Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2024-2025

Personal Information

Supervisor	Anthony Mathelier
Email	anthony.mathelier@ncmm.uio.no
Institution	Centre for Molecular Medicine Norway (NCMM), University of Oslo
Website	https://mathelierlab.com
Group	Computational Biology & Gene Regulation

Project

Computational genomics

Project Title:

CETsim: simulating transcriptional regulation

Keywords:

gene regulation, simulation, deep learning, sequence-to-activity models

Summary:

Transcriptional regulation is driven by transcription factor (TF) binding at cis-regulatory regions (CRR). TFs recognize and bind specific sequence patterns within the DNA. Multiple TFs can bind a single CRR, and their cooperativity gives rise to the complex regulation we observe in multicellular organisms. Most TFs show cell-state-specific expressions, and their availability affects their binding. Dysregulation of this complex system is the cause of many diseases, including cancer. One popular approach to understanding and predicting the effect of mutations in the non-coding regions is via machine learning models. These models (e.g., BPNet [1], AI-TAC [2]) can learn to predict molecular features (e.g., RNA-seq, ATAC-seq, or ChIP-seq signal) associated with certain genomic regions in specific cellular conditions from the static DNA sequence. It has been shown through model interpretation that the presence or absence of TF binding sites (TFBS) drives the prediction of these models. However, the validity of the existing and to-be-developed models and the validity of the interpretations are hard to assess due to the lack of ground truth knowledge or shared benchmark datasets [3]. The simulation framework we develop aims to support simulation-based development by providing a tool for assessing any sequence-to-activity model's behavior in a controlled environment. Collection, Entity, and Topic simulation (CETsim) takes a hierarchical approach where TFs are assigned to topics that build up entities (e.g., cells) and collections (e.g., samples). The binding sites of the TFs are inserted into a sequence background, and based on the relationships and assignments of the TFs, the activity of each sequence can be calculated. The framework's clear structure and modular design allow for simulating various hypotheses by the experimenting bioinformatician. The Master student's role will be to develop and assess a use case for this simulation framework. We suggest simulating single-cell ATAC datasets and assessing the prediction of deep learning models such as scBasset [4] and AI-TAC [2]. Other simulations could be investigated based on the student's preference. The Master student will collaborate closely with the PhD student developing CETsim for daily scientific and computational questions, discussions and overview. Importantly, our computational group promotes software development best practices and provides support to achieve them.

References:

1. Avsec Ž, Weilert M, Shrikumar A et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet 53, 354–366 (2021). https://doi.org/10.1038/s41588-021-00782-6 2. Maslova A, Ramirez RN, Ma K, Schmutz H, Wang C, Fox C, Ng B, Benoist C, Mostafavi S; Immunological Genome Project. Deep learning of immune cell differentiation. Proc Natl Acad Sci U S A. 2020 Oct 13;117(41):25655-25666. doi: 10.1073/pnas.2011795117. Epub 2020 Sep 25. PMID: 32978299; PMCID: PMC7568267. 3. Chen V, Yang M, Cui W, Kim JS, Talwalkar A, Ma J. Best Practices for Interpretable Machine Learning in Computational Biology. bioRxiv 2022.10.28.513978; doi: https://doi.org/10.1101/2022.10.28.513978 4. Yuan H, Kelley DR. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. Nat Methods 19, 1088–1096 (2022). https://doi.org/10.1038/s41592-022-01562-8

Expected skills:

experience in programming in Python, basic knowledge of gene regulation, basic knowledge of one (or more) experimental technique(s) to measure molecular phenotypes of gene regulation (e.g., ATAC-seq, RNA-seq, ChIP-seq)

Possibility of funding:

To be discussed

Possible continuity with PhD:

No

Comments:

Possibility to cover housing expenses.