Supplementary file of "Spatial domain identification based on variational autoencoder and single sample network"

Wei-Feng Guo
School of Electrical and Information Engineering
Zhengzhou University
Zhengzhou 450001, China
State Key Laboratory of Intelligent
Agricultural Power Equipment
Zhengzhou University
Luoyang 471100, China
guowf@zzu.edu

Yuan-Chao Wang
School of Electrical and Information Engineering
Zhengzhou University
Zhengzhou 450001, China
State Key Laboratory of Intelligent
Agricultural Power Equipment
Zhengzhou University
Luoyang 471100, China
15056097365@163.com

Song-Yan Feng
School of Electrical and Information Engineering
Zhengzhou University
Zhengzhou 450001, China
State Key Laboratory of Intelligent
Agricultural Power Equipment
Zhengzhou University
Luoyang 471100, China
fsydecomputer@outlook.com

Xin-Bo Hu
School of Electrical and Information Engineering
Zhengzhou University
Zhengzhou 450001, China
State Key Laboratory of Intelligent
Agricultural Power Equipment
Zhengzhou University
Luoyang 471100, China
1250763278@qq.com

I. Cell specific molecular action network method CSN

CSN [1] methods based on a large number of sample gene expression data, a sample specific g ene co-expression network was deduced. It provides a method to analyze gene associations at a single-cell level that can find differential gene associations just similar as differential genes.We use this method to identify the dependency and independency of two genes in a single cell, and then finds t he c hanges o f gene associations among different cell types.The algorithm is shown in the structure diagram below,Fig.S1.

For sample k, firstly, C SN d raws t wo r ectangular regions in the space near the expression value of gene x (x_k) and the expression value of gene y (y_k), and then generates the intersection of the two rectangular regions. CSN obtained the correlation between gene x and gene y by calculating the number of data samples in the two rectangular regions and the intersection(ρ_{xy}^k).

$$\rho_{xy}^k = \frac{n_{xy}^k}{n} - \frac{n_x^k}{n}\frac{n_y^k}{n} \tag{1}$$

This paper was supported by the National Natural Science Foundation of China (62002329), and Key Scientific and Technological projects of Henan Province (212102310083), and China postdoctoral foundation (2021M692915). (Corresponding author: Weifeng Guo)

n_x^k——The number of samples in the rectangle surrounding the expression value of gene x (x_k);

n_y^k——The number of samples in the rectangle surrounding the expression value of gene y (x_y);

n_{xy}^k——The number of samples in the intersection of two rectangular regions;

n——Total number of samples.

The range of the statistic is -1 to 1, and it can be proved that if x and y are independent of each other, the statistic ρ_{xy}^k approximately follows normal distribution and the mean value and standard deviation are:

$$\varphi = 0 \tag{2}$$

$$\sqrt{\frac{n_x^k n_y^k (n - n_x^k)(n - ny^k)}{n^4(n-1)}} \tag{3}$$

n_x^k and n_y^k are predetermined integers,n_x^k=n_{xy}^k=0.1n, and thus the statistic ρ_{xy}^k is only changed with n_{xy}^k.

In particular, we set n_x^k = n_y^k = 0.1n in this work, and n_x^k and n_y^k are both proportional to the sample size n. In other words, we first draw the two boxes near x_k and y_k based on the predetermined n_x^k and n_y^k, and then we can straightforwardly have the third box, which is simply the intersection of the

previous two boxes. Thus, we can obtain the value of n_{xy}^{k} by counting the plots in the third box, thereby testing the criterion of Equation (2).

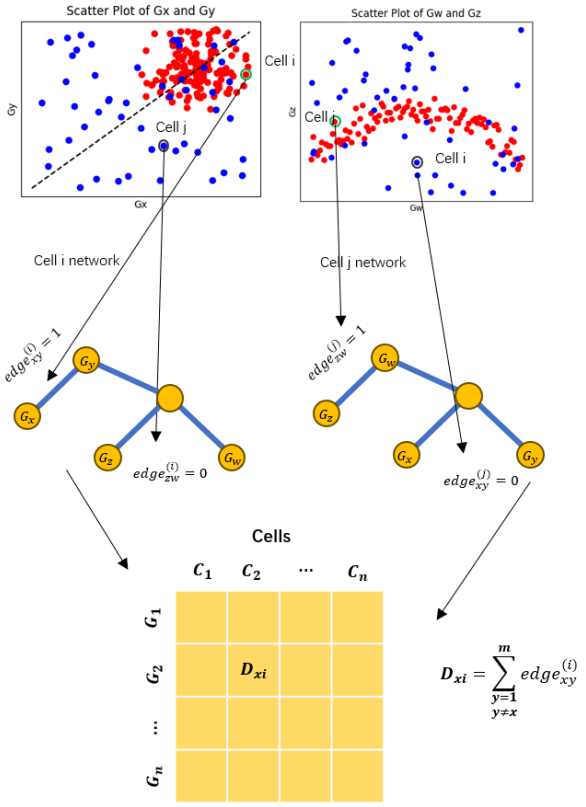

Fig. S1. CSN structure chart.

II. Variational Autoencoder VAE

stMVC [2] originally used autoencoders. In order to improve the ability of finding h eterogeneity a nd p rogression of complex diseases, this study adopted variational autoencoders to greatly improve the arithmetic accuracy of the algorithm.AE is mainly used to compress and restore data, and map the data map directly to the numerical code. VAE is mainly used to generate data, map the data to distribution, and then sample the code from the distribution.The encoder structure is shown in the frame diagram below,Fig.S2.

Assume that the gene expression data (X) follows a standard normal distribution(SN)($X - - - SN(\mu, \theta^2)$). In order to use the low-dimensional representation of the gene as a cell-specific matrix, the variational autoencoder is used to compute the 50-dimensional characteristics of the 3000 high-variant genes screened above.

The specific m ethod i s a s f ollows: t he i nput d ata X is mapped into a new distribution by the encoder (E),At the same time, two parameters of the new distribution ($\mu_x^{'}$) and ($\theta_x^{'}$) are obtained, and then the low-dimensional feature z of X is obtained, and finally t he t wo p arameters (μ_x) a nd (θ_x) of SN are simultaneously converted by the decoder (D_μ and D_θ). The training goal of the model is to maximize the edge likelihood function of the observed gene expression data. The formula can be described as:

$$p_1(x|l_3) = SN(x; \mu_x^{'}; \theta_x^{'})$$

$$SN(x; \mu_x^{'}; \theta_x^{'}) = \frac{1}{\sqrt{2\pi}\theta_x^{'}} e^{-\frac{(x-\mu_x^{'})^2}{2\theta_x^{'2}}} \tag{4}$$

$$p_2(x|l_3) = SN(x; \mu_x; \theta_x)$$

$$SN(x; \mu_x; \theta_x) = \frac{1}{\sqrt{2\pi}\theta_x} e^{-\frac{(x-\mu_x)^2}{2\theta_x^2}} \tag{5}$$

$$\log p_2(x|l_3) = E_{l_3 \sim p(l_3|x,E)(\log p(x|l_3; D_\mu, D_\theta))} \tag{6}$$

Where, for each gene (x),μ_x and θ_x each dimension represents the Mean and Dispersion of its NB distribution,the one-dimensional constant l_x is the sum of the Read counts of all its selected genes and is used to calculate the normalization factor of the cell. In this study, each neural network adopts batch normalization, using "relu" as the activation function between the two hidden layers. For the gene expression data of each SRT, after model convergence, the trained model E is used to calculate z, and z is taken as the feature space of cells.

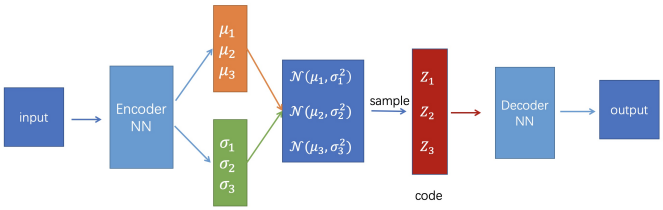

Fig. S2. VAE flow chart.

III. One Hot Coding

One hot coding is a method of converting categorical variables into numerical variables, commonly used in machine learning and deep learning.We used One hot coding, the value of discrete feature is extended to Euclidean space, and the value of discrete feature corresponds to a point in Euclidean space. Using one hot encoding for discrete features makes the distance between features more reasonable.The coding structure is shown in the frame diagram below,Fig.S3.

In this study, CSN method was used to construct a relationship network between cell genes in each cell i, and then the top 50 genes were screened out according to the connectivity degree. Then one hot coding method was used to encode N different regions. Suppose there were 3 regions in total. Then the encoding features of region 1, region 2 and region 3 are [1, 0, 0], [0, 1, 0] and [0, 0, 1] respectively. And so on, cell specific a ction n etworks [3] w ere c onstructed f or N different regions.

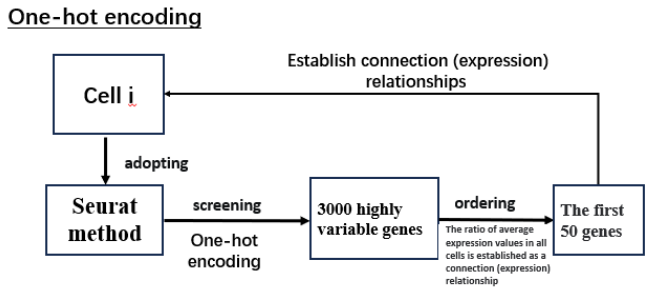

Fig. S3. One-hot encoding flow chart.

IV. EVALUATION INDEX—ARI INDEX [4]

The ARI Index [4] is an indicator used to evaluate the effectiveness of clustering algorithms, and it is a modified version of the Rand Index. Rand Index [5] cannot guarantee that the RI value of randomly partitioned clustering results is close to 0, but ARI is better suited than RI for evaluating the quality of clustering results because it automatically takes into account the randomness of the data.So, the Adjusted Rand index was proposed:

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \tag{7}$$

The E[RI] represents the expected value and the max(RI) represents the max value. Value range [-1,1], where -1 indicates that the clustering results are completely inconsistent, 0 indicates that the clustering results are equivalent to random distribution, and 1 indicates that the clustering results are completely consistent.The larger the value, the better the clustering effect.

ARI index [4] calculation needs to use the label obtained by clustering algorithm and the real label, and can use the adjustedRandIndex function in R language to calculate ARI index. It is worth noting when calculating the ARI index that in order to properly evaluate the effect of the clustering algorithm, the data set needs to be divided into a training set and a test set, and the test set is used for evaluation. Otherwise, if the ARI index [4] is calculated on all the data, the performance of the clustering algorithm may be overestimated.

$$AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \tag{8}$$

REFERENCES

[1] Hao Dai Lin Li Tao Zeng Luonan Chen. Cell-specific network constructed by single-cell rna sequencing data. *Nucleic Acids Research*, 47(11):e62, 2019.

[2] Mingqi Jiao Luonan Chen Chunman Zuo Yijian Zhang Chen Cao Jinwang Feng. Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning. *Nature Communications*, 13(5962):e2–e4, 2022.

[3] Riasat Azim and Shulin Wang. Cell-specific gene association network construction from single-cell rna sequence. *Cell Cycle*, 20(21):e2248–e2263, 2021.

[4] Hau-San WongYing Shen Shaohong Zhang. Generalized adjusted rand indices for cluster ensembles. *Pattern Recognition*, 45(6):2214–2226, 2012.

[5] Matthijs J. Warrens Hanneke van der Hoef. Understanding the adjusted rand index and other partition comparison indices based on counting object pairs. *Journal of Classification*, 39:487–509, 2022.