A deep framework for aggregating heterogeneous biological information for gene network inference

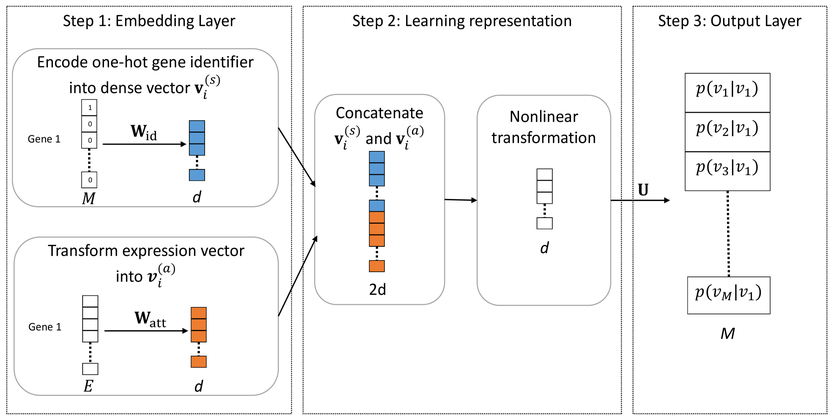
Kishan K C (kk3671@rit.edu)[1] Rui Li[1] Feng Cui[2] Anne R. Haake[1]

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

Abstract

- 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**.
- Models **complex statistical relationships** between topological patterns and gene expression, which addresses the problem of sparsity in interaction data.

GNE Architecture



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

Results on Interaction Prediction

- AUROC comparison shows that GNE outperforms other strong baselines.

Methods	Yeast		E. coli	
	AUROC	AUPR	AUROC	AUPR
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
GNE*	0.787	0.784	0.930	0.931
GNE	0.825	0.821	0.940	0.939

- Temporal holdout validation

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**

Datasets

- Gene interaction data from BioGRID interaction database and gene expression data from DREAM5 challenge.

	Interaction Network Data		Expression data
Organism	# Genes	# Interactions	# Experiments
Yeast	5,950	544,652	536
E. coli	4,511	148,340	805

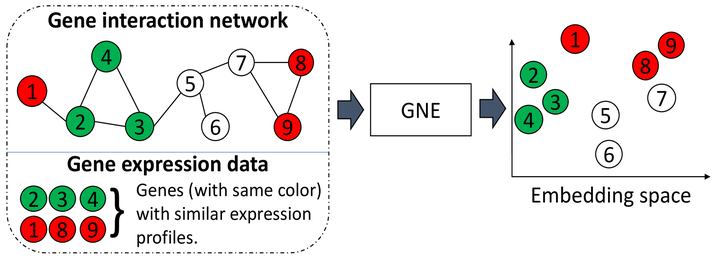
- Operons dataset from DOOR database.

Investigation of GNE's predictions

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

GNE Overview

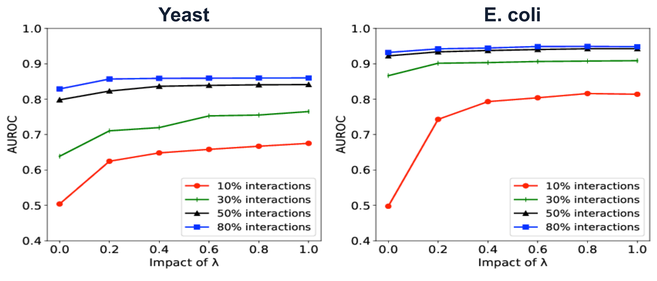
Given a gene network denoted as $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{A})$, gene network embedding aims to learn a function $\boldsymbol{f}$ that maps gene network structure and their attribute information to a $\boldsymbol{d}$-dimensional space where a gene is represented by a vector $\boldsymbol{y_i} \boldsymbol{\epsilon} \mathbb{R}^{\boldsymbol{d}}$ where $\boldsymbol{d} \ll \boldsymbol{M}$. 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.



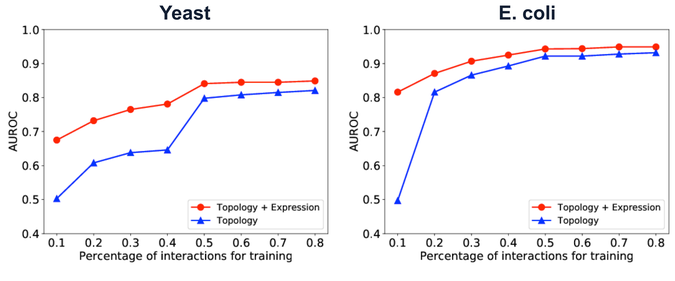
Node represents a gene and edges represent the interactions with other genes.

Sensitivity Analysis

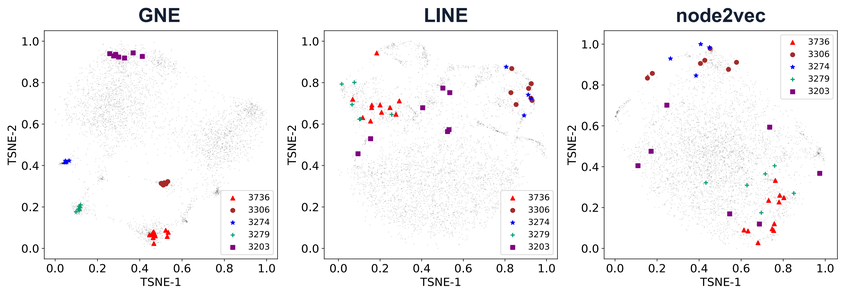
- Relative importance of topology and expression data



- Impact of network sparsity



Visualization of Embeddings



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

kishankc

kishankc.com.np