Discovering the general architecture of protein families with OverProt

MUNI SCI National Centre for Biomolecular Research

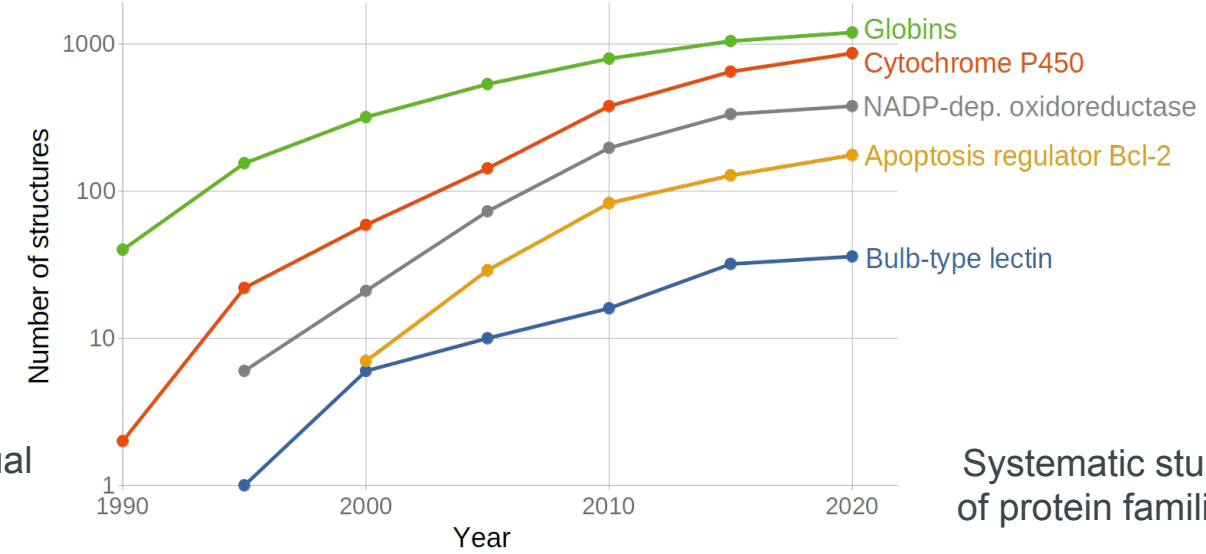
Adam Midlik, Ivana Hutařová Vařeková, Jan Hutař, Aliaksei Chareshneu, Radka Svobodová, Karel Berka

ENBIK 2022, 13 June 2022, Němčice

CEITEC

Introduction

- Protein structures → help us understand the function
  - Today > 190 000 experimental structures
- Similar protein structures form **protein families**
  - CATH database – 6 631 families
  - Families are getting bigger

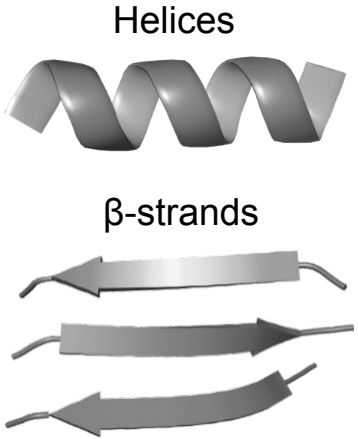


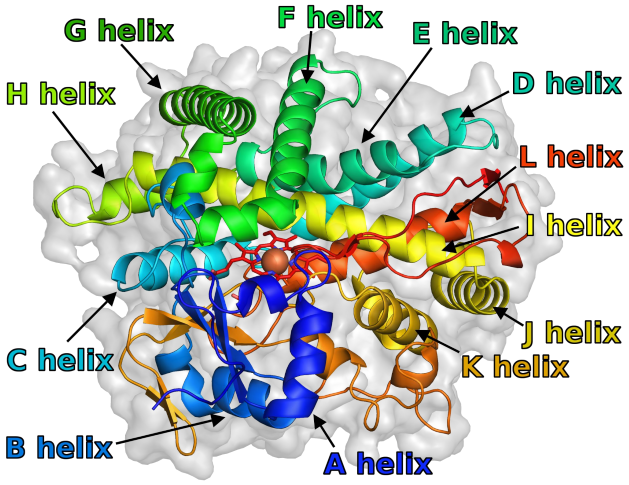
Study of individual structures

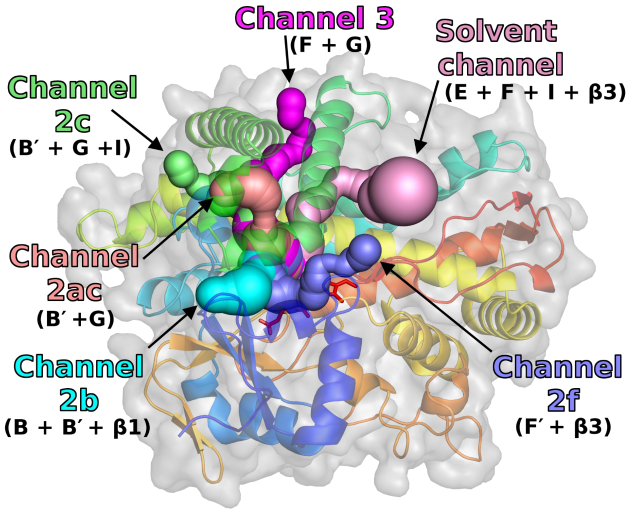
Systematic study of protein families

Secondary structure elements (SSEs)

- Characteristic arrangement in each protein family
  - ➜ “General architecture” of the family
- Can serve as landmarks
  - Help us orient in the structures
  - Help us locate the key regions (active sites, channels…)





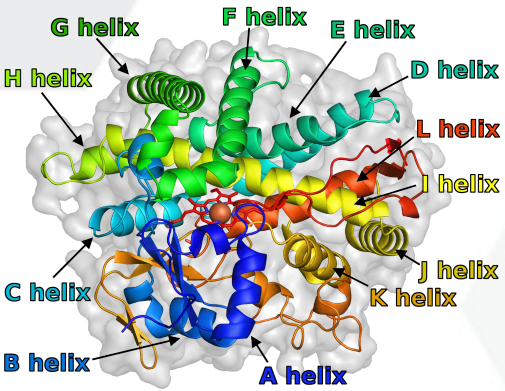


SecStrAnnotator

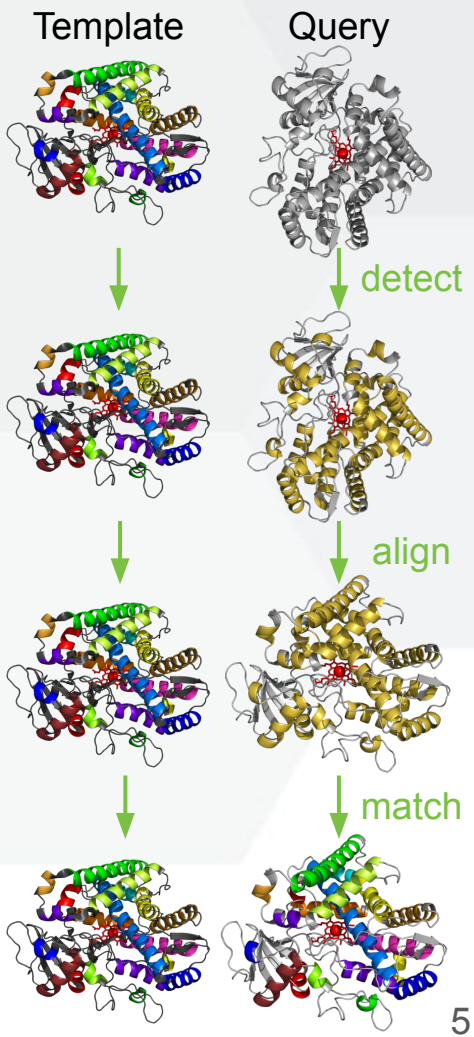
SSE annotation

= Labelling of SSEs, consistent throughout the family

- Many protein families have traditional annotations
  - Performed manually :(



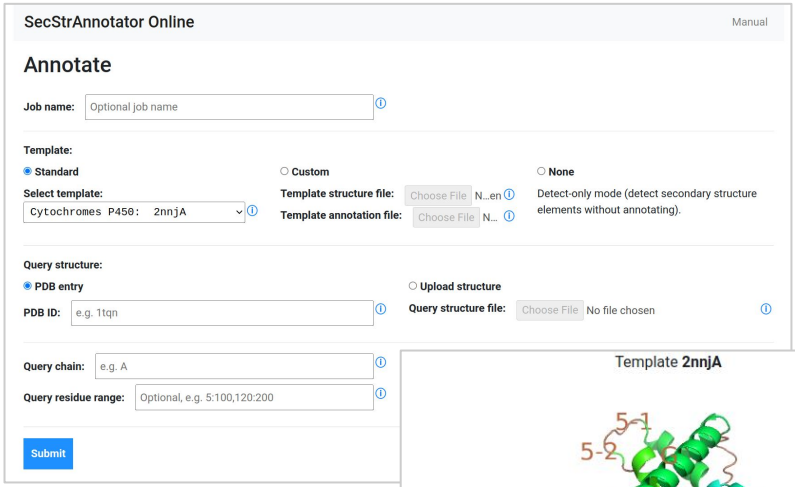
- Our solution – **SecStrAnnotator**
  - Annotation based on a provided template

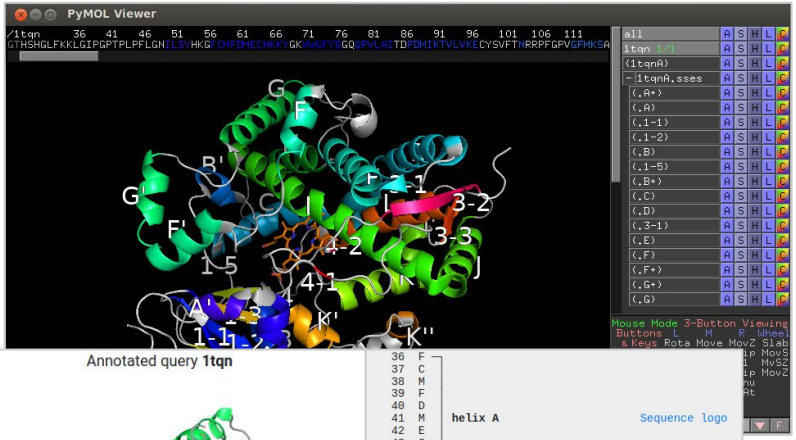


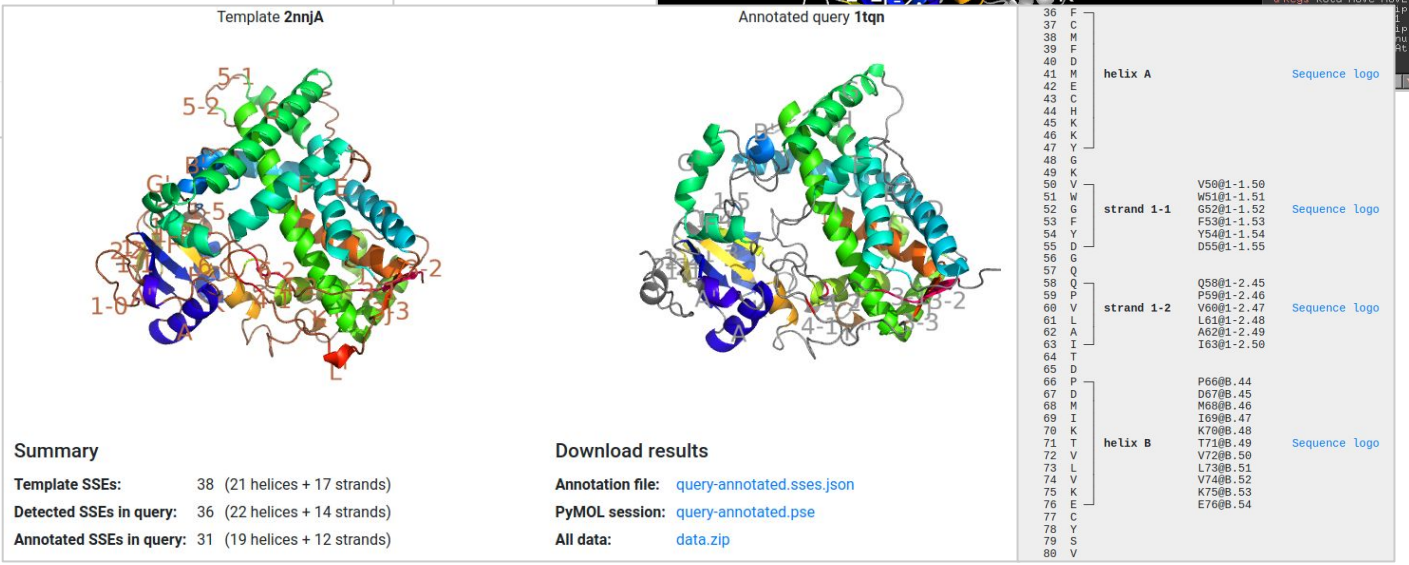
Midlik A. et al. (2019) Automated family-wide annotation of secondary structure elements. In Kister A.E. (ed.) *Protein Supersecondary Structures*, vol. 1958, pp. 47–71.

Midlik A. et al. (2021) Uncovering of cytochrome P450 anatomy by SecStrAnnotator. *Scientific Reports*, **11**, 12345.

SecStrAnnotator online – sestra.ncbr.muni.cz



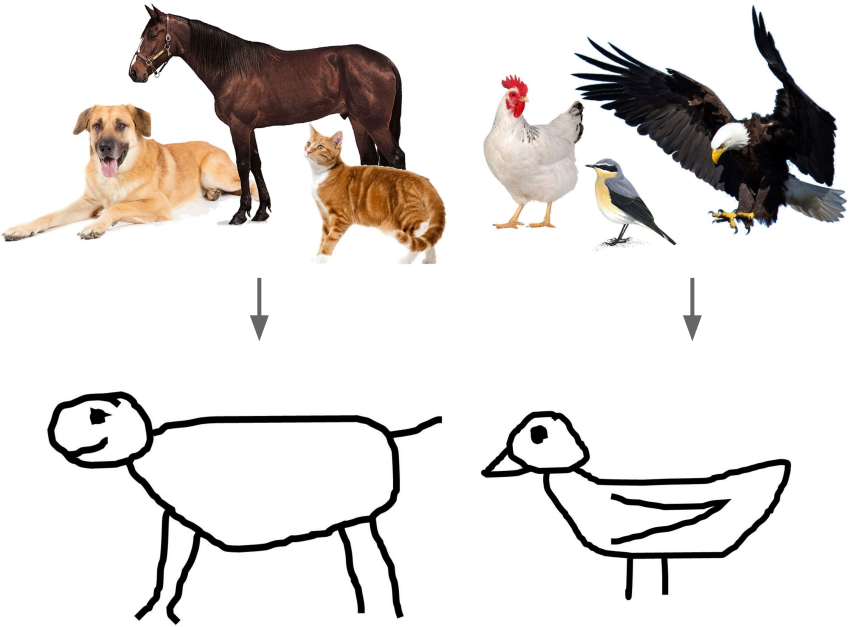




OverProt

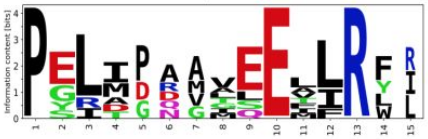
What is the “general architecture” of a family?

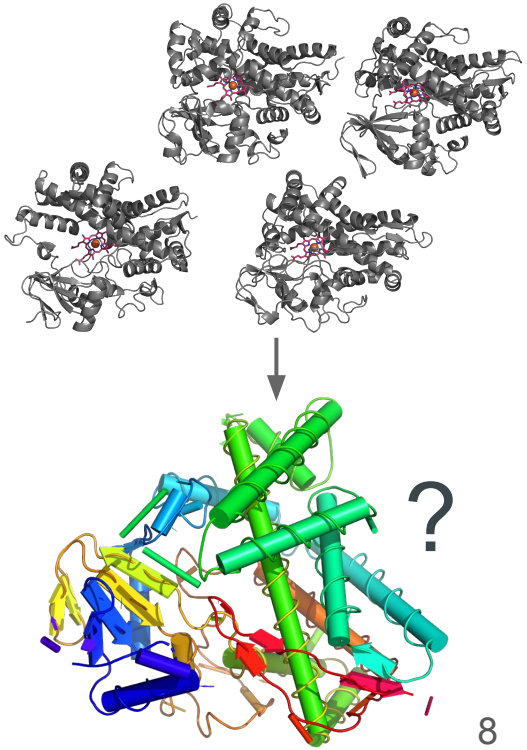
- Task: extract and visualize the characteristic features of a family



PGLIPAMVSETIRWL
ERIDAATEELLR
PELANAAEEVLRW
PYLDQVMQEILRLI
PELMGEEALRFR

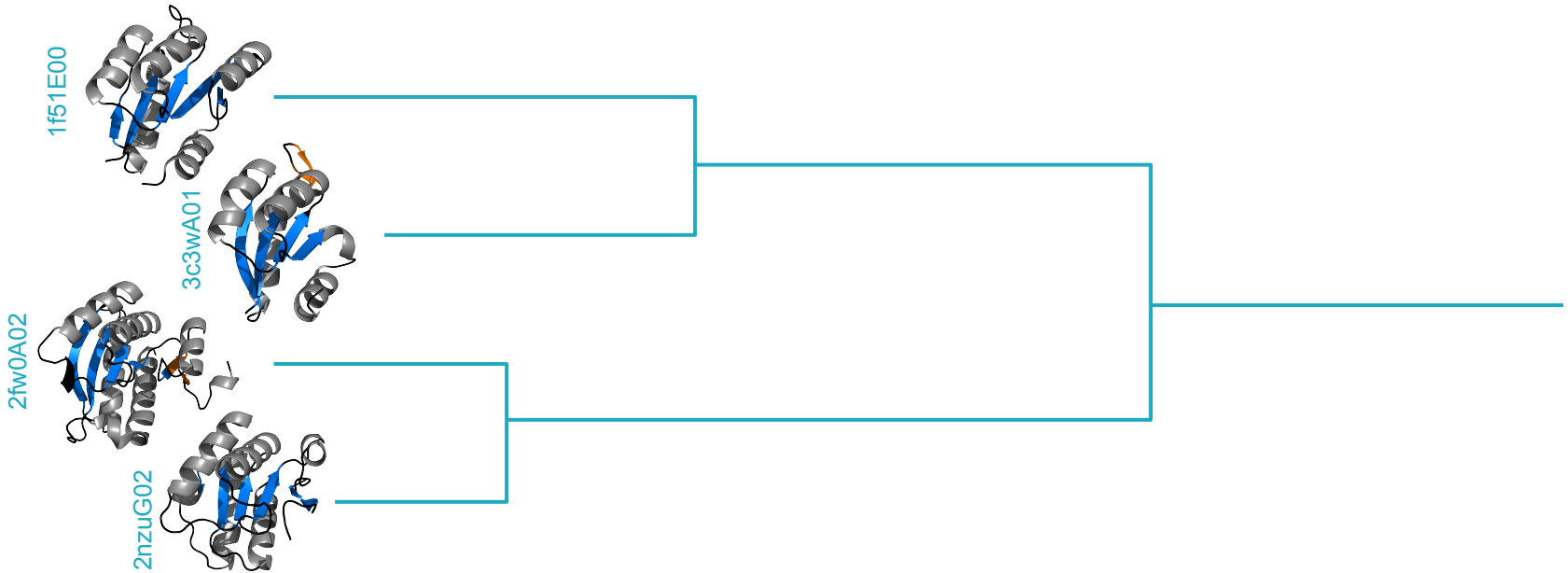
PGLIPAMVS**E**TI**R**WL
-ERIDAATE**E**LL**R**--
PELA-NAAE**E**VL**R**W-
PYLD-QVMQ**E**IL**R**LI
PELM---GE**E**AL**R**FR





Our solution: OverProt

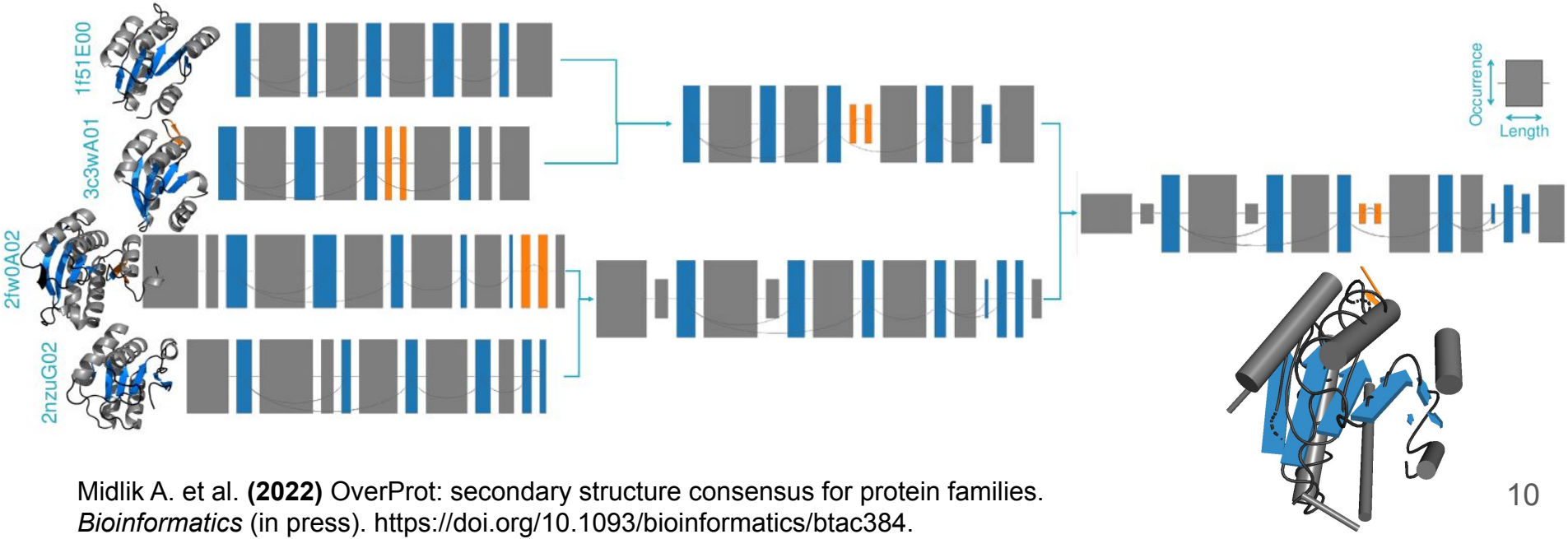
- Step 1: Align
- Step 2: Build the guide tree (agglomerative clustering of 3D domains)



Midlik A. et al. **(2022)** OverProt: secondary structure consensus for protein families. *Bioinformatics* (in press). https://doi.org/10.1093/bioinformatics/btac384.

Our solution: OverProt

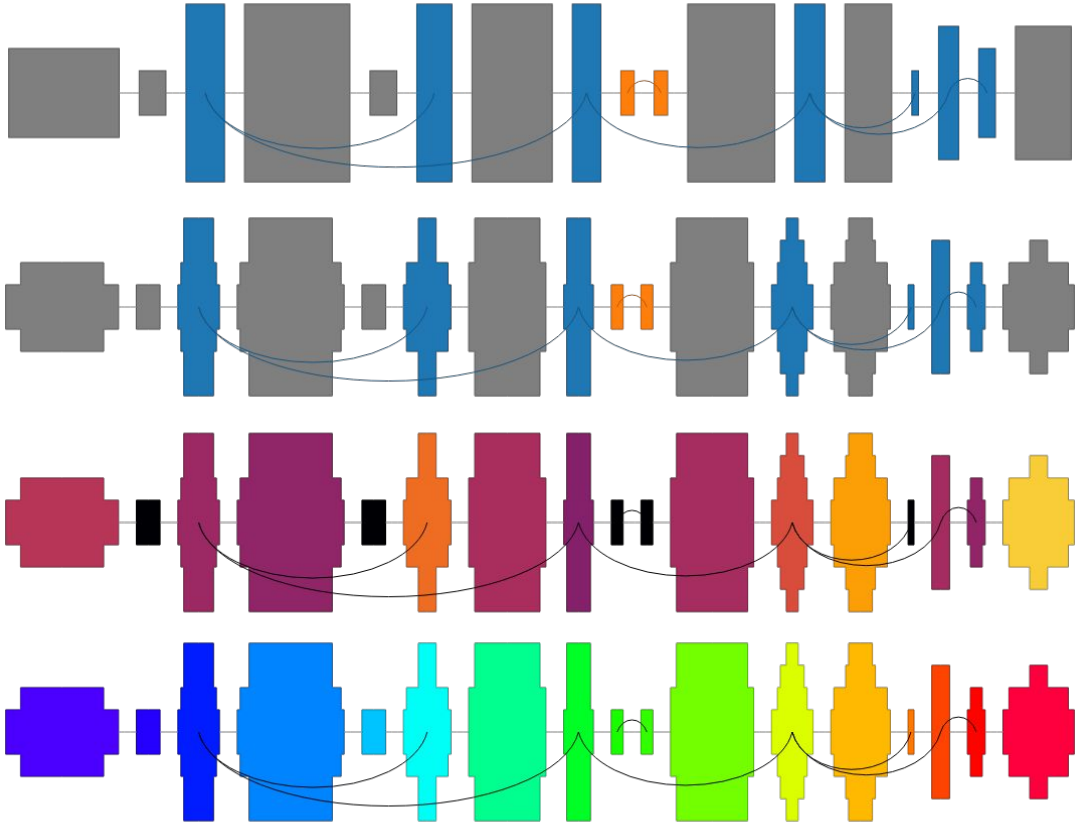
- Step 1: Align
- Step 2: Build the guide tree (agglomerative clustering of 3D domains)
- Step 3: Merge the SSEs through the guide tree



Midlik A. et al. **(2022)** OverProt: secondary structure consensus for protein families. *Bioinformatics* (in press). https://doi.org/10.1093/bioinformatics/btac384.

Alternative visualizations

- Classic
- Length distribution
- 3D variability
- Chainbow



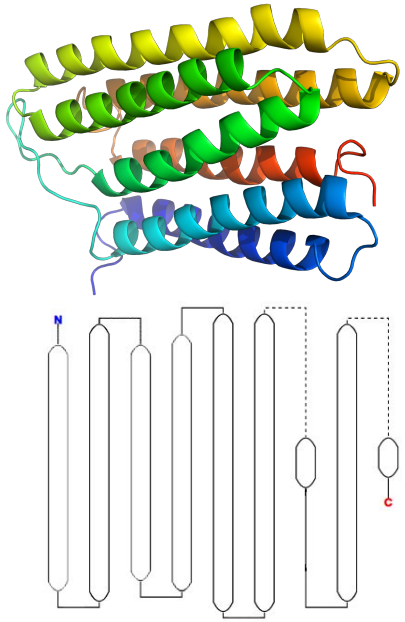
Midlik A. et al. **(2022)** OverProt: secondary structure consensus for protein families. *Bioinformatics* (in press). https://doi.org/10.1093/bioinformatics/btac384.

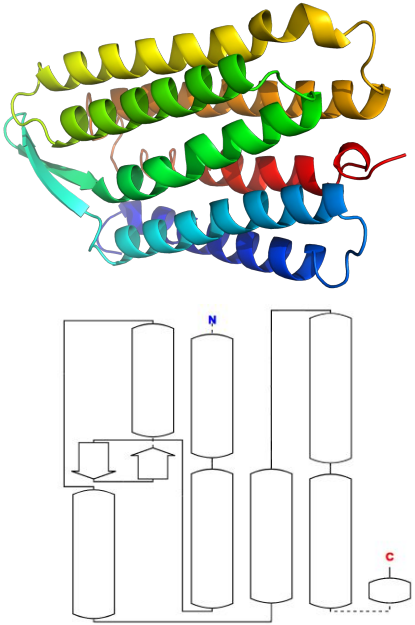
2DProts

SSE visualization in 2D diagrams

- Existing diagrams in PDBe:
  - Ignore 3D arrangement
  - Ignore protein similarity

1at9



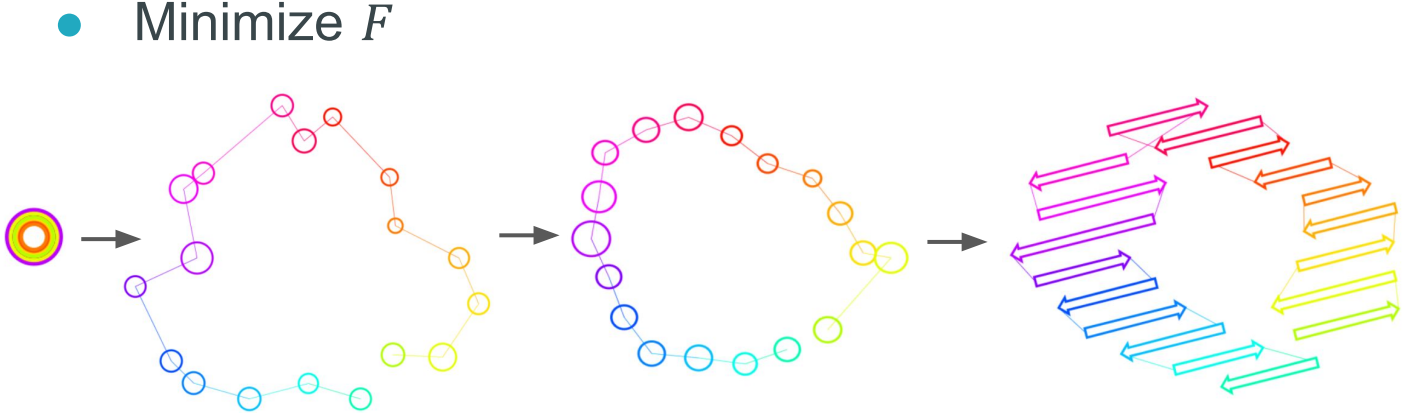


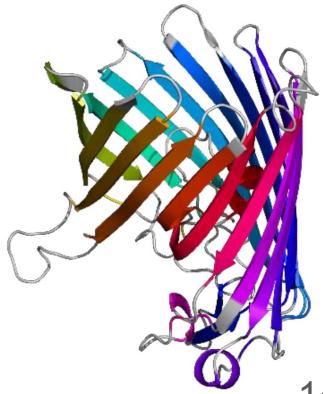
5zim

Our solution: 2DProts

- Preparation
  - OverProt creates SSE consensus for the family → annotation template
  - SecStrAnnotator annotates all protein domains in the family
- Objective function $F = \sum_{s_1,s_2 \in S_p} |d_{3D}(s_1,s_2) - d_{2D}(s_1,s_2)| \left(L(s_1) + L(s_2)\right) + \sum_{s \in S_p \cap S_0} |r(s) - r_0(s)|\, L(s) \frac{|S_p|}{20}$

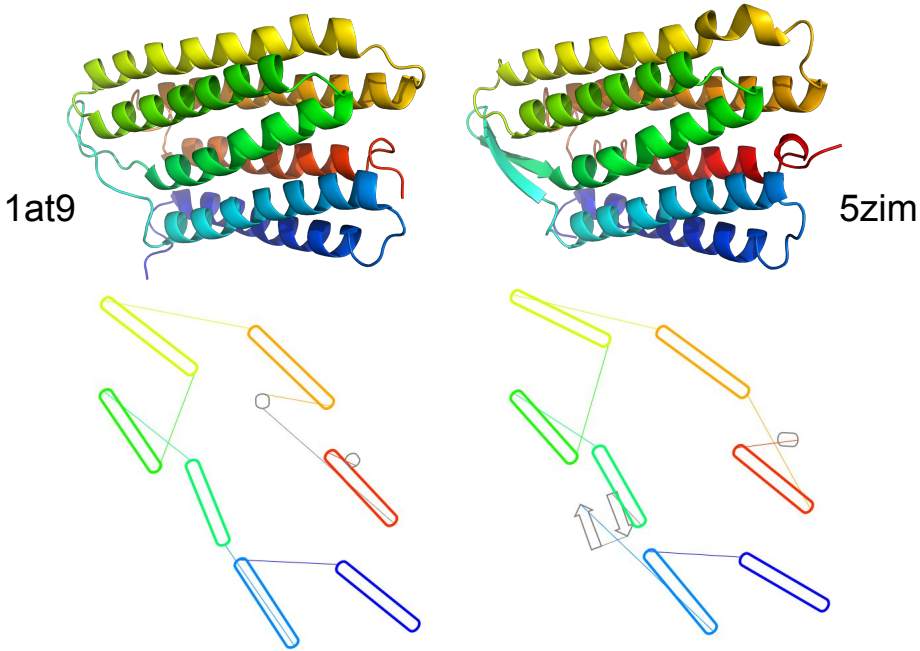
  Deviation of 2D vs 3D — Deviation of 2D vs family template 2D
- Minimize F



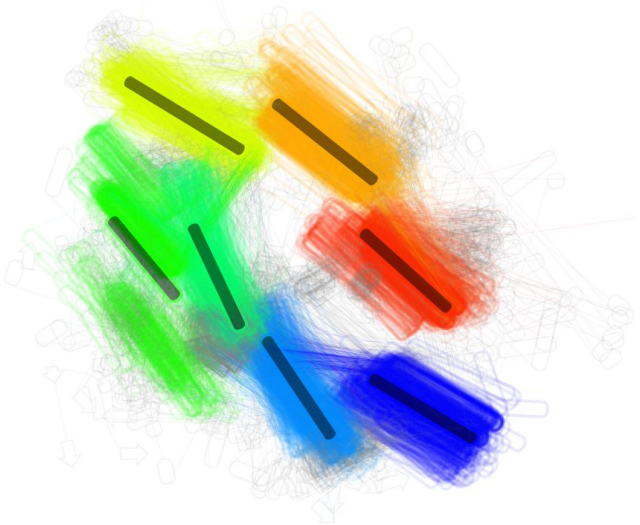


Example family

- 1.20.1070.10 Rhodopsin 7-helix transmembrane proteins



Family multi-diagram

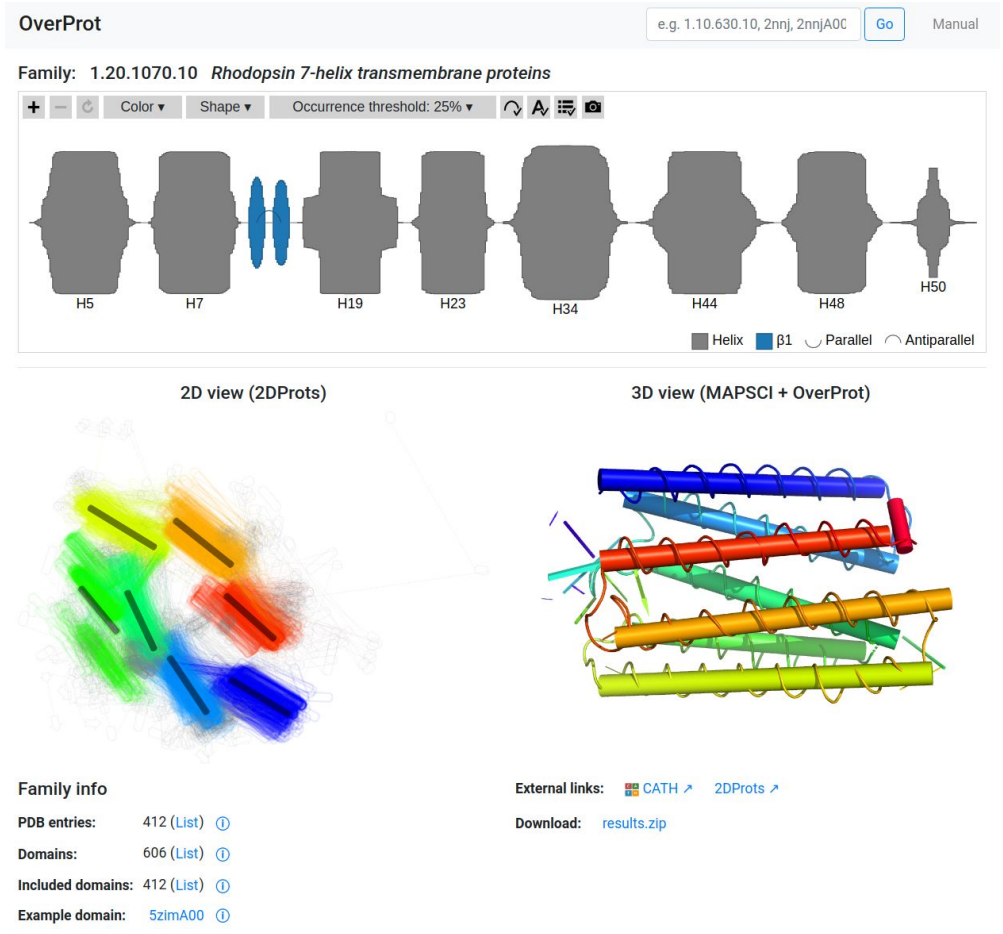


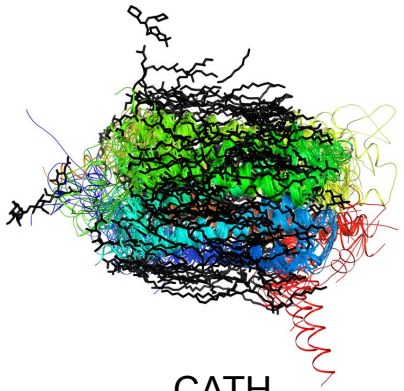
Hutařová Vařeková I. et al. (2021) 2DProts: database of family-wide protein secondary structure diagrams. *Bioinformatics*, **37**, 4599–4601.

Availability

	SecStrAnnotator	OverProt	2DProts
Desktop version	gitlab.com/midlik/SecStrAnnot2 (all platforms via .NET 6.0)	gitlab.com/midlik/overprot (all platforms via Docker/Podman)	
Online version	sestra.ncbr.muni.cz – Cytochrome P450 template – Custom template – Detect SSEs (no annotation)	overprot.ncbr.muni.cz Job calculations up to 500 protein domains	2dprots.ncbr.muni.cz Job calculations up to 30 protein domains
Database	1 family 1 855 domains	6 631 families > 470 000 domains (2.5 CPU days)	6 631 families > 470 000 domains (25 CPU days)
Integrations			CATH cathdb.info

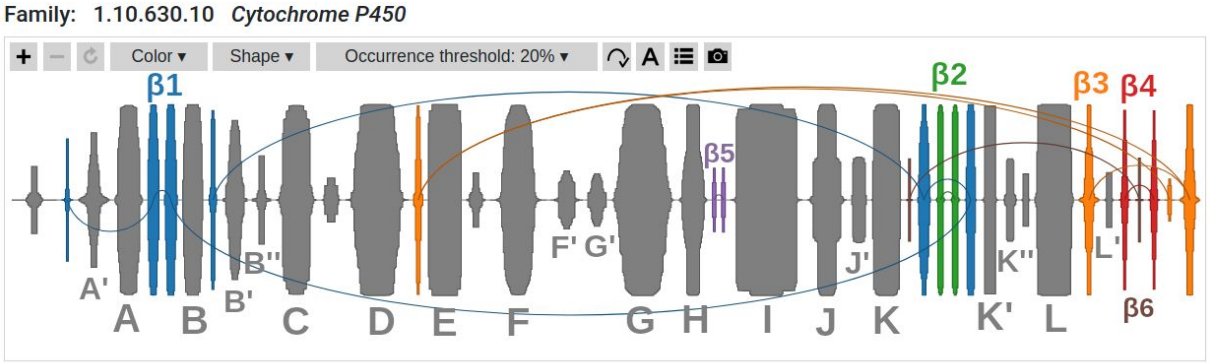
All together – overprot.ncbr.muni.cz

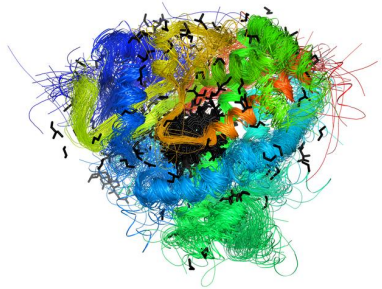




CATH superimposition

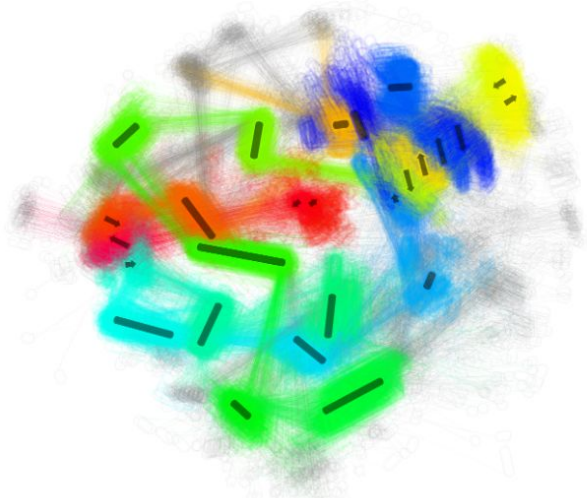
All together – overprot.ncbr.muni.cz



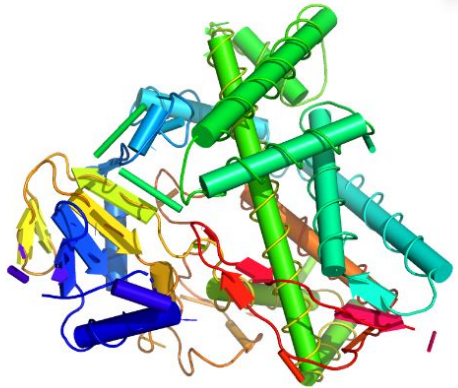


CATH superimposition

2D view (2DProts)

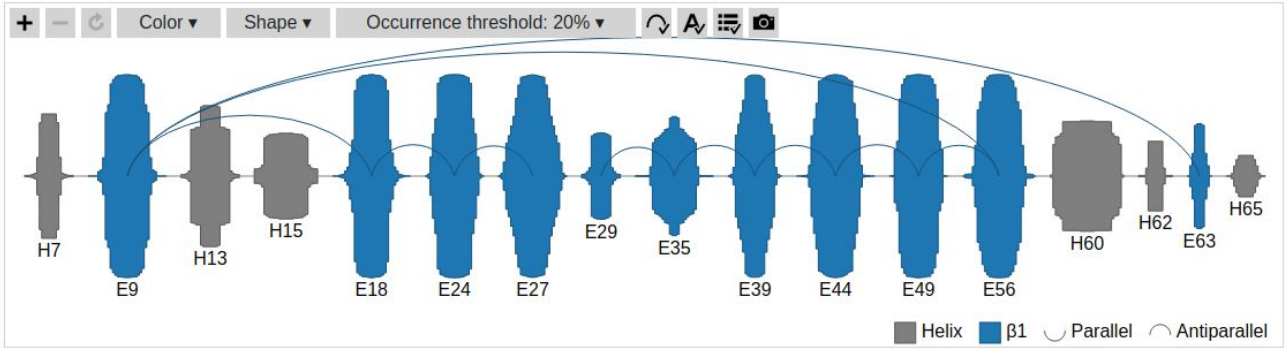


3D view (MAPSCI + OverProt)



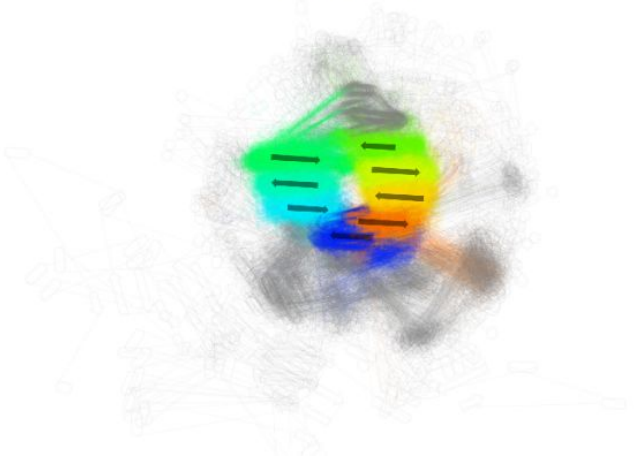
All together – overprot.ncbr.muni.cz

Family: 2.40.128.20 *Calycin beta-barrel core domain*

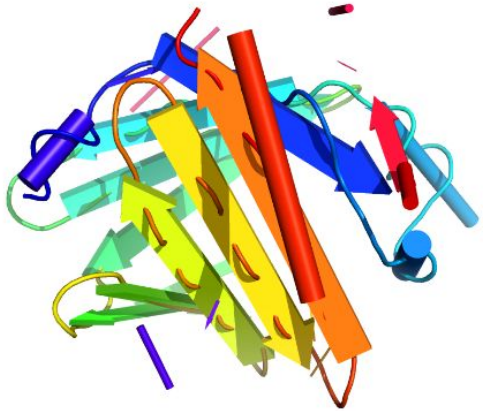


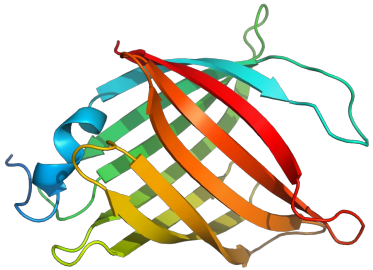
8-stranded subgroup (3k3l)

2D view (2DProts)



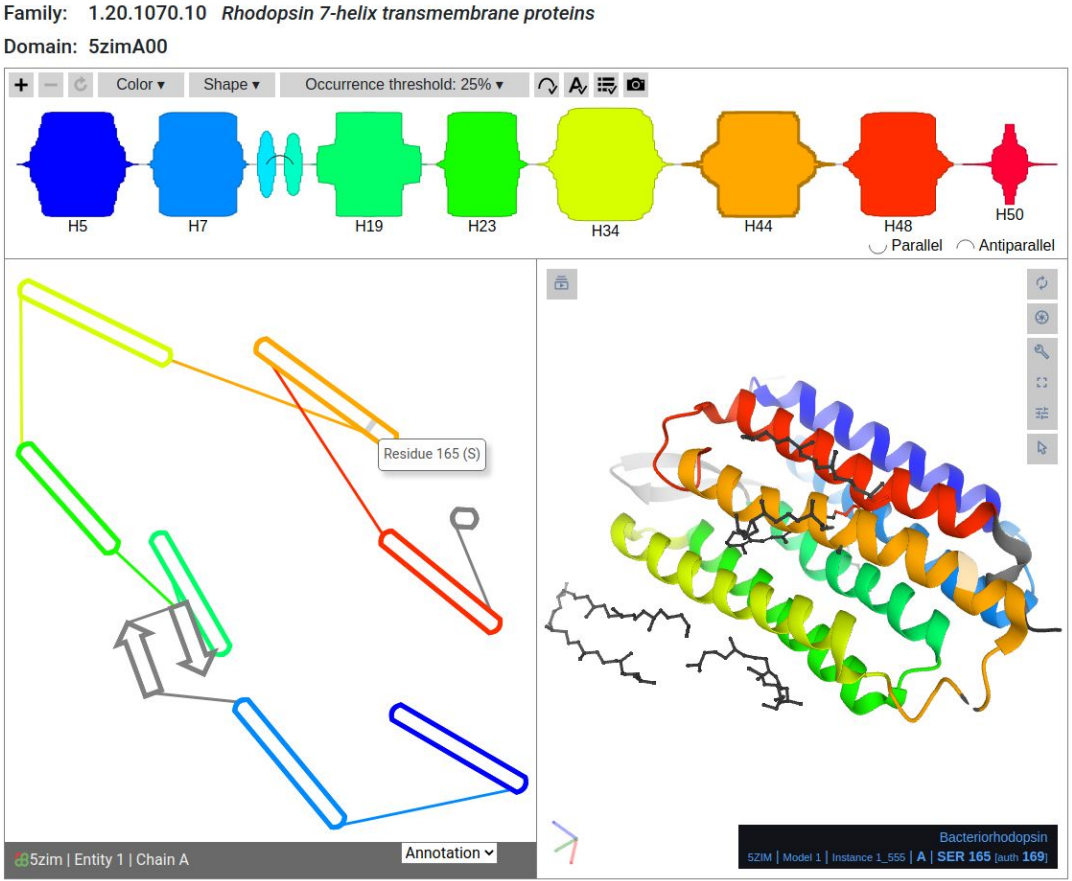
3D view (MAPSCI + OverProt)





10-stranded subgroup (3wjb)

Integrated view



← Family consensus

← Selected domain

https://overprot.ncbr.muni.cz/domain_view?family_id=1.20.1070.10&domain_id=5zimA00

Conclusion

- **SecStrAnnotator** – SSE annotation
- **OverProt** – SSE consensus + 1D visualization
- **2DProts** – SSE 2D visualization

Future plans

- **OverProt** – Integration with PDBe-KB, CATH
- **2DProts** – Inclusion of ligands
- Integration of AlphaFold DB via CATH cooperation

Acknowledgement

CEITEC
Central European Institute of Technology
BRNO | CZECH REPUBLIC

Structural bioinformatics group, Masaryk University

- Ivana Hutařová Vařeková, Jan Hutař – **2DProts**
- Aliaksei Chareshneu – **Integrated view**
- Radka Svobodová, Jaroslav Koča

MUNI SCI National Centre for Biomolecular Research

MUNI FI

Palacký University

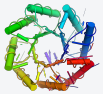
- Karel Berka

elixir CZECH REPUBLIC

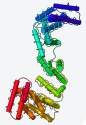
cesnet metacentrum

CATH

Thank you for your attention



overprot.ncbr.muni.cz



sestra.ncbr.muni.cz

midlik@mail.muni.cz

2dprots.ncbr.muni.cz

MUNI SCI National Centre for Biomolecular Research

CEITEC