Data Representations for Unsupervised Learning of RNA Structure

Sitara Persad *
Massachusetts Institute of Technology
spersad@mit.edu

Abstract

Beyond carrying genetic information, RNA folds into higher order structures that carry out key cell process. Computational approaches for determining RNA structure do not reflect the chemical environment of the cell and are often inaccurate. Thus, experimental assays are used to probe RNA structures in vivo and provide constraints by computing the average reactivity profile over all molecules. However, this obscures the fact that RNA can exist in multiple stable structures in the cell, which look very different from the population average.

The goals of this project are to infer the number of structures in a cell, based on experimental probing data, and to cluster this data to characterize individual structures. This data consists of (unlabelled) reads up to 1000 nucleotides long. For a test set, we mix two alleles of a small RNA riboswitch which differ by a single nucleotide, inducing stable structural changes. The data points are binary reads where a '1' represents a probable open base and a '0' represents a probable closed base.

We implement methods in learning data representations to embed these reads into two dimensions in order to visualize the structures. We implement methods of distance preservation and of minimizing reconstruction error and evaluate their performance. Finally, we propose and evaluate a modified dissimilarity metric which incorporates the distance between nearest neighbors.

To minimize reconstruction error, we implement a stacked autoencoder, mapping the data to two dimensions and found that the autoencoder separates the points corresponding to each structure almost perfectly. However, the data is so distorted in two dimensional space that it is impossible to determine the number of clusters. Thus, for unlabelled datasets, this method is unsuitable. We implement Deep Embedded Clustering, which simultaneously learns embedding as well as cluster centers and assignments. This method performs poorly, failing to separate the data and cluster accurately.

For distance preserving metrics, we analyse t-Stochastic Neighbor Embedding (t-SNE) and Multidimensional Scaling (MDS), using the Hamming distance between reads, as we expect reads generated by the same structure to have overlapping open bases and closed bases. These methods both perform poorly, failing to separate the dataset into its structural conformations. We hypothesise that this is because the chemical modification is so sparse that reads from the same structure are not very close to each other. To correct for this sparsity, we propose a new dissimilarity metric. Intuitively, if two reads are generated from the same structure, their nearest neighbors are similar as well. We implement this new metric and find that, using MDS, we can separate the data perfectly and visualize two distinct structures. Using hierarchical clustering, we recover the initial clusters with 98.5% accuracy.

We have developed a successful method for data representation to learn RNA structure heterogeneity, which will allow us to experimentally determine multiple

*

structures within a cell, and allow further investigation into their conformation in diseased and healthy cells. Future directions include studying the limits under which we can accurately cluster structures.