Abstract

Microarray is a recently developed functional genomic technology that has powerful applications in a wide array of biological research, including medical sciences, agriculture, biotechnology and environmental studies. cDNA (complementary Deoxyribo Nucleic Acid) microarray experiments make use of the fluorescently labeled cDNAs with different dyes (red and green) generated from test and control and allowed to hybridize on to target DNAs or genes spotted on the slide. From the intensity of fluorescence on the spot, level of expression of genes is inferred. The level of expression of genes from one set is compared to the other to identify differentially expressed genes. The main objectives of the present study includes developing new methods for identify differentially expressed genes in microarray data and to compare it with the available methods. For this purpose we introduced new families of asymmetric, peaked and heavy-tailed distributions and proposed these distributions as a parametric approximation of the distribution of the log-ratios of measured gene expression across genes in microarray. Other objectives are to develop a multiple hypothesis testing procedure using the generalized p-value approach and also to develop an empirical Bayesian analysis based on two component Laplace mixed model for identifying differentially expressed genes. To illustrate the applications of the model we used colon cancer dataset. We developed algorithm in $\mathbf{R}$ and $\mathbf{Maple}$ programs for computation.

Chapter 1 is an introductory chapter which gives some preliminary ideas on the biological background, principle and procedure of microarray technology, data pre-processing methods and statistical methods used in the analysis of microarray data. Further, a brief review of literature along with the summary of the work done is also given. Chapter 2 proposes new families of skew slash distributions namely, skew slash distribution generated by Cauchy kernel, skew slash distribution generated by normal kernel, skew slash t and asymmetric slash Laplace distribution. We developed these parametric models as an approximation of the distribution of the log-ratios of measured gene expression across genes. We have derived the

probability density function (*pdf*), cumulative density functions (*cdf*), characteristic function (*chf*) and mean. Further, the properties of the resulting distributions were studied. The nature of the density functions were examined and graphical studies are performed in order to demonstrate how the skew slash distributions are comparable with the standard distributions. Tail behavior of the distributions are also explored.

Applications of the newly developed skew slash family of distribution functions in the microarray gene differential expression are illustrated in Chapter 3. We developed an algorithm in $\mathbf{R}$ Statistical Package for the estimation and illustration of the proposed distributions. We compared the application of the newly developed skew slash distributions with the current models available for microarray. We found that the estimated density of the skew slash family of distributions is a better model for microarray gene expression data compared to the existing models, since it captures the skewness, peakedness and heavy tails found in microarray data. Finally we compared the newly developed slash families, the skew slash Cauchy kernel, skew slash generated by normal kernel, skew slash t and asymmetric slash Laplace (ASL) distribution to identify most suitable model among them. We computed the AIC for each model and found that ASL has the least value. This indicates that ASL performs well compared to other skew slash distributions. Computational complexity of the newly developed slash distributions prompted us to introduce another family of distributions, the asymmetric type II compound Laplace (ACL) distribution.

Chapter 4 proposes another new family of distributions which is referred to as ACL distribution for modeling microarray data. The chapter covers symmetric and asymmetric ACL distributions. The nature of the density function is examined and graphical studies are also performed in order to demonstrate how the ACL distribution is comparable with the standard type. Tail behavior of the distribution is also explored. We also developed the stress-strength reliability function for ACL and is used to compare the red and green intensity measurements.

In chapter 5 we illustrate the applications of ACL. Simulation studies for various choices of parameter values are performed to validate the algorithm developed in R package. Finally, we fit the ACL, AL, and log-normal distributions to five microarray gene expression datasets and compare them. We found that the distribution of the log-ratio of the expression values is well approximated by the ACL distribution. We developed an error model based on ACL for microarray data. The applications of stress-strength analysis in microarray gene expression studies are also illustrated. Further, we estimated the stress-strength reliability of ACL and applied it as a measure to compare the test and control populations in microarray gene expression studies. The estimate of stress-strength reliability gives, parametric insight into normalization across arrays in microarray and the proportion of genes which are differentially expressed.

In Chapter 6, we developed multiple hypothesis testing procedures based on the generalized p-value approach and Empirical Bayes methods using the mixed Laplace model to identify differentially expressed genes. The concept of GP method has been applied for the selection of differentially expressed gene in two conditions by considering an inverse Gaussian model for the gene expression data for each gene separately. The results of GP method was compared with the results of t-test by assuming unequal sample variances. Also we discussed the applications of Empirical Bayesian variable selection in identification of genes with differential expression and its prediction performance. We used a prediction rule approach based on a two component Laplace mixture model. We applied our method to publicly available cDNA microarray data set.

Keywords: Asymmetric Laplace, Asymmetric slash Laplace, Asymmetric type II compound Laplace, Bayesian model, cDNA Microarray, Empirical Bayes, Error distribution, Generalized p-value, Microarray gene expression, Mixed-Laplace, Multiple Hypothesis testing, Slash distribution, Skew slash, Skew slash Cauchy-Cauchy, Skew slash Cauchy-Laplace, Skew slash Cauchy-normal, Skew slash normal-Cauchy, Skew slash normal-Laplace, Skew slash normal-normal, Skew slash t, Stress-strength reliability.