Predicting Metabolic Pathway Activity Using Gene Expression Values in Colorectal Cancer Tissues

Kaavya Borra, Anmol Karan, Rohan Matta
Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center

Dartmouth Health

ABSTRACT

- Evaluates the ability of neural networks to predict different molecular pathways using spatial transcriptomics images as the ground truth and to identify the optimal NN model
- Preprocessing: Decoupler, single-strand Gene Set Enrichment Analysis (ssGSEA), and Scanpy, PyTorch to train and test
- Kyoto Encyclopedia of Genes and Genomes (KEGG) database, Hallmark p53, Apoptosis, WNT Beta Catenin Signaling, MTORC1 Signaling, KEGG Vascular Smooth Muscle
- Decoupler performed the best, with an average Mean Absolute Error (MAE) of 0.106
- Because ResNet18 offers promising results, we plan to utilize other model architectures, such as a VIT or GNN, and training methods to optimize performance further.

INTRODUCTION

- Caused by DNA mutations, **colorectal cancer (CRC)** is recognized by tumors in the colon
- **Gene expression** controls vital biological functions that play a role in CRC development. Gene expression changes drive the major molecular pathways involved in CRC.
- **Spatial Transcriptomics** allows researchers to map gene expression data across a tissue sample. It enables visualization and quantification of gene expression data in a WSI
- **Our goal** in this project is to explore how well neural networks can predict various molecular pathways from hematoxylin and eosin (H&E) imaging using spatial transcriptomics images as ground truth and to also find the optimal strategy for enhancing predictions with those neural networks.

METHODS

- **Data Collection & Pre-Processing**
  - Obtained 40 WSI slides of H&E stained images, and their corresponding Visium Spatial Transcriptomics data (gene expression counts) from the Dartmouth-Hitchcock Medical Center
- **Gene Set Selection**
  - KEGG database and MSigDB offer comprehensive gene sets representing CRC-related pathways. The gene sets utilized in this study include:
    - Vascular Smooth Muscle Contraction (sanity check)
    - Hallmark p53
    - Hallmark apoptosis
    - mTor Complex 1
    - WNT/β-catenin
  - **Module Scores -** Three methods were used to calculate ground truth expression values for each gene set:
    - **Decoupler** - Calculates a weighted sum of gene expression
    - **ssGSEA** - Calculates a running sum based on expression levels, adding the weight of genes in the gene set and subtracting a fixed weight for those not in the gene set
    - **Scanpy** - Subtracts the average expression of genes in a randomly generated control set from the average expression of genes in the target set
  - **Patches**:
    - For simplicity, one representative WSI was used for training, and one more for testing. Visium patches were generated from the WSIs by taking 512x512 pixel sections centered around each Visium spot.
- **Model Development**
  - **ResNet18** was the selected model architecture, as it is a commonly used CNN for transfer learning.
  - The **PyTorch** package was used for training. All models were trained using the Adam optimizer, with a batch size of 32, a learning rate of 0.001, and 10 training epochs.
  - Horizontal flips, vertical flips, and rotations were randomly applied to each patch to improve model robustness.

RESULTS

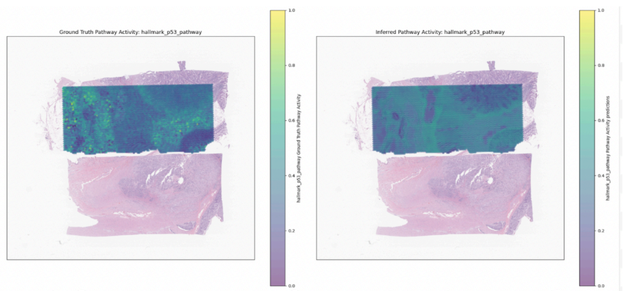

Figure 1. Scanpy scores: p53 ground truth vs.predictions

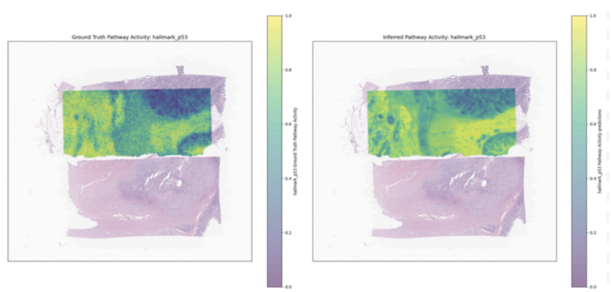

Figure 2. ssGSEA scores: p53 ground truth vs.predictions

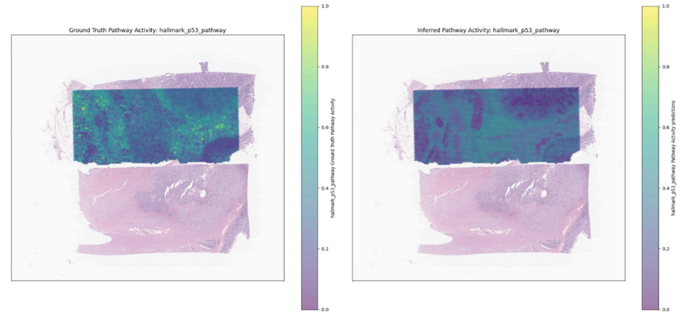

Figure 3. Decoupler scores: p53 ground truth vs.predictions

Averaging Module	Pathway mean absolute error (MAE)					
	KEGG Smooth Muscle Contraction	Hallmark p53	Hallmark Apoptosis	Hallmark WNT Beta Catenin Signaling	Hallmark MTORC1 Signaling	Average
Decoupler	0.120	0.128	0.106	0.0981	0.0786	0.106
ssGSEA	0.163	0.133	0.147	0.115	0.263	0.164
Scanpy	0.197	0.101	0.192	0.106	0.0644	0.132

Table 1. ResNet18 MAE for predicted pathway scores

CONCLUSION & NEXT STEPS

- **Potential for Clinical Impact:**
  - Make the process of acquiring Spatial Transcriptomics data more **affordable**, allowing more organization go into further research
  - Due to the lower cost, many more organizations are able to gain access to and use Spatial Transcriptomics data, making it more **accessible**
- **Limitations:**
  - **Amount of data used for training and testing the models:** Only one WSI was used to train the model, which likely caused overfitting to one patient case, reducing robustness.
  - **Hyperparameter selection:** Batch size, learning rate, and number of epochs were not tuned past our initial choice, meaning that the results of the study were not fully optimized in the current methodology..
- **Future Directions:**
  - Train and test the model **using all 40 WSIs**, and optimize hyperparameters through methods such as a coarse hyperparameter grid search.
  - Evaluating **other architectures** like a GNN or VIT in predicting these pathways
  - Train on **additional pathways** that may result in higher predictive accuracy in order to increase the project's scope and applicability

Acknowledgements: Dr. Joshua Levy, Gokul Srinivasan
Full Paper: https://tinyurl.com/3xktydfc

edit.