SecStrConsensus – outlining the common secondary structure features in protein families

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

✉ 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

Protein structures, deposited in the Protein Data Bank, can be classified into **protein families** based on their similarity. Systematic study of these families is gaining importance and can yield interesting research results.

Every protein family has a set of characteristic **secondary structure elements** (SSEs, namely helices and β-strands). Their arrangement is well defined and relatively consistent throughout the whole family. Still there are some variations and a single structure is not enough to represent the whole family of structures. A family of amino acid sequences can be compressed into a consensus sequence and visualized by a sequence logo, which shows the **essential features of the family**. For secondary structure, such an approach is not yet available, therefore we are working on its implementation.

In this work, we present our progress in the development of **SecStrConsensus** (previously Ubertemplate) – a tool for extracting the **secondary structure consensus** for a given protein family. This consensus gives an overview of the family and 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. We also suggest an interactive visualization of the consensus in a web browser.

METHODS – CLUSTERING

Older approach:

Agglomerative clustering of all SSE from all protein domains in the family

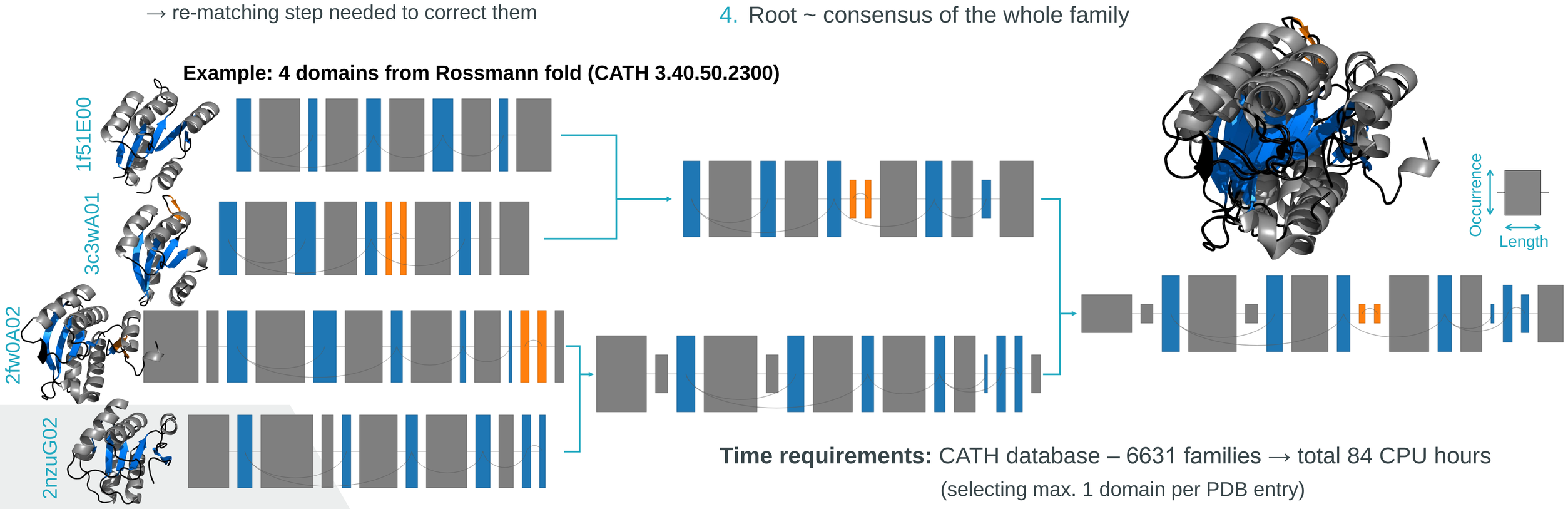
☹ Matrix of SSE distances – $(n_{domains} \cdot n_{SSEs/domain})^2$
→ time and memory issues

☹ Greedy approach → mistakes
→ re-matching step needed to correct them

Newer approach:

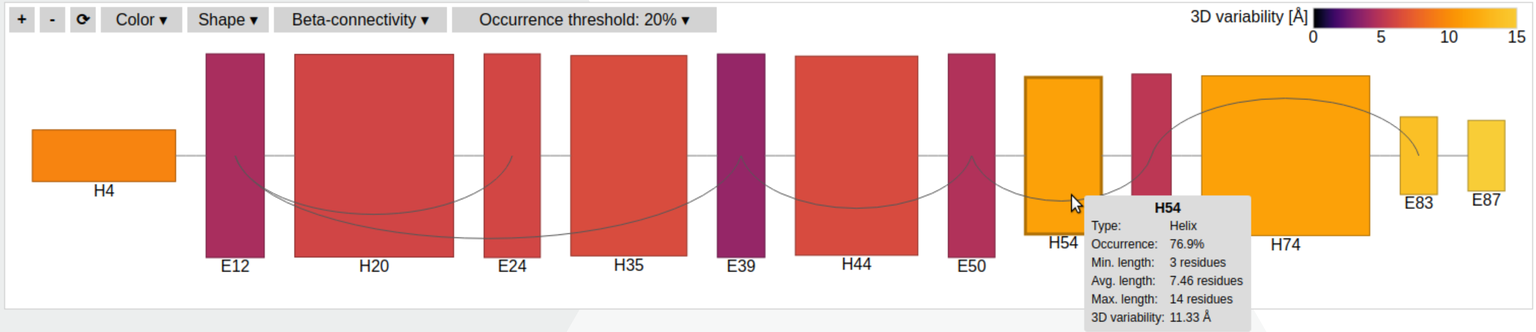
1. Agglomerative clustering of the protein domains in the family → guide tree
   Matrix of domain distances – $n_{domains}^2$ → still the most expensive step
2. Place the SSEs into the guide tree leaves (leaf ~ consensus of 1 domain)
3. In each node of the guide tree, merge consensus from the two branches
4. Root ~ consensus of the whole family

Example: 4 domains from Rossmann fold (CATH 3.40.50.2300)



Time requirements: CATH database – 6631 families → total 84 CPU hours
(selecting max. 1 domain per PDB entry)

Rossmann fold (CATH 3.40.50.2300) – whole family – 3D variability coloring



Try SecStrConsensus Viewer at https://is.muni.cz/www/midlik/secstrconsensus

VISUALIZATION

SecStrConsensus Viewer provides interactive visualization and can be easily embedded in a webpage.

In the default view, each consensus SSE is shown as a rectangle – β-strands from the same β-sheet have the same colour, while all helices are grey. Only SSEs with occurrence above 20% are shown. The arcs show 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]; European Regional Development Fund – projects ELIXIR-CZ [CZ.02.1.01/0.0/0.0/16_013/0001777]. Adam Midlik is Brno Ph.D. Talent Scholarship Holder – Funded by the Brno City Municipality.