Report

Overview

Notes: In this section, I explain the background and main outcomes of this research by adding my personal understanding.

Tumor purity is the tumor content of the sample, namely the percentage of cancer cells within the tumors, a crucial factor affecting the quality of genomic analysis. Tumor purity has practical significance, as playing important role in choosing therapies, clinical decisions, finding new gene signature and clinical variables.

There are two main approaches to estimate tumor purity: percent tumor nuclei estimation and genomic tumor purity inference. The former one is time-consuming and effected by pathologist's subjective decision. The latter one does not apply to the low tumor content samples and do not provide spatial information of the tumor cells.

This study develops a machine learning model that predicts the tumor purity from H&E stained histopathology slides and repair the deficiency of above methods. Multiple instance learning (MIL) models is one of the models which can predict tumor purity from digital histopathology slides, with advantages of convenience. This study designed a novel MIL model to predict tumor purity from H&E stained histopathology slides. Authors represent each sample as a bag of patches cropped from the sample's top and bottom slides and use the sample's genomic tumor purity as the bag label. Our MIL model has a novel 'distribution' pooling filter that produces stronger bag-level representations from patches' features than standard pooling filters like max and mean pooling. The histopathology slides in each cohort were randomly segregated at the patient level into training, validation, and test sets. Then, authors trained MIL model on the training set, chose the best set of model weights based on validation set performance, and evaluated the best model on the held-out test set.

The MIL models successfully predicted tumor purity from histopathology slides of both fresh-frozen sections and formalin-fixed paraffin-embedded sections. Besides, authors found that the top and bottom slides of a sample were significantly different in tumor purity, which showed that tumor purity varies spatially within the sample. The models also learned discriminant features for cancerous vs, normal histology and authors validated that this segmentation is correct. Lastly, these models showed outstanding ability to classified samples into tumor vs. normal.

Result

Notes: In this section, I conclude the results of this research in popular and easy-to-understand language.

1. The MIL model's tumor purity predictions correlate significantly with genomic tumor purity values. Authors used Spearman's rank correlation coefficient to evaluate model's performance. The results showed that models performed well in 8 cohorts. Subsequently, authors repeated the same analyses

between genomic tumor purity values and pathologist's percent tumor nuclei estimates. Compared with the latter one, MIL predictions are more reliable with genomic tumor purity values. Besides, authors used Fisher's z transformation-based method to assess two methods. In most cohorts (except LUSC and PRAD), MIL predictions performed better.

2. MIL model's predictions have lower mean absolute error than percent tumor nuclei estimates
   Authors also checked the mean-absolute errors to evaluate MIL models and pathologists' percent tumor nuclei estimates. The result illustrated that percent tumor nuclei estimates were generally higher than genomic tumor purity values. After performing Wilcoxon signed-rank test on absolute error values for tumor samples in the test sets, absolute error values in MIL predictions were significantly lower than ones in pathologists' tumor nuclei estimates in all cohorts except the LGG cohort, means the former is more reliable.
3. The MIL model predicts tumor purity from H&E stained slides of FFPE sections in the Singapore cohort
   Afterwards, authors evaluated MIL models' performance on slides of formalin-fixed paraffin-embedded (ffpe) sections in a local Singapore cohort. They used transfer learning and adjusted the weight of all layers in feature extractor module, because of the difference of freeze-frozen sections and ffpe sections. The result was good. However, when checking models directly, the result was not good enough, which means the difference between two preserve methods shouldn't be ignored.
4. Tumor purity varies spatially within a sample: top and bottom slides of a sample are different in tumor purity
   Authors performed trained MIL models on the top and bottom slides of a sample to check the intra-tumor heterogeneity, which is one of the reasons of therapeutic failure and drug resistance. The results demonstrated that tumor purity varies spatially within the sample.
5. Predicting the tumor purity of a sample by using both top and bottom slide is better than using only one slide
   To check whether the amount of slide used can influence the accuracy of tumor purity prediction, authors compared their absolute errors. The results showed that in most test sets, using both slides for tumor purity prediction gave better results. And in the most homogeneous tumors, there were no significant differences. Thus, authors think that predicting a sample's tumor purity using both top and bottom slides together is better.
6. Spatial tumor purity map analysis reveals the probable cause of pathologists' high percent tumor nuclei estimates
   To investigate why pathologists' percent tumor nuclei estimates were generally higher than genomic tumor purity values, authors performed error analyses repeatedly by gradually extending the region-of-interest from the highest purity region and the mean-absolute-error increased. These observations suggested that pathologists may tend to select high tumor content regions to estimate percent tumor nuclei.
7. The MIL model learns discriminant features for cancerous vs. normal tissue histology
   Features of patches cropped over the slides of the tumor and normal samples were extracted using the trained feature extractor module of the MIL model. Then, slide-level cancerous vs. normal segmentation maps were obtained by performing a clustering over the extracted feature vectors. Authors performed a qualitatively evaluation. As a result, MIL model learned discriminant features for cancerous vs. normal tissue histology.
8. The MIL model successfully classifies samples into tumor vs. normal

Authors checked our MIL model's performance in the tumor vs. normal sample classification task by using receiver operating characteristic (ROC) curve analysis and calculate the are under the ROC curve (AUC) as a standard to evaluate the accuracy of the classification. The results show that the MIL models performed well in classifying samples into tumor vs. normal.

Learning Outcome

In *Tumor Biology*, I know the concept of tumor heterogeneity, which means cancer cell perform differently in each patient or even in a specific tumor tissue. Tumor heterogeneity has many adverse effects on patients. Assuming there is a tumor tissue consisting of various type of tumor cells. Some type of them may be sensitive to the drug or treatments, but others may resist. As time passes, the sensitive cells will disappear and the other cells will be dominant in human body. By this time, the previous treatment can't work anymore. This is one of the biggest obstacles which researchers and doctors are confronting, namely drug resistance. In conclusion, to fight against cancer, we have to evaluate the tumor as thorough as possible, including its size, location, type, stage, grade of differentiation and so on, to make an optimal clinical decision on each patient.

In this paper, I know well about tumor purity and its practical significance. Tumor purity is a crucial prognostic biomarker. In my understanding, it is one of the indicators to describe tumor heterogeneity. It also affects the quality of molecular data acquisition and analysis. But the current technics can't calculate the tumor purity efficiently and accurately, so there is a limitation on its clinical application. Machine learning is a promising approach to solve this problem. In *Biomarker Screening and Its Mechanism Analysis Based on Machine Learning, Statistics and Bioinformatics*, I used machine learning model to extract features for the first time. According to our requirements, we utilized random forest algorithm to screen microRNA and test the model in support vector machines (SVM). At that time, I started to construct my knowledge system of machine learning.

After reading this paper, I understand a rigorous logic to evaluate a model and how to maximize its value. Although the initial purpose of this research is to conduct a model to predict the tumor purity, we can expand its breadth and depth. In addition to predict the different preserve methods (freeze-frozen sections and ffpe sections), we can also investigate the spatial variety of tumor purity. Moreover, evaluating the optimal approach (using both the top and bottom of the slides or only using one slide) in application is necessary too. For some problems which doesn't have a rational explanation (percent tumor nuclei estimates by different pathologists are different from genomic), the causes may be found in ingenious way (expanding the ROI gradually). On the other hand, I learn many indicators to assess the difference and accuracy, such as Spearman's rank correlation, mean-absolute errors and AUC. In conclusion, I broaden my horizon in machine learning and tumor biology by reading this paper.

I have solid background in biotechnology and biomedical. In my previous learning and research, I am familiar with bioinformatics tools and I can utilize them masterly. Admittedly, my knowledge of computer science is not strong enough. But this is exactly the reason I want to apply for your MSc Biomedical Data Science. In this age of information explosion, I firmly believe we can make good use of data to promote the development of biomedicine. I wish I can get involved in this massive undertaking. I sincerely hope that I can get your admission notice in the near future.