CancerLinker: Integrating Gene Expression For Pathway Analysis

Sol Taylor-Brill[1], Kathy Thompson[2], mentored by Anna Ritz[1]
Department of Biology[1], Department of Computer Science[2], Reed College

Abstract

Cancer often arises from dysregulated cell signaling pathways. We created a method to examine how the Wnt pathway, which controls cell proliferation, morphogenesis, and stem cell control, differs between healthy and cancerous tissues.

Methods

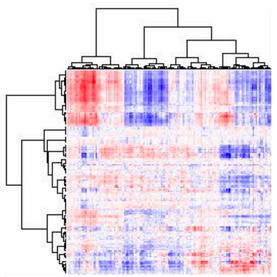
Using TCGA gene expression data we have developed a **new way of analyzing a cancer-dysregulated Wnt pathway** and its healthy counterpart based on PathLinker.

1. Developed program to **extract gene data** from GDC Commons (colorectal cancer data from COAD: 512 samples over 62804 genes) and organized into a matrix file.
2. **Measured gene expression** by calculating fold change for each gene
3. **Integrated gene expression data into PathLinker** to predict cancerous cell signaling pathways.

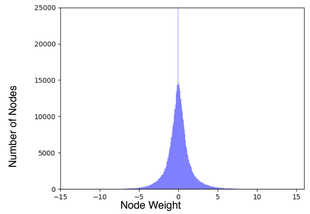
①

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

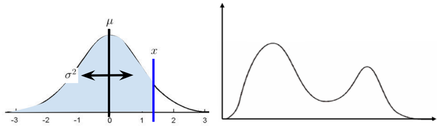
Gene expression data downloaded from GDC commons and organized into matrix file.



②



Fold-change (cancer/normal) data for each gene was log2 transformed to produce normalized distribution of node weights

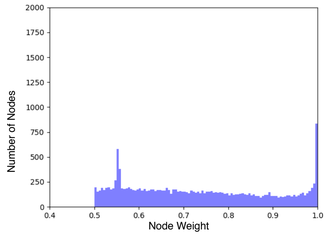


Empirical CDF

Convolution CDF

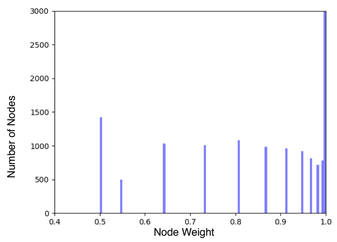
Example toy graphs illustrating the different natures of each transformation scheme

Empirical OR Convolution Transformation



Empirically Transformed

Each fold-change value calculated from a standard distribution, estimating mean and SD from under the curve



Convolution Transformed

Iteratively build a distribution by adding patients' gene values

Results

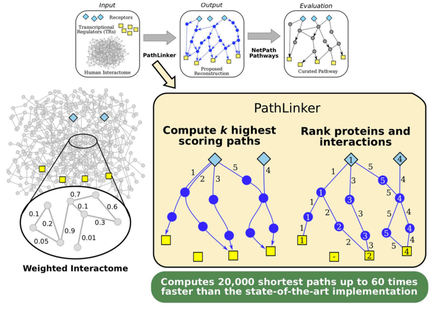
- Obtained benchmark showing our method accurately records likely-to be dysregulated proteins (catenin beta-1, low density lipoprotein receptor-related protein 6) as highly-connected
- There are more unique intermediate nodes in the convolution-transformed data (24) as compared to intermediate empirical nodes (13).
- All receptors shared between methods
- Most transcription factors conserved between methods (7/10 total)

③

$$\tilde{c}_{uv} = \beta c_{uv} + (1-\beta)c_u$$

Edge weights transformed

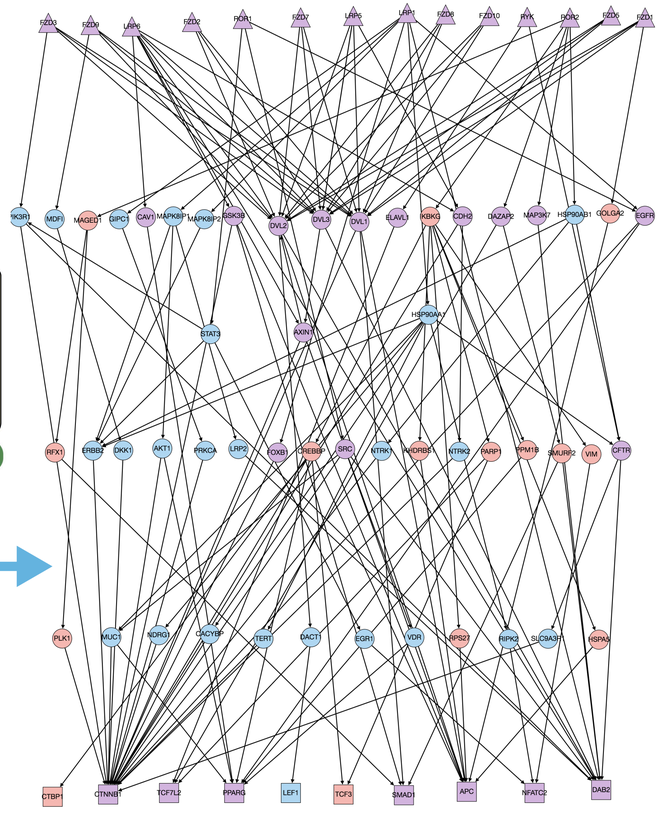
β: Tradeoff value for integrating gene expression
c_u: Cost of the individual gene (node)
c_{uv}: Edge cost from node u to node v



Run PathLinker for 100 paths on empirical and convolution transformation results

Conclusion

- In general, the proteins with the most edges leading to them were shared between both methods.
- Because the convolution model assumes less about the data distribution, it may discover more novel nodes than the empirical transformation.
- Both methods incorporated every receptor, and most highly ranked nodes in PathLinker with convolution, empirical, and the control pathways were conserved. This allows for future investigation into the differences between healthy and dysregulated paths.



Comparison of Empirical and Convolution Outputs. $\beta = 0.75$. Blue nodes are only in convolution pathways; pink nodes are only in empirical pathways; purple are shared between both. Triangles are receptors; squares are transcription factors; circles are intermediates.

Contact Information
- https://github.com/annaritz/cancer-linker
- soltb@reed.edu, katjthomp@reed.edu

References:

1. Ritz, Anna, et al."Pathways on Demand: Automated Reconstruction of Human Signaling Networks." Npj Systems Biology and Applications, vol. 2, no. 1, 2016.
2. GDC Commons (Grossman, Robert L., Heath, Allison P., Ferretti, Vincent, Varmus, Harold E., Lowy, Douglas R., Kibbe, Warren A., Staudt, Louis M. Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine375:12, 1109-1112), 2016.
3. GraphSpace (Aditya Bharadwaj, Divit P Singh, Anna Ritz, Allison N Tegge, Christopher L Poirel, Pavel Kraikivski, Neil Adames, Kurt Luther, Shiv D Kale, Jean Peccoud, John J Tyson, T M Murali; GraphSpace: stimulating interdisciplinary collaborations in network biology, Bioinformatics, Volume 33, Issue 19, 1 Pages 3134–3136), October 2017.

Acknowledgements

We would like to thank Anna Ritz, Ibrahim Youssef, and Usman Hafeez for their contributions and support, and the MJ Murdock Trust for funding this project.