MFmap: A semi-supervised generative model matching cell lines to cancer subtypes

Xiaoxiao Zhang[1,2], Maik Kschischo[1]

[1] Department of Mathematics and Technology, RheinAhrCampus, University of Applied Sciences Koblenz, 53424 Remagen, Germany
kschischo@rheinahrcampus.de
[2] Department of Informatics, Technical University of Munich, 81675 Munich, Germany

Abstract. Cell lines are widely used experimental models in cancer research. However, translating preclinical findingsinto clinical applications is limited by the discordance between cell lines used and tumours. We developedthe model fidelity map (MFmap), a semi-supervised generative model to integrate high-dimensional geneexpression, copy number variation and somatic mutation data of both tumours and cell lines into a small set of features that are highly associated with cancer subtypes, and predict the cell line subtypes simultaneously. These low-dimensional features are biologically interpretable and can be used for matching a given cell line to individual tumours. This enables cancer researchers to select the best cell line model for their experiments.The high accuracy (test set F 1 score > 90%) of MFmap cancer subtype prediction is validated in ten different cancer datasets. From an application perspective, we demonstrate how the predicted cancer subtype for cell lines can be exploited for discovering drug sensitivity differences among subtypes in glioblastoma and breast cancer. This is helpful for guiding personalised treatment decisions and could facilitate drug repurposing for cancer treatments. Thanks to its generative nature, MFmap enables the analysis of cellular status transitions during cancer progression. In particular, we show that artificially perturbing cancer samples from a baseline subtype to an aggressive subtype indeed acquires marker features unique to the targeted subtype in glioblastoma. From a methodological perspective, the newly derived loss function of our MFmap allows to jointly train a classification model and a generative model on both labelled (tumours) and unlabelled (cell lines) data in an one-step-optimisation manner. We further empirically show that the MFmap can achieve excellent classification accuracy and good generative performance simultaneously. These results show that the MFmap will be useful for many semi-supervised prediction tasks in the biomedical sciences and beyond.

This work was supported by the FOR2800 research unit funded by the Deutsche Forschungsgemeinschaft.