Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Systems Biology

Unraveling Transcription Factor Regulatory Networks in ALS: A Graph Neural Network Approach

Naufa Amirani[1,*] and Elizabeth Chang[1,*]

[1]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Building on an existing model for imputing TF-TF pairs using scATAC-seq data, DeepTFni, we expand its methods to include scRNA-seq data to infer TF-TF and TF-target gene relationships in a C9orf72-mediated ALS dataset.

Results: After modifying DeepTFni to include scRNA-seq and TF-target gene relationships, we found that the original scATAC-seq model generally remained the best performer – perhaps due to increasing model complexity. Visualization of our predicted networks showed that the models could consistently predict a TF potentially involved in neurodegenerative pathways for ALS patients, as well as non-neurodegenerative TFs for control subjects.

Contact: naufa.amirani@columbia.edu, ec3055@cumc.columbia.edu

Supplementary information: Supplementary data and code are available in the attached files and at: https://github.com/ec3055/CBMF4761

1 Introduction

Gene regulatory networks and transcription factor regulatory networks (TRNs) can offer valuable insights into the mechanisms underlying complex diseases, such as the neurodegenerative disorder amyotrophic lateral sclerosis (ALS). DeepTFni is a variational graph autoencoder (VGAE) based model that infers TRNs from scATAC-seq data, allowing it to impute transcription factor (TF) pairs (Li *et al.* 2022). In this study, we incorporate scRNA-seq data and TF-target gene relationships to expand the model. By introducing gene node types, we aim to create a more heterogeneous model capable of representing more complicated biological networks. Furthermore, we evaluate the performance of DeepTFni and our modified models using sequencing data obtained from C9orf72-mediated ALS patients to investigate potential TFs and regulatory interactions involved in the disease's pathogenesis.

2 Methods

We are expanding DeepTFni and generating three different model architectures for comparison. The first concatenates scATAC-seq with scRNA-seq (RNASEQ), the second incorporates TF-target gene relationships (GENE), and the third implements both expansions simultaneously (COMBO). The original scATAC-seq-only model (SOLO) serves as our baseline.

The dataset used in this study consists of sequencing data derived from motor cortex samples collected from C9-ALS (n=6) and control (CTR) donors (n=6) (Li *et al.* 2023). We included only cells with both scATAC-seq and scRNA-seq data before aggregating per cohort. Within these cohorts, we focused on two different cell types: excitatory neurons (ALS=3574, CTR=2870), as ALS predominantly affects motor neurons, and astrocytes (ALS=1693, CTR=2017), which are involved in neurodegeneration-related neuroinflammation processes.

scATAC-seq processing: Due to sparsity, chromosomal coordinates in the scATAC-seq data were binned in 1000bp windows with counts aggregated across patients per cohort for each cell type. Within each category, we randomly sampled 10,000 peaks and 250 cells, excluding sex chromosomes. The data was formatted in a P x C matrix, where P is peaks and C is cells. We removed peaks detected in less than 10% of cells. Using the hg19 genome reference, we then built a fasta file from the remaining peaks using the chromosome number and coordinates. TF binding site information and motifs were provided by the HOCOMOCO v11 database and scanned against the fasta file with Fimo (Kulakovskiy *et al.* 2018; Grant, Bailey and Noble 2011). If a detected TF motif had a p-value score > 1e-6, it was saved along with its location. These motifs are compared against the annotated Gencode v19 for TF promoter detection, where interaction is defined as the presence of one TF in the promoter (TSS ± 2 kb) of another TF (Frankish *et al.* 2021). A TF x TF adjacency matrix is generated where interaction is scored as 1 and no interaction as 0. The R Seurat library converts the original scATAC-seq peaks file into a dense matrix 10X HDF5 file. From this, Maestro is used to generate regulatory potential (RP) scores for an HDF5 gene score file (Wang *et al.* 2020). This transformation assumes that a scATAC-seq peak has an independent and additive effect on a given gene's expression and represents the accumulation regulation of peaks on a given gene in a given cell (Qin *et al.* 2020). Finally, the Maestro file is combined with the list of adjacency matrix TFs to generate a TF RP score matrix, forming the node feature input to the VGAE.

scRNA-seq processing: The gene expression counts for the 250 cells sampled in the scATAC-seq data were extracted from the scRNA-seq data and organized into a G x C matrix, where G is genes and C is cells. The addition of scRNA-seq data occurs after the generation of the TF RP score matrix for the scATAC-seq data, just before introducing the matrix as input to the model. This involves G row selection to match the TF RP score matrix rows and then concatenated along the C axis. This serves to augment known TF information in the node feature.

TF-target gene processing: We defined TF-target gene relationships according to hTFtarget, a comprehensive database identifying human TF-target regulations (Zhang *et al.* 2020). We limited TFs included in this study to those in the hTFtarget and HOCOMOCO databases and with known links to genes expressed in our dataset, resulting in 209 TFs. These linkages were formatted in a binary T x G matrix, where T is TF and G is the target gene. A value of 1 indicates a known relationship, while 0 indicates no relationship. Due to concerns about excessive inclusion of target genes saturating our predictions, we only kept target genes that interacted with at least half of the final TFs that comprise the initial adjacency matrix. These rows of G are then appended to the initial adjacency matrix horizontally and vertically to retain original symmetry, with G-G interactions set to 0 to focus on studying directed T-G relationships.

The encoder of the VGAE, a two-layer graph convolution network, takes this as input along with the initial adjacency matrix (our TRN skeleton) and produces latent representations for each TF node. The decoder reestablishes the representations to form the imputed TRN. To reduce any unwanted influence from the stochastic model components, k-fold splitting was implemented (k=10) and multiple runs (n=20) were instantiated with random initialization values, the average output of which is taken as the collective final imputed TRN.

Given that we had more negative than positive links in the data, we implemented majority sampling by splitting the positive links of the initial adjacency matrix into k equal subsets. A single set of positive links and the same number of randomly sampled negative links formed the test set. 1/5 of the remaining positive links randomly paired with negative links formed the validation set and all other links were used as the training set. The final matrices were evaluated on three parameters:

(1) Accuracy: number of correctly predicted links out of all links.
(2) Precision: fraction of true positive links among all links predicted as positive.
(3) Recall: fraction of correctly retrieved positive links among all positive links

For the following analyses, we separate our data into two categories: ALS vs CTR patients. We then further subset by cell type: Astrocyte vs Excitatory neurons.

3 Results

For our excitatory neuron data, we find that the RNASEQ model performs the best, with an average accuracy score of 85.77% (Table 1). For our astrocyte data, we observe the SOLO model performs the best with an average accuracy score of 83.16% (Table 2). Overall, there does not appear to be meaningful differences between the accuracies of ALS and CTR data. Across all models, there was a strong propensity to accurately predict no interaction (true negatives). Within the incorrect predictions, there is a trend of predicting connections where there are none (false positives) that increase proportionately to model complexity (SOLO < RNASEQ < GENE < COMBO). Indeed, across Tables 1 and 2 we observe recall is generally higher than precision. This indicates that the introduced complexity makes the overall input too noisy for the current model, or perhaps the current VGAE architecture needs to be further modified to accommodate the new data.

Table 1. Final results of the DeepTFni models for excitatory neurons

No	Model architecture	Accuracy	Precision	Recall
1	ALS-SOLO	86.67%	45.45%	60.00%
2	ALS-RNASEQ	90.48%	61.90%	52.00%
3	ALS-GENE	87.02%	17.36%	11.80%
4	ALS-COMBO	87.02%	17.36%	11.80%
1	CTR-SOLO	89.01%	48.39%	78.95%
2	CTR-RNASEQ	92.31%	61.90%	68.42%
3	CTR-GENE	76.19%	28.81%	68.00%
4	CTR-COMBO	70.48%	23.94%	68.00%

Table 2. Final results of the DeepTFni models for astrocytes

No	Model architecture	Accuracy	Precision	Recall
1	ALS-SOLO	83.16%	41.30%	79.17%
2	ALS-RNASEQ	73.95%	29.27%	75.00%
3	ALS-GENE	68.12%	22.92%	61.11%
4	ALS-COMBO	61.41%	18.67%	58.33%
1	CTR-SOLO	86.26%	49.41%	91.30%
2	CTR-RNASEQ	79.24%	37.62%	82.61%
3	CTR-GENE	67.67%	21.36%	57.89%
4	CTR-COMBO	60.67%	17.74%	57.89%

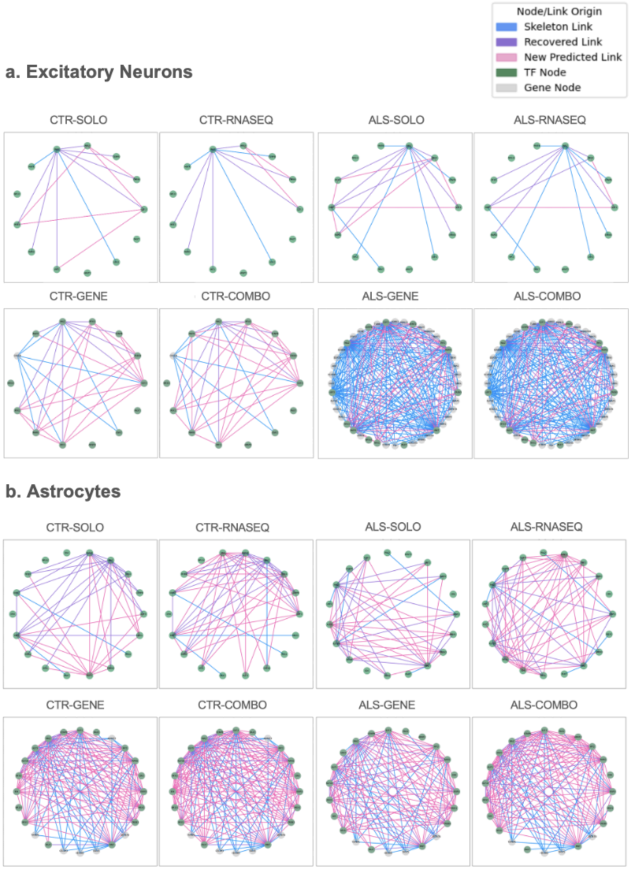

Fig. 1. Neural network graphs generated by each model. TF nodes (green) are present in all four model types with the addition of gene nodes (gray) in GENE/COMBO. The output of each model is an imputed TRN that recovers the links present in the initial TRN skeleton (purple), fails to recapitulate them (blue), or predicts new ones (pink). **a:** Networks for excitatory neurons in CTR vs ALS. **b:** Networks for astrocytes in CTR vs ALS.

We observe a general sparsity of detected TFs and connections across all models for our excitatory neuron data (Fig 1a). Our ALS data mirrors the results of the CTR data for the SOLO and RNASEQ graphs but we observe robust GENE and COMBO graphs, suggesting that the addition of target gene information boosts the model's ability to predict new connections. When comparing the proportion of recovered, predicted, and lost connections to the TRN skeleton for each network generated, we see that RNASEQ tends to predict fewer new connections than SOLO but is adept at recovering skeletal links. Conversely, GENE and COMBO predict more new links but recover fewer skeletal links. We see this particularly pronounced in ALS, where although the model fails to recover many TF-gene connections from the skeleton, it predicts many TF-TF links. This discrepancy is likely because of the large number of genes detected by the model for ALS, which is consistent with the expected changes in motor neuron activity associated with the disease. For our astrocyte data, we find that ALS and CTR results are similar and, like the ALS excitatory data, generate many connections when target genes are included (Fig 1b).

Within the true positives of the data, a consistently predicted TF in our ALS data is MAZ. For our CTR data, we found that E2F3, CREB, RFX3, RARG, IRF2, and RFX1 were consistently predicted. MAZ has been previously associated with Alzheimer's disease, implicating its involvement in neurodegenerative pathways (Jordan-Sciutto *et al.* 2000). The TFs identified in our CTR data are involved in a variety of functions – including cell cycle regulation, embryonic development, and immune response – which aligns with our expectations for healthy patients.

4 Discussion and conclusions

In this study, we applied DeepTFni to a C9orf72-mediated ALS dataset. We used the original scATAC-seq model as our baseline and subsequently enhanced it with the inclusion of scRNA-seq, TF-target gene relationships, and both. We found that while RNASEQ was the best performer for excitatory neuron data, SOLO performed the best in astrocyte data. Evaluation of the GENE and COMBO graphs revealed consistent patterns within each cell type for each cohort, suggesting that the addition of TF-target gene relationships alone is sufficient to introduce gene node types to the network, with only marginal improvements from integrating scRNA-seq as well. We visualized the predicted networks and found that they could consistently predict a TF potentially involved in neurodegeneration in the ALS cohort, as well as non-neurodegenerative TFs in the CTR cohort. Furthermore, these models appear to perform better at recovering TF-TF links than TF-gene links across both cell types. This is particularly evident in ALS excitatory neurons, supporting the need for further refinement of the VGAE architecture to improve TF-target gene relationship inference.

Future directions of this work include expanding our analysis to include additional cell types – such as microglia and oligodendrocytes – that have previously been shown to be affected by ALS. Additionally, incorporating sequencing data from frontal cortex samples and a cohort of C9orf72-mediated frontotemporal dementia patients may further enrich our understanding of neurodegenerative diseases. Enhancing the graph network by adding directional information would also allow the model to reflect the nature of regulatory relationships. Overall, this study lays the groundwork for integrating scRNA-seq and TF-target relationships into DeepTFni, highlighting the potential of multi omics data in deciphering the regulatory landscape underlying complex diseases.

Acknowledgements

We extend our gratitude to Professor Itsik Pe'er and our Teaching Assistants, Stella Park and Christina Floristean, for their guidance and feedback on this project. Additionally, we appreciate their continuous support and efforts in teaching Computational Genomics throughout this semester.

Funding

No funding was received for this work.

Conflict of Interest: none declared.

References

Frankish A, Diekhans M, Jungreis I *et al.* GENCODE 2021. *Nucleic Acids Research* 2021;**49**:D916–23.

Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;**27**:1017–8.

Jordan-Sciutto KL, Dragich JM, Caltagarone J *et al.* Fetal Alz-50 Clone 1 (FAC1) Protein Interacts with the Myc-Associated Zinc Finger Protein (ZF87/MAZ) and Alters Its Transcriptional Activity. *Biochemistry* 2000;**39**:3206–15.

Kulakovskiy IV, Vorontsov IE, Yevshin IS *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research* 2018;**46**:D252–9.

Li H, Sun Y, Hong H *et al.* Inferring transcription factor regulatory networks from single-cell ATAC-seq data based on graph neural networks. *Nat Mach Intell* 2022;**4**:389–400.

Li J, Jaiswal MK, Chien J-F *et al.* Divergent single cell transcriptome and epigenome alterations in ALS and FTD patients with C9orf72 mutation. *Nat Commun* 2023;**14**:5714.

Qin Q, Fan J, Zheng R *et al.* Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biology* 2020;**21**:32.

Wang C, Sun D, Huang X *et al.* Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biology* 2020;**21**:198.

Zhang Q, Liu W, Zhang H-M *et al.* hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. *Genomics, Proteomics & Bioinformatics* 2020;**18**:120–8.