Protein structural domain prediction via machine learning approach

D. Iakovlev*, A. Kobchenko, E. Semina

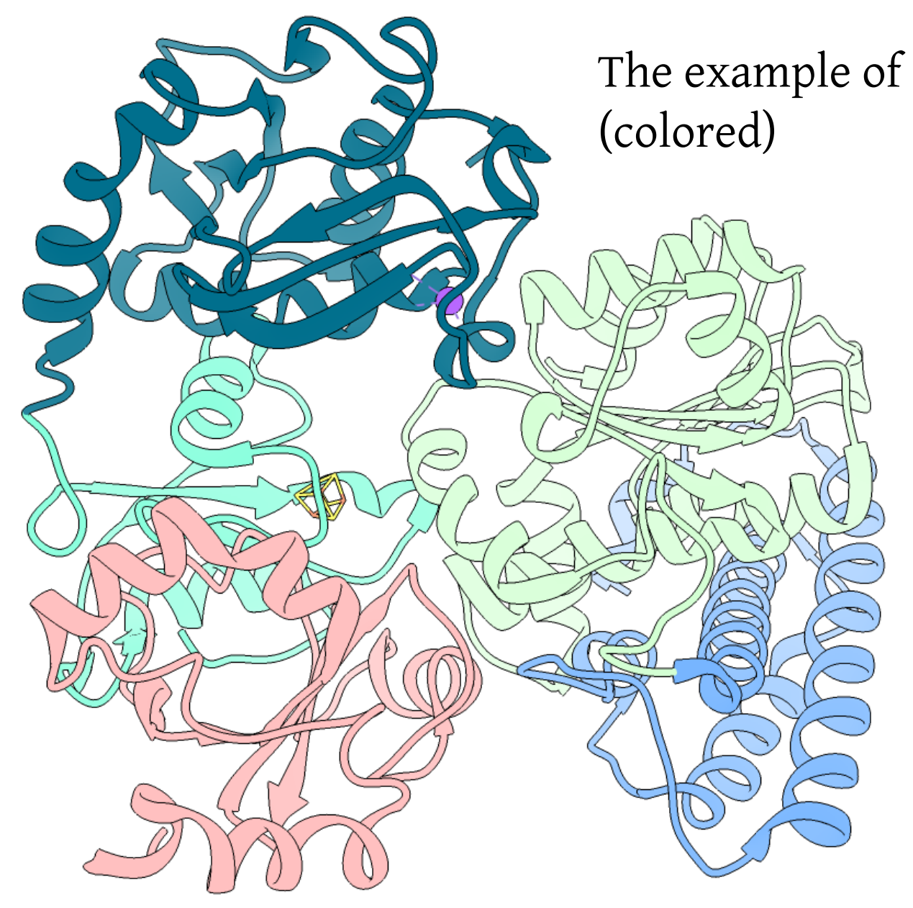
** Institute of chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia*

Novosibirsk State University
*THE REAL SCIENCE

Motivation and Aim

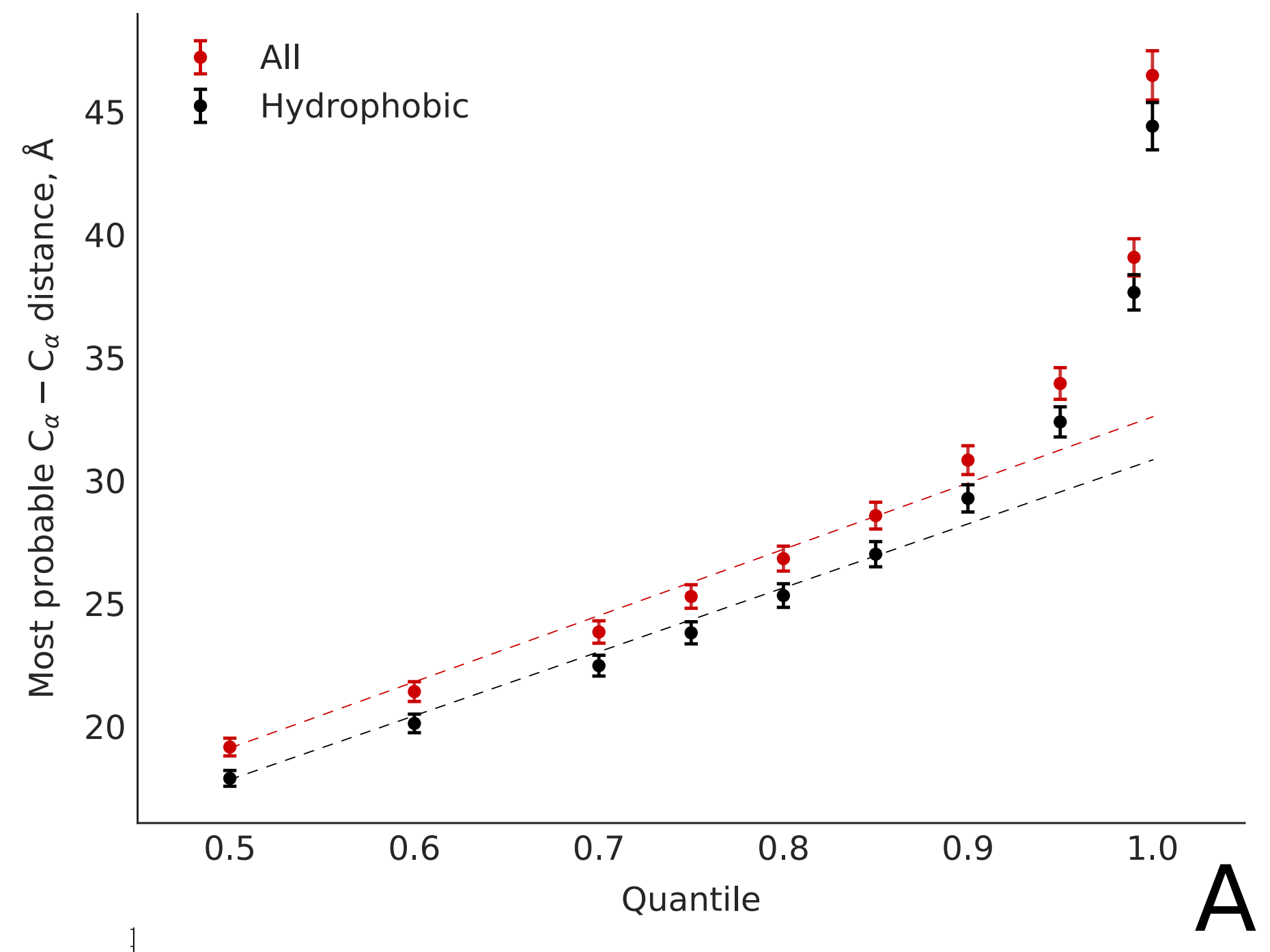
Amount of solved protein structures in databases such as PDB is growing incredibly fast, making manual investigations in this field more and more challenging. One of a basic and, usually, manual steps of protein analysis is a structural domains annotation. A concept sometimes taken as a rough working definition of a structural domain is that, if excised, the domain should remain folded as a stable structure [1].

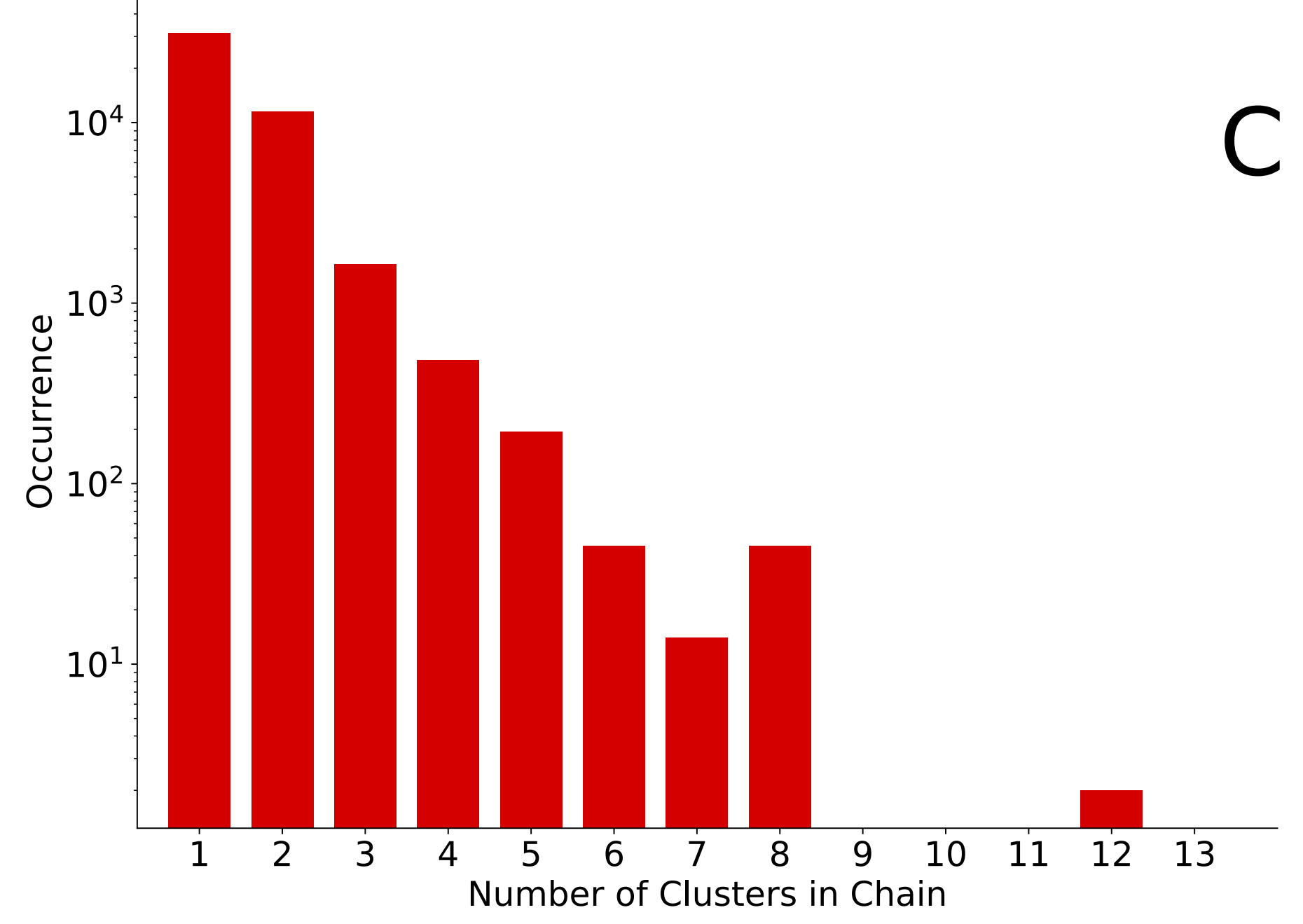
Therefore, residues in protein domains are distributed more dense than averagely in protein and would be detectable by clustering algorithms. Despite there are many tools for protein structure analysis and visualization, no one of them can automatically split protein into domains using only a structural information (e.g. a PDB file). Using protein architecture database such as CATH or SCOP2 is also difficult if we deal with protein, 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.



The example of protein domains (colored)

Initial Dataset Analysis

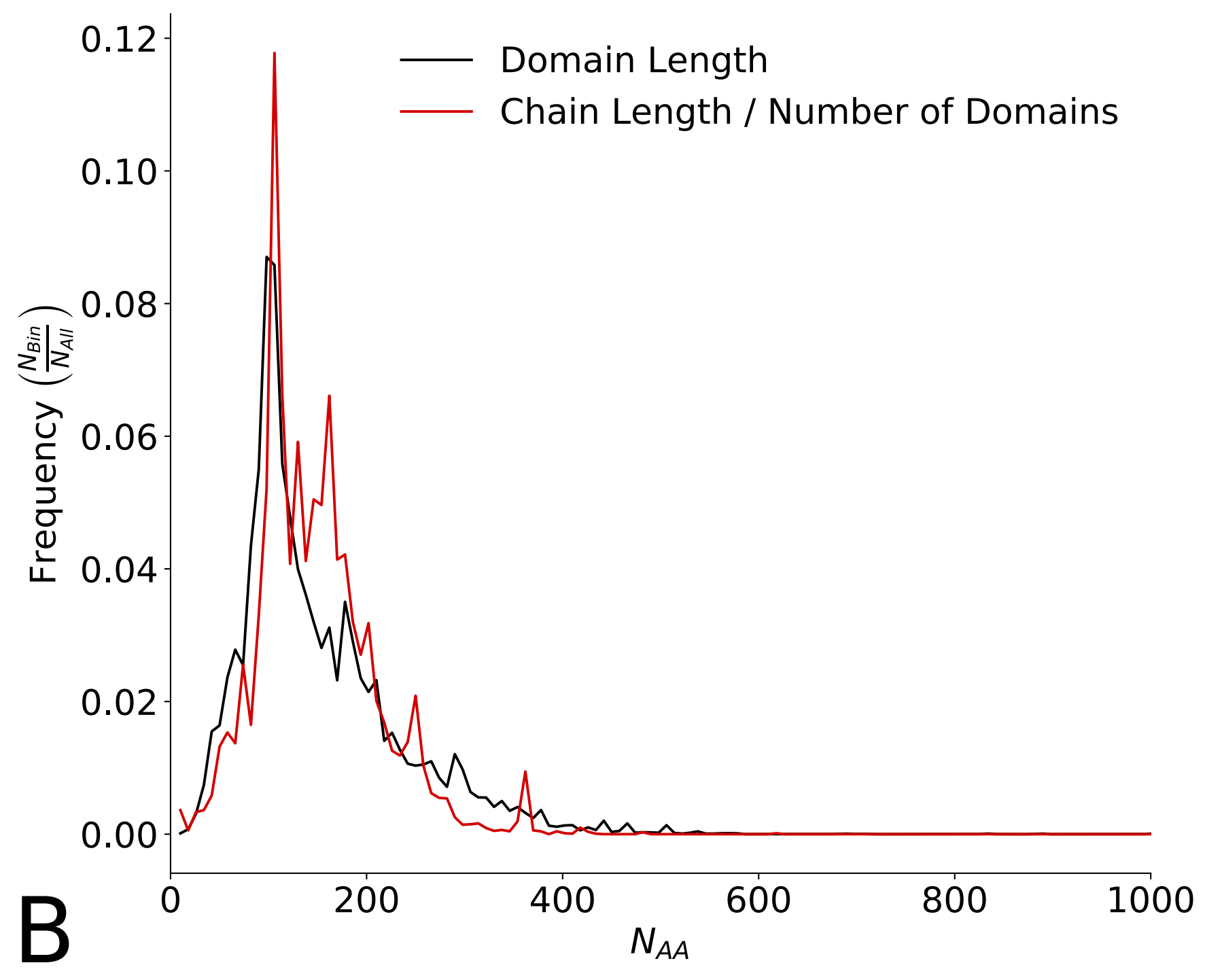




Finally, number of domains per chain are distributed very unevenly: about 75 % of chains contain only one domain and about 25 % of chains are two-domained (*fig. C*).

As the number of domains is highly correlated with square root of chain length (*fig. D*), this feature is the first obvious candidate for input parameter for further neural network.

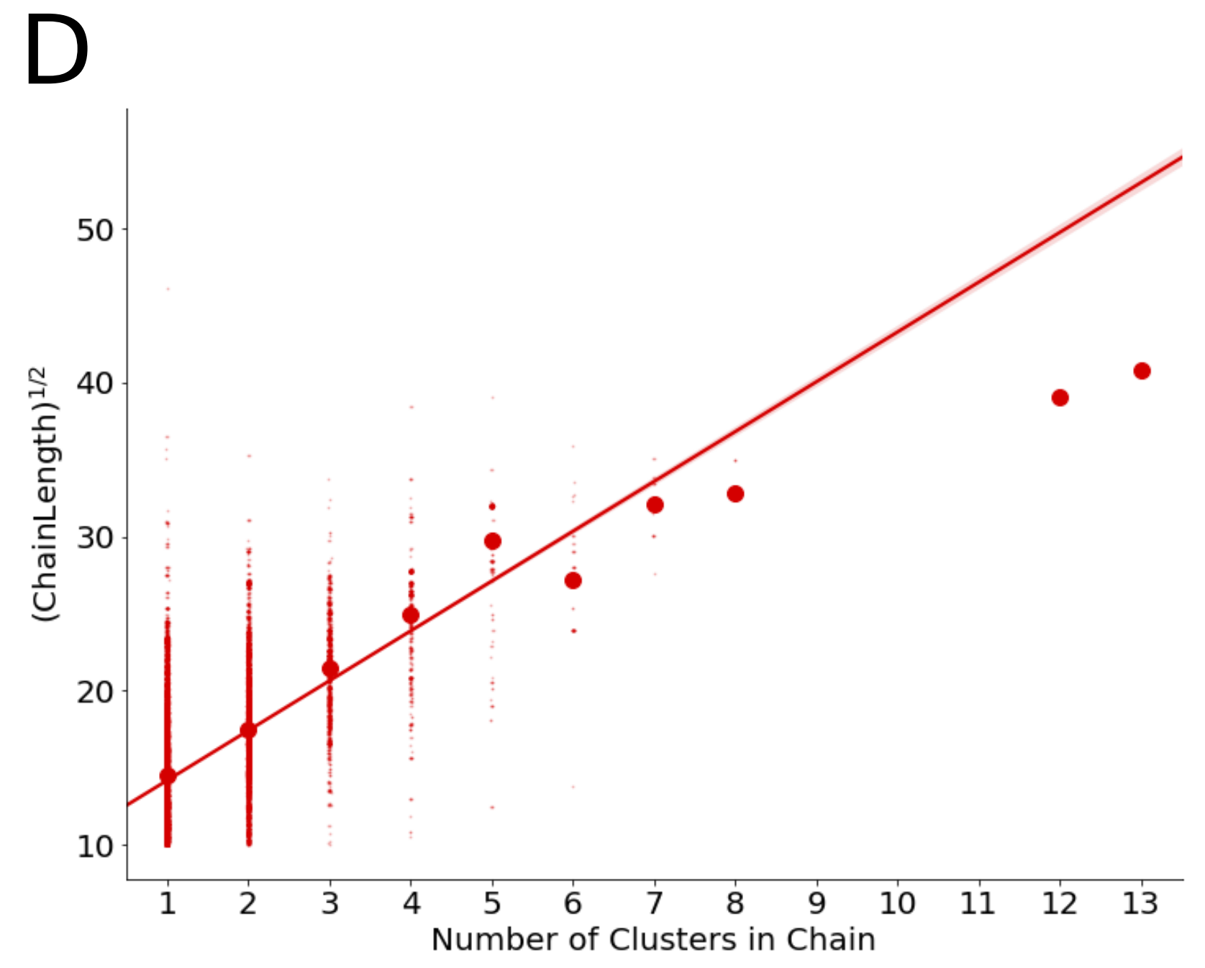
Some other interesting features are chain volume and chain minimum volume enclosing ellipsoid parameters.



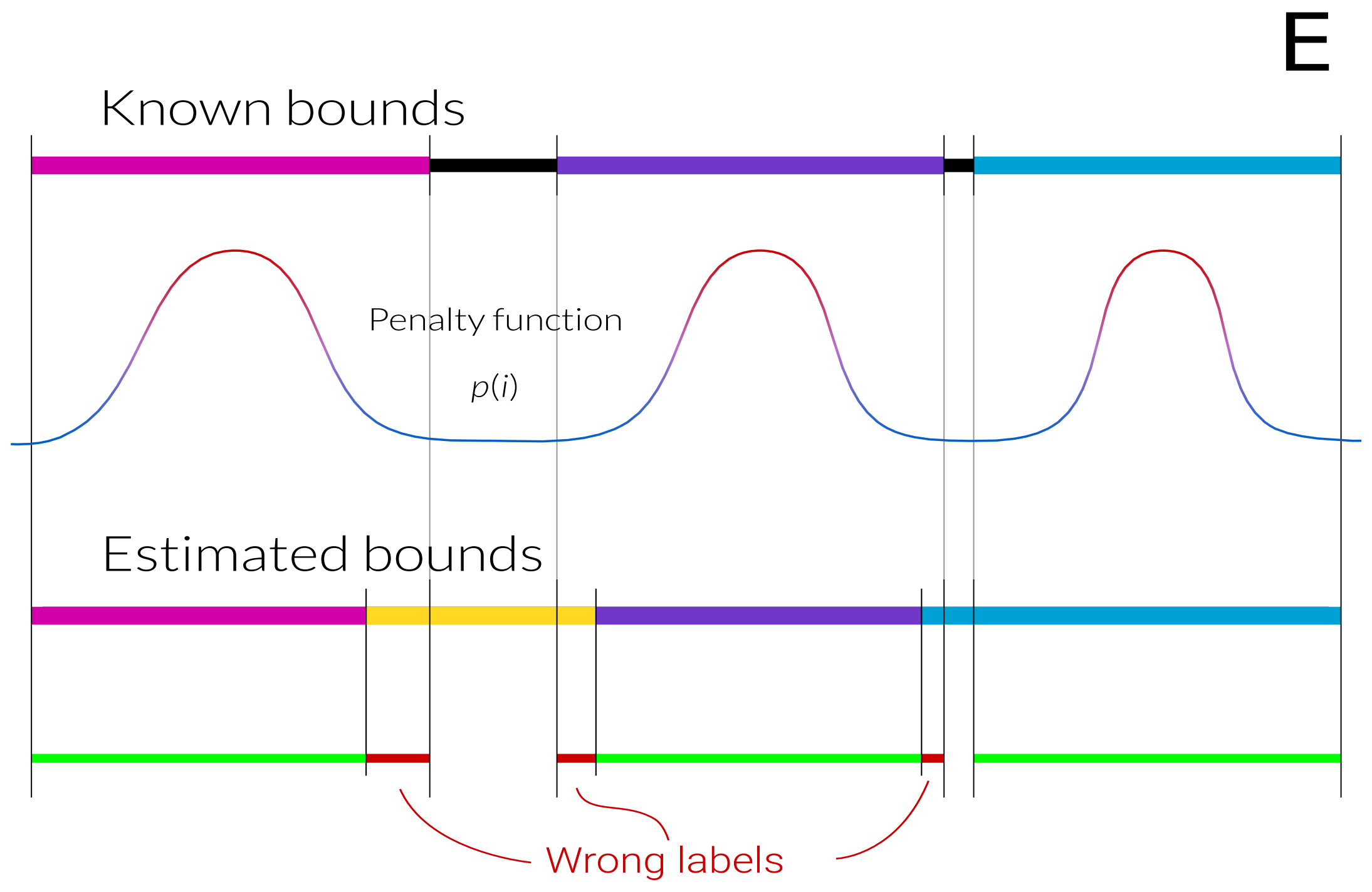
The dataset of clustered protein chains was collected from CATH DB (cathdb.info), yielding about 130 000 entries with domain boundaries. During dataset analysis we found some patterns in domain distribution and features.

Firstly, the hydrophobic residues are distributed more densely than whole amino acid set in the chain (*fig. A*).

Than, number of amino acids in domains are distributed very similar to whole chain length to number of domains in the chain ratio (*fig. B*)

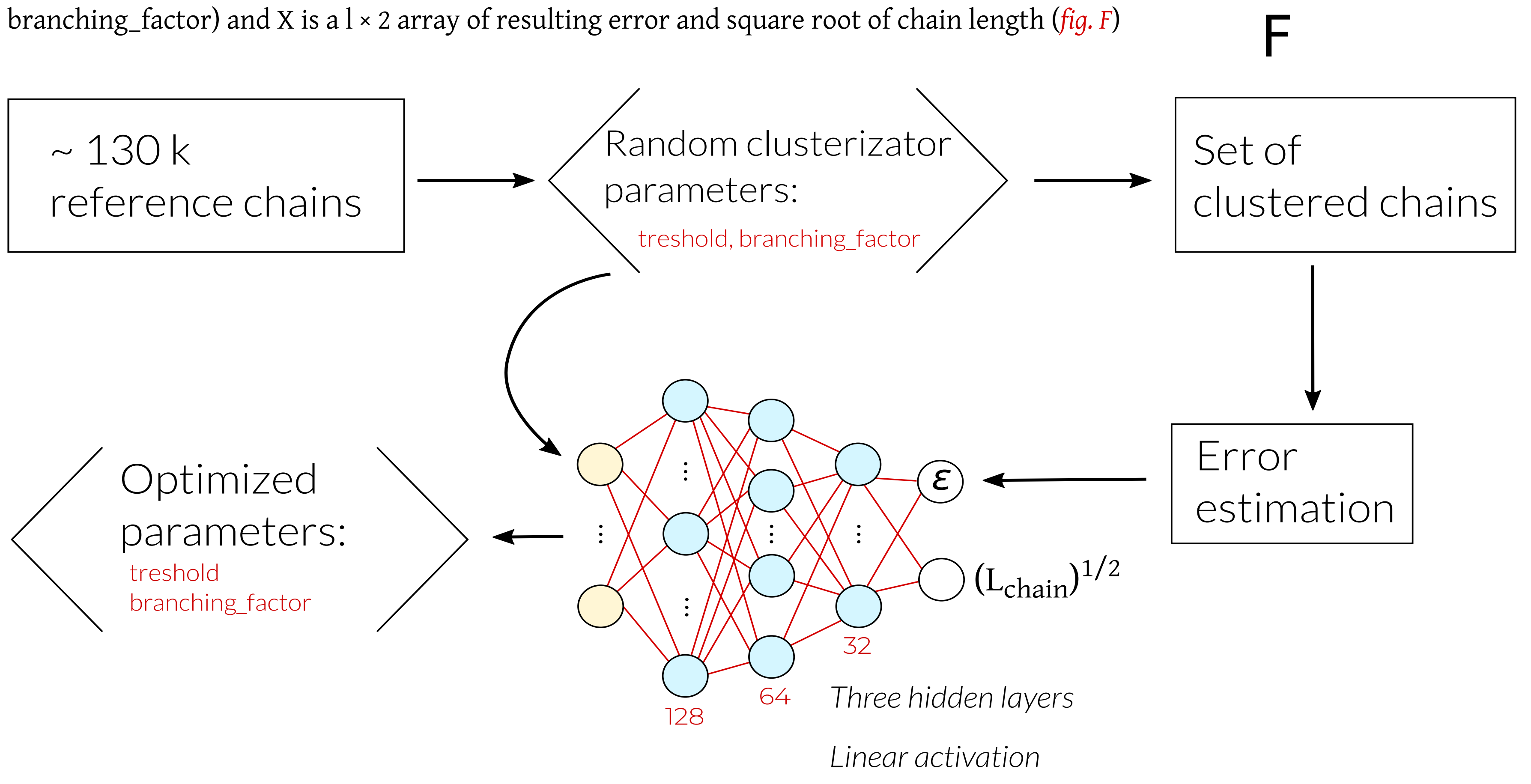


Methods and Algorithms



Scoring function: error of domain prediction is estimated by following technique (*fig. E*).
1. Dividing resulting label vector into reference region.
2. Calculating vector of incorrect labels $D = (0, 0, ..., 0, 1, 1, ..., 0, 0)$ with shape $N_{\text{residues}} \times 1$ and where i^{th} position is 1 if corresponding label differs from the most frequent in reference region.
3. Caltulating vector of penalty weights P normalized to 1. In this work we used a binominal distribution for each chunk (*see fig. E*).
4. Clustering error ε is a dot product of vectors described above: $\varepsilon = D \cdot P$

Clustering algorithms: in this work two clustering algoritms were used:
1. BIRCH [2], and 2. Ising model [1]. Both algorithms parameters could be described as `(dist_matrix, treshold, branching factor)` -> `label_vector`. Ising algorithm have shown better results and is used in final clustering tool.
Clustering hypermarameter optimization: for optimization of clustering parameters we used deep neural network with three hidden layers of neurons with linear activation, trained on a precalculated set XY, where Y is a l × 2 array of uniformely distributed parameters (treshold, branching_factor) and X is a l × 2 array of resulting error and square root of chain length (*fig. F*)


F

~ 130 k reference chains

Random clusterizator parameters: treshold, branching_factor

Set of clustered chains

Error estimation

Optimized parameters: treshold branching_factor

Literature

1. Taylor W.R. (1999) Protein structural domain identification. Protein Engineering. 12(3): 203–216
2. 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. pp. 103–114.

Availability

The protein clustering visualization tool is available as a web service at

protein-clustering.ru

Conclusion

Naive protein clusterization based only on atom coordinates does not provide good domain prediction. But adding some more parameters (*e. g.* chain length) gives much better results. Our protein clusterization will be subsequently improved to highlight the most predictive polypeptide chain features, hepling us to split protein into domains.

Aknowledgements

We express our gratitude to Yuri Vyatkin for formulation of the problem and curation during our investigation.