CJSJ

A Deep Learning Model for Protein Abundance Prediction from RNA Data with Manifold-Preserving Regularization

Alice Feng, Shaoheng Liang, Ken Chen Ph.D.

***Abstract*—A key challenge in single-cell multiomics study is to quantify the relationship between mRNA level and protein abundance. This relationship is complicated by the dynamic nature of mRNA and protein. In this paper, a deep learning regression model was proposed to predict protein abundance from mRNA expression data. However, overfitting was identified as a major source of error. Because different modalities of the same sample should concord to the same cell population structure, we invented a manifold-preserving regularization term to reduce overfitting induced by noise specific to training. By applying our model on CITE-seq data for 25 cell-surface proteins representing well-characterized markers, we observed an improvement of up to 30% on testing error. Thus, manifold-preserving regularization helps distill true mRNA-protein relationships from noisy data. We expect it to be generally applicable to other multiomics applications.**

I. Introduction

Single-cell multiomics technologies provide co-assayed measurements from the same cells and can help learn relationships among different omics. Protein and RNA are important in understanding how cells work, and studying their relationship provides further insight into disease progression. A key challenge in single-cell multiomics study is to quantify the relationship between mRNA level and protein abundance [1,2], which is complicated by the dynamics in synthesis, splicing and degradation of mRNAs, and modification, folding and transportation of proteins. Furthermore, co-assayed data are limited, as they are usually more costly and compromise throughput and read depth. Hence, it is desirable to find a computational method that quantifies the relationship between proteins and RNA and predicts the protein profile with certain accuracy.

Overparameterized deep learning models are ideal for modeling complicated relationships, including that of mRNAs and proteins. However, overfitting occurs when a model fits to not only the true relationship of the predictors and responses, but also training data-specific noise (such as technical noise from sample preparation and sequencing). To reduce the sensitivity to noise, regularization is a common way to shrink the solution space using heuristics.

For single-cell data, manifold is an important characteristic that represents the cell population structure in a biospecimen. The manifold of a single-cell dataset is defined by the pairwise similarities of cells that characterize clusters, subtypes, and trajectories. It has been widely used in dimension reduction and trajectory inference [3,4,5]. Because different modalities of the same sample largely agree, we hypothesize that the majority of discrepancies between the manifolds of mRNA and protein data are technical noise. To mitigate overfitting, we introduce Manifold-Preserving Regularization (MPR), which suppresses the noise to mitigate overfitting (Fig. 1).

In this paper, a deep learning regression model was built to map the relationship between protein and gene expression levels. For a specific protein, although its coding gene can be a predictor, other genes also play important roles in the complicated regulations of its production and degradation. Thus, we use all genes as predictors. Results show that MPR increases the accuracy of models, compared to unregularized L2-regularized models. The genes with high weights in the resulting model are indeed in related biological pathways.

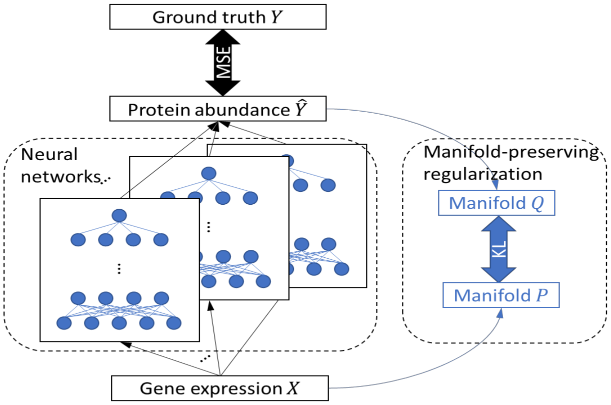

Figure 1. Architecture of Deep Neural Network with MPR

II. Methods

We used a CITE-seq dataset with co-assayed bone marrow mononuclear cells from two healthy donors [6]. We used 14,468 cells from Donor 1 as the training data, and, to test if the model can generalize, 16,204 cells from Donor 2 as the testing data. Measured in the datasets are 25 proteins and 17,009 genes. Twenty-seven cell types were identified by the original publication.

First, we preprocessed the data, including data cleaning and normalization. For the protein data, we normalized, log-transformed, and scaled the data. For the RNA data, we first normalized and performed log transform using the same process as with the protein data. Based on our analysis of the

* Alice is with the Harker School, San Jose, CA. Shaoheng Liang and Dr. Ken Chen are with Depart of Bioinformatics and Computational Biology at the University of Texas MD Anderson Cancer Center, Houston. TX.

data, we decided the most effective data cleaning method was removing genes with constant or similar values across cell samples. As a common practice in single-cell data analysis, we used a cutoff of dispersion at 0.5 and mean at 0.5, 1,555 highly variable genes are selected for Donor 1 (and later also used on Donor 2). Finally, we also scaled the RNA data.

Then the goal of our regression analysis is to identify a function of the RNA expression data **X** so that the function value is close to the protein abundance value **Y**, let the function be $f(\mathbf{X},\beta)$ where β is the function parameters, we need to optimize the function form and the parameters so that the prediction $\hat{\mathbf{Y}}=f(\mathbf{X},\beta)$ is as close to **Y** as possible.

We selected a 5-layer neural network with 1500 nodes per hidden layer and a batch size of 64 after experimenting with the number of layers, number of nodes in the first layer, and the batch size. Each protein has its own neural network. Mean squared error (MSE) is used as the loss function. We also measured the Overfitting Factor of each model defined as $1 - \sqrt{\text{Training MSE} \mid \text{Testing MSE}}$.

Based on our hypothesis that matching the manifold of RNA and protein will mitigate overfitting, we used manifold information as a regularization in fitting the model (Fig.1). The manifolds for **X** and **Y** are denoted **P** and **Q**, defined as

$$p_{ij} = \frac{\exp\left(-\left(\left\|x_i - x_j\right\| - \rho_i\right)/\sigma_i\right)}{\sum_{k\neq l}\exp(-(\|x_k - x_l\| - \rho_i)/\sigma_i)}$$

and

$$q_{ij} = \frac{\left(1 + \left(\left\|\hat{y}_i - \hat{y}_j\right\| - \tau_i\right)/\sigma_i\right)^{-1}}{\sum_{k\neq l}(1 + (\|\widehat{y_k} - \hat{y}_l\| - \tau_i)/\sigma_i)^{-1}}$$

where $\rho_i = m \in \left\|x_i - x_j\right\|, \tau_i = m \in \left\|\hat{y}_i - \hat{y}_j\right\|$, and σ_i is selected to ensure

$$\sum_j \exp\left(-\left(\left\|x_i - x_j\right\| - \rho_i\right)/\sigma_i\right) = k$$

The regularization term is defined as the Kullback-Liebler (KL) divergence of *P* and *Q* [7], also called relative entropy, measuring the difference between two distributions.

III. Results and Discussion

First, we experimented with deep learning architectures as shown in Tab. 1. Drop-out layers are added to reduce overfitting. Generally, the results are not sensitive to the network architectures. Architectures trained with larger batch sizes tend to have smaller Overfitting Factors. The model with the lowest MSE was model #6: 1500 nodes in the first layer and trained with batch size 64, which is used throughout the rest of this paper.

TABLE I. MSE Comparison Among Different Architectures

Model #	Nodes	Batch Size	MSE	Overfitting Factor
1	500	16	0.372	0.415
2	500	64	0.356	0.279
3	1000	16	0.362	0.472
4	1000	64	0.358	0.314
5	1500	16	0.365	0.505
6	1500	64	0.356	0.344

We trained and tested deep learning models without any regularization on each of the 25 proteins individually. The Overfitting Factor is nearly 1 for all proteins, suggesting that a near-perfect fit was achieved on training data, but does not generalize to testing data. Overfitting is indeed a major problem which necessitates regularization.

We then experimented with L2-regularization and MPR. Both improved the Overfitting Factor significantly. MPR reduced the Overfitting Factor by 25% on average and up to 50% for certain proteins, and outperformed L2 for every protein (Fig 2).

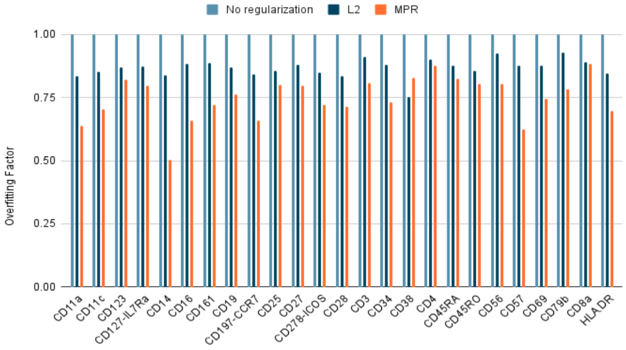

Figure 2. Overfitting Factor Comparison among No Regularization, L2, MPR

For MSE, MPR consistently reduced the model's testing MSE by 15% on average and up to 30% for certain proteins (Fig. 3). MPR performed preferably for 16 of the 25 proteins compared to L2-regularization, especially for proteins harder to predict from gene expression data, including CD123, CD25, CD38, CD4, CD56, CD79B, and CD8A. When compared with state-of-the-art model cTP-net[2], MRP outperformed cTP-net by up to 15%.

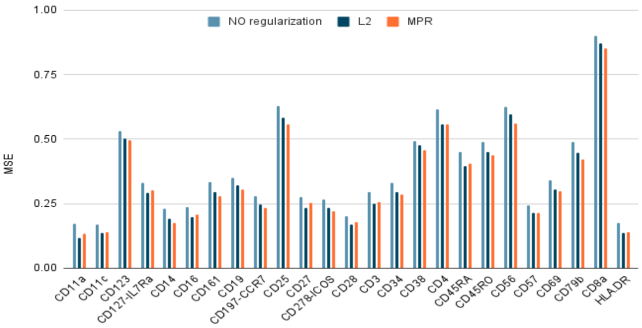

Figure 3. MSE Comparison among No Regularization, L2, MPR

Lastly, we performed gene set enrichment analysis on the top ten predictors for each protein. For one example, MPR identified CD19, a B-cell marker to be significantly related to B cell activation.

IV. Conclusion

In this study, a deep learning model was built to accurately predict protein abundance from RNA expression levels and quality the relationship between mRNA and protein. MPR was able to effectively mitigate overfitting in models characterizing relationships of modalities in multiomic data. It validates our hypothesis that the majority of discrepancies between the manifolds are noise and do not generalize across samples. MPR helps to obtain a more robust gene list that is closely related to biological processes. We expect it to be useful in other scenarios including predicting mRNA levels from assay for ATAC-seq chromatin accessibility profiles. Utility may also expand to non-single-cell data, such as bulk RNA sequencing and phenotypical characterization in The Cancer Genome Atlas (TCGA).

V. References

[1] Wu KE, Yost KE, Chang HY, Zou J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. Proc Natl Acad Sci. 2021;118.

[2] Zhou Z, Ye C, Wang J, Zhang NR. Surface protein imputation from single cell transcriptomes by deep neural networks. Nat Commun. 2020;11:651.

[3] Hao Y, Hao S, Andersen-Nissen E, A, et al. Integrated analysis of mutimodal single-cell data. Cell. 2021;184:3573-3587.e29.

[4] Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–605.

[5] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat. 2020. http://arxiv.org/abs/1802.03426.

[6] Cao J, Spielmann M, Qiu X, Huang X, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566:496–502.

[7] Liang S, Mohanty, et al. Single-cell manifold-preserving feature selction for detecting rare cell populations. Nat Comput Sci. 2021;1:374–84.