Data Representations for Unsupervised Learning of RNA Structure

Sitara Persad
MIT 6.860/9.520 Project

Introduction

Beyond carrying genetic information, RNA folds into higher order structures that carry out key cell process. RNA structure prediction algorithms do not reflect the chemical environment of the cell, which affects accuracy. Experimental data constrain these predictions, but compute only average signal over all RNA molecules, obscuring the structures which perform different functions.
We study low dimensional data representation via distance preservation and minimal reconstruction error to identify and cluster individual structures. We implement t-SNE, multi-dimensional scaling (MDS), autoencoders and deep embedded clustering and evaluate their performance on a test molecule.

RNA Structure Probing

Dimethyl Sulfate probes RNA structure *in vivo* by modifying the Watson-Crick position of A and C bases. TGIRT enzyme converts modifications to mutations that are sequenced in the DNA. [1].

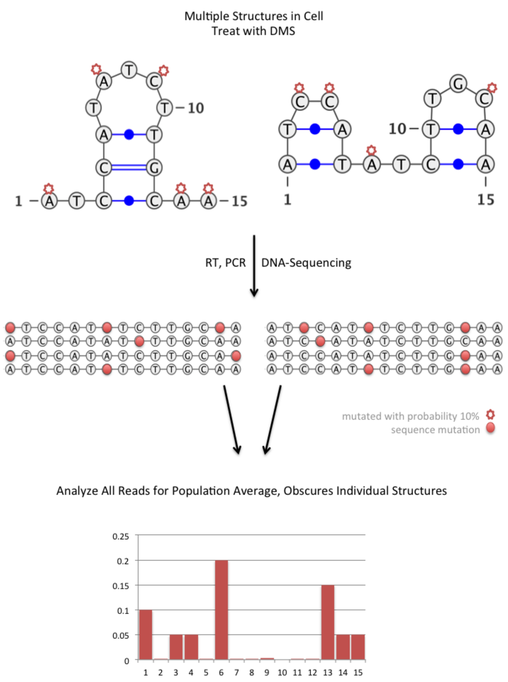


Figure: Experimental Assay for Determining RNA Structure Constraints

Method

1. Synthesize two structural forms of a small RNA molecule which differ at a single genetic locus as a **test molecule** (structure membership typically unknown).

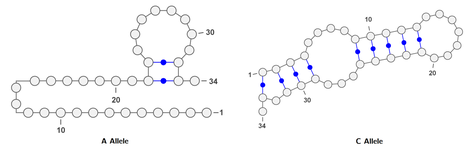
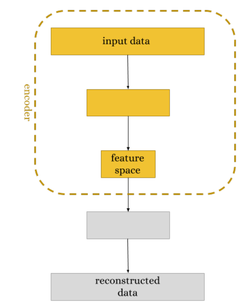


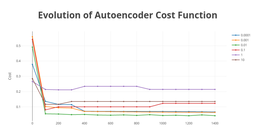
Figure: Multiple Structures Mixed at Ratio 1:3 and Probed In Vitro

2. Identify **unique reads**.
3. Learn lower-dimensional representation of reads for clustering, using distance preservation methods and reconstruction algorithms.

Autoencoder



(a) Autoencoder Architecture



(b) Autoencoder converges to minimal reconstruction cost

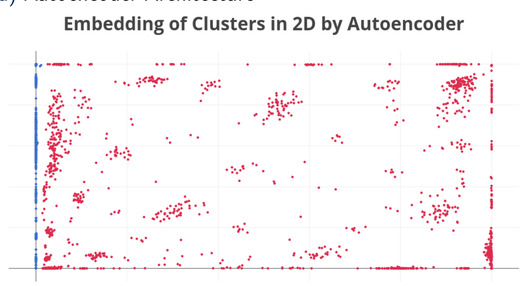


Figure: Autoencoder perfectly separates data but obscures number of true clusters

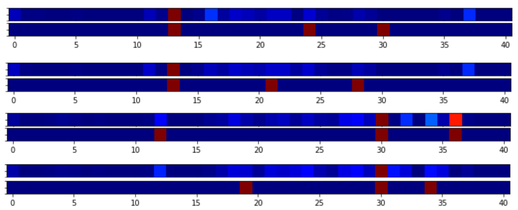
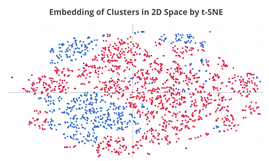


Figure: Sample reconstruction (above) from original (below)

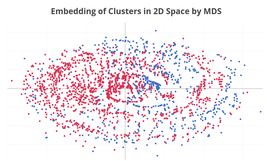
Distance Preservation

D is the matrix of Hamming distances between reads.

1. t-SNE [2] minimizes divergence between distribution of embedded points, y, and distribution defined original dissimilarity matrix.
2. MDS [3] minimizes the difference between the distance between points in the embedded space and original dissimilarity matrix.



(a) tSNE Fails to Separate Clusters in 2D



(b) MDS Fails to Separate Clusters in 2D

Distance Preservation + kNN

We propose a hybrid method combining k-nearest neighbors.

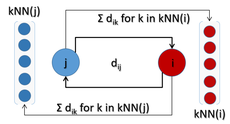


Figure: Include average distance from x_i to kNN of x_j

$$\delta_{ij} = D_{ij} + \alpha(\Sigma_{k \in KNN(j)} D_{ik} + \Sigma_{k \in KNN(i)} D_{jk})$$

(a) tSNE with 5NN Fails to Separate Clusters in 2D

(b) MDS with 5NN Separates Clusters Well in 2D

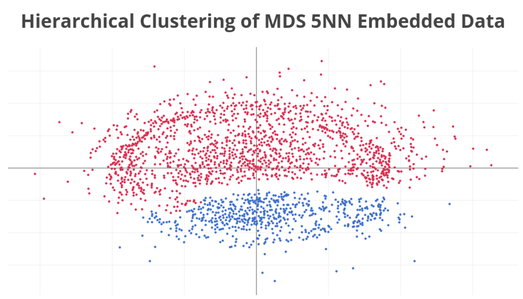


Figure: Hierarchical Clustering Recovers Structures with 98.5% Accuracy

Deep Embedded Clustering

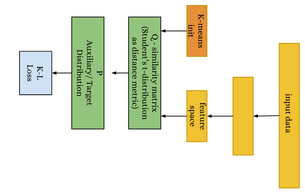


Figure: DEC uses autoencoder + distance preservation from a cluster exemplar to simultaneously learn encoding and assignments

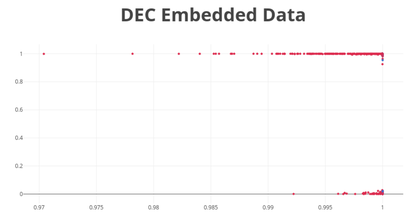


Figure: DEC fails to separate data, achieving 66.3% accuracy.

Conclusion

t-SNE and MDS cannot separate the sparse sequence reads. However, if we include nearest neighbours, we can visualize the number of structures using MDS and perform clustering.
Autoencoders separate the data in 2D, but do not allow us to infer the number of structures present. Deep Embedded Clustering does not separate the data, and fails to produce accurate clustering.
Next steps are to benchmark longer molecules, molecules with increased similarity, and structures mixed in smaller ratios

References

[1] M. et al Zubradt.
Dms-mapseq for genome-wide or targeted rna structure probing in vivo.
Nature methods, 14(1):75–82, 2017.

[2] L. et al Maaten.
Visualizing data using t-sne.
Journal of Machine Learning Research, 9(Nov), 2008.

[3] J Kruskal.
Nonmetric multidimensional scaling: a numerical method.
Psychometrika, 1964.

[4] J. et al Xie.
Unsupervised deep embedding for clustering analysis.
In *Int'l Conference on Machine Learning (ICML)*, 2016.