Aarhus University

Bioinformatics Master Thesis

Bioinformatics Research Center & Department of Molecular Medicine

# Sequence-based neural networks applied to denoising of single-cell RNAseq data

Author:
Jens Brogård Stenbye
Stdnr: 201505646

Supervisors:
Lasse Maretty Sørensen
Palle Villesen

April 30, 2022

# Contents

# 1 Acknowledgements

I would like to thank the following people who have been integral to me completing this project. First and foremost a massive thanks to my supervisor Lasse Maretty Sørensen not only for his insight and supervision regarding the project but also his kindness and understanding of the troubles i encountered during the project.

Furthermore i would like to extend a special thanks to Christian Storm and Palle Villesen from BiRC for their help, patience and understanding in the administrative issues i encountered.

Finally i would like to thank my family and also my good friend Erik Gunnersen whose support have been invaluable during the making of this thesis.

# 2 Abstract

Single-cell RNA-sequencing(scRNAseq) offers many exciting analysis opportunities. Unlike bulk RNA-sequencing, scRNAseq gives an expression profile for each individual cell in a sample. This allows for identification of rare sample cell-subtypes, analysis of cell differentiation among other interesting analysis. The sequencing depth of each cell in scRNAseq is however very small, leading to many dropout events and general overdispersed data.

Several methods have been developed in order to denoise scRNAseq data, imputing the dropout events and correcting counts. In recent years, several neural network models have been developed in the field of scRNAseq denoising. Common for all current modelling approaches is that they only utilize the scRNAseq gene-counts themselves to denoise the counts.

In this thesis we argue that the inherent sequence context of each gene holds expression information, and that this could be utilized in inferring the denoised gene-counts. To that end we develop the Sequence- and Count-based Encoder and Linear Decoder(SCELD) model which utilizes sequence information in addition to the count-matrix. We also implement the DeepCountAutoencoder(DCA) an existing scRNA-denoising model as a benchmark model.

In general the SCELD was slightly outperformed by DCA across the several denoising metrics. Especially when investigating the ability to retain the clustering of cells when trained on a corrupted dataset did SCELD perform poorly while DCA was able to reproduce the cell clusters. The choice of benchmarking data and the model design may not have been optimal however, and we argue that further research is needed to properly evaluate the use of sequence information in scRNAseq denoising.