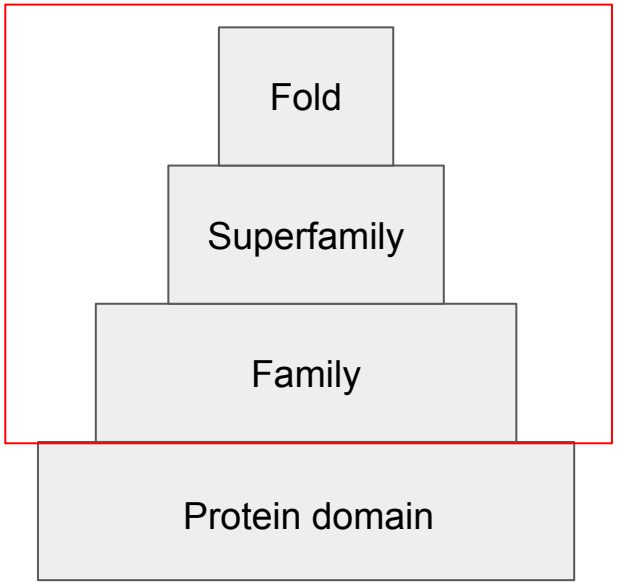
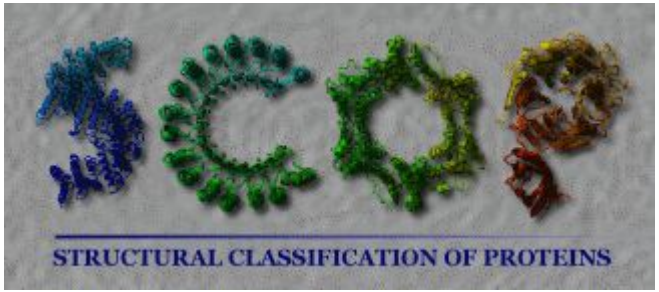
A Practical Introduction to Reproducible Computational Workflows

Practical: SCOP Class Prediction

Problem definition

- **Toy problem:** classification of proteins

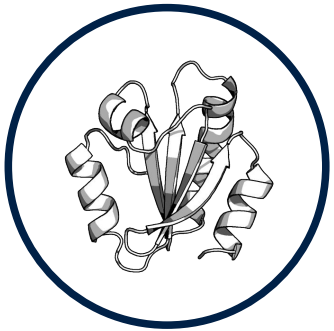




Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540

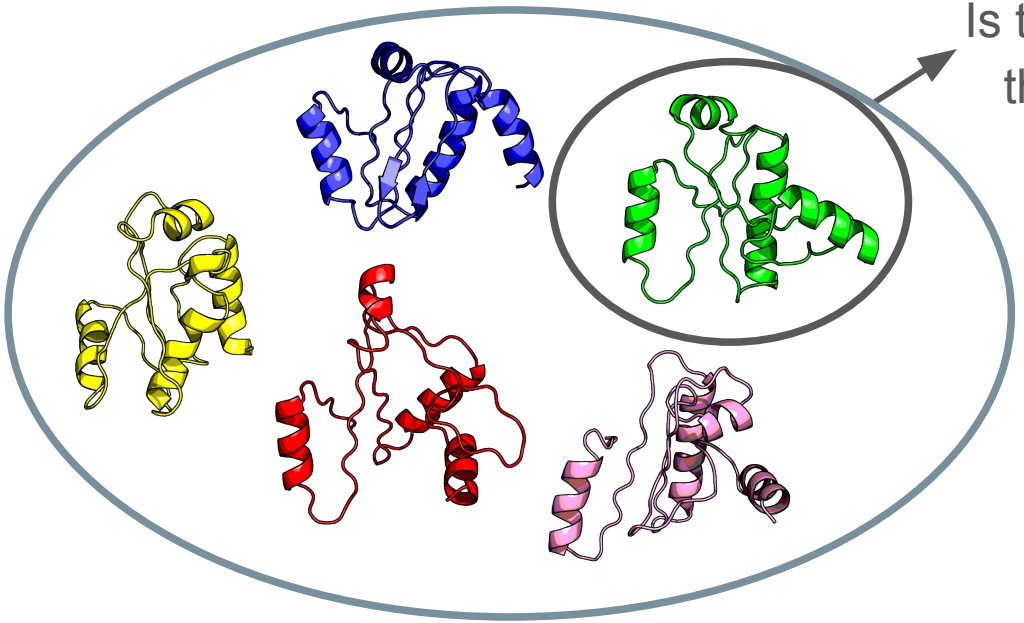
Dataset

- Astral data set (06.02.2016 build): contains 13,710 proteins belonging to 4,535 distinct folds



Target protein

Template proteins



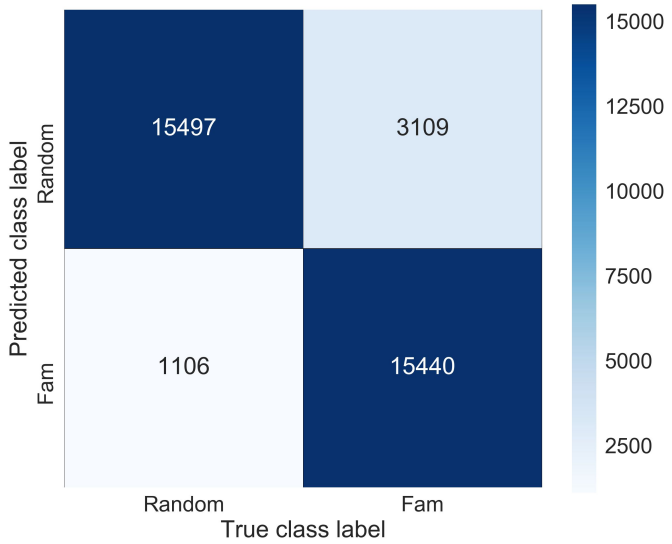
Is the template from the same SCOP classification?

Inputs:
8 pairwise sequence-based features between target and template protein

Results

- Question: Family or not?
- Results: 88.0% accuracy

Confusion matrix



ROC curve