Contents

1 Introduction **7**

1.1 Objectives 8

1.2 Motivation 8

1.3 Contribution 10

2 Background **11**

2.1 Microarray Technology 11

2.2 Gene Expression Data Analysis 12

2.3 Data Mining 13

2.4 Data Mining in Gene Expression Data Analysis 14

3 Related Works **16**

3.1 Proximity Measures 16

3.1.1 Euclidean Distance 18

3.1.2 Cosine Similarity 18

3.1.3 Pearson Correlation Coefficient 18

3.1.4 Spearman Rank Order Correlation 19

3.1.5 Jackknife Correlation 19

3.1.6 BioSim 20

3.1.7 Discussion 21

3.2 Clustering Gene Expression Data using Prior Knowledge . . . 22

3.2.1 GO Semantic Similarity Measures 23

3.2.2 Clustering approaches using GO annotations as prior knowledge 24

3.2.3 Discussion 26

4 **Proposed Measure-PNCSIM** **27**
4.1 Conceptual Framework 27
4.2 Properties 30
4.3 Validation 30
4.3.1 Consistency of PNCSIM under normal and non normal distributions 31
4.3.2 Biological Validation 33
4.4 Discussion 36

5 **Integration of PNCSIM and Prior Knowledge** **38**
5.1 Proposed Framework 38
5.1.1 Finding the Semantic Similarity and Homogeneity . . . 40
5.2 Validation 42
5.2.1 Finding the Sequence Similarity 44
5.3 Discussion 47

6 **Conclusion and Future Work** **51**

List of Tables

3.1 Comparison of the existing correlation measures 22

4.1 Symbols used 28

4.2 p-value of some of the clusters for yeast sporulation and human datasets using diana algorithm 35

4.3 p-value of some of the clusters for yeast sporulation and human datasets using clara algorithm 36

4.4 p-value of some of the clusters for yeast sporulation and human datasets using pam algorithm 37

5.1 Average Semantic Similarity and Homogeneity of clusters . . . 42

5.2 Result Summary of clusters for Sequence Similarity using Diana Algorithm 45

5.3 Result Summary of clusters for Sequence Similarity using Clara Algorithm 46

5.4 Result Summary of clusters for Sequence Similarity using Pam Algorithm 47

5.5 Average Sequence Similarity of clusters 48

5.6 Average phylogenetic distance of the clusters 50

List of Figures

1.1 Partial Positive and Negative Correlation 9

3.1 Basic Architecture of BioSim 20

4.1 Conceptual framework of PNCSIM 29
4.2 Comparisons showing (a) normal (b) log-normal (c) beta-distributions on different correlation measures 32

5.1 Proposed Complete Framework 39
5.2 Comparison between semantic similarity and homogeneity for the three algorithms 43
5.3 Phylogram of some clusters for the three algorithms 49