GNE: A deep learning framework for gene network inference by aggregating biological information

Kishan K C (kk3671@rit.edu)[1] Rui Li[1] Feng Cui[2] Qi Yu[1] Anne R. Haake[1]

[1]Golisano College of Computing and Information Sciences [2]Thomas H. Gosnell School of Life Sciences

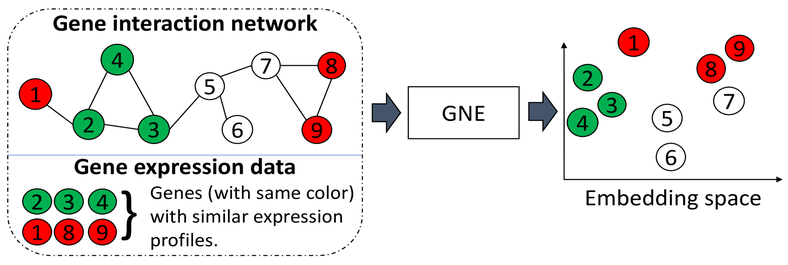
Rochester Institute of Technology, Rochester, New York, USA

Overview

- Understanding functional aspects of genes or proteins is crucial to provide insights into underlying biological phenomena for different health and disease conditions.
- Often intractable through biological experiments.
- Topological landscape of gene interactions provides the support for understanding such phenomena.
- Sparse connectivity between the genes
- We propose ***Gene Network Embedding (GNE)***, a deep neural network architecture to learn lower dimensional representation for each gene, by integrating the topological properties of gene interaction network with additional information such as expression data.
- Outperforms strong baselines.

Gene Network Embedding (GNE)

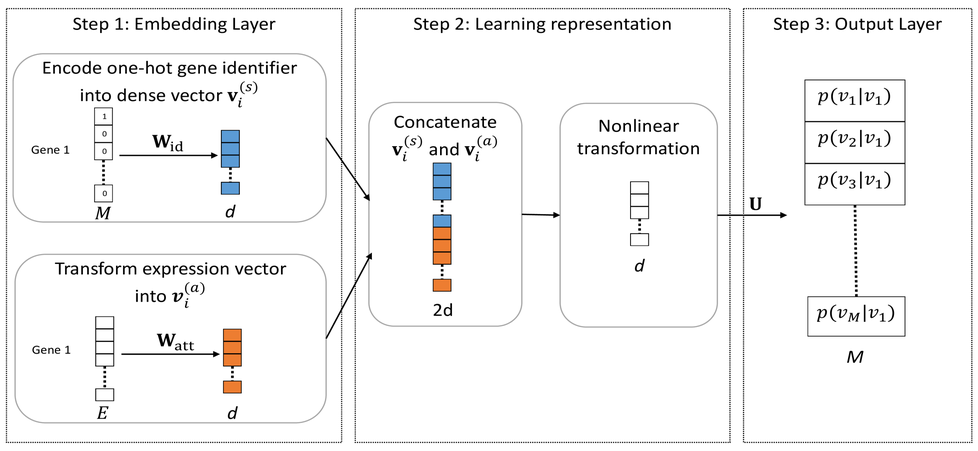
Given a gene network denoted as $\boldsymbol{G = (V, E, A)}$, gene network embedding aims to learn a function $\boldsymbol{f}$ that maps gene network structure and their attribute information to a $\boldsymbol{d}$-dimensional space where a gene is represented by a vector $\boldsymbol{y_i \epsilon \mathbb{R}^d}$ where $\boldsymbol{d \ll M}$. The low dimensional vectors $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$ for genes $\boldsymbol{v_i}$ and $\boldsymbol{v_j}$ preserve their relationships in terms of the network topological structure and attribute proximity.



- Node represents a gene and edges represent the interactions with other genes.

Overview of GNE

- Models the complex statistical relationship between ***topological properties*** and ***expression data*** via nonlinear transformation of fused representation.



Assuming a gene network with expression data as node attributes:

1. Obtain dense representation $\boldsymbol{v_i^{(s)}}$ of topological properties of a gene through topological encoder
2. Obtain expression representation $\boldsymbol{v_i^{(a)}}$ of a gene by passing expression data through expression encoder
3. Get the joint representation $\boldsymbol{v_i^{(s)} + \lambda v_i^{(a)}}$
4. Transform the representations using nonlinear layers
5. Predict probability of interactions
6. Update the parameters of encoders, hidden layers by applying gradient descent to optimize maximum likelihood loss

Acknowledgements

This material is based upon work supported by National Science under Grant NSF-1062422.

github.com/kckishan/GNE Access paper from bioRxiv kishan_kc07

Quantitative results

Result on interaction prediction

- AUROC comparison shows that GNE outperforms other strong baselines.

Methods	Yeast		E. coli	
	AUROC	AUPR	AUROC	AUPR
Correlation	0.582	0.579	0.537	0.557
Isomap	0.507	0.588	0.559	0.672
LINE	0.726	0.686	0.897	0.851
node2vec	0.739	0.708	0.912	0.862
Isomap+	0.653	0.652	0.644	0.649
LINE+	0.745	0.713	0.899	0.856
node2vec+	0.751	0.716	0.871	0.826
GNE (topology only)	0.787	0.784	0.930	0.931
GNE	**0.825**	**0.821**	**0.940**	**0.939**

Temporal Holdout Validation

- Temporal holdout validation with two versions of interaction data: 2017 and 2018 version
- Model trained on 2017 version and tested on 2018 version

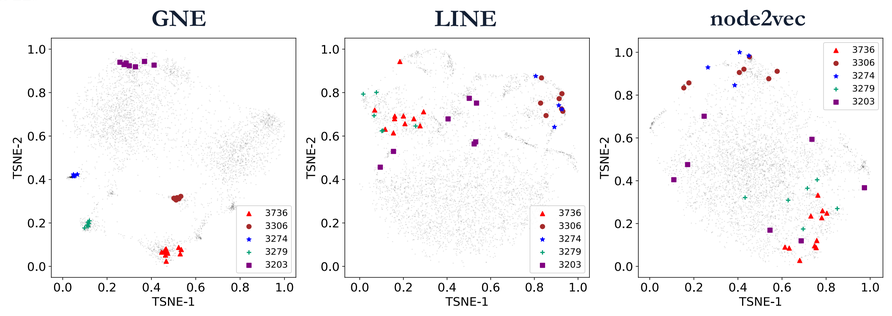
Methods	**Yeast**		**E. coli**	
	AUROC	**AUPR**	**AUROC**	**AUPR**
LINE	0.620	0.611	0.569	0.598
node2vec	0.640	0.609	0.587	0.599
GNE	**0.710**	**0.683**	**0.653**	**0.658**

- Predicts gene interactions more accurately
- Integration of expression data improves the interaction prediction.

Dataset	**Probability**		**Gene i**	**Gene j**	**Experimental Evidence code**
	Topology	**Topology + Expression**			
Yeast	0.287	0.677	TFC8	DHH1	Affinity Capture-RNA
	0.394	0.730	SYH1	DHH1	Affinity Capture-RNA
	0.413	0.746	CPR7	DHH1	Affinity Capture-RNA
E. coli	0.014	0.944	ATPB	RFBC	Affinity Capture-MS
	0.012	0.941	NARQ	CYDB	Affinity Capture-MS
	0.013	0.937	PCNB	PAND	Affinity Capture-MS

Qualitative results

- Learned embeddings is projected into 2D space using t-SNE package for visualization.



- Our method learns similar representation for genes within same operon.
- Two-sample KS test shows that genes within the same operon have significantly similar vector representation than expected by chance.

Sensitivity Analysis

- Integration of expression data with topological properties improves the performance.
- No significant improvement when number of (training) interactions increases (> 50%).