Predicting DNA methylation states from pathology images using Deep Learning

Pierre-Louis Cedoz; Quinlan Jung

Department of Biomedical Informatics, Stanford University, Stanford, CA
Department of Computer Science, Stanford University, Stanford, CA

Introduction

DNA methylation is an important mechanism regulating gene transcription, and its role in cancer has been extensively studied. While assays exist to take DNA methylation measurements, they remain extremely expensive and are not performed systematically. On the other hand, pathology images are part of the common procedure in cancer treatment and are relatively cheap. Many studies have focussed on the analysis of DNA methylation in cancer and extracting morphometric features from pathology images but none of them have tried to link these data types.

Project Statement

In this project, we examined the correlation between pathology images and DNA methylation and the opportunity to predict the methylation profile of a patient from the pathology images using deep learning. We used convolutional neural networks on whole slide images to predict the methylation state of the patients and we evaluated our models by computing their classification accuracy on an independent dataset of pathology images.

Dataset

- Pathology images and DNA methylation gathered from The Cancer Genome Atlas (TCGA) data portal.
- Focussed on Glioblastoma Multiforme (GBM) and Lower Grade Glioma (LGG) cancer types
- Labelled each pathology image with the patient's methylation profile for a set of genes. This methylation profile was computed using a computational tool called *MethylMix* developed in Pr Gevaert's Lab.
- For each sample in the dataset, we had the methylation states for 768 genes and the dataset comprised 932 patients.

Data Preprocessing

Whole Slide Images (WSI) are very wide, so it is impossible to feed them directly into the CNN because of memory constraints. To solve this issue, we divided the images into smaller patches and input them into our convolutional neural network (CNN). Each pathology image contained a tissue sample on a slide background, with the tissue being the object of interest. In order to give the CNN the most relevant input, we only kept the tiles that contained at least 90% of tissue using the following steps:

- Apply a grayscale filter: we used a nonlinear luma component (Y') that is calculated directly from gamma-compressed primary intensities as a weighted sum: Y' = 0.299 R + 0.587 G + 0.114 B
- Take the 8-bit image complement
- Perform hysteresis thresholding with an experimentally-chosen high threshold of 100 and a low threshold of 50

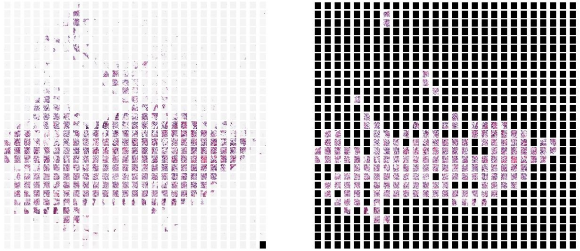
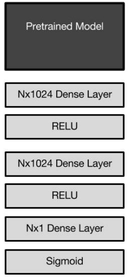


Figure 1 - Left: *A typical cell section,* ***RIght:*** *A filtered image. Black tiles do not contain enough tissue, so they are not included as inputs in our neural network.*

Transfer Learning



Our neural network consisted of a pretrained model attached to custom cells that we trained to specialize in our problem domain:

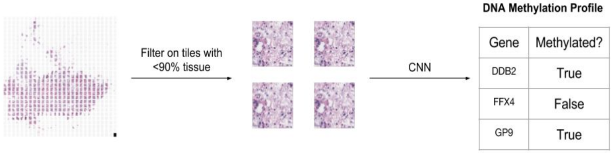
Pretrained model: The pretrained model was initialized with ImageNet weights and was not trained any further to avoid overfitting since our dataset was small. The model took as input a 224x224x3 picture and output a feature representation of the image. We examined the results of Inception V3 and Resnet 50.

Custom Layers: The custom layers consisted of 2 Dense-RELU cells with Dropout connected to a Dense-Sigmoid cell. The final output was {0,1}, signalling the methylation state of a particular gene. These layers were trained using the Adam Optimizer and Binary Cross Entropy loss.

Adam Optimizer: lr=1e-4, B1=0.9, B2=0.999, eps=1e-08, decay=1e-6

The Deep Learning Pipeline

- Divide whole slide images into small tiles
- Filter out the tiles that contain less than 90% of tissue
- Label each image with the methylation profile for a set of genes: 1 for a hyper-methylation and 0 for no methylation. Here, we focussed on gene DDB2
- Training time: Label each patch with the label of the whole slide image and train a CNN to predict this label
- Testing time: Split test image into patches and the model predicts the methylation state for every patches. Then we used majority voting to classify the whole image.



Gene	Methylated?
DDB2	True
FFX4	False
GP9	True

Figure 2 - *Deep Learning pipeline: a filtering step followed by a prediction step*

Results and Discussion

The preprocessing pipeline was computationally expensive and therefore we were only able to preprocess 38 images. We held out 30% of the data as a test set and trained on the remaining 26 images, with a balanced set of labels. The learning curves (loss and accuracy) are plotted on ***Figure 3*** in the case of the ResNet model. The loss was the binary Cross Entropy loss ***(Equation 1)***

Equation 1 - *Binary cross-entropy loss*

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$$

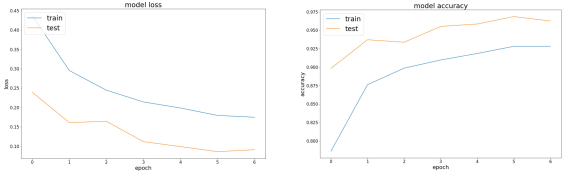


Figure 3 - *Learning curves for the ResNet architecture. Test accuracy is higher than train accuracy because of random effects in the dropout layer during training.*

Then, we computed the classification accuracy of our models on the test set consisting of 12 new images. We obtained an accuracy of 84% with the ResNet architecture and 75 % with the Inception architecture. This proves that the methylation state of a gene can be predicted from the pathology images.

Conclusion and Future Work

In this project, we have shown that it is possible to bridge the gap between pathology images and DNA methylation data. Thanks to deep learning, we could avoid the burden of DNA-methylation profiling and extract all relevant information automatically from the pathology images. However, we have only focussed on one gene so we would need to investigate the generalizability of the model to other genes. For the next steps, we plan to reduce the computation time by running the preprocessing on GPUs. We will also experiment with new network architectures and train a classifier from earlier activations in the pretrained network.