Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

Study of CITE-seq protein expression prediction method based on LSTM

Wenrui Zhao *

University of Michigan, Ann Arbor, Michigan, US

* Corresponding Author: zwenrui@umich.edu

Abstract. CITE-seq consists of mRNA sequencing and labeled antibody sequencing. This paper predicts labeled antibody sequences based on LSTM neural network. Research shows that LSTM algorithm has a reliable application in CITE-seq protein expression prediction technology, and the prediction accuracy rate reaches 95%. The LSTM algorithm proposed in this paper is relatively accurate and can simulate the gene sequence of proteins well. The CITE-seq data training model learns the potential relationship between RNA and protein, realizes the prediction of protein expression by using scRNA-seq data, greatly reduces the cost of CITE-seq test experiment, and improves the experimental efficiency.

Keywords: CITE-seq; mRNA sequencing; labeled antibody sequencing; LSTM algorithm; neural network.

1. Introduction

At present, multimodal single-cell data are increasingly available, but data analysis methods are still scarce. Due to the small size of a single unit, measurements are sparse and noisy. Differences in the depth of molecular sampling between cells (sequencing depth) and the technical effects of batching cells (batch effects) often override biological differences. When analyzing multimodal data, different feature Spaces must be considered, as well as shared and unique variations between modes and between batches. In addition, current pipelines for single-cell data analysis treat cells as static snapshots, even when potentially dynamic biological processes are present. Considering temporal dynamics and how states change over time is an open challenge in single-cell data science.

In general, genetic information goes from DNA to RNA to proteins. DNA must be accessible (ATAC data) to produce RNA (GEX data), which in turn is used as a template to produce proteins (ADT data). These processes are regulated by feedback: for example, a protein may bind DNA to prevent the production of more RNA. This genetic regulation is the basis for dynamic cellular processes that enable organisms to develop and adapt to changing environments. In single-cell data science, dynamic processes have been modeled by so-called pseudo-time algorithms that capture the progress of biological processes. However, the generalization of these algorithms to consider both pseudo-time and real-time remains an open question, as shown in Figure 1.

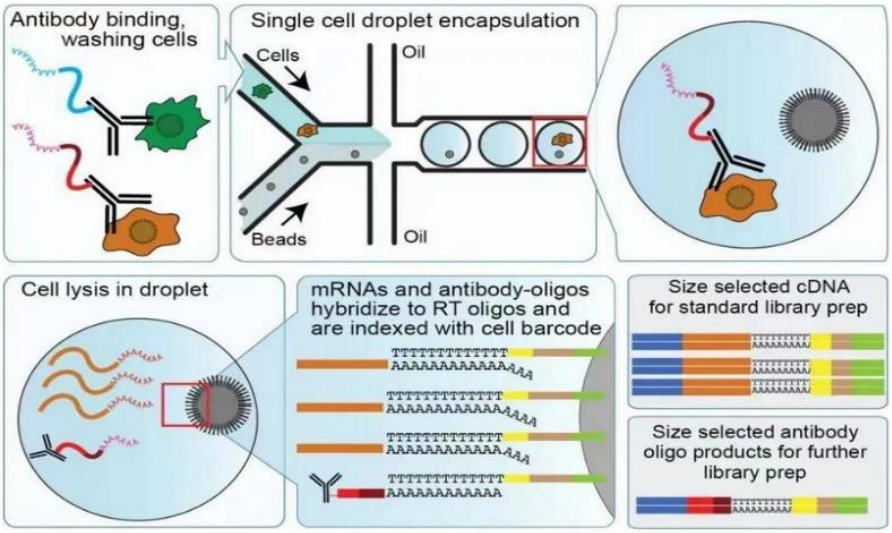

Figure 1. Size selected cDNA and antibody oligo products

CITE-seq is a single-cell multi-omics technique that allows simultaneous analysis of RNA and protein expression in single cells and has been widely used in biomedical research. However, CITE-seq data generation costs are too high. The researchers developed TotalVI and Seurat 4 to learn RNA-protein relationships, borrow information from large reference datasets, and predict protein expression directly using RNA-SEQ data. However, when studying complex problems, it is necessary to integrate multiple CITE-seq data sets with incomplete overlap of proteins, which is easy to produce batch effect and affect the prediction ability of the model. Therefore, the authors developed sciPENN to provide greater computational efficiency, model robustness, and prediction accuracy.

Therefore, this paper proposes a neural network LSTM based on time series, considering time and real-time, to accurately predict and grasp the genetic regulatory process, and make up for the insufficient consideration of time in the analysis of presequencing multi-source data. Understanding how a single genome generates the diversity of cell states is key to gaining insights into the mechanisms of how tissues function or malfunction in health and disease, helping to address questions in single-cell biology that over time, predict how gene regulation may affect the differentiation of blood and immune cells as they mature.

2. Methodology

For Multiome samples: LSTM-based models predict gene expression given chromatin accessibility. For CITEseq samples: LSTM-based model, given gene expression, predicts protein levels. LSTM neural network model was proposed by Sepp Hochreiter and Bergen Schmidhuber in the mid-1990s. The memory module was introduced through the input gate, the forget gate and the output gate, so that the new model could learn the long-term dependence in the time series data. In this way, the problems of gradient disappearance and gradient explosion of common RNN methods are solved, and the computational efficiency of the model is effectively improved. The structure diagram of the LSTM neural network model is output through input gates, forgetting gates, output gates, and representative memory units and candidate memory units respectively, which are adapted to time series data, as shown in Figure 2.

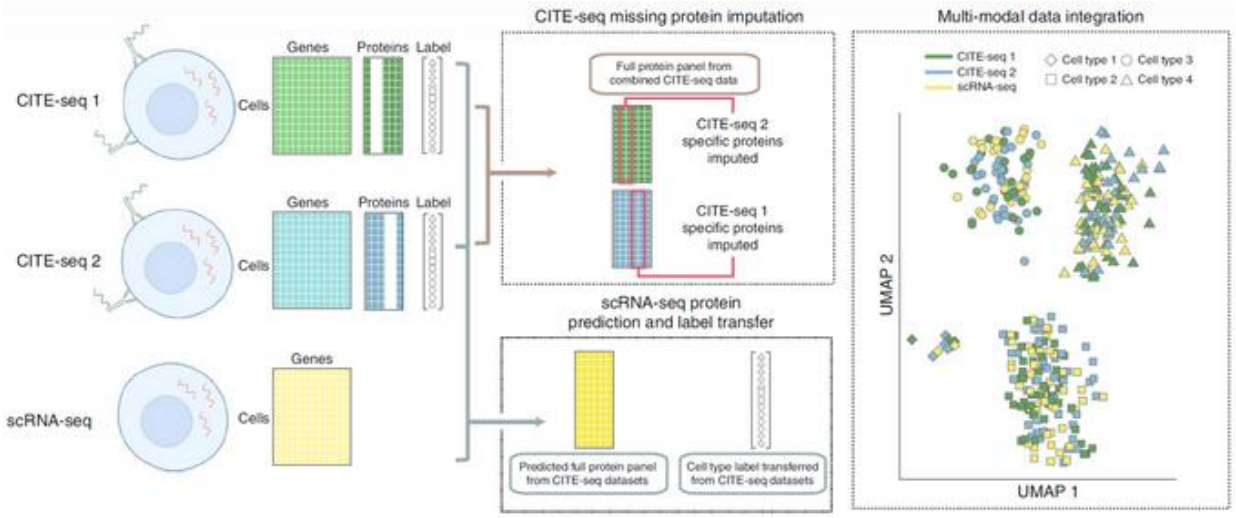

Figure 2. Typical LSTM series

Supports transcriptome and epitope cell index (CITE-seq) with single-mode single-cell RNA sequencing (scRNA-seq) data integration, protein expression prediction (scRNA-seq). LSTM neural network regulates the degree of memory unit forgetting and the degree of candidate memory unit addition through forgetting gate ft and input gate it, respectively. ft and it are defined as follows:

$$f_t = sigmoid(W_f x_t + U_f h_{t-1} + b_f) \tag{1}$$

$$i_t = sigmoid(W_i x_t + U_i h_{t-1} + b_i) \tag{2}$$

Where ht-1 represents the hidden layer state at the time of T-1, when t=1, ht-1=0; And respectively represent the input weight matrix, the last time active value weight matrix and the bias matrix in the forgetting gate. And respectively represent the input weight matrix in the input gate, the last time active value weight matrix and the bias matrix. The LSTM neural network updates the memory unit by forgetting part of the existing memory and adding candidate memory. The calculation formula of the memory unit and candidate memory unit at T-moment is as follows.

$$C_t = i_t \otimes C_t' + f_t \otimes C_{t-1} \tag{3}$$

$$C_t' = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{4}$$

Where the tensor product is represented; tanh represents the hyperbolic tangent function; And respectively represent the input weight matrix, the last time hidden layer weight matrix and the bias matrix in the candidate memory unit. The output gate ot at time t is defined as follows:

$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$

In the formula, neutralization represents the input weight matrix, the hidden layer weight matrix and the bias matrix in the output gate respectively.

$$h_t = o_t \otimes \tanh(C_t) \tag{6}$$

After ht is calculated according to formulas (1)~(6), the output yt of LSTM neural network can be calculated by the following formula.

$$y_t = sigmoid(W_y h_t) \tag{7}$$

3. Experiment

This paper uses python language to complete code writing and application of experimental data. Table 1 gives specific experimental application information.

Table 1. Experimental configuration

Name	Configure
System	64 bit, Windows10 pro
CPU handle	Intel Core i5-10210U
RAM	8GB
EDITOR	PyCharm2020.2.2
language	Python3.6

To illustrate the performance of prediction models based on deep learning, the two most widely used metrics for continuous variables are selected: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

3.1. Accuracy

Accuracy is the ratio of the correctly identified value to the value of the overall experiment.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

The formula is explained as follows:

True Positive (TP): Predicts the number of positive classes to be positive.

True Negative (TN): predicts the number of negative classes.

False Positive (FP): indicates that the number of negative classes is predicted to be positive, and Type I error is generated.

False Negative (FN): indicates that the number of positive classes is predicted to be negative, and Type II error is missed.

3.2. MAE

MAE represents the average absolute error between the predicted value and the observed value. MAE is a commonly used regression loss function that represents the average error margin of the predicted value without considering the direction of the error. The average absolute error is calculated as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m} | \hat{y}_i - y_i | \tag{9}$$

Where, and correspond respectively to the predicted and true values in the experiment.

3.3. RMSE

The root mean square error is calculated as follows:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2} \tag{10}$$

The MAE value represents the average error between the predicted result and the actual value, and all individual differences are equally weighted during MAE calculation. During RMSE calculation, the individual differences between the predicted results and the corresponding observed values are squared, and then the average value is taken on the sample, and the square root of the average value is used as the final result. RMSE gives relatively high weights to large errors, as shown in Figure 3.

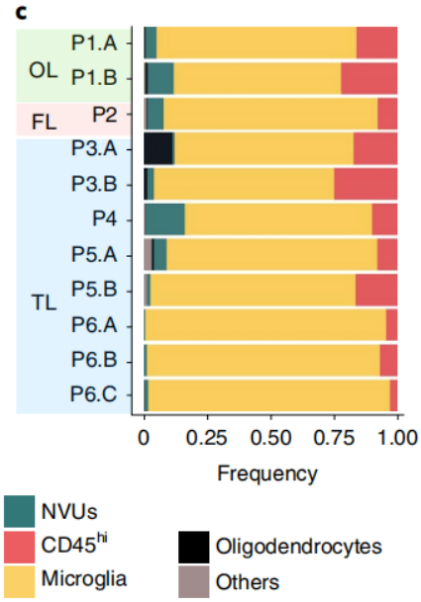

Figure 3. Weights of 11 variables contributing to 5-dimension principal components

Table 2. The 5 sets of test results

variable	1 (73.15%)	2 (12.76%)	3 (3.93%)	4 (3.22%)	5 (2.78%)
Series1	0.2249	0.5896	-0.0722	-0.2413	-0.1534
Series2	-0.3416	-0.2445	0.0270	0.1697	0.0356
Series3	0.0067	0.1109	0.9928	0.0335	-0.0098
Series4	-0.3639	0.1277	-0.0158	-0.0399	-0.0565
Series5	0.1421	-0.3479	0.0717	-0.8840	-0.0836
Series6	-0.2928	-0.4471	0.0367	-0.0338	-0.3525
Series7	-0.3580	0.1560	-0.0177	-0.2294	0.2183
Series8	-0.3287	0.1930	0.0019	-0.2222	0.6481
Series9	-0.3793	-0.0375	0.0054	-0.0566	0.1020
Series10	-0.2896	0.4184	-0.0362	-0.1488	-0.5049
Series11	-0.3611	0.0710	0.0016	0.0183	-0.3281

Currently, Biotek can perform 10x Genomics based single-cell transcriptome experiments and analyses on a variety of tissue types, including tissue samples, blood samples, and cell samples. A good sample preparation process and transportation process are crucial to the success of this experiment, so we have summarized some sample preparation process and transportation precautions, hoping to help local (Beijing, Guangzhou and Shanghai) and overseas customers with their research projects, as shown in Table 2.

Sample preparation of fresh samples such as tissue, blood and cells. Fresh tissue: Take at least 250mm3 volume of the sample, use 1640 medium, PBS buffer or normal saline to clean 2-3 times (customers can also choose the appropriate medium to clean according to their own organization; Tissues that are not easily contaminated with blood, such as tumors, can also not be cleaned). After the initial processing of the sample is completed within 30 minutes, the sample is placed in 5 mL of tissue preservation solution (Miltenyi Biotec), and the tissue needs to be completely immersed in the tissue preservation solution. The preservation solution should be pre-cooled at 4℃ in advance.

Fresh blood: Take a sample volume of at least 3mL and quickly place it in an EDTA anticoagulant tube for storage at 4℃. Digest cell suspension and cell line: take at least 1×106 cells (calculated according to cell concentration), and place them in PBS buffer containing 10% serum pre-cooled at 4℃ in advance, and store them at 4℃

4. Result and Discussion

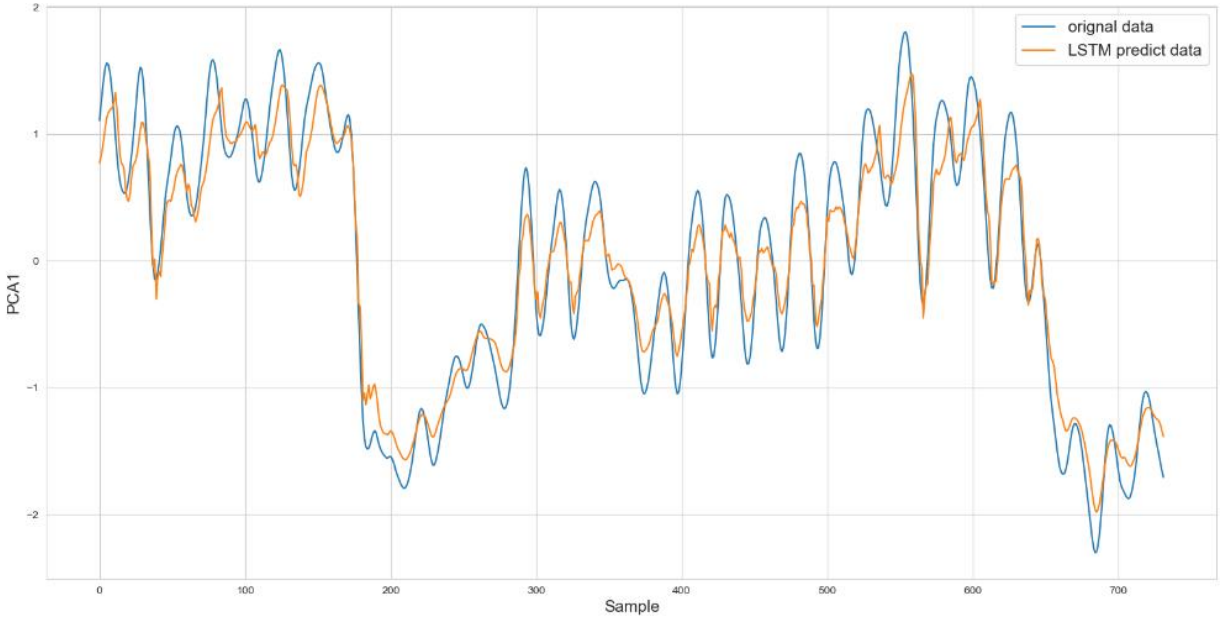

Figure 4. The fitting results

From the above Figure 4 fitting results, it can be seen that the LSTM algorithm proposed in this paper is relatively accurate and can simulate the gene sequence of proteins well.

5. Conclusions

The CITE-seq data training model learns the potential relationship between RNA and protein, realizes the prediction of protein expression by using scRNA-seq data, greatly reduces the cost of CITE-seq test experiment, and improves the experimental efficiency. More interestingly, the idea of CITE-seq low-dimensional embedding achieves better mixing of different data in embedding, which effectively improves the accuracy of subsequent sciPENN prediction of protein expression.

References

[1] Musu Y, Liang C, Deng M. Clustering single cell CITE-seq data with a canonical correlation based deep learning method [J]. Cold Spring Harbor Laboratory, 2021. DOI: 10.1101/2021. 09. 07. 459236.

[2] A F X, A S W, A X D, et al. Ensemble learning models that predict surface protein abundance from single-cell multimodal omics data – ScienceDirect [J]. Methods, 2020. DOI: 10.1016/j. ymeth. 2020. 10. 001.

[3] Liu X, Gosline S J, Pflieger L T, et al. Knowledge-based classification of fine-grained immune cell types in single-cell RNA-Seq data with ImmClassifier [J]. Briefings in Bioinformatics, 2021. DOI: 10. 1101/2020. 03. 23. 002758.

[4] Meng X, Yao J. Impact of classification difficulty on the weight matrices spectra in Deep Learning and application to early-stopping [J]. 2021. DOI: 10. 48550/arXiv. 2111. 13331.