Creation and visualization of secondary structure consensus for protein families

Adam Midlik [1,2], Ivana Hutařová Vařeková [1,2,3], Jan Hutař [1,2], 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, Regional Centre of Advanced Technologies and Materials, 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 currently missing, therefore we work its implementation.

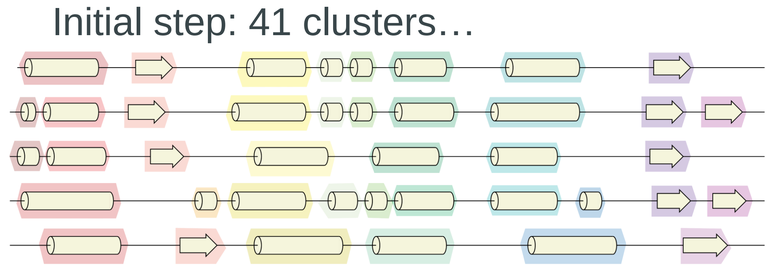
Our tool **Ubertemplate** produces a **secondary structure consensus**, which gives an overview of the family and can also be used as an annotation template for our previously developed program SecStrAnnotator. This allows annotation of SSEs in any family and unlocks the possibility of automated annotation of the key regions (e.g. active sites and channels) based on their position relative to the SSEs.

METHODS – CLUSTERING

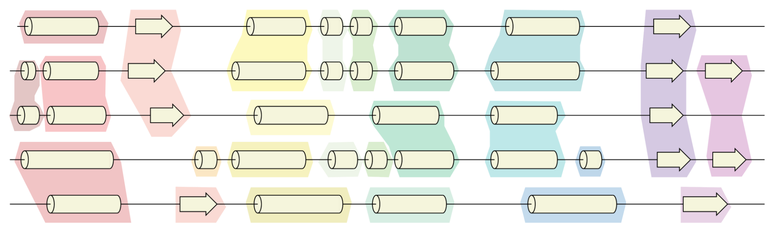
The algorithm is based on agglomerative clustering and consists of 3 main steps:

1. **Initialization:** Each SSE in each protein domain in the family represents an individual cluster containing 1 SSE.
2. **Iterative step:** The two nearest clusters are merged into a larger cluster. Repeat this step while the following constraints can be fulfilled:
   - Each cluster contains only helices or only β-strands.
   - Each cluster contains at most one SSE from each protein domain.
   - The clusters are partially ordered (“if A is before B then B isn’t before A”).
3. **Final re-matching:** The average SSE position of each cluster from step 2 is taken as a seed. Then all SSEs are matched to their nearest seed (considering also seed weight (cluster size)). This is repeated several times to obtain the best clustering. Each final cluster represent one **consensus SSE**. Occurrence of the consensus SSE is given by the cluster size.

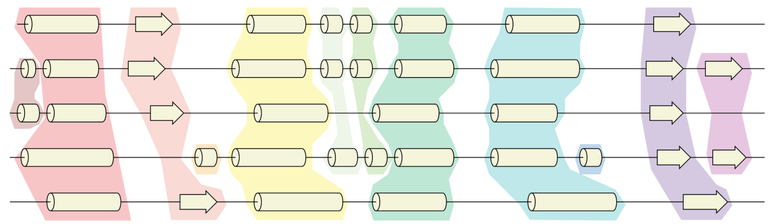
Initial step: 41 clusters…



…During iteration: 25 clusters…



…Final step: 12 clusters

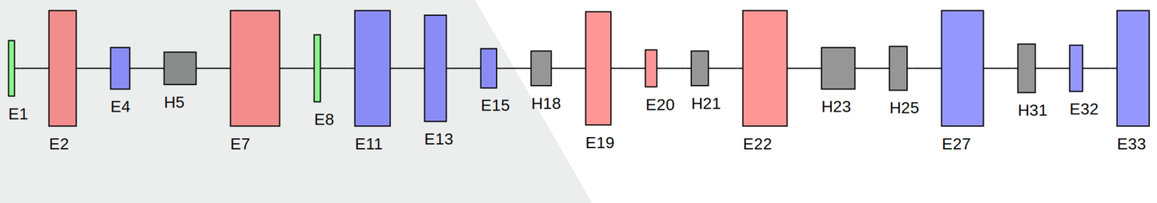


METHODS – VISUALIZATION

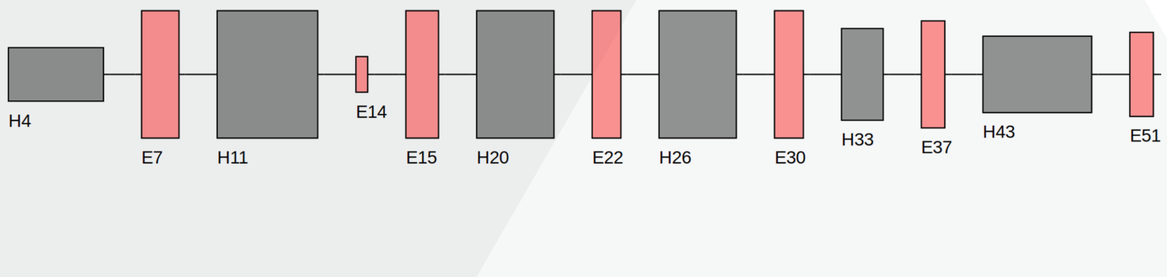
Each consensus SSE is shown as a rectangle – β-strands from the same β-sheet have the same colour, while all helices are grey. Only SSEs with occurrence above 20% are shown. Several alternative modes of visualization are available.

RESULTS – SELECTED PROTEIN FAMILIES

Immunoglobulin (CATH 2.60.40.10) – two β-sheets

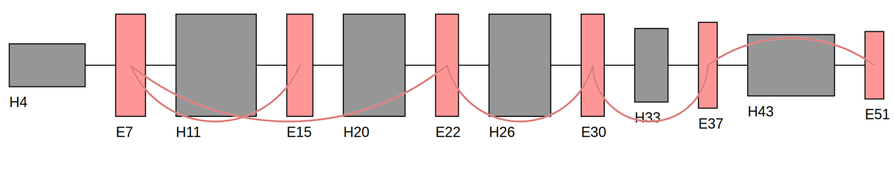


Rossmann fold (CATH 3.40.50.2300) – a β-sheet with helices around

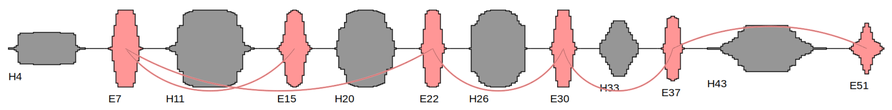


Alternative visualizations (Rossmann fold):

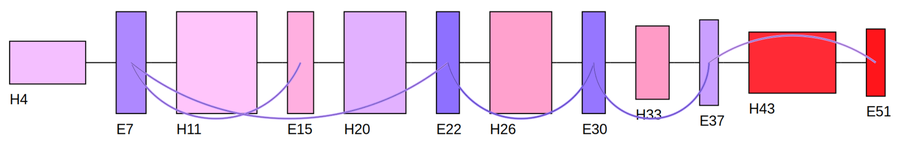
- β-connectivity (up = parallel, down = antiparallel)



- Distribution of SSE length (number of residues)



- Spatial variability (conserved – variable)



ACKNOWLEDGEMENT

This research was supported by Ministry of Education, Youth and Sports of Czech Republic as a part of projects CEITEC 2020 (LQ1601), ELIXIR CZ research infrastructure project (LM2015047), and ELIXIR-CZ: Budování kapacit (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.