Zastosowanie metod sztucznej inteligencji, w szczególności głębokiego uczenia się, w analizie danych molekularnych

prof. dr inż. Jacek Koronacki
Instytut Podstaw Informatyki
j.koronacki@ipipan.waw.pl

dr inż. Michał Dramiński
Instytut Podstaw Informatyki
m.draminski@ipipan.waw.pl

1. Opis projektu

Podstawowym problemem, jaki napotykamy analizując dane molekularne, w szczególności epigenetyczne i genomiczne obszary regulatorowe, jest ich wielki wymiar. W Zespole Biologii Obliczeniowej IPI PAN zaimplementowano własne algorytmy [1] i z powodzeniem stosuje się je do wykrywania ukrytej struktury takich danych [2, 3] oraz rozwiązywania zadań klasyfikacji i predykcji. Projekt będzie koncentrował się na odkrywaniu epigenetycznych sieci regulatorowych w kontekście specyficznego fenotypu przy użyciu rozwijanych algorytmów sztucznej inteligencji i ich implementacji. Sieci pozwolą na odkrycie przyczynowo skutkowych zależności pomiędzy dziedziczonymi modyfikacjami genetycznymi niezwiązanymi z sekwencją DNA a molekułami wchodzącymi z nimi w interakcje. Specyfika tych interakcji wpływa na zmiany ekspresji genów i jeżeli należą one do jednej ścieżki sygnałowej to mogą przekładać się na powstanie stanu chorobotwórczego. Liczba wszystkich interakcji jest tak duża, że bez odpowiednich dziedzinowo specyficznych metod AI, niemożliwa jest ich analiza w skończonym czasie. Do rozwiązania tego problemu obok opracowanych już algorytmów chcemy użyć głębokich (np. konwolucyjnych) sieci neuronowych [4, 5, 6]. To ostatnie podejście staje się powoli dominującym w wymienionym obszarze, ale jeszcze bardzo wiele pozostaje do zbadania i wynalezienia nowych rozwiązań. W projekcie nacisk zostanie położony zarówno na jego aspekty obliczeniowe, jak i na uzyskiwanie nowych biologicznie ważnych wyników o jasnej interpretacji biologicznej.

2. Wymagania (oczekiwania)

Z uwagi na wielowątkowy i multidyscyplinarny charakter kompetencji osób pracujących w Zespole Biologii Obliczeniowej (http://zbo.ipipan.waw.pl), jesteśmy otwarci na współpracę z osobami zainteresowanymi pracą naukową i posiadającymi pewne doświadczenie oraz umiejętności w jednej (lub więcej niż jednej) z następujących dziedzin: **statystyka, uczenie maszynowe, przetwarzanie i analiza danych jak również biologia, chemia, fizyka.**

Kwalifikacje kandydatów:

- tytuł magistra w zakresie informatyki, matematyki, fizyki, bioinformatyki lub dziedzin pokrewnych,
- dobra znajomość i doświadczenie w analizie danych z wykorzystaniem narzędzi statystycznych i uczenia maszynowego,
- znajomość popularnych języków skryptowych używanych w analizie danych np. R/RStudio/Python,
- przynajmniej podstawowa znajomość programowania w jednym języku obiektowym (C/C++/C#/Java etc.),
- podstawowa znajomość obsługi i administrowania systemem Linux,
- dobra znajomość j. angielskiego,
- wysoka motywacja, umiejętność analitycznego myślenia i pracy w zespole.

Oczekiwania dodatkowe, podnoszące ocenę kandydata:

- wiedza z zakresu biologii molekularnej a zwłaszcza genetyki,
- chęć doskonalenia swojej wiedzy w dziedzinach, których podstawowa znajomość jest niezbędna do wielopłaszczyznowego analizowania danych w zespole.

Literatura

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.