Attention-fusion Model for Multi-Omics (AMMO) Data Integration in Lung Adenocarcinoma

Wentao Li, Amgad Muneer, Muhammad Waqas, Xiaobo Zhou, Jia Wu

Abstract. The multi-omics integration gives a whole new perspective into pathway analysis to reveal the complicated nature of cellular systems. While the understanding of interactions among different omics data remains unknown, current methods do not consider the unique and similar properties. In this paper, we propose Attention-fusion Model for Multi-Omics (AMMO), a robust method that addresses this challenge through domain separation. Our proposed attention-based approach inherently captures the similarities and differences across various omics modalities, enhancing the interpretability of the integrated data. Our proposed method can achieve a state-of-the-art C-index of 0.8017 in overall survival prediction in TCGA-LUAD data with the diverse types of omics data: DNA Methylation, exon expression RNA Seq (HiC), and protein expression (RPPA). We also demonstrated the performance increase by adding more modalities with the ablation test, the results confirmed our assumption of improving model performance by including more modalities to our method.

Keywords: Multi-omics, Attention model, Single cell.

REFERENCE

Motivation

The recent advancements of biotechnology have enabled the collection of extensive research data derived from different biological processes. This includes data from genomes, transcriptomics, proteomics, and metabolomics [1,2]. However, focusing on one single modality of these molecular data may not adequately capture the underlying biological relationships among the different layers.

In cancer studies, multi-omics integration can provide several advantages over a single modality, where they can better measure intra-tumor heterogeneity [3], lead to robust biomarker discovery and validation [4], and uncover key regulatory mechanisms [5]. Different omics modalities often exhibit both common and unique patterns in their expression and effects on the human body [6]. These modalities' common or interacting patterns can help identify the biomarker and therapeutic targets. Therefore, combining both unique and similar patterns can offer a comprehensive understanding of cancer system biology [7–10].

Method

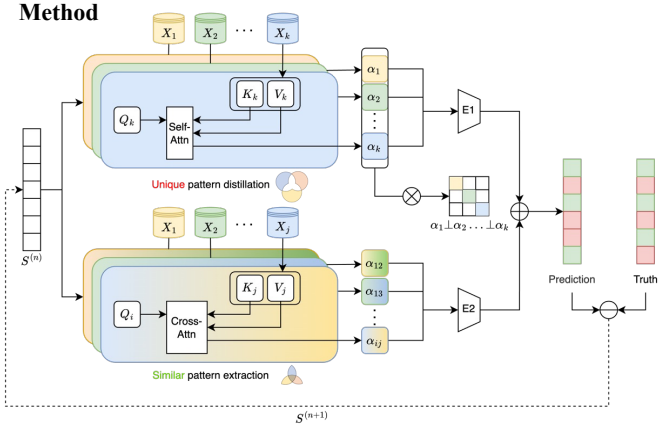

Figure 1. Flowchart of Attention-fusion Model for Multi-Omics (AMMO).

We proposed an end-to-end trainable attention-based mixed-fusion approach. The pro-posed approach extracts unique and common patterns from different omics modalities and fuses them to obtain improved analysis performance. The model will start with an initial shared similarity latency $S^{(0)}$. Then this latency will enter two modules at the same time, similar pattern extraction and unique pattern distillation.

Extracting the similarity

To calculate the cross-attention, we use the scaled dot-product attention mechanism. The cross-attention score α_{ij} is computed as follows:

$$\alpha_{ij}(S^{(n)}, X_i, X_j; \theta) = softmax\left(\frac{Q_j K_i}{\sqrt{d_k}}\right) V_i$$

The output will contain the similarity latency between omics modality i and j, and assembles as the updated similarity latency across the modalities with a linear layer $f(\cdot)$, denotes as $S^{(n)} = f(S^{(n)}, \boldsymbol{\alpha}; \theta)$, where $\boldsymbol{\alpha} = \{\alpha_{ij}\}_{i \neq j}$.

Preserving the uniqueness

Similar to the cross-attention score calculation, the multi-head self-attention score α_i is computed as follows:

$$\alpha_i(S^{(n)}, X_i; \theta) = softmax\left(\frac{Q_i K_i}{\sqrt{d_k}}\right) V_i$$

TCGA-LUAD Dataset

The sample sizes for the three omics modalities are as follows: 492 samples for DNA methylation, 576 samples for exon expression RNA-Seq, and 365 samples for protein expression (RPPA). The feature dimensions for the DNA methylation, RNA-Seq, and RPPA data are 485,577, 239,322, and 276, respectively.

Ablation test and benchmark results

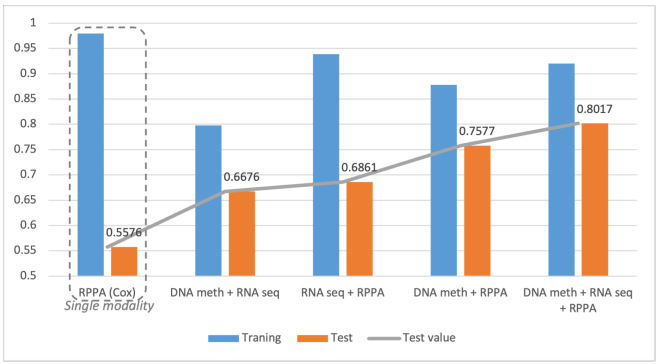

Figure 2. Comparative C-index values in ablation test.

Table 1. Comparative C-index values in benchmark.

Study	Year	Method	Reported best C-Indices on OS
Song et al., [11]	2019	Cox	0.723
Li et al., [12]	2019	LASSO+Cox	0.695
Zhang et al., [13]	2020	LASSO+Cox	**0.796**
Wen et al., [14]	2022	Cox	0.631
Peng et al., [15]	2023	Deep learning survival	0.740
Rączkowska et al., [16]	2022	DL+CNN plus H&E slides	0.723
Our study	2024	An attention-based model	**0.802**