Topic 4: Development of tumour micro-environment transcriptomic hypoxia signatures using autoencoders

Tutor: Delmiro Fernandez-Reyes

Hypoxia (low oxygen levels) in the tumour microenvironment plays a pivotal role in tumour progression and resistance to therapy, making it a crucial aspect to explore and understand in cancer research. In this project we will focus on developing transcriptomic hypoxia signatures of the tumour microenvironment utilising autoencoders (AEs). In the literature review we will develop a systematic understanding of how transcriptomic hypoxia signatures have been previously identified using machine learning and utilised in cancer research. We will also systematically identify datasets, methodologies, and results of previous studies, laying a foundation for the research project. The literature review will also focus on gaining an understanding of autoencoders and identifying relevant previous work of their application for linking tumour microenvironment with cancer prognosis.

**Identification of transcriptomic hypoxia signature from tumour microenvironment by machine learning: a systematic review**

Abstract

This systematic review comprehensively analyses the application of machine learning in identifying transcriptomic hypoxia signatures (HSs) within the tumour microenvironment, a key factor in understanding tumour progression, tailoring treatment strategies, and potentially improving patient outcomes. Certain HSs have been validated to stratify patients into different risk groups and predict overall survival (OS) (Liu et al., 2020). Traditional methods for identifying HSs, including in vivo–derived analyses using hypoxia markers(Marotta et al., 2011), meta-analysis of microarray-based gene signatures(Seigneuric et al., 2007), and gene function and co-expression patterns analysis (Buffa et al., 2010), provided foundational insights. However, since 2014 (Lalonde et al., 2014), machine learning has revolutionized this field. Its ability to efficiently analyse complex, large datasets has significantly enhanced the accuracy and depth of HS identification. This review synthesizes findings from a comprehensive analysis of literature focused on using machine learning methods to identify HSs in the tumour microenvironment, adhering to strict topic relevance criteria. This review reveals that unsupervised clustering algorithms are predominant in this area. Besides, while unsupervised clustering is well-established, this review highlights an untapped potential in the application of autoencoders (AEs). The insights gained from this review could pave the way for the HS identification for a broader range of tumour types and the development of novel identification methods.

Keywords: hypoxia signature, tumour, cancer, microenvironment, machine learning, prognosis, autoencoder

Introduction

A HS, a set of genes differentially expressed under hypoxic conditions in the tumour microenvironment, plays a pivotal role in measuring tumour hypoxia. Hypoxia, prevalent in various tumours or cancers, is linked with increased tumour aggressiveness, reduced therapy response, especially to radiotherapy (Toustrup et al., 2012), and an overall poorer clinical prognosis(Tawk et al., 2016). Despite of the importance of hypoxia in tumours, its detection remains to be a challenge. Using identified HSs as biomarkers, the intratumoural hypoxia level can be inferred by measuring the expression level of certain HSs. Moreover, HS can also be use to predict more valuable information, including response to immunotherapy (Hong et al., 2021). However, traditional HSs identification has faced significant limitations.

Machine learning, including techniques such as autoencoders (AE), presents a transformative approach in advancing the identification of HSs. AEs, specialized in unsupervised learning tasks like dimensionality reduction and feature learning, offer potential solutions for identifying gene signatures with increased precision. Compared to traditional methods, machine learning requires less empirical gene signature knowledge, automates the signature selection process, and efficiently handles larger, more complex gene datasets. This review is necessitated by the need to systematically examine these advanced machine learning techniques, assessing their impact and potential in refining HS identification.

This review aims to systematically analyse and synthesize all the existing literature on the application of machine learning methods for identifying HSs in the tumour microenvironment. Given the nascent nature of this research area, our scope will be broad and inclusive, covering studies across various tumour types and machine learning techniques.

To effectively collate and assess the findings from the selected papers, the narrative synthesis approach is employed. This method will allow qualitatively summarization and explanation of the diverse range of methodologies and outcomes observed in the studies. The narrative synthesis is particularly suitable given the expected heterogeneity in study designs, machine learning algorithms used, and types of tumours examined. By synthesizing this information narratively, we aim to draw comprehensive insights into the current state of machine learning applications in HS identification and highlight potential directions for future research.

Objectives

The primary objective of this review is to explore and synthesize current methodologies and findings in the field of tumour microenvironment transcriptomics, particularly focusing on the development of HSs using AEs. The research questions guiding this review are:

1. How is machine learning being utilized to develop transcriptomic HSs in the tumour microenvironment?
2. How effective are AEs in comparison to other machine learning techniques in identifying and characterizing HSs?

Methodology

1. Identification of Studies
   1. Searching Strategies

In this systematic review, we employed an "Iterative and Expert-Informed Search Strategy" to meticulously gather a comprehensive set of studies on machine learning applications in identifying HSs within tumour microenvironments. This approach involved dynamically refining search terms based on initial findings and ongoing insights.

We primarily utilized PubMed, noted for its repository of peer-reviewed biomedical literature, ensuring access to high-quality scientific studies. Recognizing its limitations in encompassing newly published or grey literature, we supplemented our search with Google Scholar. While Google Scholar offers a broader range of literature, including grey and non-peer-reviewed articles, it may include sources with varying degrees of rigor.

Our search terms, carefully selected to cover critical aspects of our research theme, included (hypoxia), (signature), (transcriptomic), ((cancer) or (tumour)), ((micro-environment) or (TME) or (microenvironment)), and ((develop) or (autoencoders) or (AE) or (machine learning) or (deep learning)). Given the field's novelty, we imposed no date restrictions, with relevant studies dating back to 2014. This strategy yielded 90 papers from PubMed and 13 from Google Scholar, augmented by three expert-recommended papers, enriching the search's breadth and depth. For full transparency and reproducibility, the complete search methodology and iterations are detailed in Appendix A.

* 1. Inclusion/Exclusion Criteria and Screening

After the removal of duplicates, the screening of papers was conducted by the author. The initial stage involved evaluating titles and abstracts against the inclusion criteria outlined in Table 1. This criteria was subsequently refined based on the outcomes of the first screening, leading to the inclusion of studies that employed statistical methods for identifying HSs.

The decision to incorporate statistical methods alongside machine learning in this review is justified for several reasons. Primarily, the distinction between statistics and machine learning is often blurred, as both disciplines share fundamental methodologies. This overlap can make it challenging to categorize certain techniques exclusively under one field. Furthermore, the inclusion of statistical methods broadens the scope of the review, encompassing a more extensive range of data analysis approaches. This approach is critical, as a preliminary survey of the literature indicated a prevalent use of statistical methods in contrast to the more limited application of pure machine learning techniques. Thus, this review aims to bridge this gap, offering a holistic understanding of the interplay between statistical and machine learning methodologies in the context of hypoxia signature identification.

Papers were included in the review based on their alignment with the refined inclusion criteria, which encompassed topic relevance, utilization of specified technologies, and validation of results. This process resulted in the selection of 17 papers. Additionally, consultations with field experts led to the identification of three more pertinent publications, bringing the total count to 20 papers included in the review.

|  |  |
| --- | --- |
| **Inclusion Criteria** | **Reason** |
| Focus on machine learning or statistical methods in signatures identification | Ensures that the paper contributes to the understanding of how machine learning or statistical methods are used in the identification of signatures, which is central to the review's aim. |
| Focus on HSs in tumour microenvironment | Directly aligns with the review's objective to understand how HSs are identified in the context of tumour microenvironments. |
| Development of signature conducted on real gene profiles | Ensures that the study's findings are based on empirical data, enhancing the practical applicability and relevance of the research. |
| Validation of signature development | Validation of the developed signatures is crucial to ascertain their reliability and utility in real-world applications. |
|  | |
| **Exclusion Criteria** | **Reason** |
| Not focusing on HSs (e.g., immune signatures) | Studies focusing on other types of signatures are not relevant to the specific focus on HSs and thus fall outside the scope of the review. |
| Not focusing on gene signatures (e.g., HIF) | Studies not cantered on gene signatures do not contribute to the understanding of transcriptomic HSs, which is the core interest of the review. |
| Non-tumour studies (e.g., DSS-induced colitis) | The review is specifically interested in tumour-related hypoxia, so studies on other diseases are not within its scope. |

Table 1. Inclusion and exclusion criteria

1. Coding and Analysing
   1. Data extraction and Critical Appraisal

The data extraction strategy employed in this systematic review was meticulously designed based on the CHARMS checklist (Moons et al., 2014). This approach ensured that only terms relevant to the review's focus were included, with non-relevant terms being removed or adjusted appropriately. The extracted data was stored in a separate document for clarity and ease of analysis. Key data elements extracted included TUMOUR type, data source, data size, model development, performance, evaluation results, and interpretative and discussion data.

The extracted data were then rigorously assessed by the author following the Critical Appraisal guidelines provided by the PROBAST tool (Wolff et al., 2019) . This appraisal aimed to evaluate the risk of bias and the applicability of the studies included in the review.

The critical appraisal process revealed a significant level of bias within the studies: approximately 40% were found to have a high risk of bias, with 30% demonstrating significant concerns regarding applicability. Notably, the most substantial risk of bias was concentrated in the domains of data and analysis assessment.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Author, Year** | **Risk of Bias** | | | | **Applicability** | | | **Overall** | |
| 1. Participants | 2. Predictors | 3. Outcome | 4. Analysis | 1. Participants | 2. Predictors | 3. Outcome | **Risk of Bias** | **Applicability** |
| Jill M. Brooks, 2019 | + | + | + | + | + | + | + | + | + |
| Edian, 2013 | + | + | + | + | + | + | + | + | + |
| Cheng-Peng Gui, 2021 | - | + | + | ? | + | + | + | - | + |
| Yifan Liu, 2020 | + | + | ? | ? | + | + | + | ? | + |
| Emilie, 2014 | + | + | + | + | + | + | + | + | + |
| Jia Li, 2022 | + | + | + | - | + | + | - | - | - |
| Run Shi , 2021 | + | + | + | + | - | + | + | + | - |
| Jun Shao, 2021 | - | + | + | + | + | + | + | - | + |
| Baohui Zhang, 2020 | - | + | + | - | + | + | + | - | + |
| Qiangnu Zhang, 2021 | + | + | + | + | + | ? | + | + | ? |
| Fanhong Zeng, 2021 | + | + | + | + | + | + | + | + | + |
| Brian Lane, 2022 | - | - | + | ? | + | + | - | - | - |
| Ke Wang, 2022 | + | ? | + | + | + | + | + | ? | + |
| Emilie Lalonde, 2014 | + | + | + | + | + | - | + | + | - |
| Dongjie Chen, 2021 | + | + | + | + | + | + | - | + | - |
| Yanhong Shou, 2021 | + | + | + | + | + | + | + | + | + |
| Xiangqian Zhang, 2023 | - | + | + | + | + | + | + | - | + |
| Chenyu Nie, 2022 | + | + | + | + | - | + | + | + | - |
| Jinman Zhong, 2021 | - | + | + | + | + | + | + | - | + |
| Xia Yang, 2021 | + | - | + | + | + | + | + | - | + |

**Table 2.** Risk of bias and applicability assessment

* 1. Synthesis

1. Review management
   1. Quality assurance (critical discussion of the methodology)

Findings

Discussion

Conclusion

Buffa, F. M., Harris, A. L., West, C. M., & Miller, C. J. (2010). Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene [Article]. *British Journal of Cancer*, *102*(2), 428-435. <https://doi.org/10.1038/sj.bjc.6605450>

Hong, S., Zhang, Y., Cao, M., Lin, A., Yang, Q., Zhang, J., Luo, P., & Guo, L. (2021). Hypoxic Characteristic Genes Predict Response to Immunotherapy for Urothelial Carcinoma [Original Research]. *Frontiers in Cell and Developmental Biology*, *9*. <https://www.frontiersin.org/articles/10.3389/fcell.2021.762478>

Lalonde, E., Ishkanian, A. S., Sykes, J., Fraser, M., Ross-Adams, H., Erho, N., Dunning, M. J., Halim, S., Lamb, A. D., Moon, N. C., Zafarana, G., Warren, A. Y., Meng, X., Thoms, J., Grzadkowski, M. R., Berlin, A., Have, C. L., Ramnarine, V. R., Yao, C. Q., . . . Bristow, R. G. (2014). Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *The Lancet Oncology*, *15*(13), 1521-1532. <https://doi.org/https://doi.org/10.1016/S1470-2045(14)71021-6>

Liu, Y., Wu, J., Huang, W., Weng, S., Wang, B., Chen, Y., & Wang, H. (2020). Development and validation of a hypoxia-immune-based microenvironment gene signature for risk stratification in gastric cancer. *Journal of Translational Medicine*, *18*(1), 201. <https://doi.org/10.1186/s12967-020-02366-0>

Marotta, D., Karar, J., Jenkins, W. T., Kumanova, M., Jenkins, K. W., Tobias, J. W., Baldwin, D., Hatzigeorgiou, A., Alexiou, P., Evans, S. M., Alarcon, R., Maity, A., Koch, C., & Koumenis, C. (2011). In vivo profiling of hypoxic gene expression in gliomas using the hypoxia marker EF5 and laser-capture microdissection [Article]. *Cancer Research*, *71*(3), 779-789. <https://doi.org/10.1158/0008-5472.CAN-10-3061>

Moons, K. G. M., de Groot, J. A. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLOS Medicine*, *11*(10), e1001744. <https://doi.org/10.1371/journal.pmed.1001744>

Seigneuric, R., Starmans, M. H. W., Fung, G., Krishnapuram, B., Nuyten, D. S. A., van Erk, A., Magagnin, M. G., Rouschop, K. M., Krishnan, S., Rao, R. B., Evelo, C. T. A., Begg, A. C., Wouters, B. G., & Lambin, P. (2007). Impact of supervised gene signatures of early hypoxia on patient survival [Article]. *Radiotherapy and Oncology*, *83*(3), 374-382. <https://doi.org/10.1016/j.radonc.2007.05.002>

Tawk, B., Schwager, C., Deffaa, O., Dyckhoff, G., Warta, R., Linge, A., Krause, M., Weichert, W., Baumann, M., Herold-Mende, C., Debus, J., & Abdollahi, A. (2016). Comparative analysis of transcriptomics based hypoxia signatures in head- and neck squamous cell carcinoma. *Radiotherapy and Oncology*, *118*(2), 350-358. <https://doi.org/https://doi.org/10.1016/j.radonc.2015.11.027>

Toustrup, K., Sørensen, B. S., Alsner, J., & Overgaard, J. (2012). Hypoxia Gene Expression Signatures as Prognostic and Predictive Markers in Head and Neck Radiotherapy. *Seminars in Radiation Oncology*, *22*(2), 119-127. <https://doi.org/https://doi.org/10.1016/j.semradonc.2011.12.006>

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*, *170*(1), 51-58. <https://doi.org/10.7326/M18-1376>