EnsembleRegNet: Leveraging Ensemble Encoder-Decoder and Multiple-Layer Perceptron Bagging for Predicting Gene Regulatory Networks from Single-Cell RNA-Seq Data

Duaa Mohammad Alawad[1], Ataur Katebi[2,3] , Md Tamjidul Hoque[1]

email: dmalawad@uno.edu, arkatebi@gmail.com, thoque@uno.edu

[1]Department of Computer Science, University of New Orleans, New Orleans, LA, USA.

[2]Department of Bioengineering, Northeastern University, Boston, MA, USA.

[3]Center for Theoretical Biological Physics, Northeastern University, Boston, MA, USA.

THE UNIVERSITY of NEW ORLEANS

Bioinformatics & Machine Learning Lab

Introduction

- Biological processes are regulated by underlying genes and their interactions that form gene regulatory networks (GRNs).
- Dysregulation of these GRNs can cause complex diseases such as cancer, Alzheimer's, and diabetes.
- Single-cell transcriptome profiling provides detailed data from thousands of cells. This helps to identify key gene regulatory networks and modules in different cell types, enhancing our understanding of various biological mechanisms.

Materials

Table 1: Three Benchmark scRNA-seq Datasets Sourced from GEO

GEO accession	Title	Summary
GSE60361	cerebral cortex cells from mice	3005 single cells from 33 males and 34 females
GSE115978	human skin cutaneous melanoma	7186 single cells from 31 melanoma tumors
GSE103322	head and neck squamous cell carcinoma	5902 single cells from 18 patients with oral cavity tumors

Performance Evaluation Metrics –part 1

There are mainly two categories of clustering validation:

1. Internal clustering validation

Metric	Formula
Root-mean-square standard deviation (RMSSTD)	$RMSSTD = \sqrt{\frac{\sum_i \sum_{x \in C_i} \|\|x - c_i\|\|^2}{p \sum_i (n_i - 1)}}$ Where: •c_i is the centroid of cluster (i). •n_i is the number of data points in cluster (i). •P is the total number of data points.
Modified Hubert Γ statistic	$\Gamma = \frac{2}{n(n-1)} \sum_{x \in D} \sum_{y \in D} d(x,y) \cdot d_{x \in C_i, y \in C_j}(c_i, c_j)$ Where: •d is a distance measure between two data points. •D is the dataset. • C_i and C_j are clusters. • c_i and c_i are the centroids of the clusters C_i and C_j respectively. •n is the number of data points. The term $d_{x \in C_i,\ y \in C_j}(c_i, c_j)$ represents the distance between centroids of the clusters to which x and y belong, respectively.
Silhouette index score	$S = \frac{1}{NC} \sum_i \frac{1}{n_i} \sum_{x \in C_i} \frac{b(x) - a(x)}{\max[b(x), a(x)]}$ Where: •(a) is the average distance from the i^{th} data point to the other data points in the same cluster. •$b(x)$ is the smallest average distance from the i^{th} data point to data points in a different cluster, minimized over clusters. •C_i is the i^{th} cluster. •ni is the number of data points in cluster C_i . •NC is the number of clusters.
R-squared index	$R^2 = \frac{\sum_{x \in D} \|\|x - c\|\|^2 - \sum_i \sum_{i \in C_i} \|\|x - c_i\|\|^2}{\sum_{x \in D} \|\|x - c\|\|^2}$ Where: •D is the dataset. •C_i is the i^{th} cluster. •c is the centroid of the entire dataset D. •c_i is the centroid of the i^{th} cluster.

Methods

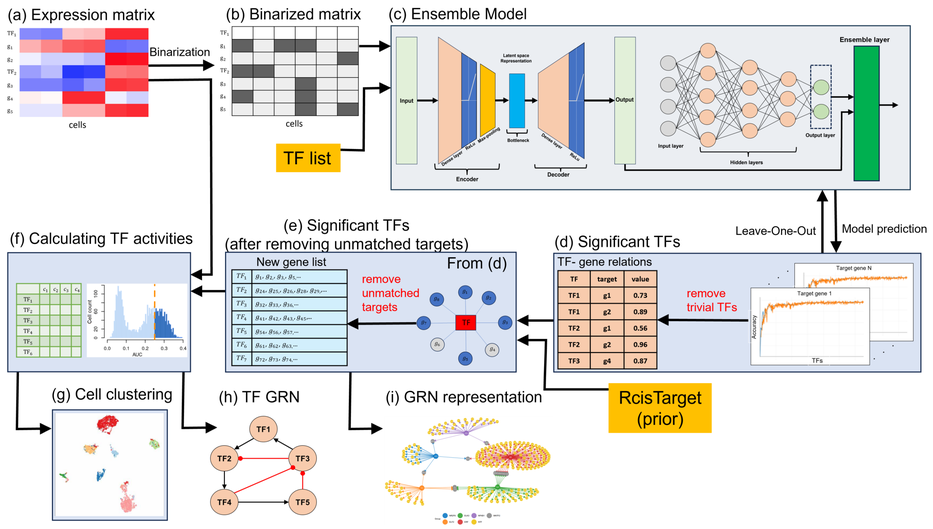


Figure 1. EnsembleRegNet framework for Predicting Gene Regulatory Networks from Single-Cell RNA-Seq Data.

1. **Data Preprocessing**: This initial phase is crucial for ensuring the scRNA-seq data is of optimal quality and adequately prepared for sophisticated analyses, such as deducing gene regulatory networks and deciphering cellular diversity.

2. **Ensemble Models**: This involves the implementation of multiple fully connected Multi-Layer Perceptron (MLPs) with an encoder-decoder framework. The primary objective here is to discern the interactions between transcription factors (TFs) and non-TFs or downstream targets. Each model in the ensemble is dedicated to predicting a specific feature gene.

3. **RcisTarget Utilization**: This part involves the application of the RcisTarget package. Its primary function is to determine if the TFs can authentically bind to the motifs of the target genes, also known as regulons, by analyzing motif enrichment scores.

4. **AUCell Analysis**: This stage employs AUCell to assess the relative activity of the regulons in individual scRNA-seq data samples. The analysis relies on the Area Under the Curve (AUC) score matrix to draw inferences.

5. **Cell Clustering**: This process facilitates an in-depth exploration of gene regulatory networks on a cell-type basis, allowing for a more nuanced understanding.

6. **Network Visualization**: The final step involves graphically representing the intricate interplay within the gene regulatory networks as inferred from scRNA-seq data. This visualization is key to interpreting data, formulating hypotheses, and effectively communicating research findings

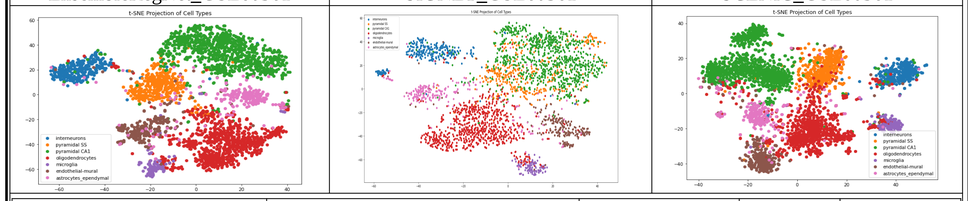
Performance Evaluation Metrics –part 2

2. External clustering validation.

Metric	Formula
the adjusted rand index (ARI)	$ARI = \frac{RI - E[RI]}{1 - E(RI)}$ Where: • Expected_RI is the expected value of the Rand Index of two random clustering. • Max_RI is the maximum possible value of the Rand Index, which is 1.
pair sets index (PSI).	$\begin{Bmatrix} \frac{S - E(s)}{K - E(s)}, & S \geq E(S), K > 1 \\ 0, & S < E(S), \end{Bmatrix}$, $S = \sum_{i=1}^{K} \frac{n_{ij}}{\max(n_i, m_j)}$ $e_{ij} = \sum_{i=1}^{K} \frac{n_i \times m_j}{N}$, $E(S) = \sum_{i=1}^{K} \frac{e_{ij}}{\max(n_i, m_j)}$ Where: •n_{ij} is the number of items in cluster i of the first clustering. •m_j is the number of items in cluster j of the second clustering. •n_{ij} is the number of items common to cluster i of the first clustering and cluster j of the second clustering.

Results

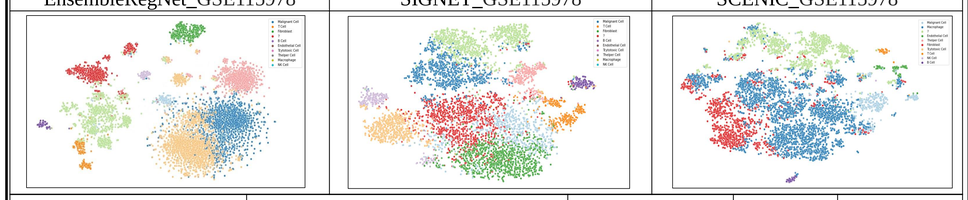
EnsembleRegNet_GSE60361 | SIGNET_GSE60361 | SCENIC_GSE60361



	Measure	EnsembleRegNet	SIGNET	SCENIC
External clustering validation	the adjusted rand index (ARI)	**0.012**	0.0090	0.0091
	normalized mutual information (NMI)	**0.021**	0.0258	0.035
	F-score	**0.069**	0.0767	0.087
	Normalized Van Dongen (NVD) score	**0.747**	0.744	0.223
	Pair Sets Index (PSI) score	**0.489**	0.477	0.484
Internal clustering validation	RMSSTD	0.0159	**0.012**	**0.012**
	Modified Hubert Γ statistic	**127.080**	91.266	116.120
	Silhouette index	**0.497**	0.471	0.458
	R-squared index	**0.978**	0.958	0.975

Table 2. Comparison of EnsembleRegNet performance with the existing method using GSE60361 datasets. The best score values are bold-faced.

EnsembleRegNet_GSE115978 | SIGNET_GSE115978 | SCENIC_GSE115978



	Measure	EnsembleRegNet	SIGNET	SCENIC
External clustering validation	the adjusted rand index (ARI)	**0.022**	0.0121	0.007
	normalized mutual information (NMI)	**0.031**	0.024	0.019
	F-score	**0.113**	0.095	0.0189
	Normalized Van Dongen (NVD) score	**0.748**	0.754	0.893
	Pair Sets Index (PSI) score	**0.515**	0.511	0.507
Internal clustering validation	RMSSTD	**0.126**	0.175	0.186
	Modified Hubert Γ statistic	5680.086	9383.478	**9553.364**
	Silhouette index	**0.4304**	0.412	0.417
	R-squared index	**0.924**	0.9123	0.903

Table 3. Comparison of EnsembleRegNet performance with the existing method using GSE115978 datasets. The best score values are bold-faced.

Network Visualization

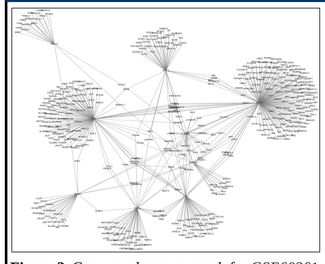


Figure 2. Gene regulatory network for GSE60361 dataset.

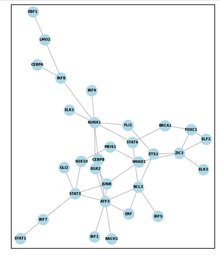


Figure 3. Transcription factor gene regulatory network for GSE60361 dataset.

Conclusion

- EnsembleRegNet achieved more compact and accurate cell classification than widely used methods, such as SCENIC and SIGNET.
- EnsembleRegNet is sensitive to important regulatory modules during various biological processes, especially in rare cells, where new biology is more likely to be discovered with single-cell data.

Acknowledgments

We gratefully acknowledge LBRN's support in providing the computational resources.