Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study

1. Introduction

Tumor purity is the proportion of cancer cells in the tumor tissue. Tumor purity will affect the pathological evaluation of cancer and the selection of samples for genomic analysis. Therefore, tumor purity plays a vital role in cancer clinical detection and cancer genomic research. This research developed a MIL（Multiple Instance Learning）model that can predict tumor purity based on digital histopathology slides stained by H&E. These predictions are highly consistent with genomic tumor purity, which are inferred from genomic data and accepted as the gold standard.

2. Implementation

1) A novel MIL model predicts sample-level tumor purity from H&E stained digital histopathology slides.

2) Spatial tumor purity map which shows the variation of tumor purity over the slide.

3) Learning discriminant features for cancerous and normal histology without requiring annotations from pathologists.

4) Tumor vs. normal sample classification.

3. Significance

1) Provide sample selection basis for high-throughput genome sequencing

High throughput genome sequencing(A massively parallel sequencing technology) is essential for cancer research[1], but its samples need to have sufficient tumor content[2]. The MIL model proposed by this research can be used for high-throughput sample selection for genome analysis, which will help reduce the workload of pathologists and reduce inter-observer variability.

2) Promoting of tumor purity prediction method

The results of tumor purity prediction of the MIL model are highly consistent with the genomic tumor purity. At the same time, it only requires clinician histopathology slides, and almost no manual steps are required, which is cost-effective. At the same time, the MIL model can provide information about the spatial organization of the tumor microenvironment, which is conducive to cancer detection and prevention.

3) Promoting of machine learning method in tumor purity prediction

The MIL model does not require pixel-level annotations. It represents the sample as a bag of patches cut out from the sample slide, and uses the sample-level label as the bag label[3]. Sample-level labels are weak labels and only provide aggregate information instead of pixel-level information. However, they can be easily collected from pathology reports, electronic health records, or different data models.

4. Method

1) Dataset

The dataset used in this research are H&E stained fresh-frozen section histopathology slides and corresponding genomic sequencing data for ten different cohorts in TCGA[4] and one Singapore cohort in East Asian.

The datasets were segregated randomly at the patient level into training, validation, and test sets, which had similar tumor purity. The training set was used to train the machine learning model, the validation set was used to choose the best model, and the test set was used to evaluate the best model.

2) Preprocess

The sample's top and bottom slides are cropped into many patches, and these patches are collected to form a bag(each bag consists of 200 patches in the study). Then, the bags are imported into the model to generate bag level label as the tumor purity of the sample.

3) Model

The novel MIL model proposed in this study consists of three modules: feature extractor module, MIL pooling filter, and bag-level representation transformation module.

The feature extractor is a 18-layer residual network[5], which can eliminate the problem of network vanishing gradient and network degradation. The pooling filter is a distribution pooling filter[6], which can produce stronger bag-level representations. The bag-level representation transformation module is a 3-layer neural network.

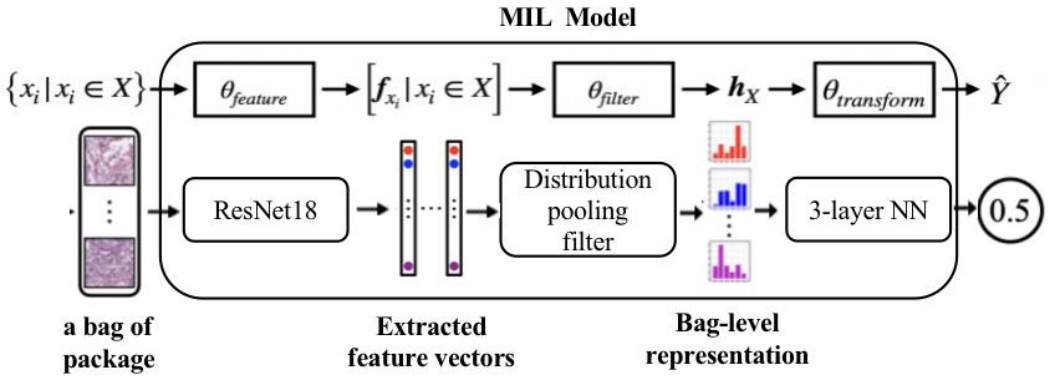

Figure 1:The process of tumor purity prediction

As the fig.1 shows, The feature extractor module extracts a feature vector from each patch in the bag. The genomic tumor purity were taken as the original label of the sample. Then, the MIL pooling filter module aggregates the extracted feature vectors and obtains a bag level representation. Lastly, the bag level representation transformation module transforms bag level representation into predicted bag label as the tumor purity of the sample.

4) Evaluation

To evaluate the performance of the model, a correlation analysis between the genomic tumor purity (obtained from ABSOLUTE[7]) and the MIL model prediction was performed. The Spearman rank correlation coefficient is used as a performance metric. The results shows that the MIL model's tumor purity predictions significantly with genomic tumor purity values.

In addition, this study also examined the average absolute error between the genomic tumor purity and the MIL model prediction. The results shows that compared with the pathologist's estimate of the percent of tumor nuclei, the MIL prediction lower mean absolute error and higher Spearman's correlation coefficient than pathologists' percent tumor nuclei estimates.

5. Conclusion / recommendation

This research proposed a MIL model to predict the tumor purity of any H&E stained histopathology slides and obtained successful result. However, the model mainly tested on the high tumor content sample, checking the model on the low tumor content sample will strengthen the applicability of the model. Besides, the model are deep learning based, training the model with larger cohorts would help to improve the model performance.

6. **Reference**

[1] Schuster, S. C. Next-generation sequencing transforms today's biology. Nature methods 5, 16–18 (2008).

[2] Smits, A. J. et al. The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. Modern Pathology 27, 168–174 (2014).

[3] Quellec, G., Cazuguel, G., Cochener, B. & Lamard, M. Multiple-instance learning for medical image and video analysis. IEEE reviews in biomedical engineering 10, 213–234 (2017).

[4]Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. Nature communications 6, 1–12 (2015).

[5]He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition in Proceedings of the IEEE conference on computer vision and pattern recognition (2016), 770–778.

[6]Oner, Mustafa Umit, et al. Distribution Based MIL Pooling Filters are Superior to Point Estimate Based Counterparts (2020).

[7]Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology 30, 413–421 (2012).