Supplementary: Graph-based Hub Gene Selection Technique using Protein Interaction Information: Application to Sample Classification

Pratik Dutta, *Student Member, IEEE,* Sriparna Saha, *Member, IEEE*, and Saurabh Gulati

I. Complexity Analysis of Modified Goldberg Algorithm

The complexity of the proposed modified Goldberg algorithm (refer Algorithm-1) is analyzed as follows:

1) Line 1 and line 2 of Algorithm-1 require constant time.
2) Line 3, contains a *while* loop, which executes $\lceil \log((m+1)n(n-1)) \rceil = \mathcal{O}(\log(n))$. Here, n is the number of nodes (genes) and m is the number of edges.
3) For each execution, two steps are dominating
   a) *Creation of network*:- For assigning weights to all m edges, the time complexity is $\mathcal{O}(m)$. Then for n nodes, we construct edges between source(s) and sink (t) and assign weights. This step takes $\mathcal{O}(n)$ time. Hence, the time complexity for constructing the network is $\mathcal{O}(m+n) = \mathcal{O}(m)$
   b) *Finding min-cut(S,T)*: For finding min-cut of the network, it takes $\mathcal{O}(nm\log(n))$ [1].

Therefore, the general time complexity for finding dense subgraphs (DSG) from a induced subgraph ($\mathcal{N}_i^{\mathcal{T}}$) is

$$\begin{aligned} T(n) &\leq C_1 \log(n)[C_2 m + C_3 nm \log(n)] \\ &\leq C_1 \log(n) C_3 nm \log(n) \\ &\leq Cnm[\log(n)]^2 \\ T(n) &= \mathcal{O}(nm[\log(n)]^2) \end{aligned} \tag{1}$$

As there are K induced subgraphs, total time complexity of the proposed modified Goldberg algorithm becomes $\mathcal{O}(Knm[\log(n)]^2)$.

TABLE I: Parameter setting for our proposed clustering approach

Parameters	Values
Number of generations	50
Population size	35
Probability of crossover	0.8
Mutation strength	0.2
μ_n	< 0.7
μ_i	$0.7 \geq$ and < 0.85
μ_d	≥ 0.85
K_{min}	2
K_{max}	$\sqrt{N}$, N is the number of genes

TABLE II: Paramaterization of the Binary Classifiers

Classifiers	Parameters	Description
*k*NN	n_neighbors = 3	Number of neighbors
	weights = 'uniform'	All points in each neighborhood are weighted equally.
Random Forest	n_estimators: 50	Number of trees in the forest
Support Vector Machine (SVM)	kernel= 'rbf'	Radial basis function kernel
Artificial Neural network (ANN)	hidden_layer_sizes = (h1,h2) h1, h2 = (input_layer_size + output_layer_size)/2	Number of neurons in the i^{th} hidden layer.
	activation : 'relu'	Activation function for the hidden layer.
	solver: 'adam'	Solver for weight optimization

TABLE III: Disease-Gene Association Table

Dataset	List of the selected genes which belong to the top 10 disease-related genes
B-cell chronic lymphocytic leukemia (B-CLL)	TP53, ATM, CCND1, P2RX7, IGHV3-21, ARL11, BCL2, ITGA4, MTHFR, CD5
Interstitial lung disease (ILD)	DCAF7, ABCA3, SFTPC
Prostate cancer	AR, ERG, TGFB1, TP53, IL6, ATM NFE2L2, TP53
Colon epithelial biopsies of ulcerative colitis (GDS3268)	NOD2, IL23R, HLA-DRB1, TNF, IL10 IL1B, TNFSF15, NKX2-3, ICAM1, STAT3
Myelodysplastic syndrome (GDS3795)	KMT2A, BCL2L10, YWHAE, U2AF1, TP53 NLRP2, TET2, PAFAH1B1, NBN, HPGDS
Pediatric Acute Leukemia with early relapse (GDS4206)	FLT3, KMT2A, NUP98, WT1, KIT NPM1, RUNX1T1, NSD1, PTPN11

TABLE IV: Summary of the nine binary classification datasets (six NCBI GEO datasets and three simulated datasets) used in this study

Dataset	Samples	Description
B-cell chronic lymphocytic leukemia (B-CLL)[2]	21	11 stable and 10 progressive
Interstitial lung disease (ILD)[3]	29	6 normal and 23 disease-related
Prostate cancer[4]	104	34 normal and 70 disease-related
Colon epithelial biopsies of ulcerative colitis (GDS3268)[5]	202	73 normal and 129 diseased
Myelodysplastic syndrome (GDS3795)[6]	200	183 diseased and 17 healthy control
Pediatric Acute Leukemia with early relapse (GDS4206)[7]	197	40 relapsed and 157 non-relapsed
Simulated_Dataset_1	500	250 target and 250 control
Simulated_Dataset_2	700	350 target and 350 control
Simulated_Dataset_3	800	400 target and 400 control

References

[1] A. Goldberg and R. Tarjan, "Solving minimum-cost flow problems by successive approximation," in *Proceedings of the nineteenth annual ACM symposium on Theory of computing*. ACM, 1987, pp. 7–18.

[2] S. Fält, M. Merup, G. Gahrton, B. Lambert, and A. Wennborg, "Identification of progression markers in b-cll by gene expression profiling," *Experimental hematology*, vol. 33, no. 8, pp. 883–893, 2005.

[3] J.-H. Cho, R. Gelinas, K. Wang, A. Etheridge, M. G. Piper, K. Batte, D. Dakhlallah, J. Price, D. Bornman, S. Zhang *et al.*, "Systems biology of interstitial lung diseases: integration of mrna and microrna expression changes," *BMC medical genomics*, vol. 4, no. 1, p. 8, 2011.

TABLE V: Summarization of Performance Metrics for B-CLL Dataset; *DSG*: Dense Sub-graph module number, *3NN*: 3 Nearest Neighbors, *RF*: Random Forest, *SVM*: Support Vector Machine, *Sen*: Sensitivity, *Spec*: Specificity, *FM*: F-measure, *MCC*: Matthews correlation coefficient

DSG	CL	3NN					RF					SVM				
		Sen	Spec	Prec	FM	MCC	Sen	Spec	Prec	FM	MCC	Sen	Spec	Prec	FM	MCC
Baseline 1	**N**	0.45	0.58	0.44	0.44	0.02	0.63	0.38	0.38	0.47	0.01	0.40	0.73	0.524	0.583	0.06
	P	0.58	0.4	0.58	0.58	0.02	0.38	0.63	0.63	0.48	0.01	0.73	0.40	0.314	0.488	0.06
Baseline 2	**N**	0.45	0.70	0.63	0.53	0.15	0.67	0.47	0.33	0.44	0.12	0.70	0.36	0.5	0.58	0.07
	P	0.70	0.45	0.54	0.61	0.15	0.47	0.67	0.78	0.58	0.12	0.37	0.70	0.57	0.44	0.07
DSG 1	**N**	0.70	0.36	0.5	0.58	0.07	0.90	0.27	0.53	0.67	0.22	0.72	0.43	0.38	0.50	0.14
	P	0.37	0.70	0.57	0.44	0.07	0.27	0.90	0.75	0.40	0.22	0.43	0.72	0.75	0.55	0.14
DSG 2	**N**	0.75	0.46	0.46	0.57	0.21	0.75	0.62	0.55	0.63	0.35	0.72	0.36	0.36	0.48	0.07
	P	0.46	0.75	0.75	0.57	0.21	0.62	0.75	0.80	0.70	0.35	0.36	0.72	0.72	0.47	0.07
DSG 3	**N**	0.60	0.45	0.50	0.55	0.06	0.70	0.36	0.5	0.58	0.07	0.25	0.77	0.40	0.31	0.02
	P	0.45	0.60	0.56	0.50	0.06	0.37	0.70	0.57	0.44	0.07	0.77	0.25	0.63	0.69	0.02
DSG 4	**N**	0.78	0.34	0.47	0.58	0.12	0.64	0.43	0.69	0.67	0.06	0.70	0.45	0.54	0.61	0.16
	P	0.34	0.78	0.67	0.45	0.12	0.43	0.64	0.37	0.40	0.06	0.45	0.70	0.63	0.53	0.16
DSG 5	**N**	0.70	0.36	0.5	0.58	0.07	0.86	0.36	0.40	0.55	0.22	0.42	0.78	0.72	0.53	0.20
	P	0.37	0.70	0.57	0.44	0.07	0.36	0.86	0.83	0.50	0.22	0.78	0.42	0.50	0.61	0.20
DSG 6	**N**	0.64	0.50	0.58	0.61	0.14	0.72	0.36	0.36	0.48	0.07	0.72	0.43	0.38	0.50	0.14
	P	0.50	0.64	0.56	0.53	0.14	0.36	0.72	0.72	0.47	0.07	0.43	0.72	0.75	0.55	0.14
DSG 7	**N**	0.40	0.83	0.86	0.55	0.23	0.70	0.45	0.54	0.61	0.16	0.86	0.36	0.40	0.55	0.22
	P	0.83	0.40	0.36	0.50	0.23	0.45	0.70	0.63	0.53	0.16	0.36	0.86	0.83	0.50	0.22
DSG 8	**N**	0.70	0.36	0.5	0.58	0.07	0.70	0.45	0.53	0.61	0.16	0.70	0.36	0.5	0.58	0.07
	P	0.37	0.70	0.57	0.44	0.07	0.45	0.70	0.63	0.53	0.16	0.37	0.70	0.57	0.44	0.07
DSG 9	**N**	0.72	0.36	0.36	0.48	0.07	0.63	0.38	0.38	0.47	0.01	0.60	0.45	0.50	0.55	0.06
	P	0.36	0.72	0.72	0.47	0.07	0.38	0.63	0.63	0.48	0.01	0.45	0.60	0.56	0.50	0.06
HG 3	**N**	0.66	0.77	0.8	0.72	0.44	0.727	0.40	0.573	0.64	0.145	0.47	0.75	0.88	0.61	0.17
	P	0.77	0.66	0.636	0.70	0.44	0.40	0.727	0.573	0.471	0.145	0.75	0.471	0.25	0.375	0.17
HG 5	**N**	0.667	0.417	0.462	0.545	0.08	0.778	0.667	0.636	0.70	0.44	0.667	0.667	0.727	0.695	0.33
	P	0.416	0.667	0.625	0.50	0.08	0.667	0.778	0.80	0.727	0.44	0.667	0.667	0.60	0.632	0.33
HG 17	**N**	0.667	0.583	0.545	0.60	0.248	0.455	0.70	0.625	0.526	0.159	0.556	0.50	0.454	0.50	0.05
	P	0.583	0.667	0.70	0.636	0.248	0.70	0.454	0.538	0.609	0.159	0.50	0.556	0.60	0.545	0.05
HG 33	**N**	0.462	0.75	0.75	0.571	0.211	0.571	0.714	0.80	0.667	0.27	0.40	0.667	0.75	0.522	0.06
	P	0.75	0.462	0.462	0.571	0.211	0.714	0.571	0.454	0.556	0.27	0.667	0.40	0.307	0.421	0.06
HG 55	**N**	0.80	0.273	0.50	0.615	0.085	0.778	0.25	0.437	0.56	0.032	0.385	0.625	0.625	0.476	0.009
	P	0.273	0.80	0.60	0.375	0.085	0.25	0.778	0.60	0.354	0.032	0.625	0.385	0.385	0.471	0.009

TABLE VI: Comparison of Different Performance Metrics for Simulated Datasets

		Ge et. al (2016)				Radovic et. al (2017)				Kang et. al (2017)				Proposed Method			
Dataset	Performance metrics	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN
Simulated_Dataset_1	Accuracy	0.67	0.80	0.70	0.69	0.68	0.81	0.76	0.80	0.67	0.68	0.71	0.78	**0.85**	**0.82**	**0.82**	**0.90**
	Precision	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.79	0.78	0.80	0.80	**0.80**	**0.80**	**0.80**	**0.85**
	F-measure	0.73	0.80	0.74	0.62	0.73	0.70	0.78	0.62	0.65	0.63	0.70	0.73	**0.82**	**0.84**	**0.81**	**0.81**
	MCC	-0.18	-0.01	-0.16	-0.01	-0.17	-0.01	-0.09	-0.06	-0.04	-0.14	-0.10	-0.12	**-0.02**	**-0.01**	**0.09**	**0.12**
Simulated_Dataset_2	Accuracy	0.66	0.79	0.71	0.73	0.66	0.71	0.72	0.70	0.67	0.70	0.66	0.70	**0.80**	**0.82**	**0.85**	**0.91**
	Precision	0.80	0.79	0.80	0.60	0.80	0.80	0.80	0.70	0.66	0.73	0.70	0.63	**0.80**	**0.80**	**0.80**	**0.80**
	F-measure	0.72	0.81	0.75	0.71	0.72	0.79	0.75	0.61	0.66	0.63	0.70	0.71	**0.81**	**0.81**	**0.80**	**0.84**
	MCC	-0.19	-0.06	-0.15	-0.01	-0.19	-0.03	-0.19	0.27	-0.17	-0.03	-0.19	-0.19	**0.12**	**0.08**	**0.12**	**0.32**
Simulated_Dataset_3	Accuracy	0.70	0.80	0.69	0.80	0.73	0.71	0.68	0.80	0.71	0.70	0.69	0.65	**0.80**	**0.77**	**0.82**	**0.88**
	Precision	0.80	0.72	0.80	0.70	0.80	0.80	0.80	0.70	0.74	0.63	0.79	0.80	**0.81**	**0.80**	**0.80**	**0.80**
	F-measure	0.75	0.80	0.74	0.80	0.76	0.71	0.73	0.79	0.76	0.73	0.73	0.79	**0.81**	**0.80**	**0.81**	**0.83**
	MCC	-0.16	-0.01	-0.16	-0.03	-0.13	-0.01	-0.18	-0.05	-0.04	-0.17	-0.03	-0.14	**0.24**	**0.22**	**0.24**	**0.32**

TABLE VII: Summarization of Performance Metrics for ILD Dataset; *DSG*: Dense Sub-graph module number, *3NN*: 3 Nearest Neighbors, *RF*: Random Forest, *SVM*: Support Vector Machine, *Sen*: Sensitivity, *Spec*: Specificity, *FM*: F-measure, *MCC*: Matthews correlation coefficient

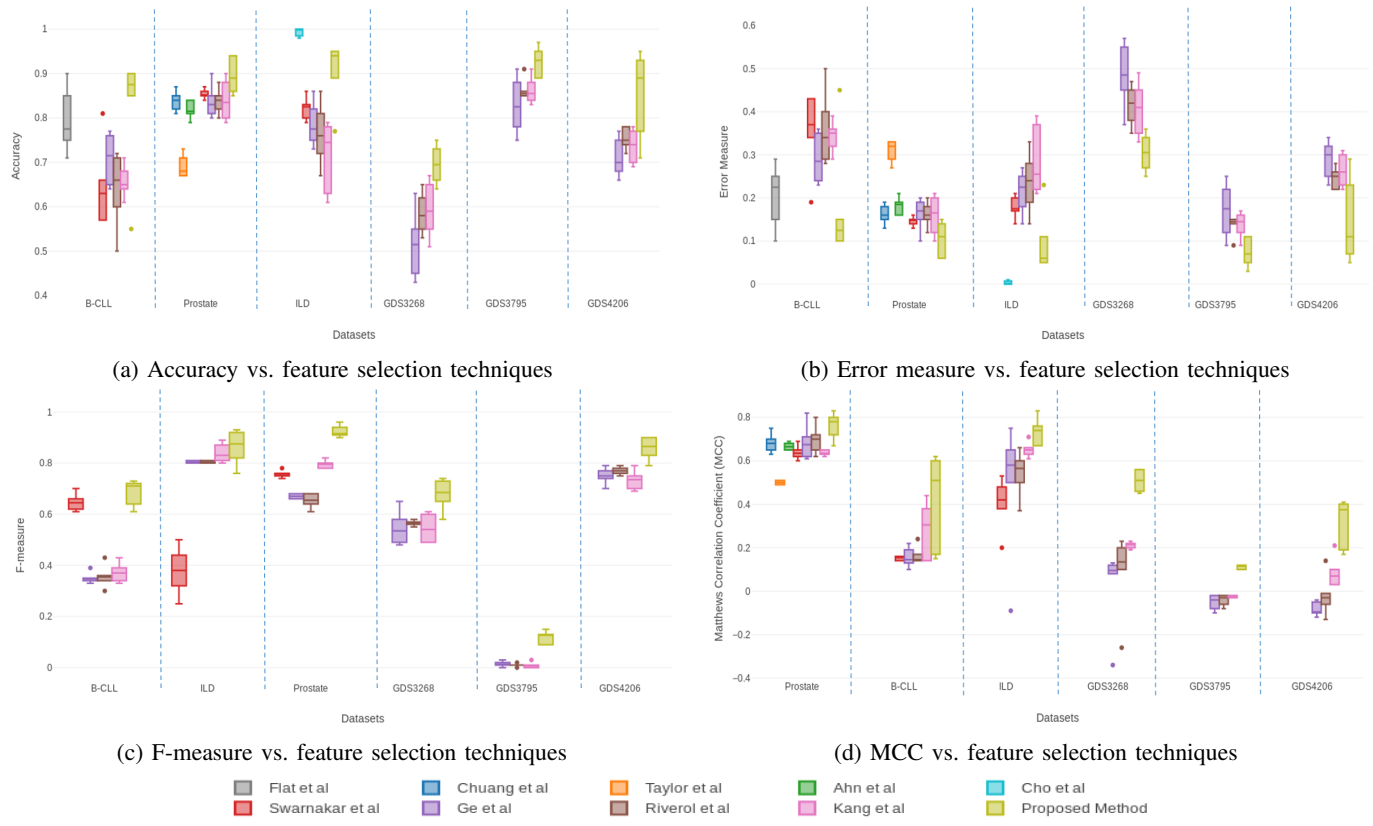
DSG	CL	3NN					RF					SVM				
		Sen	Spec	Prec	FM	MCC	Sen	Spec	Prec	FM	MCC	Sen	Spec	Prec	FM	MCC
Baseline 1	**N**	0.45	0.64	0.67	0.53	0.08	0.47	0.58	0.62	0.53	0.05	0.78	0.27	0.50	0.61	0.06
	P	0.64	0.44	0.41	0.50	0.07	0.58	0.47	0.44	0.50	0.05	0.27	0.78	0.57	0.36	0.06
Baseline 2	**N**	0.45	0.71	0.83	0.59	0.15	0.48	0.67	0.85	0.61	0.12	0.45	0.64	0.67	0.53	0.08
	P	0.71	0.45	0.29	0.42	0.15	0.67	0.48	0.25	0.36	0.12	0.64	0.44	0.41	0.50	0.07
HG 3	**N**	0.86	0.86	0.95	0.90	0.67	0.90	0.87	0.95	0.93	0.75	0.87	0.77	0.82	0.85	0.65
	P	0.86	0.86	0.67	0.90	0.67	0.87	0.90	0.78	0.82	0.75	0.77	0.87	0.83	0.80	0.65
HG 5	**N**	0.818	0.444	0.474	0.60	0.27	0.928	0.60	0.68	0.78	0.56	0.75	0.471	0.50	0.60	0.224
	P	0.444	0.818	0.80	0.571	0.27	0.60	0.928	0.90	0.72	0.56	0.471	0.75	0.73	0.57	0.224
HG 17	**N**	0.81	0.85	0.87	0.84	0.66	0.83	0.91	0.94	0.88	0.73	0.56	0.91	0.91	0.67	0.44
	P	0.85	0.81	0.79	0.81	0.66	0.91	0.83	0.77	0.83	0.73	0.91	0.56	0.53	0.67	0.44
HG 33	**N**	0.42	0.80	0.80	0.55	0.22	0.47	0.83	0.80	0.59	0.32	0.41	0.92	0.88	0.56	0.36
	P	0.80	0.42	0.42	0.55	0.22	0.83	0.47	0.53	0.65	0.32	0.92	0.41	0.52	0.67	0.36

[4] X. Ren, Y. Wang, X.-S. Zhang, and Q. Jin, "ipcc: a novel feature extraction method for accurate disease class discovery and prediction," *Nucleic acids research*, vol. 41, no. 14, pp. e143–e143, 2013.

[5] C. L. Noble, A. R. Abbas, J. Cornelius, C. W. Lees, G.-T. Ho, K. Toy, Z. Modrusan, N. Pal, F. Zhong, S. Chalasani *et al.*, "Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis,"

TABLE VIII: Summarization of Performance Metrics for Prostate Dataset; *DSG*: Dense Sub-graph module number, *3NN*: 3 Nearest Neighbors, *RF*: Random Forest, *SVM*: Support Vector Machine, *Sen*: Sensitivity, *Spec*: Specificity, *FM*: F-measure, *MCC*: Matthews correlation coefficient

DSG	CL	3NN					RF					SVM				
		Sen	Spec	Prec	FM	MCC	Sen	Spec	Prec	FM	MCC	Sen	Spec	Prec	FM	MCC
Baseline 1	**N**	0.89	0.36	0.89	0.89	0.23	0.88	0.36	0.89	0.88	0.22	0.78	0.36	0.77	0.77	0.13
	P	0.36	0.89	0.33	0.34	0.23	0.36	0.88	0.31	0.33	0.22	0.36	0.78	0.37	0.36	0.13
Baseline 2	**N**	0.81	0.50	0.80	0.81	0.31	0.80	0.50	0.88	0.84	0.26	0.75	0.75	0.85	0.79	0.48
	P	0.0	0.81	0.52	0.51	0.31	0.50	0.80	0.34	0.40	0.26	0.75	0.75	0.61	0.67	0.48
HG 3	**N**	0.87	0.86	0.94	0.90	0.69	0.89	0.79	0.92	0.91	0.67	0.92	0.83	0.96	0.94	0.70
	P	0.86	0.87	0.71	0.78	0.69	0.79	0.89	0.73	0.76	0.67	0.83	0.92	0.68	0.75	0.70
HG 5	**N**	0.90	0.58	0.80	0.84	0.52	0.87	0.59	0.84	0.86	0.46	0.81	0.76	0.90	0.85	0.54
	P	0.58	0.90	0.75	0.66	0.52	0.59	0.87	0.63	0.61	0.46	0.76	0.81	0.61	0.68	0.54
HG 17	**N**	0.69	0.77	0.83	0.76	0.45	0.60	0.81	0.85	0.70	0.39	0.62	0.90	0.91	0.73	0.50
	P	0.77	0.69	0.60	0.67	0.45	0.81	0.60	0.53	0.64	0.39	0.90	0.62	0.58	0.71	0.50
HG 33	**N**	0.81	0.59	0.80	0.81	0.41	0.78	0.73	0.80	0.79	0.51	0.78	0.68	0.77	0.78	0.47
	P	0.59	0.81	0.61	0.60	0.41	0.73	0.78	0.71	0.72	0.51	0.68	0.78	0.70	0.69	0.47
HG 55	**N**	0.80	0.59	0.76	0.78	0.40	0.82	0.47	0.63	0.71	0.31	0.75	0.70	0.70	0.72	0.44
	P	0.59	0.80	0.64	0.61	0.40	0.47	0.82	0.70	0.56	0.31	0.70	0.75	0.74	0.72	0.44
HG 85	**N**	0.56	0.73	0.76	0.64	0.28	0.58	0.75	0.76	0.66	0.33	0.67	0.74	0.76	0.71	0.41
	P	0.73	0.56	0.52	0.61	0.28	0.75	0.58	0.57	0.65	0.33	0.74	0.67	0.65	0.69	0.41



(a) Accuracy vs. feature selection techniques

(b) Error measure vs. feature selection techniques

(c) F-measure vs. feature selection techniques

(d) MCC vs. feature selection techniques

Fig. 1: Boxplots of the proposed graph-based feature selection method against different stat-of-the-art methods for accuracy, error measure, F-measure and MCC.

TABLE IX: p-values produced by Welch's t-test comparing our proposed method with other algorithms

	Falt et al (2006)	Chuang et al. (2007)	Taylor at al.(2009)	Ahn et al. (2011)	Cho et al. (2011)	Swarnakar et al. (2015)	Ge et al (2016)	Riverol et al. (2017)	Kang et al. (2017)
B-CLL	4.42E-31	-	-	-	-	5.84E-54	1.41E-43	4.92E-50	3.89E-51
ILD	-	-	-	-	1.66E-29	1.88E-37	1.91E-35	3.54E-37	7.34E-45
Prostate	-	3.21E-35	2.86E-54	3.02E-40	-	3.95E-35	1.05E-28	6.63E-33	2.73E-27
GDS3268	-	-	-	-	-	-	5.15E-43	1.13E-39	2.52E-34
GDS3765	-	-	-	-	-	-	8.57E-32	5.19E-29	3.07E-32
GDS4206	-	-	-	-	-	-	1.59E-47	5.74E-49	1.45E-50

Gut, 2008.

[6] A. Pellagatti, M. Cazzola, A. Giagounidis, J. Perry, L. Malcovati,

TABLE X: Biological functional analysis of different dense sub graph modules and hub gene modules for B-CLL dataset. DSG/HG: Name/ ID of the Dense Subgraph/ Hub gene modules, HG%: Percentage of Genes of the dense subgraph or hub gene modules related to the particular Gene Ontology(GO) term, Genome%: Percentage of Genes of the Gene Ontology Consortium related to the particular Gene Ontology(GO) term

DSG / HG	Gene ontology name/EA ID	Enrichment analysis name	Gene names	p-value	HG%	Genome%
DSG 1	GO:0065007	biological regulation	LCK,RAP1GAP,CASP2,IER3,PTPRD,MAPK9,PTPN2, STAT2,PRKD1,PRKDC,PRKCH,GUCA1A,PTPRS, GSTA2,DDB2,IL3,PIK3R1,ERF,QSOX1,GSK3B, SSX2B,ZNF23,PAK2,ERCC4,TGFB2,GAB1,POLR2H	4.29E-04	**87.09%**	57.88%
	GO:0071704	organic substance metabolic process	LCK,CASP2,PTPRD,UBA1,MAPK9,PTPN2,STAT2, PRKD1,PRKDC,PRKCH,PTPRS,GSTA2,DDB2,IL3, PIK3R1,ERF,QSOX1,GSK3B,SSX2B,ZNF23,PAK2, ERCC4,TGFB2,GSTA3,GAB1,POLR2H	1.22E-05	**83.87%**	44.85%
	GO:1901564	organonitrogen compound metabolic process	LCK,CASP2,PTPRD,UBA1,MAPK9,PTPN2,PRKD1, PRKDC,PRKCH,PTPRS,GSTA2,DDB2,IL3,PIK3R1, QSOX1,GSK3B,PAK2,TGFB2,GSTA3	3.74E-05	**61.29%**	26.01%
	GO:0048518	positive regulation of biological process	LCK,RAP1GAP,CASP2,IER3,PTPRD,MAPK9, PTPN2,PRKD1,PRKDC,PRKCH,PTPRS,GSTA2, IL3,PIK3R1,GSK3B,PAK2,TGFB2,GAB1	2.68E-04	**58.06%**	24.43%
DSG 2	GO:0009987	cellular process	CYP2A13,FLT1,TAF1,EGF,PAK1,JAK3,DUSP8, EGFR,IFNA5,PAK1,MAP2K6,IL2,HYAL1,RBM5, NRAS,AKT1,MRPL28,IL10,MRPL28,HIST1H2BJ, FOXN3,PAK2,BLM,SUMO1,CDC6,CYP2A7,GRB2, TNFRSF1A,IFNA6	8.47E-04	**96.67%**	71.13%
	GO:0050794	regulation of cellular process	FLT1,TAF1,EGF,PAK1,JAK3,DUSP8,EGFR,IFNA5, PAK1,MAP2K6,IL2,HYAL1,RBM5,NRAS,AKT1, IL10,HIST1H2BJ,FOXN3,PAK2,BLM,SUMO1, SERPINB4,CDC6,GRB2,TNFRSF1A,IFNA6	6.23E-05	**86.67%**	50.74%
	GO:0043170	macromolecule, metabolic process	FLT1,TAF1,EGF,PAK1,JAK3,DUSP8,EGFR, PAK1,MAP2K6,IL2,HYAL1,RBM5,NRAS, AKT1,MRPL28,IL10,MRPL28,HIST1H2BJ, FOXN3,PAK2,BLM,SUMO1,CDC6,GRB2	2.62E-06	**80.00%**	36.58%
HG3	GO:0008150	biological_process	CCNC,CASP10,CCR1,ADRB1,EPHA2,E2F4,BAX, KLK3,INS,SLA,MSH4,ERCC4,PRKCD,RBBP4,SRPK1, NKX2-1, NOS2,TGFB2,BAK1,CDK1,CDKN1C,FGF4, CDK8,CDH11,POLD2,SUMO1,PTPN11,RAD9A, MAP2K1,PSMB1,MLH1,PARP1	4.43E-47	**99.39%**	82.53%
	GO:0050794	regulation of cellular process	ANXA3,FAS,PIK3R1,APOB,ERF,TAF13,CDC23, MARK2,CUX1,DAP3,CDKN1A,NR2F1,RAD52, ADAM10,MTA1,IL1RN,RPS6KB1,PAX8,RAD51, PTN,PECAM1,CCNF,JUP,EFNB1,FGFR3	5.85E-113	**91.92%**	50.73%
	GO:0071704	organic substance metabolic process	PPP2CA,CSF1,IGFBP3,CFLAR,NR4A1,UBA1, HSF1,ARAF,CCNA2,MYB,OSM,MAPK9,CEACAM5, DUSP4,ERCC5,PTPN2,ACVRL1,SKP2,IL2,IL16, CDC25C,KDR,SKP1,ANGPT1,PSMA3,FN1,PLD1	3.38E-66,	**78.20%**	44.85%
	GO:0048519	negative regulation of biological process	PTN,CCNF,FGFR3,AIM2,PSMD2,HGF,EDNRB,ESRRB, TGFB1,TRAF2,IFIT5,RARA,TP53BP1,ADRA2A,AKT2, PSMB2,HIST1H2BJ,HIF1A,E2F1,BCL2L1,TRADD, ARHGAP1,SQSTM1,EPHA4,EPHB2,BDNF,EIF4EBP1, DDB1,	1.29E-97	**62.65%**	23.43%

M. Della Porta, M. Jädersten, S. Killick, A. Verma, C. Norbury *et al.*, "Deregulated gene expression pathways in myelodysplastic syndrome hematopoietic stem cells," *Leukemia*, vol. 24, no. 4, p. 756, 2010.

[7] L. H. Meyer, S. M. Eckhoff, M. Queudeville, J. M. Kraus, M. Giordan, J. Stursberg, A. Zangrando, E. Vendramini, A. Möricke, M. Zimmermann *et al.*, "Early relapse in all is identified by time to leukemia in nod/scid mice and is characterized by a gene signature involving survival pathways," *Cancer cell*, vol. 19, no. 2, pp. 206–217, 2011.