Application of artificial intelligence methods, in particular deep learning, in the analysis of molecular data

prof. dr inż. Jacek Koronacki
Institute of Computer Science, PAS, Warsaw
j.koronacki@ipipan.waw.pl

dr inż. Michał Draminski
Institute of Computer Science PAS, Warsaw
m.draminski@ipipan.waw.pl

1. Project Description

The well-known challenge posed by molecular data, in particular epigenetic and genomic regulatory areas, is their huge dimensions. In the Computational Biology Lab of the IPI PAS, the algorithms [1] have been implemented and are successfully used to detect the hidden structure of such data [2, 3] and to solve classification and prediction tasks. The project will focus on the discovery of epigenetic regulatory networks in the context of a specific phenotype using the developed artificial intelligence algorithms and their implementations. The networks will allow the discovery of causally effective relationships between inherited genetic modifications (not related to the DNA sequence) and molecules interacting with them. The specificity of these interactions affects changes in gene expression and if they belong to one signal pathway, they can translate into the formation of a disease-causing state.
The number of all interactions is so large that without appropriate specific AI methods, it is impossible to analyze them in a reasonable time. To solve this problem, apart from the already developed algorithms, we want to use deep (e.g. convolutional) neural networks [4, 5, 6]. This last approach slowly becomes dominant in the mentioned area, but there is still a lot to explore and invent. The project's emphasis will be both on its computational aspects and on obtaining new biologically important results with a clear biological interpretation.

2. Requirements

Due to the multidisciplinary character of research conducted in the Computational Biology Lab (http://zbo.ipipan.waw.pl) in general, there are welcome persons having some experience in one (or more) of such domains: **statistics, machine learning, data processing and analysis as well as biology**, **chemistry**.

Qualifications requested and eligibility:

- Master of Science in one of the following: Computer Science, Mathematics, Physics, Bioinformatics, or one of the like scientific disciplines
- The candidate should have good communication skills and excellent study merits, and good skills in oral and written English.

The candidate should also have:

- good skills and some experience in data analysis with use of statistical/machine learning tools,
- knowledge of script languages used in statistical analysis e.g. R/RStudio/Python,
- knowledge of one or more programming languages (C/C++/C#/Java etc.),
- basic experience in administration of Linux OS.

The following specific qualifications are welcome but not required:
Knowledge of:

- molecular biology, genetics in particular,
- learning database structures and retrieving information from databases.

References

1. Draminski M., Koronacki J. (2018). rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. Journal of Statistical Software vol. 85(12), doi:10.18637/jss.v085.i12.
2. Dabrowski M.J., Draminski M., Diamanti K., Stepniak K., Mozolewska M.A., Teisseyre P., Koronacki J., Komorowski J., Kaminska B. & Wojtas B. (2018). Unveiling new interdependencies between significant DNA methylation sites, gene expression profiles and glioma patients survival. Scientific Reports vol. 8, Article number: 4390, doi:10.1038/s41598-018-22829-1.
3. Dabrowski M.J., Dziedzic A., Guzik R., Draminski M., Wojtas B., Stępniak K., Gielniewski B., Koronacki J., Kamińska B. "Genome-wide mapping of DNA methylation variants affecting gene expression levels in gliomas with respect to their grade and IDH gene mutation status." Abstracts of papers presented at the meeting on "The Biology of Genomes" May 7-May 11, 2019. Cold Spring Harbor, USA (2019):69
4. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular systems biology. 2016 Jul 1;12(7):878.
5. Ainscough BJ, Barnell EK, Ronning P, Campbell KM, Wagner AH, Fehniger TA, Dunn GP, Uppaluri R, Govindan R, Rohan TE, Griffith M. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nature genetics. 2018 Dec;50(12):1735.
6. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MT, Cheng C, Fan X, Gerstein M, Cheng AS. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nature genetics. 2017 Oct;49(10):1428.