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:

Deep learning approach to decipher the cis-regulatory code in breast cancer patients

Keywords:

cancer genomics, gene regulation, deep learning

Summary:

Breast cancer represents a spectrum of diseases controlled by distinct molecular mechanisms. Breast cancer is classified into six intrinsic subtypes characterized by different gene expression signatures reflective of distinct biological aspects of tumor presentation, function, and identity1. While such molecular classifications focus on protein-coding genes to define gene expression signatures, they neglect the critical role of noncoding regulatory elements in cancer development. Indeed, developmental/typical enhancers that control growth-related genes can be aberrantly activated in cancer cells2,3. We aim to shed light on the critical cis-regulatory code activated in cancer cells to rewire the regulatory networks. Recently, researchers used RNA-seq and ATAC-seq data to identify the cis-regulatory regions (promoters and enhancers) activated in cancer patients. Data on the activity of promoters and enhancers across samples provide the means to decipher the underlying molecular mechanisms that trigger the specific activation of these regions. In this project, we will use a deep learning (DL) approach based on convolutional neural networks to predict promoter and enhancer activities across samples from their DNA sequences. Such deep learning frameworks have successfully deciphered the cis-regulatory motif syntax from ATAC-seq or CAGE data4–6. Specifically, the selected candidate will use and adapt the scBasset framework7, initially developed for scATAC-seq data, to automatically identify both the TF binding profiles/motifs and their combinatorics acting upon the enhancers and promoters. Advantageously, using the latent variables in the sequence embedding also learns sample embedding, which can be used to cluster the samples. Finally, the enhancers and promoters important for the clusterization can be extracted from the CNNs, which will determine the coordinated activities of regulatory regions in each group of patients. We plan to apply the DL approach to promoters and enhancers separately, and the corresponding latent representation of these two spaces will then be combined following similar approaches to Polarbear8 and multiVI9. As opposed to approaches solely relying on gene expression, this approach will assess the activity of transcriptional regulatory regions, thus providing an innovative concept. Moreover, this project's machine-learning method will provide molecular cis-regulatory signatures (relationship between promoter and enhancer activities) and sample relationships. Finally, this approach offers nucleotide-resolution feature discovery to shed light on the TF binding code at enhancers and promoters specific to the breast cancer subtypes cis-regulatory signatures. We anticipate that the envisioned approach can potentially reveal new breast cancer subgroups that cannot be discovered by relying solely on protein-coding gene expression. Identifying molecular and sample relationships provides an innovative way to identify new biomarkers by focusing not only on active genes (and promoters) but also on enhancers.

References:

1. Russnes, H. G., Lingjærde, O. C., Børresen-Dale, A.-L. & Caldas, C. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. The American Journal of Pathology 187, 2152–2162 (2017). 2. Murakawa, Y. et al. Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. Trends in Genetics 32, 76–88 (2016). 3. Sur, I. & Taipale, J. The role of enhancers in cancer. Nature Reviews Cancer 16, 483–493 (2016). 4. González-Blas, C. B. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nature Methods 16, 397 (2019). 5. Maslova, A. et al. Deep

learning of immune cell differentiation. PNAS 117, 25655–25666 (2020). 6. Einarsson, H. et al. Promoter sequence and architecture determine expression variability and confer robustness to genetic variants. http://biorxiv.org/lookup/doi/10.1101/2021.10.29.466407 (2021) doi:10.1101/2021.10.29.466407. 7. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. Nat Methods 1–9 (2022) doi:10.1038/s41592-022-01562-8. 8. Zhang, R., Meng-Papaxanthos, L., Vert, J.-P. & Noble, W. S. Semi-supervised single-cell cross-modality translation using Polarbear. http://biorxiv.org/lookup/doi/10.1101/2021.11.18.467517 (2021) doi:10.1101/2021.11.18.467517. 9. Ashuach, T., Gabitto, M. I., Jordan, M. I. & Yosef, N. MultiVI: deep generative model for the integration of multi-modal data. 2021.08.20.457057 Preprint at https://doi.org/10.1101/2021.08.20.457057 (2021).

Expected skills:

experience in programming in Python, basic knowledge of gene regulation, interest in cancer genomics

Possibility of funding:

To be discussed

Possible continuity with PhD:

No

Comments:

Support for housing expenses could be provided. The work will be conducted at the Centre for Molecular Medicine Norway (NCMM), Oslo, Norway, in the Computational Biology & Gene Regulation group led by Anthony Mathelier. The group combines complementary experimental and computational biology expertise and will provide additional mentorship and support. The group meets weekly for research or journal club presentations, and provides software quality and code review meetings.