Contents

	Page No
CHAPTER 1	
Introduction	1
1.1 Microarray Data Analysis	1-2
1.2 Coherent Pattern Identification: A Clustering Problem	2-3
1.3 Existing Approaches	3-3
1.4 Motivation	3-3
1.5 Aim of our work	4-4
1.5.1 Objectives of our work	4-4
1.6 Organization of the Dissertation	4-4
CHAPTER 2	
Background of the work	5
2.1 Proximity measurement for gene expression data	5-6
2.2 Comparison of Proximity measures	6-6
2.3 Discussion	6-7
CHAPTER 3	
Existing work done on clustering	8
3.1Partitional Clustering	8-8
3.1.1 K-Means Algorithm	8-8
3.1.2 QT Clustering	8-9
3.2 Hierarchical Clustering	9-9
3.2.1 UPGMA (Unweighted Pair Group Method with Arithmetic mean)	9-9
3.3 Density based Clustering	10-10
3.3.1 DHC	10-10
3.4 Model-based methods	10-10
3.4.1SOM (Self-Organizing Map)	10-11
3.4.2 SOTA (Self Organizing Tree Algorithm)	11-11
3.5 Graph Based clustering	11-11
3.5.1 CLICK (Cluster Identification via Connectivity Kernels)	12-12
3.5.2 CAST (Cluster Affinity Search Technique)	12-13
3.6 Gene Ontology based clustering	13-13
3.7 Discussion	13-14
CHAPTER 4	
Introduction to Gene Ontology (GO)	15

4.1 GO SIMILARITYOF TWO GENES 16-17
4.2 Discussion 17-17

CHAPTER 5

Proposed Expression based Clustering Method 18

5.1 Pair Wise Changing Tendency Measure (PWCTM) 18-19
5.1.1 Basics of PWCTM 19-21
5.2. Our graph based clustering approach (DClique) 21-23
5.2.1 Our Proposed algorithm (DClique) 23-25
5.2.2. Analysis of the algorithm 25-26

CHAPTER 6

Experimental Results 27

6.1 Comparison of our proposed similarity measure, PWCTM 27-28
6.2 Comparison and validation of our proposed algorithm, DClique 28-34

CHAPTER 7

GO based clustering 35

7.1 Common Information Content (CIC) 35-35
7.2 Construction of the GO similarity matrix 35-37

CHAPTER 8

Results of GO clustering 38-47

CHAPTER 9

Conclusion and future work 48-48

List of Figures

Figure Number	Description	Page
Figure 1	Gene Expression matrix	2
Figure 2	The structure of the GO is illustrated on some of the paths of term GO:0060491 to its root term. Note that it is possible for a term to have multiple parents.	15
Figure 3	graph representation of similarity matrix	22
Figure 4	clusters obtained by applying K-Means on dataset1 using PWCTM as similarity measure	27
Figure 5	clusters obtained by applying K-Means on dataset 1 using PCC as similarity measure	27
Figure 6	clusters obtained by applying DClique on dataset 1	29
Figure 7	clusters obtained by applying DClique on dataset 2	29
Figure 8	clusters obtained by applying DClique on dataset 3	30
Figure 9	Construction of GOSimilarity Matrix	35
Figure 10	Figure shows how the GOSimilarity matrix is used in clustering process.	36
Figure 11	clusters obtained by applying GOClustering on yeast dataset GO weight 100	39
Figure 12	clusters obtained by applying GOClustering on Drosophila dataset GO weight 100	40
Figure 13	clusters obtained by applying GOClustering on yeast dataset GO weight 40	40
Figure 14	clusters obtained by applying GOClustering on dataset4 GO weight 40	41

List of Tables

Table Number	Description	Page
Table 1	Description of datasets used for our experiment	28
Table 2	Comparison of p-values for k-means clustering on dataset1 using PWCTM and PCC	28
Table 3	P-value comparison of CLICK, CC and DClique clusters on dataset 1	31
Table 4	P-value comparison of CLICK, CC and DClique clusters on dataset 2	31
Table 5	P-value comparison of CLICK, CC and DClique clusters on dataset 3	32
Table 6	Q-value of cluster1 obtained from applying DClique on dataset 1	33
Table 7	Q-value of cluster2 obtained from dataset 1 by applying DClique	34
Table 8	Percentage weight of expression network of the clusters obtained from dataset 1 by applying DClique.	34
Table 9	Description of the datasets used in our experiments	38
Table 10	p values of clusters (best four in terms of p value) for different datasets	41