

PPIGCF: A Protein–Protein Interaction-Based Gene Correlation Filter for Optimal Gene Selection

Soumen Kumar Pati ^{1,†}, Manan Kumar Gupta ^{1,†}, Ayan Banerjee ^{2,†}, Saurav Mallik ^{3,4,5}
and Zhongming Zhao ^{3,6,*}

¹ Department of Bioinformatics, Maulana Abul Kalam Azad University of Technology, Haringhata 741249, West Bengal, India

² Department of Computer Science and Engineering, Jalpaiguri Govt. Engineering College, Jalpaiguri 735102, West Bengal, India

³ Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; smallik@arizona.edu

⁴ Department of Environmental Health, Harvard T H Chan School of Public Health, Boston, MA 02115, USA

⁵ Department of Pharmacology & Toxicology, University of Arizona, Tucson, AZ 85721, USA

⁶ Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

* Correspondence: zhongming.zhao@uth.tmc.edu

† These authors contributed equally to this work.

Supplementary Materials

Table S1: Gene classification based on GO similarity

GO ID	CC	Leukemia	Colon	DLBCL	Lung	Prostate
GO:0003674	Molecular function	712	223	797	1392	1400
GO:0005764	Lysosome	729	222	791	1091	1400
GO:0005783	endoplasmic reticulum	793	225	393	1398	1400
GO:0005794	Golgi apparatus	792	220	765	1456	1400
GO:0005886	plasma membrane	761	211	987	1154	1400
GO:0008150	Biological process	994	229	1567	1393	1400
GO:0016021	An integral component of the membrane	345	222	723	1241	1400
GO:0005737	Cytoplasm	1093	223	696	1397	1400
GO:0015630	Microtubule cytoskeleton	793	221	799	1345	1400

Table S2: Gene reduction based on PPI network

GO ID	Leukemia		Colon		DLBCL		Lung		Prostate	
	# genes	IIG	#genes	IIG	# genes	IIG	#genes	IIG	# genes	IIG
GO:0003674	712	39	223	23	797	139	1392	34	1400	186
GO:0005764	729	135	222	22	791	118	1091	86	1400	137
GO:0005783	793	61	225	25	393	91	1398	186	1400	194
GO:0005794	792	43	220	20	765	54	1456	137	1400	139
GO:0005886	761	77	211	21	987	131	1154	194	1400	118
GO:0008150	994	131	229	29	1567	103	1393	84	1400	131
GO:0016021	345	23	222	22	723	89	1241	61	1400	103
GO:0005737	1093	59	223	23	696	24	1397	7	1400	166
GO:0015630	793	97	221	21	799	166	1345	30	1400	131

Table S3: Elimination of genes through PCC and NSCC

GO ID	Leukemia			Colon			DLBCL		
	#gene	PCC	NSCC	#gene	PCC	NSCC	#gene	PCC	NSCC
GO:0003674	473	25	4	182	23	3	558	68	86
GO:0005764	434	20	55	171	22	2	573	95	38
GO:0005783	632	41	66	157	25	5	202	75	53
GO:0005794	643	19	37	200	20	2	511	78	85

GO:0005886	584	73	73	190	21	10	856	97	47
GO:0008150	763	72	79	154	29	9	364	28	50
GO:0016021	222	89	81	162	22	12	534	93	17
GO:0005737	144	84	5	197	23	3	572	10	21
GO:0015630	496	68	87	102	21	12	533	14	19
GO ID	Lung			Prostate					
	#gene	PCC	NSCC	#gene	PCC	NSCC			
GO:0003674	358	96	74	357	186	34			
GO:0005764	1105	98	12	1003	137	86			
GO:0005783	212	39	40	1211	194	186			
GO:0005794	1139	57	55	318	139	137			
GO:0005886	960	97	6	962	118	194			
GO:0008150	309	64	31	319	131	84			
GO:0016021	1180	83	96	1170	103	61			
GO:0005737	1139	70	51	380	166	7			
GO:0015630	315	30	46	325	131	30			

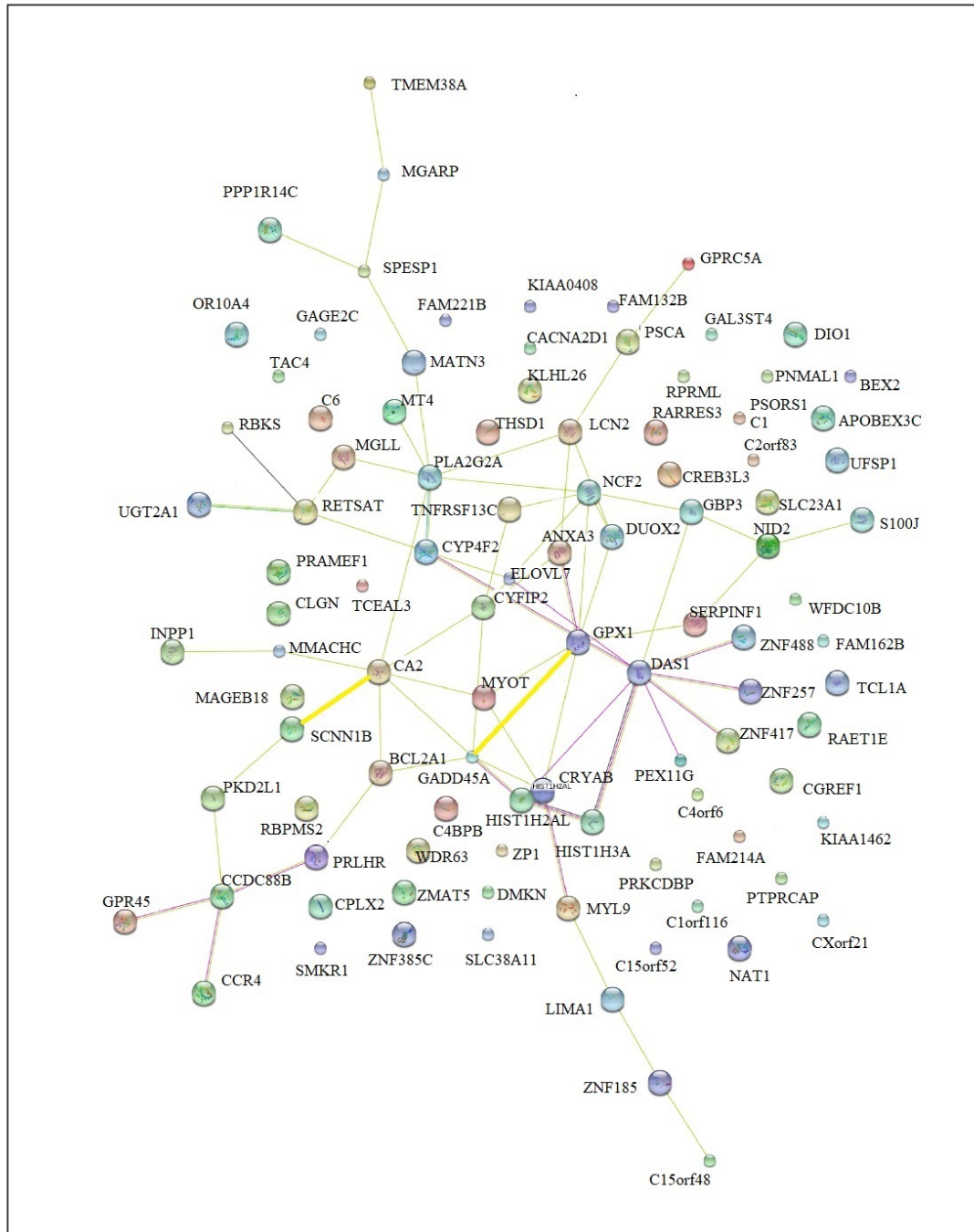
Table S4: Elimination of genes with respect to IC value

GO ID	Leukemia			Colon			DLBCL		
	#genes	< δ	#final genes	#genes	< δ	#final genes	#genes	< δ	#final genes
GO:0003674	644	97	547	174	33	141	504	130	374
GO:0005764	459	9	450	176	32	144	540	63	477
GO:0005783	627	138	489	170	35	135	164	4	160
GO:0005794	687	159	528	178	30	148	548	139	409
GO:0005886	538	121	417	159	31	128	712	145	567
GO:0008150	712	45	667	162	39	123	386	96	190

GO:0016021	152	7	145	166	32	134	534	71	463
GO:0005737	955	126	829	174	33	141	641	58	583
GO:0015630	541	85	456	167	31	136	600	182	418

GO ID	Lung			Prostate		
	#genes	< δ	#final genes	#gene	< δ	#final genes
GO:0003674	188	114	74	358	139	219
GO:0005764	895	194	601	1005	118	887
GO:0005783	133	56	77	212	91	121
GO:0005794	207	145	62	319	54	265
GO:0005886	857	94	763	960	131	829
GO:0008150	214	69	145	309	103	206
GO:0016021	1001	101	900	180	79	101
GO:0005737	269	100	169	390	24	366
GO:0015630	239	62	177	315	166	149

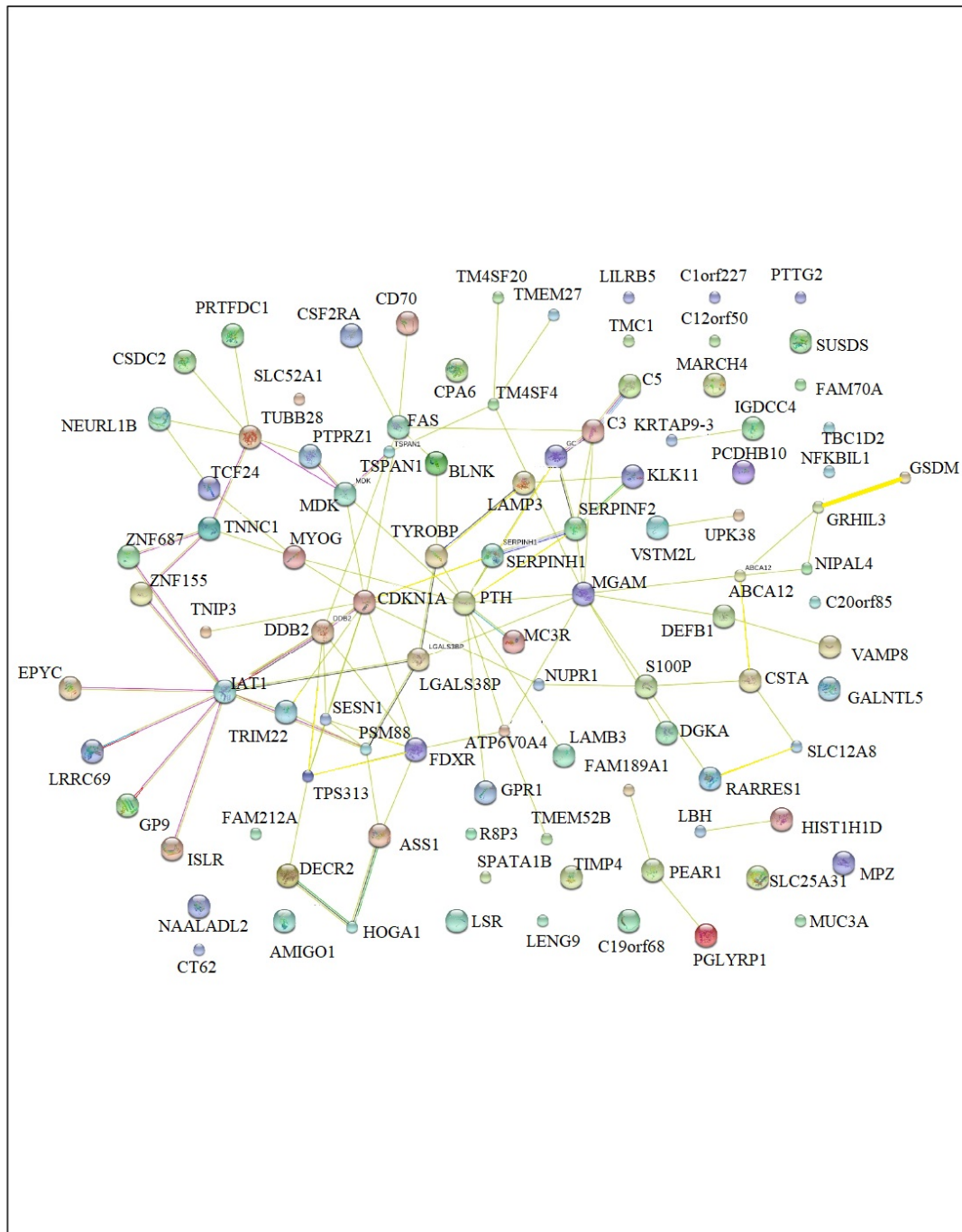
proteins: 100
interactions: 2
expected interactions: 2 (p-value: 0.637548993431305)



<http://version10.string-db.org/10/p/2126463091>

(a)

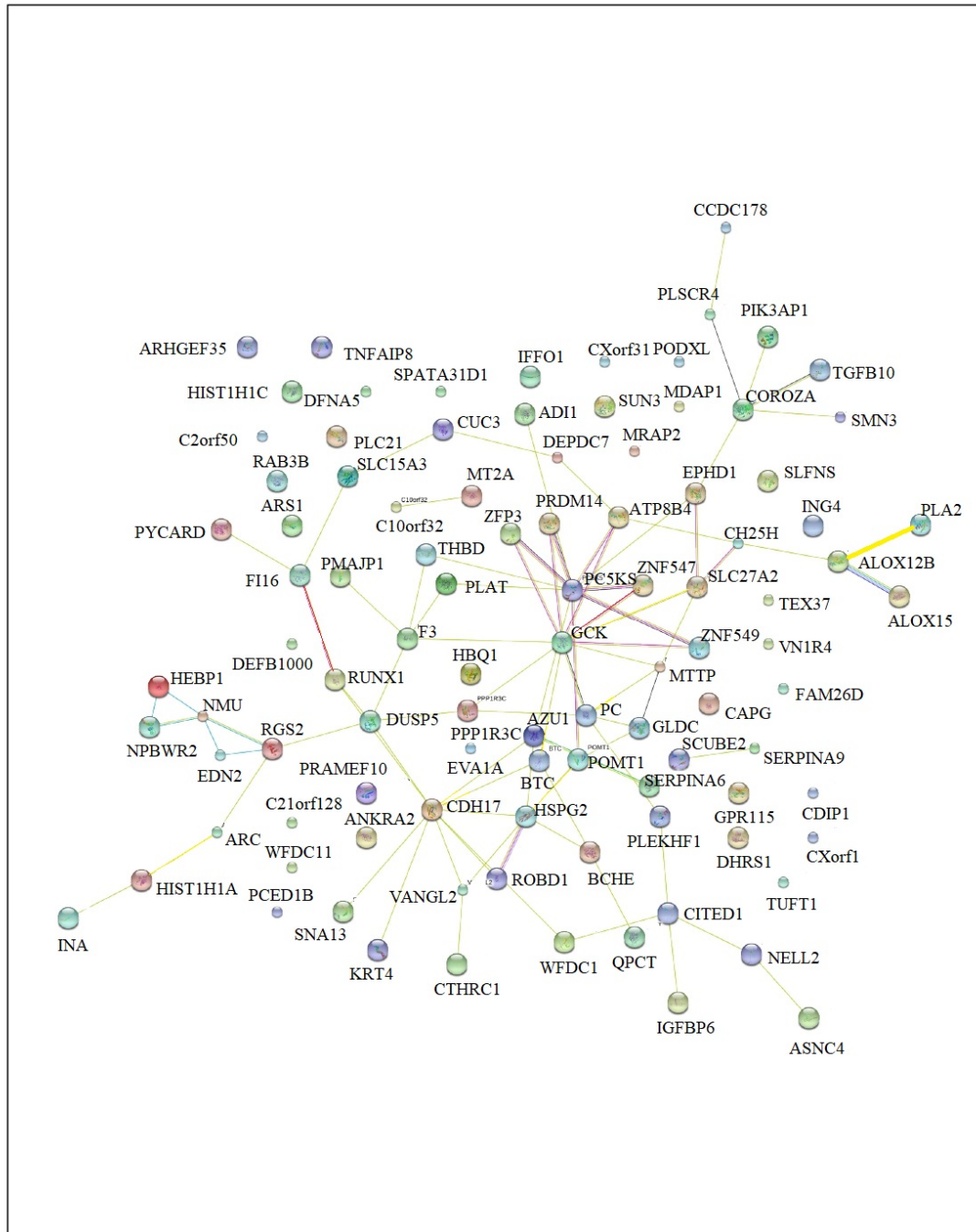
proteins: 100
interactions: 2
expected interactions: 1 (p-value: 0.573393582116955)



<http://version10.string-db.org/10/p/1257463092>

(b)

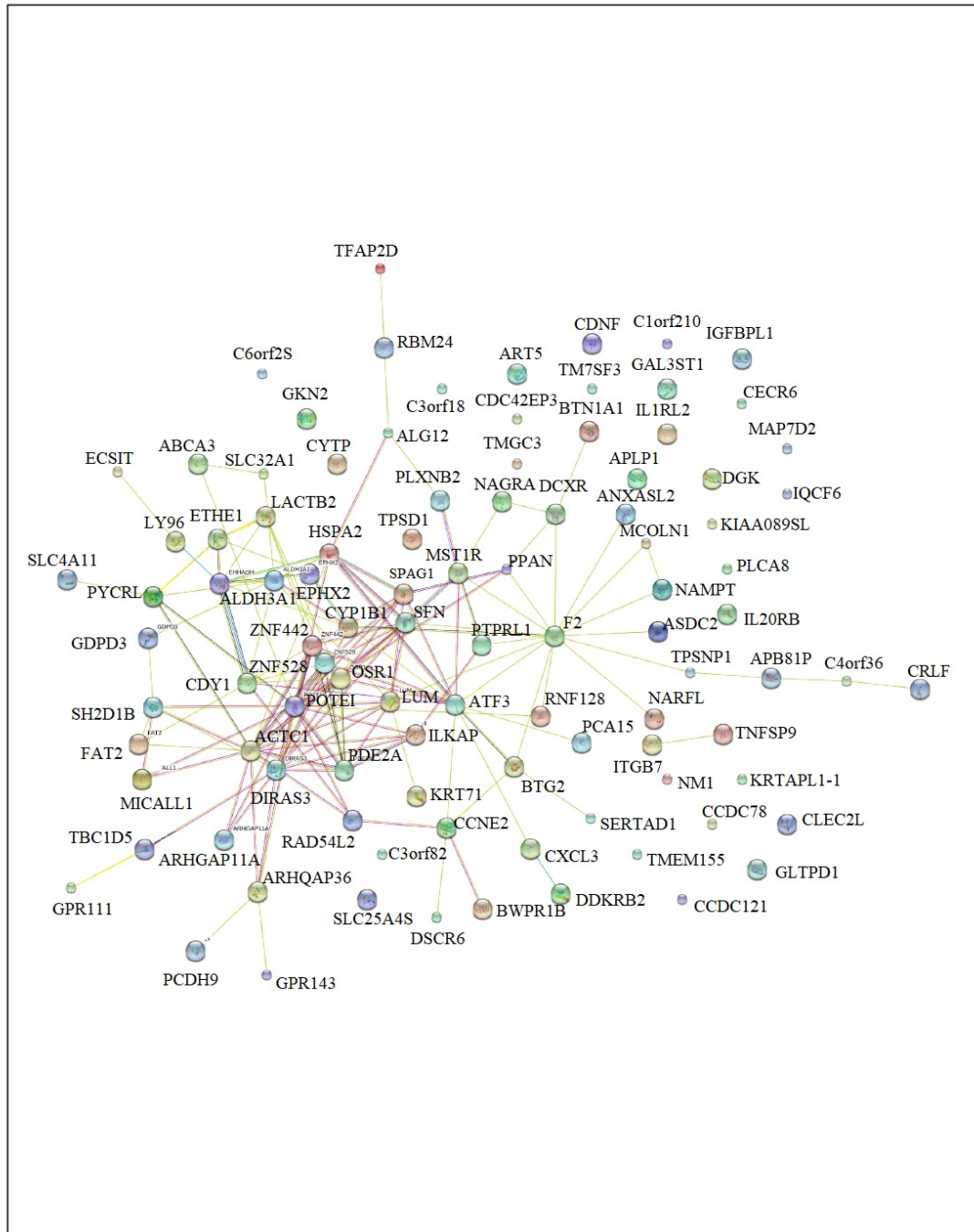
proteins: 100
interactions: 12
expected interactions: 7 (p-value: 0.0942060734484242)



<http://version10.string-db.org/10/p/7463463097>

(c)

proteins: 100
interactions: 1
expected interactions: 1 (p-value: 0.807600234904562)



<http://version10.string-db.org/10/p/2831463107>

(d)

