Multimodal Transcriptomic/Histology AI for PDAC Precursor Detection

Type: Master Thesis
Supervisor: Prof. Peter Schüffler (https://schuefflerlab.org)
Starting date: asap
Contact: peter.schueffler@tum.de

Summary

The Schuefflerlab for Computational Pathology at the TUM Institute for Pathology is offering a CIT Master's thesis in the field of medical machine learning and pathology AI (artificial intelligence). The study explores the potential unimodal and multimodal deep learning models to predict precursors of pancreatic ductal adenocarcinoma (PDAC) from histology and/or spatial transcriptomics data.

Background

Pancreatic ductal adenocarcinoma (PDAC) has several precursors, such as *acinar to ductal metaplasia* (ADM) and *mucinous tubular complexes* (MTC), but their exact progression pattern is still unknown. Spatial transcriptomics is a technique that quantifies the local transcriptome of a tissue sample in high resolution, allowing for detailed investigation of PDAC tissue and its precursors on a molecular level. However, spatial transcriptomics is slow and expensive, limiting its application to small tissue samples. In this work, we explore the potential of deep learning (DL) in histology to detect precursors of PDAC and related expression profiles. We consider DL as a replacement technology (*can DL in histology predict the precursors similar as spatial transcriptomics can?*) and as an enhancement technology (*can a multimodal histology/RNA model improve the precursor prediction?*). A publication is aimed at the end of the thesis, depending on the results.

Dataset

We will use transcriptomics data from five whole slide images (WSI) of murine PDAC precursor tissue of the pancreas, totaling over 8000 annotated tissue spots. Figure 1 illustrates an excerpt of the data set. Each tissue spot has **two types of annotations:** The first type is a hand-drawn annotation of the precursor class, *ADM* and *MTC*, as well as other tissue types such as stroma. The second annotation type is the *Visium Spatial Gene Expression* (10X Genomics) spatial transcriptomics data, a feature vector with RNA counts of over 18,000 genes[1]. Seven selected genes have been found previously to correlate with the precursor classes (Keratin 19 (Krt19) for ductal cells, Amylase (Amy1) for acinar cells, further Claudin 10 (Cldn10), Claudin 18 (Cldn18), Tetraspanin 8 (Tspan8), Cathepsin L (Ctsl), Mesothelin (Msln)). Further code to load the medium data is available on GitHub[2].

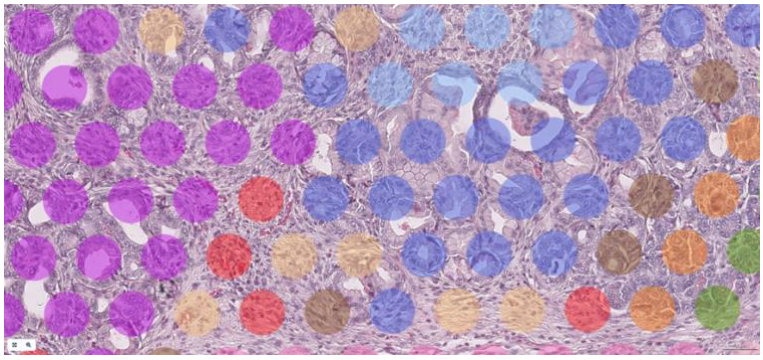

Figure 1: Example of Visium (10X Genomics) expression data. RNA profiles are measured in circles with 55 µm diameter. Colors correspond to tissue type classes or expression levels.

[1] See https://www.10xgenomics.com/products/spatial-gene-expression
[2] See https://github.com/thommetz/Visium_Preneopl

Tasks

T1: To test if unimodal deep learning models trained on histology image patches can predict the precursor classes, and the RNA expression of selected RNA molecules.

T2: To test if multimodal deep learning models trained on histology image patches and RNA expression profiles can improve the precursor class prediction.

For all tasks, image patches will be used directly to train convolutional neural networks (CNNs), and will be embedded in a pre-trained pathology space (e.g., using UNI[1]).

Methods

Data preparation: the student will extract tissue patches from the Visium transcriptomics data and associate them with their transcriptomic expression profile and the annotated tissue class. This will serve as the training and test set for the follow up experiments. The data set will be described and explored (e.g. class distributions, data outliers, etc.). Data augmentation strategies and data splits for 5-fold CV will be defined.

Modelling: The student will design and implement experimental setups for all tasks:

- DL Models (e.g. ResNet50) to predict the precursor class from patches or their embeddings
- DL Models (e.g. ResNet50) to predict the expression levels of selected genes from patches or their embeddings
- Unimodal DL Models (e.g. ResNet50) to predict the precursor class from the expression levels of selected genes
- Multimodal DL Models (e.g. ResNet50) to predict the precursor class from patches or their embeddings combined with the expression levels of selected genes

All training will be performed on our internal GPU-based computer cluster. Python (ScanPy, SquidPy, PyTorch) will be used to process data and train models. Data visualization and result plotting is a major part of this work.

Requirements

This is a machine learning study with components of bioinformatics, data science and computer vision. Candidates should be comfortable in learning new things and concepts if needed.

- Strong understanding of concepts of machine learning and deep learning
- Basic understanding of bioinformatics and computer vision
- Experience in coding in Python
- Fun working in an interdisciplinary field (cancer research, bioinformatics, machine learning)
- Fluent in German or English

References

1. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. *Nat Med*. 2024;30(3):850-862. doi:10.1038/s41591-024-02857-3