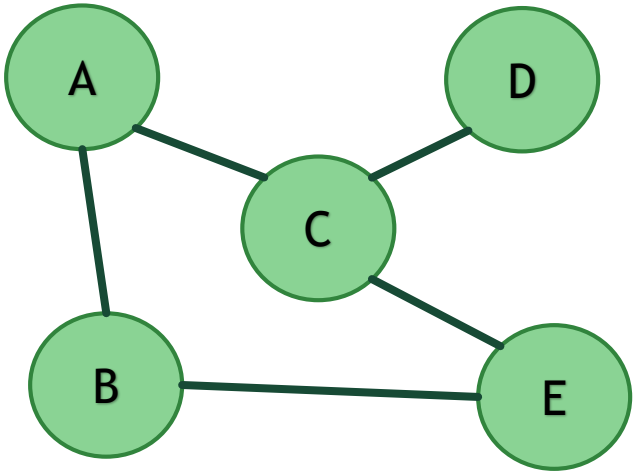
Active subgraph identification for Pancreatic adenocarcinoma

Beethika Tripathi (CS15S004)

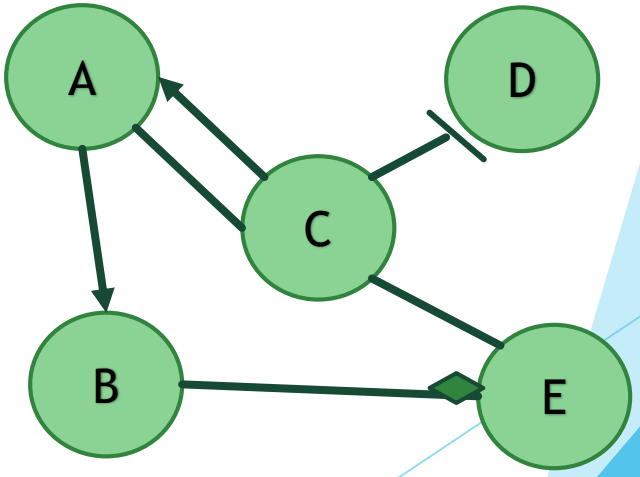
Malvika Sudhakar (BT15D306)

Introduction

- Protein protein interaction (PPI) network : all possible interactions between all proteins in an organism.
- These networks are mostly undirected.
  - Doesn't give information about type of interaction.
  - Directionality of interaction.



Undirected graph



Directed graph

Protein protein interaction (PPI) network

- Different types of interaction
  - Activation
  - Inhibition
  - Binding
  - Post translational modification
  - Catalysis
- Applications
  - Identifying potential drug targets
  - Functionally important pathways
  - Genes involved in diseases
  - etc.

Active sub network(ASN)

- Active sub network(ASN) is a subset of the original network consisting of only significant nodes.
- Gene expression data is the amount of mRNA produced in a cell
- Differential expression tells us the change in expression of genes in tumor sample with respect to normal sample.
  - Up regulation
  - Down regulation
  - No change
- 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

- Pancreatic adenocarcinoma arises when the secreting cells of the pancreas start to divide uncontrollably to form a tumor.
- One of the top 5 cancer types leading to death
- 53070 new cases are predicted in 2016
- 41780 deaths due to pancreatic cancer are predicted in USA

Estimated Deaths

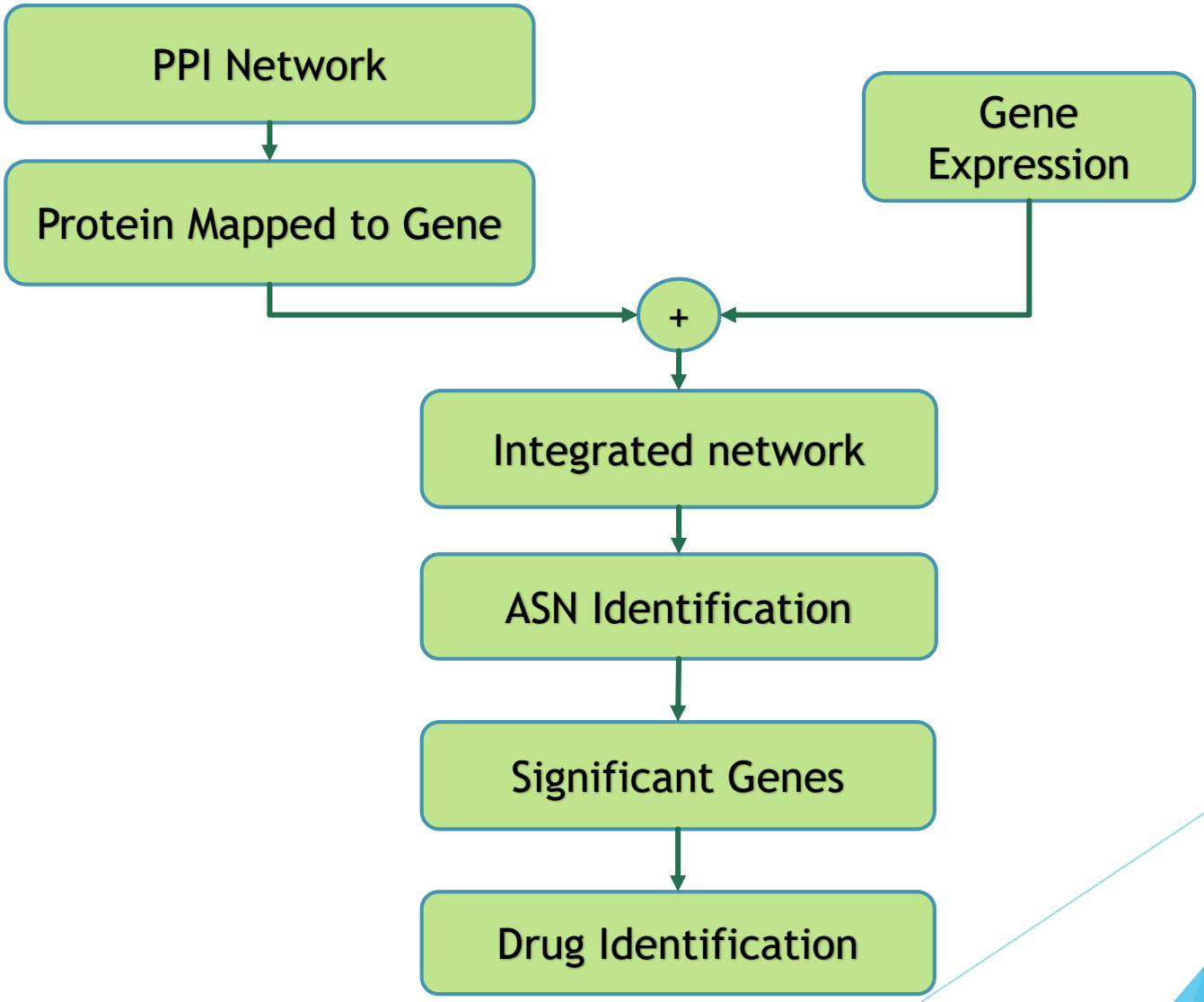
Male	Female
Lung & bronchus 85,920 (27%)	Lung & bronchus 72,160 (26%)
Prostate 26,120 (8%)	Breast 40,450 (14%)
Colon & rectum 26,020 (8%)	Colon & rectum 23,170 (8%)
Pancreas 21,450 (7%)	Pancreas 20,330 (7%)
Liver & intrahepatic bile duct 18,280 (6%)	Ovary 14,240 (5%)
Leukemia 14,130 (4%)	Uterine corpus 10,470 (4%)
Esophagus 12,720 (4%)	Leukemia 10,270 (4%)
Urinary bladder 11,820 (4%)	Liver & intrahepatic bile duct 8,890 (3%)
Non-Hodgkin lymphoma 11,520 (4%)	Non-Hodgkin lymphoma 8,630 (3%)
Brain & other nervous system 9,440 (3%)	Brain & other nervous system 6,610 (2%)
All sites 314,290 (100%)	All sites 281,400 (100%)

Estimates of cancer deaths for 2016(Siegel, Miller, & Jemal, 2016)

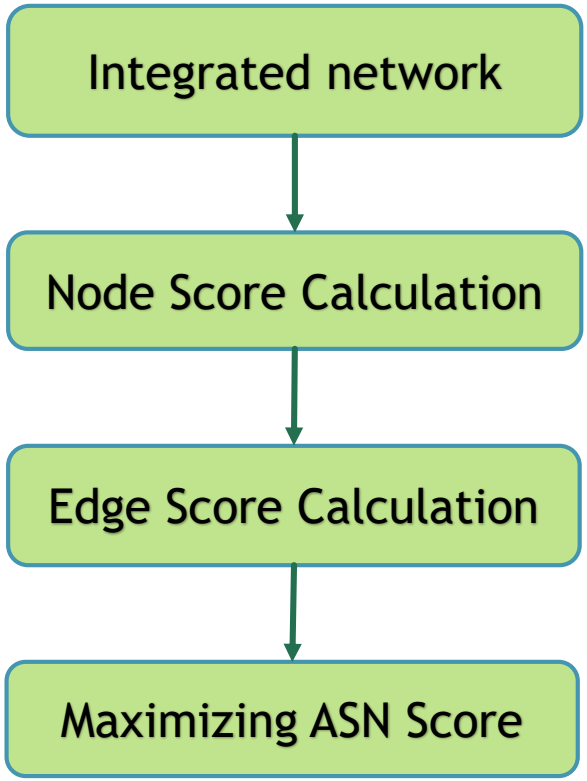
Objectives

- 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



Methodology



Node score

- Profit score

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

  $(W_n) = -1 \times \log(p_n)$

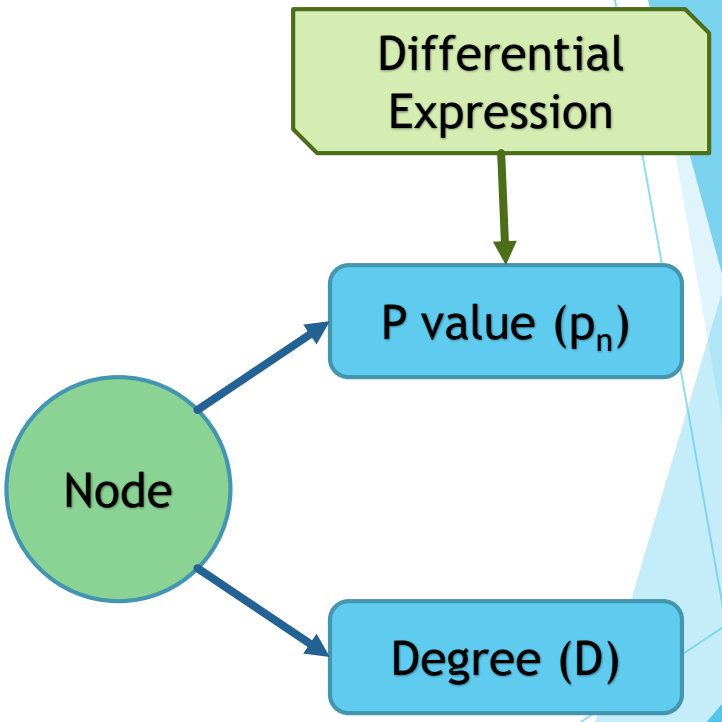
  $(W_n) = \log(\text{threshold})$

  where, threshold = pvalue

  = 0.05

- Cost score

  $C_n = \log(D)$



Edge score

- $Confidence\ score(C_e) = \frac{combined\ \ score}{\max(combined\ scores)}$
- Consistency Score(W_e)

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

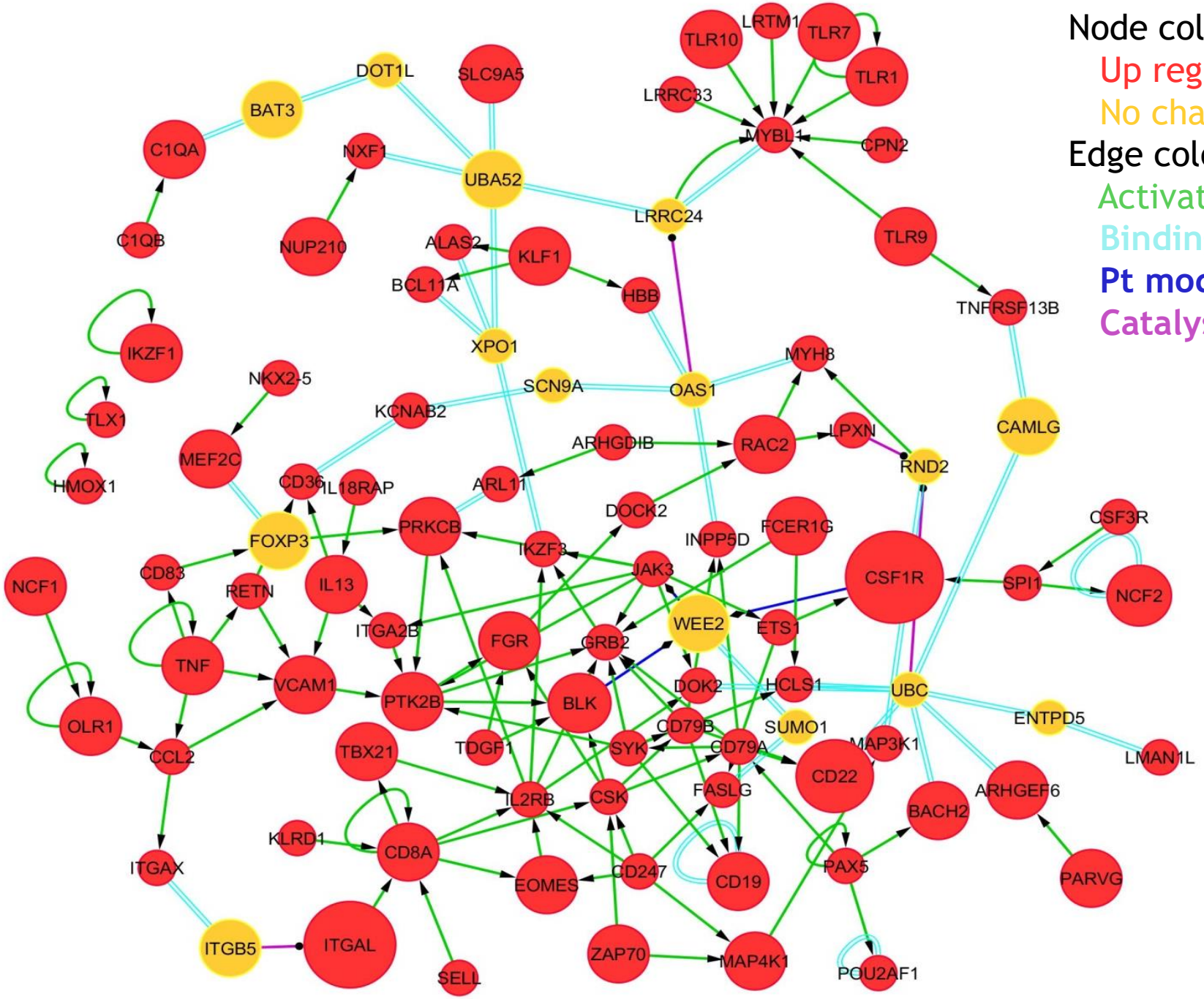
- Edge Score $(S_e) = W_e \times C_e$

Mixed Integer Programming

- Objective function
  - $S = \sum_{n \in V} x_n \times (W'_n - \beta_n \times C'_n) + \sum_{e \in E} x_e \times S'_e$
- Constraints
  - $x_{n(i)} \leq \sum_j x_{e(i,j)}$
  - $x_{e(i,j)} \leq x_{n(i)}$
  - $x_{e(i,j)} \leq x_{n(j)}$

Results

- Nodes = 98
- Edges = 154
- Objective value = 9.15834


Node color:
Up regulated
No change
Edge color:
Activation
Binding
Pt mod
Catalysis
TLR10
LRTM1
TLR7
TLR1
LRRC33
MYBL1
CPN2
DOT1L
SLC9A5
BAT3
C1QA
NXF1
UBA52
LRRC24
C1QB
NUP210
ALAS2
KLF1
BCL11A
HBB
TLR9
TNFRSF13B
IKZF1
XPO1
MYH8
NKX2-5
SCN9A
OAS1
TLX1
KCNAB2
CAMLG
MEF2C
ARHGDIB
RAC2
LPXN
RND2
HMOX1
CD36
IL18RAP
ARL11
DOCK2
CSF3R
FOXP3
PRKCB
IKZF3
INPP5D
FCER1G
CD83
JAK3
CSF1R
SPI1
NCF2
NCF1
RETN
IL13
ITGA2B
WEE2
ETS1
FGR
GRB2
TNF
VCAM1
PTK2B
BLK
DOK2
HCLS1
UBC
ENTPD5
OLR1
CCL2
SUMO1
CD79B
MAP3K1
LMAN1L
TBX21
TDGF1
SYK
CD79A
CD22
IL2RB
CSK
FASLG
BACH2
ARHGEF6
KLRD1
CD8A
PAX5
ITGAX
EOMES
CD247
CD19
PARVG
ITGB5
ITGAL
ZAP70
MAP4K1
SELL
POU2AF1

Analysis of significant genes

- Literature mining for the top 20 scoring nodes ->TNF,CSF1R,TLR7.
- Pancreatic Cancer Gene Database (PCGDB) -> MAP3K1
- PCGDB ∩ differentially expressed gene = GNG7, MAP3K1
- PCGDB(120 genes) ∩ PPI network = 101 genes
- PCGDB(120 genes) ∩ PPI network ∩ differentially expressed gene = MAP3K1
- 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.

Analysis of effect of τ

- No significant change in the ASN size was observed

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

Identification of drugs

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 size of PPI network -> lack of information may miss important interactions
  - Undirected PPI network in STRING = 8548003 edges
  - Directed PPI network in STRING = 2045558 edges.
  - For e.g., KRAS is a well known driver gene in pancreatic cancer which is not present in the directed PPI network.
- Lack of experimental evidence
  - Shortlisted genes can be used for understanding mechanism in tumor formation.
- Priority given to differentially expressed genes than change in functional impact due to mutations in cancer cells.
  - No change is expression is penalized.

Conclusion

- Gene expression data + directed PPI network -> identification of ASN
- Can be used to identify genes involved in important pathways.
- ASN identified consisted of 98 nodes with 154 edges.
- 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(SYK, TNF and PRKCB) for cancer therapy and few drugs(BAY613606, Lenalidomide and Bryostapin) were identified.

Datasets

- Protein protein network -> STRING database
- RNA-seq data for pancreatic cancer -> TCGA
- gene2ensemble data file -> NCBI
- Molecular drugs and their potential targets -> CancerDR
- List of significant genes in pancreatic cancer -> Pancreatic Cancer Gene Database (PCGDB)

References

- 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
- 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
- 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–S40. 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
- 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

Thank you! ☺