DOI 10.18699/BGRSSB-2018-077

Protein structural domain prediction via machine learning approach

D. Iakovlev*, A. Kobchenko, E. Semina
Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia
** e-mail: yakovlevd@niboch.nsc.ru*

Key words: protein domain, protein visualization, machine learning, clustering

Motivation and Aim: Amount of solved protein structures in databases such as PDB is growing incredibly fast, making manual investigations in this field more and more challenging. One of a basic and, usually, manual steps of protein analysis is a structural domains annotation. A concept sometimes taken as a rough working definition of a structural domain is that, if excised, the domain should remain folded as a stable structure [1]. Therefore, residues in protein domains are distributed more densely than averagely in protein and would be detectable by clustering algorithms. Despite there are many tools for protein structure analysis and visualization, no one of them can automatically split protein into domains using only structural information (e. g. a PDB file). Using protein architecture database such as CATH or SCOP2 is also difficult if we deal with proteins that do not have annotated homologues in there. Some methods for automatic detection of protein domains have been already developed earlier [1] but we improved them using modern algorithms and computational approaches.
Methods and Algorithms: Training sample of structural domains boundaries was obtained from CATH database [2]. Boundaries were tested to not cross secondary structure features (α-helices and β-sheets), yielding a set of approximately 130000 marked up polypeptide chains. To predict structure domains we used clusterization of C_α atoms of polypeptide chain by BIRCH [3] algorithm with some *ad hoc* modifications: handling of helix/sheet integrity and amino acid hydrophobicity. To improve clustering precision neural network was used for optimization of hyperparameters: BIRCH branching factor and threshold and post-clustering division factor. Quality of domain prediction was estimated by scoring function which matches each predicted domain with known ones and counts a proportion of correctly predicted amino acid residues.
Results: A new machine learning-based method for protein structural domain prediction was implemented. Predictive power of the clustering model was proved by cross-validation technique.
Availability: A web service for protein domain prediction and visualization is available at protein-clustering.ru
Acknowledgements: We express our gratitude to Yuri Vyatkin for formulation of the problem and curation during our investigation.

References

1. Taylor W.R. (1999) Protein structural domain identification. Protein Engineering. 12(3):203-216.
2. Sillitoe I. et al. (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 43(D1):D376-D381.
3. Zhang T., Ramakrishnan R., Livny M. (1996) BIRCH: an efficient data clustering method for very large databases. Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD ‘96:103-114.