

Contents

1	Introduction	1
1.1	Protein protein interaction(PPI) networks	1
1.2	What is a protein complex ?	3
2	Literature Survey	4
2.1	Approaches	4
2.1.1	Graph-theoretic approach	4
2.1.2	Flow simulation-based	4
2.1.3	Spectral clustering-based	4
2.1.4	Supervised clustering	5
2.1.5	Core attachment-based	5
2.1.6	Swarm Intelligence-Based Approaches	5
2.2	Methods for Protein Complex detection	5
2.2.1	Clique	5
2.2.2	Markov Clustering Method	5
2.2.3	MCODE	6
2.2.4	COACH	6
2.2.5	DPC	6
2.2.6	ClusterONE	6
3	Evaluation Metrics	7
3.1	Precision, Recall and F-measure	7
3.2	Co-localization and Gene Ontology(GO) semantic similarity score:	7
4	Method 1: Identifying Protein complexes in protein protein interaction network based on core attachment approach incorporating Gene Expression Profile	9
4.1	Motivation	9
4.2	Method and algorithm	10
4.2.1	Terminologies	10
4.2.2	Complex Identification Procedure:	10
5	Method 2: Weighted edge based clustering to identify protein complexes by incorporating Gene Expression Profile	14
5.1	Motivation	14
5.2	Preliminaries	14
5.2.1	PPI Networks	14
5.2.2	GEP	14
5.2.3	Edge Clustering Coefficient	15
5.2.4	Weighted Edge:	15
5.3	Method	15
6	Experimental Output	17
6.1	Input data	17
6.2	Parameter sensitivity and observations from CAG	18
6.3	Parameter sensitivity and observation of WEC	18
6.4	Comparison with other methods	19
6.4.1	Precision,Recall and F-measure Score	19
6.4.2	Perfect Match	20
6.4.3	Co-localization score	21
6.4.4	Gene Ontology(GO) semantic similarity score	21

7	Disussion and Conclusion	23
7.1	Discussion	23
7.2	Conclusion	24
	Appendices	25
A	Methods used for comparison with our methods and their parameter settings	26
A.1	METHODS:	26
A.2	Parameter Settings:	26
B	Some of the protein complexes detected by our Methods CAG and WEC	27
C	Source of the binary executable programs and software used	28

List of Figures

1.1	Overview of the Collins PPI network with 9074 interactions and 1622 distinct proteins . .	2
5.1	algorithm for detecting protein complexes	16
6.1	Illustration of the change of f-measure value obtained by CAG with changing density threshold value T_d and similarity threshold value T_s on the Collins network data using SGD as a gold standard data.	18
6.2	The parameter sensitivity of method WEC with changing parameter values on the collins network. The x-axis represents the changing value of Threshold weight T_{weight} and the f-measure is represented by the y-axis. The lines $a = 0, a = 1... etc.$ represents the value so the balance factor T_a	18
6.3	The illustration of the comparison of our methods CAG and WEC with MCODE,CoAch, ClusterONE, SPICi and DPC in term of f-measure on the Gavin Network	19
6.4	The illustration of the comparison of our methods CAG and WEC with MCODE,CoAch, ClusterONE, SPICi and DPC in term of f-measure on the Collins Network	20
6.5	The number of perfect and good complexes predicted by MCODE, CoAch, ClusterONE, DPC, SPICi, CAG and WEC on Collins network.	21
6.6	The co-localization score of MCODE, CoAch, ClusterONE, SPICi, DPC, CAG and WEC on using the Gavin network	22
6.7	The GO semantic similarity score of MCODE, CoAch, ClusterONE, SPICi, DPC, CAG and WEC on using the Gavin network	22
B.2	The complexes detected from CYC2008 by both CAG and WEC with their Gene Ontology annotations.	27

List of Tables

6.1	Detail of the Collins and Gavin PPI network	17
6.2	Showing the values of f-measure at different values of similarity threshold T_s and density threshold (T_d)	18
6.3	The observations recorded for WEC on Collins and Gavin network	19
6.4	The f-measure values obtained by MCODE, CoAch, ClusterONE, SPICi,DPC and CAG using both Collins and Gavin network data.	20
7.1	Showing the proteins in the predicted and the real RSC complex	23
7.2	Showing the proteins in the real SWI/SNF complex	23