DADIMPUTE: DATA DIFFUSION ON IMPUTING THE DROPOUT EVENTS IN THE SCRNASEQ PROFILE

Zican Zhu

Faculty of computer science, McGill University

*Supervised by **Amin Emad**, faculty of computer and electrical engineering*

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) technology enables scientists to approach problems that used to be seemingly inaccessible, however it suffers from a variety of biases and noises including 'dropouts', where undetected RNA molecules get assigned to zero (or near zero) counts. To address this, we developed DaDImpute (Data Diffusion on imputing scRNAseq), a novel computational approach incorporating information on the inherent cell-cell relationship as well as external information on gene-gene relationship estimate the true values of the false-zero counts. We corroborate DadImpute's ability to effectively recuperate the missing value and recover the gene expression patterns, by imputing masked glioblastoma data and comparing the result with the original values.

I. INTRODUCTION

The pioneering single-cell RNA sequencing technology provides an unprecedented high resolution of single-cell heterogeneity on a global scale, for an array of biological investigations. However, the technology is prone to high variation of the gene-gene expression and various noises introduced into the system, one of which comes from the so-called "dropout" events, a distinctive feature of scRNA-seq data, describing the presence of zero or near-zero inflated counts which can severely obscure the true gene-gene relationship.[1] The methods currently available to correct the dropout data often suffer from bias towards known (potentially erroneous) values, leading to error propagation in the dataset. In this paper, we present DaDImpute, a method that introduces in the prior known information on protein-protein interaction (PPI) as well as exploits the inherent cell-cell relationship from the scRNAseq profile to recover the true gene expression level. To validate DaDImpute's robustness and accuracy to rescue the sparse scRNAseq dataset, we ran the program on real glioblastoma data in which a proportion of non-zero counts were randomly masked. We compared the results with that produced by MAGIC and netSmooth to show DaDImpute is potent to recover the original data in a reasonable range.

II. BACKGROUND

Traditional RNA sequencing methods measure the RNAs of an ensemble of cells, yielding a bulk

average of the measurements instead of representing single cells. However, transcriptome information in any individual cell reflects only the transcriptional level of a subset of genes even they share nearly identical genomic information. Furthermore, all genes are not expressed in every type of cells, and each type of cell express a unique transcriptome[2]..

The scRNAseq technique has thrived in the years as it supports sequencing an entire transcriptomic profile at the level of individual-cell therefore its potential to identify rare events that are undetectable by analysing a pooled sample of cells [3]

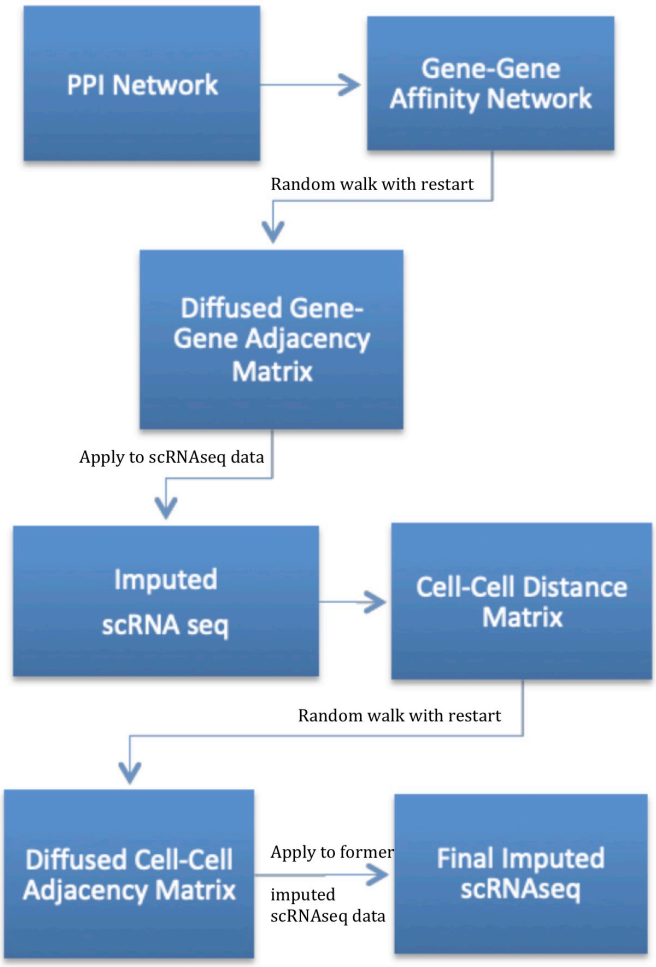
However, scRNA-seq profiling does not develop without challenges: it has a relatively higher noise level than bulk RNA-seq, typically, "dropout" events occur where the expression of a gene is not captured and assigned to a false-zero, due to the stochastic nature of the scRNA-seq technique [2] (only a fraction of total transcriptome may be detected, thus producing bias for lowly expressed genes).

The current methods designed to deal with this issue often study the measured information in the data: DrImpute[4] aggregates similar cells based on clustering, then impute by averaging the expression values from similar cells, in which process the single-cell resolution is compromised; MAGIC[5] relies on local averaging as well but by data diffusion across similar cells it finds a cell's closest neighbours. Both methods suffer from the bias towards known (potentially erroneous) values, leading to error propagation in the dataset. Another approach proposed is netSmooth[6], which also uses graph diffusion but on prior knowledge of protein-protein interaction (PPI) to infer gene-gene similarity, then used this to estimate the true value of the under-sampled counts. This method, however, neglects the information obtained from the scRNAseq data itself.

The PPI database can be accounted to infer gene-gene interaction/ co-expression activity because interacting proteins are more likely to be co-expressed[7], several studies in yeast have observed that protein pairs encoded by co-expressed genes interact more frequently than random pairs [8] [9].

III. METHOD

Overview of the algorithm



(FIGURE.1 Overview of DaDImpute algorithm)

We firstly projected protein-protein interaction scores onto corresponding encoding gene-gene pairs, which is then used to form a gene-gene adjacency matrix. We applied iterative random walk with restart (RWR) to this matrix until convergence. We then performed the first imputation by multiplying the diffused adjacency matrix to the left of the scRNAseq matrix (gene by cell). Next, a similar smoothed cell-cell adjacency matrix was constructed using cell-cell similarity embedded in the imputed matrix from earlier. We measured the Euclidean distance based on the transcription counts as a proxy to represent the similarity between cells. We multiplied again the resulted cell-cell adjacency matrix to right of the formerly imputed matrix, giving the final imputed scRNAseq. Necessary normalisations have been done during the process. Furthermore, the restarting probability for the RWR process was designed as a parameter and could be optimised.

Construction of the adjacency matrix

We used string-db[10] to generate the weighted PPI graph from which the gene-gene interaction was derived. For each protein pair, string-db provides an experiment score which indicates how often the interaction between the proteins were observed from experiments. Using this experiment score we constructed a protein-protein adjacency matrix. We then used the protein to gene mapping file that BioMart[11] provided to convert the proteins names to genes. We fed the whole matrix to the RWR step including those proteins that are not in the map, they still provide information on other genes. They were pruned right after RWR before degree-normalisation. Since we could only infer the gene-gene relationship for genes that are present in both the input scRNAseq profile and the PPI network, we subsetted the original expression matrix to contain only the genes in the PPI (18190 genes for human).

Construction of the cell-cell adjacency matrix

For each two cell samples in the given expression matrix from scRNAseq (gene by cell and m by n), we calculate the Euclidean distance between them by measuring the expression difference:

$$\boldsymbol{Distance}(\boldsymbol{C_i}, \boldsymbol{C_j}) = \sqrt{\sum_{\boldsymbol{k=0}}^{\boldsymbol{k=m}} \{(\boldsymbol{C_i}, \boldsymbol{G_k}) - (\boldsymbol{C_j}, \boldsymbol{G_k})\}^2}$$

Where $\boldsymbol{C_x}$ and G_x represent the *xth* cell and gene respectively. $(\boldsymbol{C_x}, \boldsymbol{G_y})$ is the expression level of *xth* cell and *yth* gene.

Because cell-cell similarity is negatively related to the cell-cell distance, we represent cell-cell affinity by the invert of cell-cell distance. To get rid of infinity similarity score by 0 distance, we additionally add 1 to the distance, so that infinitely far cells get 0 similarity and cells that are nearly identical get a high similarity score:

$$\boldsymbol{Sim}(\boldsymbol{C_i}, \boldsymbol{C_j}) = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{Distance}(\boldsymbol{C_i}, \boldsymbol{C_j})}$$

Finally construct the cell-cell adjacency matrix using the similarity scores. Normalise the result matrix to have row sums of 1.

The Normalisation process

We defined the $\boldsymbol{M_{i,j}}$ entry in the adjacency matrix $\boldsymbol{M}$ to be the edge weight hence the probability to transit from $\boldsymbol{node_i}$ to $\boldsymbol{node_j}$, thus we needed to make all the edges coming out from each node add up to a total weight of one. The formula we adopted to row-normalise the matrix(i.e. to make the row sums to one) follows:

$$M_{(i,j)} = \frac{M_{(i,j)}}{\sum_{k=0}^{k=n} M_{(i,k)}}$$

Note for those row comprised of all zeros, this step is skipped.

The random walks with restarts process

The RWR algorithm iteratively explores the network's structure to estimate the proximity between two nodes[12]. At each iteration of the random walk, a conceptual walker either jumps to a fixed restart point by a restart probability α or walks to a neighbour node by the probability of (1- α), in the case the walker continues to explore the network, the probability of to transit to a neighbouring node is given by the edge weight. Mathematically, we can represent each iteration by:

$$\boldsymbol{v^{t+1}} = (\mathbf{1}-\boldsymbol{\alpha})\mathbf{A}\boldsymbol{v^{t}} + \boldsymbol{\alpha v^{0}}$$

Where $\boldsymbol{v^{t}}_{[m\times 1]}$ is a vector representing the probability that the walker is at each node at time t. $\boldsymbol{v^{0}}_{[m\times 1]}$ is the vector of the probability to restart at each point. $\boldsymbol{A}_{[m\times m]}$ is the adjacent matrix where $A_{i,j}$ is the transition probability from node i to node j at each step. α is the restart probability and is between 0 and 1.

We then separately choose node i to be the restart point, assign 1 to $\boldsymbol{v^{0}}_{i}$ and 0 to other entries in $\boldsymbol{v^{0}}$. Through RWRs, information on gene-gene relationship was shared between nodes via data diffusion, and the result vector $\boldsymbol{v^{t}}$ after convergence contained the affinity scores each node has to the respective restart node.

To further simply the calculation, we stack the RWRs equation for each choice of restart node together and instead calculate the following:

$$\boldsymbol{M^{t+1}} = (\mathbf{1}-\boldsymbol{\alpha})\mathbf{A}\boldsymbol{M^{t}} + \boldsymbol{\alpha M^{0}}$$

where each row i in $\boldsymbol{M}$ is the respective vector. If we set the restart nodes by order, we get an identity matrix for $\boldsymbol{M^{0}}$.

The converged matrices are then to be applied to the input expression matrix $\boldsymbol{E}_{[m\times n]}$:

$$\boldsymbol{E}_{[m\times n]}{}' = \boldsymbol{M}_{g[m\times m]}\boldsymbol{E}_{[m\times n]}\boldsymbol{M}_{c[n\times n]}$$

Where $\boldsymbol{M_c}$ is the resulting matrix after RWRs on PPI network and $\boldsymbol{M_c}$ the cell-cell similarity matrix. $\boldsymbol{E'}$ is the imputed matrix and the ultimate output of DaDImpute.

Evaluation

The input scRNAseq profile we used to test and evaluate the different method is the glioblastoma[13] datasets (GSE57872) obtained from *conquer*[14], a repository of uniformly processed scRNA-seq datasets. Furthermore, we randomly removed (20%, 50%, 90%, 99.5%) proportion of non-zero entries in the dataset, then feed it to MAGIC, netSmooth and our program. Note for both MAGIC and netSmooth, the default and optimal parameters were chosen by inherent methods embedded in their programs. For DaDImpute, we twisted the restarting probability for the two RWRs to make the walker diffuse further or more close the restart node [15].

IV. Result

We imputed the masked glioblastoma datasets of 864 cells and 18190 genes using MAGIC, netSmooth and DaDImpute, then analysed the result matrices by comparing the Spearman's correlation coefficient rho between the original value of the masked counts and the values post imputation. Here the original values are treated as the "ground truth" and the correlation between those and the imputed values should reflect how well each method managed to recover the "true value".

DaDImpute out-performs netSmooth in terms of the recovery whereas MAGIC generally better recovers the masked data than DaDImpute

For all datasets masked ranging from 20% to 99.8%, DaDImpute were able to recover a larger portion of the "ground truth" data than netSmooth with a almost doubled Spearman's correlation coefficient (rho). From FIGURE.3 in *supplemental figures*, we can see that except for extremely sparse case, MAGIC was able to recover entries most correlated with the "true values".

DaDImpute is robust with extremely sparse datasets

DaDimpute has shown to be able to recover the masked non-zero entries with relatively high Spearman's correlation coefficient on matrices that are extremely sparse: for more than 90% non-zero data masked, DaDImpute was able to produce a rho of 0.5, whereas netSmooth got around 0.2 and MAGIC was able to recover the entries with rho of 0.6 up until the sparsity of 99.5% where the rho dropped to 0.26. (FIGURE.2, supplemental figures)

DaDImpute reduces the sparsity of the zero counts but also near-zeros

We investigated and realised that MAGIC produced an abundant of near-zero value and the resulting matrix is still relatively sparse (14.75% sparsity) in terms of the near-zero values (defined by us as smaller than one). NetSmooth also produced a relative sparse matrix (17.33% values being smaller than 1) whereas DaDImpute reduced the sparsity to 5.92%. We further found MAGIC also produces false negative counts after imputation which did not seem to be explained and dealt with. This issue was present with the vignette provided by them.

DaDImpute the total transcriptome expression up to a higher level

It is reasonable to expect the total transcriptome level would be increased post imputation, as we are bringing up the values of the dropout events. By analyzing the nature of the imputed matrix of the 20% masked dataset, we observed that the mean value was brought back to a higher value (27tpm) closer to that of the original non-masked, non-imputed dataset (34.9 tpm), but MAGIC yielded a matrix with much lower transcriptome counts (2.78 tpm) than the original.

V. DISCUSSION

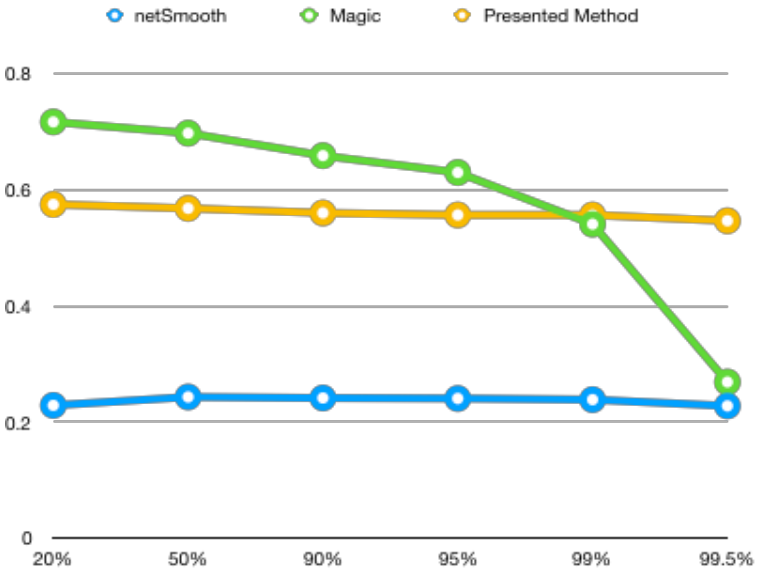
The single-cell RNA sequencing (scRNA-seq) technology provides a higher resolution of cellular differences on the whole genome profile which has allowed biological scientists to approach problems that used to be seemingly inaccessible. However, the result profile suffers from a variety of noises including the dropout event zero-inflated counts are present. which can severely obscure the true gene-gene relationship. The methods currently available either correct the dropout data solely on the non-zeros in the data which leads to error propagation in the dataset, or strongly rely on external information. To address this, we incorporate information from the data as well as external information from PPI to

estimate the true values of the false-zero counts and developed DaDImpute, which has shown to be able to imputed data correlate with the "true value" better than netSmooth, though may not as good as MAGIC does. On the other hand, MAGIC does present risks to produce negative counts and does not improve the sparsity of near-zeros (defined as smaller than one) significantly, whereas DaDImpute is able to reduce that to a relative low number. The information on the PPI network introduced by DaDImpute also in theory reduces the effect of the error propagation of the potentially erroneous data in MAGIC. The method is also very robust with extremely sparse data relative to both netSmooth and MAGIC. Finally, the method is versatile and may be incorporated into other experiments where a high quality PPI network is available to de-noise the system.

VI. Supplemental Figures

FIGURE.2
Data analysis of the imputed data of 20% masked glioblastoma dataset.
Here, the mean of MAGIC imputed data drops significantly and both MAGIC and netSmooth retain a good amount of near-zero counts.

	Raw	MAGIC	netSmooth	DaDImpute
Counts of zeros (<0)	2442137	365	360	0
Sparsity (zero)	15.54%	0.00%	12.74%	0.00%
Sparsity (<=1)	23.79%	14.75%	17.33%	5.92%
Mean	34.94791	2.78016	35.8749	26.99202
Rho	N/A	0.629703	0.2433598	0.5753443



(FIGURE.3)
The Spearman's correalation coefficient of the original value of the masked data and the imputed data by netSmooth, MAGIC and DaDImpute.
We see that MAGIC and DaDImpute imputed data correlate with the "true value" better than netSmooth.
DaDImpute is still robust with extremely spaese dataset where MAGIC's rho values drops to below 3.

VII. BIBLIOGRAPHY

[1] Oh, S., & Song, S. (2018). Bayesian Modeling Approaches for Temporal Dynamics in RNA-seq Data. New Insights into Bayesian Inference. doi:10.5772/intechopen.73062

[2] O. Stegle, S. A. Teichmann, and J. C. Marioni, Computational and analytical challenges in single-cell transcriptomics, *Nature News*, 28-Jan-2015. [Online]. Available: https://www.nature.com/articles/nrg3833. [Accessed: 08-Dec-2018].

[3] B. Hwang, J. H. Lee, and D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, *Experimental & Molecular Medicine*, vol. 50, no. 8, 2018.

[4] Kwak, I., Gong, W., Koyano-Nakagawa, N., & Garry, D. (2017). DrImpute: Imputing dropout events in single cell RNA sequencing data. doi:10.1101/181479

[5] van Dijk D, Nainys J, Sharma R, et al.: Magic: A diffusion-based imputation method reveals gene- gene interactions in single-cell rna-sequencing data. bioRxiv. 2017.doi: https://doi.org/10.1101/111591

[6] Ronen, J., & Akalin, A. (2018). NetSmooth: Network-smoothing based imputation for single cell RNA-seq. F1000Research, 7, 8. doi:10.12688/f1000research.13511.

[7] Bhardwaj, N., & Lu, H. (2005). Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics,21(11), 2730-2738. doi:10.1093/bioinformatics/bti398

[8] Grigoriev, A. 2001A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. Nucleic Acids Res. 293513–3519

[9] Ge, H., et al. 2001 Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat. Genet 29482–486

[10] Szklarczyk D, Morris JH, Cook H, et al.: The STRING database in 2017: quality- controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017; 45(D1): D362–D368.

[11] Guberman, J. M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., . . . Kasprzyk, A. (2011). BioMart Central Portal: An open database network for the biological community. Database,2011(0). doi:10.1093/database/bar041

[12] Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., . . . Baudot, A. (2018). Random walk with restart on multiplex and

[13] Patel AP, Tirosh I, Trombetta JJ, et al.: Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344(6190): 1396–1401.

[14] Soneson C, Robinson MD: Bias, robustness and scalability in differential expression analysis of single-cell rna-seq data. bioRxiv. 2017.

[15] Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., . . . Baudot, A. (2018). Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics. doi:10.1093/bioinformatics/bty637