

List of Figures

1.1	Relationship between genes, proteins and phenotypic traits.	1
2.1	DNA structure (source- http://theinvestigation.yolasite.com/dna-structure.php , accessed on-14/12/2017)	10
2.2	Coding and non-coding regions of DNA (source- https://scienceoveracuppa.com/tag/fragile-x-syndrome/ , accessed on-14/12/2017)	10
2.3	Central dogma-Protein synthesis from DNA (source- http://fourier.eng.hmc.edu/bioinformatics/intro/node8.html , accesesed on 14/12/2017)	11
2.4	PPI network where the interaction between proteins i and j is shown as a thick edge in the graph (source- http://benthamscience.com/journal/abstracts.php , accesesed on 03/02/2016)	15
3.1	Performance of CNCM for various α threshold values for Gavin_2002 dataset.	42
3.2	Precision, Recall and F-measure of CNCM and other algorithms on the Gavin_2002 dataset using MIPS as benchmark.	43
3.3	Precision, Recall and F-measure of CNCM and other algorithms on the Gavin_2006 dataset using MIPS as benchmark.	44
3.4	Precision, Recall and F-measure of CNCM and other algorithms on the Krogan_2006 dataset using MIPS as benchmark.	45
3.5	Precision, Recall and F-measure of CNCM and other algorithms on the Tong_2004 dataset using MIPS as benchmark.	46
3.6	Average F-measure using Bader's overlapping scheme with MIPS as benchmark on Krogan_2006 dataset (Higher value implies better performance)	47
3.7	Co-localization score for four yeast datasets using Huh et al., localization dataset (Higher value implies better performance).	48
3.8	Co-localization score over four yeast datasets using Kumar et al. localization data.	49
3.9	Performance measures in terms of Sn, PPV and Acc of DCRS compared with other methods over HPRD dataset.	59
3.10	Performance measures of DCRS compared with other methods over HPRD dataset and variations in results using two different set of initial population.	60

4.1	Performance indices obtained at varying thresholds using HPRD dataset.	68
4.2	Comparing Sensitivity of CSC with other algorithms on DIP dataset using MIPS as benchmark.	71
4.3	Comparing Positive Predictive Value of CSC with other algorithms on DIP dataset using MIPS as benchmark.	72
4.4	Comparing Accuracy of CSC with other algorithms on DIP dataset using MIPS as benchmark.	73
4.5	Comparing Sensitivity of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.	74
4.6	Comparing Positive Predictive Value of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.	75
4.7	Comparing Accuracy of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.	76
4.8	Comparing Sensitivity, Positive Predictive Value and Accuracy of CSC with other algorithms on DIP dataset using CYC2008 as benchmark with varying α and β thresholds.	77
4.9	Comparing Sensitivity of CSC with other algorithms on HPRD dataset.	80
4.10	Comparing Positive Predictive Value of CSC with other algorithms on HPRD dataset.	81
4.11	Comparing Accuracy of CSC with other algorithms on HPRD dataset.	82
4.12	Disease gene-Member genes analysis framework	82
4.13	Protein complex members along with its association links	83
4.14	Performance indices obtained by varying α threshold on HPRD dataset.	85
4.15	Performance indices of ComFiR obtained at varying thresholds of β on HPRD dataset.	88
4.16	Positive Predictive Value and Sensitivity of ComFiR and other methods on DIP dataset.	89
4.17	Accuracy of ComFiR and other methods on DIP dataset	90
4.18	Comparison of ComFiR and other methods on DIP dataset in terms of Precision and F –measure at overlapping threshold of 0.2.	91
4.19	Positive Predictive Value and Sensitivity of ComFiR and other methods on HPRD dataset.	92
4.20	Accuracy of ComFiR and other methods on HPRD dataset	93
4.21	Complex Ranking steps for a given disease query.	95
4.22	Genes found in top five complexes using ComFiR method on Alzheimer’s disease (source-STRING tool)	96
5.1	Cancer cell growth leading to tumor and its spreading (source- http://www.unc.edu/depts/our/hhmi/hhmi-ft_learning_modules/cancermodule/pages/cancer.html , accessed on-20.09.2017)	101
5.2	Chemokine signalling pathway (source- http://www.kegg.jp/kegg/kegg1.html).107	

5.3	Tryptophan metabolism (source- http://www.kegg.jp/kegg/kegg1.html)	108
5.4	Common genes found among the five modules in nonmetastasis stage.	114
5.5	Common genes found among the two modules in metastasis stage. . .	114
5.6	Expression trend among module members in nonmetastasis stage. . .	117
5.7	Expression trend among module members in metastasis stage. . . .	118
5.8	Common genes found among the two modules in metastasis stage. . .	118
6.1	Genes g_1 and g_2 are shown connected with dotted line since Pearson correlation value between them is less than 0.5	125
6.2	Genes g_1 and g_2 are shown to be connected via three subset of edges.	126
6.3	Conceptual framework of module extraction method	132
6.4	Conceptual framework of module correspondence between control and disease stage	136
6.5	Expression patterns of six causal genes and new suspected genes . .	147
6.6	Visual representation of control module and disease module along with their hub genes using STRING tool.	151

List of Tables

2.1	Summary of PPI detection techniques	14
2.2	Datasets Used	23
3.1	Summary of some protein complex finding techniques	35
3.2	Comparison with MCODE, FAG-EC, FT, TFit, OCG, QCUT, ClusterONE and GMFTP in terms of p-value for Gavin_2002 dataset.	51
3.3	Comparison with MCODE, FAG-EC, FT, TFit, OCG, QCUT, ClusterONE and GMFTP in terms of p-value for Gavin_2006 dataset.	52
3.4	Comparison with MCODE, FAG-EC, FT, TFit, OCG, QCUT, ClusterONE and GMFTP in terms of p-value for Krogan_2006 dataset.	53
3.5	Comparison with MCODE, FAG-EC, FT, TFit, OCG, QCUT, ClusterONE and GMFTP in terms of p-value for Tong_2004 dataset.	54
4.1	Summary of protein complex finding techniques based on a combination of both topological and functional information	64
4.2	Alzheimer associated complex (Association of disease gene with other genes in the complex)	78
4.3	Pathway associated with each member of Alzheimer associated complexes	79
4.4	List of complexes associated with Alzheimer' s Disease	94
5.1	Number of gene samples drawn at varying variance threshold, V_{th}	104
5.2	p-value of top 3 modules obtained at different threshold value of $CCft$ and $SemSim_{th}$ in metastasis stage.	109
5.3	Comparison based on p-value of top 3 modules obtained in metastasis stage.	109
5.4	Non-Metastasis modules	111
5.5	Non-Metastasis modules	112
5.6	Metastasis modules	113
5.7	Expression value of genes involved in common pathways in both the stages	115
6.1	Gene expression value for dummy genes g_1 and g_2 for seven samples	125

6.2	Labels for dummy genes g_1 and g_2 for seven samples using our edge finding approach	126
6.3	p-values of modules obtained using different thresholds in both control and disease stages	133
6.4	Best p-value reported by modules obtained using different thresholds in both control and disease stage using the method given in [133] .	133
6.5	Best p-value reported using my proposed method and other existing works.	134
6.6	Best p-value reported by modules obtained using CC and BiMax biclustering technique in both control and disease stage	135
6.7	Number of common pathways for all the 22 control modules w.r.t. the 7 disease modules	137
6.8	Causal genes along with their interacting partners among each control module	139
6.9	Causal genes along with their interacting partners among each control module	140
6.10	Causal genes along with their interacting partners among each control module	141
6.11	Causal genes along with their interacting partners among each control module	142
6.12	Causal genes along with their interacting partners among each control module	143
6.13	Causal genes along with their interacting partners among each control module	144
6.14	Causal genes along with their interacting partners among each control module	145
6.15	Hub genes based on radially and pagerank measure for each module in control stage along with their associations with causal and suspected genes.	149
6.16	Hub genes based on radially and pagerank measure for each module in disease stage along with their associations with causal and suspected genes.	150
A.1	List of genes associated with Parkinson's Disease and the potential biomarkers along with their source of information	175
A.2	List of genes associated with Parkinson's Disease and the potential biomarkers along with their source of information	176

List of Algorithms

1	Steps involved in CNCM Algorithm	40
2	CSC Algorithm for complex formation	69
3	ComFiR Algorithm steps for complex formation	86
4	Network module extraction algorithm	106
5	Algorithm for module extraction from gene gene network	130

Symbols and Meanings

Symbol	Meaning
$G(V, E)$	PPI network with $V = \{v_1, v_2, \dots, v_n\}$ vertices and $E = \{e_1, e_2, \dots, e_m\}$ edges.
<i>cluster</i>	Set of clusters (complexes) without removing redundant complexes.
$d(v_i)$	degree of v_i .
$CCf(v_i)$	Clustering coefficient of v_i .
$Conn_t(v_i, G')$	Connectivity of v_i with G' .
$CCfT$	Clustering coefficient threshold.
$NgCp(C_i)$	Neighbor of complex, C_i .
$Ng(v_i, v_j)$	Neighbor of v_i and v_j .
α	Connectivity threshold.
<i>RemList</i>	Data structure to keep track of nodes during complex finding.
<i>pC</i>	Data structure to store elements while growing a cluster.
$HC(v_i, v_j)$	HConfidence value between v_i and v_j .
$sd = (v_i, v_j)$	Seed pair consisting of nodes v_i and v_j during unsupervised complex finding.
RbI_{v_i}	Reachability Index of node v_i .
$Supp(G')$	Contribution of subgraph G' .
β	Semantic similarity threshold.
h_t	HConfidence threshold .
<i>secCluster</i>	set of clusters before removal of redundant clusters.
<i>CndS</i>	Data structure to record candidate seed pair.
<i>HCS</i>	HConfidence store to keep record of the HConfidence values of each pair of nodes.
sd_p	Selected seed pair from the set of candidate seed pair, <i>CndS</i> .
<i>NodeExpcluster</i>	Data structure to keep track of nodes during cluster expansion.
<i>pC1</i>	Data structure similar to <i>pC</i> but only with an extra node, v_m to calculate contribution function.

Symbols and Meanings

Symbol	Meaning
$hcount$	Counter to maintain the record for every pair of nodes with HConfidence values.
$ccount$	Counter to maintain record of $Cluster$.
$acount$	Counter to maintain record for $secCluster$.
S_{sd_p}	Seed pair consisting of nodes v_i and v_j during semi-supervised complex finding process.
ms_{v_a, v_b}	Matching scores between nodes v_a and v_b .
$MMgS(v_i, v_j)$	Maximum Matching Score between a pair of nodes v_i and v_j .
$NSSm$	Semantic similarity score matrix.
$D_{gs} = \{g_{d_1}, g_{d_2}, \dots, g_{d_p}\}$	set of disease genes.
DGN	Disease gene association matrix.
SC_j	Relevance score of complex C_j .
BCM	Protein complex benchmark matrix.
$AMMgS$	Data structure to store maximum matching score of v_i and v_j .
$scount$	Counter to maintain track of seed node.
$ActualCluster$	Set of unique clusters.
$rs_{g_{d_i}}$	Relevance score of disease gene g_{d_i} .
$xcount$	Counter to count the number of disease genes.
$rcount$	Counter for ranking disease associated complexes.
$pvalue_{C_i}$	p-value of complex C_i .
$DN_{C_{v_i}}$	Direct neighbors of node v_i within C_i .
$contbn_{v_i}$	Node v_i contribution in terms of direct neighbors.
Cl_{contbn}	Cluster contribution (of all nodes within a cluster).
$Rbty_{v_i}$	Reachability of a node v_i in a cluster.
$RbyContbn$	Reachability contribution.

Symbol	Meaning
$A = \{G, E\}$	Adjacency matrix representation of network with V vertices and E edges.
V_{th}	Variance threshold.
PCC_{th}	Pearson correlation coefficient threshold.
Adj_{nm}	Adjacency network representation of non metastasis stage.
Adj_m	Adjacency network representation of metastasis stage.
M_i	i^{th} module.
$SemSim(g_a, g_b)$	Semantic similarity between genes g_a and g_b .
$SemSim_{th}$	Semantic similarity threshold.
$SemSimM$	Semantic similarity score matrix.
$Modules = \{M_1, M_2, \dots, M_N\}$	Set of N modules.
$RemList$	Data structure to keep track of nodes during module extraction.
$mCount$	Counter to keep track of number of modules .
GM	Gene Expression matrix with a number of genes (rows) at b conditions (columns).
D_m	Discretized gene expression matrix with a rows and $b \times (b - 1)/2$
η	Threshold based on the standard deviation of the expression matrix.
Γ	Threshold used to calculate whether two genes g_1 and g_2 are connected or not based on the absolute difference between their expression values.
PCC	Pearson correlation coefficient.

Symbols and Meanings

Symbol	Meaning
$T_{SbOM_{ab}}$	Topological Subspace Overlap Metric for a pair of nodes a and b .
αA	similarity between neighbors.
βB	direct edge connectivity.
$E_{g_a g_i}^p$	edge between genes g_a and g_i at p^{th} condition.
V_g	total number of genes in GM .
sm	total number of samples in GM .
Adj_{g_a, g_b}	adjacency matrix having value 1 or 0 depending on whether g_a and g_b are connected or not.
n_{th}	Threshold used for assigning values to Adj_{g_a, g_b}
$Nodedistance(g_v, g_k)$	Distance between nodes g_v and g_k .
$N_{CN}(g_v)$	Constrained neighbor set of g_v .
$CnS(g_v)$	Constrained node score.
$SI(g_v, g_k)$	Simpson Index for g_v and g_k .
$N(g_v)$	Neighbors of g_v .
γ	Threshold used during seed selection based on CnS .
m_i	Network module
$L = \{c_{g_1}, c_{g_2}, \dots, c_{g_k}\}$	Set of causal genes associated with the disease.
δ	Threshold used during selection of SI between genes.
$partialCluster$	Data structure to store elements while growing a cluster.
$T_{SbOM} = \{V, E\}$	TSOM network.
$Modules = \{C_1, C_2, \dots, C_N\}$	Set of modules.
M_{c_i}	Modules from the control stage.
M_{d_j}	Modules from the disease stage.
$MR(c_i, d_j)$	Matching ratio between control module, c_i and disease module d_j .