Genetic Regulatory Network Inference Utilizing Transformers

Yash Patel (Mentor: Dr. Yue Wang, 2024 IICD SRP)

Gene expression is the process in which genes are transcribed into mRNA and then translated into proteins. Gene expression levels can be measured by the mRNA count or the protein count. The expression of certain genes may influence the expression of other genes (mutual regulation) or itself (autoregulation). The regulatory relationships between genes form a gene regulatory network (GRN).

Determining the GRN structure is valuable in biology, as it provides insights into complex biological processes that can be leveraged to advance treatments, such as cancer therapies. Establishing the GRN with current technologies is challenging because biochemical methods cannot directly observe the inside of a living cell at multiple time points. Instead, single-cell gene expression data have been collected and used in various methodologies to infer the GRN structure. Many inference methods utilize single-cell expression data at one time point, but there are fewer methods that handle data over multiple time points, where the joint distribution is unknown. Recently, experiments have measured single-cell gene expression levels at multiple time points without the joint distribution being known. Such data at different time points are generally treated as separate, resulting in the time information being wasted. We aim to develop new methods for this type of data.

There have been many non-deep learning GRN inference methods. Most inference methods convert the GRN inference problem into a feature selection problem and solve it using regression or other approaches (e.g., ARACNe, GENIE3, BiXGBoost). Other methods (e.g., DBN, WENDY, Dictys) build various models of gene expression and use optimization to find the best parameters that represent the GRN. Recently, several deep learning GRN inference methods have been proposed. The involved neural network structures include RNNs (e.g., DA-RNN), transformers (e.g., STGRNS), CNNs (e.g., 3DCEMA), and autoencoders (e.g., DeepSEM). However, some of these methods do not process the data in a biologically meaningful way.

This work aims to improve the accuracy of current GRN inference methods based on single-cell RNA expression levels over multiple time points, using deep learning structures in a biologically meaningful way. One proposed method is to utilize a transformer neural network, where the input is an inferred GRN, and the output is a refined prediction of the true GRN. By using the initial guess as the input rather than the expression data, it may be possible to increase the accuracy and robustness of GRN inferences from single-cell RNA expression levels over multiple time points. Another proposed method is to approach GRN inference as a link prediction problem within the graph neural network framework. Although a few papers have explored this idea, we plan to incorporate other inference methods to build a biologically meaningful graph neural network to infer the GRN.