Summer Review 3

Towards Gene Expression Convolutions using Gene Interaction Graphs

School of Engineering and Applied Sciences Harvard University
Cambridge, MA, USA

Reviewed by : Arshdeep Sekhon

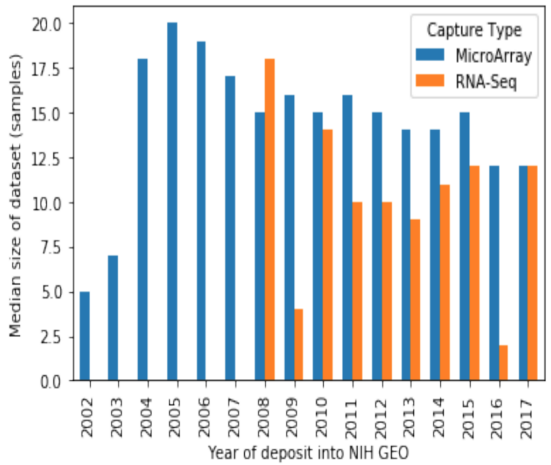
[1]Department of Computer Science, University of Virginia
https://qdata.github.io/deep2Read/

GEO database

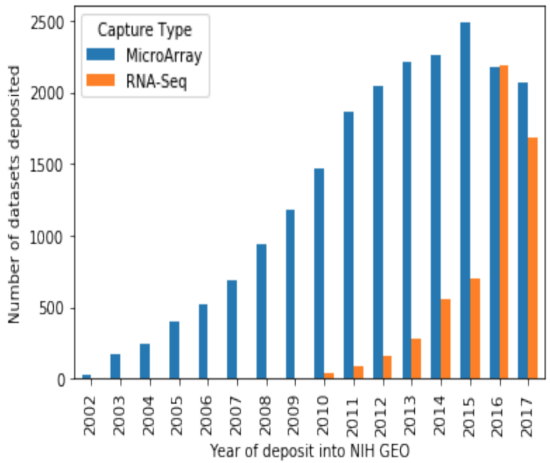
DataSet	Title	Organism(s)	Platform	Series	Samples
GDS6248	Diet-induced obesity model: liver	*Mus musculus*	GPL6887	GSE39549	51
GDS6247	Diet-induced obesity model: white adipose tissue	*Mus musculus*	GPL6887	GSE39549	40
GDS6177	Acute alcohol consumption effect on whole blood (control group): time course	*Homo sapiens*	GPL570	GSE20489	25
GDS6176	Caspase-1 deficiency effect on lipid-loaded intestines	*Mus musculus*	GPL11533	GSE32515	18
GDS6100	MicroRNA-135b overexpression effect on prostate cancer cell line: time course	*Homo sapiens*	GPL10558	GSE57820	12
GDS6083	Chronic lymphocytic leukemia cells response to the neutralization of inhibitor of apoptosis proteins	*Homo sapiens*	GPL570	GSE62533	12
GDS6082	Sendai virus infection effect on monocytic cell line: dose response	*Homo sapiens*	GPL10558	GSE67198	11
GDS6064	Arthritic tarsal joints induced by collagen: time course	*Mus musculus*	GPL6246	GSE61140	15
GDS6063	Influenza A effect on plasmacytoid dendritic cells	*Homo sapiens*	GPL10558	GSE68849	10
GDS6016	Transcription factor engrailed-2 loss-of-function model of autism spectrum disorder: hippocampus	*Mus musculus*	GPL7202	GSE51612	6

Figure: GEO data

(a) Median dataset size

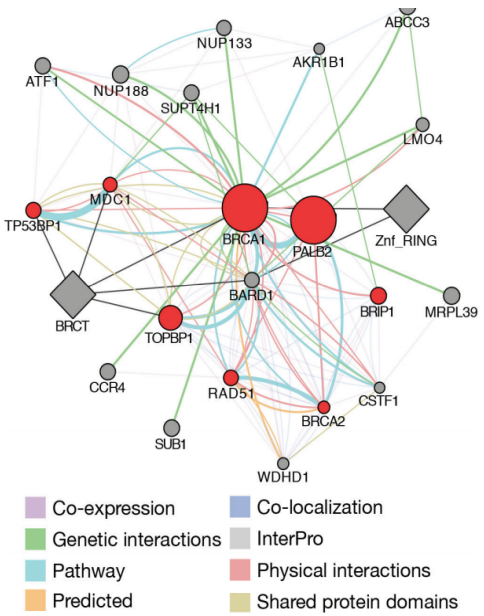


(b) Number of datasets added



Prior Information: GeneMANIA

- GeneMANIA uses a database of organism-specific weighted networks to construct the resulting composite network.
- The database includes over 1800 networks, containing over 500 million interactions for 8 organisms
- The networks are organized into groups such as co-expression, where edges are derived from expression profiles, and shared protein domains, where edges represent genes that encode proteins with similar domains.



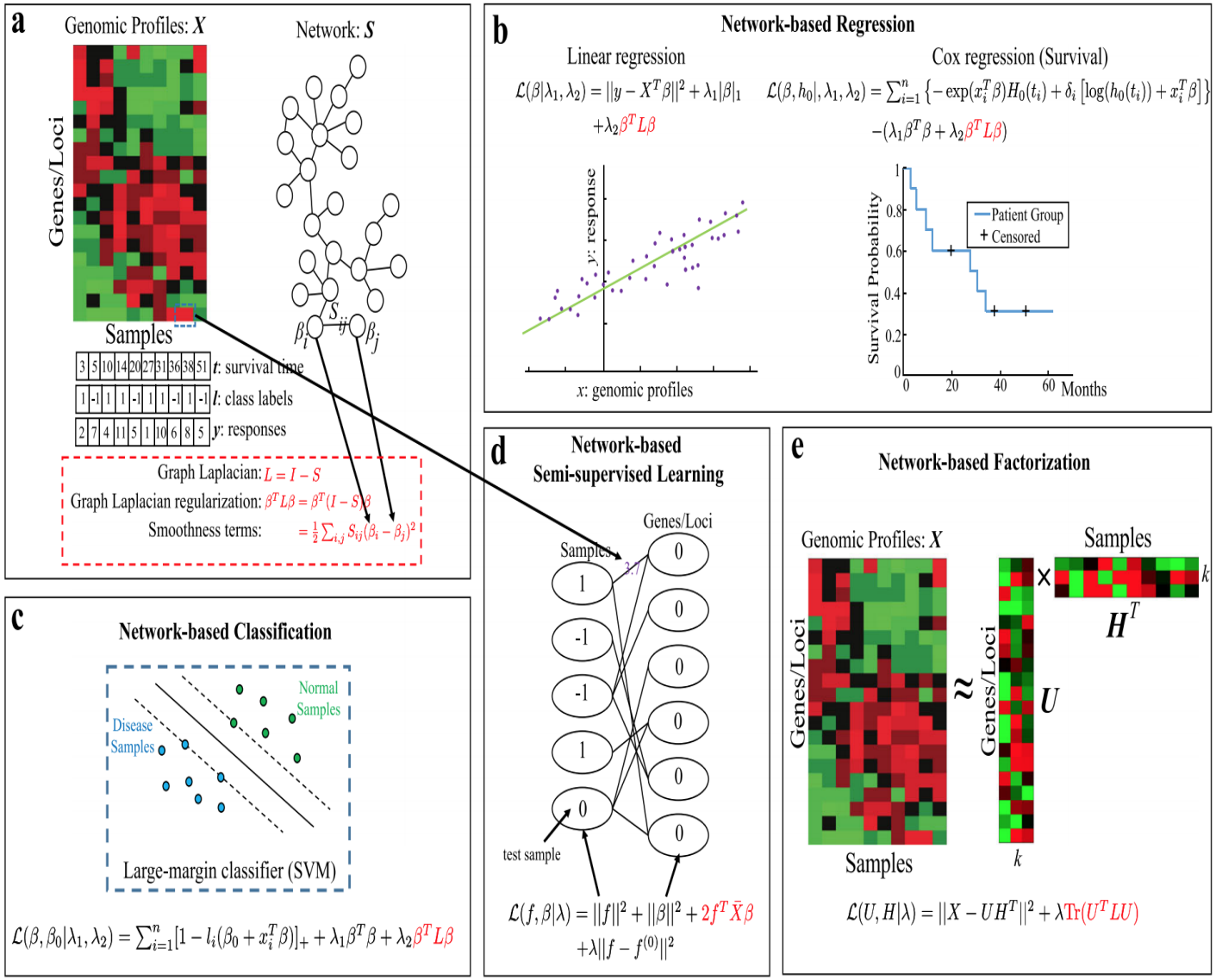


Fig. 3 Model-based integration of whole-genomic profiles and a molecular network. **a** The patient genomic profiles ***X*** along with the clinical information: the survival time, two patient subgroups for classification and treatment response of each individual patient are shown. The network ***S*** is typically integrated into the genomic profile analysis with a graph Laplacian regularization. The formulas of the graph Laplacian and its regularization are shown below. The graph Laplacian regularization can be rewritten as summation of pairwise smoothness terms that promote smoothness among the connected genomic features in the network. **b** The network-based linear regression and Cox regression models are illustrated in the figure with the graph Laplacian regularization term added to the original cost functions. **c** Network-based

Gene Graph Convolutions

- Graph Convolution usually used where data is in the form of graphs: citation networks, etc
- Graphs complementary to the main task in Gene Expression
- Use graphs to bias the model
- with low number of samples, known relationships between variables can avoid spurious relationships.

Background: Graph Laplacian

- The graph Laplacian regularization is a summation of smoothness terms on the variables to encourage similar coefficients on the genes or other genomic features that are connected in the network

Previous Work: Network *Regularized* Sparse Logistic Regression

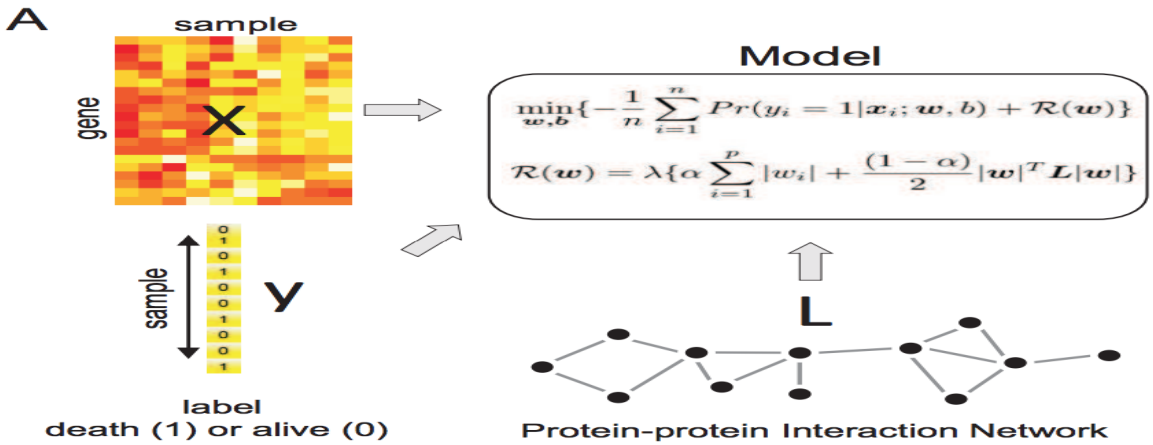


Figure: NSLR

L is the normalized Laplacian matrix encoding a prior network[1]

$$\lambda\sum_{i=1}^{p}|w_i| + \eta\sum_{i\sim j}A_{ij}\left(\frac{\text{sign}(\hat{w}_i)w_i}{\sqrt{d_i}} - \frac{\text{sign}(\hat{w}_j)w_j}{\sqrt{d_j}}\right)^2$$

Graph Convolution

- Extract information from the neighbor nodes in a graph
- Graph convolutions are generalisation of convolutions, and easiest to define in spectral domain
- euclidean vs non euclidean data

Graph convolution

- Laplacian $L = D - A$
- D: degree matrix, A : adjacency matrix
- for convolution, euclidean shift-invariant definition not applicable since the structure isnt shift-invariant
- use the spectral definition (Convolution is element-wise multiplication in the Fourier domain)
- $L^{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$

Graph Convolution

- Consider $X^{l+1} = \sigma(AX^l\theta^l)$
- Issues:
  - sum up all the feature vectors of all neighboring nodes but not the node itself : $\hat{A} = A + I$
  - A is typically not normalized and therefore the multiplication with A will completely change the scale of the feature vectors[2]
  - Symmetric Normalization: $\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}$

[2] https://tkipf.github.io/graph-convolutional-networks/

Graph Convolution

$$\theta \star X^l \approx D'^{-1/2} A' D'^{-1/2} X^l \theta = \tilde{A} X^l \theta \tag{1}$$

- l is layer, n is the number of nodes, o is output feature size, c is input feature size
- $A' = A + I_N$ A is adjacency matrix
- $D'_i = \sum_j A'_{ij}$
- where $X^l \in \mathbf{R}^{n \times c}$
- $\tilde{A} \in \mathbf{R}^{n \times n}$
- $\theta \in \mathbf{R}^{c \times o}$

Graph Convolution for gene networks

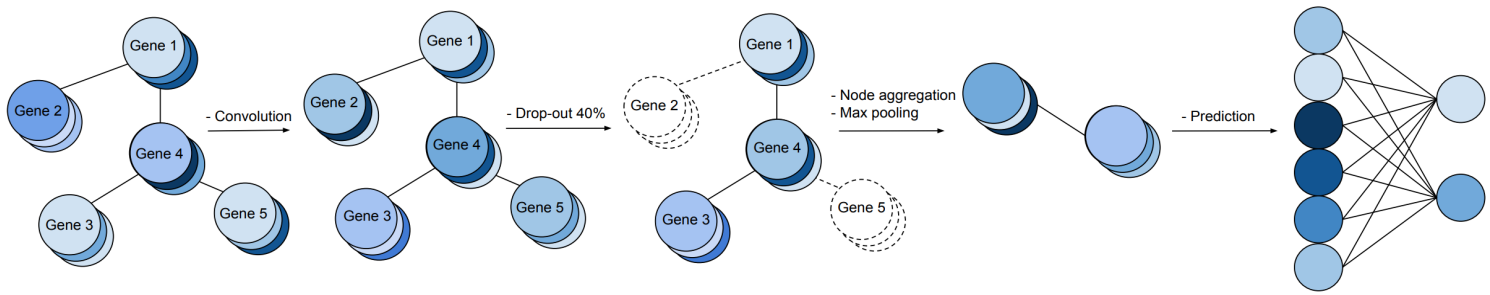


Figure 2. An overview of the Graph Convolutional Network applied to gene expression data. At first each gene is embedded in a graph where neighbors are extracted from prior biological knowledge. After each convolution, the genes are aggregated together based on their connectivity. Finally, a prediction is made from the remaining nodes.

Figure: GCN

- not possible to have multiple types of interactions: all nodes are aggregated before any transformation
- possible to have different sets of parameters for different types of interactions
- But, genes do not have such interactions.

GCN

- add skip connection at each conv layer: add neighborhood as well as node itself.

$$X^l = Aggregate(\sigma(\tilde{A}X^l\theta_1 + X^l\theta_2)) \quad (2)$$

- *Aggregate* is hierarchical clustering.
- Dropout encourages the model to spread information across all nodes and not rely on a single node.

Experiment 1: PANCAN

- $10,459$ RNA Seq samples from TCGA
- 16300 genes for each sample
- Each sample has some cancer subtype or healthy label. (Actual labels)
- Option 1: But, DNNs don't work on this setting of actual labels.
- Option 2: Use a smal subset of genes relevant for cancer subtype detection or trait, etc.
- Option 2 doesn't work because
  - No assumption about which gene is relevant.
  - "we cannot guarantee any complex relationship is important to solve the task".
- Option 3: Select a specific gene and binarize the label: predict using the rest of the genes.

Experiment : PANCAN

Figure 3. This plot illustrates the construction of the single gene inference task by adding nodes based on their distance from the node we want to predict. This graph is for the gene S100A8. The higher the distance the lighter the color.

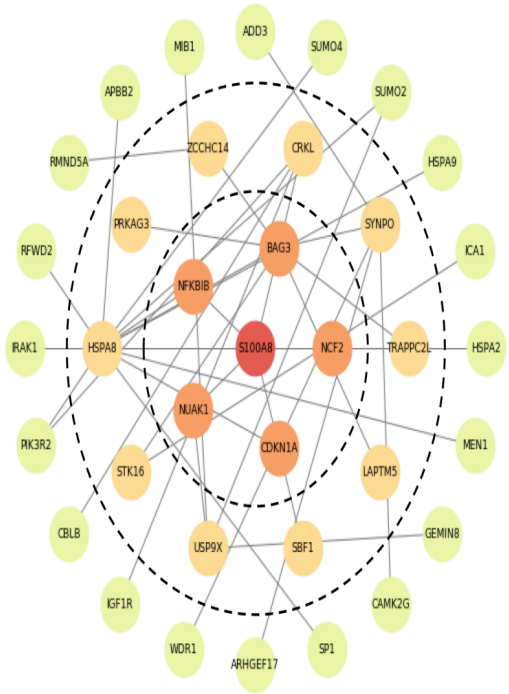


Figure: Adding genes based on distance

- predict $+1/-1$ by using the expression values of closest neighbors
- add more successively till all covered

Experiment: Add graph information

Use GeneMANIA and RegNetwork: two types of public database graphs.

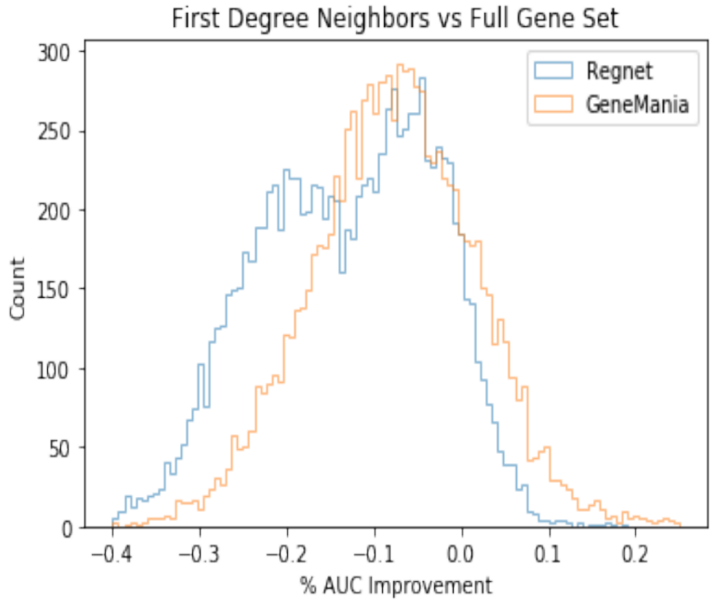


Figure 4. For each graph we train two MLPs to predict each of the 16k genes. One uses all genes and the other uses only the first degree neighbors in the graph. We show the difference in AUC between the models. If a gene has no neighbors then the model predicts 50%. Genes with a %AUC improvement > 0 were better predicted when only considering the first degree neighbors.

Figure: Results

Results: Quality of graphs

- Genes with a $> 20\%$ AUC improvement are a minority.
- Gene Mania outperforms RegNetwork
- RegNetwork has twice as many edges per node
- Indicating simply merging does not lead to improvement.

Robustness to Noise

Fig in paper(Low resolution)