



HAL
open science

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

► To cite this version:

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. ECCB2022- 21st European Conference on Computational Biology, Sep 2022, Sitges, Spain. 10.48546/WORKFLOWHUB.WORKFLOW.390.1 . hal-03789541

HAL Id: hal-03789541

<https://hal.science/hal-03789541v1>

Submitted on 27 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

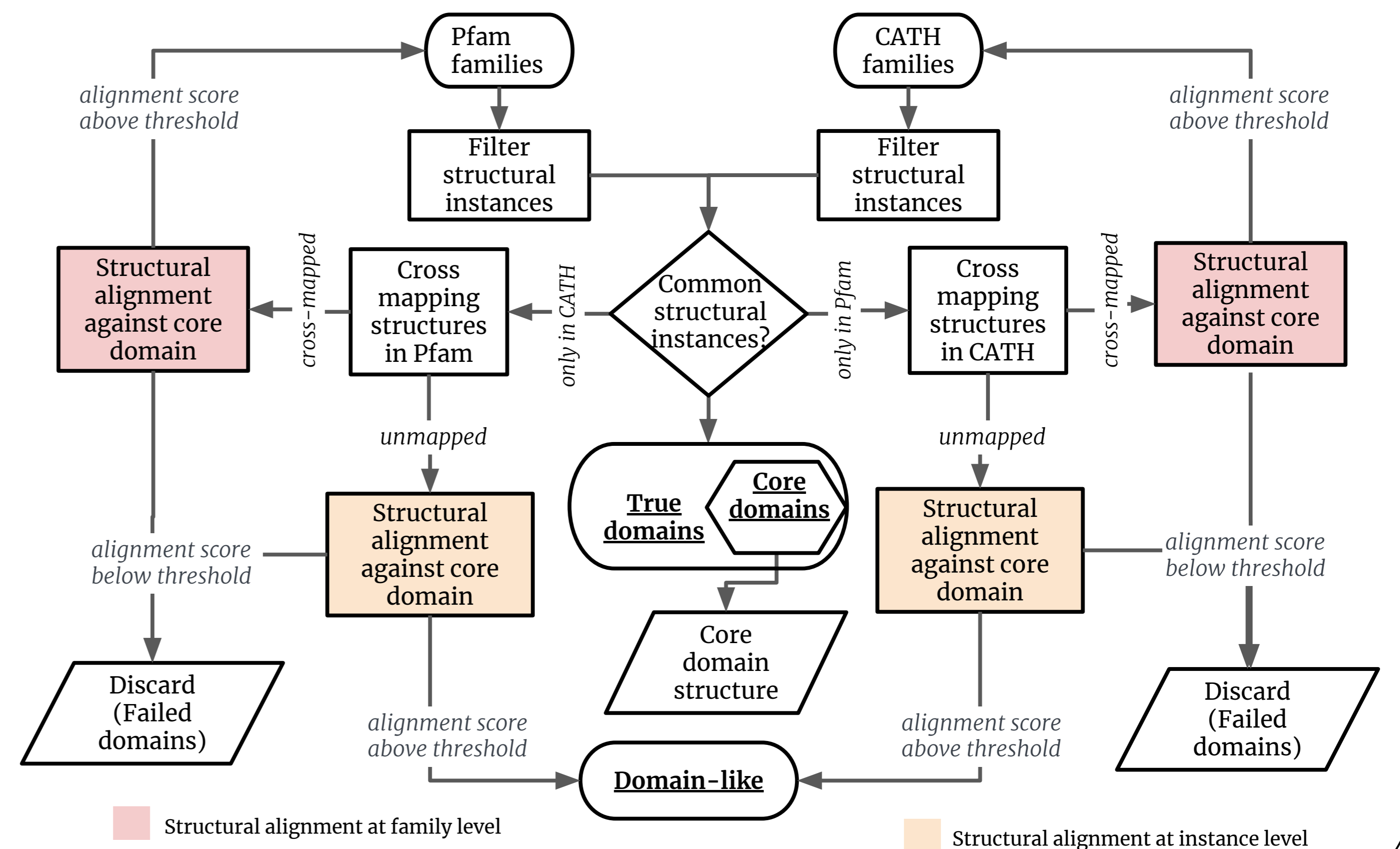
Hrishikesh Dhondge¹, Isaure Chauvot de Beauchêne¹, Marie-Dominique Devignes¹
¹Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Overview

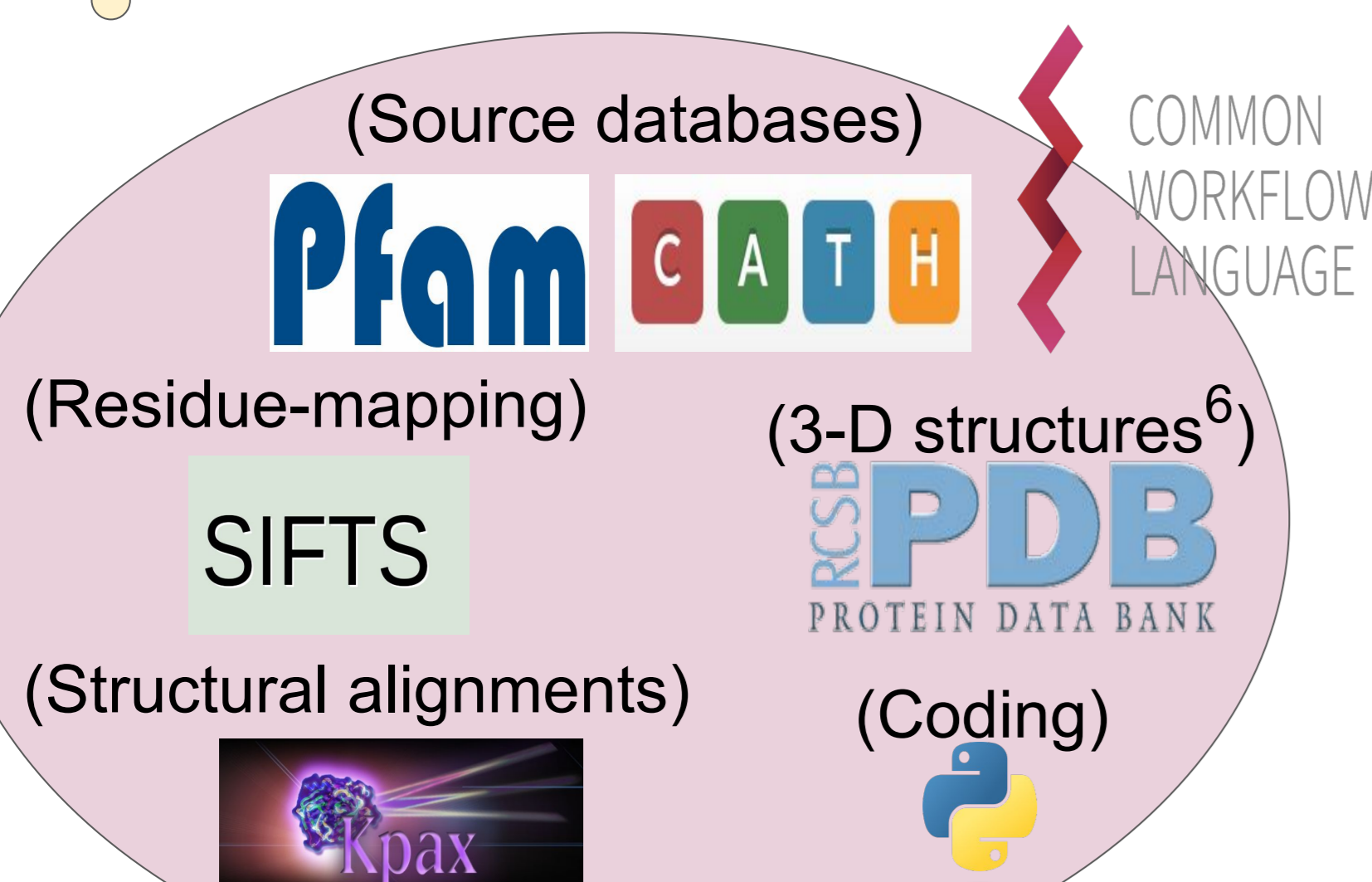
Protein domains are building blocks of proteins, essential for understanding structure-function relationships. However the diversity of domain databases makes it difficult to capture a clear definition for a given type of domain and to enumerate its true instances.

CroMaSt (**C**ross-**M**apper for **S**tructural 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, Pfam² (sequence-based) and CATH³ (structure-based).

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



Implementation

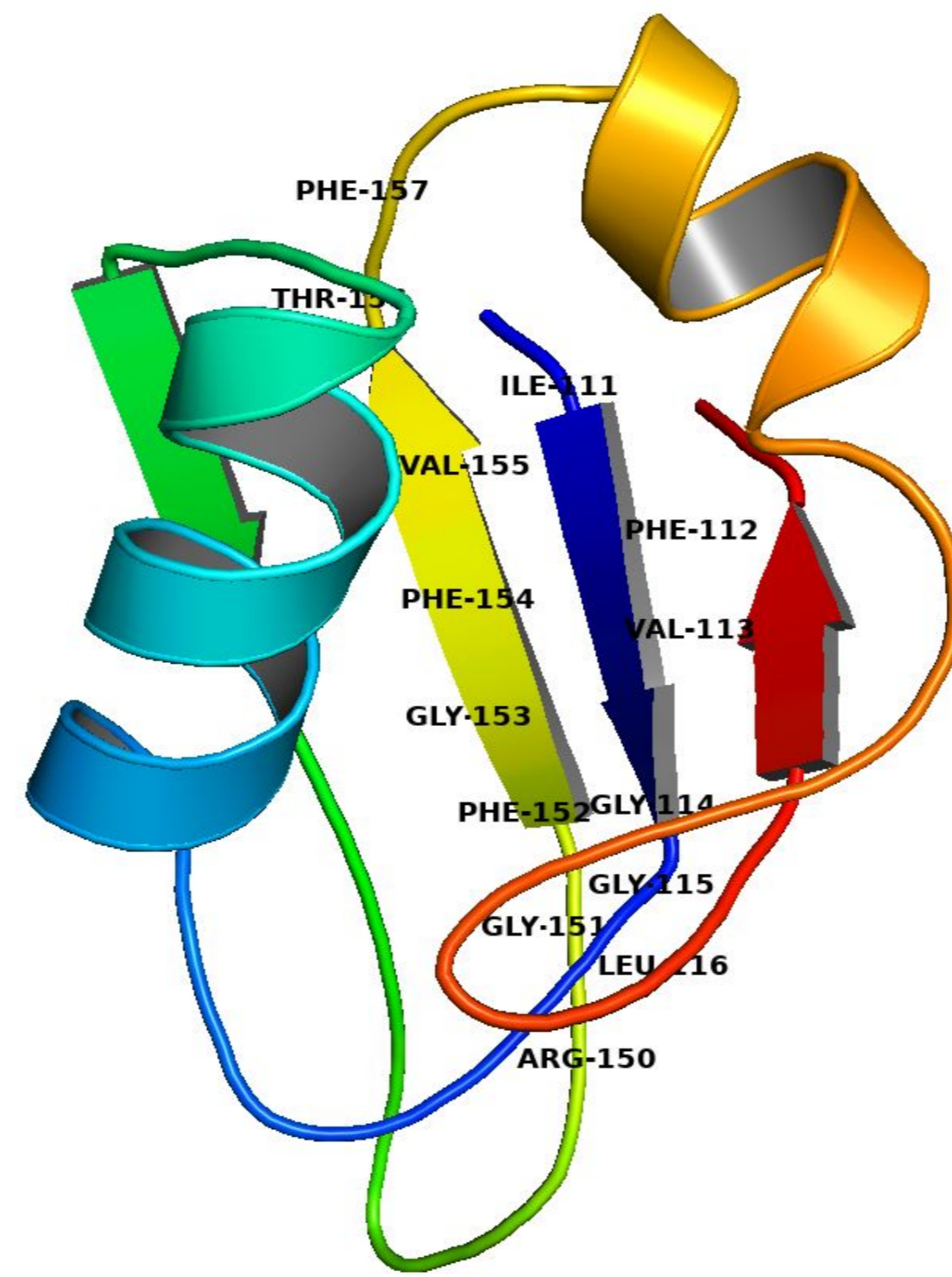


Main difficulties:

- Residue-mapping (SIFTS⁴)
- Cross-mapping (Source db)
- Structural alignments (Kpax⁵)
- Computing average structures (3-D coordinates)

Structural instance

2MSS_A (110-184)



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

Results

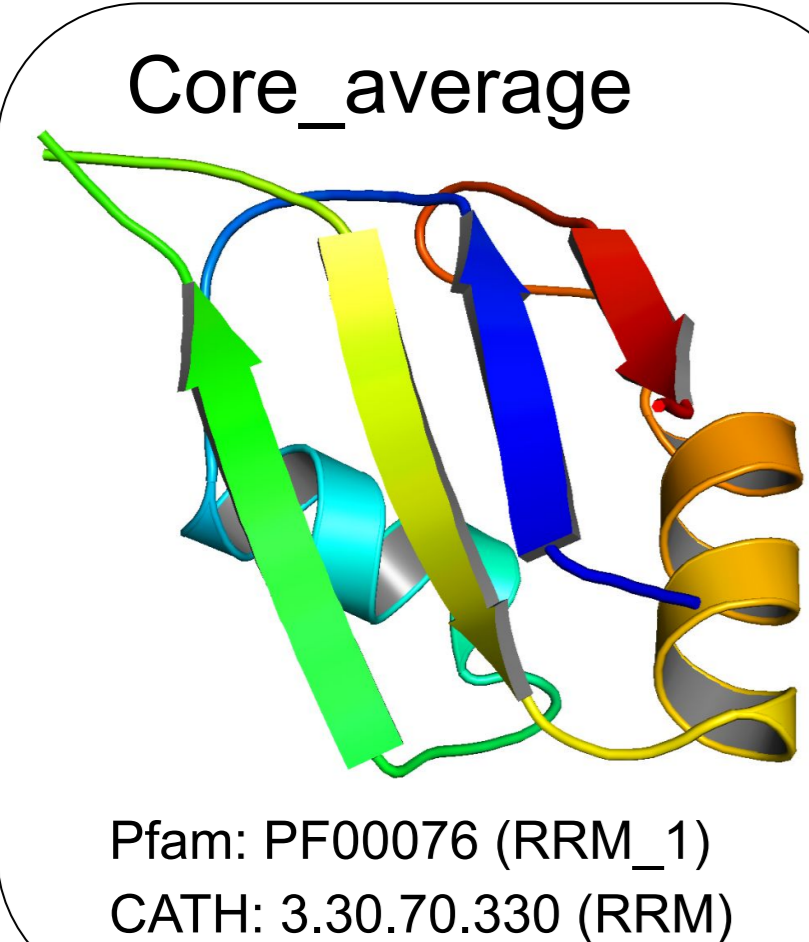
Steps	Iteration 1		Iteration 2	
	Pfam	CATH	Pfam	CATH
Starting Families	1	1	14	0
StIs filtered on domain length	1147	1527	96	80*
Obsolete and inconsistent entries	3	323	0	0
=>Residue-mapped StIs	1144	1204	96	-
Common StIs (Core & True)	886	886	80	80
Remaining StIs (not common)	258	318	16	0
Cross-mapped 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

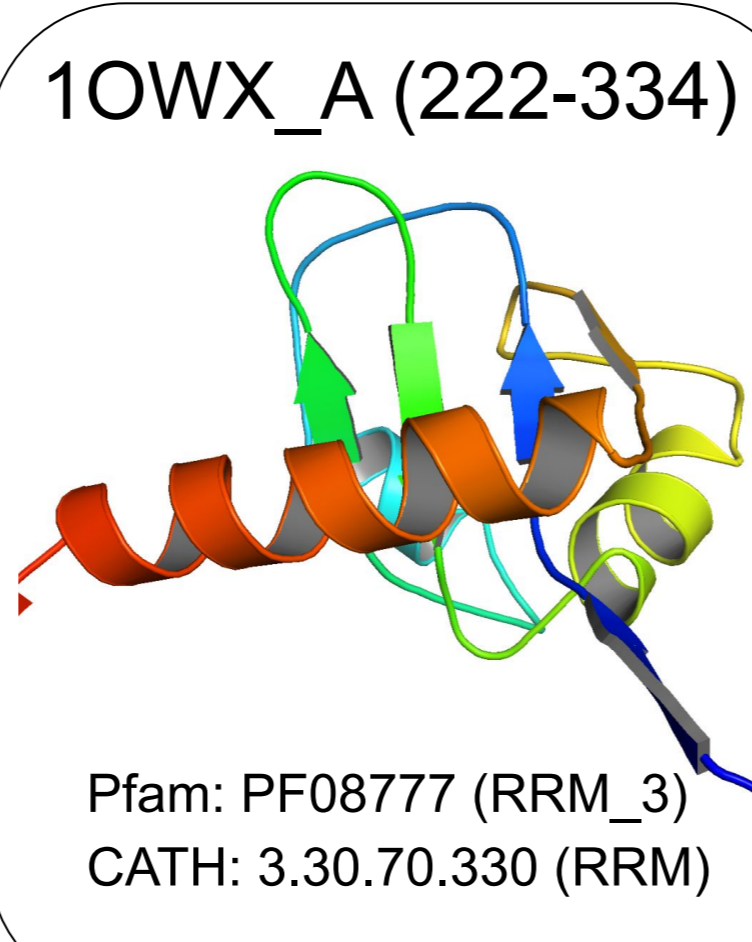
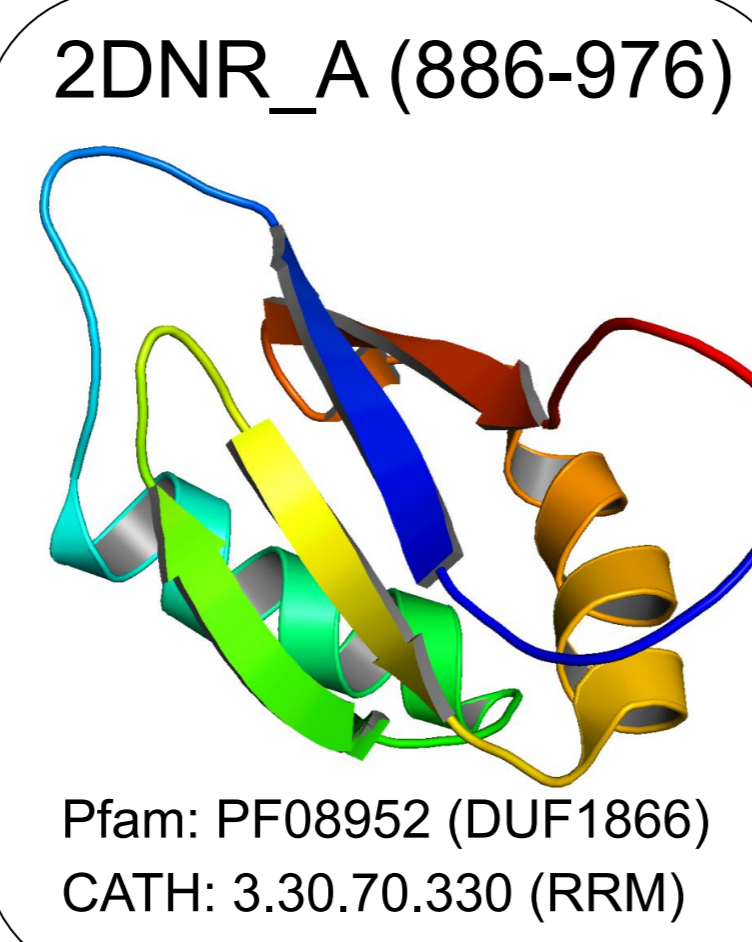
*These StI entries are cross-mapped and properly aligned at the family level from previous iteration

Results Visualization

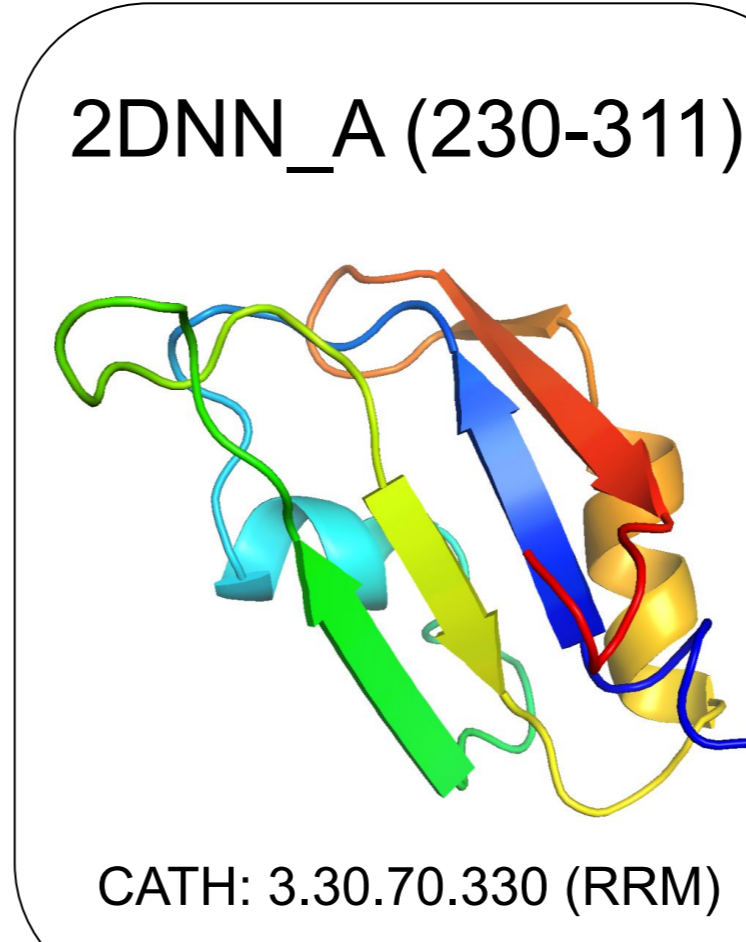
Core domain



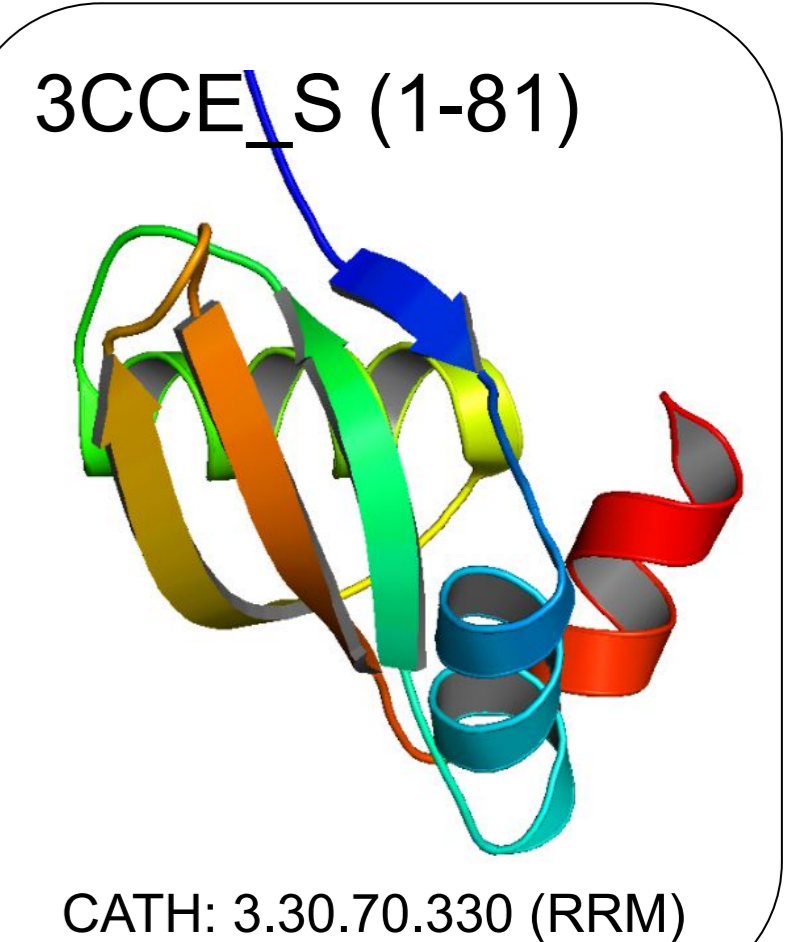
True-domains



Domain-like



Failed domains



All 3-D structure visualizations were generated using PyMol⁷

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



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.

