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

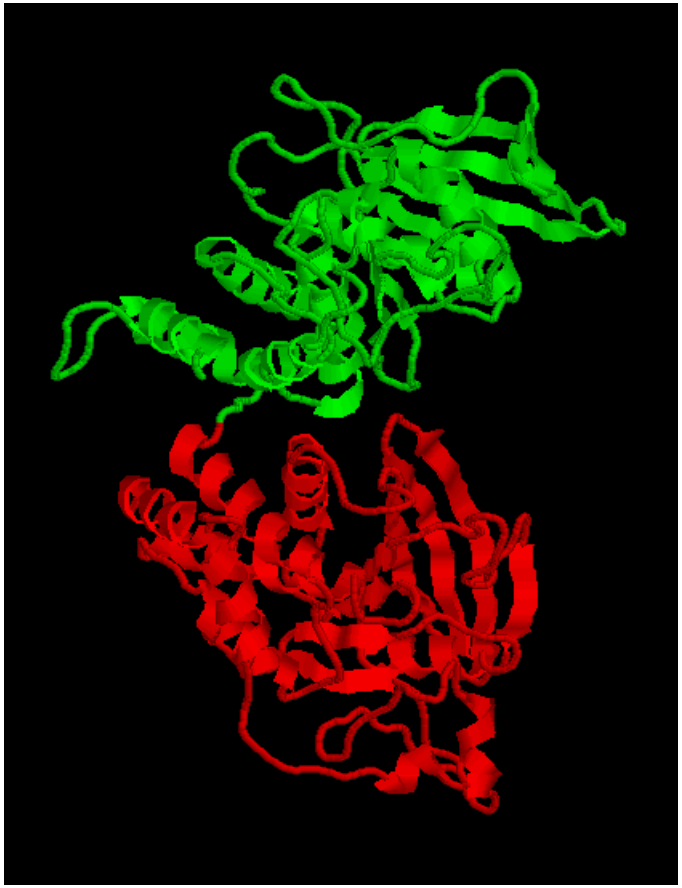
Proteins normally are multidomain (average 300 aa)

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 (2020) *Nucleic Acids Research*

Domain databases

SMART

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

SMART MODE:
NORMAL
GENOMIC

Simple
Modular
Architecture
Research
Tool

keywords...
Search SMART

HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK

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

Sequence SMART | Reset

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

Architecture query | Reset

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

SORL_HUMAN Examples: #1, #2

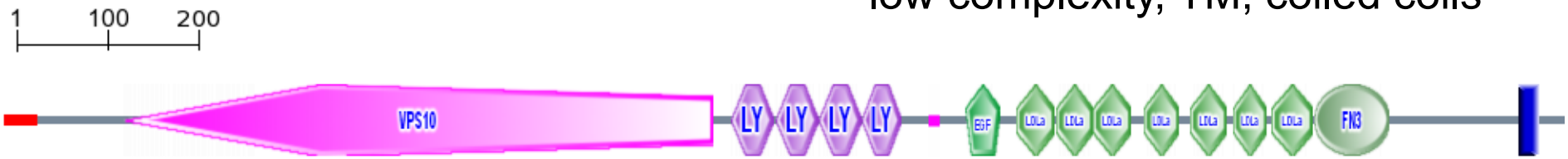
Protein sequence

Examples: #1, #2

Sequence SMART Reset

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: [ ]

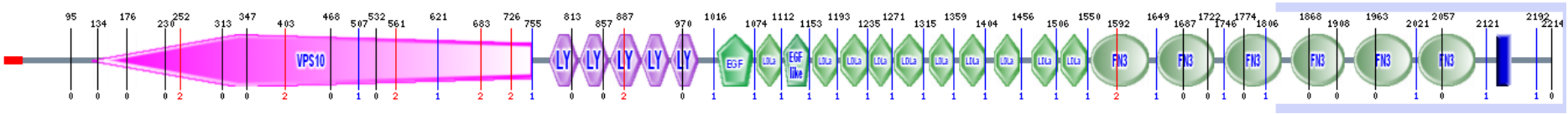
Taxonomic tree of query results.

- 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



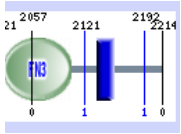
Domain databases SMART

Protein
Descrip
Species
Domain inventec
Represe 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)

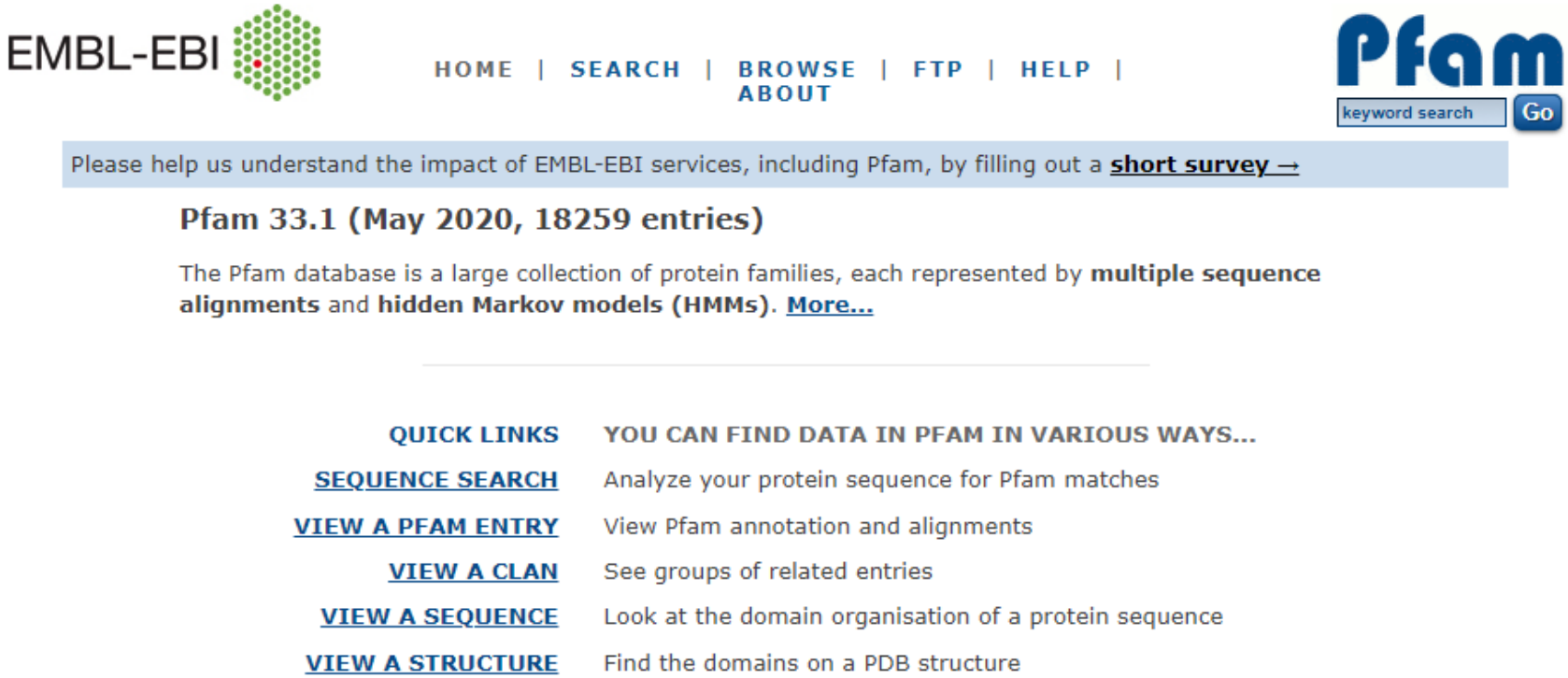
3D Structures of EGF domains in PDB

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*

...

Mistri et al (2021) *Nucleic Acids Research*

Domain databases
PFAM

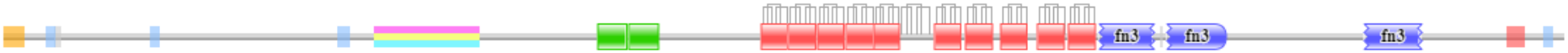
This is the summary of UniProt entry SORL_HUMAN (Q92673).

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
PFAM

Family: *fn3* (PF00041)

3078 architectures | 58087 sequences | 10 interactions | 1986 species | 274 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Summary: Fibronectin type III domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Fibronectin type III domain | Pfam | InterPro

Wikipedia rules!

This is the Wikipedia entry entitled "Fibronectin type III domain". More...

Fibronectin type

The **Fibronectin type III**
fibronectin protein in whic
long and possesses a beta
widely distributed in anim

Human proteins c

ABI3BP; ANKFN1; ASTN2;
CNTN5; CNTN6; COL12A1
EGFLAM; EPHA1; EPHA10;
FANK1; FLRT1; FLRT2; FL
HCFC1; HCFC2; HUGO; IF
IL31RA; IL6R; IL6ST; IL7
LRIT1; LRRN1; LRRN3; M
MYOM3; NCAM1; NCAM2;
PTPRB; PTPRC; PTPRD; PT
RIMBP2; ROBO1; ROBO2;
TRIM36; TRIM42; TRIM46

See also[edit]

- Monobody, an engineer

References[edit]

- Bazan, J. F. (1990). "St
National Academy of S
- Little, E.; Bork, P.; Doo
Journal of molecular ev
- Kornblihtt, A. R.; Umeza
from a single gene". The EMBO journal **4** (7): 1755–1759. PMC 554414. PMID 2992939.

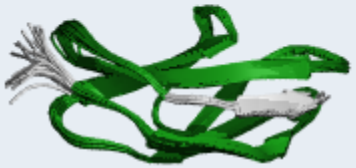
This page is based on a Wikipedia article. The text is available under the Creative Commons Attribution/Share-Alike License.

Summary

PDB entry 1k85

Solution structure of the fibronectin type iii domain from bacillus circulans wl-12 chitinase a1.

Experiment type:	NMR
Deposition date:	23-OCT-01
Authors:	Jee, J.G., Ikegami, T., Hashimoto, M., Kawabata, T., Ikeguchi, M., Watanabe, T., Shirakawa, M.
Species:	n/a
PubMed reference:	11600504



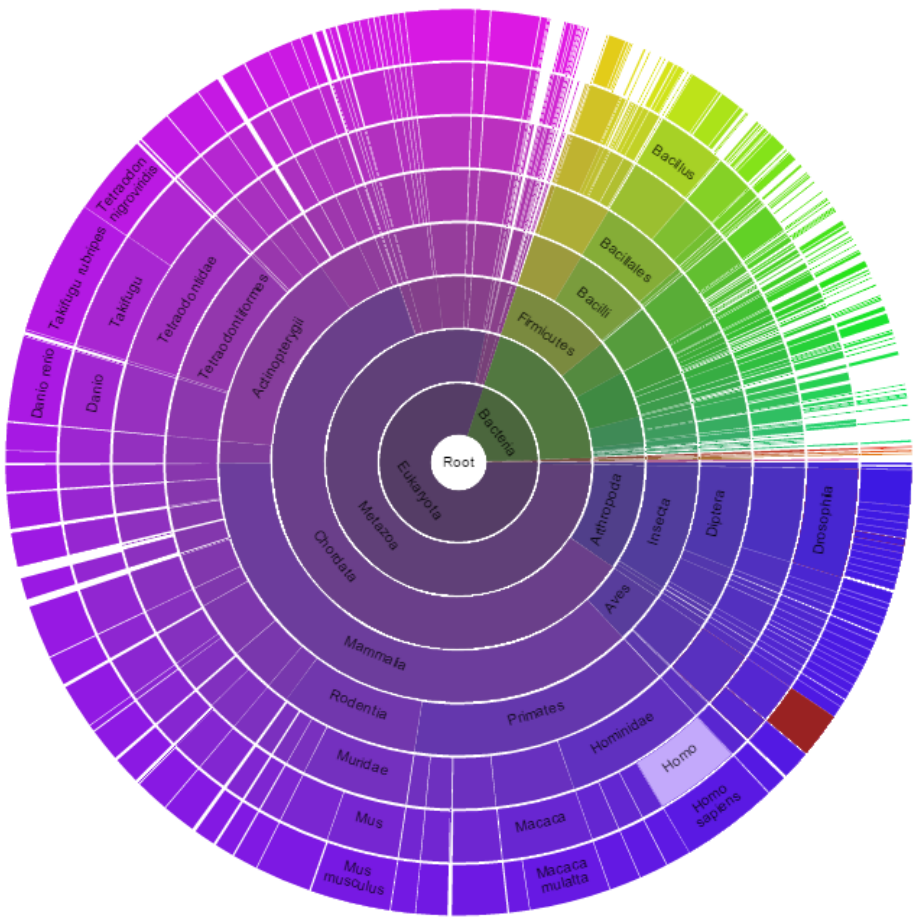
PDB entry 1k85: SOLUTION STRUCTURE OF THE FIBRONECTIN TYPE III DOMAIN FROM BACILLUS CIRCULANS WL-12 CHITINASE A1.
Enlarge image.

Domain databases
PFAM

Species distribution

Sunburst | Tree

This visualisation provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the adjacent tab. **More...**



Sunburst controls Hide

Homo

Root
- Eukaryota
  - Metazoa
    - Chordata
      - Mammalia
        - Primates
          - Hominidae
            - Homo

Weight segments by...

- number of sequences
- number of species

Change the size of the sunburst

Small — Large

Colour assignments

Archea	Eukaryota
Bacteria	Other sequences
Viruses	Unclassified
Viroids	Unclassified sequence

Selections

Align selected sequences to HMM
Generate a FASTA-format file
Clear selection

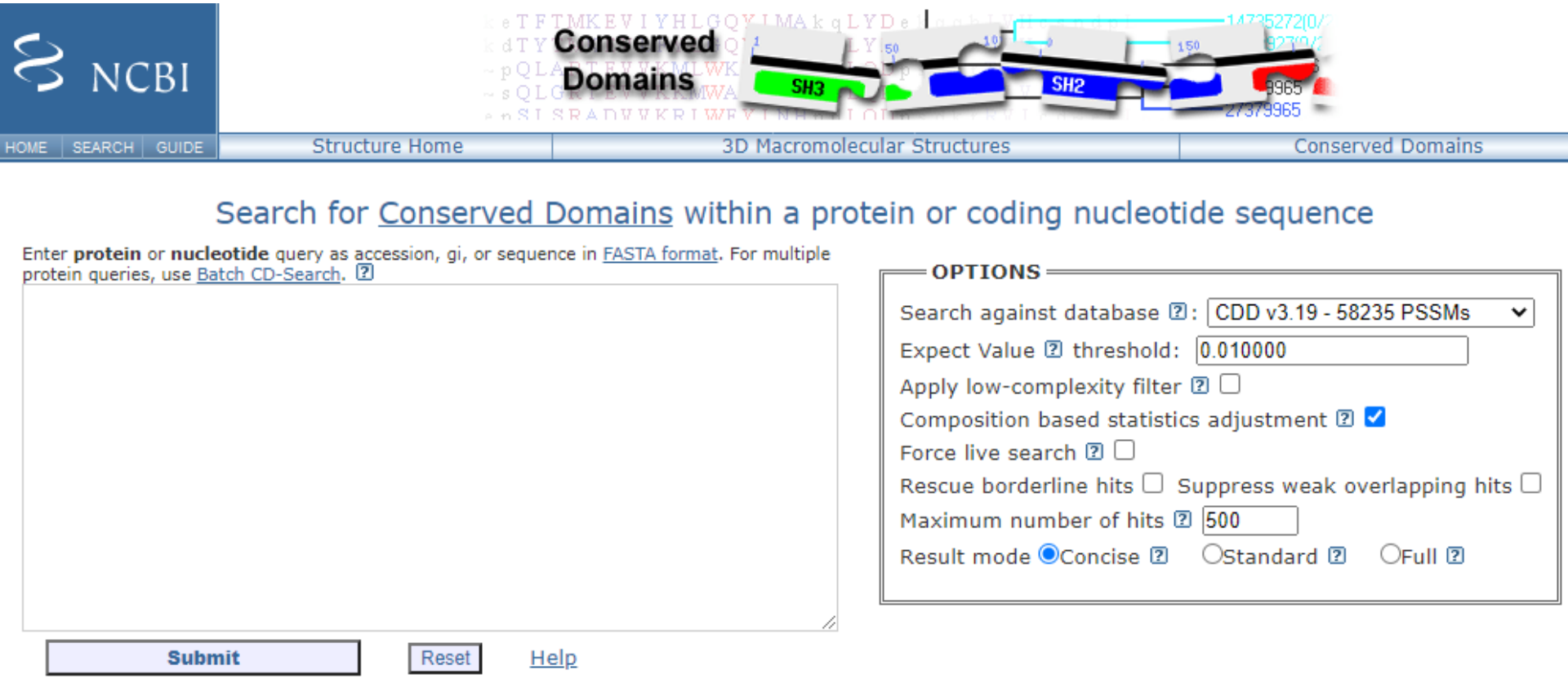
Currently selected:

- 274 sequences
- 1 species

Note: selection tools show results in pop-up windows. Please disable pop-up blockers.

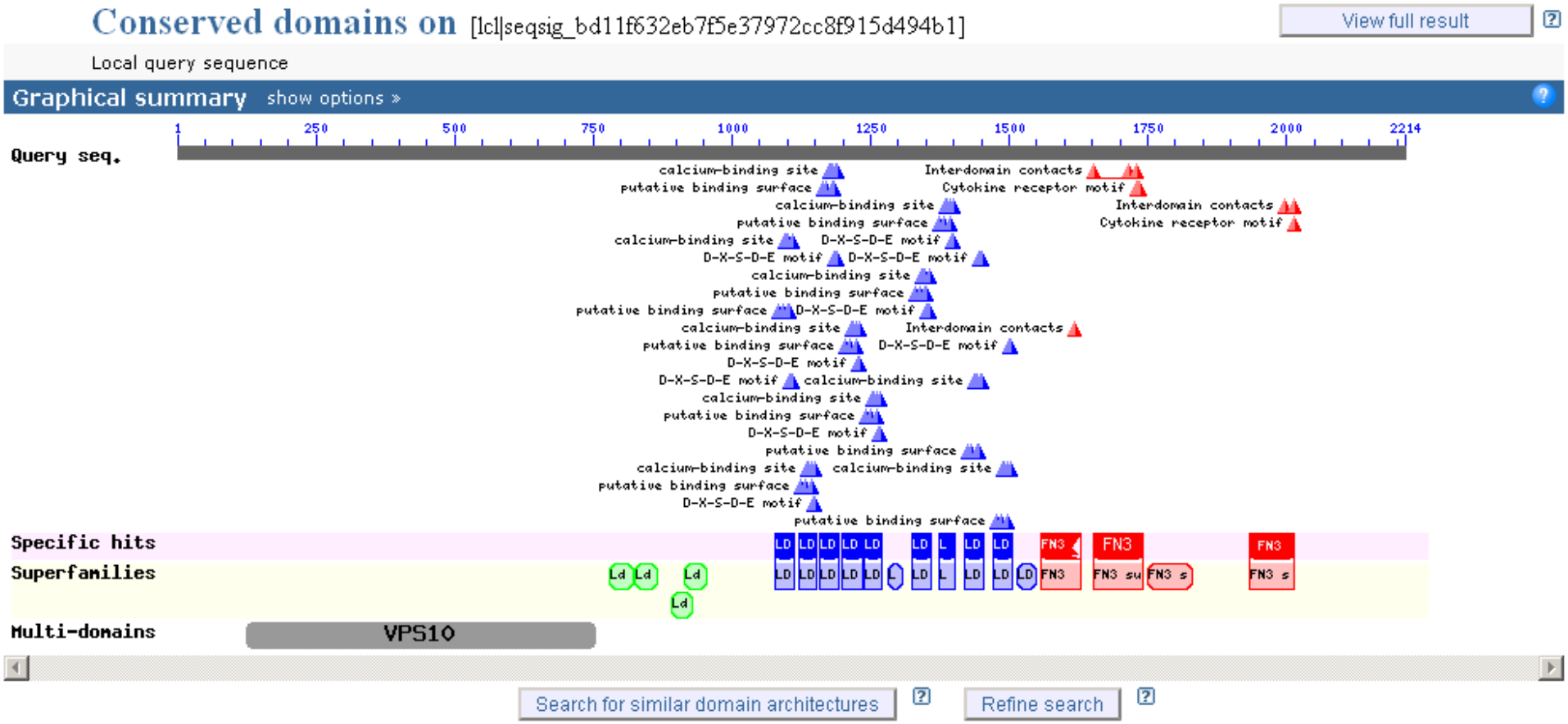
Domain databases
CDD

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



Lu et al (2020) *Nucleic Acids Res*

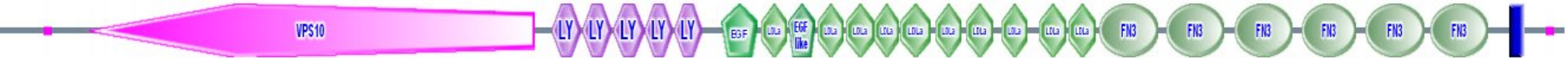
Domain databases
CDD



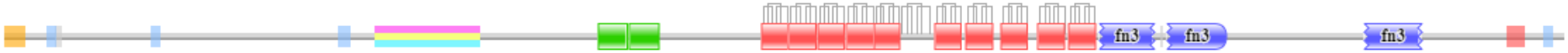
Domain databases

SORLA/SORL1 from *Homo sapiens*

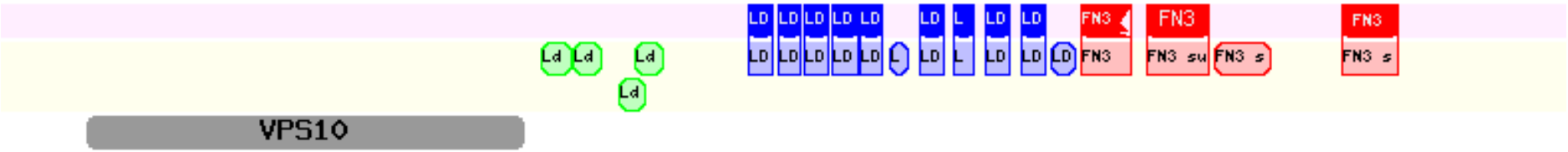
SMART



PFAM



CDD



Exercise 1

Examine a UniProt Entry and find related PDBs

• Let's see whether human myosin X (UniProt id Q9HD67) or its homologs have a solved structure. Go to PDB Advanced Search page:

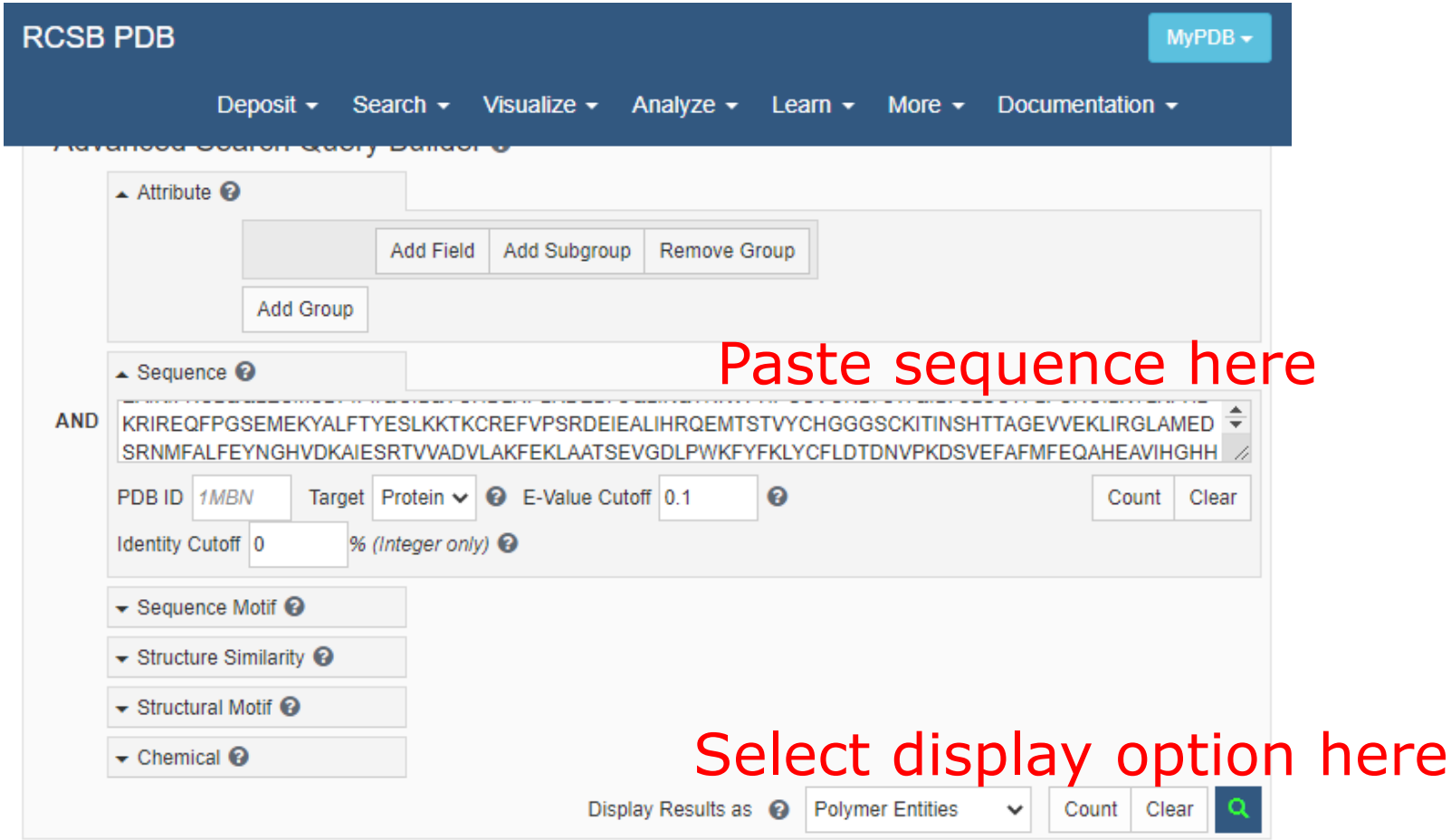
Menu > Search > Advanced Search
https://www.rcsb.org/search/advanced

• Obtain from UniProt the protein sequence "Q9HD67" and paste it the Sequence window (only sequence – no header).

• In "Display Results as" select option "Polymer entities"

Exercise 1

Examine a UniProt Entry and find related PDBs



Exercise 1

Examine a UniProt Entry and find related PDBs

- Considering that your query was a human myosin X, can you interpret the first three hits? Which part of your query was matched? Which protein was hit in the database?

- What about the 4th hit?

Exercise 1

Examine a UniProt Entry and find related PDBs

- Considering that your query was a human myosin X, can you interpret the first three hits? Which part of your query was matched? Which protein was hit in the database?

- What about the 4th hit?

- Can you find a hit to a protein that is not human myosin X? Which part of your query was matched?

Exercise 2

Analyse domain predictions with PFAM

- Let's look at the domains predicted for human myosin X. Go to PFAM: http://pfam.xfam.org/

- Select the option VIEW A SEQUENCE

- Type in the window the UniProt id of the protein sequence "Q9HD67" and hit the Go button.

- Compare the positions of the domains predicted with the ranges of the BLAST matches in PDB from the previous exercise.

Which domains were matched in the human myosin X by each of those hits?

Exercise 3

Examine domains in Chimera

•Open the structure of the 3rd hit (3PZD) in Chimera

Now colour the fragments corresponding to the PFAM domains MyTH4 (in orange), RAS associated (in pink) and FERM_M (in blue).

How do the PFAM annotations fit the structure?

How many more domains can you identify visually?

•Chain B in this structure is a small peptide.

Which part of the human myosin X is interacting with this peptide in relation to the domains you have coloured? And what about the glycerol?