

Inferred GRNs and Cancer

Michael Dempsey


Motivating Questions:

- How can we create a network based solely off gene expression data?
- What information can we glean from inferred networks, particularly when it comes to cancer research?

Inferred Networks...

- Inferring gene regulatory networks (GRNs) from expression data
- BC3NET algorithm

Bagging Statistical Network Inference from Large-Scale Gene Expression Data

Ricardo de Matos Simoes, Frank Emmert-Streib 

Published: March 30, 2012 • <https://doi.org/10.1371/journal.pone.0033624>

What is BC₃NET?

- Bagging C3NET (BC3NET)

“The underlying procedure of BC3NET is to generate an ensemble of bootstrap datasets from which an ensemble of networks is inferred by using C3NET”

What is C₃NET?

- Identify significant maximum mutual information network

Inferring the Conservative Causal Core of Gene Regulatory Networks

Article in BMC Systems Biology · September 2010

DOI: 10.1186/1752-0509-4-132 · Source: PubMed

Gene Expression Data  Adjacency Matrix

What is C₃NET?

Algorithm 1 Principle steps of our inference algorithm C₃NET.

```
1:  $A$ : initiate adjacency matrix,  $A_{ij} = 0$  for all  $i, j \in V$ 
2:  $C$ : initiate connectivity matrix,  $C_{ij} = 1$  for all  $i, j \in V$ 
3: estimate mutual information  $I_{ij}$  for all  $i, j \in V$ 
4: repeat
5:   Set  $C_{ij} = 0$  if  $I_{ij} = 0$  is not statistically significant
   (hypothesis test)
6: until all pairs  $i \neq j$  are tested
7: for all  $i \in V$  do
8:    $N_s(i) = \{j : C_{ij} = 1 \text{ and } j \neq i\}$ 
9:   if  $N_s(i) \neq \emptyset$ 
10:     $j_c(i) = \arg \max_{j \in N_s(i)} \{I_{ij}\}$ 
11:   else
12:     $j_c(i) = \emptyset$ 
13:   endif
14: end for
15: for all  $i \in V$  do
16:   if  $j_c(i) = \emptyset$ 
17:     $A_{ij_c(i)} = A_{j_c(i)i} = 1$ 
18:   endif
19: end for
20: return adjacency matrix  $A$ 
```

Inferring the Conservative Causal Core of Gene Regulatory Networks

Article in BMC Systems Biology · September 2010

DOI: 10.1186/1752-0509-4-132 · Source: PubMed

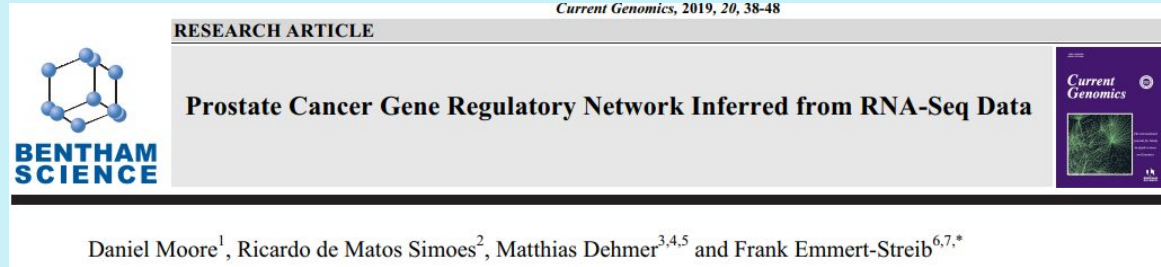
“The availability of this type of data [large-scale expression data] ushered the development of methods inferring and analyzing gene regulatory networks”

What is C₃NET?

- Step 1: Mutual information for each gene pair is calculated
 - Values are determined to be significant based off a threshold
- Step 2: For each gene, the most significant link is selected

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

- Gene expression data from The Cancer Genome Atlas (TCGA)
- 333 samples



Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Data Collection and Preprocessing

- cBioPortal

<input type="checkbox"/> Prostate Adenocarcinoma (MSK, PNAS 2014)	104 samples
<input type="checkbox"/> Prostate Adenocarcinoma (MSK/DFCI, Nature Genetics 2018)	1013 samples
<input type="checkbox"/> Prostate Adenocarcinoma (SMMU, Eur Urol 2017)	65 samples
<input checked="" type="checkbox"/> Prostate Adenocarcinoma (TCGA, Cell 2015)	333 samples
<input type="checkbox"/> Prostate Adenocarcinoma (TCGA, Firehose Legacy)	501 samples
<input type="checkbox"/> Prostate Adenocarcinoma (TCGA, GDC)	501 samples

- Found the average gene expression levels for each gene
 - Removed the lower quartile
- 15,376 genes used

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Data Collection and Preprocessing

[7]:

	Hugo_Symbol	TCGA-2A-A8VL-01	TCGA-2A-A8VO-01	TCGA-2A-A8VT-01	TCGA-2A-A8VV-01	TCGA-2A-A8W1-01	TCGA-2A-A8W3-01	TCGA-CH-5737-01	TCGA-CH-5738-01	TCGA-CH-5739-01	...	TCGA-YL-A8SQ-01	TCGA-A8S
10092	DUX4L5_653545_1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000000e+00	0.0000	0.0000	...	0.0000	0
16738	SNORD115-41_100033815_1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000000e+00	0.0000	0.0000	...	0.0000	0
16737	SNORD115-40_100033814_1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000000e+00	0.0000	0.0000	...	0.0000	0
16736	SNORD115-3_100033440_1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000000e+00	0.0000	0.0000	...	0.0000	0
11101	MRGPRG_386746_1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000000e+00	0.0000	0.0000	...	0.0000	0
...
5359	EEF2_1938_1	73802.2657	98047.6733	68504.7794	131972.2320	113594.1447	99169.9039	9.903228e+04	98289.0157	108054.2568	...	91775.2747	75180
5351	EEF1A1_1915_1	94756.9516	128735.2579	47778.6765	154442.6403	58509.7391	148416.6837	9.429255e+04	111491.2190	117451.9078	...	102367.0330	62089
174	ACPP_55_1	322984.0371	58250.1171	31068.7500	186556.5361	30452.6030	13681.4557	3.742105e+05	159862.8420	34364.8649	...	61764.2857	111736
9146	KLK2_3817_1	211670.4428	266434.2706	76452.2059	195684.6688	141385.9049	109252.5046	1.815106e+05	107073.6385	130960.6399	...	204940.6593	194498
9147	KLK3_354_1	586951.5963	628137.4305	119260.6618	603737.2350	737810.8842	244040.0736	1.041554e+06	260975.7962	251408.3506	...	521491.7582	725025

20502 rows × 335 columns

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Data Collection and Preprocessing

[14]:

	Hugo_Symbol	TCGA-2A-A8VL-01	TCGA-2A-A8VO-01	TCGA-2A-A8VT-01	TCGA-2A-A8VV-01	TCGA-2A-A8W1-01	TCGA-2A-A8W3-01	TCGA-CH-5737-01	TCGA-CH-5738-01	TCGA-CH-5739-01	...	TCGA-A8SC
665	ANKRD20A3_441425_1	3.343614	3.280343	2.737039	1.885402	2.294170	1.744808	1.478007	2.924005	1.910549	...	3.461
13755	POMC_5443_1	2.737480	1.200898	1.301961	1.885402	0.635942	1.302967	0.871879	2.117532	1.156787	...	1.598
19908	ZMAT4_79698_1	2.555155	0.284577	2.257462	2.959919	-0.057205	1.185187	0.178732	1.789022	1.716389	...	0.499
30	ABCA10_10349_1	2.427322	3.305035	2.783560	2.381840	2.746155	1.878335	1.719171	2.636325	1.785372	...	2.659
19368	VPREB3_29802_1	0.000000	0.690093	1.196586	1.038119	1.957698	1.878335	1.277344	2.562215	1.562241	...	3.229
...
5359	EEF2_1938_1	11.209145	11.493209	11.134659	11.790347	11.640387	11.504590	11.503201	11.495668	11.590389	...	11.427
5351	EEF1A1_1915_1	11.459070	11.765513	10.774335	11.947578	10.976948	11.907779	11.454157	11.621701	11.673784	...	11.536
174	ACPP_55_1	12.685358	10.972501	10.343958	12.136490	10.323927	9.523797	12.832574	11.982071	10.444790	...	11.037
9146	KLK2_3817_1	12.262786	12.492883	11.244421	12.184260	11.859248	11.601417	12.109069	11.581272	11.782652	...	12.230
9147	KLK3_354_1	13.282698	13.350514	11.689067	13.310894	13.511443	12.405088	13.856224	12.472183	12.434834	...	13.164

15376 rows × 335 columns

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Running BC₃NET

- Using R

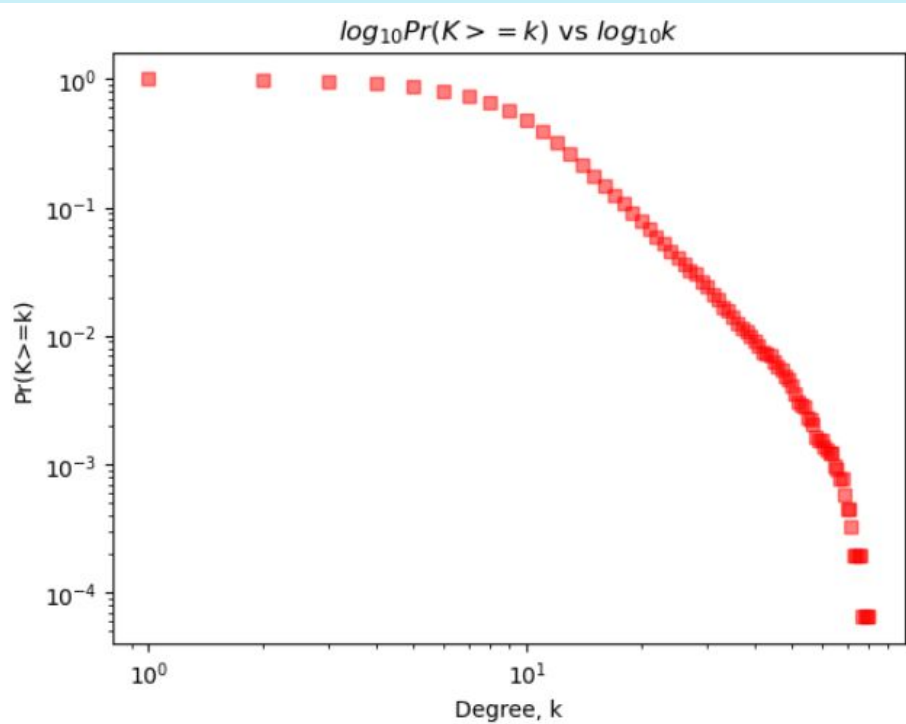
```
bc3net(dataset, boot=100, estimator="pearson", disc="equalwidth", mtc1=TRUE,  
alpha1=0.05, nullit=NA, null=c(), adj1="bonferroni", mtc2=TRUE,  
alpha2=0.05, adj2="bonferroni",  
weighted=TRUE, igraph=TRUE, verbose=FALSE)
```

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Running BC₃NET

	from	to	weight
1	FAM95B1_10013303...	ANKRD20A3_441425...	1
2	VAMP5_10791_1	POMC_5443_1	0.45
3	EFNB1_1947_1	POMC_5443_1	0.03
4	TRIP6_7205_1	POMC_5443_1	0.29
5	RARRES2_5919_1	POMC_5443_1	0.1
6	EFEMP2_30008_1	POMC_5443_1	0.03
7	CDH22_64405_1	ZMAT4_79698_1	0.1
8	GALNT14_79623_1	ZMAT4_79698_1	0.04
9	NOVA1_4857_1	ZMAT4_79698_1	0.03
10	FZD10_11211_1	ZMAT4_79698_1	0.03
11	RSPO3_84870_1	ZMAT4_79698_1	0.04
12	ZNF439_90594_1	ZMAT4_79698_1	0.03
13	SLITRK5_26050_1	ZMAT4_79698_1	0.04
14	RGMA_56963_1	ZMAT4_79698_1	0.06
15	NUDT3_11165_1	ZMAT4_79698_1	0.03
16	NGFR_4804_1	ZMAT4_79698_1	0.06
17	TFEB_7942_1	ZMAT4_79698_1	0.08
18	DHRS3_9249_1	ZMAT4_79698_1	0.03

Inferring a Prostate Cancer Gene Regulatory Network (GRN) Analyzing Data



Gene Regulatory Network:

number of nodes, $n = 15376$

number of edges, $m = 81068$

mean degree, $\langle k \rangle = 10.54$

clustering coefficient, $C = 0.08$

mean geodesic distance, $\langle \ell \rangle = 4.93$

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Comparing Results

Paper

- Nodes: 15,376 genes
- Edges: 82,579 edges
- MGD: 4.98
- Edge Density: 6.99×10^{-4}
- Max Degree: 96

Recreation

- Nodes: 15,376 genes
- Edges: 81,068 edges
- MGD: 4.93
- Edge Density: 6.86×10^{-4}
- Max Degree: 80

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Comparing Results

Paper

	Gene Symbols	Entrez ID	Degree
	PIK3C2A	5286	96
	SCAF11	9169	90
	AURKAIP1	54998	86
	RIF1	55183	85
	NDUFA13	51079	84
	NOSIP	51070	79
	NAA10	8260	78
	CLTB	1212	75
	ASXL2	55252	75

Recreation

	Gene	Degree
0	HELZ	80
1	TRIP6	76
2	CCDC124	76
3	TAOK1	72
4	NDUFB7	72
5	GSTP1	71
6	BPTF	71
7	FAM156A	69
8	CCNL2	69

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Comparing Results

Paper

	Gene Symbols	Entrez ID	Degree
	PIK3C2A	5286	96
	SCAF11	9169	90
	AURKAIP1	54998	86
	RIF1	55183	85
	NDUFA13	51079	84
	NOSIP	51070	79
	NAA10	8260	78
	CLTB	1212	75
	ASXL2	55252	75

Recreation

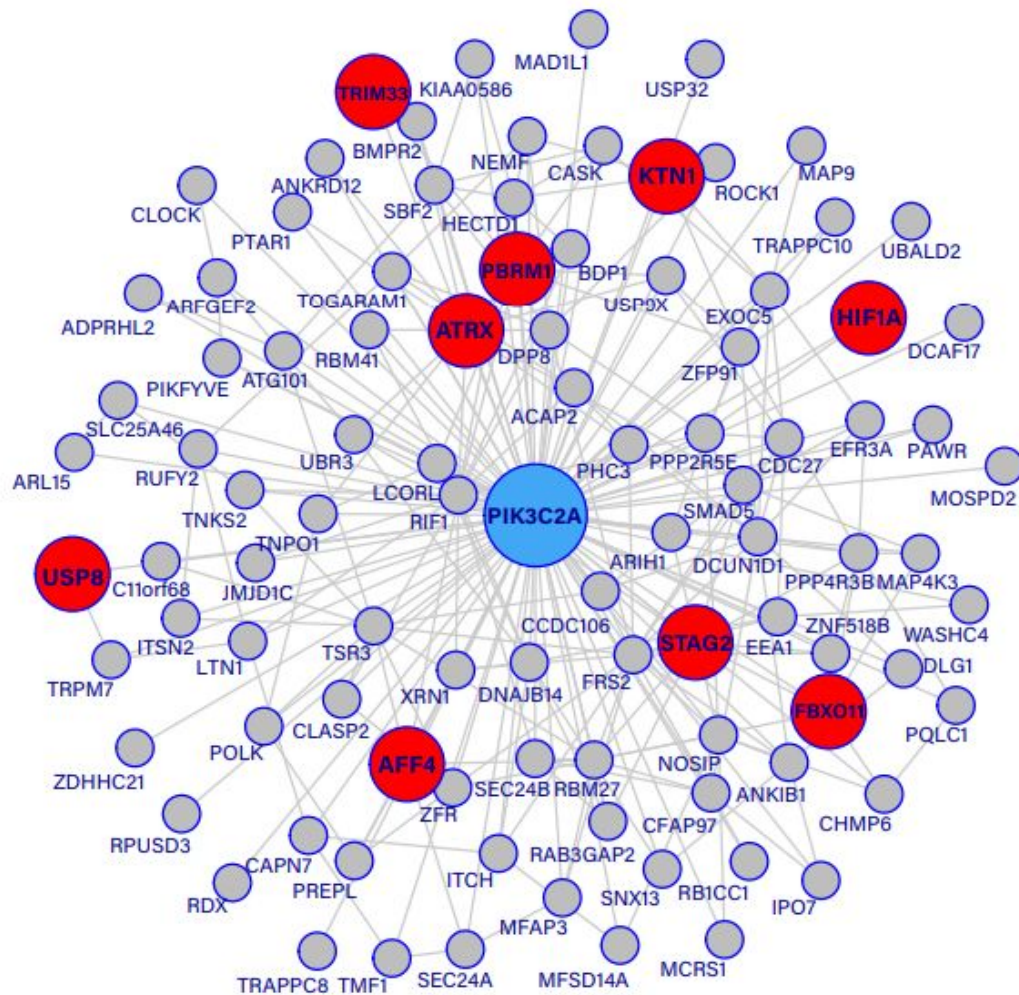
	Gene	Degree
0	PIK3C2A	56
1	SCAF1	31
2	AURKAIP1_54998_1	57
3	RIF1	56
4	NDUFA13	57
5	NOSIP	47
6	NAA10	53
7	CLTB	31
8	ASXL2	38

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

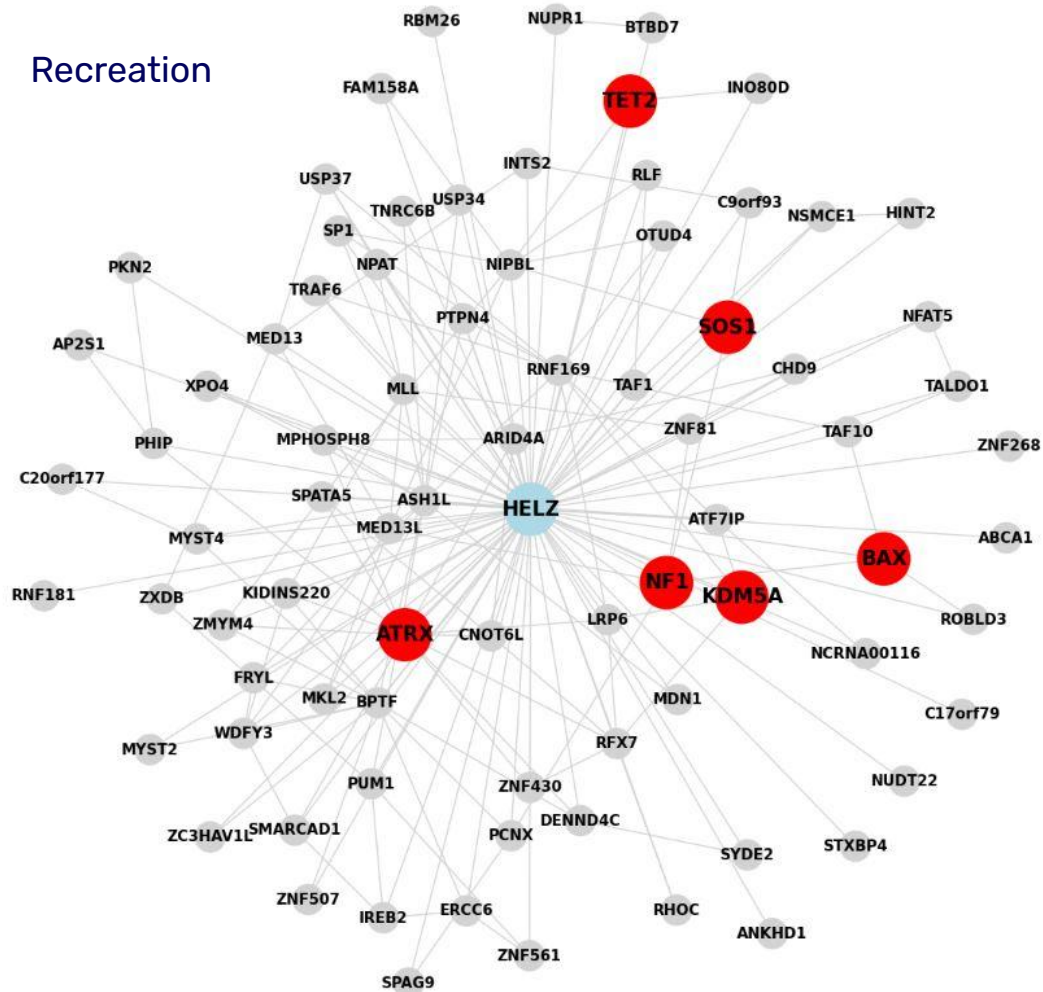
- Tier 1 Subset of Cancer Gene Census
 - “... documented activity relevant to cancer, along with evidence of mutations in cancer which change the activity of the gene product in a way that promotes oncogenic transformation.”



Paper



Recreation



Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Null Modeling

- How much of the structure of the GRN is driven purely by degree structure?

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Random Models

- **Erdős-Rényi (ER):** edges are independent and identically distributed (iid) with probability p
- **Configuration Model:** Given a degree sequence / degree distribution
- **Modular:** Given a specified modular structure and (possibly) a specified degree sequence

Inferring a Prostate Cancer Gene Regulatory Network (GRN)

Random Models

- **Erdős-Rényi (ER):** edges are independent and identically distributed (iid) with probability p
- **Configuration Model:** Given a degree sequence / degree distribution
- **Modular:** Given a specified modular structure and (possibly) a specified degree sequence

Inferring a Prostate Cancer Gene Regulatory Network (GRN) Using the Configuration Model

configuration_model

`configuration_model` (*deg_sequence*, *create_using=None*, *seed=None*) [\[source\]](#)

Return a random graph with the given degree sequence.


The configuration model generates a random pseudograph (graph with parallel edges and self loops) by randomly assigning edges to match the given degree sequence.


Inferring a Prostate Cancer Gene Regulatory Network (GRN) Using the Configuration Model

UpasanaDutta98 / ConfigModel_MCMC Public

 Notifications

 Code

 Issues

 Pull requests

 Actions

 Projects

 Security

 Insights

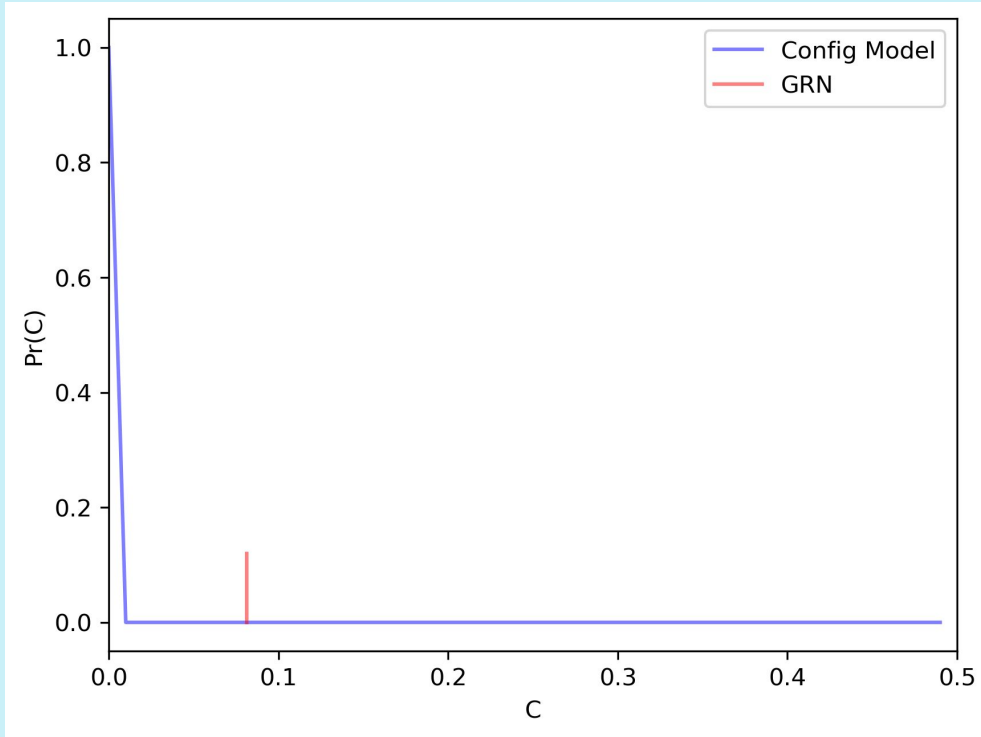
```
import ConfigModel_MCMC as CM
# Specify the graph space and create a new object
allow_loops = True
allow_multi = False
is_vertex_labeled = True
mcmc_object = CM.MCMC(G, allow_loops, allow_multi, is_vertex_labeled)

# Get a new graph (G2) from the Configuration model
G2 = mcmc_object.get_graph()
```

```
self_loops = list(nx.selfloop_edges(G))
print("Self loops:", self_loops)

Self loops: [('TMEM88', 'TMEM88'), ('FGF13', 'FGF13'), ('QSOX1', 'QSOX1'), ('SNAP47', 'SNAP47')]
```

Inferring a Prostate Cancer Gene Regulatory Network (GRN) Using the Configuration Model



- Degree structure cannot explain the clustering coefficient

Thank you!