One application of machine learning in cancer

WU SIYI

Abstract This report first presented an overview of the causes of carcinogenesis in biology and the role of omics data in cancer research, while tumour purity is inferred from different types of omics data and contributes to pathologic evaluation. The application of machine learning in tumour purity is also mentioned in this report.

1 Introduction

The past few decades have seen tremendous progress in understanding cancer at the molecular level. It is now known that the progressive transformation of normal cells into malignant cancer cells requires the sequential acquisition of mutations arising from genomic damage[1]. Cancer cells are infinitely proliferating, transformable and prone to metastasis and they require a large supply of nutrients and oxygen to survive. Machine learning(ML), which was first paraphrased by Arthur Samuel[2]， enables computers to learn from data without being explicitly programmed and make accurate predictions. Therefore, this technology can be used to model cancer progression and improve treatments. And appropriate levels of validation are required in order to increase prediction accuracy.

2 Causes of carcinogenesis in biology

Random mutations in the genes which control proliferation or apoptosis are responsible for cancer. Most of these mutations are not inherited but arise spontaneously as a consequence of chemical damage to DNA resulting in the altered function of crucial genes[1]. DNA as a molecule depends on the formation and breaking of chemical bonds. It is therefore often susceptible to chemical damage. This damage can be the consequence of endogenous processes such as errors in replication of DNA. In addition, exogenous damage including UV radiation and chemical carcinogens is also a significant factor. In general, carcinogenesis involves processes by which genes that control proliferation and cell death suffer mutational damage, resulting in the activation of oncogenes and the inactivation of tumour suppressor genes.

3 Abundance of omics data in cancer

With the rapid development of high-throughput sequencing technologies, the explosion of biomedical data like omics has greatly contributed to the understanding of cancer at the molecular level. The NGS-based massively parallel sequencing enabled researchers to characterize the mutational landscapes of a given tissue. For instance, in the project “The Cancer Genome Atlas” (TCGA), the scientists employed the NGS coupled with bioinformatics analysis to discover somatic mutational landscape[3]. High-throughput genomics is also essential for cancer research and hold the promise of personalized medicine[4]. The epigenomics studies have played an integral role in uncovering the disease-associated epigenetic markers[5]. Proteomics can be used to find suitable cancer biomarkers by comparing and analyzing the differences between the proteomes of diseased individuals and normal individuals. While a single type of omics study can reveal a great deal of information about cancer at a unidirectional level, the complexity of cancer requires the need for integrated multi-omics research.

4 Key indicator in cancer: Tumour purity

Omics data provide a means of quantifying tumour purity. Tumour purity represents the fraction of cancer cells in a tumour. It was estimated either by expert pathologists reviewing tumour sections[6] or in silico (using epigenomic, genomic, or transcriptomic profiles)[7]. The purity of the tumour is of great clinical importance. For instance, Low tumour purity is associated with poor prognosis in colon cancer[8] . The purity of the tumour also has an impact on both the acquisition and analysis of high throughput data. Besides, spatial tumour purity maps can help better understand the tumour microenvironment as a key determinant in tumour formation and therapeutic response[9].

5 Machine learning in tumour purity

Machine learning generally consists of two processes: (i) the process of learning from data to obtain a model is called “learning” or “training”. (ii) the process of using the learned model to make predictions is called "testing". The ultimate goal of machine learning is to make the learned model can be used to perform classification, prediction, estimation or any other similar task. Based on how a method learns from the data, the ML techniques can be broadly categorized into supervised and unsupervised approaches. Supervised algorithms learn patterns from a limited amount of annotated training data, and then use the knowledge gained to classify the remaining test data. Instead，the unsupervised algorithm first defines patterns of objects in a pool of data with unknown annotations or attributes or features, and then uses the knowledge gained to classify the remaining data[10].

ML methods are now also being applied to multi-omics data to investigate and interpret the relationships between data and phenotypes[11][12]. These methods can discover and identify patterns from complicated data while they can effectively predict the future outcome of cancer[13]. Large initiatives such as TCGA have allowed the use of omics data for the training of ML algorithms. For instance, scientists have successfully used deep Multiple Instance Learning (MIL) model to predict tumour purity from histopathology slides in different TCGA cohorts and obtained spatially resolved tumour purity maps[9]. The innovation of this experiment is the use of a novel ‘distribution’ pooling filter，which is superior to point-estimate based MIL pooling filters, like max-pooling or mean-pooling[14]. Moreover, it should be mentioned that in spite of the claims that this model can result in accurate tumour purity, the sample size, the sample quality as well as the careful feature selection schemes are of great importance for effective models. In addition, in order to obtain accurate results for predictive model, internal and external validation should be performed in the study that would enable the extraction of more accurate and reliable predictions while it would minimize any bias.

6 Conclusion

Humans have some explanations for the causes and processes of cancer at a certain extent, while the vast number of omics data generated by cancer play an important role in future research such as supporting the quantification of tumour purity. When dealing with massive amounts of data, machine learning has the ability to detect key features from complicated datasets in order to build reasonable predictive models, which contribute to applications in cancer research. Concerning the future of ML in cancer still should be studied for overcoming the different limitations.

References

[1] Bertram, J. S. (2000). The molecular biology of cancer. *Molecular aspects of medicine*, *21*(6), 167-223.

[2] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229.

[3] Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., ... & Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *502*(7471), 333-339.

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

[5] Chakraborty, S., Hosen, M., Ahmed, M., & Shekhar, H. U. (2018). Onco-multi-OMICS approach: a new frontier in cancer research. *BioMed research international*, *2018*.

[6] Smits, A. J., Kummer, J. A., De Bruin, P. C., Bol, M., Van Den Tweel, J. G., Seldenrijk, K. A., ... & Vink, A. (2014). The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Modern Pathology*, *27*(2), 168-174.

[7] Yadav, V. K., & De, S. (2015). An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Briefings in bioinformatics*, *16*(2), 232-241.

[8] Mao, Y., Feng, Q., Zheng, P., Yang, L., Liu, T., Xu, Y., ... & Xu, J. (2018). Low tumor purity is associated with poor prognosis, heavy mutation burden, and intense immune phenotype in colon cancer. *Cancer management and research*, *10*, 3569.

[9] Oner, M. U., Chen, J., Revkov, E., James, A., Heng, S. Y., Kaya, A. N., ... & Lee, H. K. (2021). Obtaining Spatially Resolved Tumour Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study. *bioRxiv*.

[10] Mahmud, M., Kaiser, M. S., McGinnity, T. M., & Hussain, A. (2021). Deep learning in mining biological data. *Cognitive Computation*, *13*(1), 1-33.

[11] Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, *17*(2), 167-177.

[12] Kim, M., & Tagkopoulos, I. (2018). Data integration and predictive modeling methods for multi-omics datasets. *Molecular omics*, *14*(1), 8-25.

[13] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.

[14] Oner, M. U., Kye-Jet, J. M. S., Lee, H. K., & Sung, W. K. (2020). Studying The Effect of MIL Pooling Filters on MIL Tasks. *arXiv preprint arXiv:2006.01561*.