OverProt: Overview of Secondary Structure Consensus in Protein Families

Adam Midlik [1,2], Ivana Hutařová Vařeková [1,2,3], Jan Hutař [1,2,3], Jaroslav Koča [1,2], Radka Svobodová [1,2], Karel Berka [4]

✉ midlik@mail.muni.cz

[1] CEITEC – Central European Institute of Technology, Masaryk University, Kamenice 5, Brno
[2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, Brno
[3] Faculty of Informatics, Masaryk University, Botanická 68a, Brno
[4] Department of Physical Chemistry, Faculty of Science, Palacký University, 17. listopadu 12, Olomouc

INTRODUCTION

Secondary structure elements (SSEs) provide a deep insight into the architecture of a protein. Therefore, a figure depicting a sequence of SSEs for individual proteins is used in key structural databases (PDBe, RCSB PDB). Similarly, for a whole protein family, a consensus of SSEs can be constructed. This consensus shows the general protein fold of the family and its structural variation. It can also be used as an annotation template for our previously developed program SecStrAnnotator. This allows annotation of SSEs in any family and unlocks the possibility of automated annotation of the key regions (e.g. active sites and channels) based on their position relative to the SSEs.

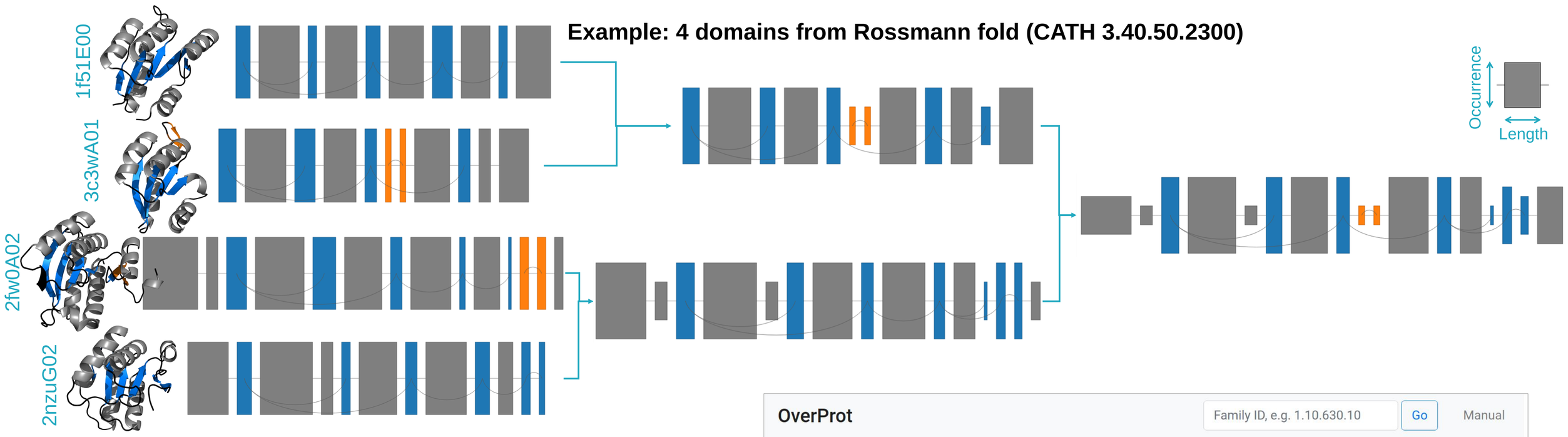
Our OverProt algorithm allows fully automated construction of the SSE consensus. The OverProt Server offers the precomputed SSE consensus for each CATH protein family.

ALGORITHM

In order to construct the SSE consensus, the SSEs from individual protein family members must be matched together. Some proteins within a family often miss some SSEs, therefore the matching is not straightforward. Current version of the algorithm allows consensus computation even for the largest families. All 6631 CATH families can be processed in 84 CPU hours in total.

OverProt algorithm outline:

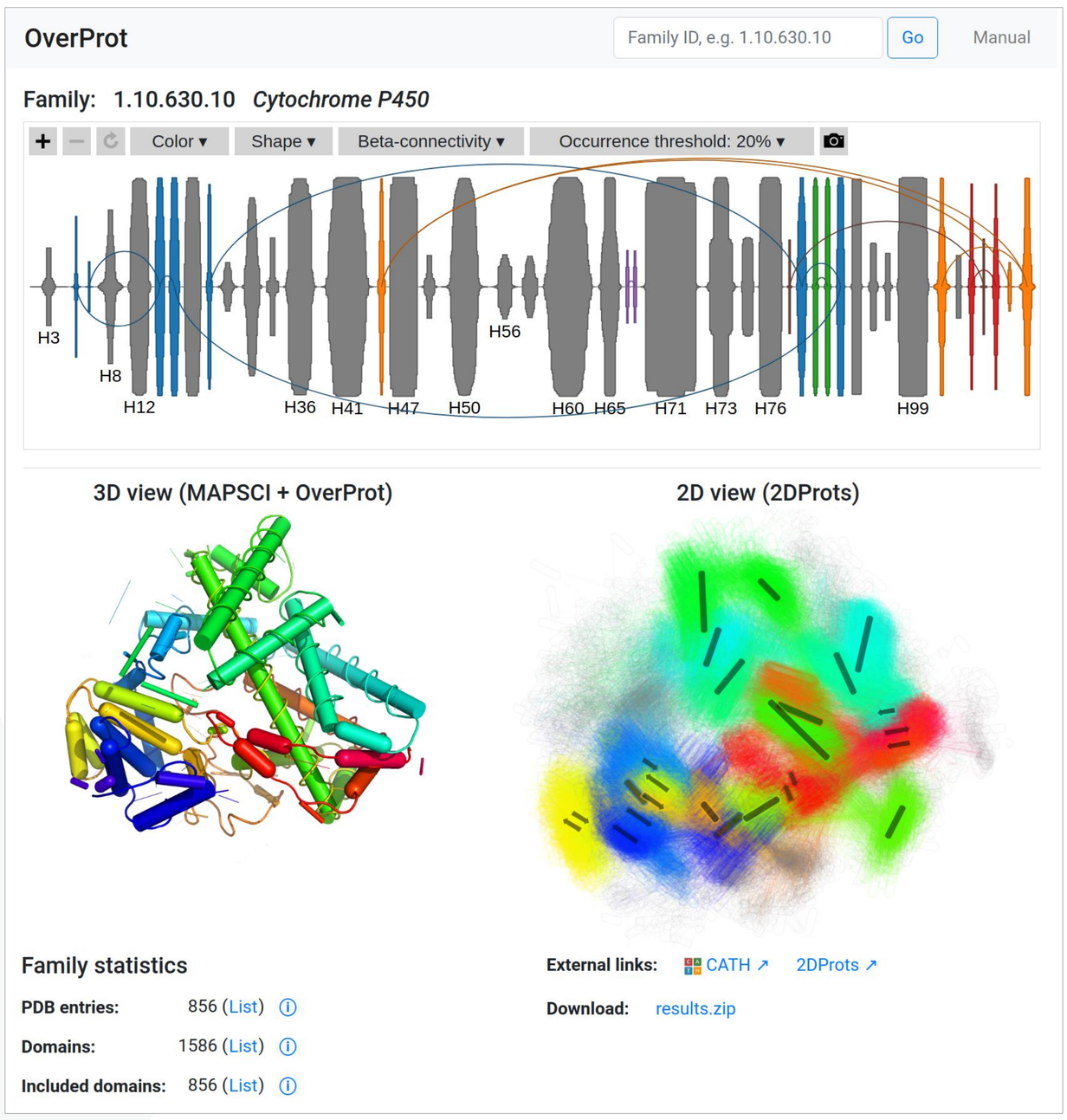
1. Remove duplicates: when a family contains multiple protein domains from the same PDB entry, select just the first domain
2. Agglomerative clustering of the domains in the family → guide tree
3. Place the SSEs into the guide tree leaves (leaf ~ consensus of 1 domain)
4. In each node of the guide tree, merge consensus from the two branches
5. Tree root ~ consensus of the whole family



WEB SERVER

The OverProt Server (https://overprot.ncbr.muni.cz) provides precomputed consensus for each of the 6631 CATH families. User-defined queries allow consensus computation for custom protein families (up to 500 protein domains).

The consensus is visualized by the interactive OverProt Viewer. The height of each consensus SSE shows its occurrence, its shape reveals the length distribution. β-strands from the same β-sheet have the same colour, helices are grey. The SSEs with low occurrence (below 20%) are hidden. The arcs show the connectivity of the β-strands. Several alternative modes of visualization are available.



ACKNOWLEDGEMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project CEITEC 2020 [LQ1601]; ELIXIR-CZ research infrastructure project including access to computing and storage facilities [LM2018131].