Surge EndTerm Report

July 30, 2022

INDIAN INSTITUE OF TECHNOLOGY
kanpur

"Evaluation of the accuracy of DestVI, CARD and AutoGeneS on cell-type Deconvolution"

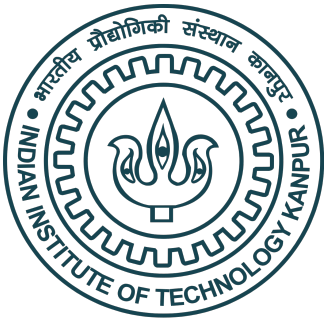
Submitted by

Milan Anand Raj

SURGE No.: 2230226
Department of Biological Sciences and Bioengineering
IIT Kanpur, U.P

Under the Guidance of

Dr. Hamim Zafar
Department of Computer Science and Engineering
IIT Kanpur (U.P)



Contents

Keywords: *Negative Binomial Distribution, Deconvolution and Neural Framework*

Acknowledgements

First and foremost, I am grateful to my college, Indian Institute of Technology Kanpur, and my Project Supervisor and Mentor Professor Hamim Zafar for giving me the opportunity to pursue a research internship under the prestigious SURGE program 2022 and. I am deeply indebted to my Professor for helping me to explore my interests and encouraging me to work on new domains. Professor Hamim Zafar has opened new doors of opportunity for me and helped me explore the domain of Computational Biology. I am thankful to him for guiding and mentoring by project and for providing assistance whenever I was facing difficulties in some way.

I am thankful to Ajita Shree(PhD student) for helping me to understand the concepts and theoretical aspects related to the project. She has constantly guided and supported me in the face of difficulties.

I would also like to express my gratitude for the Department of Computer Science and Engineering, IIT Kanpur for providing me an environment, that helped me develop an interest in research and further studies.

Lastly, I would like to thank my parents and my friends for having faith in me, constantly supporting me, for helping me through the difficult days and for always inspiring me to work harder.

Milan Anand Raj
200584
Third Year Undergraduate
Department of Biological Sciences and Bioengineering
IIT Kanpur

CERTIFICATE

This is to certify that the project entitled "Evaluation of the accuracy of DestVI, CARD and AutoGeneS on cell-type Deconvolution" submitted by Milan Anand Raj (S2230226) as a part of Summer Undergraduate Research and Graduate Excellence 2022 offered by the Indian Institute of Technology, Kanpur, is a Bonafide record of the work done by him under my guidance and supervision at the Indian Institute of Technology, Kanpur from 13th May, 2022 to 14th July, 2022.

Professor Hamim Zafar
Assistant Professor
Department of Computer Science and Engineering
IIT Kanpur

Keywords: *Negative Binomial Distribution, Deconvolution and Neural Framework*

Abstract

Spatial Transcriptomics is an overarching term for a range of methods designed for assigning cell types(identified by mRNA readouts) to their locations in the histological sections. This method can also be used to determine subcellular localisation of mRNA molecules.

The Stahl method implies positioning individual tissue samples on the arrays of spatially barcoded reverse transcription primers able to capture mRNA with oligo(dT) tails. Besides oligo(dT) tails and spatial barcode, which indicates the x and y position on the arrayed slide, the probe contains a cleavage site, amplification and sequencing handle, and unique molecular identifier.

In the broader meaning of this term, spatial transcriptomics include methods that can be divided into five principal approaches to resolving spatial distribution of transcripts. They are microdissection techniques, Fluorescent techniques in situ hybridisation methods, in situ sequencing, in situ capture protocols and in silico approaches.

What we do at our labs is in silico analysis. We are provided with the Single-cell data and the spatial transcriptomic data. Single cell data contain the information regarding the distribution of mRNA counts in specific cells. These data are mainly generated by 10X genomics. The spatial transcriptomic data contain RNA distribution at all the spots in the tissue section.

We run those datasets on the published methods like DestVI, Stereoscope, Seurat, CARD, DSTG, Autogenes and many more. DestVI and Autogenes, and almost all other methods, have two model $sc-model$ and $st-model$. DestVI posits that for each gene g and each cell n , the number of observed transcripts follows a negative binomial distribution.The distribution is parametrized as (r_{ng}, p_g) with mean $\frac{(r_{ng}p_g)}{1-p_g}$ and where p_g is the gene specific parameter determining the mean-variance relationship at each spot.Parameter $r_{ng} = l_n\rho_{ng}$ of the negative binomial depends on the type assigned to c_n, and its overall number of detected molecules l_n and a low dimensional latent vector γ_n which captures the variability within its respective cell types. A neural framework maps γ_n and c_n to ρ_{ng}. The other st-model also relies on Negative Binomial distribution. Finally, we deconvolute the sc-model with the st-model getting the final disttribution at each spot.

Keywords: *Negative Binomial Distribution, Deconvolution and Neural Framework*

1 Introduction

Spatial transcriptomics opens up new opportunities to define the organization of cellular niches and crosstalk that modulate cellular function [7]. . In particular, this emerging technology helped study the organization of complex tissues such as the mouse brain[3] and the human heart [2]. The research of human pathologies, such as the structure of tumors, is also an important avenue for spatial transcriptomics[4] since the tumor microenvironment consists of a rich milieu of cell types and states that are organized in different anatomical niches.

Pseudo-bulk spatial transcriptomic measurements (Slide-Seq , 10x Visium) are appealing technologies as they provide measurements of the whole transcriptome, although the spatial resolution, in current versions, is limited to cell aggregates (10 microns for Slide-Seq and 55 microns for 10x Visium). Depending on the density of the tissue, a single bead spot of 10x Visium may have a large number of cells, emphasizing the need for deconvolution of spots to obtain a better resolution view of their cellular content. To overcome this limitation of current leading genome-wide spatial transcriptomics experimental protocols, these datasets are often matched with a single cell RNA-sequencing (scRNA-seq) dataset from the same tissue. The convention for analyzing such pairs of datasets (as implemented by all existing pipelines, including DestVI [5], AutoGeneS[1] and CARD [6]) is to apply a two-step process. First, a dictionary of cell types is inferred from the scRNA-seq data; then, the proportion of each cell type within each spot is estimated using a linear model. We have tried to calculate the accuracy of DestVI, AutogeneS and CARD for cell type deconvolution on 6 Visium mouse brain datasets, 2 simulated data from DestVI and Cell2Location and 9 other simulated datasets.

2 Methods

2.1 Model of single-cell RNA sequencing data

Let designate a cell in the scRNA-seq dataset. We assume that each cell is annotated with cell-type c_n label , but those labels are broad enough such that the introduction of continuous covariates into γ_n into the model helps explain additional variance in gene expression (i.e., within-cell-type variation). For example, $c_n$represents a discrete cell type (e.g., B cells or CD8 T cells) while $\gamma_n$is a continuous vector summarizing a sub-cell state of interest (e.g., B cell activation, CD8 T cell exhaustion). In the following, we assume that $c_n$is observed (e.g., obtained via clustering) and that γ_n , however, is unobserved and treated as a latent variable. We posit the following generative model for our data:

$$\gamma_n \sim Normal(0, I) \tag{1}$$

$$x_{ng} \sim NegativeBinomial(l_n f^g(c_n, \gamma_n), \rho_g) \tag{2}$$

,where $l_n$is the library size of cell n ,f is a two-layers neural network and $p$is a G-dimensional vector. f takes as input the concatenation of the one-hot encoding of c_n , as well as the scalar γ, and outputs a G-dimensional vector. f has an softplus non-linearity at its output to ensure positivity.

2.2 Model of the spatial transcriptomics data

In the spatial data, we assume that the gene expression of each spot is the combination of multiple cells, each potentially being from different cell types. A standard modeling assumption is that a spot s has for expression x_s the sum of individual cells . Similarly, let us assume each spot has C cells, and

that each cell $n$in spot s is generated from latent variables (c_{ns}, γ_{ns}). We then have a distribution of gene expression:

$$x_{nsg} \sim NegativeBinomial(l_s \alpha_g f^g(c_{ns}, \gamma_{ns}), p_g) \tag{3}$$

,with $l_s$is a spot specific scaling factor and $\alpha_g$is a correction term for the difference in experimental assays. From the standard property of the rate-shape parameterization of the negative binomial distribution, the distribution of the total gene expression x_{sg} in spot s and gene g is:

$$x_{sg} \sim NegativeBinomial(l_s \alpha_g \sum f^g(c_{ns}, \gamma_{ns}), p_g) \tag{4}$$

Keywords:*Negative Binomial Distribution, Deconvolution and Neural Framework*

3 Results

We evaluated our methods (i.e., DestVI, CARD and AutogeneS) on 6 Visium mouse brain datasets, 2 simulated data from DestVI and Cell2Location and 9 other simulated datasets. We first filtered 2000 highly variable genes from among all the gene types present. Thereafter, we find the intersection of the 2000 highly variables genes from single-cell data with the genes in spatial data. We, then, proceeded with the cell type deconvolution of the model.

For examining the accuracy of the model, we calculated F1 score, Precision and Recall. In order to set the prediction and truth cell type proportion to binary values, we set an array of thresholds varying from 0.1 to 0.001. We also calculated the average KL divergence and JS divergence. To calculate the correlation of predicted matrix with the true matrix, we also calculated cell-wise Pearson-correlation. Averaging the pearson correlation values helped us to see the average performance of the model in terms of correlation. The higher the correlation, the better the accuracy. Contrarily, higher divergence implies lower accuracy. We also attempted to visualize the results of Visium mouse brain datasets of all the three methods with Giotto tool in *R*. For Giotto plot, we first had to do the Kmeans clustering of the results from the methods and then Giotto helped plot them on their spatial locations according to the cluster they fall in. This helped clear visualization of cell-type distribution in mouse brain tissue samples.

Figure 1: Spatial Mapping of CARD result

Figure 2: Spatial Mapping of DestVI result

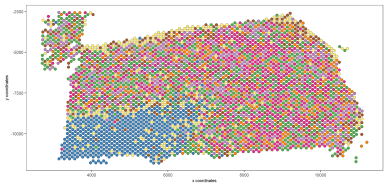


Figure 3: Spatial Mapping of Autogenes result

4 Summary

In this study, we evaluated the performance of 3 integration methods capable of combining spatial transcriptomics data and single-cell transcriptomics data. We have also provided a benchmark pipeline (Pipeline for above three methods).

References

[1] Hananeh Aliee and Fabian J Theis. Autogenes: Automatic gene selection using multi-objective optimization for rna-seq deconvolution. *Cell Systems*, 12(7):706–715, 2021.

[2] Michaela Asp, Fredrik Salmén, Patrik L Ståhl, Sanja Vickovic, Ulrika Felldin, Marie Löfling, José Fernandez Navarro, Jonas Maaskola, Maria J Eriksson, Bengt Persson, et al. Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Scientific reports*, 7(1):1–10, 2017.

[3] Simone Codeluppi, Lars E Borm, Amit Zeisel, Gioele La Manno, Josina A van Lunteren, Camilla I Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmfish. *Nature methods*, 15(11):932–935, 2018.

[4] MV Hunter, R Moncada, JM Weiss, I Yanai, and RM White. Spatial transcriptomics reveals the architecture of the tumor/microenvironment interface. *Nat. Commun*, 12:6278, 2020.

[5] Romain Lopez, Baoguo Li, Hadas Keren-Shaul, Pierre Boyeau, Merav Kedmi, David Pilzer, Adam Jelinski, Ido Yofe, Eyal David, Allon Wagner, et al. Destvi identifies continuums of cell types in spatial transcriptomics data. *Nature biotechnology*, pages 1–10, 2022.

[6] Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, pages 1–11, 2022.

[7] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145–1160, 2016.