Chapter 8

Concluding Remarks

Advancements in microarray technology have facilitated the generation of huge amount of gene expression data. Solutions to the biclustering problem is best suited for the analysis of this biological data for the following reasons: 1) genes are responsive to a small subset of conditions, 2) a gene may be responsible for more than one biological activity and 3) a condition or a group of conditions may trigger the expression of genes responsible for more than one activity.

The main contribution of this thesis is to develop a novel approach to biclustering gene expression data using mutual information. Unlike other similarity measures like distance measures and correlation coefficient, mutual information is a more general measure as it is able to extract both linear and nonlinear relationships. Though mutual information has been used earlier as a measure of similarity by traditional clustering algorithms, none of the biclustering algorithms have used it.

In our work we have proposed a set of biclustering algorithms which use mutual information as a measure of similarity. Through extensive experimentation on synthetic datasets and available real datasets we have demonstrated the utility of our work. The biclusters extracted by our algorithms are statistically more significant than the biclusters extracted by other algorithms. Algorithms are simple yet very effective (as exhibited by
the experimental results) thereby demonstrating the strength of mutual information as a similarity measure. Time analysis of our algorithms and Tables 5.1, 6.1 and 7.1 show that the proposed algorithms scale well with the size of the data.

In MRB and GenBiClus, we have used kernel density estimator to estimate mutual information as the bin method cannot be used to compute the contribution of each condition to the mutual information. In BRC, both the bin method and the kernel density estimator can be used to estimate the mutual information. However, we have presented the results with kernel density estimator as it provides better estimates as discussed in Chapter 4.

**Future Work:** It will be interesting to see if machine learning tools like Discriminant Analysis, Mixture Models and Expectation Maximization which have been used successfully for the classification problem can provide better solutions to the problem. The major challenge in doing so would be to handle different sets of conditions for different biclusters and the overlapping nature of biclusters. Another direction of work would be to improve the results by generating ensembles for biclusters. The only work closely related to this is ensembles by Gullo et al. [F. 09] and Wang et al. [WLDJ11]. However, they create ensemble of biclusters which do not overlap either on genes or on conditions.