Matching Cells Between Modalities with a Cross Attention Network

Xinyao Fan
xinyao.fan@stat.ubc.ca

Ning Shen
ning.shen@stat.ubc.ca

1. Introduction

Human body is composed of approximately 37 trillion cells [1] that are organized into tissues, organs, and systems, performing vital functions of life. As a human develops from a single fertilized egg cell, all of our cells contain the same DNA sequence. Yet, cells differ in behaviors, functions, and proliferation rates. Therefore, understanding cell-to-cell heterogeneity is the key to gaining mechanistic insight into how tissues function or malfunction in health and disease.

Regions of DNA called genes are transcribed into messenger RNA (mRNA) molecules and subsequently translated into proteins, which perform a vast amount of inter- and intra-cellular functions within living organisms. Thus, cellular heterogeneity can be characterized by the regulation of gene expression which is affected by various molecular mechanisms, such as DNA methylation and chromatin accessibility. All these different types of data existing in individual cells can provide massive information on cellular state hence are critical to biomedical research.

With recent advances in experimental techniques, it is now possible to measure several of these modalities in the same cell. These technologies have enabled the study of biology at an unprecedented scale and resolution. One of the research problem of interest is to match multi-modal profiles measured from the same cell when the correspondences are hidden. As most existing single-cell datasets measure a single modality, aligning modalities while preserving underlying biology is of importance for it enables leveraging complementary layers of information measured independently. This task is also formalized in the NeurIPS Competition from 2021, Multimodal Single-cell Data Integration Challenge (openproblems.bio/neurips_2021). The organizers of the competition have selflessly generated a processed, annotated, and formatted large-scale dataset of the human bone marrow specifically designed for benchmarking studies [3].

While the amount of available public multi-modal data is increasing, methodologies tailored to this specific task are still scarce. Previous literature contributing to this specific topic is very limited. The winning method in the competition builds upon variational autoencoders to learn joint cell embeddings with a cross-linked structure to account for different cell type resolution from modalities [2]. The second place method is a shallow MLP encoder applied on preprocessed latent vectors [2]. However, neither of the two methods takes into account the spatial correlations existing between features within and across modalities.

Motivated by the above discussions, we propose a multimodality CNN+attention network (Fig. 1) to jointly model inter-modality relationship and intra-modality relationship of gene expression and chromatin accessibility. For each modality, the CNN module outputs features of regional information, which serves as embeddings and are fed into attention modules. The incorporated self- and cross-attention modules dynamically highlight relevant features within and across modalities respectively. Analyzing and visualizing the learnt attention may help us to understand how regions in the chromosome interact and further shed light on the significance of different regions of DNA in cell identity and regulation of downstream genetic processes.

2. Data Preprocessing

Due to the limited time allowed for the project, we have decided to focus on one of the benchmark datasets provided by the challenge, 70,000 cells processed via the 10x Multiome assay. The assay captures chromatin accessibility (based on the Assay for Transposase-Accessible Chromatin, ATAC) and nucleus RNA gene expression (GEX) on a single-cell level. These two modalities are referred to as **ATAC** and **GEX** henthforth.

The dataset is presented as two cell-by-feature matrices, each representing a modality. ATAC data appear as binary values indicating if a DNA region is accessible, while GEX data appear as integer values indicating gene expression counts. As We apply a TF-IDF transformation to the ATAC data as a normalization step. For GEX data, we adopt common practice in genetics, that is, normalizing the counts per cell (dividing the UMI counts by the size factors) and a log1p transformation [4].

3. Proposed Approach

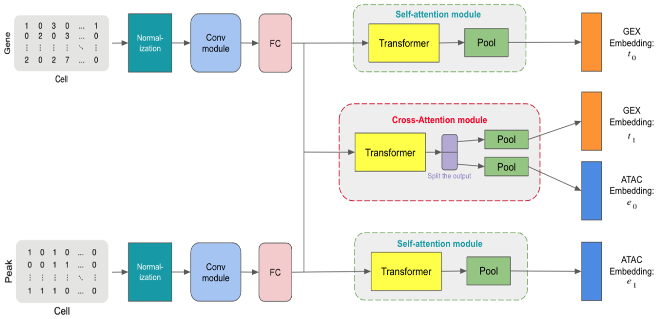

Figure 1. The proposed multi-modality cross attention network consisting of the self-attention module (shown in the green dashed blocks) and the cross-attention module (shown in the red dashed blocks).

3.1. Overview

The network mainly consists of three types of modules: 1-D convolution modules, self-attention modules and cross-attention modules. The architecture of the network is shown in Fig. 1. We first apply 1-D convolutional layers to each modalities to extract local features representing regional information. Then the extracted features are feed into self-attention modules for two modalities respectively to model the intra-modality relationships (section 3.2). Also, we use the Cross-Attention module to model the inter- and intra-relationships between modalities (section 3.3). For a given cell, the network produces an embedding for each modality concatenated from self- and cross-attention module outputs. We expect this network to have a better feature-discriminative ability compared to previous methods as it consider both intra- and inter-relationships between modalities.

The design of 1-D convolution modules is similar to [5]. Denote the number of cells as n. We apply $f^{(g)}$ filters on the 1-D GEX vector and obtain a 2-D output matrix

$$\mathbf{G} = [g_1; \ldots ; g_{d^{(g)}}],$$

where $g_i \in \mathcal{R}^{f^{(g)} \times 1}, i = 1, \ldots , d^{(g)}$. Similarly, $f^{(a)}$ filters are applied to ATAC data to output feature map in the matrix form

$$\mathbf{A} = [a_1; \ldots ; a_{d^{(a)}}],$$

where $a_j \in \mathcal{R}^{f^{(a)} \times 1}$, $j = 1, \ldots , d^{(a)}$.

The output of convolution modules $\mathbf{G}$ and $\mathbf{A}$ is followed by a fully connected (FC) layer to project them into the same dimension, denote as p. Denote the output of the FC layers as $\mathbf{G}' = [g'_1; \ldots ; g'_{d^{(g)}}]$ and $\mathbf{A}' = [a'_1; \ldots ; a'_{d^{(a)}}]$, where $g'_i, a'_j \in \mathcal{R}^{p \times 1}$ for $i = 1, \ldots , d^{(g)}$ and $j = 1, \ldots , d^{(a)}$.

3.2. Self-attention module

We introduce two self-attention modules to model the intra-modality relationships for GEX and ATAC respectively. In each module, a Transformer unit [6] is applied to implement the attention function (Fig. 2). Briefly, the unit consists of a multi-head self attention layer and feed forward layer, where each sub-layer has a residual connection and is followed by a layer-normalization step.

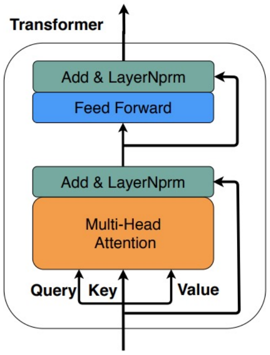

Figure 2. The Transformer unit used in the Attention modules. It consists of two layers, the multi-head self-attention sub-layer and feed-forward layer.

For a given cell, inputs of the self-attention units in Fig. 1 are $\mathbf{G}'$ and $\mathbf{A}'$. By visualize the $d^{(g)} \times d^{(g)}$ and $d^{(a)} \times d^{(a)}$ attention matrices, we may examine the region-region relevance for each modalities respectively. Average pooling layers are applied to the outputs of the two self-attention modules, producing embeddings e_0 for GEX and t_0 for ATAC.

3.3. Cross-attention module

We exploit the inter-modality relationships among cells using the cross-attention module proposed by [7]. A brief description is as follows.

Input: The input of this model is the stacked regions of two modalities $\mathbf{Y} = \begin{pmatrix} \mathbf{G}' \\ \mathbf{A}' \end{pmatrix} \in \mathcal{R}^{(d^{(g)}+d^{(a)}) \times p}$.

Module: The query, key and value matrices are denoted as $\mathbf{K}_Y = \mathbf{Y}\mathbf{W}^K = \begin{pmatrix} \mathbf{K}_G \\ \mathbf{K}_A \end{pmatrix}$, $\mathbf{Q}_Y = \mathbf{Y}\mathbf{W}^Q = \begin{pmatrix} \mathbf{Q}_G \\ \mathbf{Q}_A \end{pmatrix}$, and $\mathbf{V}_Y = \mathbf{Y}\mathbf{W}^V = \begin{pmatrix} \mathbf{V}_G \\ \mathbf{V}_A \end{pmatrix}$ respectively. The cross attention proposed by [7] is carried out as below (for simplicity, we omit the softmax and scaled function) :

$$\mathbf{Q}_Y \mathbf{K}_Y^T \mathbf{V}_Y = \begin{pmatrix} \mathbf{Q}_G \mathbf{K}_G^T \mathbf{V}_G + \mathbf{Q}_G \mathbf{K}_A^T \mathbf{V}_A \\ \mathbf{Q}_A \mathbf{K}_A^T \mathbf{V}_A + \mathbf{Q}_A \mathbf{K}_G^T \mathbf{V}_G \end{pmatrix} \tag{1}$$

Output: Denote the output features from Eq.(1) as

$$\mathbf{G}'' = [g_1''; \ldots; g_{d^{(g)}}''] = \mathbf{Q}_G\mathbf{K}_G^T\mathbf{V}_G + \mathbf{Q}_G\mathbf{K}_A^T\mathbf{V}_A \qquad (2)$$

$$\mathbf{A}'' = [a_1''; \ldots; a_{d^{(a)}}''] = \mathbf{Q}_A\mathbf{K}_A^T\mathbf{V}_A + \mathbf{Q}_A\mathbf{K}_G^T\mathbf{V}_G \qquad (3)$$

This result shows that the output of the multi-head sublayer in this Transformer module takes the intermodality and intra-modality relationships into consideration simultaneously. Similar to the Self-Attention module, we pass the $\mathbf{G}''$ and $\mathbf{A}''$ to average pooling layers and get another pair of embeddings, e_1 and t_1.

3.4. Loss function

For a given GEX-ATAC pair, the proposed network produces two pairs of embeddings, denote as (e_0, t_0) and (e_1, t_1) respectively. The embeddings are scaled to have a unit norm, and similarity scores are used for measuring the similarity between the two embeddings. Same with [7], for a given pair, the similarity scores are defined as the weighted summation of inner products between two pairs of embeddings as follows,

$$S(E,T) = e_0 \cdot t_0 + \alpha(e_1 \cdot t_1)$$

where α is a hyper-parameter that controls the effect of the self attention and cross-attention module in the network.

The loss function adopted in the model is a bi-directional triplet ranking loss.

$$\mathcal{L} = \max[0, m - S(E,T) + S(E,\hat{T})] \\ + \max[0, m - S(E,T) + S(\hat{E},T)]$$

where m denotes the margin, (E,T) denotes the true matched GEX-ATAC pair; $\hat{E}$, $\hat{T}$ denotes the hard negatives in a mini-batch. $\hat{E} = \arg\max_{x \neq E} S(x,T)$ and $\hat{T} = \arg\max_{y \neq A} S(E,y)$. This loss encourages the similarity scores of matched pairs to be larger than those of mismatched pairs.

4. Evaluation Metrics

We adopt the evaluation metrics used in the competition [3]. In the matching task, the jointly profiled cells are presented as two sets of unmatched singly profiled cells. The algorithmic goal is assign to each cell in modality GEX a probability distribution across all cells in modality ATAC in order to place high probability on the true matched cell. For n cells, we have a (n,n) matrix of non-negative values where each row sums to 1. Note the matrix is enforced to have at most 1000 non-zero values per row to accommodate the memory requirements. The first metric is the sum of the probabilities assigned to the correct matching (Fig. 4). We also consider the area under the precision recall curve (AUPR) as a side metric.

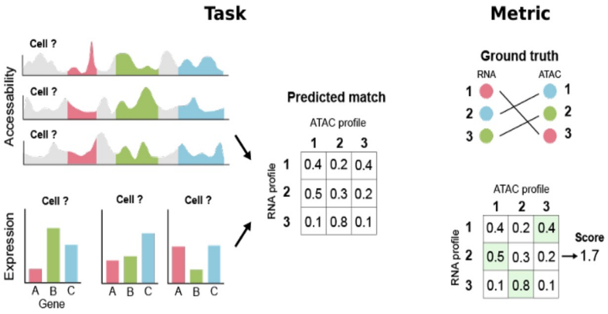

Figure 3. Conceptual figures of the three tasks and evaluation metrics used in the competition. The task was the matching of cells across modalities, evaluated by a match probability score. The plot is adopted from [2].

References

[1] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, Soledad Perez-Amodio, Pierluigi Strippoli, and Silvia Canaider. An estimation of the number of cells in the human body. *Annals of Human Biology*, 40(6):463–471, 2013. PMID: 23829164. 1

[2] Christopher Lance, Malte D Luecken, Daniel B Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Christine Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, et al. Multimodal single cell data integration challenge: results and lessons learned. *bioRxiv*, 2022. 1, 3

[3] Malte D Luecken, Daniel Bernard Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T Chen, Louise Deconinck, Angela M Detweiler, Alejandro A Granados, et al. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1, 3

[4] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. 1

[5] Milad Mostavi, Yu-Chiao Chiu, Yufei Huang, and Yidong Chen. Convolutional neural network models for cancer type prediction based on gene expression. *BMC medical genomics*, 13(5):1–13, 2020. 2

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[7] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 2, 3