2010
2011

FACULTY OF SCIENCES

*Master of Statistics: Bioinformatics*

# Masterproef

## *Joint modeling of phenotypic variables and gene expression data for compound screening*

Promotor :
Prof. dr. Ziv SHKEDY
Prof. dr. Dan LIN


Md. Fazlul Karim Patwary


*Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics*

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

# FACULTY OF SCIENCES

*Master of Statistics: Bioinformatics*

# Masterproef

*Joint modeling of phenotypic variables and gene expression data for compound screening*

Promotor :
Prof. dr. Ziv SHKEDY
Prof. dr. Dan LIN


## Md. Fazlul Karim Patwary


*Master Thesis nominated to obtain the degree of Master of Statistics , specialization Bioinformatics*

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

# Certification

I declare that this thesis was written by me under the guidance and counsel of my supervisors.

…….. …… Date……

**Md. Fazlul Karim Patwary** Student

We certify that this is the true thesis report written by **Md. Fazlul Karim Patwary** under our supervision and we thus permit its presentation for assessment.

…….. …… Date……

**Prof. dr. Ziv Shkedy** Internal Supervisor

…….. …… Date……

**Prof. dr. Dan Lin** Internal Supervisor

*Dedicated to*

*My*

*Beloved wife and kids*

***Acknowledgements***

What is Bioinformatics? When I have applied for an admission into this university and also when started was a little bit blur to me. But I had a great interest to know a new area which is related with informatics. Now, things are pretty clear to me and this credit goes to Prof. dr. Ziv Shkedy, Prof Tomasz Burzykowski, Prof. Dan Lin and others. Their lectures, especially prof. Ziv, gave me in-dept idea on what is bioinformatics. It is a great opportunity to convey my deepest indebtedness to them. I express my deep feelings and gratitude to my supervisors prof. dr. Ziv Shkedy and prof. dr. Dan Lin not only for supervising my work but also for many more helpful suggestions they extended me without which it would have not been possible for me to complete my research.

I am also grateful to all of my respected teachers and staffs of the Censtat, UHasselt and Institutue of Information Technology, Jahangirnagar University, Bangladesh where I am in job. I desire to express my especial thanks to Atiq, Joy (the little sister), Veronica, Angelica Moira and many more for their co-operation during the entire two years of my study in Belgium.

On a more personal level, I would like to express my heartiest honour to my mother and beloved wife and kids whose affections and encouragement enabled me to finish this research work specially my wife who suffered very much due to absence. I am also thankful to my family members, specially my father-in-law, mother-in-low and younger brothers (Shohel and Candan) for their patience and understanding during the period of my study in UHasselt, Belgium.

Finally, my gratitude to Prof. Abdul Bayes and Prof. dr. Sharif Enamul Kabir who helped me for necessary fund for this study period. I am solely responsible for errors and omissions in this dissertation, if any.

*Md. Fazlul Karim Patwary*

**Abstract**

*In the invention of new effective drug, experiment has to be conducted quickly and efficiently. Effectiveness of a compound can be evaluated by measuring the molecular changes and gene expression is a good measure to observe the molecular change. So, gene expression can help us in selecting the relevant genes as biomarker. Using these biomarker, one can screen out that can help us in screening out the less effective compounds quickly in the drug development process. In this research, focus was concentrated on the identification of genomic biomarkers, based on gene expression for $IC_{50}$ values. Joint model and information theoretic approach were used to identify and evaluate gene-specific biomarkers. In addition to the evaluation of gene-specific biomarker, supervised principal components analysis was used to construct joint biomarker using the information from a potential set of genes. Within the framework of supervised principal components analysis, three different approaches for selecting the set of genes were used. These are selecting genes on the basis of adjusted association, inclusion of a gene in the set if it increases the coefficient of association and on the basis of factor loadings. The last two approaches gave good results and almost same gain.*

*Keywords: Biomarker, $IC_{50}$, Joint model, Supervised Principal Component Analysis*

# Contents

## 1. Introduction

In the invention of new drug, experiment is done by treating the subjects with group of compounds where each group contains large number of compounds. These compounds are almost identical except some differences in their chemical structure. So, selection of the best compound by screening out less effective or non-effective compound needs huge amount time in a clinical trial. Besides this, there is increasing public pressure for new promising drugs for marketing as rapidly as possible. Now-a-days, an increasing number of new drugs have well-defined mechanism of action at the molecular levels hence it is feasible to measure the effect of these drugs on the relevant biomarker quickly rather than some long-term clinical endpoint. (Molenberghs *et al.,* 2008). Expression of a gene is one kind of molecular measures which provides information on the process of transforming DNA into a functional gene product like protein or RNA. This gene expression can help us in selecting the relevant genes as biomarker that can help us in screening out the less effective compounds quickly in the drug development process.

In the drug development process, microarray experiment is one which consists measurement of gene expression along with the phenotypic response for different group of compound as treatment. The main purpose of this experiment is to find genes those are differentially expressed genes in response to the treatment of interest. When the phenotypic response to the treatment is known and researchers want to identify the underlying mechanisms, the genetic pathways that are affected by it. Some tools in microarray experiments can also be used to observe the activity of thousands of genes simultaneously that can be used to predict an outcome of interest (Sanden, 2008; Amaratunga and Cabrera, 2004). Such genes are labelled as genomic biomarkers. "Biomarkers play an increasingly important role in improving the effectiveness of drug research and development in pharmaceutical industries. In both pre-clinical and clinical trials, biomarkers have the potential to encourage innovation, improve efficiency, save costs and time, and gain research organizations a valuable advantage over their competitors" (Lin et al., 2010).

In the early days, we needed tools to build classifier using gene biomarkers which can classify sample in some specified groups or in clusters. But then purposes of biomarkers extended to not only to prediction of the sample but also to look at insight into the biological processes associated with the response of interest. Moreover, a treatment sometimes affects

gene expression as well as responses. Hence, it is crucial to have a tool to construct candidate biomarker considering these issues.

In this particular drug-development experiment, after the treating drug to the patients, $IC_{50}$ values (half maximal inhibitory concentration), a measure of effectiveness of a drug, was measured as a response. For each $IC_{50}$ values that associated to a treatment/group, there is a gene expression data. Treatment affects both the response and the gene expression data. Several authors have discussed the issue of using such gene expression data to predict a specific response in different ways. Buyse et al. (2000) proposed joint modelling considering the above issues and in this research that proposed gene-specific joint model were applied to identify the gene biomarkers for the response of interest. This gene-specific model helps us identify and evaluate the quality of each gene in the array as a biomarker for the response.

Having a lot of gene biomarkers, the main of interest is to evaluate a joint biomarker for which information from all the genes or most informative genes is need to use simultaneously. This joint biomarker is necessary because for further analysis, such as use regression, it is hard to adopt a lot of genes/features simultaneously. Several authors developed different approaches to solve this summarisation problem. An efficient method proposed by Bair et al. (2006) is Supervised Principal Component Analysis (SPCA). This method involves estimation of some new orthogonal predictors using linear combination of the vectors of gene expression and hence reduction in number of predictors. In this research, construction of joint biomarker is based on supervised principal components analysis (SPCA) were used.

### 1.1 Objective

The main objective of this thesis is to identify the set of genes that can be used to screening out compounds with the help of joint modelling technique.

### 1.2 Organization of the study

Following the Section 1, Methods and materials discuss in section 2, results of the study present in Section 3, finally discussion and conclusion are included in Section 4.

## 2. Methods and materials

### 2.1 Data Description

A dose-response study, from which data were found, was performed in a behavioural experiment. A data file consisting $IC_{50}$ values for of several clinical families of drugs called group of compound or cluster. In each family, there are several types of drugs among them they are almost common. There are something changed in the chemical structure to the existing drug that made them different drugs. The idea behind this is to identify the efficient drug after treating the patient by it. For each member of a family, i.e. drug, there is a gene expression microarray data regarded as sample or true end point in the experiment.

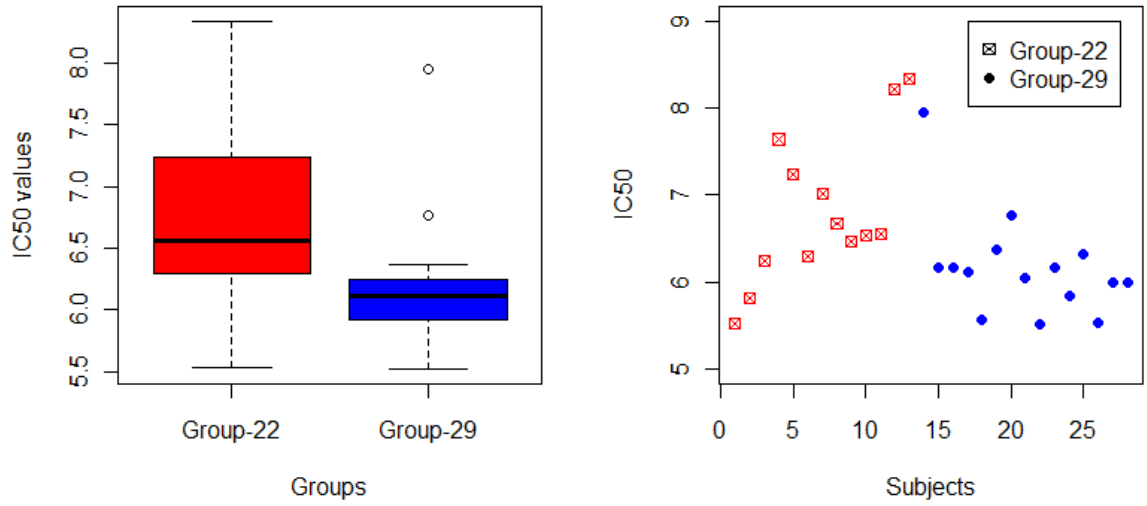


*Figure 1: Plot of IC50 values for two compound groups*

In this setting, there are two families which are numbered cluster 22 and 29 with their phenotype data i.e. surrogate end-point ($IC_{50}$) as well as gene expression data for each sample point. For the ease of the description later, these will be called G-22 and G-29. G-22 has 13 and G-29 has 15 samples respectively. For each sample there total number of features is 7722 and one $IC_{50}$ value.

Figure 1 shows the $IC_{50}$ values of drugs for the compound families. In the left panel, box plot of $IC_{50}$ values for the groups G-22 and G-29, their mean values are 6.81 and 6.17 respectively and value of t test statistic for the hypothesis of mean difference is 2.283 with p-value 0.03294. So, there are significant difference differences among the $IC_{50}$ values of two groups. Scatter plot in the right shows that $IC_{50}$ values of both the groups also supports this argument too because most of the points of G-22 are at top than that of G-29.

### 2.2 Gene-specific biomarker

There are three types of biomarkers for early drug development studies such as *therapeutic* and *prognostic* biomarkers, and biomarkers that are both *therapeutic and prognostic*. Therapeutic biomarkers are genes that are differentially expressed with respect to the treatment and thus can be used to predict the effect of the treatment on the response of interest. Prognostic biomarkers are genes, expression levels of which are correlated with the response, after adjustment for treatment. These genes have ability to explain some of its underlying genetic causes. (Sanden, 2008). A joint model for the gene expressions and the response which allow us to identify three types of genes i.e. to construct biomarkers described above.

Let $X_{ij}$ be the gene expression of the jth gene, j = 1, . . . ,m of the i-th subject, i = 1, . . . , n, and $Y_i$ be the $IC_{50}$ value of the drug applied in the subject. Let $Z_i$ be the vector indicating the treatment/compound family of i-th drug as follows:

$$Z_i = \begin{cases} 1 & if\ drug\ is\ in\ the\ 22\ compound\ family \\ 0 & otherwise \end{cases}$$

Then gene expression can be expressed with the following linear model in equation 1 assuming the relationship between gene expression and treatment and also response/$IC_{50}$ can be expressed as equation 2

$$X_{ij} = \mu_j + \beta_j Z_i + \varepsilon_{ij} \quad j = 1,2,\dots m \tag{1}$$

$$Y_i = \mu_Y + \alpha Z_i + \varepsilon_i \tag{2}$$

$\beta_j$ is a gene-specific parameter vector for gene j and $\alpha$ is the parameter denoting the treatment effect upon the response.

Following Buyse et al. (2000), Abel et al. (2010) and Lin et al. (2010) defined the gene-specific joint model in which the linear predictors of the $IC_{50}$ values and the gene expression are given by

$$E(X_{ij}|Z_i) = \mu_j + \beta_j Z_i, \qquad j = 1,..m; \quad i = 1,..n, \tag{3}$$
$$E(Y_i|Z_i) = \mu_Y + \alpha Z_i$$

The above joint model (3) is a gene-specific model and usual practice is to fit this model separately for each gene. This procedure is often termed "gene-by-gene" analysis. It is further assumed that the two outcomes are normally distributed

$$\begin{pmatrix} Y_i \\ X_{ij} \end{pmatrix} \sim N\left( \begin{pmatrix} E(Y_i|Z_i) \\ E(X_{ij}|Z_i) \end{pmatrix}, \sum_j = \begin{pmatrix} \sigma_{YY} & \sigma_{jY} \\ \sigma_{jY} & \sigma_{jj} \end{pmatrix} \right) \quad (4)$$

In the context of surrogate-marker evaluation in randomized clinical trials, Buyse and Molenberghs (1998) proposed the adjusted association as a measure of association, a coefficient derived from the covariance matrix of gene-specific joint model (4):

$$\rho_j = \frac{\sigma_{jY}}{\sqrt{\sigma_{jj}\ \sigma_{YY}}} \quad (5)$$

Here, $\rho_j = 1$ indicates a deterministic relationship between the gene expression and the response. i.e. a perfect prediction of $IC_{50}$ value is possible using this gene expression after adjusting treatment effect. Indeed, $\rho_j$ can be equal to 1 even if the gene is not differentially expressed among the treatment. So it is not necessary that genes have to be differentially expressed to be a good predictor for the response. (Dan Lin et al, 2010).

Estimation of the correlation between gene expressions and outcomes described in (5) is computationally complex. For the same purpose, Alonso and Molenberghs (2007) suggested a method called information-theoretic approach (ITA) that is simultaneously conceptually elegant and computationally simple. By adapting that method our two models can be describes as

$$E(Y_i) = Z_i\beta \quad (6)$$

$$E(Y_i|X_{ij}) = \mu_j + Z_i\beta + \gamma_j X_{ij}, \qquad j = 1,..m; \quad i = 1,..n \quad (7)$$

where in (7), the additional variable compared to (6) is specific gene with the coefficient $\gamma_j$, the effect of the j-th gene on the outcome which may indicate the potential biomarker. Upon fitting (6)–(7), the degree of association can be measured by

$$R_{hj}^2 = 1 - exp\left(\frac{-G^2}{n}\right) \quad (8)$$

Where $G^2$ denotes the likelihood ratio statistics to compare models (6) and (7) and n is the sample size. $R_{hj}^2$ is also called measures of uncertainty and it gives identical results as

squared adjusted association $\rho^2$ in (5) . If there is a strong interaction between Z and X then according to Buyse et al.(2000), an extra interaction term require to be introduced as follows:

$$E(Y_i|X_{ij}) = \mu_j + Z_i\beta + \gamma_j X_{ij} + \delta_j X_{ij} Z_i, \qquad j = 1,..m; \quad i = 1,..n, \tag{9}$$

$$E(Y_i) = Z_i\beta \tag{10}$$

But if the interaction term is not statistically significant then Buyse et al. (2000) suggest to use model which described in (7) and (8).

### 2.2.1Testing for Biomarkers

At this stage we have to test the following hypothesis

$$H_{10}: \beta_j = 0 \qquad H_{20}: \rho_j = 0 \qquad \text{and} \qquad H_{30}: \alpha = 0 \tag{11}$$
$$H_{1A}: \beta_j \neq 0 \qquad H_{2A}: \rho_j \neq 0 \qquad\qquad H_{3A}: \alpha \neq 0$$

Sanden (2008) described that the following situation may hold for a gene

a. If $H_{10}$ is rejected but $H_{20}$ is accepted and at the same time $H_{30}$ then this gene is potential therapeutic biomarker.
b. If $H_{10}$ is accepted but $H_{20}$ is rejected then this gene is potential prognostic biomarker. If $\rho_j > 0$ then it is called up-regulated prognostic biomarker and if $\rho_j < 0$ then it is called down-regulated prognostic biomarker.
c. If both $H_{10}$ and $H_{20}$ are rejected then this gene is potential therapeutic as well as potential prognostic biomarker.

### 2.3 Joint Biomarkers in Microarray Experiments

In the previous section gene-specific models were discussed which helps us identify and evaluate the quality of each gene in the array as a biomarker for the response. To create a joint biomarker, information from all the genes or most informative genes need to use simultaneously. Since numbers of genes are greater than the number of response, making a regression approach to summarize information into one linear predictor is no longer feasible. Several authors developed different approaches to solve this summarisation problem. Hastie and Tribshirani (2003) use Ridge regression with a regularization parameter in the genomic settings which is capable of accommodating large number of correlated predictors. Abdi (2003) employed Partial Least Squares regression to predict a set of dependent variables from

a (very) large set of independent variables that combines features from PCA and multiple regression. Hastie et al. (2000) proposed a "gene shaving" method which identifies subsets of genes with coherent expression patterns and large variation across conditions. In order to combine information about gene expression level from all genes another method proposed by Bair et al. (2006) which is Supervised Principal Component Analysis (SPCA). It involves estimating some new orthogonal predictors using linear combination of the vectors of gene expression X matrix. One or more predictors here can be used to regress the response. Our analysis to for joint biomarker is based on supervised principal components analysis (SPCA) only.

### 2.3.1 Supervised Principal Component Analysis

Let $X_{(N\times p)}$ is a gene expressions matrix consists of p genes (features) measured on N samples (patients) and $y_N$ vector of measured outcome. Assume that measured outcome is quantitative variable and columns of X matrix centered with mean 0. Then the singular value decomposition of X can be written as

$$X = WDV^T$$

where

$W$ is an $N\times N$ unitary matrix which is a eigenvector of $XX^T$ ,

D is an $N\times p$ rectangular diagonal matrix containing singular values $d_j$ on the diagonal,

$V^T$, the conjugate transpose of $V$, is an $p\times p$ unitary matrix which is a eigenvector of $X^TX$ and

Then is $U = W^TX$ where columns of $U: u_1, u_2, \cdots, u_p$ are the principal components whose are the vector of size N. In the SPCA methods, only the first principal component is used that consists the following three steps:

Step 1: Fit one of the gene-specific models and estimate the association measure.

Step 2: Form a reduced expression matrix consisting of only those genes whose gene specific association measure exceeds a threshold level.

Step 3: Let $X_R$ be the reduced matrix. Compute for each matrix the first principal component in a regression model to predict the outcome

First principal component is used because if variation of some gene expressions is strongly related to the outcome then first principal component will be highly correlated with the outcome and first principal component will be effective to predict the outcome. On the other hand, if variation of those gene expressions is related with some other biological process which is not related with the outcome then outcome might be highly correlated with second or some higher-order principal component and there may be some other genes that are related with the outcome (Bair et al. 2006).

If we use reduced number of genes (say k) using some threshold instead of all p genes then $X_{(N\times p)}$ matrix will be reduced to $X_R = X_{(N\times k)}$ and we will have $U: u_{R.1}, u_{R,2}, \cdots, u_{R.k}$ principal components where first principal component $u_{R.1} = w_{R.1}^T . X_R$ is like a linear combination of the columns of $X_R$ .

If we regress first principal component $u_{R.1}$ to regress y then

$$\hat{y}^{spc.R} = \bar{y} + \hat{\gamma}. u_{R.1}$$

$$= \bar{y} + \hat{\gamma}. w_{R.1}^T . X_R$$

$$= \bar{y} + \hat{\beta}_R . X_R$$

Where $\hat{\beta}_R = \hat{\gamma}. w_{R.1}^T$

Hence linear regression model estimate can be viewed as a restricted linear model estimate using all the predictors in $X_R$

Once first principal component is found for a gene expression data set X, let $U(X_R)$ then it can be considered the best joint biomarker and evaluation of this joint biomarker can be done using the same joint model or information-theoretic approach applicable for single gene as follows:

$$\begin{pmatrix} Y_i \\ U(X_R) \end{pmatrix} \sim N\left( \begin{pmatrix} E(Y_i|Z_i) \\ E(U(X_R)|Z_i) \end{pmatrix}, \sum_j = \begin{pmatrix} \sigma_{YY} & \sigma_{UY} \\ \sigma_{UY} & \sigma_{UU} \end{pmatrix} \right) \quad (12)$$

From (12), the conditional distribution of $Y_i$ is given by

$$Y_i, U(X_R) \sim N\left(\mu_Y - \sigma_{UY}\sigma_{UU}^{-1}\mu_U + \sigma_{UY}\sigma_{UU}^{-1}U(X_R)_i, \sigma_{YY} - \sigma_{UY}^2\sigma_{UU}^{-1}\right) \quad (13)$$

and the model can be written as

$$Y_i = \delta_0 + \delta_1 U(X_R) + \widetilde{\varepsilon_i} \quad (14)$$

where

$$\delta_0 = \mu_Y - \sigma_{UY}\sigma_{UU}^{-1}\mu_U$$
$$\delta_0 = \sigma_{UY}\sigma_{UU}^{-1}$$
$$\widetilde{\varepsilon_i} \sim N(0, \sigma_{YY} - \sigma_{UY}^2\sigma_{UU}^{-1})$$

Hence conditional distribution (7) with the following model is similar to the underlying model described in Bair et al (2006). The above conditional model can be used to calculate adjusted association by using the model either (5) or (8).

### 2.3.2 Leave-one-out cross validation (LOOCV)

Leave-one-out cross-validation can be use to obtain a reliable estimates of the measured association. The procedure of LOOCV is as follows: let there are n cell lines

a. Remove i-th cell line from the very beginning
b. Estimate first principal component based on remaining n-1 cell lines
c. Use this first principal component as a predictor in the linear regression according to (14)
d. Compute the value of the PC for the removed cell line and use the regression line to predict the $IC_{50}$

## 3. Results

### 3.1 Gene-specific biomarker

Applying the methods described above on the given data: $IC_{50}$ values of treatment group and gene expressions, results are presented with some logical discussions. Gene specific models were used to identify a set of genes as a possible biomarker for the $IC_{50}$ values.. The aim of the analysis is to identify specific genes, which can be used as genomic biomarkers i.e. therapeutic biomarker, prognostic biomarker and both therapeutic and prognostic biomarker. In other words, it can be used to predict the $IC_{50}$ values and/or whose expression is related to treatment.

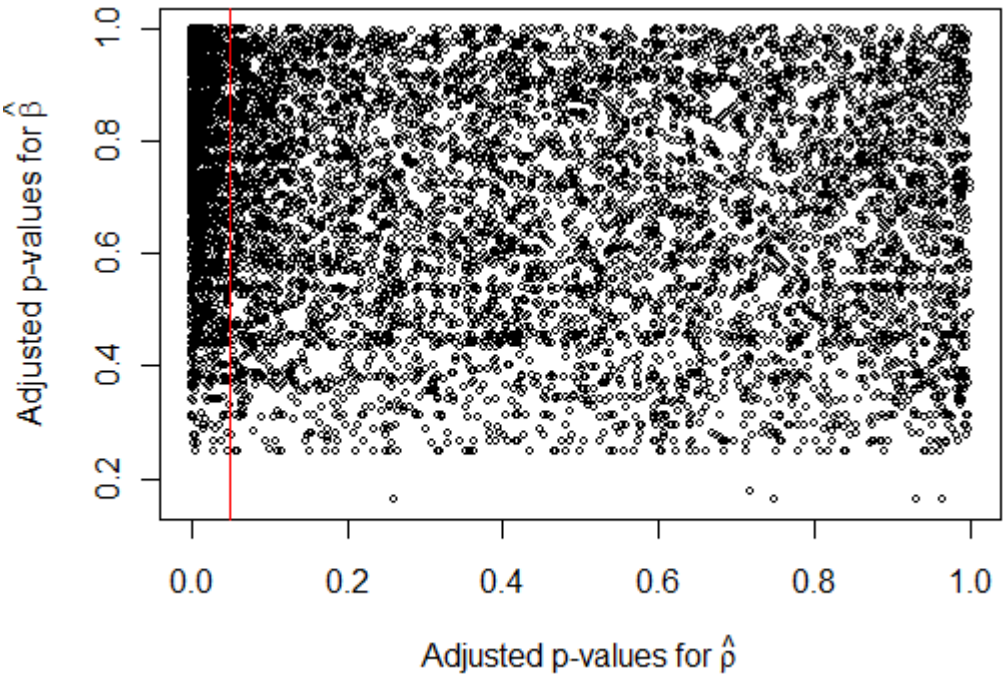


*Figure 2: Scatter plot of BH Adjusted P-values*

After fitting the joint model (4) on the data, to find the therapeutic and/or prognostic biomarkers null hypothesis in (6), $H_{20}: \rho_j = 0$, were tested to confirm correlation between $IC_{50}$ values and gene expression (5) and null hypothesis $H_{10}: \beta_j = 0$ were tested to confirm that gene is differentially expressed between the two groups of compounds. Note that, all the p-values found from the models were adjusted by multiplicity correction to control the False Discovery Rate (FDR). In Table 1 number of genes that are accepted or rejected for both the hypothesis is presented which reflects in Figure 2 where each spot represent a gene and all genes left to the red horizontal line have adjusted p-value of their estimated association parameter is less than 0.05. Similarly, there is no gene for which adjusted p-values of their estimated coefficient β are less than 0.05.

Table 1: Cross table of accepted and rejected genes for the hypothesis stated in (6) after adjusted p-values using BH Procedure

| | | $H_o: \rho_j = 0$ | | |
|---|---|---|---|---|
| | | Not rejected | Rejected | Total |
| $H_o: \beta_j = 0$ | Not rejected | 6239 | 1483 | 7722 |
| | Rejected | 0 | 0 | 0 |
| | Total | 6239 | 1483 | 7722 |

For the null hypothesis $H_{30}: \alpha = 0$, the model $E(Y_i|Z_i) = \mu_Y + \alpha Z_i$ in (3) was fitted independently and p-value of the estimated coefficient $\alpha$ is 0.0271 means null hypothesis is significant i.e. treatment effect on the outcome $IC_{50}$ is significant. The null hypothesis of no treatment effect on gene-expression, $H_{10}: \beta_j = 0$, is rejected for no genes after controlling False Discovery Rate (FDR) using Benjamini and Hochberg (1995) multiple testing procedure (BH Procedure) at level of 5%. So we have not found any therapeutic biomarker in this experiment. In other words, there are no genes that can be used to predict the effect of the drugs on $IC_{50}$ values.

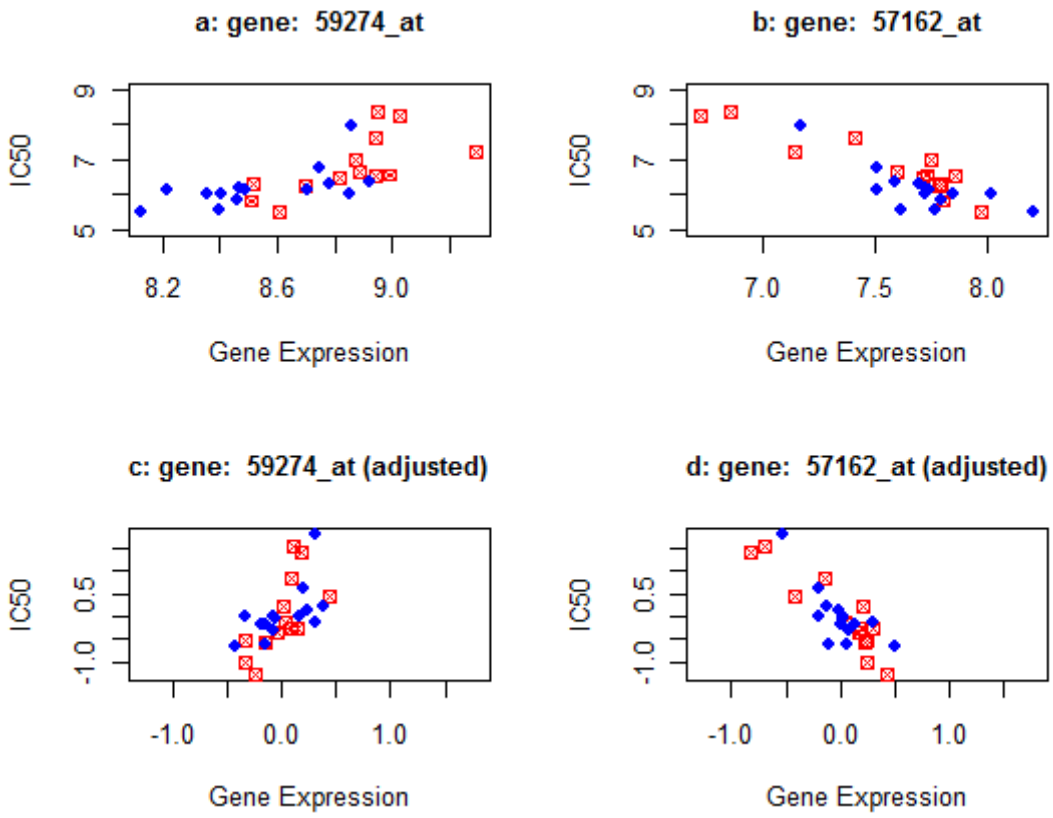


Figure 3: Two not differentially expressed but correlated genes

The null hypothesis $H_{20}: \rho_j = 0$ were rejected for 1483 genes. So these 1483 genes can be treated as prognostic biomarker. In Figure 3, such two genes were plotted as an example where circles and squares used to distinguish the groups. Scatter plot in panel a and b for the response ($IC_{50}$ values) versus gene expression and panel c and d are the residual plots of them respectively. Residuals were used for both the $IC_{50}$ values and gene expressions which can be regarded as adjusted values after removing the effect of treatments though effect of the treatment on gene expression is not significant. These genes are prognostic biomarker and can be use to explain genetic causes.

Model based on (6) to (7) were estimated and found no genes that have significant coefficient for its interaction term. So, according to Buyse et al. (2000) we can use model (9) instead of (7). Since the result of $R_{hj}^2$ in (8) are expected to be the same with squared of adjusted association $\rho_j$ in (5), so here $\rho_j$ were used as an estimation of $R_{hj}^2$. Figure 4 shows the distribution of $R^2$ where its reflects that most of the values of R2 are near to zero and only a few have higher value. In Table 2, top 20 genes with respect to adjusted p-value are listed

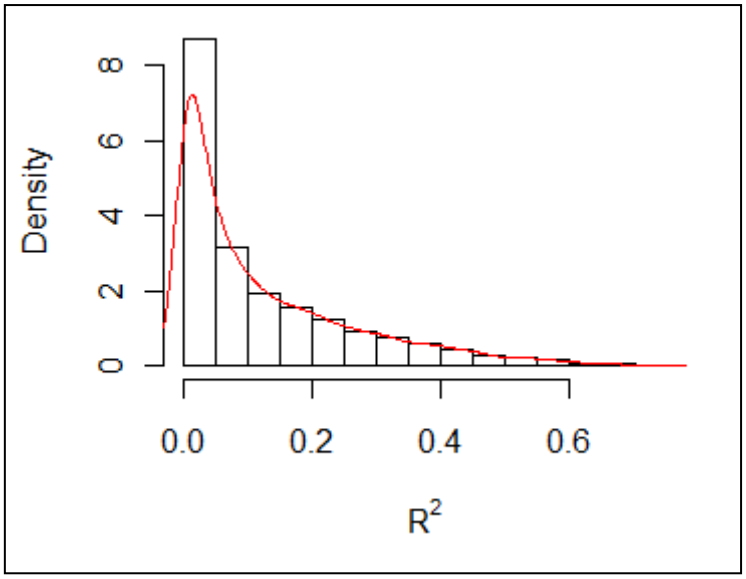


*Figure 4: Distribution of $R^2$*

Table 2: Results for top 20 genes. $R^2$ , Association measure based on adjusted correlation; $raw_p$: Raw *p*-values; $adj_p$: adjusted *p*-values

| | Gene Id | $R^2$ | Raw p | Adj P |
|---|---|---|---|---|
| 1 | 57162_at | 0.74177 | <0.00001 | 0.00003 |
| 2 | 7692_at | 0.70746 | <0.00001 | 0.00007 |
| 3 | 23286_at | 0.70128 | <0.00001 | 0.00007 |
| 4 | 57037_at | 0.68791 | <0.00001 | 0.00007 |
| 5 | 3491_at | 0.68362 | <0.00001 | 0.00007 |
| 6 | 5864_at | 0.68305 | <0.00001 | 0.00007 |
| 7 | 7057_at | 0.68038 | <0.00001 | 0.00007 |
| 8 | 4681_at | 0.67203 | <0.00001 | 0.00008 |
| 9 | 64864_at | 0.66856 | <0.00001 | 0.00008 |
| 10 | 8986_at | 0.66643 | <0.00001 | 0.00008 |
| 11 | 7003_at | 0.66620 | <0.00001 | 0.00008 |
| 12 | 5095_at | 0.65644 | <0.00001 | 0.00011 |
| 13 | 9454_at | 0.65220 | <0.00001 | 0.00011 |
| 14 | 9235_at | 0.65129 | <0.00001 | 0.00011 |
| 15 | 58480_at | 0.65093 | <0.00001 | 0.00011 |
| 16 | 388552_at | 0.64993 | <0.00001 | 0.00011 |
| 17 | 6450_at | 0.64789 | <0.00001 | 0.00011 |
| 18 | 7076_at | 0.64658 | <0.00001 | 0.00011 |
| 19 | 55624_at | 0.63570 | <0.00001 | 0.00014 |
| 20 | 10769_at | 0.63487 | <0.00001 | 0.00014 |

### 3.2 Joint Biomarkers using Supervised Principal Component Analysis

As described in the discussion, the gene specific models are used to identify optimum set of genes as possible biomarkers for the response. There is also a method that is called joint biomarkers that can be identify by using a set of methods and one of these is supervised principal component analysis (SPCA) with the expectation that more information will be gained from the joint biomarkers by constructing one kind of score of genes. In the previous section, results for gene specific models have been presented where 1483 genes were identified as prognostic genes. Here results from SPCA are presented along with set of genes as joint biomarkers.

In the gene specific model, 1483 prognostic genes were identified using their significance of the adjusted correlation with $IC_{50}$ values. If we rank them according to their significance of adjusted correlation i.e. adjusted p-values or according to any other thresholds then principal component of the subset of top k genes regarded as a possible biomarker. Here first principal component was considered which is justified in the discussions.

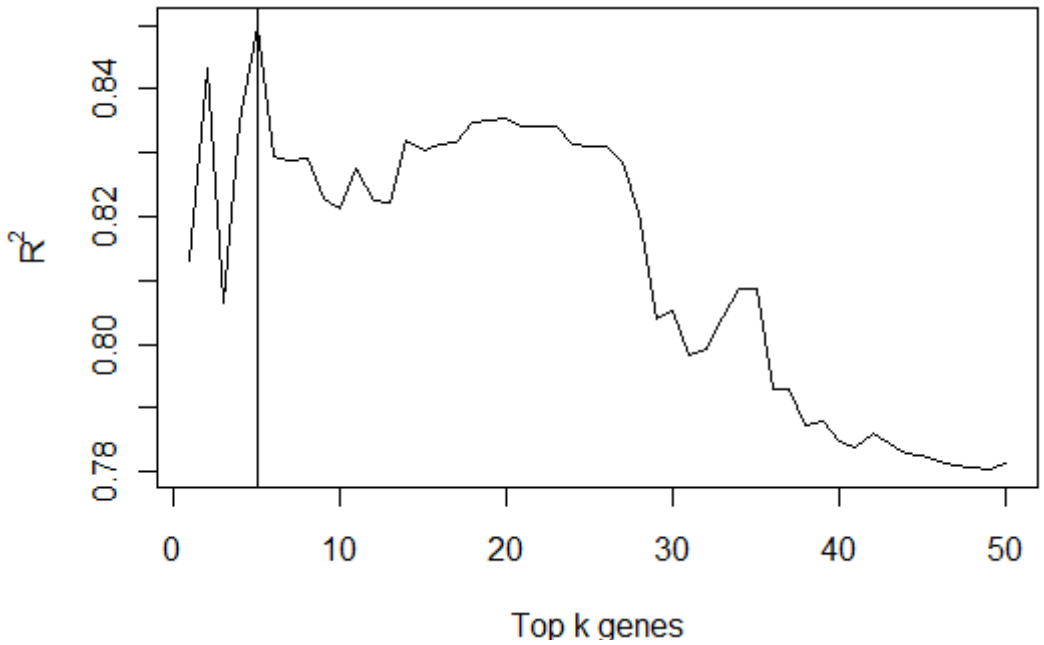


Figure 5: Plot of the $R^2_{hcv}$, association based on leave-one-out cross-validation for top genes selected based on $R^2_h$. For k=6, $R^2_{hcv}$ consists maximum value 0.8499

Following Tilahun et al. (2010), first approach among three approaches which involves taking a set of genes which are top *k* genes selected on the basis of their significance of adjusted correlation: $R_h^2$. Then first principal component were considered as a joint biomarker. This procedure applied for different size of genes set i.e. (k=2, 3, …., n). Figure 5 shows adjusted correlation between the first principal component of top k genes and response ($IC_{50}$). Maximum correlation occurred for k=23. But before that it has an irregular fluctuation, also fluctuation exists after that point though measured association decreases slowly. In Table 3, this measured association with association of one-leave-out cross validation and bootstrap permutation p-values were presented.

Table 3: Measured associations between response and first principal components of top genes selected based on gene specific adjusted correlation. $R_h^2$: *adjusted association* $R_{hcv}^2$: leave-one-out cross validation association and $P_{val}$: permutation v-values.

| Top | $R_h^2$ | $R_{hcv}^2$ | $P_{val}$ |
|---|---|---|---|
| 2 | 0.7761 | 0.8132 | <0.001 |
| 3 | 0.8107 | 0.8433 | <0.001 |
| 4 | 0.7650 | 0.8064 | <0.001 |
| 5 | 0.7995 | 0.8348 | <0.001 |
| 6 | 0.8198 | 0.8499 | <0.001 |
| 10 | 0.8133 | 0.8227 | <0.001 |
| 20 | 0.8314 | 0.8350 | <0.001 |
| 30 | 0.8181 | 0.8038 | <0.001 |
| 40 | 0.8018 | 0.7880 | <0.001 |

In the second approach, a gene has been considered if it contributes in increasing the correlation between the gene profile and the response. If we observe Figure 6 and Table 4, we can easily comment that using this gene profiling approach, 34 genes were considered in the construction of the gene profile that gives an $R_h^2$ value of 0.8931, which is higher than taking any gene set in the first procedures,

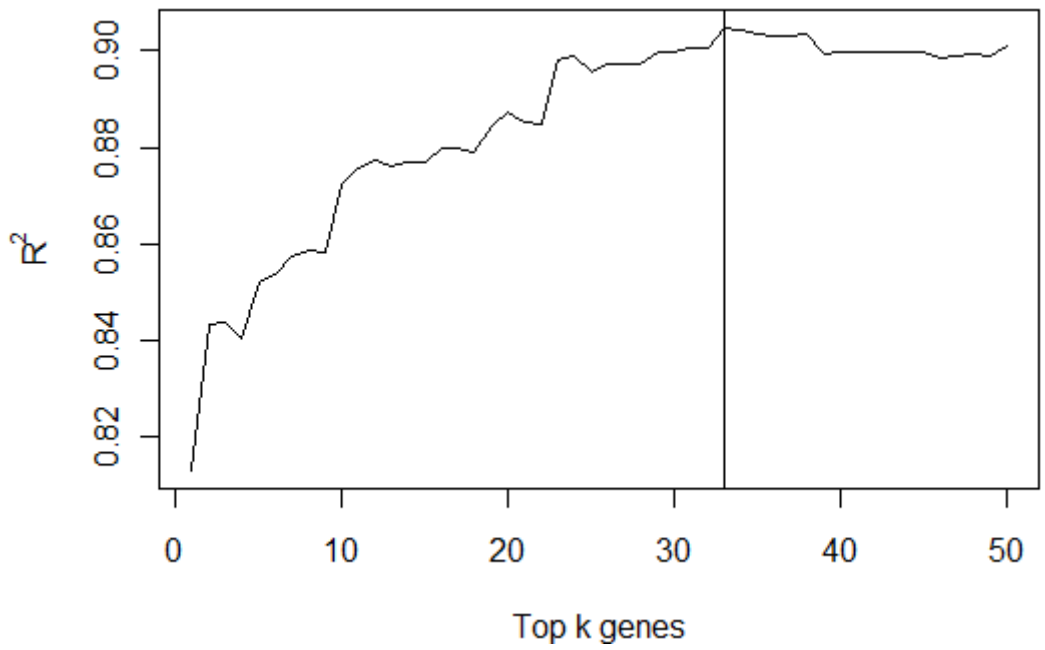


Figure 6: Plot of the $R^2_{hcv}$, association based on leave-one-out cross-validation for top genes selected based on contribution in the increase of $R^2_h$. For k=34, maximum $R^2_h$=0.8931

Table 4. Measured associations between response and first principal components of top genes selected based on their contribution on $R^2_h$: adjusted association $R^2_{hcv}$: leave-one-out cross validation association and $P_{val}$: permutation v-values.

| Top | $R^2_h$ | $R^2_{hcv}$ | $P_{val}$ |
|---|---|---|---|
| 2 | 0.7761 | 0.8132 | <0.001 |
| 3 | 0.8107 | 0.8433 | <0.001 |
| 4 | 0.8119 | 0.8437 | <0.001 |
| 5 | 0.8154 | 0.8403 | <0.001 |
| 10 | 0.8419 | 0.8585 | <0.001 |
| 15 | 0.8587 | 0.8770 | <0.001 |
| 20 | 0.8698 | 0.8842 | <0.001 |
| 25 | 0.8843 | 0.8989 | <0.001 |
| 30 | 0.8903 | 0.8995 | <0.001 |
| 31 | 0.8910 | 0.8997 | <0.001 |
| 32 | 0.8914 | 0.9003 | <0.001 |
| 33 | 0.8914 | 0.9003 | <0.001 |
| 34 | 0.8931 | 0.9046 | <0.001 |
| 35 | 0.8932 | 0.9044 | <0.001 |

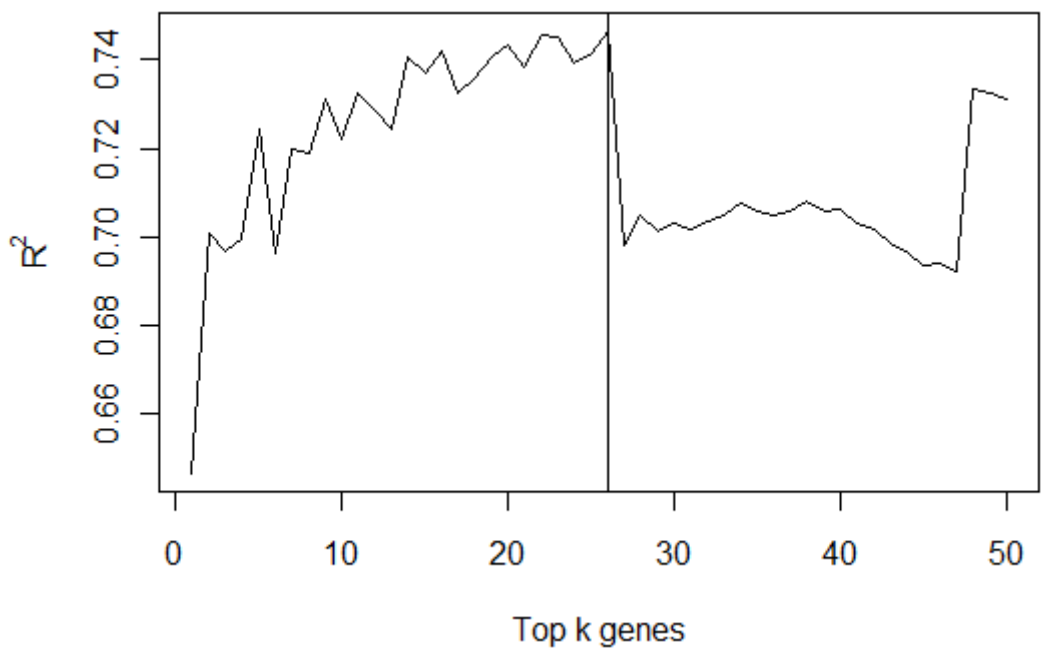


Figure 7: Plot of the $R^2_{hcv}$, association based on leave-one-out cross-validation for top genes selected based on contribution in the increase of $R^2_h$. For k=27, maximum $R^2_h$=0.8878

The third approach involves three steps as described in the methodology. 300 top genes were selected based on the gene specific adjusted association and then was constructed the first principal component. Genes were re-ranked based on their loadings of the first principal component i.e. absolute value of loadings and then joint biomarkers were constructed based on the top genes. Here factor loading can be considered weights in the ranking procedure where a gene having higher loading received higher weight.

Table 5: Measured associations between response and first principal components of top genes selected based on weights. $R^2_h$: adjusted association $R^2_{hcv}$: leave-one-out cross validation association and $P_{val}$: permutation v-values.

| Top | $R^2_h$ | $R^2_{hcv}$ | $P_{val}$ |
|---|---|---|---|
| 2 | 0.7761 | 0.8132 | <0.001 |
| 3 | 0.8107 | 0.8433 | <0.001 |
| 4 | 0.8119 | 0.8437 | <0.001 |
| 5 | 0.8154 | 0.8403 | <0.001 |
| 10 | 0.8419 | 0.8585 | <0.001 |
| 15 | 0.8587 | 0.8770 | <0.001 |
| 20 | 0.8698 | 0.8842 | <0.001 |
| 25 | 0.8843 | 0.8989 | <0.001 |
| 27 | 0.8878 | 0.8972 | <0.001 |
| 29 | 0.8886 | 0.8971 | <0.001 |
| 30 | 0.8903 | 0.8995 | <0.001 |

Figure 7 and Table 5 represent the results from the third procedure where gene set consisting 27 genes gave maximum $R^2_{hcv}$ for which $R^2_h$=0.8878 which is less than the second procedure.

## 4. Discussion and Conclusion

The main objective of this paper is to study joint model of phenotypic variables and gene expression data which can be used for compound screening. The purpose is to identify and evaluate genes as a biomarkers that that can be used to predict the surrogate end-point i.e. the response and as well as to identify a set of genes whose information used to construct a combined score i.e. joint biomarker that has maximum association with the response. In this study, there is only response ($IC_{50}$) as a phenotypic variable instead of more than one phenotypic variable.

Two modelling approaches: joint modelling and information-theoretic approach have been applied to select and evaluate genes that are strongly correlated with $IC_{50}$. These approaches involved measuring the linear association between the IC50 values with the gene expression by considering the treatment effect on both of them. According to the theoretical explanations both the approaches yielded similar results. But information-theoretic approach consists less calculating complexity than joint model approach. Furthermore, it is possible to distinguish between genes with positive and negative association with the response, directly from the model and information theoretic approach can readily applicable to non-normal settings, such as binary and time-to-event (Tilahun et al. 2010).

There was no differentially expressed gene that is no gene found to be significant that can identified as therapeutic biomarkers. Besides this, a lot of prognostic biomarkers found whose increased or decreased expression level affect on the value of $IC_{50}$. These facts insisted to advance the research for further use of other techniques.

Having some genes as prognostic biomarkers from the gene specific model, the interest gone through the use of supervised principal components analysis to construct a joint biomarker from a set of prognostic biomarkers. Since number of prognostic biomarkers is large enough and interest was on the less number of genes those information will construct such a joint biomarkers which association with the response will be maximum. For this purpose, two additional approaches, other than selecting top k genes based on adjuster association, were used to select small set of genes within the framework of supervised principal components

analysis. Selecting genes based on adjusted association always gives lower measures of association between joint biomarker and response than the two other alternative selection procedures where 23 genes identified that are able to construct a good biomarker. On the other hand, gene selection based on second alternative, i.e. a gene will be included in the set if and only if its inclusion results in increase in the magnitude of the association of the joint biomarker with the response produced a set of 34 genes to construct a good biomarker. The third approach or gene selection method is to create a set of top k genes whose absolute values of the first principal component are larger. This approach gave almost same result as of second alternative method but it produced a prognostic genes set of size 27 that joint biomarker maximum associated with the response. Permutation tests for all the measured association between joint biomarker and response were performed and showed that all the measured associations are highly significant.

In conclusion, joint biomarker has a great impact on $IC_{50}$ and hence it can be used to predict the response of interest effectively.

## References

Amaratunga, D. and Cabrera, J. (2004) *Exploration and Analysis of DNA Microarray and Protein Array Data*, New York: Wiley.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57(1)**, 289-300.

Buyse, M. and Molenberghs, G. (1998) The Validation of Surrogate Endpoints in Randomized Experiments, *Biometrics*, **54**, 186–201.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000) The Validation of Surrogate Endpoints in Meta-analyses of Randomized Experiments. *Biostatistics*, **1**, 49–67.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000), The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1(1)**, 49–67

Lin, D., Shkedy Z., et al (2010), Selection and evaluation of gene-specific biomarkers in pre-clinical and clinical microarray experiments. *Online Journal of Bioinformatics,* **11(1),** 106-127.

*Molenberghs, G., Burzykowski, T.,* et al (2008) The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials, *Journal of Statistical Planning and Infererence,* **138(2),** 432-449.

Suzy Van S. (2008) *Statistical Methods for Microarray-based Analysis of Gene-expression, Classification, and Biomarker Validation,* PhD dissertation, Interuniversity Institute for Biostatistics and statistical Bioinformatics, University of Hasselt, Belgium.

Tilahun, A., Lin, D., Shkedy Z., et al (2010) Genomic Biomarkers for Depression: Feature-Specific and Joint Biomarkers. *American Statistical Association Statistics in Biopharmaceutical Research* , **2(3)**, 419-434.

# Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
**Joint modeling of phenotypic variables and gene expression data for compound screening**

Richting: **Master of Statistics-Bioinformatics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

**Patwary, Md. Fazlul Karim**

Datum: **12/09/2011**