PREDICTION OF THE PROTEIN CONFORMATION VIA PATTERN RECOGNITION AND CONSTRAINT SATISFACTION METHODS

B.Matkarimov*[1], RTakhanov[2]

[1]Nazarbayev University Research and Innovation System, Astana, Kazakhstan; *bmatkarimov@nu.edu.kz;
[2]Institute of Science and Technology, Austria

INTRODUCTION.

The prediction of the native structure of proteins given their amino acid sequence is one of the central problems in modern computational biology/biophysics/biochemistry and computer science. Today there are more than 80000 3D structures of various biomacromolecules in the open access, e.g. in Protein Data Bank (PDB), and these databases have exponential growth rate. The project goal is to design and implement computing experiments related to biomacromolecular folding. This includes the development of high performance bioinformatics software.

MATERIALS AND METHODS.

Most of the methods for studying this problem are knowledge-based, i.e. statistics rather than physical modeling. Knowledge-based models are divided on two main parts, namely comparative modeling and threading. Unlike comparative modeling, threading is used when there are no already resolved homologs of target protein. The approach that we develop also belongs to this second class of models; therefore, we tackle the important problem of protein dihedral angles prediction. The latter is interpreted as a sequence labeling problem; therefore, we develop a novel statistical approach called pattern-based conditional random field.

RESULTS AND DISCUSSION.

We introduced a new statistical model based on patterns and developed key learning and inference algorithms for it. All algorithms are implemented in C++, and used to train the model on data from PDB. For a protein dihedral angles prediction problem, we achieved state-of-the-art values of prediction accuracy.

CONCLUSIONS.

Application of conditional random fields in bioinformatics is a relatively new topic. This research shows that the potential of this approach is far from being exhausted.

ACKNOWLEDGMENTS.

This work is funded by the grant of the Ministry of Education and Science of the Republic of Kazakhstan.

REFERENCES.

1. Takhanov R., Kolmogorov V. (2013). Inference algorithms for pattern-based CRFs on sequence data, Proceedings of International Conference on Machine Learning (ICML), Atlanta, GA, USA. JMLR: W&CP, Vol. 28.
2. Ten V., Isembergenov N., Matkarimov B. (2013). Approach to control of hybrid renewable power system on the basis of AE-method using genetic algorithm, Proceedings of the International Conference on Machine Learning and Applications, December 4 – 7, 2013, Miami, FL, USA.