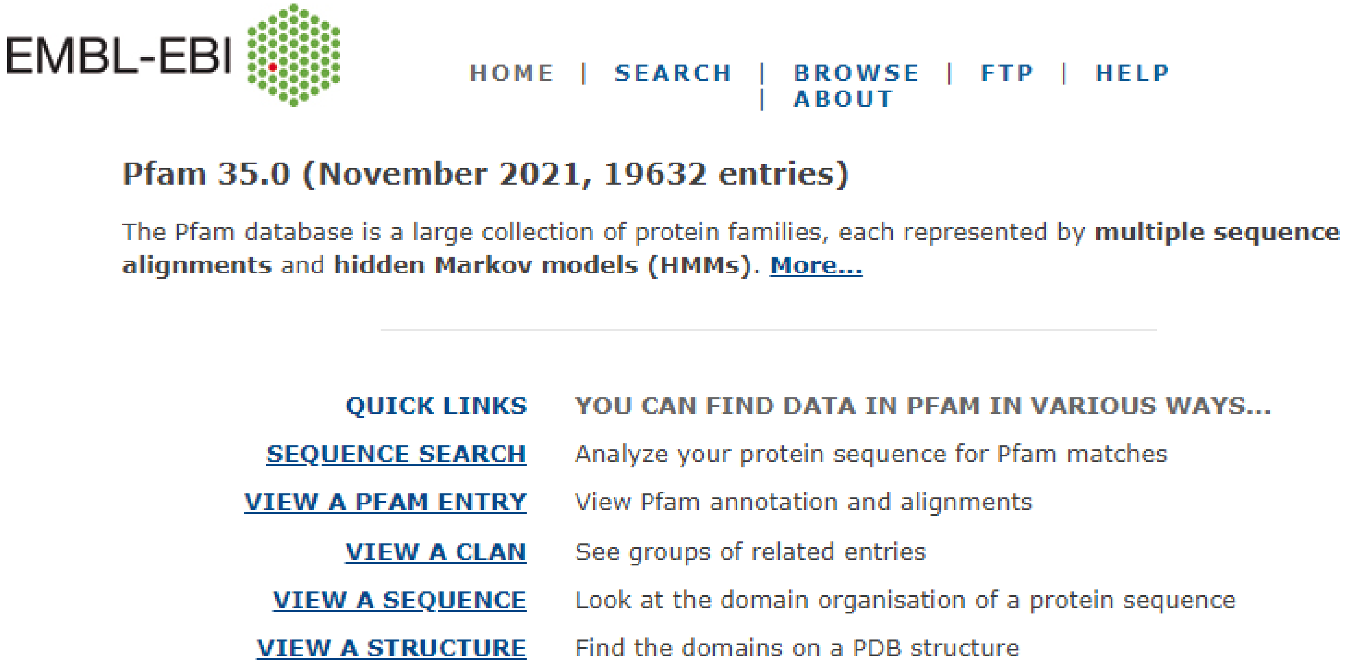
Filter:

	PDB ID	Title
	1a3p	ROLE OF THE 6-20 DISULFIDE BRIDGE IN THE STRUCTURE AND ACTIVITY OF EPIDERMAL GROWTH FACTOR, NMR, 20 STRUCTURES
	1adx	FIFTH EGF-LIKE DOMAIN OF THROMBOMODULIN (TMEGF5), NMR, 14 STRUCTURES
	1cqe	PROSTAGLANDIN H2 SYNTHASE-1 COMPLEX WITH FLURBIPROFEN
	1cvu	CRYSTAL STRUCTURE OF ARACHIDONIC ACID BOUND TO THE CYCLOOXYGENASE ACTIVE SITE OF COX-2
	1cvw	Crystal structure of active site-inhibited human coagulation factor VIIA (DES-GLA)
	1cx2	CYCLOOXYGENASE-2 (PROSTAGLANDIN SYNTHASE-2) COMPLEXED WITH A SELECTIVE INHIBITOR, SC-558
	1ddx	CRYSTAL STRUCTURE OF A MIXTURE OF ARACHIDONIC ACID AND PROSTAGLANDIN BOUND TO THE CYCLOOXYGENASE ACTIVE SITE OF COX-2: PROSTAGLANDIN STRUCTURE
	1diy	CRYSTAL STRUCTURE OF ARACHIDONIC ACID BOUND IN THE CYCLOOXYGENASE ACTIVE SITE OF PGHS-1
	1dqb	NMR STRUCTURE OF THROMBOMODULIN EGF(4-5)
	1dx5	Crystal structure of the thrombin-thrombomodulin complex

Showing 1 to 10 of 458 structures

Domain databases
PFAM (until Jan 2023)

Erik Sonnhammer/Ewan Birney/Alex Bateman
http://pfam.xfam.org/



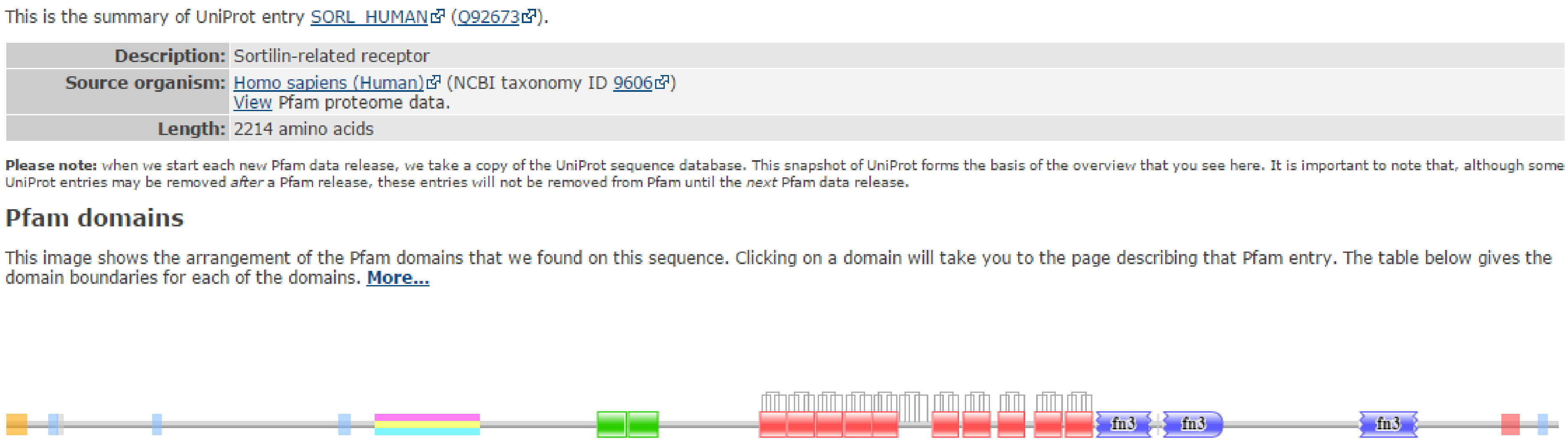
Pfam 35.0 (November 2021, 19632 entries)

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

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure

Sonnhammer et al (1997) *Proteins*
...
Mistry et al (2021) *Nucleic Acids Research*

Domain databases
PFAM



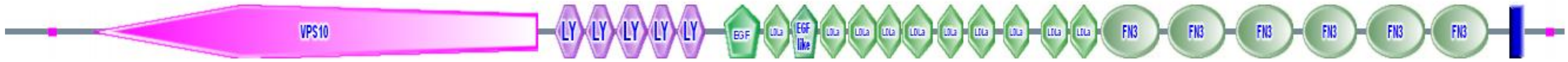
This is the summary of UniProt entry SORL_HUMAN (Q92673).

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.

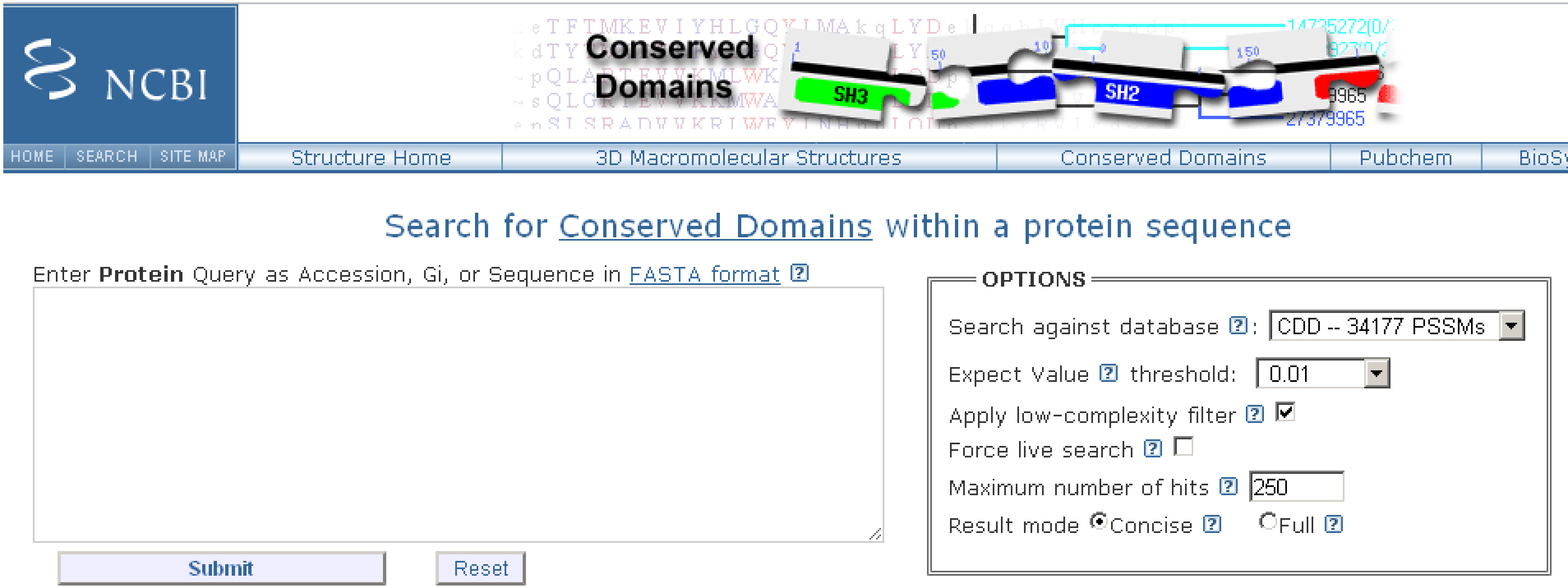
Pfam domains

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



Search for Conserved Domains within a protein sequence

Enter **Protein** Query as Accession, Gi, or Sequence in FASTA format ?

OPTIONS

Search against database ?: CDD -- 34177 PSSMs

Expect Value ? threshold: 0.01

Apply low-complexity filter ? ☑

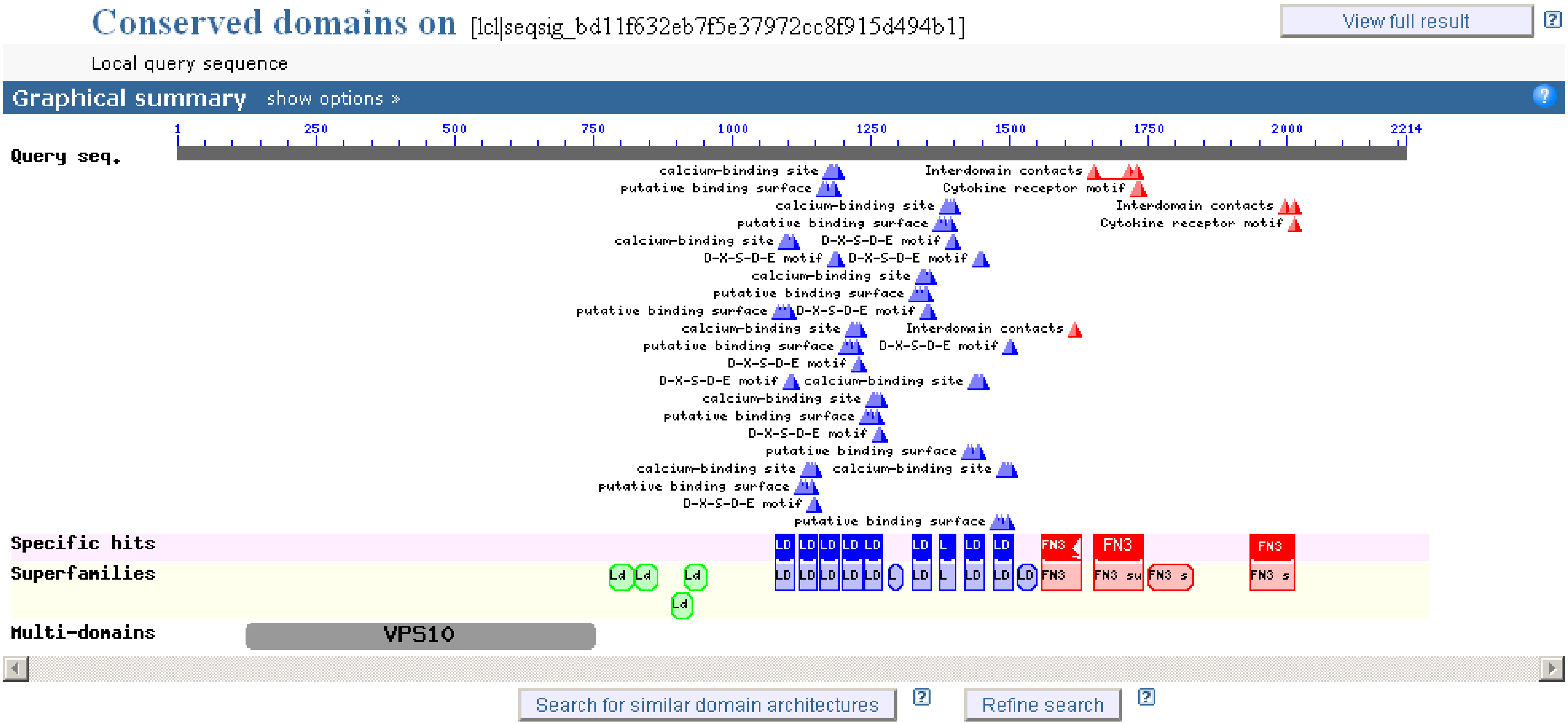
Force live search ? □

Maximum number of hits ? 250

Result mode ● Concise ? ○ Full ?

Wang et al (2022) *Nucleic Acids Res*

Domain databases
CDD



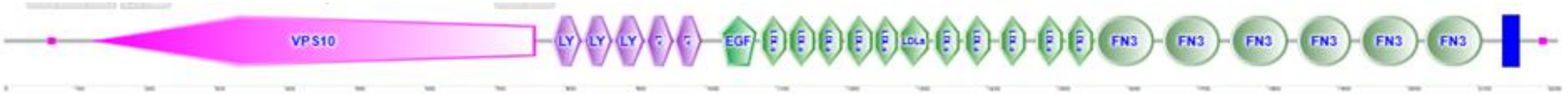
Conserved domains on [lcl|seqsig_bd11f632eb7f5e37972cc8f915d494b1]

Local query sequence

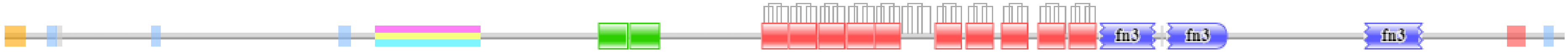
Domain databases

SORLA/SORL1 from *Homo sapiens*

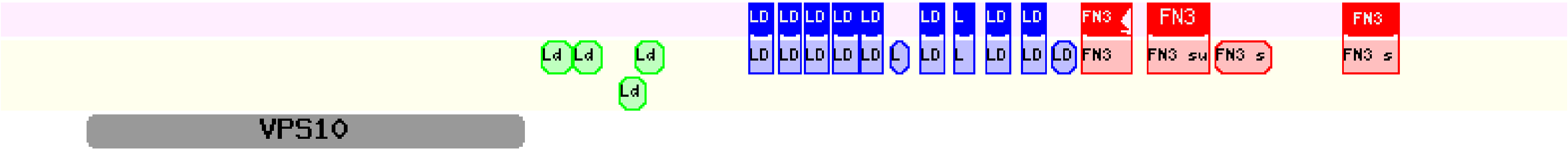
SMART



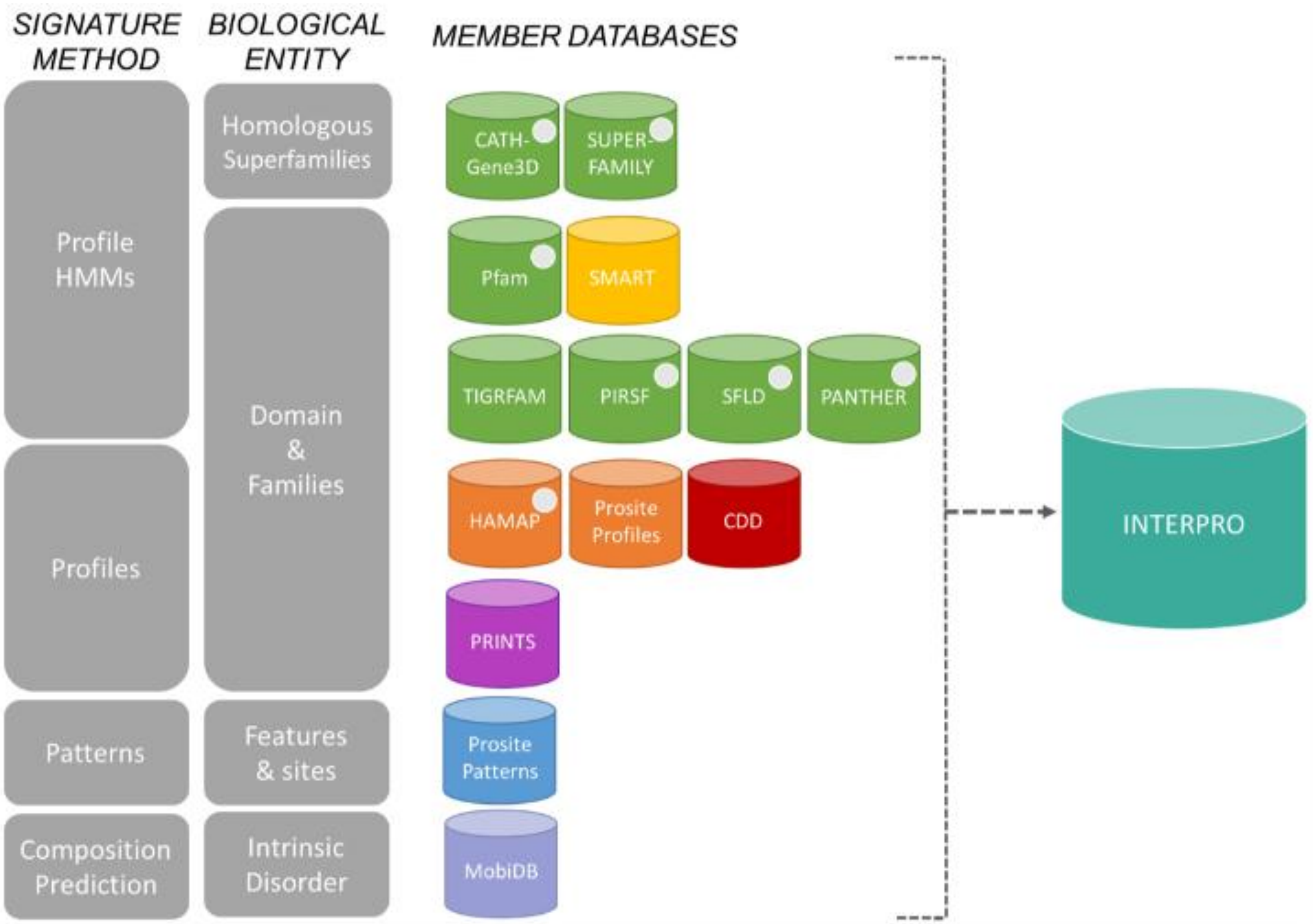
PFAM



CDD



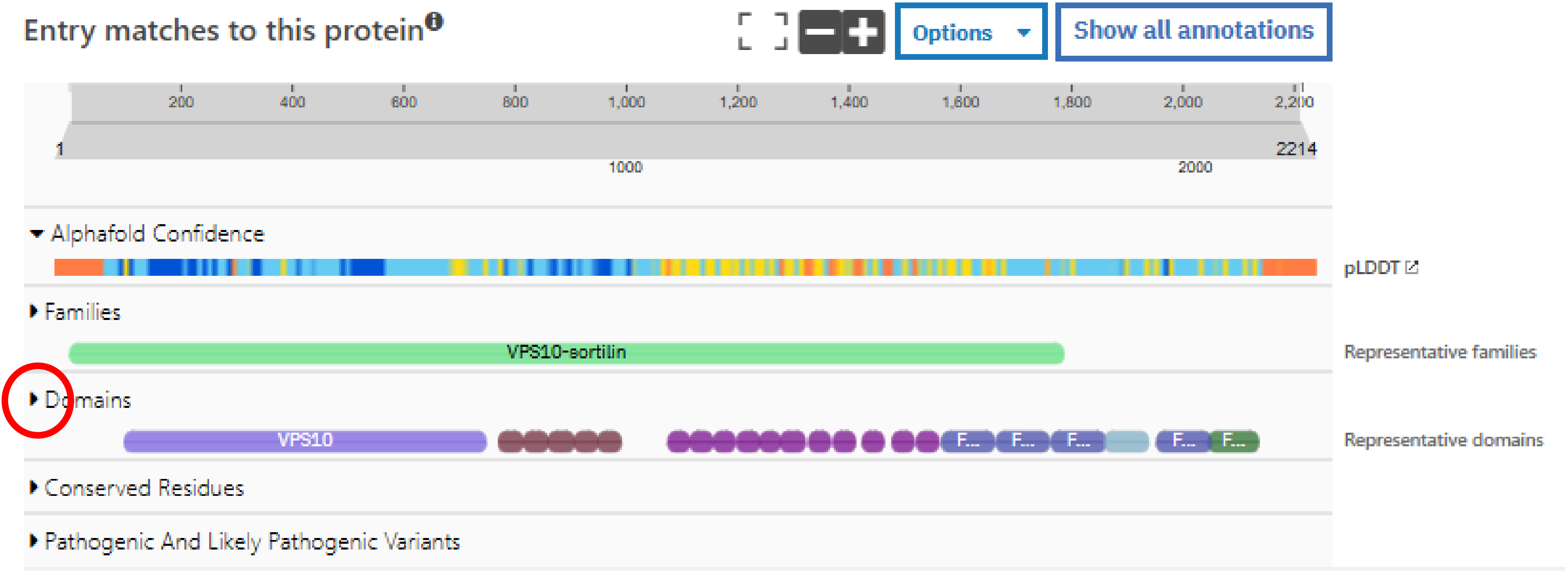
InterPro

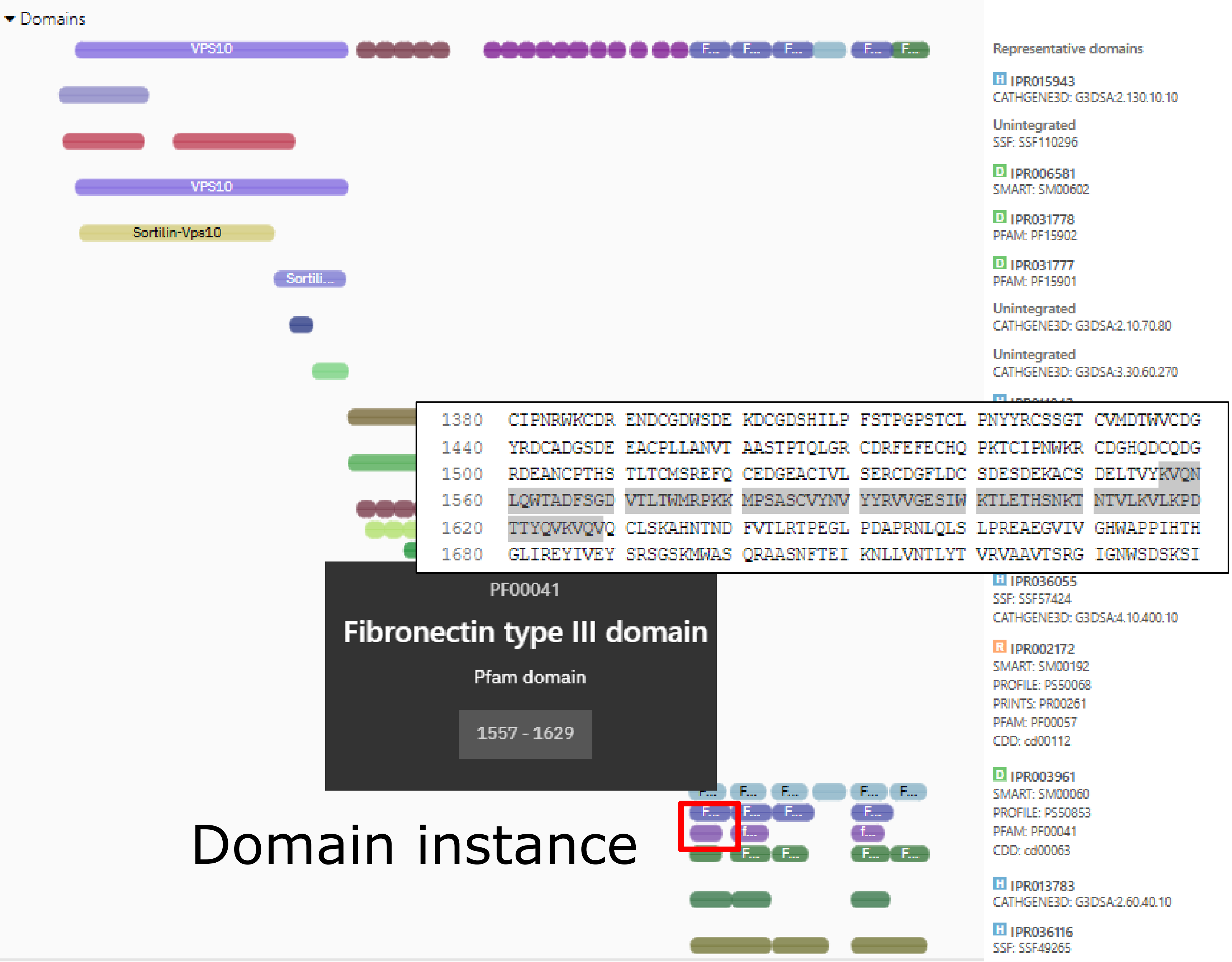

SIGNATURE METHOD
BIOLOGICAL ENTITY
MEMBER DATABASES
Profile HMMs
Profiles
Patterns
Composition Prediction
Homologous Superfamilies
Domain & Families
Features & sites
Intrinsic Disorder
CATH-Gene3D
SUPER-FAMILY
Pfam
SMART
TIGRFAM
PIRSF
SFLD
PANTHER
HAMAP
Prosite Profiles
CDD
PRINTS
Prosite Patterns
MobiDB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

InterPro

SORLA/SORL1 from *Homo sapiens*

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





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

IPR036055
SSF: SSF57424
CATHGENE3D: G3DSA:4.10.400.10

IPR002172
SMART: SM00192
PROFILE: PS50068
PRINTS: PR00261
PFAM: PF00057
CDD: cd00112

IPR003961
SMART: SM00060
PROFILE: PS50853
PFAM: PF00041
CDD: cd00063

IPR013783
CATHGENE3D: G3DSA:2.60.40.10

IPR036116
SSF: SSF49265

```
1380   CIPNRWKCDR ENDCGDWSDE KDCGDSHILP FSTPGPSTCL PNYYRCSSGT CVMDTWVCDG
1440   YRDCADGSDE EACPLLANVT AASTPTQLGR CDRFEFECHQ PKTCIPNWKR CDGHQDCQDG
1500   RDEANCPTHS TLTCMSREFQ CEDGEACIVL SERCDGFLDC SDESDEKACS DELTVYKVQN
1560   LQWTADFSGD VTLTWMRPKK MPSASCVYNV YYRVVGESIW KTLETHSNKT NTVLKVLKPD
1620   TTYQVKVQVQ CLSKAHNTND FVTLRTPEGL PDAPRNLQLS LPREAEGVIV GHWAPPIHTH
1680   GLIREYIVEY SRSGSKMWAS QRAASNFTEI KNLLVNTLYT VRVAAVTSRG IGNWSDSKSI
```

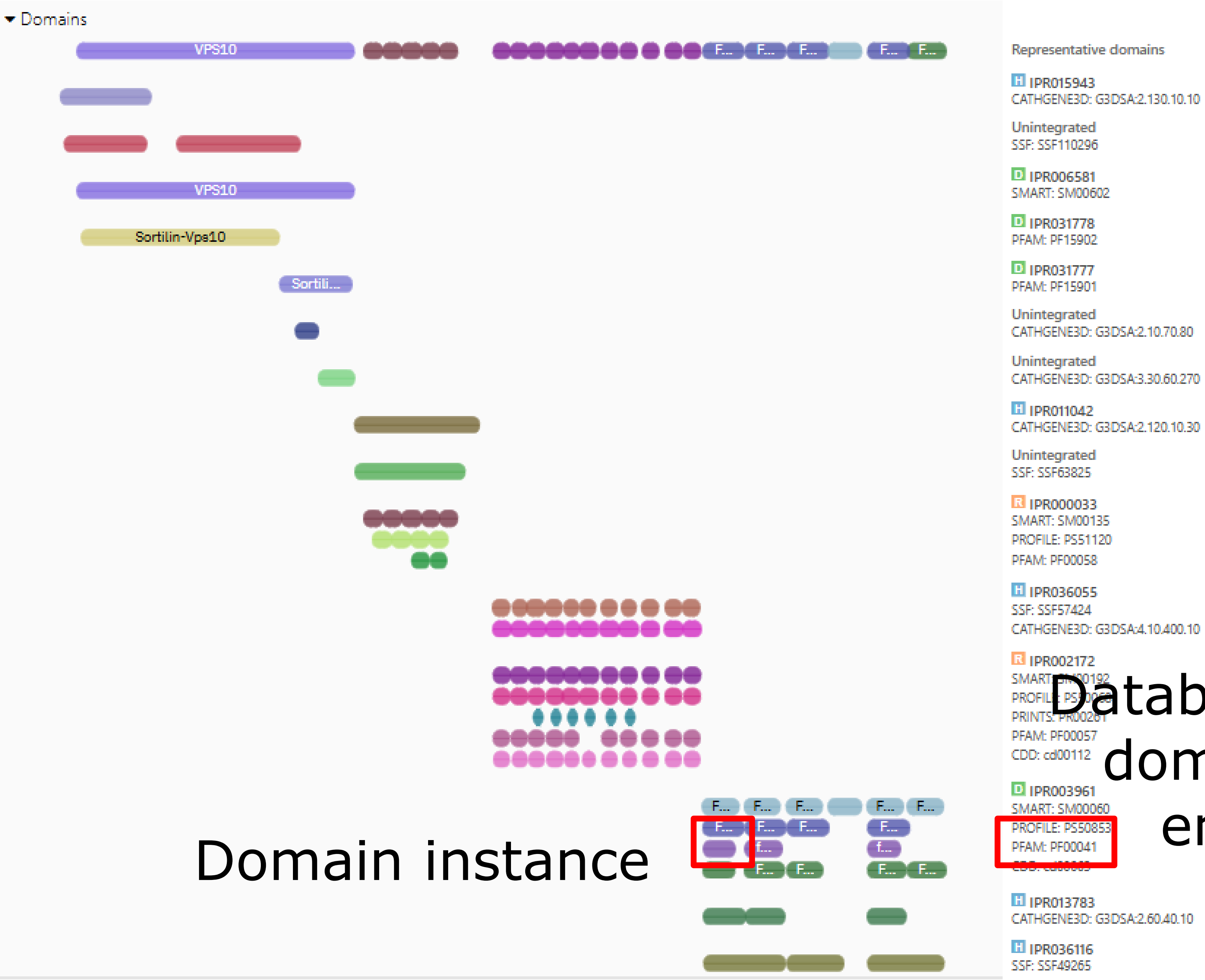
PF00041

Fibronectin type III domain

Pfam domain

1557 - 1629

Domain instance



Representative domains

H IPR015943
CATHGENE3D: G3DSA:2.130.10.10

Unintegrated
SSF: SSF110296

D IPR006581
SMART: SM00602

D IPR031778
PFAM: PF15902

D IPR031777
PFAM: PF15901

Unintegrated
CATHGENE3D: G3DSA:2.10.70.80

Unintegrated
CATHGENE3D: G3DSA:3.30.60.270

H IPR011042
CATHGENE3D: G3DSA:2.120.10.30

Unintegrated
SSF: SSF63825

R IPR000033
SMART: SM00135
PROFILE: PS51120
PFAM: PF00058

H IPR036055
SSF: SSF57424
CATHGENE3D: G3DSA:4.10.400.10

R IPR002172
SMART: SM00192
PROFILE: PS50068
PRINTS: PR00261
PFAM: PF00057
CDD: cd00112

D IPR003961
SMART: SM00060
PROFILE: PS50853
PFAM: PF00041
CDD: cd00063

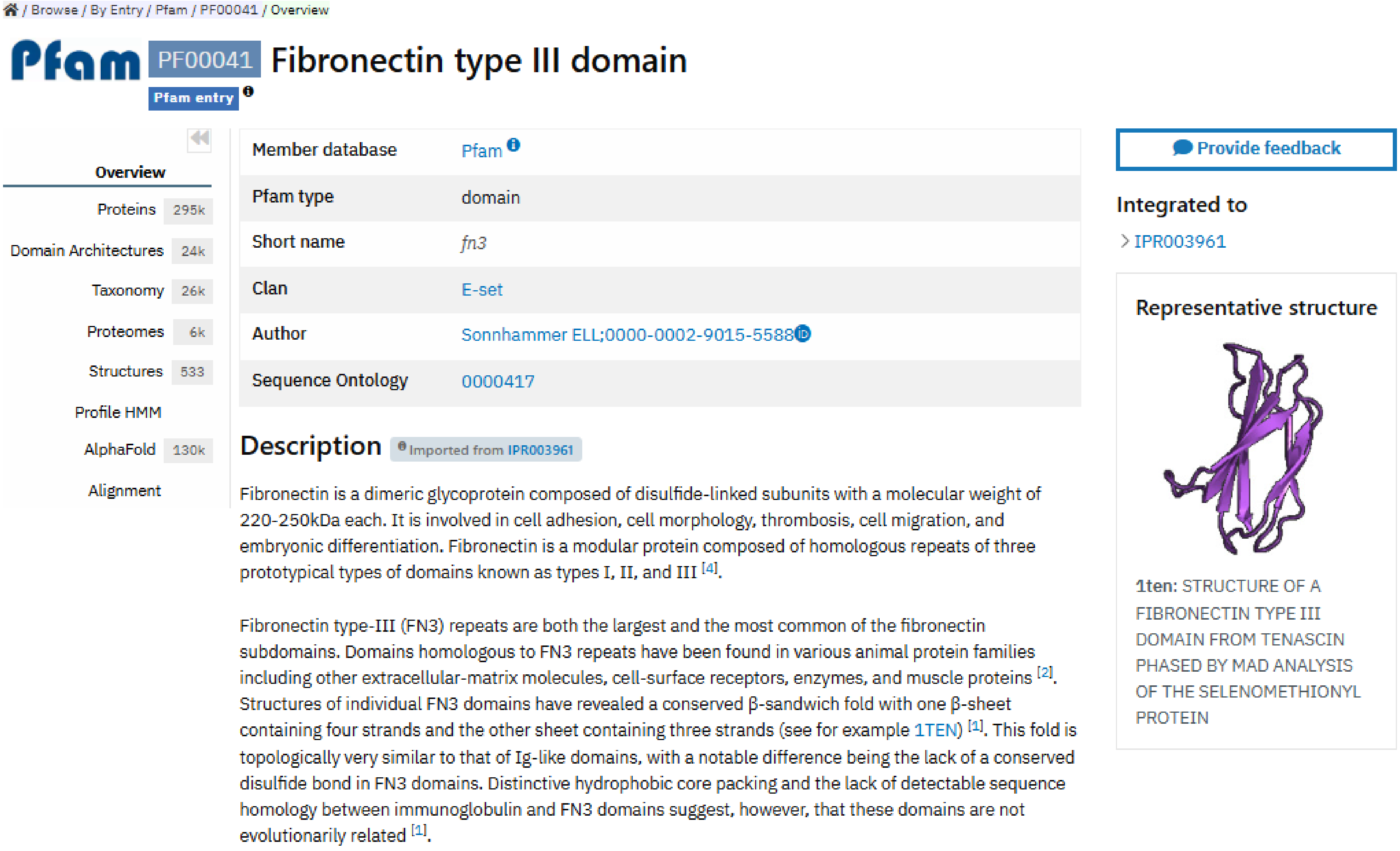
H IPR013783
CATHGENE3D: G3DSA:2.60.40.10

H IPR036116
SSF: SSF49265

Database domain entry

Domain instance

InterPro



🏠 / Browse / By Entry / Pfam / PF00041 / Overview

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].

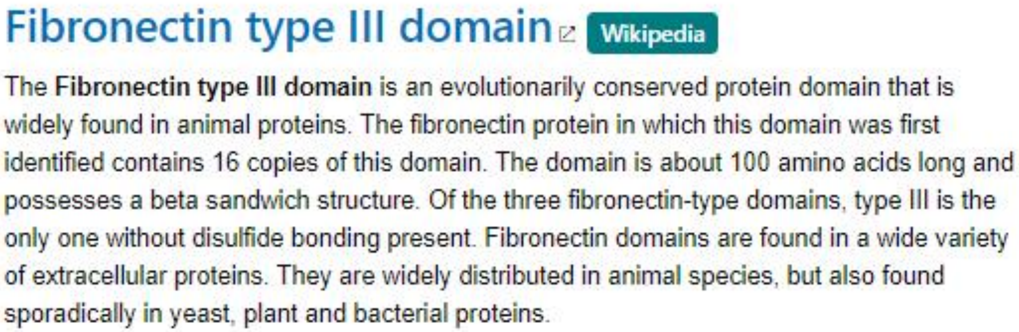
Integrated to

› IPR003961

Representative structure

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

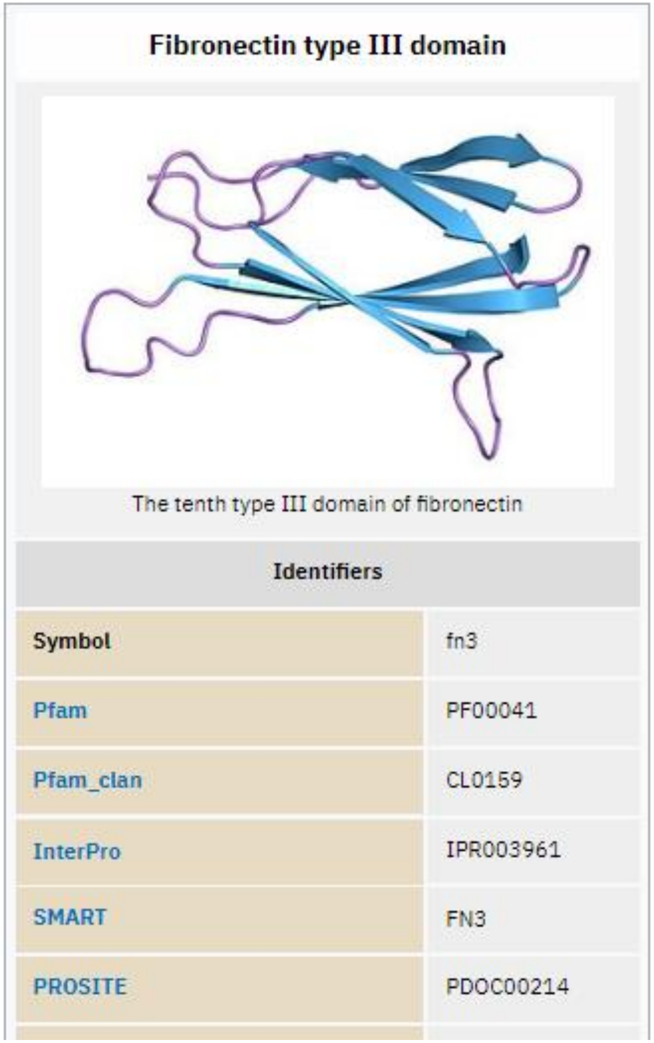
InterPro



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.

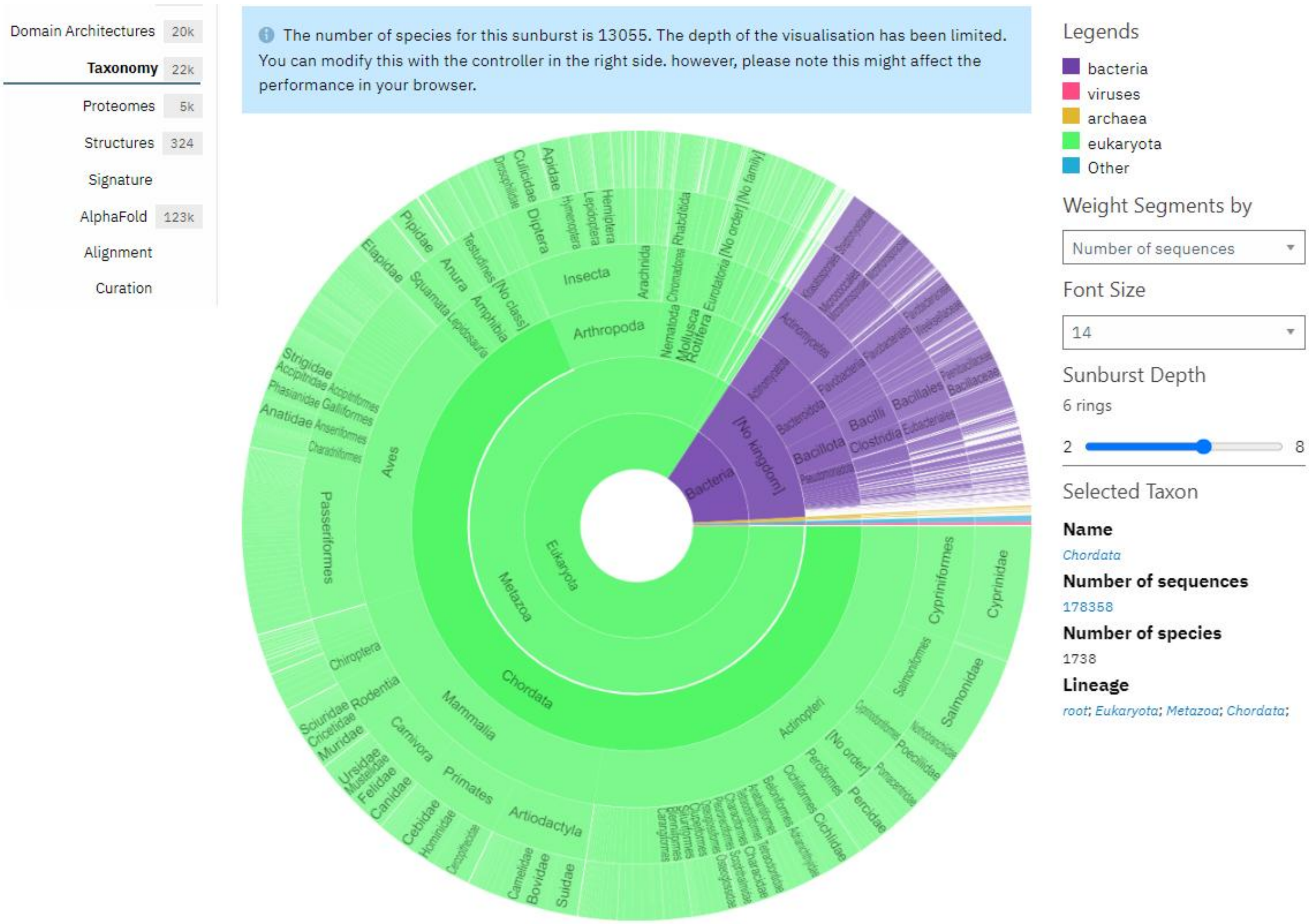


Fibronectin type III domain

The tenth type III domain of fibronectin

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

InterPro



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

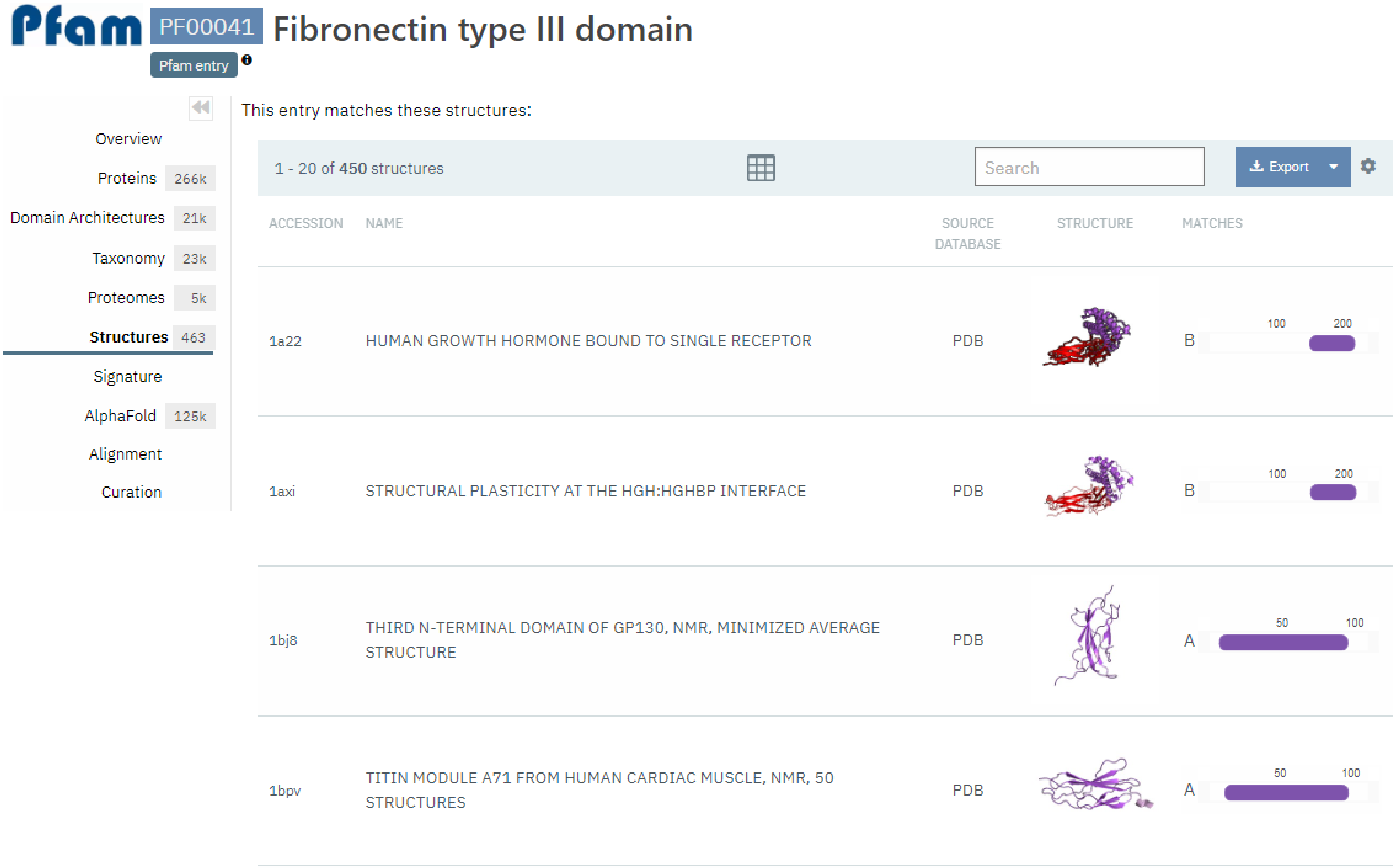
Name
Chordata

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:

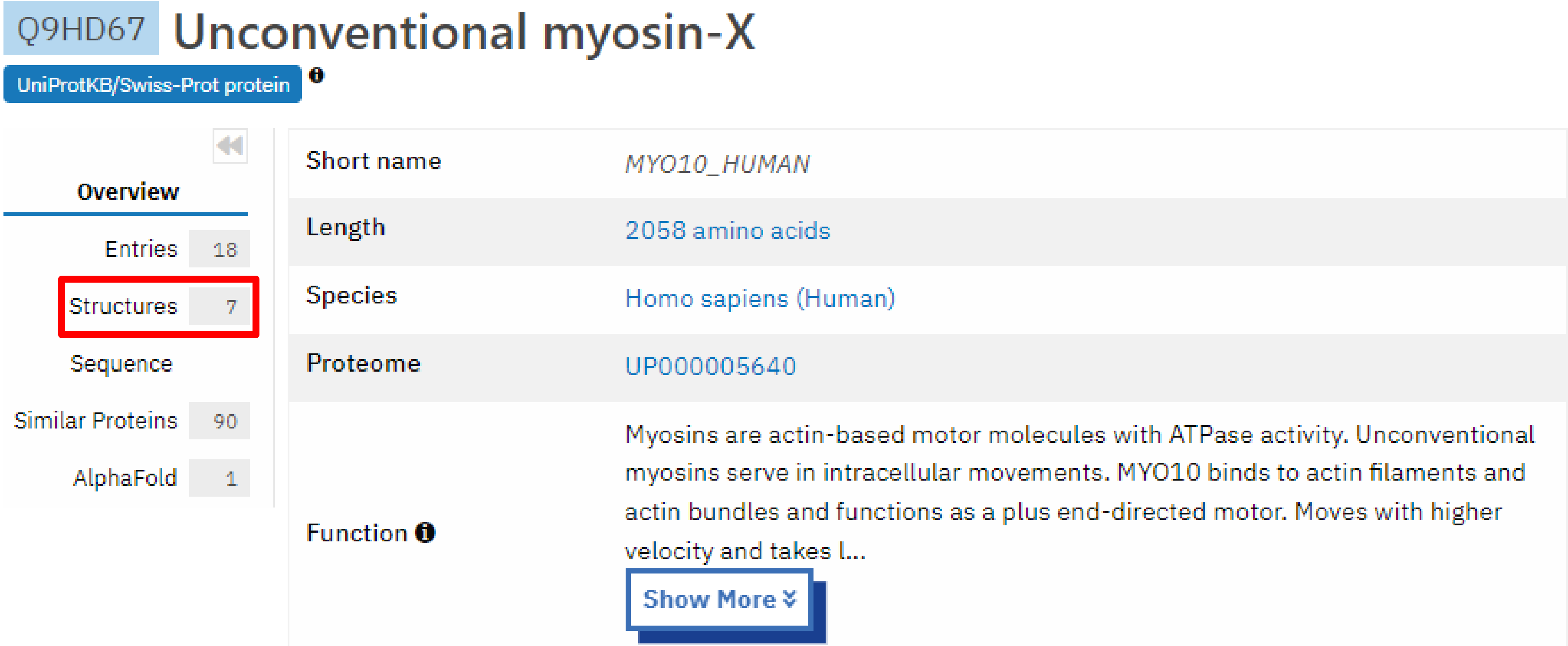
1 - 20 of **450** structures

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

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



Q9HD67 Unconventional myosin-X

UniProtKB/Swiss-Prot protein

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...

Exercise 1

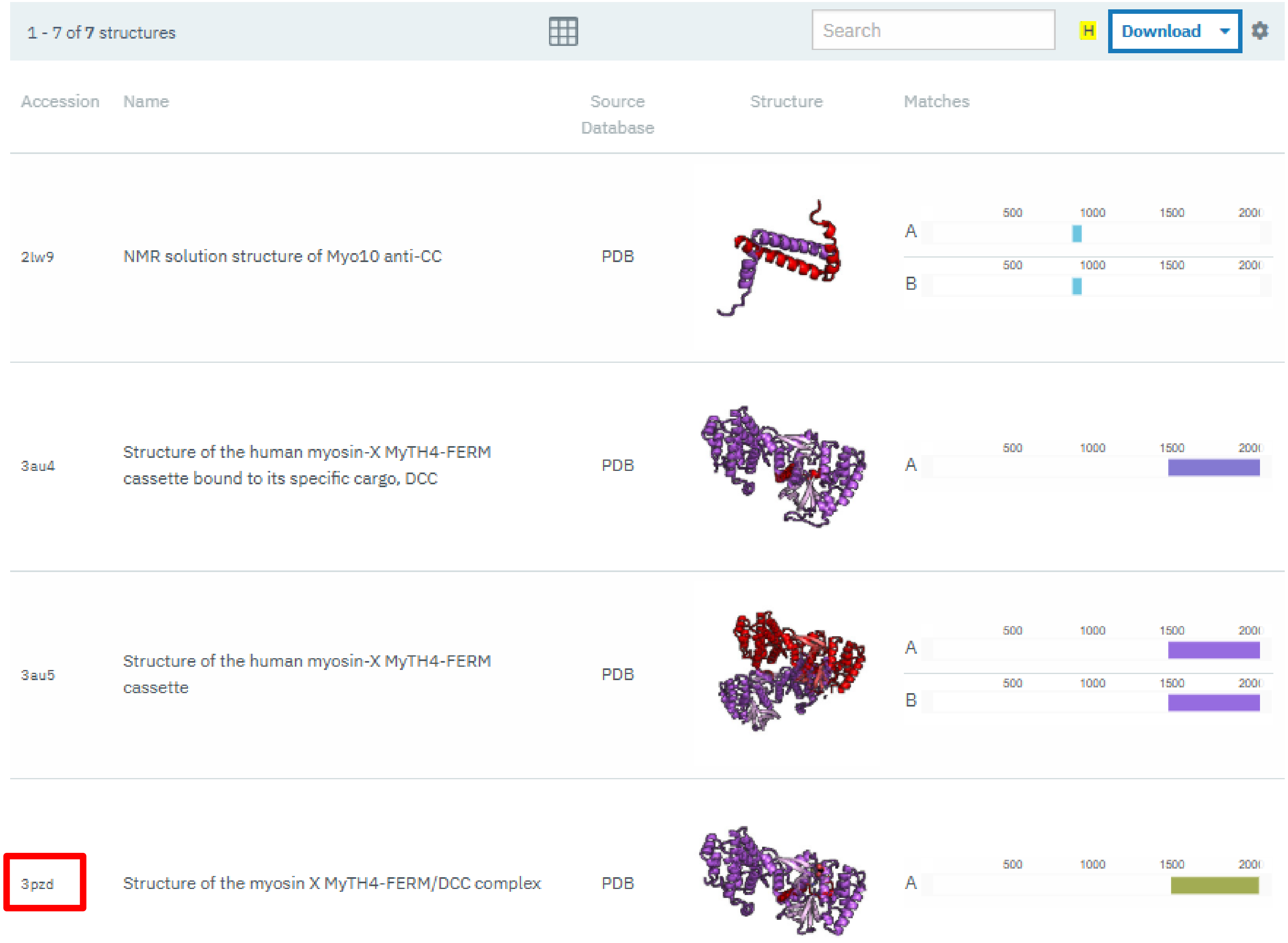
Find structures in the PDB for human myosin X

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

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

Exercise 2

Compare domain predictions to structure



Accession	Name	Source Database	Structure	Matches
2lw9	NMR solution structure of Myo10 anti-CC	PDB		A B
3au4	Structure of the human myosin-X MyTH4-FERM cassette bound to its specific cargo, DCC	PDB		A
3au5	Structure of the human myosin-X MyTH4-FERM cassette	PDB		A B
3pzd	Structure of the myosin X MyTH4-FERM/DCC complex	PDB		A

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 the C-terminal PH-like domain (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?