In Silico Discovery of ACC Cancer Biomarkers: Applying Link Prediction to a Purpose-Built Hypergraph

Pieter Derdeyn, BA[1,2], Kendyl Douglas, BA[1], David Schneider[1], Chris Yoo, PhD[1,3]
[1]Systems Imagination, Inc, Tempe, Arizona, [2]University of California, Irvine, [3]Arizona State University, Tempe, Arizona

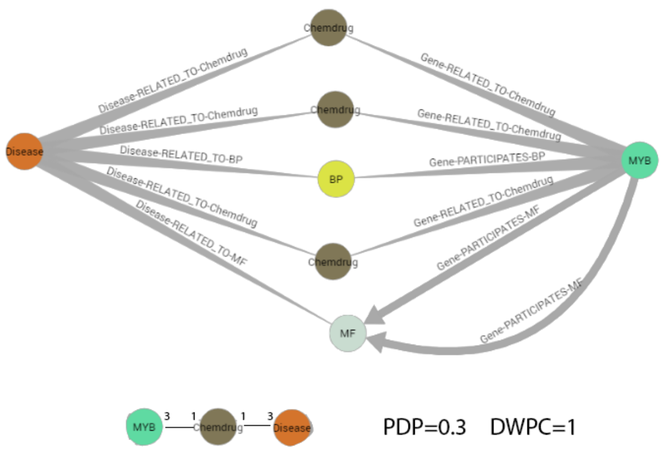
Abstract

Adenoid Cystic Carcinoma (ACC) is a rare but aggressive cancer. 89% of patients survive for the first 5 years, but only 40% survive for 15 years or longer. Since ACC is uncommon, application of high throughput big data analyses that have made progress in other cancers to ACC data has been challenging. We used a hypergraph database made up of data from dozens of state of the art cancer research efforts to identify unknown insights related to ACC. To generate high likelihood hypotheses, we have applied link prediction, pioneered in social network analyses to this purpose-built hypergraph. With this analysis, we were able to predict several high scoring gene fusions that may be implicated in ACC. We discuss the significance of this approach and suggest further research from the findings.

Introduction

Graphs have proven a useful structure for both representing and analyzing data in social networks and maps [4,5]. For example, Facebook uses link prediction to recommend friends while Netflix uses link prediction to recommend movies [1]. Here, we have used a purpose-built hypergraph, a generalization of a graph, to represent ACC tumor biology and organize disparate knowledge from diverse open source cancer research efforts (Figures 1). In doing so, we applied link prediction to generated previously unknown data driven hypotheses.

Figure 1: A selection of some of the largest data sources integrated into our hypergraph database



$$PDP(path) = \prod_{d \in D_{path}} d^{-w}$$

$$DWPC(metapath) = \sum_{path \in Paths} PDP(path)$$

Figure 2: A derivation of the degree weighted path count feature from Himmelstein et al (2)

Methodology

A modification of Himmelstein et al. [3] was implemented to enable link prediction to be applied to our very large data set with the following process:

1) Mine features relating to pairs of nodes and their connectivity
2) Use supervised learning to train a model that can classify whether or not these nodes are connected or not
3) Calculate degrees of nodes from source to target node
4) Calculate degree weighted path counts (DWPC) between nodes (see Figure 2)
5) Calculate prior likelihood of an edge between the nodes
6) Calculate probability score based on features in steps #3-5

Analysis and Results

The algorithms ran on our hypergraph database which contains over 700,000 nodes and 12 million+ edges to make up one of the largest representations of cancer-specific knowledge (Figure 3).

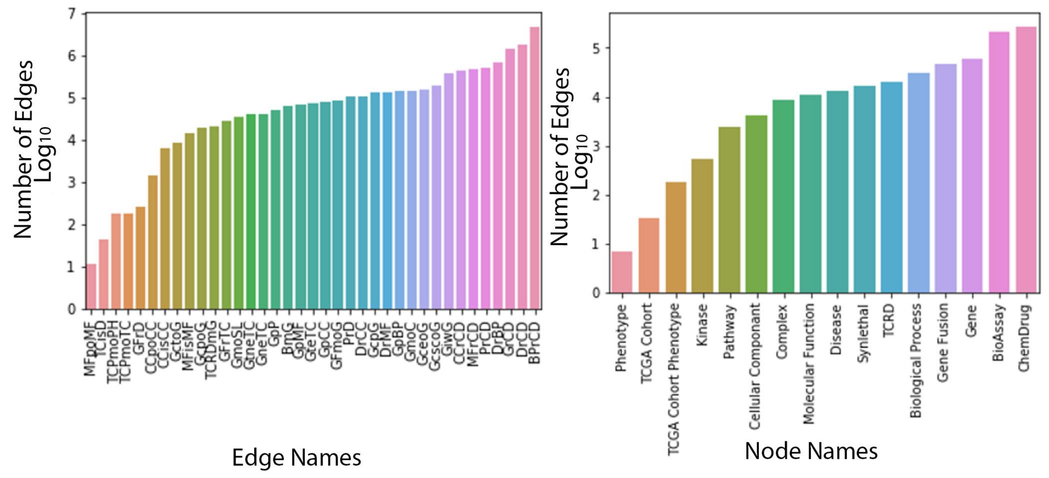


Figure 3: The number of nodes and edges in hypergraph database.

Analysis and Results

We generated features for all gene pairs that were previously reported in a database of gene fusions [2]. We then ran the feature matrix through multiple supervised learning algorithms including random forest, logistic regression, decision trees, and xg boost. We also built a neural network with mxnet and keras, which ended up having the best performance as measured by its true positive rate. Based on the neural network's superior performance, we ran the same neural network on gene pairs that were not present as pairs in the dataset of gene fusions used to develop the performance characteristics of the analyses. This resulted in a list of high likelihood predictions of novel or unreported gene fusion pairs, the top 10 of which is displayed in Table 1. One of the top 10 fusion pairs is represented in a simplified network diagram in Figure 4.

Table 1. The 10 highest likelihood predicted gene pairs

Gene 1	Gene 2	Probability of Gene Fusion
EWSR1	HMGA2	0.929894619
BBS9	KMT2A	0.928350421
IQCJ	KMT2A	0.927711269
CYP11B1	KMT2A	0.926647616
KMT2A	TCIRG1	0.912818918
KMT2A	VEPH1	0.911844923
CCR6	KMT2A	0.873986434
KCNQ1	KMT2A	0.834963505
ACSL1	KMT2A	0.834097153
EWSR1	RUNX1T1	0.832868597

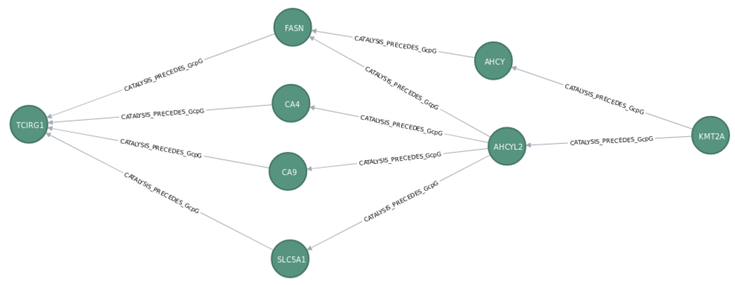


Figure 4: An example of a path between TCIRG1 and KMT2A, the fifth highest likely gene fusion prediction.

Conclusion and Discussion

This approach represents an expansion of the possibilities of potential fusion proteins important in ACC tumor biology. Since the known fusion proteins were used in a training set and removed from the test set, the analysis resulted in linked genes that resemble known fusion proteins in our network. The link prediction approach resulted in an overrepresentation of certain genes such as KMT2A and EWSR1. This is likely due to the fact that EWSR1 is often found in several fusion proteins in the literature and KMT2A is a transcription factor binding protein reported often in the literature. Basic literature searching of these two genes revealed potentially promising cancer-implicated functions that could be important in soft tissue tumors and thus ACC tumor biology. Also potentially significant HMGA2 is a transcriptional regulator and often an oncogene. These *in silico* results can now be used to validate potential hypotheses relevant for ACC tumor biology.

References

[1] Backstrom, Lars, and Jure Leskovec. "Supervised Random Walks: Predicting and Recommending Links in Social Networks." *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining - WSDM 11*, 2011, doi:10.1145/1935826.1935914.

[2] *Chimerdb*, 203.255.191.229:8080/chimerdbv31/mchimerkb.cdb.

[3] Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*

[4] Lu LY, Zhou T (2011) Link prediction in complex networks: A survey. Physica a-Statistical Mechanics and Its Applications 390: 1150–1170.

[5] Sun, Y., & Barber, R. (2009). Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks *, 1–31.