Reducing Noise in Single-Cell Data with Generative Networks

Jaison Jain[1], Ben Foulon[2], and Daniel Ben-Isvy[1]
[1]Center for Computational Molecular Biology, Brown University
[2]School of Engineering, Brown University

CENTER FOR Computational Molecular Biology

BROWN Computer Science

Introduction

- Single-cell RNA sequencing (scRNA-seq) data holds immense promise for understanding gene expression at the cellular level
- However, this data is very noisy due to the low RNA capture rate of scRNA-seq methods
- While denoising algorithms exist (van Dijk et al., Huang et al.), most of these techniques do not take full advantage of the sparsity and inherent nonlinearities in scRNA-seq data
- The deep count autoencoder (DCA) model proposed by Eraslan et al. successfully addresses these problems:
  - DCA explicitly models data sparsity by using a zero-inflated negative binomial (ZINB) likelihood function
  - The ability of neural networks to learn complex nonlinear functions allows DCA to successfully model nonlinear gene-gene dependencies

Methods

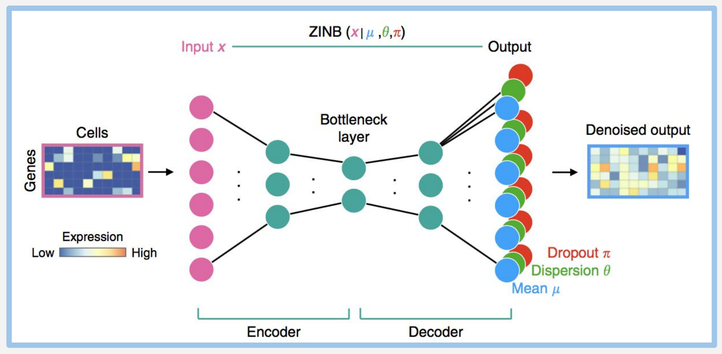

FIGURE 1: Model architecture, from Eraslan et al. The internal layers have sizes 64, 32, and 64 with ReLU activation. The mean and dropout output layers have exponential activations, and the dispersion output layer has sigmoid activation. The model was trained to maximize the likelihood of the input data according to the ZINB distribution.

$$\mathrm{NB}(x;\mu,\theta) = \frac{\Gamma(x+\theta)}{\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{x}$$

$$\mathrm{ZINB}(x;\pi,\mu,\theta) = \pi\delta_0(x) + (1-\pi)\mathrm{NB}(x;\mu,\theta)$$

EQUATION 1: The negative binomial (NB) and zero-inflated negative binomial (ZINB) distributions, from Eraslan et al.

- Simulated scRNA-seq data was generated with the Splatter R package (Zappia, Phipson, & Oshlack)

Results

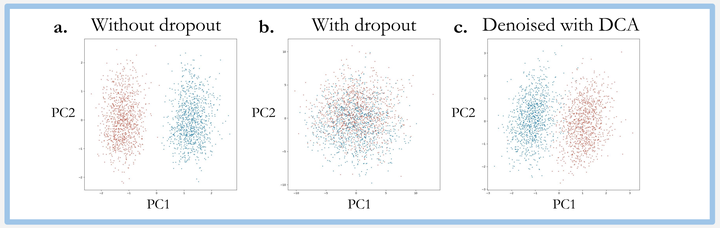

FIGURE 2: Principal component (PC) plots of simulated scRNA-seq data of 2000 cells with 198 genes, distributed into two groups. **a.** With no dropout noise, cell types can be distinguished along the top two PCs. **b.** With 65% dropout noise, cell types are no longer distinguishable. **c.** Denoising with DCA restores distinguishability.

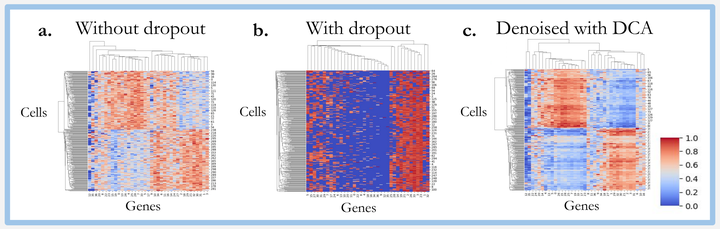

FIGURE 3: Heatmap of the gene expression matrix for two cell types with 198 genes. **a.** With no dropout noise, gene expression patterns are markedly different between cell types. **b.** With 65% dropout noise, gene expression patterns appear far more similar. **c.** Denoising with DCA recovers more differentiable gene expression patterns.

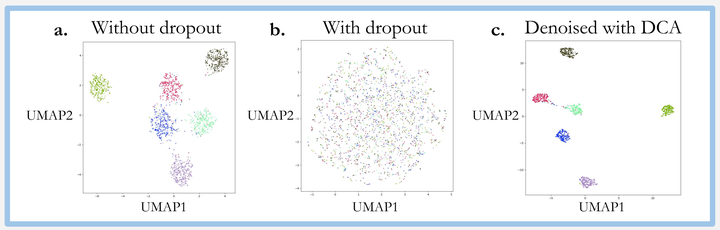

FIGURE 4: UMAP charts of simulated scRNA-seq data of 2000 cells with 195 genes, distributed into six groups. **a.** With no dropout noise, cell types can be distinguished along the two UMAP axes. **b.** With 35% dropout noise, cell types are no longer distinguishable. **c.** Denoising with DCA restores distinguishability.

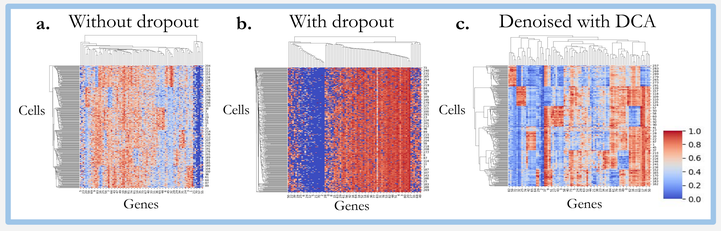

FIGURE 5: Heatmap of the gene expression matrix for six cell types with 195 genes. **a.** With no dropout noise, gene expression patterns are markedly different between cell types. **b.** With 35% dropout noise, gene expression patterns appear far more similar. **c.** Denoising with DCA recovers more differentiable gene expression patterns.

Results

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}$$

EQUATION 2: An expression representing the Adjusted Rand Index (ARI), a measure of clustering accuracy (Yeung & Ruzzo). n_{ij} is the number of objects found in both class i and cluster j.

- After clustering datasets with the K-means algorithm, we achieved an ARI of 0.91 and 0.97 for the denoised 2-group and 6-group datasets, respectively.

Discussion

- Limitations:
  - While DCA can effectively denoise scRNA-seq data from one experiment, it does not correct for batch effects, or non-biological variation between measured gene expression levels in different experiments
  - Simulated data has cell type labels (hidden during training), but it is much harder to determine if DCA correctly clusters real datasets without known cell types
  - Gene-gene correlations differ across cell types, meaning that DCA must be re-trained on datasets composed of distinct sets of cell types.
- Future directions:
  - Validate the accuracy of DCA on real data with known cell types
  - Expand DCA to correct for batch effects, which could allow joint clustering of cell types across comparable datasets

References

1. Eraslan, G. et al. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 10, 390 (2019).
2. van Dijk, D. et al. MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv* (2017).
3. Huang, M. et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542 (2018).
4. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 18, 174 (2017).
5. Yeung, K. & Ruzzo, W. Details of the Adjusted Rand Index and clustering algorithms. *Bioinformatics* (2001).