Report - Spatially Resolved Tumor Purity Maps

Zhiling Yan

December 7, 2021

1 Introduction

The tumor micro-environment is complex, comprising various cellular populations. The cancer cell fraction is an essential factor in genomic analyses and in interpreting the clinical properties of the tumor [1]. So how to accurate assess the percentage of cancer cells is an important issue, and the cancer cell fraction within the tumor is called tumor purity [2].

There are two traditional approaches to solve the issue: percent tumor nuclei estimation [3, 4] and genomic tumor purity inference. Pathologists applied different types of genomic data, such as somatic copy number data [5], somatic mutations data [6] and gene expression data [7, 8] to infer the tumor purity. However, the percentage of tumor nuclei estimated by pathologists was usually higher than genomic tumor purity values, while genomic tumor purity inference lost the spatial information of the tumor micro-environment, which is important in therapeutic response and spatial-omics analyses [9, 10].

Recently, with the success of machine learning, patch-based models and multiple instance learning (MIL) models have been applied in the assessment of tumor purity. With rarely available and expensive pixel-level annotations, patch-based models extracted patch features using support vector machines (SVMs) [11] or deep neural networks (DNNs) [12, 13], then predict tumor cellularity and estimated tumor purity.

According to the discussion above, Mustafa Umit Oner et al [14] propose a novel MIL model to obtain Spatially Resolved Tumor Purity Maps (SRTPMs) and to predict tumor purity from HE stained histopathology slides, and this model is called SRTPMs later in this report. Unlike pixel-level annotations providing whether each cell is cancerous or normal, sample's genomic tumor purity is used as the bag label. A residual learning framework is implemented to extract features from a bag of patches cropped from samples' slides. Then they design a novel pooling filter superior to standard pooling filters, which will be specifically discussed later. The last module transforms the representation and predicts the tumor purity. Their experiments show that SRTPMs successfully classifies samples into tumor vs. normal with highly consistent with the golden standard, and with a higher correlation and lower mean-absolute-error with genomic tumor purity values.

2 Model and Experiment

As shown in Figure 1, the model SRTPMs consists of three modules, including feature extractor module, MIL pooling filter module and bag-level representation transformation module. **Feature Extractor.** SRTPMs implements ResNet18 [15] model, which is a residual learning framework that explicitly reformulates the layers as learning residual functions with reference to the layer inputs. The objective is to predict bag label Y for a given bag of instances $X = \{x_i | x_i \in I, i = 1, 2, ..., N\}$, where N is the number of instances per bag. With *feature extractor model*, X is reflected to $F_X = \left\{f_{x_i} | f_{x_i} \in \mathbb{R}^J, i = 1, 2, ..., N\right\}$. **Distribution Pooling Filter.** This module obtains $F_X \in \mathbb{R}^{JN}$ from *feature extractor* and applies kernel density estimation to estimate marginal distributions.

$$\tilde{p}_X^j(\nu) = \sum_{i=1}^{N} \beta_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\nu - \alpha_i f_{x_i}^j)^2}$$

Attention weights are fixed to $\alpha_i = 1, \forall i$, $\beta_i = \frac{1}{N}, \forall i$ and the standard deviation $\sigma = 0.05$. **Bag-level Representation Transformation** : In this module, featured vectors are transformed by three

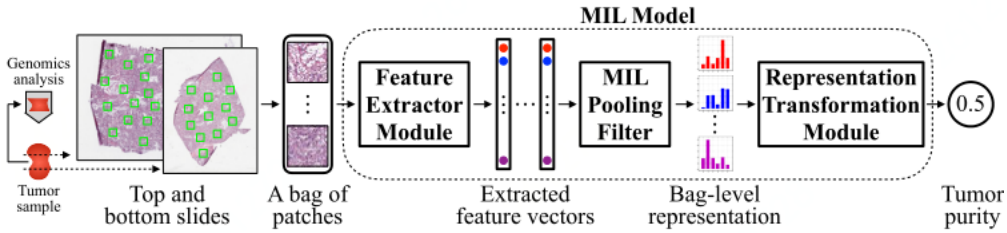

Figure 1: The structure of the novel MIL model.

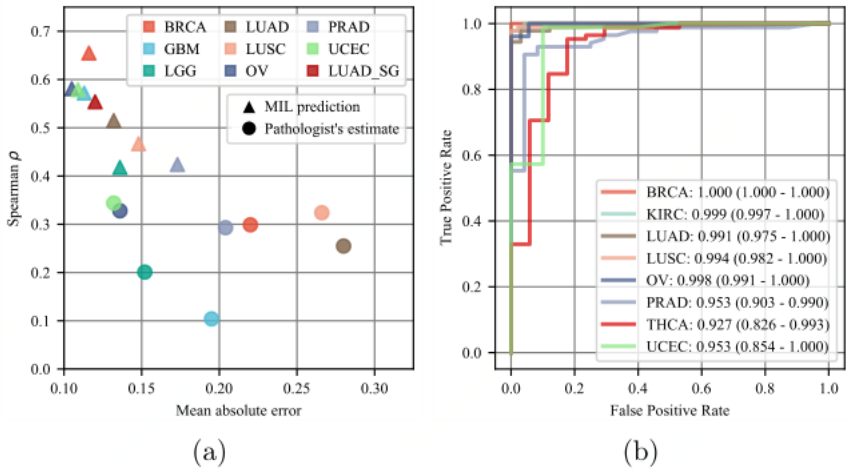

Figure 2: (a)MIL models' predictions achieve lower mean absolute error and higher Spearman's correlation coefficient than percent tumor nuclei estimates. (b) MIL models successfully classified samples into tumor vs. normal in all cohorts.

layers of feed-forward neural network affiliated with activation function ReLU and dropout.

The SRTPMs model successfully assessed tumor purity in eight TCGA cohorts and in a local Singapore cohort. In the local Singapore cohort, the given slides are formalin-fixed paraffin-embedded (ffpe) sections, which requires extra process compared with fresh-frozen sections. As shown in Figure 2(b), SRTPMs classified samples into tumor vs. normal with high precision. This model obtained tumor purity predictions of all samples of each cohort and conducted receiver operating characteristic (ROC) curve analysis. Then authors calculated the area under the ROC curve, which is a significant metric, and apply 95% confidence interval (CI) in analysis. The threshold is settled as 0.927, and the highest AUC value is obtained in BRCA cohort, the largest cohort with 559 patients. Besides of the strong ability of classification, SRTPMs performs well in tagging sample-level genomic tumor purity, by learning the discriminant features for cancerous vs. normal tissue histology.

In addition, they conducted correlation analyses using Spearman's rank correlation coefficient as the performance metric, and found significant correlation between the MIL models' predictions and the genomic tumor purity values. Moreover, MIL models' predictions show lower mean absolute error than percent tumor nuclei estimates (Figure 2(a)) using the Wilcoxon signed-rank test [16].

3 Conclusion

The novel MIL model aims to predict tumor purity with weak labels. The system has a distribution pooling filter that produces stronger bag-level representations with marginal distribution estimation. The model shows a good performance in all cohorts including fresh-frozen sections and formalin-fixed paraffin-embedded (ffpe) sections with reservation of spacial information, high correlation with golden standard and lower mean absolute error.

References

[1] S. Haider, S. Tyekucheva, D. Prandi, N. S. Fox, J. Ahn, A. W. Xu, A. Pantazi, P. J. Park, P. W. Laird, C. Sander, *et al.*, "Systematic assessment of tumor purity and its clinical implications," *JCO precision oncology*, vol. 4, pp. 995–1005, 2020.

[2] D. Aran, M. Sirota, and A. J. Butte, "Systematic pan-cancer analysis of tumour purity," *Nature communications*, vol. 6, no. 1, pp. 1–12, 2015.

[3] X. Zheng, N. Zhang, H.-J. Wu, and H. Wu, "Estimating and accounting for tumor purity in the analysis of dna methylation data from cancer studies," *Genome biology*, vol. 18, no. 1, pp. 1–14, 2017.

[4] L. Bao, M. Pu, and K. Messer, "Abscn-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data," *Bioinformatics*, vol. 30, no. 8, pp. 1056–1063, 2014.

[5] S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, *et al.*, "Absolute quantification of somatic dna alterations in human cancer," *Nature biotechnology*, vol. 30, no. 5, pp. 413–421, 2012.

[6] F. Favero, T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund, "Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data," *Annals of Oncology*, vol. 26, no. 1, pp. 64–70, 2015.

[7] K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Treviño, H. Shen, P. W. Laird, D. A. Levine, *et al.*, "Inferring tumour purity and stromal and immune cell admixture from expression data," *Nature communications*, vol. 4, no. 1, pp. 1–11, 2013.

[8] Y. Li, D. M. Umbach, A. Bingham, Q.-J. Li, Y. Zhuang, and L. Li, "Putative biomarkers for predicting tumor sample purity based on gene expression data," *BMC genomics*, vol. 20, no. 1, pp. 1–12, 2019.

[9] V. Svensson, S. A. Teichmann, and O. Stegle, "Spatialde: identification of spatially variable genes," *Nature methods*, vol. 15, no. 5, pp. 343–346, 2018.

[10] J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, *et al.*, "Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region," *Science*, vol. 362, no. 6416, 2018.

[11] P. W. Hamilton, Y. Wang, C. Boyd, J. A. James, M. B. Loughrey, J. P. Hougton, D. P. Boyle, P. Kelly, P. Maxwell, D. McCleary, *et al.*, "Automated tumor analysis for molecular profiling in lung cancer," *Oncotarget*, vol. 6, no. 29, p. 27938, 2015.

[12] Z. Pei, S. Cao, L. Lu, and W. Chen, "Direct cellularity estimation on breast cancer histopathology images using transfer learning," *Computational and mathematical methods in medicine*, vol. 2019, 2019.

[13] A. Rakhlin, A. Tiulpin, A. A. Shvets, A. A. Kalinin, V. I. Iglovikov, and S. Nikolenko, "Breast tumor cellularity assessment using deep neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[14] M. U. Oner, J. Chen, E. Revkov, A. James, S. Y. Heng, A. N. Kaya, J. J. S. Alvarez, A. Takano, X. M. Cheng, T. K. H. Lim, *et al.*, "Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study," *bioRxiv*, 2021.

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

[16] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*, pp. 196–202, Springer, 1992.