# Linear Models and Biomarker Search with Microarray Data

**E.M. Rivera[1], M.L. Sánchez-Peña[1], C.E. Isaza[1,2], J. Seguel[3], M. Cabrera-Ríos[1]**
[1]Bio IE Lab, Industrial Engineering Department, University of Puerto Rico - Mayagüez, Mayagüez, PR, USA
[2]Biology Department, Universidad Autónoma de Nuevo León, Sn. Nicolás de los Garza, NL, MEX
[3]Electrical &Computer Engineering Department, University of Puerto Rico - Mayagüez, Mayagüez, PR, USA

**Abstract –** *High throughput biological experiments such as DNA Microarrays are very powerful tools to understand and characterize multiple illnesses. These types of experiments, however, have also been described as large, complex, expensive and hard to analyze. For these reasons, analyses with linear assumptions are frequently bypassed for more sophisticated procedures with higher complexity. In this work, a search procedure for potential biomarkers using data from microarray experiments is proposed under purely linear assumptions. The method shows a high discrimination rate and does not require the adjustment of parameters by the user, thus preserving analysis objectivity and repeatability. A case study in the identification of potential biomarkers for cervix cancer is presented to illustrate the application of the proposed procedure.*

**Keywords:** Cancer biomarkers, Microarray Experiments,

## 1 Introduction

The search for genes whose measured change in expression behavior is an indication of a tissue being in a particular state (e.g. in a state of cancer vs. a state of health) is an important research objective in biology and the medical sciences. These genes are known as biomarkers. Microarray experiments play an important role in the identification of this type of genes. In the successful identification of potential biomarker genes, lies an important characterization of the cell in the presence of cancer. This can lead to enhance disease diagnosis and prognosis capabilities.

Based on our own experience with microarray data, the following challenges regarding microarray experiments can be identified: (1) the available data is highly dimensional in terms of the number of genes to be studied (~104) while showing a scarce number of replicates, (2) there is a rather large variation across replicates, (3) the data is not normally distributed and does not exhibit homogeneous variances, (4) there is a considerable number of missing observations in the majority of experiments, (5) the data is commonly found already being normalized or nonlinearly transformed. All of these complicate the detection of potential biomarkers.

Furthermore, when it comes to data analyses, the following are also important challenges: (i) there is no standard way to compare results for gene selection or identification between studies, (ii) even with the same data (and sometimes with the same technique) different researchers end up with different screening of genes [Ein-dor, et al. 2005] thereby leading to a large number of potential biomarkers to be investigated, the research of which could prove lengthy and very expensive.

Truly integrated work across disciplines is not frequent in most microarray analysis works. Biology and Medicine experts are usually left with the burden of using coded analysis tools with a series of parameters -of statistical, computational or mathematical nature/ that significantly affect the outcome of the software packages [Pan, 2002]. This leads to issues in results reproducibility and comparability between studies.

These challenges motivate the search for microarray analysis techniques from which consistent results can be achieved across several experiments and users, particularly for the identification of potential biomarkers.

The purpose of this work is to introduce an approach to identify potential biomarkers from the analysis of microarray experiments based solely on linear models and assumptions. Although an initial purpose on the design of the method was to establish a baseline of comparison for the many sophisticated methods with underlying nonlinear assumptions, it soon became apparent that a very effective strategy might be based on linearity.

## 2 The Analysis Strategy

Figure 1 schematically shows the strategy proposed in this work. Each step is explained below.

**Step 1: Microarray Experiment**. The process begins with a microarray experiment with m1 tissues in state one (Healthy) and m2 tissues in state two (Cancer) characterized in n genes. In the intersection of each of the n genes with each of the m1+m2 tissues, the relative expression of that particular gene in the selected tissue is quantified.

**Step 2: Represent each gene with multiple performance measures.** In this work, the use of a p_values is advocated to represent each gene. A p_value can be computed from the application of a statistical comparison test, like the Mann-Whitney nonparametric test for difference of medians. A different p_value for the same gene can be obtained by removing a couple of tissues from the microarray experiment under analysis. In a comparison of medians, a low p_value indicates a high probability for the medians to be significantly different.

**Step 3: Apply Data Envelopment Analysis.** Data Envelopment Analysis (DEA) finds the convex envelop of a particular data set consistently and without the need of varying parameters manually. If, for example, two p_values were used to represent each of the n genes in the experiment, then DEA can be used to find the envelope conformed by the dominating genes following the minimization direction of both p_values. Finding such envelope is done through the application of a linear programming formulation, which is the first instance where linearity becomes useful.

**Step 4: Select genes in a series of efficient frontiers.** The envelopes found through DEA are formally known as efficient frontiers. When an efficient frontier is found, then the solutions lying on it can be removed (as a layer of an onion), to then find the efficient frontier right underneath it. Following this scheme, several layers can be chosen containing different numbers of genes. These genes, having been found through the minimization of their p_values, are the most likely candidates to be biomarkers. These will be referred to as efficient genes.

**Step 5: Create an experimental design to vary the efficient genes.** An experimental design using as controllable variables the presence of the genes can be constructed. Each variable can take a value of 0 or 1 (0 for absence of the gene). This experimental design will prescribe a limited number of runs to measure a particular response of interest. In this case, one run corresponds to a combination of efficient genes.

**Step 6: At each experimental design point, measure classification performance through linear discriminant analysis.** Using the experimental design from the previous step, at each combination of efficient genes it is possible to obtain a measure of classification performance using a linear classifier through linear discriminant analysis. A linear classifier of this kind will always converge to the same position, thus preserving consistent results. At this point, then, a complete experimental design relating the classification rate with the absence or presence of the potential biomarkers is available.

**Step 7: Fit a 1st order linear regression model.** With the complete experimental design, it is possible to fit a 1st order linear regression model. This model will relate classification performance (response) to the absence or presence of the efficient genes (independent variables).

**Step 8: Apply integer linear programming to choose the potential biomarkers that maximize classification performance.** An optimization problem can be set up in this stage. This problem entails finding the combination of efficient genes –recall that each gene is represented by a variable that can take values of 0 or 1 to indicate absence or presence of that gene-, that maximizes the classification performance, i.e. choose the genes that maximize the regression model from the previous step.

This procedure, as it was explained, uses only linear models. Because of the techniques chosen in the strategy, the results are consistent. Furthermore, the selected genes do not depend upon the setting of any parameters by the user. This favors the repeatability and auditability of the analysis.

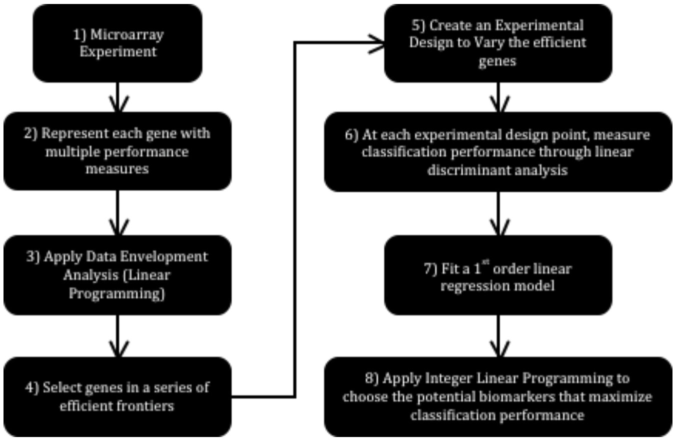
1) Microarray Experiment
2) Represent each gene with multiple performance measures
3) Apply Data Envelopment Analysis (Linear Programming)
4) Select genes in a series of efficient frontiers
5) Create an Experimental Design to Vary the efficient genes
6) At each experimental design point, measure classification performance through linear discriminant analysis
7) Fit a 1st order linear regression model
8) Apply Integer Linear Programming to choose the potential biomarkers that maximize classification performance

**Figure 1**. Analysis Strategy based on Linear Models

## 3 Case Study on Cervix Cancer

This case study helps to illustrate the application and the performance of the proposed procedure.

**Step 1**. The microarray database under analysis is related to cervix cancer and was compiled by Wong et al [3]. The database consists of 8 healthy tissues and 25 cervix cancer tissues, all of them with expression level readings for 10,690 genes.

**Step 2**. The Mann-Whitney nonparametric two-sided test for comparison of medians was used to generate two different p_values per gene, following a leave-one-tissue-out strategy, which focuses on extracting a particular tissue associated with one state. By removing a vector, a replicate is deleted from the set, thereby forcing a p_value that is different to the original one. Thus, two different p_values are effectively created. The selection of the tissue to be removed to create a distinct matrix is performed randomly as a first approach.

**Step 3.** The Data Envelopment Analysis model used for this case study was the Banks-Charnes-Cooper (BCC) model [4].

This is a linear programming model with the following associated formulations:

$$\begin{array}{ll} \text{Find} & \boldsymbol{\mu}, \boldsymbol{\nu}, \mu_0^+, \mu_0^- \quad \text{to} \\ \text{Maximize} & \boldsymbol{\mu}^T \mathbf{Y}_0^{\max} + \mu_0^+ - \mu_0^- \\ \text{Subject to} & \end{array}$$

$$\begin{gathered} \boldsymbol{\nu}^T \mathbf{Y}_0^{\min} = 1 \\ \boldsymbol{\mu}^T \mathbf{Y}_j^{\max} - \boldsymbol{\nu}^T \mathbf{Y}_j^{\min} + \mu_0^+ - \mu_0^- \le 0 \quad j = 1,\dots,n \\ \boldsymbol{\mu}^T \ge \varepsilon \cdot \mathbf{1} \\ \boldsymbol{\nu}^T \ge \varepsilon \cdot \mathbf{1} \\ \mu_0^+, \mu_0^- \ge 0 \end{gathered}$$

$$\begin{array}{ll} \text{Find} & \boldsymbol{\nu}, \boldsymbol{\mu}, \nu_0^+, \nu_0^- \quad \text{to} \\ \text{Minimize} & \boldsymbol{\nu}^T \mathbf{Y}_0^{\min} + \nu_0^+ - \nu_0^- \\ \text{Subject to} & \end{array}$$

$$\begin{gathered} \boldsymbol{\mu}^T \mathbf{Y}_0^{\max} = 1 \\ \boldsymbol{\nu}^T \mathbf{Y}_j^{\min} - \boldsymbol{\mu}^T \mathbf{Y}_j^{\max} + \nu_0^+ - \nu_0^- \ge 0 \quad j = 1,\dots,n \\ \boldsymbol{\nu}^T \ge \varepsilon \cdot \mathbf{1} \\ \boldsymbol{\mu}^T \ge \varepsilon \cdot \mathbf{1} \\ \nu_0^+, \nu_0^- \ge 0 \end{gathered}$$

The optimal values of the decision variables correspond to the interceptor and the partial first derivatives (with respect of each performance measure involved) of a supporting hyperplane lying on top of extreme points of the data set under analysis. At the end of the analysis, a piece-wise frontier is distinguishable as shown in Figure 3.

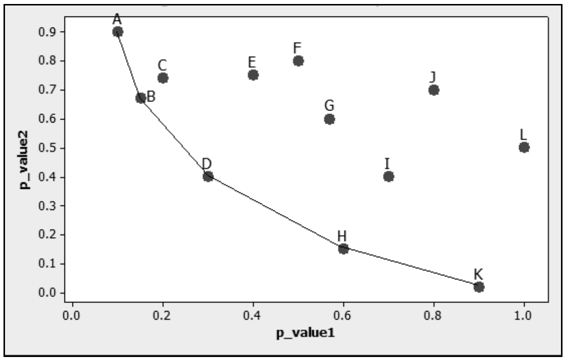

**Figure 3.** Representation of genes characterized through two different p_values. Only the case with 2 p_values has a convenient graphical representation, but the analysis can be extended to as many dimensions as performance measures selected.

**Step 4.** The first ten frontiers were kept for this analysis containing a total of 28 genes. It is important to note the discrimination rate shown by the method already at this point: a reduction of four orders of magnitude in the number of genes to analyze.

**Step 5.** A composite experimental design involving 28 binary variables (one per gene in the shortlist from the previous step), was used. Three different experimental designs form the composite with 123 runs. The first design is an orthogonal array consisting on 47 runs with between 10 to 18 genes each; the second design has 48 runs with between 1 to 26 genes generated randomly; and the third design consisted of 28 runs, each with only one gene. Figures 4, 5 and 6 show the resulting designs.

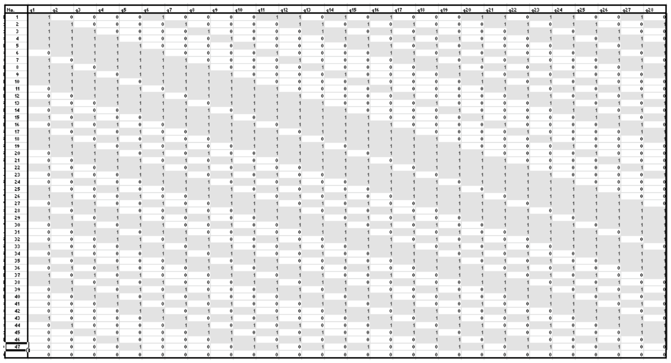

**Figure 5**. Design of Experiment 1. Shaded in gray are the values of 1.

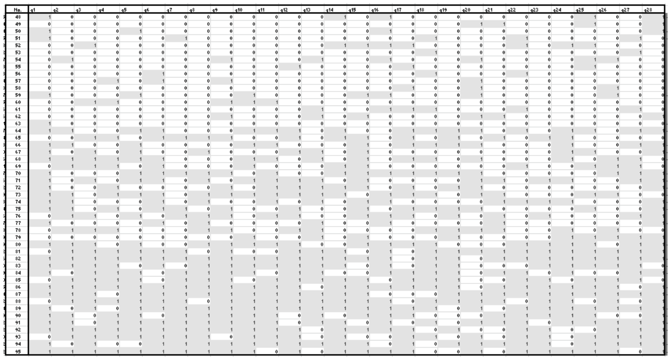

**Figure 6**. Design of Experiment 2 (Runs 1-16: 20% of total number of genes, runs 16-32: 50% of total number of genes, runs 33-48: 80% of total number of genes). Shaded in gray are the values of 1.

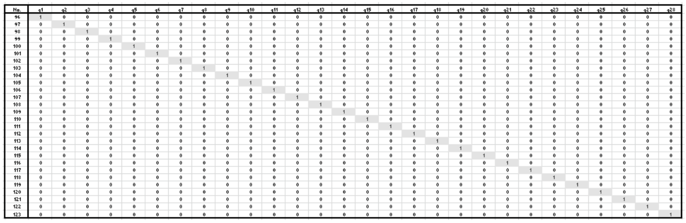

**Figure 7**. Design of Experiment 3. Shaded in gray are the values of 1.

**Step 6**. A linear discriminant analysis was carried out using the combination of genes prescribed by each run of the composite design to record the classification performance of a linear classifier.

**Step 7**. With the experimental design complete, a linear regression of the classification performance as a function of the presence or absence of the 28 genes is built as shown in Table 1.

| Variable | Coefficient Symbol | Regression Coefficient |
|---|---|---|
| | b0 | 0.8868 |
| g1 | **b1** | 0.0152 |
| g2 | **b2** | 0.0027 |
| g3 | **b3** | 0.0097 |
| g4 | **b4** | 0.0146 |
| g5 | **b5** | 0.0030 |
| g6 | **b6** | 0.0083 |
| g7 | **b7** | -0.0034 |
| g8 | **b8** | 0.0051 |
| g9 | **b9** | 0.0001 |
| g10 | **b10** | 0.0054 |
| g11 | **b11** | 0.0008 |
| g12 | **b12** | -0.0020 |
| g13 | **b13** | 0.0120 |
| g14 | **b14** | -0.0027 |
| g15 | **b15** | 0.0138 |
| g16 | **b16** | 0.0089 |
| g17 | **b17** | 0.0166 |
| g18 | **b18** | 0.0145 |
| g19 | **b19** | 0.0089 |
| g20 | **b20** | 0.0120 |
| g21 | **b21** | 0.0137 |
| g22 | **b22** | 0.0105 |
| g23 | **b23** | -0.0068 |
| g24 | **b24** | -0.0025 |
| g25 | **b25** | 0.0093 |
| g26 | **b26** | 0.0050 |
| g27 | **b27** | 0.0079 |
| g28 | **b28** | 0.0158 |

**Table 1.** Linear Regression Model using 123 experimental designs.

**Step 8**. Using the linear regression model from Table 1, the optimization model is to find the combination of genes (through the use of binary variables) to maximize the predicted classification performance. Such optimization resulted in the identification of 23 important genes, that is, potential cervix cancer biomarkers. These are shown in Table 2.

Currently, our group is working on the validation of these potential biomarkers, as well as on their representation in a hierarchical list or a relationship network.

| Index | Frontier | Accession Number | Optimization Selection |
|---|---|---|---|
| 1 | 1 | AA488645 | **X** |
| 2 | 2 | H22826 | **X** |
| 3 | 3 | AI553969 | **X** |
| 4 | 3 | T71316 | **X** |
| 5 | 3 | AA243749 | **X** |
| 6 | 3 | AA460827 | **X** |
| 7 | 4 | AA454831 | |
| 8 | 4 | AA913408,<br>AA913864 | **X** |
| 9 | 5 | AA487237 | **X** |
| 10 | 5 | AA446565 | **X** |
| 11 | 6 | H23187 | **X** |
| 12 | 7 | AI221445 | |
| 13 | 7 | R36086 | **X** |
| 14 | 7 | AA282537 | |
| 15 | 8 | N93686 | **X** |
| 16 | 8 | R91078 | **X** |
| 17 | 8 | R44822 | **X** |
| 18 | 9 | AI334914 | **X** |
| 19 | 9 | R93394 | **X** |
| 20 | 9 | AA621155 | **x** |
| 21 | 9 | AA705112 | **x** |
| 22 | 9 | R52794 | **x** |
| 23 | 10 | AA424344 | |
| 24 | 10 | H69876 | |
| 25 | 10 | H55909 | **x** |
| 26 | 10 | W74657 | **x** |
| 27 | 10 | AI017398 | **x** |
| 28 | 10 | H99699 | **x** |

**Table 2.** The procedure selected 23 potential biomarkers through the maximization of the expected classification performance.

## 4 Conclusions

In this work, a strategy to detect potential biomarkers from the analysis of microarray experiments is proposed. The strategy is based solely on linear models and assumptions. Its consistent convergence and lack of parameter setting by the users, make this method a very competitive and attractive one for repeatability and auditability. This is especially important in high throughput experiments and in a highly interdisciplinary field like bioinformatics. A case study involving the analysis of a microarray database on cervix cancer was presented to demonstrate the capabilities of the strategy. Indeed, in this case study it was possible to discriminate among more than 10,000 genes to converge to 23 potential cervix cancer biomarkers. These are currently under analysis for validation in our research group.

## 5 Acknowledgement

This work was made possible thanks to the NIH-MARC grant "Assisting Bioinformatics Efforts at Minority Institutions" PAR-03-026 and BioSEI UPRM grant 330103080301.

## 6 References

[1] Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005 Jan 15;21(2):171-178.

[2] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics. 2002 Apr 1;18(4):546-554.

[3] Wong YF, Selvanayagam ZE, Wei N, Porter J, Vittal R, Hu R, Lin Y, Liao J, Shih JW, Cheung TH, Lo KW, Yim SF, Yip SK, Ngong DT, Siu N, Chan LK, Chan CS, Kong T, Kutlina E, McKinnon RD, Denhardt DT, Chin KV, Chung TK. Expression Genomics of Cervical Cancer: Molecular Classification and Prediction of Radiotherapy Response by DNA Microarray. Clin Cancer Res. 2003 Nov 15;9(15):5486-92.

[4] Charnes A, Cooper WW, Lewin AY, Seiford LM. Data Envelopment Analysis: Theory, Methodology, and Applications. Boston MA, USA: Kluwer Academic Publishers.1993.