CroMaSt: A workflow for domain family curation through cross-mapping of structural instances between protein domain databases

Hrishikesh Dhondge, Isaure Chauvot de Beauchêne, Marie-Dominique Devignes

CroMaSt: A workflow for domain family curation through cross-mapping of structural instances between protein domain databases

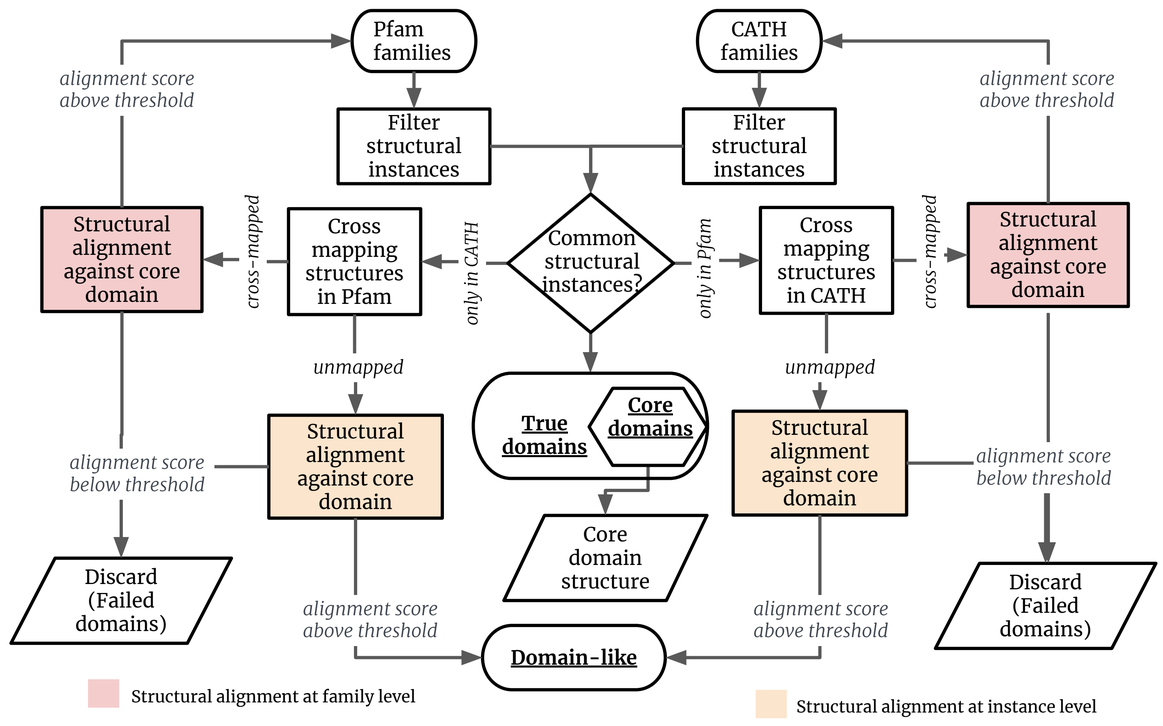
Hrishikesh Dhondge[1], Isaure Chauvot de Beauchêne[1], Marie-Dominique Devignes[1]
[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Overview

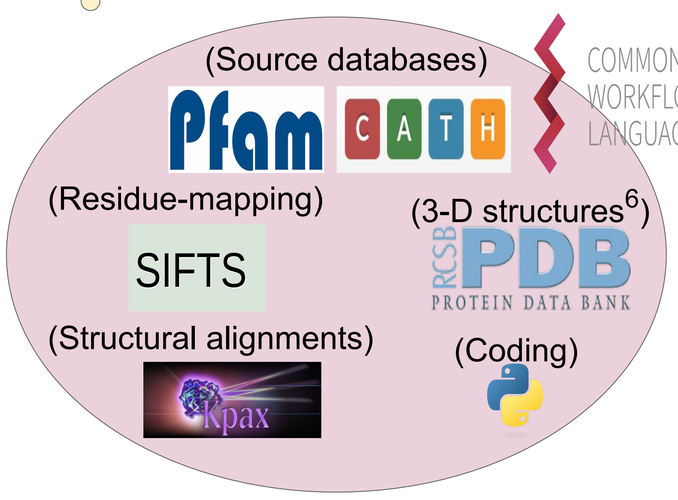
Protein domains are building blocks of proteins, essential for understanding structure-function relationships. However the diversity of domain databases makes it difficult to capture a clear definition for a given type of domain and to enumerate its true instances.

CroMaSt (**Cro**ss-**Ma**pper for **St**ructural domains) is an automated iterative workflow to clarify domain definition by cross-mapping of domain structural instances between domain databases. It classifies all structural instances of a given domain type into 3 different categories (core, true and domain-like). CroMaSt is developed in Common Workflow Language (CWL)[1] and takes advantage of 2 well-known and widely used domain databases, Pfam[2] (sequence-based) and CATH[3] (structure-based).

Category	Starting family member	Cross-mapped	Structurally well aligned
Core	✓	✓	✓
True	-	✓	✓
Domain-like	-	-	✓



Implementation

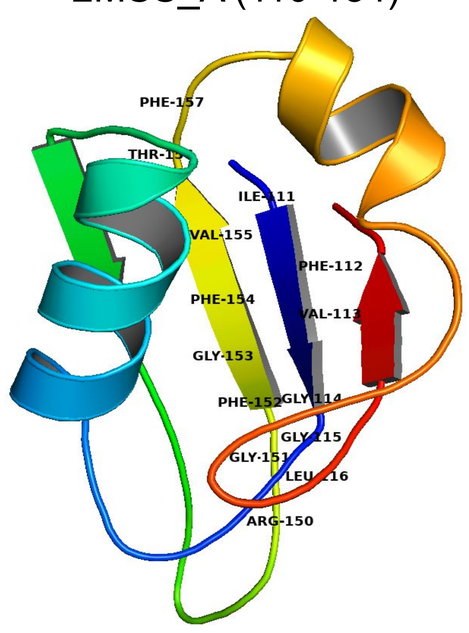


Main difficulties:

- Residue-mapping (SIFTS[4])
- Cross-mapping (Source db)
- Structural alignments (Kpax[5])
- Computing average structures (3-D coordinates)

Structural instance

2MSS_A (110-184)



Pfam: PF00076 (RRM_1)
CATH: 3.30.70.330 (RRM)

Results

Steps	Iteration 1		Iteration 2	
	Pfam	CATH	Pfam	CATH
Starting Families	1	1	14	0
StIs filtered on domain length	1147	1527	96	80*
Obsolete and inconsistent entries	3	323	0	0
=>Residue-mapped StIs	1144	1204	96	–
Common StIs (**Core** & **True**)	886	886	80	80
Remaining StIs (not common)	258	318	16	0
Cross-mapped StIs	0	244	0	0
Properly aligned at family level	–	80	–	–
Not properly aligned at family level	–	164	–	–
Not cross-mapped StIs (unmapped)	258	74	16	0
Properly aligned at instance level (**Domain-like**)	255	74	15	–
Not properly aligned at instance level	3	0	1	–
Failed structural instances	3	164	1	0
New families found	14	0	0	0

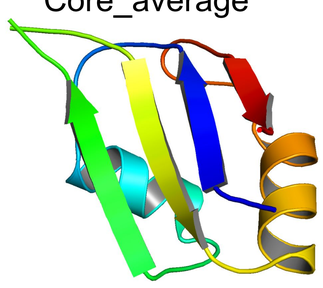
StI: Structural instance

*These StI entries are cross-mapped and properly aligned at the family level from previous iteration

Results Visualization

Core domain

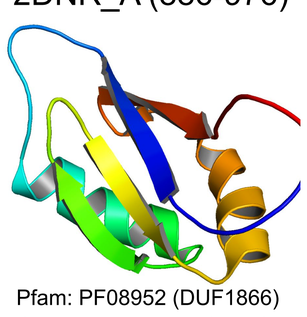
Core_average



Pfam: PF00076 (RRM_1)
CATH: 3.30.70.330 (RRM)

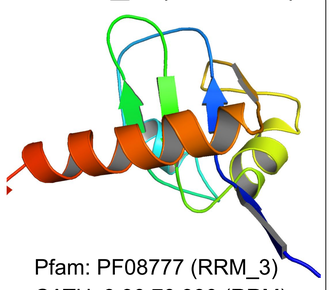
True-domains

2DNR_A (886-976)



Pfam: PF08952 (DUF1866)
CATH: 3.30.70.330 (RRM)

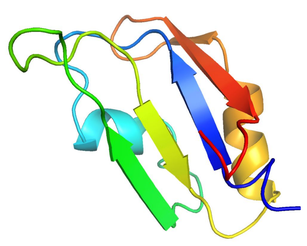
1OWX_A (222-334)



Pfam: PF08777 (RRM_3)
CATH: 3.30.70.330 (RRM)

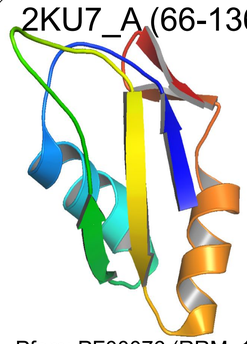
Domain-like

2DNN_A (230-311)



CATH: 3.30.70.330 (RRM)

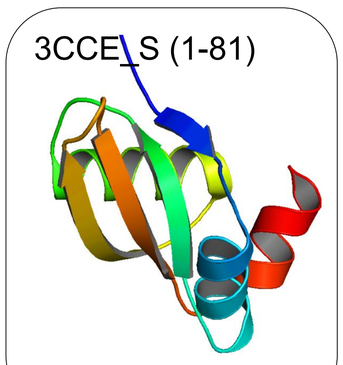
2KU7_A (66-136)



Pfam: PF00076 (RRM_1)

Failed domains

3CCE_S (1-81)



CATH: 3.30.70.330 (RRM)

All 3-D structure visualizations were generated using PyMol[7]

Conclusion & Perspectives

- Generic, can be used for other domain types
- Protein design: Provides the prototype for given domain type
- Database curation: Explores “allowed” domain diversity and detects inconsistencies

Visualize Complete Workflow

References

1. Michael R. Crusoe, doi:10.1145/3486897; 2022
2. Mistry, Jaina et al. doi:10.1093/nar/gkaa913; 2021
3. Sillitoe, Ian et al. doi:10.1093/nar/gkaa1079; 2021
4. Dana, Jose M et al. doi:10.1093/nar/gky1114; 2019
5. Ritchie, David W. doi:10.1093/bioinformatics/btw300; 2016
6. Burley, Stephen K et al. doi:10.1093/nar/gkaa1038; 2021
7. PyMOL, Version 2.5.2, Schrödinger, LLC

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813239.