PROTEIN^ S CONFORMATIONAL STRUCTURE PREDICTION VIA PATTERN RECOGNITION AND CONSTRAINT SATISFACTION METHODS

B. Matkarimov[1*], R. Takhanov[2]

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

Introduction. The prediction of protein's native structure given its 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 at the open access, e.g. in Protein Data Bank (PDB), and these databases have exponential growth rate. Project goal is to design and implement computing experiments related to biomacromolecular folding problem. This includes the development of high performance bioinformatics software.

Materials and methods. Most of the methods for this problem are knowledge-based, i.e. based on statistics rather than physical modeling. Knowledge-based models themselves are divided on two main parts, namely comparative modeling and threading. Unlike comparative modeling, threading is used in case when there are no already resolved homologs of target protein. The approach that we develop in current research also belongs to this second class of models. For this purpose we tackle the important problem of protein dihedral angles prediction. The last is interpreted as sequence labeling problem, and for it we develop a novel statistical approach called pattern-based conditional random field. We also generalize pattern-based approach to include non-local interactions between patterns and discuss kernelization of our models.

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 protein dihedral angles prediction problem we achieved state-of-the-art values of prediction accuracy.

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

Acknowledgments. This work funded by the grant of the Ministry of Education and Sciences of the Republic of Kazakhstan.

References.

1. Rustem Takhanov, Vladimir Kolmogorov, Inference algorithms for pattern-based CRFs on sequence data // Proceedings of International Conference on Machine Learning (ICML), 2013.
2. Mazouzi A., Vigouroux A., Aikeshev B., Brooks P.J., Saparbaev M.K., Morera S., Ishchenko A.A. Insight into mechanisms of 3"-5" exonuclease activity and removal of bulky 8,5"-cyclopurine adducts by apurinic/apyrimidinic endonucleases // Proceedings of the National Academy of Sciences of the United States of America .- 2013 .- Vol. 110, №33 .- P. 3071-3080.
3. Rustem Takhanov, Vladimir Kolmogorov, Combining pattern-based CRFs and weighted context-free grammars // electronic preprint http://arxiv.org/pdf/1404.5475
4. I.Talhaoui, S. Couve, L. Gros, A. Ishchenko, B. Matkarimov, M. Saparbaev. Aberrant repair initiated by mismatch-specific thymine-DNA glycosylases provides a mechanism for the mutational bias observed in CpG islands, Nucleic Acids Research, 2014, 42(10): 6300-6313.