scTranslate: Learning to Translate Between Epigenetic and Transcriptomic Single-Cell Assays

Benjamin J. Lengerich*[1], Michael Kleyman*[1], Andreas R. Pfenning[1], Eric P. Xing[1,2]

{blengeri, mkleyman} @cs.cmu.edu

1) Carnegie Mellon University 2) Petuum, Inc.

Guiding Questions

Is there a latent space which summarizes **both** epigenetic and transcriptomic cell state?

Can we use this to translate between single cell RNA-seq data and single cell ATAC-seq data?

Motivation

1. Impute missing assay from single assay data

Majority of single cell data is scRNA-seq or scATAC-seq. Ability to infer the other assay would provide understanding without needing to perform expensive dual assays.

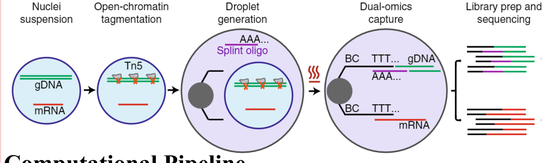
2. Cell type-specific gene regulatory mechanisms

Interpretable translator could link changes in open chromatin events to gene expression events.

3. Learn a highly accurate latent space to define cell state from multiple views.

Data

Snare-seq: Dual scRNA-seq and scATAC-seq on the same adult mouse cerebra cortex cells (Chen et al.2019)



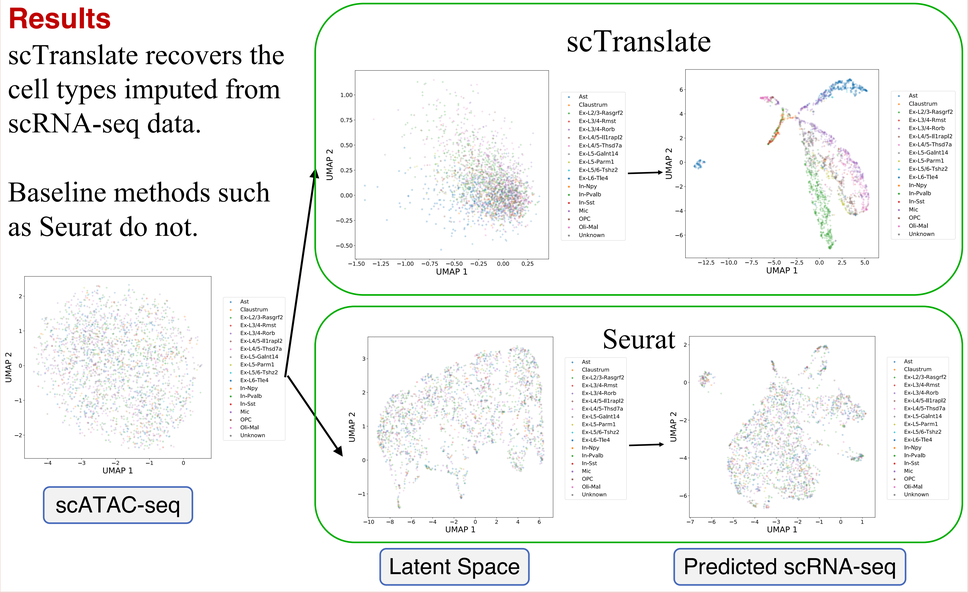
Computational Pipeline

Computational Step	Doublet Removal	scRNA-seq Normalization	scATAC-seq QC	scATAC-seq peak calling	Cell type inference
Software Package	*SCDS*	*Linnorm*	*SnapATAC*	*MACS*	*Scanpy*

Results

scTranslate recovers the cell types imputed from scRNA-seq data.

Baseline methods such as Seurat do not.



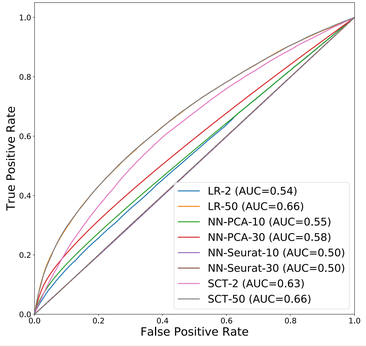
Predictive performance

Takeaways: SCT performs as well as non-bottlenecked LR, and the latent space can be significantly optimized. Indicates that there exists a small latent space governing both epigenetic and transcriptomic state.

Metric: ROC of scATAC-seq (mean of 1000 peaks)

Baseline Models:

- LR-n: Logistic Regression from n PCs of RNA-seq
- NN-PCA-k: k-Nearest Neighbors in 50 PCs of RNA-seq
- NN-Seurat-k: k-Nearest Neighbors
- SCT-n: scTranslate with n-dimensional latent space



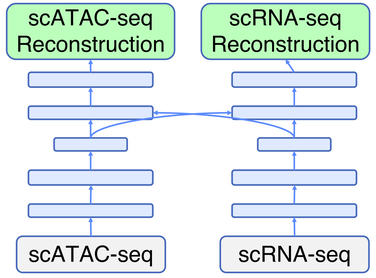
Model

Multi-modal Linear VAE

Losses:

- Translation
- Latent
- Reconstruction

Design allows for semi-supervised training



Open Questions

1. How much transferability of translators is there between replicates, different tissues, and different assay platforms?
2. How similar will cell types need to be perform transfer learning where we train on one cell type and predict on the other?
3. How much will semi-supervised training help?
4. What loss functions would be optimal for the translation architecture?

Summary

- For the first time, we have data to translate between epigenetic and transcriptomic state at the level of individual cells.
- Can investigate cell type-specific regulatory patterns.
- Translation is technically challenging because assays are sparse and have many sources of experimental noise.
- Simple models tend to outperform complex ones, especially in this setting of limited data. Generalizability is the highest priority.

References

Chen, S., Lake, B.B. and Zhang, K., 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, pp.1-6.

Bais, A.S. and Kostka, D., 2019. scds: Computational Annotation of Doublets in Single Cell RNA Sequencing Data. *bioRxiv*, p.564021.

Yip, S.H., Wang, P., Kocher, J.P.A., Sham, P.C. and Wang, J., 2017. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic acids research*, *45*(22), pp.e179-e179.

Fang, R., Preissl, S., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Mukamel, E.A., Zhang, Y., Behrens, M.M. and Ecker, J., 2019. Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *bioRxiv*, p.615179.

Wolf, F.A., Angerer, P. and Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, *19*(1), p.15.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W.M., Hao, Y., Stoeckius, M., Smibert, P. and Satija, R., 2019. Comprehensive Integration of Single-Cell Data. *Cell*.

Liu, T., 2015. MACS–Model-based Analysis of ChIP-Seq.