DNA microarray experiments raise numerous statistical questions in different fields as diverse as image analysis, experimental design, hypothesis testing, cluster analysis and distribution theory etc. Noise creeps into microarray experiments at each stage from the preparation of tissue samples to the extraction of data. In order to measure gene expression changes accurately, it is important to take into account the random and systematic variations that occur in every microarray experiment. The greatest challenge to array technology lies in the analysis of gene expression data to identify which genes are differentially expressed across tissue samples or experimental conditions. The ability to measure gene expression enmasee has resulted in data with number of variables $p$ far exceeding the number of samples $N$. Standard statistical methodologies do not work well or even at all when $N < p$. Modifications of existing methodologies or development of new methodologies is needed for the analysis of microarray data.

Usually in microarray data most genes are expressed at very low levels and only few genes are expressed at high intensity. The main objectives of the research work undertaken were to develop tools specific to microarray data analysis in identification of differentially expressed genes and to have a comparative study with the existing methods, to study the use of statistical classification and dimension reduction techniques in identifying corregulated genes/samples and to study the distribution of gene expression intensities across genes.

The thesis is organized in six Chapters. Chapter 1 discusses the biological background, design of microarray chip technology and statistical issues in analysis of microarray data. A summary of data pre-processing methods, background correction and variance stabilization methods are reviewed in Chapter 2. This Chapter also reviews the existing approaches in microarray data analysis and the frequently used visualization tools of gene expression data. Chapter 3 proposes a new method to identify differentially expressed genes by generalized p value technique. The concept of generalized p–value(GP) method has been applied for the selection of differentially expressed gene in two conditions by considering a lognormal model for the gene expression data for each gene separately. We also considered the generalized p–value test with shrinkage(GPS) by combining information across genes. The results of GP and GPS methods were compared with the results of t–test approach by assuming unequal sample variances, bootstrap t–test and moderated t–test. We have applied the methods to two publicly available datasets from spike-in experi-
ment and Apo AI experiment. The numerical results show that the procedure based on generalized p–value and generalized p–value method with shrinkage is as good as or superior to the best of the alternatives.

The application of Bayesian variable selection in identification of genes with differential expression and its prediction performance is discussed in Chapter 4. In the present study we applied the gene selection step by Bayesian variable selection methodology by considering latent variables followed by the Fuzzy C-means clustering method for partitioning the samples into two groups. Bayesian methods were used in gene selection where the criterion for identifying the differentially expressed gene is based on the posterior probability of differential expression. Fuzzy C-means clustering method have been applied for the first time to classify gene expression data. We have also conducted a simulation study to validate the procedure of variable selection and to investigate the power of detecting genes that are differentially expressed. Chapter 5 is concerned with the study on the use of statistical classification and dimension reduction techniques in identifying corregulated genes/samples. In this study the mathematical derivation of SVD analysis are described and the application of SVD as a preprocessing technique for clustering gene expression data are studied. In order to study the importance of singular value decomposition for identification of pattern of gene expression we complemented our study with synthetic dataset. The study on distribution of log ratio of gene expression across genes is discussed in Chapter 6. The asymmetric Laplace distribution have been applied to the publicly available gene expression dataset and got a reasonable fit to the gene expression data and greatly improved upon the Normal and lognormal distribution. The estimation procedure of parameters of asymmetric Laplace and the interpretation for the suitability of the distribution to gene expression data are discussed. We also developed an autoregressive process of order one with asymmetric-Laplace marginal distribution (ALAR(1)) for temporal gene expression data. The sample path and histogram confirms the suitability of the process for modelling gene expression data which are often heavy tailed.

The description of datasets used in the study and the algorithms developed for computation are given in Appendix.

**Keywords:** Microarray, gene expression, Generalized p value, Multiple hypothesis testing, False Discovery Rate, Bayesian variable selection, Principal component analysis, Partial Least Squares, Distribution theory, asymmetric Laplace, autoregressive process, bootstrapping, simulation