Chapter 6

6 Conclusions and Future Work

6.1 Conclusions

In this thesis, classification and clustering techniques using ensemble approach have been applied for gene expression data analysis and protein interaction networks.

An empirical study of various existing supervised classifiers and their ensembles is first performed to identify their pros and cons. The idea was to design a cost effective ensemble method that will not be influenced by the biasness of the base classifiers. Hence an ensemble method is developed for combining the base classifiers from different classification families into an ensemble, based on a simple estimation of each classifier's class performance. An improvement in the classification accuracy is observed as compared to the best single model in the combination. However, the performance of the ensemble found to be average when compared to Bagging and Boosting. So using the combination method proposed, the results of the proposed ensemble were combined with the results of the best performing ensemble methods into a Meta-Ensemble. The experimental results obtained over nine cancer datasets were significantly better than those obtained by using Boosting, Bagging or Stacking and it establishes the effectiveness of the proposed model.

Next, a method involving supervised ensemble methods is devised to improve upon the analysis of the results that were obtained from the previous work on cancer datasets. In this method, it is proposed to integrate existing biological knowledge, such as the GO database, so to assist in the cluster validation process of cancer datasets. The clustering results arrived at through the ensemble approach are then validated using external validity measures such as semantic and sequence similarity measures to ascertain whether the clusters obtained are biologically significant. Internal validity measures such as Dunn index, silhouette width and the homogeneity index are used to evaluate the visual separation of the clusters obtained from a clustering algorithm. Since the dataset is highly correlated, additional stability measures and biological validation measures are used in the form of biological homogeneity index and biological stability index to obtain biologically relevant clusters. The approach was tested on several benchmark cancer datasets and the experimental results have been found to be satisfactory, establishing the fact that generating high-quality clusters leads to a marked improvement in the biological relevance of the clusters.

The final contribution is an ensemble approach for protein complex identification to improve the robustness and stability of the protein complexes. The PPI network data is taken as input by *n* base clustering algorithms to generate *n* individual complexes. The process of identifying the node of the highest degree from the next set of complexes and the process of identifying protein complexes through consensus and through the use of GO similarity is repeated for each corresponding complex. The end result of this phase is that there will be interacting protein complexes with overlapping proteins. This is followed by the validation phase that evaluates the protein complexes predicted by comparing them to a set of gold standard protein complexes and also verifies the biological relevance of the predicted protein complexes.

The central idea of ensemble learning is to employ multiple learners and then combine their predictions for better accuracy. An ensemble has the capability of achieving consistently well performing results for any dataset by deriving individual benefits of participating algorithms in terms of noise handling, cluster overlap detection and possesses the ability to detect complexes with minimum prior

knowledge. The design of an ensemble involves two main issues – the diversity of the algorithms and the integration of the outputs of the base algorithms. Diversity is the degree to which classifiers disagree in the errors they make and can be achieved by employing various strategies such as employing different classifier models, different feature subsets and different training data sets. Diversity, therefore, allows the voted accuracy to be greater than the accuracy of any single classifier. A consensus of all the individual base classifiers and base clustering algorithms gives more weight to the decision arrived at by a majority of the algorithms and also takes care of the possible errors made during clustering by a single algorithm.

It can be concluded that the ensemble approach has distinct advantages over any of the existing clustering algorithms, whether for clustering gene data or for protein complex detection. Integration of biological information such as GO improves the biological relevance of the predicted clusters.

6.2 Future Work

The work presented in this thesis can be extended in diverse directions. Some ideas for future work are listed below.

At present there is a dearth of tools / software packages for ensemble creation for the purpose of clustering gene expression data. The development of an integrated software package or data mining tool having the capability of feature selection, ensemble creation using the three approaches (of supervised, unsupervised and semi-supervised), with a large collection of combination rules and validity measures, would contribute to growth in this field. The capability of this tool could be extended to combine multiple sources of biological information to extract non-trivial patterns of high biological significance.

A recent trend in the analysis of gene expression is by using novel bio-technologies very different from the mature DNA microarray technology. One such technology is RNA Sequencing (RNA-Seq) that has revolutionized the exploration of gene expression data. RNA sequencing workflow, from sample preparation to data

analysis, allows one to discover and profile the transcriptome in any organism, has some advantages over microarray technology such as:

- RNA-Seq technology does not require species-specific or transcript-specific probes. It enables unbiased detection of novel transcripts and other changes, such as single nucleotide variants, indels (small insertions and deletions), and gene fusions.

- Microarray technology cannot detect novel transcripts or other previously unidentified changes and is effective only for detecting expression of known genes and transcripts.

- With array hybridization technology, gene expression measurement is limited by background at the low end and signal saturation at the high end. RNA-Seq technology, on the other hand, quantifies discrete, digital sequencing read counts, offering a broader dynamic range.

- Compared to microarrays, RNA-Seq technology offers increased specificity and sensitivity, for enhanced detection of genes, transcripts, and differential expression. Sequencing coverage depth can easily be increased to detect rare transcripts, single transcripts per cell, or weakly expressed genes.

Use of these technologies with ensemble techniques may lead to greater accuracy in the analysis of gene expression data.

In protein-protein interaction networks, proteins are represented as nodes and interactions are represented as edges. This makes it difficult to define the distance between two proteins. Moreover, most methods search for complexes in the "dense" regions of the network causing them to miss complexes of low densities, that is, "sparse" in the network [348]. An alternate way would be to augment topological information with biological information so that small complexes of two or three proteins are not discarded.

Every cluster algorithm has its own nuances, being highly dependent on specific initial conditions, seed selection or parameter setting. An ensemble method is

desirable as it can identify larger, denser clusters with improved biological significance. It will be also desirable to perform validation during complex formation and not after, with an aim to check the "goodness" of the protein complex based on cohesiveness.

Membrane protein complexes are formed by the physical interactions among membrane proteins and they constitute approximately 30% of the proteomes of organisms. The study of membrane proteins and their complexes is crucial in understanding diseases and aiding new drug discoveries [349]. Conventional techniques are unable to mine membrane proteins because they are not as stable entities as their soluble counterparts [350]. Sophisticated algorithms need to be developed specifically to mine membrane complexes from membrane sub-interactomes.