A Project Report on

Active subgraph identification for Pancreatic adenocarcinoma

Submitted by

Malvika Sudhakar
BT15D306

Beethika Tripathi
CS15S004

IIT Madras,

Chennai, 600036

April 2016

Contents

List of tables

List of figures

Abstract

Cancer is one of the leading causes of death today of which pancreatic cancer is estimated to contribute to 7% of deaths, in both males and females for the year 2016. Diseases are caused by changes in protein interaction. We aim to identify genes in *Homo sapiens*, which are active in cancer cells by studying protein-protein interaction (PPI) network.

To do this we have integrated differential expression data into directed PPI network(STRING) for more meaningful active sub graphs(Lai, Eckenrode, & She, 2009) identification. Gene expression data from RNA-seq experiments (TCGA) was used to calculate differential expression in pancreatic adenocarcinoma, which was further used to scoring the subnetwork. Mixed Integer Linear Programming(MILP) was used for identifying active subgraphs(Gaire et al., 2013). Further potential drugs for treatment of pancreatic cancer were identified using the subgraph.

Active sub graphs was identified for pancreatic cancer type, which included genes such as TNF and MAP3K1. We identified SYK, TNF and PRKCB to be potential drug targets, and BAY613606, Lenalidomide and Bryostapin drugs may be used for treatment of pancreatic cancer type. Active sub graphs can be used for identification of significant genes in a caner type, which can be used as molecular targets for treatment.

Introduction

Protein protein interaction (PPI) network gives an overview of all the possible interactions between all the given proteins in an organism. These networks have been used for identifying potential drug targets(Chu & Chen, 2008; Hormozdiari, Salari, Bafna, & Sahinalp, 2010), functionally important pathways(Gitter, Klein-Seetharaman, Gupta, & Bar-Joseph, 2011; Segal, Wang, & Koller, 2003), genes involved in diseases(Navlakha & Kingsford, 2010; Vanunu, Magger, Ruppin, Shlomi, & Sharan, 2010), etc.

Most PPI networks are undirected which doesn’t give much information about the functionality or effect of one gene on another. Introducing directionality along with gene expression data can help in reducing the effect of false positives and false negatives in identifying active sub networks.

Active sub network(ASN) is a subset of the original network consisting of only significant nodes as identified by experiments(Ideker, Ozier, Schwikowski, & Siegel, 2002). Since we are interested in PPI and identification of significant proteins, the change in expression levels of the genes coding for them in tumor compared to normal is good criteria for selecting significant nodes.

Pancreatic cancer has been estimated to be responsible for 7% of deaths in both men and women(Siegel, Miller, & Jemal, 2016), making it one of the top 5 cancer types leading to death. For the year 2016, 53070 new cases are predicted along with 41780 deaths due to pancreatic cancer in USA(Siegel et al., 2016).

The proteins selected in ASN are significant for pancreatic cancer. Hence, targeting these proteins may be useful.

Objectives

- Identification of ASN for pancreatic cancer type using directed PPI network and gene expression data.
- Change in the size of the network based on the value of τ (threshold for false discovery rate).
- Identification of probable drug targets and molecular drugs acting on them.

Methodology

ASN were identified using the method published by Gaire et al.(2013).

Datasets

Protein protein network was downloaded from STRING database containing the detailed information about the type of activity (STRING actions database) between proteins and their scores. The different types of interactions are activation, inhibition, binding, post translational modifications and catalysis. The resultant network is directed and genome wide.

Gene expression information for pancreatic cancer was downloaded from TCGA in the form of RNA-seq counts for both, normal and cancer tissues in humans. This information is processed using edgeR to get differentially expressed genes in tumors. This data contains gene names, expression levels and p value.

The PPI network contains only Ensemble proteins IDs, while gene expression data contains Entrez gene IDs and HUGO gene symbols. To map these two data onto each other we downloaded gene2ensemble data file from NCBI.

CancerDR database of molecular drugs and their potential targets was downloaded. Pancreatic Cancer Gene Database (PCGDB) was used for validation.

Calculation of scores

<u>Confidence score for PPI</u>

The score for a protein interaction is from STRING is an integer representing the confidence of an interaction occurring, given 2 proteins. This is combined score derived from neighborhood, fusion, co-occurrence, co expression, experimental data, database and text mining research papers. This was converted to confidence score between (0, 1].

$$Confidence\ score(C_e) = \frac{combined\ \ score}{\max(combined\ scores)}$$

Node scores

Nodes of an active subnetwork should have:

- low p-values(Ideker et al., 2002)
- not be highly connected(Dezso et al., 2009)

Dittrich et al’s node scoring method(Dittrich, Klau, Rosenwald, Dandekar, & Müller, 2008) was used that includes low p-values in the networks (profit) and penalises highly connected nodes (cost).

Profit score

Profit score (W_n) = $(a - 1) \times (\log(p_n) - \log(\tau))$

Where,

a = (0, 1] is a shape parameter of the beta distribution fitted for a dataset representing a signal to noise ratio

τ = threshold to control the size of the ASN (false discovery rate).

p_n = p value of a node

The value of (a - 1) acts as a scaling factor and if a single p-value dataset is used, a can be assigned a constant value close to zero without affecting the resulting ASN. Since a = 0 and τ can be ignored after applying a p-value threshold on experimental dataset, node score is

$W_n = 1\times \log(pn)$ for all the genes above p-value threshold

and

$W_n = -1 \times |\text{Constant}|$ for the genes not above p-value threshold

Cost score

If D is the degree of a node n, we assign a cost C_n to the node to penalize highly connected nodes.

$C_n = \log(D)$.

Since the profit and the cost scores of a node do not have the same scale, we scale these values to a range of [-1, 1] to obtain standard scores as,

$$W'_n = \frac{W_n}{\max(|W_n|)}$$

$$C'_n = \frac{C_n}{\max(|C_n|)}$$

Edge Scores

An edge score needs to be high if it is consistent with DE dataset and penalized otherwise. This is done as defined below:

Sr. no.	Activity	Consistency	Consistency Score(W$_e$)
1	$n1 \rightarrow n2$	$(n1 \uparrow \wedge\ n2 \uparrow) \vee (n1 \downarrow \wedge\ n2 \downarrow)$	2
		$(n1\ unchanged) \vee (n2\ unchanged)$	-1
		$(n1 \uparrow \wedge\ n2 \downarrow) \vee (n1 \downarrow \wedge\ n2 \uparrow)$	-2
2	$n1 \dashv n2$	$(n1 \uparrow \wedge\ n2 \uparrow) \vee (n1 \downarrow \wedge\ n2 \downarrow)$	-2
		$(n1\ unchanged) \vee (n2\ unchanged)$	-1
		$(n1 \uparrow \wedge\ n2 \downarrow) \vee (n1 \downarrow \wedge\ n2 \uparrow)$	2
3	$(n1 \multimap n2)$ $\vee (n1 - n2)$	$(n1 \uparrow \wedge\ n2 \uparrow) \vee (n1 \downarrow \wedge\ n2 \downarrow)$	-1
		$(n1\ unchanged) \vee (n2\ unchanged)$	-1
		$(n1 \uparrow \wedge\ n2 \downarrow) \vee (n1 \downarrow \wedge\ n2 \uparrow)$	-1

Table 1 : List of edge activity and consistency score

$$\text{Edge Score } (\mathrm{S_e}) = \mathrm{W_e} \times \mathrm{C_e}$$

Where,

C_e = (0, 1] is the confidence score of the edge

$$Standard\ edge\ score\ (S'_e) = \frac{S_e}{\max(|S_e|)}$$

Sub network Score

Based on the standard node and edge scores, the score of a subnetwork G’-is obtained.

$$S = \sum_{n \in V} x_n \times (W'_n - \beta_n \times C'_n) + \sum_{e \in E} x_e \times S'_e$$

Where,

β = 0 for positive scoring nodes

Or

β = 1 for negative scoring nodes

We used GUROBI optimizer to maximize the sub network score using mixed integer linear programming (MILP) model. Cytoscape was used to create the ASN. The selected list of proteins was later compared to molecular targets in CancerDR(Kumar et al., 2013) database to identify potential target site and their corresponding drugs.

Results and Discussion

The resulting ASN identified consisted of 98 nodes with 154 edges between them. The objective value returned was 9.15834. We conducted literature mining for the top 20 scoring nodes, out of which only one, TNF, was identified as significant gene for this cancer.

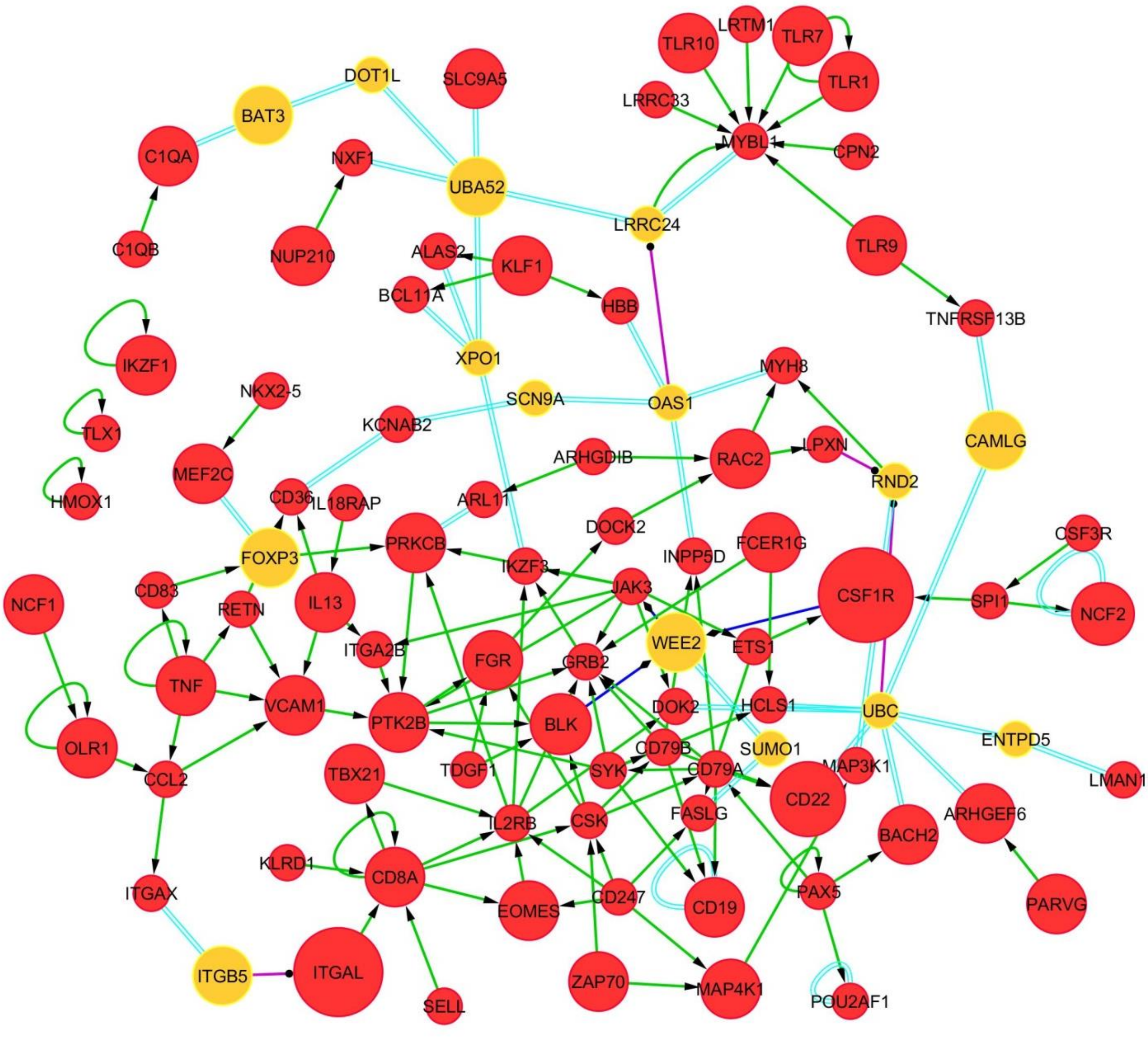

Figure 1 : ASN for pancreatic cancer. Red nodes signify unregulated genes. Yellow signifies no change. Size of node determines the relative node score. The arrow head point towards the target node in a given edges. Color code for edge : Green = activation; sky blue = binding; dark blue = protein modification; purple = catalysis.

Further all the selected nodes were compared to the gene list in Pancreatic Cancer Gene Database (PCGDB)* and only one gene (MAP3K1) was found to be common. This is because the other genes in PCGDB were not differentially expressed and hence the nodes were given low score. Out of the entire differentially expressed gene list in our dataset, GNG7 and MAP3K1, were found in PCGDB. But only MAP3K1 was present in ASN because GNG7 in not present in PPI network. Even though 101 genes out of 120 genes in PCGDB are found in PPI network, only 1 is differentially expressed.

No down regulated gene from the 32 down regulated genes were selected in the ASN. None of the 2328 inhibition edges out of 2045558 were elected either. Three disjoint nodes with self activation edges are also identified. This is probably because these nodes are high scoring and increase the overall value of the objective function.

ASN was generated using different τ (threshold for false discovery rate) to see effect on the network size. While the size of ASN reduced with decreasing τ values, no significant difference was observed. Given below are the details:

Sr. no.	τ	No. of nodes	No. of edges	Objective value
1	0.05	98	153	89.382
2	0.005	98	153	88.887
3	0.0005	96	151	82.211
4	0.00005	93	147	76.526
5	0.000005	92	146	73.511

Table 2 : Effect of τ on ASN size

The genes present in the ASN were compared to CancerDR database for identification on possible drug targets. Three targets were identified out of which TNF was also present in the list significant genes in PCGDB.

Sr. no.	Gene ID	Molecular Target (HUGO symbol)	Drug	Clinical Status	Activity
1	6850	SYK	BAY613606	Experimental	Tyrosine kinase
2	7124	TNF	Lenalidomide	Approved	Angiogenesis
3	5579	PRKCB	Bryostatin	Phase II	Angiogenesis, apoptosis

Table 3 : List of probable drugs for pancreatic cancer

Limitations

The selection of ASN is largely limited by the PPI network used. An undirected PPI network in STRING contains 8548003 edges while a directed PPI network in STRING has only 2045558 edges. This leaves out loads of important proteins whose interactions are not known. For eg., KRAS is a well known driver gene in pancreatic cancer which is not present in the PPI network.

The research on significant genes in pancreatic cancer is relatively less compared to other cancer types such as lung cancer and breast cancer. Experimental validation for the list of top significant

*PCGDB : www.bioinformatics.org/pcgdb/

genes was not found due to lack of research. Further research may prove them to be actively involved in pancreatic cancer. These shortlisted genes can be used for understanding mechanism in tumor formation.

The identification of significant genes gives importance to differentially expressed genes. Not all genes in cancer express differentially. Some genes have a differential functional impact in cancer cells due to mutations. This effect is not captured by the scoring function. On the other hand no change is expression is penalized.

Conclusion

Integration of gene expression data along with directed PPI network results in identification of ASN which can be used to identify genes involved in important pathways. For pancreatic cancer the ASN identified consisted of 98 nodes with 154 edges. This methods biggest limitation would be the lack all biological PPI in the PPI network used. Another major limitation would be that genes that are not differentially expressed but are significant to cancer type are not identified. The false discovery rate can be controlled by the value of τ, but no significant change in ASN size was observed. The ASN was searched for potential molecular targets for cancer therapy and few drugs were identified.

References

Chu, L.-H., & Chen, B.-S. (2008). Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. *BMC Systems Biology*, *2*(1), 56. http://doi.org/10.1186/1752-0509-2-56

Dezso, Z., Nikolsky, Y., Nikolskaya, T., Miller, J., Cherba, D., Webb, C., & Bugrim, A. (2009). Identifying disease-specific genes based on their topological significance in protein networks. *BMC Systems Biology*, *3*(1), 36. http://doi.org/10.1186/1752-0509-3-36

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., & Müller, T. (2008). Identifying functional modules in protein-protein interaction networks: An integrated exact approach. *Bioinformatics*, *24*(13). http://doi.org/10.1093/bioinformatics/btn161

Gaire, R. K., Smith, L., Humbert, P., Bailey, J., Stuckey, P. J., & Haviv, I. (2013). Discovery and analysis of consistent active sub-networks in cancers. *BMC Bioinformatics*, *14 Suppl 2*(Suppl 2), S7. http://doi.org/10.1186/1471-2105-14-S2-S7

Gitter, A., Klein-Seetharaman, J., Gupta, A., & Bar-Joseph, Z. (2011). Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Research*, *39*(4). http://doi.org/10.1093/nar/gkq1207

Hormozdiari, F., Salari, R., Bafna, V., & Sahinalp, S. C. (2010). Protein-protein interaction network evaluation for identifying potential drug targets. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *17*(5), 669–684. http://doi.org/10.1089/cmb.2009.0032

Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)*, *18 Suppl 1*, S233–S240. http://doi.org/10.1093/bioinformatics/18.suppl_1.S233

Kumar, R., Chaudhary, K., Gupta, S., Singh, H., Kumar, S., Gautam, A., … Raghava, G. P. S. (2013). CancerDR: cancer drug resistance database. *Scientific Reports*, *3*, 1445. http://doi.org/10.1038/srep01445

Lai, Y., Eckenrode, S. E., & She, J.-X. (2009). A statistical framework for integrating two microarray data sets in differential expression analysis. *BMC Bioinformatics*, *10 Suppl 1*(Suppl 1), S23. http://doi.org/10.1186/1471-2105-10-S1-S23

Navlakha, S., & Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, *26*(8), 1057–1063. http://doi.org/10.1093/bioinformatics/btq076

Segal, E., Wang, H., & Koller, D. (2003). Discovering molecular pathways from protein interaction and gene expression data. In *Bioinformatics* (Vol. 19). http://doi.org/10.1093/bioinformatics/btg1037

Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*, *00*(0), 1–20. http://doi.org/10.3322/caac.21332

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., & Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, *6*(1). http://doi.org/10.1371/journal.pcbi.1000641