

GIN-NA: Gene Interaction Network Analysis

Arzucan Ozgur¹ and Dragomir R. Radev^{1,2} CLAIR Lab



¹Department of Electrical Engineering and Computer Science, ²School of Information, University of Michigan, Ann Arbor, MI

System Description

- A system for analysing molecule interaction networks
 - Interactions retrieved from the MiMI database
 - Uses the Network module of Clairlib
 - Molecule-Specific Networks:
 - Network of interactions in the neighborhood of a molecule
 - Disease-Specific Networks:
 - Network of interactions in the neighborhood of known disease genes
 - Prostate Cancer, Type 1 Diabetes, Type 2 Diabetes, Bipolar Disorder
 - General network statistics
 - Ranking molecules based on centrality and second neighbor statistics

General Network Statistics

Ranking Molecules: Second Neighbor Statistics

- Rank second neighbors based on the number of common immediate neighbors
 AR JUN MYC
 - AR and MYC not directly connected
 - They have 45 common neighbors such as JUN.



- Diameter
- Degree Statistics:
 - Average degree
 - Power-law degree distribution
 - (exponent and r-squared statistics)
- Clustering:
 - Watts Strogatz clustering coefficient:
 - How well connected a molecule's neighbors are.
 - (# of interactions among a molecule's neighbors) / (# of possible interactions among them)
 - Newman clustering coefficient:
 - Measures the density of triangles (local clusters) in the network.
 - 3 x (# of triangles in the network) / (# of connected triples in the network)
 - Shortest Path Statistics:
 - Clairlib average undirected shortest path:
 - Average of the shortest paths between all molecules
 - Ferrer average undirected shortest path:
 - Average of all the average path lengths between all molecules
 - Harmonic mean geodesic distance:
 - Harmonic mean of the shortest paths between all pairs of molecules

Ranking Molecules based on Network Centrality



 $logP(\mathbf{k})$ $P(\mathbf{k}) \sim \mathbf{k}^{-\gamma}$ power-law

log(k)

Molecule-Specific Networks



Disease-Specific Networks

Network analysis methods to predict gene-disease associations.

- Centrality: Importance of a node in the graph
 - Degree: The more neighbors a node has, the more important it is. (most central node: x)
 - Betweenness: The more shortest paths pass through the node, the more important it is. Control of a node over the information flow of the network.

(most central node: y)

• Closeness: The closer a node to the other nodes, the more important it is. (most central node: y)



Disease-Specific Networks for the Driving Biological Problems

2		Gene	miera	cuon	Network	NC #BI
ne Molecu	e-specinic	Network Dise	ase-Specific in	IECWOIKS A	bout	
iseas	e-Sp	ecific l	Netwo	ork: Pi	rostate	Seed Disease Genes
ancer	-					• AR
ancer						BRCA2
eneral S	tatist	ics				<u>CHEK2</u> CD82
						PTEN
Nodes: 613						• <u>MXI1</u> - KI E6
 Edges: 603 Diameter: 5 	4					• ELAC2
Average de	gree: 19	.69				• HP1
 Degree sta 	tistics:					HPC1 PCAP
o Powe o Newn	r law exp nan powe	onent: 2.36 r-so er law exponent:	1.52 Error	0.02		• MSR1
 Clustering: 						RNASEL
 Watts 	Strogat	z clustering coef	ficient 0.25	20		MAU1L1 FPHB2
o Newn o Clairli	b averad	ening coefficient te undirected sh	ortest path:	2.48		
 Ferre 	r averag	e undirected sho	ortest path: 2	2.47		Links
 Harm 	onic mea	an geodesic dis	tance: 2.31			
lost Cen	tral 1	0 Nodes (Inferred	1 Diseas	e Genes)	MIMI
001 0011	indi i	o modeo (Diocut	ie Genes,	Information about the seed Prostate Cancer
Degree Co	entrality	Betweenness	Centrality	Closeness	Centrality	genes on MiMI
Molecule	Score	Molecule	Score	Molecule	Score	Cytoscape
TAF1	0.384	TAF1	0.159	TAF1	0.594	Cycocapo
HNF4A	0.255	AR	0.087	HNF4A	0.545	Visualize the Prostate Cancer specific network
AR	0.245	HNE44	0.080	MYC	0.535	
	0.240	CINC 9/A			0.000	on choscape
MYC	0.229	PTEN	0.074	AR	0.534	<u>on Cwoscape</u>
MYC PTEN	0.229	PTEN MYC	0.074	AR TP53	0.534 0.524	<u>on cytoscale</u>
MYC PTEN E2F4	0.240	PTEN MYC BRCA2	0.074 0.035 0.035	AR TP53 PTEN	0.534 0.524 0.514	<u>on cytoscale</u>
MYC PTEN E2F4 MAX	0.229 0.217 0.181 0.178	PTEN MYC BRCA2 E2F4	0.074 0.035 0.035 0.031	AR TP53 PTEN E2F4	0.534 0.524 0.514 0.512	<u>on cytoscale</u>
MYC PTEN E2F4 MAX E2F1	0.229 0.217 0.181 0.178 0.150	PTEN MYC BRCA2 E2E4 MAD1L1	0.074 0.035 0.035 0.031 0.024	AR TP53 PTEN E2F4 RBL2	0.534 0.524 0.514 0.512 0.507	<u>on cytoscope</u>

Home Molec	GIN - Ile-Specific	– Gene I Network Disea	nterac se-Specific N	ction N letworks	letwo	rk	NCTBI
Diseas	e-Sp	ecific N	etwo	rk: Bip	olar	S	eed Disease Genes
DISOID General S Edges: 12 Edges: 12 Diameter Average d Degree st OPow New Clustering Wat Clustering OW Clair Ferr OHAR	er Statisti o 38 6 9 gree: 7.7 tistics: er law exp man clust ib averago man clust ib averago r average nonic mea	CS 4 onent: 2.32 r-sg r law exponent: z clustering coefficient: e undirected shor undirected shor n geodesic dista) Nodes (In	uared: 0.96 1.93, Error 0.0836 0.0836 trest path: test path: 2 ncce: 2.49	0.05 896 2.68 Disease	Gene	Li M ge Cy S) Vir	SLC6A4 TPH2 DRD4 SLC6A3 DAOA DINBP1 NRG1 DISC1 BDNF inks iiMi formation about the seed Bipolar Disorder enes on MiMI sytoscape isualize the Bipolar Disorder specific network
Degree (entrality	Betweenness	Centrality	Closeness (Centrality	on	n Cytoscape
Molecule	Score	Molecule	Score	Molecule	Score		
DISC1	0.458	DISC1	0.335	DAOA	159,500		
TAF1	0.323	TAF1	0.177	TDH	106.333		
DTNBP1	0.210	DTNBP1	0.121	ANXA8L2	106.333		
HNF4A	0.207	HNF4A	0.081	DISC1	0.557		
MYC	0.201	DRD4	0.069	TAF1	0.537		
MAX	0.154	MYC	0.051	MYC	0.497		
E2F4	0.141	E2F4	0.048	HNF4A	0.478		
DRD4	0.113	BDNF	0.041	RPS25	0.471		

Home Molecul	GIN - e-Specific	- Gene				k I
Disease	e-Sp	ecific N	Vetwo	ork: Tv	pe 1	Seed Disease Genes
Diabete	s					• OAS1
General S	tatisti	cs				• <u>FOAP3</u> • <u>ITPR3</u> • <u>DTDN22</u>
	uusu					• IDDM1
 Nodes: 490 Edges: 229 	3					• <u>IL6</u> • <u>HNE1A</u>
 Diameter: 5 Average de 	gree: 9.3	6				Links
Degree stat O Powe	tistics:	- onent: 2 31 r.s	coupred: 0.9	a1		Links
 Newn 	nan powe	r law exponen	t: 1.92, Erro	or: 0.04		MiMI
 Clustering: Watts 	Strogat	z clustering co	efficient: 0.4	4913		Information about the seed Type 1 D
 Newn Clairli 	han clust	ering coefficier	nt: 0.0926	2 39		genes on Millini
 Ferrer 	r average	undirected sh	ortest path:	2.39		Cytoscape
o Harm	onic mea	n geodesic dis	stance: 2.25	2		Visualize the Type 1 Diabetes speci on Cytoscape
Most Cen	tral 10) Nodes (Inferred	d Disease	Gene	<u>un cytoscape</u>
Degree C	entrality	Betweennes	s Centralit	y Closeness	Centrality	
Molecule	Score	Molecule	Score	Molecule	Score	
HNE1A	0.703	HNE1A	0.532	HNE1A	0.734	
HNF4A	0.511	HNF4A	0.243	HNF4A	0.643	
E254	100 B 100 B 100 B	IIPRO	0.048	<u>EP300</u>	0.490	
E2F4 MYC	0 147	PLCB1	0.038	MYC	0 4 9 5	
E2F4 MYC ITPR3	0.147	PLCB1 MYC	0.038	MYC PLCB1	0.495	
E2F4 MYC ITPR3 PLCG1	0.143	PLCB1 MYC E2F4	0.038 0.036 0.035	MYC PLCB1 E2F4	0.495 0.494 0.492	



Disease	e-Specif	ic Netwo	ork: Ty	pe 2	Seed Disease Genes
Diabete	S				IGFBP2
					CDKAL1 CDKN2A
General S	tatistics				CDKN2B
a Nadaa: 442					• FTO • PPAPC
 Nodes: 443 Edges: 272 	0				 EPARG SLC30A8
 Diameter: 5 	-				HHEX
 Average deg 	ree: 12.28				<u>TCF7L2</u> KONUM
 Degree stat Power 	istics: law exponent: 2	35 r-squared: 0	92		• KONDII
 Newm 	an power law ex	ponent: 1.69, Em	or: 0.03		Links
 Clustering: 					
 Watts 	Strogatz cluster	ing coefficient: 0. afficient: 0.1063	3106		MiMI
 Clairli 	b average undirect	ted shortest path	1: 2.47		before the short the second True O Distance
 Ferrer 	average undirect	ed shortest path	2.46		denes on MiMI
		sic distance: 2.2	9		
 Harmo 	nic mean geode		-		
• Harmo	tral 10 Nod	es (Inferre	d Disease	Genes)	Cytoscape
o Harmo Most Cent	tral 10 Nod	es (Inferre	d Disease	Genes)	Cytoscape <u>Visualize the Type 2 Diabetes specific networ</u>
o Harmo Most Cent Degree Co	tral 10 Nod	es (Inferre)	d Disease	e Genes) Centrality	Cytoscape Visualize the Type 2 Diabetes specific networ on Cytoscape
o Harmo Most Cent Degree Co Molecule	tral 10 Nod entrality Betwee Score Molecu	es (Inferre enness Centrali ile Score	d Disease	e Genes) Centrality Score	Cytoscape Visualize the Type 2 Diabetes specific networ on Cytoscape
Most Cent Degree Co Molecule	tral 10 Nod entrality Betwee Score Molect 0.450 TAF1	es (Inferre enness Centrali ile Score 0.239	d Disease	Centrality Score 0.623	Cytoscape Visualize the Type 2 Diabetes specific networ on Cytoscape
Most Cent Degree Co Molecule TAF1 CDKN2A	tral 10 Nod entrality Betwee Score Molect 0.450 TAF1 0.430 CDKN2	enness Centrali Ile Score 0.239 A 0.230	d Disease ty Closeness (Molecule TAF1 CDKN2A	Centrality Score 0.623 0.605	Cytoscape <u>Visualize the Type 2 Diabetes specific networ</u> on Cytoscape
Most Cent Degree C Molecule TAF1 CDKN2A HNF4A	tral 10 Nod entrality Betwee Score Molect 0.450 TAF1 0.430 CDKN2 0.269 HNF4A	enness Centrali ile Score 0.239 A 0.230 0.105	d Disease y Closeness Molecule TAF1 CDKN2A HNF4A	Genes) Centrality Score 0.623 0.605 0.557	Cytoscape <u>Visualize the Type 2 Diabetes specific networ</u> on Cytoscape
Most Cent Degree C Molecule TAF1 CDKN2A HNF4A PPARG	Score Molect 0.450 TAF1 0.430 CDKN2 0.269 HNF4A 0.251 PPARC	es (Inferred enness Centrali ile Score 0.239 A 0.230 0.105 2 0.091	d Disease	Genes) Score 0.623 0.605 0.557 0.552	Cytoscape <u>Visualize the Type 2 Diabetes specific networ</u> on Cytoscape
Most Cent Degree C Molecule TAF1 CDKN2A HNF4A PPARG MYC	Score Molect 0.450 TAF1 0.430 CDKN2 0.269 HNF4A 0.251 PPARC 0.244 TCF7L2	es (Inferred enness Centrali ile Score 0.239 A 0.230 0.105 2 0.091 2 0.061	d Disease y Closeness Molecule TAF1 CDKN2A HNF4A MYC E2F1	Genes) Score 0.623 0.605 0.557 0.552 0.515	Cytoscape <u>Visualize the Type 2 Diabetes specific networ</u> on Cytoscape
Most Cent Degree C Molecule TAF1 CDKN2A HNF4A PPARG MYC E2F1	Score Molect 0.450 TAF1 0.430 CDKN2 0.269 HNF4A 0.251 PPARC 0.244 TCF7L2 0.165 MYC	es (Inferred enness Centrali ile Score 0.239 A 0.230 0.105 2 0.091 2 0.061 0.054	d Disease y Closeness (Molecule TAF1 CDKN2A HNF4A MYC E2F1 SMAD4	Centrality Score 0.623 0.605 0.557 0.552 0.515 0.504	Cytoscape <u>Visualize the Type 2 Diabetes specific networ</u> on Cytoscape
Most Cent Degree C Molecule TAF1 CDKN2A HNF4A PPARG MYC E2F1 TCF7L2	Inc. Inteal geode trail 10 Nod entrality Betwee Score Molect 0.450 TAF1 0.430 CDKN2 0.269 HNF4A 0.251 PPARC 0.244 TCF7L2 0.165 MYC 0.154 KCNJ1	es (Inferred enness Centrali Ile Score 0.239 A 0.230 0.105 2 0.091 2 0.061 0.054 1 0.026	d Disease y Closeness (Molecule TAF1 CDKN2A HNF4A MYC E2F1 SMAD4 JUN	Centrality Score 0.623 0.605 0.557 0.552 0.515 0.504 0.499	Cytoscape <u>Visualize the Type 2 Diabetes specific networ</u> on Cytoscape
Most Cent Degree C Molecule TAF1 CDKN2A HNF4A PPARG MYC E2F1 TCF7L2 JUN	Inc. Inteal geode trail 10 Nod entrality Betwee Score Molecu 0.450 TAF1 0.430 CDKN2 0.269 HNF4A 0.251 PPARC 0.244 TCF7L2 0.165 MYC 0.154 KCNJ11 0.143 JUN	enness Centrali Ile Score 0.239 A 0.230 0.105 0.091 0.061 0.054 1 0.026 0.024	d Disease y Closeness (Molecule TAF1 CDKN2A HNE4A MYC E2F1 SMAD4 JUN HNRNPR	Centrality Score 0.623 0.605 0.557 0.552 0.515 0.504 0.499 0.499	Cytoscape <u>Visualize the Type 2 Diabetes specific networ</u> on Cytoscape



References:

- A. Ozgur, T. Vu, G. Erkan, and D. R. Radev. "Identifying gene-disease associationsusing centrality on a literature mined gene interaction network", Bioinformatics, Vol. 24, Num. 13, pp. i277-i285, 2008.

- D. R. Radev, M. Hodges, A. Fader, M. Joseph, J. Gerrish, M. Schaller, J. dePeri, and B. Gibson. Clairlib documentation v1.03. Technical Report CSE-TR-536-07, University of Michigan. Department of EECS, 2007.

- Scott *et al*. Genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science. 2007; 316(5829):1341-5. 20.

- Serretti and Mandelli. The genetics of bipolar disorder: genome 'hot regions,' genes, new potential candidates and future directions. Mol Psychiatry. 2008 Aug;13(8):742-71.

Acknowledgements

This work was supported by National Institutes of Health: Grant #U54 DA021519.

We would like to thank Jing Gao, Glenn Tarcea, Aaron Bookvich, Terry Weymouth, and H. V. Jagadish for their

contributions to the project.