

Table of Contents

List of Tables	vi
List of Figures	viii
Glossary of Terms / Abbreviations	xi
List of Notations	xvi
Chapter 1	1
1 Introduction	2
1.1 Cluster Analysis	3
1.2 Gene Expression Clustering	4
1.3 Microarray Datasets – Gene, Cancer Data and Protein Interaction Data.....	5
1.4 Motivation	7
1.5 Contributions	9
1.6 Organization of the Thesis	11
Chapter 2	13
2 Gene Expression Data Analysis	14
2.1 Background of Molecular Biology.....	14
2.2 Microarray Data – Generation and Analysis.....	17
2.3 Gene Expression Data	20
2.4 Data Mining in Gene Expression Data Analysis.....	21
2.4.1 Supervised Approach	22
2.4.2 Unsupervised Approach	23
2.4.3 Discussion	25
2.5 Ensemble Approach	27
2.5.1 Supervised Ensembles.....	28
2.5.2 Unsupervised Ensembles.....	36
2.5.3 Semi-Supervised Learning	46

2.5.4 Discussion	52
2.6 K-Fold Cross Validation	53
2.7 Discussion	55
2.8 Validity Measures.....	55
2.8.1 External Validity Measures	56
2.8.2 Internal Validity Measures	59
2.9 Protein-Protein Interaction (PPI) Data	62
2.10 Analysis of PPI Data Using Data Mining Techniques	63
2.10.1 Properties of PPI networks	64
2.10.2 PPI Network and Protein Complexes.....	64
2.11 Discussion	67
Chapter 3.....	69
3 Classification of Microarray Cancer Data Using Ensemble Approach	70
3.1 Introduction	70
3.2 Background	71
3.3 Construction of Ensembles.....	72
3.3.1 Classifier Selection.....	72
3.3.2 Fusion	73
3.4 Existing Methods.....	73
3.4.1 Different Classifier Models	74
3.4.2 Different Feature Subsets	74
3.4.3 Different Training Sets.....	74
3.4.4 Different Combination Schemes	74
3.5 Ensemble Combination Methods	75
3.5.1 Weighting Methods	75
3.5.2 Combining Continuous Outputs.....	76
3.6 Related Work.....	78
3.7 Base Classifiers	80
3.7.1 J48	81
3.7.2 IBk.....	81
3.7.3 Naïve Bayes (NB)	82

3.8 Ensemble Methods	82
3.8.1 Bagging	82
3.8.2 Boosting	83
3.8.3 Stacked Generalization.....	83
3.9 Motivation	85
3.10 Experimental Design Methodology.....	86
3.11 Software Used For Comparison	87
3.12 The Proposed SD-EnClass	87
3.12.1 Distinct Training Sample Selection.....	88
3.12.2 Working of the SD-Enclass.....	89
3.12.3 Performance Analysis of the Proposed Model	92
3.13 Meta-Ensemble.....	94
3.14 Statistical Significance Between the Classifiers	97
3.15 Discussion	99
Chapter 4.....	101
4 Cluster Analysis of Cancer Data Using Similarity Measures	102
4.1 Introduction	103
4.2 Related Work.....	104
4.2.1 Partitioning Approach	104
4.2.2 Hierarchical Approach	105
4.2.3 Model-Based Approach.....	105
4.3 Validation	106
4.3.1 External Validation	106
4.3.2 Internal Validation.....	112
4.3.3 Stability Measures	114
4.3.4 Biological Measures	114
4.4 Motivation	115
4.5 Methodology	116
4.6 Dataset Description	118
4.7 Results	118
4.7.1 Results of External Validation	118

4.7.2 Results of Internal Validation	121
4.7.3 Results of Stability Measures	123
4.7.4 Results of Biological Validation	125
4.8 Discussion	127
Chapter 5.....	129
5 Complex Detection from PPI Data Using Unsupervised Methods	130
5.1 Introduction	131
5.2 Related Work.....	132
5.2.1 Taxonomy of Existing Clustering Methods	132
5.2.2 Review of Existing Work.....	133
5.3 Motivation for an Empirical Study.....	135
5.4 Terminology	135
5.5 Selection of Unsupervised Methods.....	136
5.6 Overview of the Approaches	137
5.6.1 MCL (Markov Clustering)	137
5.6.2 MCODE (Molecular Complex Detection)	138
5.6.3 RNSC (Restricted Neighbourhoods Search Clustering)	138
5.6.4 CFinder (Clique Finder)	139
5.6.5 RRW (Repeated Random Walk).....	139
5.6.6 AP (Affinity Propagation).....	140
5.6.7 CMC (Clustering based on Maximum Cliques).....	140
5.6.8 ClusterONE (Cluster Overlapping Neighbourhood Expansion).....	141
5.7 Evaluation of Protein Complexes.....	141
5.7.1 Evaluation Measures	142
5.7.2 Gold Standard for Protein Complexes	144
5.8 Comparison of Selected Algorithms	144
5.8.1 Parameter Settings for Each Algorithm	145
5.8.2 Quality Scores for MIPS Gold Standard	146
5.8.3 Quality Score for SGD Gold Standard.....	147
5.9 Biological Coherence of Predicted Complexes.....	149
5.9.1 Co-Localization Similarity	149
5.9.2 GO Semantic Similarity	150

5.10 Statistical Significance	151
5.10.1 Statistical Evaluation of Predicted Complexes	152
5.10.2 Statistical significance of ClusterONE and MCL	153
5.11 Motivation for an Ensemble Framework.....	154
5.12 Protein Complex Detection Ensemble (PCDEN).....	155
5.13 Experimental Evaluation	158
5.13.1 Results	158
5.14 Discussion	165
Chapter 6.....	167
6 Conclusions and Future Work	168
6.1 Conclusions	168
6.2 Future Work	170
Bibliography	173
List of Publications.....	204
Appendix	207