Feasibility of Determining the Presence of Mutations from Tissue Slides

Emre Sevgen

April 2019

1 Background and Goals

CSMD3 is a mutation that affects roughly half (43.6%) of lung cancer cases recorded on Genomic Data Commons as part of the 'TCGA-LUAD' project. Furthermore, it has a noticeable impact on survival.

We use it as a preliminary test system to see whether it is feasible to detect mutations from tissue slides i.e. whether there is any visual signal that can be detected due to a mutation. If so, a more sophisticated approach could be warranted, and lead to a robust detector or pre-screening tool for multiple mutations.

2 Conclusions

A simple CNN trained on a dataset of 180 slides (adding up to roughly 36,000 tiled images, close to 150,000 with data augmentation) performs significantly better ($p < 0.022$) than random on detecting the presence of a CSMD3 mutation based on a binomial test. We conclude that it is possible to determine the presence of mutations from histological samples.

The ROC curve for the CNN is provided below (Fig. 1).

3 Limitations

There was no selection done on type of mutation. There are synonymous mutations in the training set, and it also likely includes a range of silent mutations. We can also expect that some mutations will have a milder phenotype than nonsense or frameshift mutations.

There is no guarantee that the cause of whatever is observed is the mutation, and not something that simply correlates with the mutation. Therefore, if inspecting tiles that are flagged as positive yields a visual sign of mutation, there is no guarantee that a CSMD3 mutation is the direct cause.

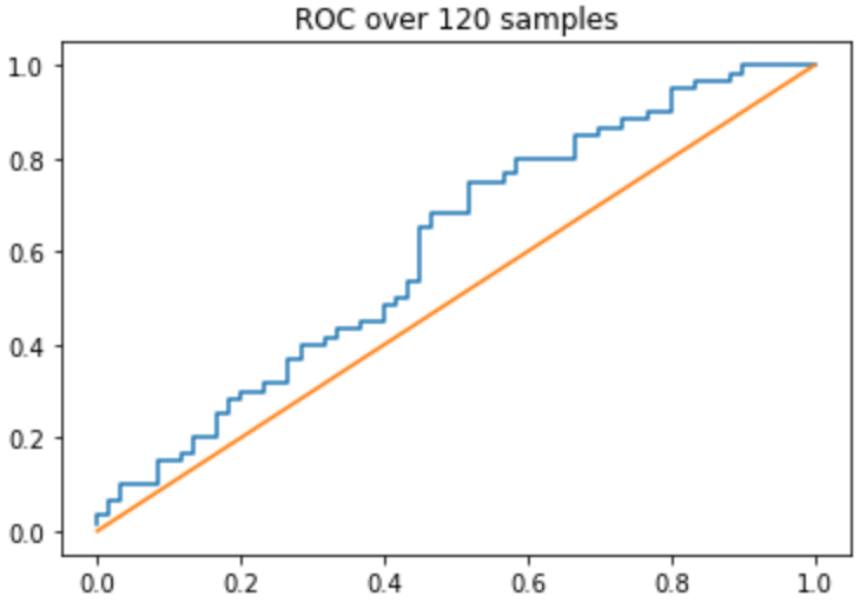

Figure 1: AUC = 0.6067, Accuracy at threshold 0.4 is due to random chance with 2.21% probability and is significant with 97.79% probability according to a simple binomial test.

4 Methods

Slides from the Genomic Data Commons were obtained using the gdc-client tool. The manifests for CSMD3+ and CSMD3- datasets were created by using the online filtering tool to first limit to TCGA-LUAD dataset, and then to mutation positive and negative groups and selecting the first 100 slides for each group ranked by file size, for a total of 200 slides. These slides were split into 180-20 for training and testing (90-10 positive and 90-10 negative).

A separate validation dataset of 120 slides (60 positive, 60 negative) was obtained from Genomic Data Commons for final validation.

Slide images were loaded using OpenSlide Python.

A 4 layer CNN of 32,48,64,80 maps, ending in two dense layers with a 0.1 dropout between CNN layers and 0.5 dropout between the dense layers was built using PyTorch 1.0 and trained on two NVIDIA 1080Ti GPUs as the model.

Each full slide was tiled into 1024x1024 images, converted to 8-bit grayscale, and were filtered based on their average pixel values to exclude empty regions of the slide. The threshold for removal was >200 average pixel value out of 255 possible for the final training and classification. The remaining tiles were downsampled to 256x256 and randomly flipped horizontally or vertically during training (final input being 25% normal, 25% v-flipped, 25% h-flipped and 25% hv-flipped).

Only the maximum output of the entire slide set of tiles for a single slide was used as output for a slide. The reasoning behind is that we did not want to assume that the whole slide would hold a phenotype. Rather, we wanted to account for the possibility that only a few tiles have a signal and the rest are indistinguishable from tiles belonging to the negative set. Another advantage to this approach, if it works, is the possibility of narrowing down what exactly the network is seeing to make its decisions. Finally, this approach is simply agnostic to the number of slides after thresholding, which makes it easy to work with.

Dropout was used both for regularization, and to ensure that different slides contributed the maximum output for classification during training, especially for negative samples.

Models were trained using the Adam optimizer with Binary Cross-entropy loss. A range of learning rates were used, with learning rate decay. Beta1, Beta2 and Epsilon were left at their default values of 0.9, 0.999 and 1e-8 respectively.

The network was not trained to full convergence, nor a proper hyperparameter optimization was performed, due to time constraints. However, this is acceptable as the goal is not to build a finalized, high accuracy detector, only to determine the feasibility of the approach.

5 Future Work

This approach can be easily extended to classify multiple mutations. An interesting point that arises is that some (or most) tissue slides would be positive for multiple mutations. Therefore, one should allow for multiple positives, turning it into a multi-label classification problem.

A pre-trained, larger network can be potentially fine-tuned / partially trained for higher accuracy with shorter training time compared to training a network from scratch.

6 Remarks

Data loading remains a bottleneck, even when parallelized. For any serious model-building, pre-processing the images into a faster readable or accessible format is recommended.

A systematic sweep of learning rate would be beneficial, as training was quite sensitive to this parameter.