Predlogi tem za zaključna dela 2024-2025

Tomaž Curk, september 2024

Splošne teme

Raziskovalno se ukvarjam z:

- bioinformatiko,
- odkrivanje znanj iz podatkov,
- zlivanjem podatkov,
- avtomatizacijo zlivanja podatkov v relacijskih bazah,
- prikazovanjem in interpretacijo rezultatov zlivanja podatkov,
- modeliranjem interakcij protein-RNA,
- uporabo globokih nevronskih mrež za analizo genomskih podatkov,
- vizualizacijo podatkov,
- odkrivanjem in prikazovanjem trendov v znanstveni literaturi,
- priporočilnimi sistemi in njihovimi aplikacijami v bioinformatiki in tudi v turizmu (modeliranje in napovedovanje cestnega prometa, letalskega prometa, obiskov in ocenjevanje ponudnikov turističnih storitev, itd.).

Če vas našteta področja zanimajo, potem se zglasite pri meni, po dogovoru na tomaz.curk@fri.uni-lj.si. Skupaj bomo oblikovali temo glede na vaše interese, predznanje in stopnjo študija (diploma, magisterij). Predlagate lahko tudi lastno temo oz. raziskovalni ali aplikativni problem.

Primere preteklih zaključnih del pod mojim mentorstvom lahko prelistate tukaj, moje objavljene znanstvene članke pa tukaj.

Teme v sodelovanju z NIB

Na področju bioinformatike tesno sodelujemo z Nacionalnim inštitutom za biologijo (NIB). Navedenih je nekaj konkretnih tem in potencialnih somentorjev iz NIB.

- **Toward fully explainable deep learning of plant phenotypes based on interaction networks**

Contact:
Jan Zrimec, jan.zrimec@nib.si
Carissa Bleker, carissa.bleker@nib.si

Key resources:
[1] https://skm.nib.si/
[2] https://doi.org/10.1093/gigascience/giz022, https://www.cell.com/fulltext/S0092-8674(16)30667-5
[3] https://arxiv.org/abs/1704.03165
[4] https://doi.org/10.1016/j.bpr.2023.100118

Development of explainable models of plant phenotypes is a key prerequisite to understand the mechanisms behind the response of plants to environmental and disease stress, and thus help us develop more resilient crops. Here we seek a highly motivated student with programming experience to develop and test deep neural networks (DNNs) that integrate prior knowledge from molecular interaction networks [1]. The plant model Arabidopsis thaliana will be used, with the option to include population-scale genetic variation and gene

expression data [2]. The work will include: (i) processing and analysis of training data including genomic and transcriptomic data, (ii) setting up and testing different DNN architectures and training parameters, including graph neural networks, (ii) testing different strategies for integrating interaction networks within DNN models, such as learning latent representations for the structural identity of nodes [3], or using a structure reflecting gene regulation [4] or the flow of genetic information (gene, transcript, protein, protein complex, phenotype layers), and benchmarking them against standard non-prior-based DNNs, and finally, (iv) interpreting the trained models to determine whether this is a viable strategy for enriching biological networks (e.g. prediction of interaction strength or identifying new edges) as well as to improve our understanding of how the network is rewired in different tissues or perturbations.

- **Interpretation of genomic variation effects on tissue and condition-specific gene expression using deep learning**

(taken by AZ)

Contact:
Jan Zrimec, jan.zrimec@nib.si

Key literature:
[1] https://www.nature.com/articles/s41467-020-19921-4
[2] https://www.cell.com/fulltext/S0092-8674(16)30667-5
[3] https://www.pnas.org/doi/10.1073/pnas.2311219120
[4] https://doi.org/10.1073/pnas.2216698120

Understanding the genetic evolutionary code governing gene expression and environmental phenotypes is an important challenge in biotechnology, especially if we want to understand mechanisms behind environmental acclimation and stress responses that can help us to develop more resilient crops. A highly motivated student with programming experience is sought to aid in developing and interpreting deep neural network (DNN) predictive models of molecular phenotypes (e.g. tissue, perturbation or environmental condition) in the plant model organism Arabidopsis thaliana. These DNNs will be trained on gene-specific DNA sequences as input [1]. The work will include: (i) processing and analysis of omics training data including genomics and transcriptomics [1,2], (ii) model training and hyperparameter optimization pipelines, (iii) development or implementation of tools to interpret the learned representations of DNA motifs or nucleotide variations from DNNs [3,4], (iv) analysis of tissue and/or condition-specific models and interpretation as well as comparative analysis of the tissue/condition-specific DNA regulatory grammar. This will hopefully enable us to discover and contrast the major tissue-specific determinants of gene expression.

- **Using large scale systems-biology simulations to generate benchmarking data for integrative multi-omics analysis**

Contact:
Anže Županič, anze.zupanic@nib.si

Key literature:
[paper1]: https://skm.nib.si/ (PSS knowledge network)
[paper2]: https://doi.org/10.1371/journal.pcbi.1005752
[paper3]: https://doi.org/10.1093/bioinformatics/btr373
[paper4]: https://www.nature.com/articles/s42256-020-0218-x

With the advances in high-throughput biological data generation technologies the amount of unbiased large scale datasets (genomics, transcriptomics, proteomics, metabolomics, etc) has tremendously increased, but our ability to mechanistically interpret the data has not followed. While several multi-omic integration methodologies and tools have been developed, their usefulness and credibility is hard to determine without benchmarking data where the true molecular mechanisms underlying the measured entities are known.

A highly motivated student with programming experience is sought to aid in developing synthetic multiomic benchmarking datasets based on networks of known molecular interactions, and benchmarking existing multiomic integration tools. The work will include: (i) constraining the kinetic parameters of the Plant Stress Signalling Model (highly curated knowledge network available at NIB) according to current biochemical knowledge and literature data, (ii) simulating time series data that will serve as benchmarking with additional noise components, (iii) choosing appropriate measures for data integration success (how much the inferred interactions concur with the underlying knowledge network) and (iv) implement and benchmark existing methods for multiomics data analysis and network inference.

* Generated datasets to be added here: https://openebench.bsc.es/vre/home/ for future benchmarking