CONSISTENT DETECTION OF CANCER BIOMARKERS WITH LINEAR MODELS

E.M. Rivera¹, Z. Irizarry¹, M.L. Sánchez-Peña¹, C.E. Isaza^{1,2}, J. Seguel³, M. Cabrera-Ríos¹

¹Bio IE Lab, Industrial Engineering Department, University of Puerto Rico - Mayagüez, PR

²Biology Department, Autonomous University of Nuevo León, Sn. Nicolás de los Garza, NL, MEX

³Electrical & Computer Engineering Department, University of Puerto Rico - Mayagüez, PR

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, as well as a validation study for colon cancer is presented to illustrate the application of the proposed procedure.

Keywords: Cancer biomarkers; Microarray Experiments.

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.

Analysis Strategy

Fig. 1 schematically shows the strategy proposed in this work. Each step is explained below. Maintaining the Integrity of the Specifications

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.

Figure 1. Analysis Strategy based on Linear Models

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:

$$\begin{aligned}
 &\text{Find } \mu, \mathbf{v}, \mu_0^+, \mu_0^- \quad \text{to} \\
 &\text{Maximize } \mu^T \mathbf{Y}_0^{\max} + \mu_0^+ - \mu_0^- \\
 &\text{Subject to} \\
 &\quad \mathbf{v}^T \mathbf{Y}_0^{\min} = 1 \\
 &\quad \mu^T \mathbf{Y}_j^{\max} - \mathbf{v}^T \mathbf{Y}_j^{\min} + \mu_0^+ - \mu_0^- \leq 0 \quad j = 1, \dots, n \\
 &\quad \mu^T \geq \varepsilon \cdot \mathbf{1} \\
 &\quad \mathbf{v}^T \geq \varepsilon \cdot \mathbf{1} \\
 &\quad \mu_0^+, \mu_0^- \geq 0
 \end{aligned}$$

$$\begin{aligned}
&\text{Find} && \mathbf{v}, \boldsymbol{\mu}, v_0^+, v_0^- && \text{to} \\
&\text{Minimize} && \mathbf{v}^T \mathbf{Y}_0^{\min} + v_0^+ - v_0^- \\
&\text{Subject to} && \boldsymbol{\mu}^T \mathbf{Y}_0^{\max} = 1 \\
&&& \mathbf{v}^T \mathbf{Y}_j^{\min} - \boldsymbol{\mu}^T \mathbf{Y}_j^{\max} + v_0^+ - v_0^- \geq 0 \quad j = 1, \dots, n \\
&&& \mathbf{v}^T \geq \varepsilon \cdot \mathbf{1} \\
&&& \boldsymbol{\mu}^T \geq \varepsilon \cdot \mathbf{1} \\
&&& v_0^+, v_0^- \geq 0
\end{aligned}$$

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 hyper plane 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 Fig. 2.

Figure 2. 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. In this step a composite experimental design involving 28 binary variables (one per gene in the shortlist from the previous step), was created and used. In it are a series of combinations (also called runs) that represent the importance of specific genes (named potential biomarkers) and their expression in cancerous or healthy tissues. The experimental design consisted of 122 different runs in which each gene could take a value of 0 (absence) or 1 (presence). In each run the contribution of each gene to detect each tissue is quantified, may it be a set of genes or a gene by itself. Three different experimental designs form the total of 122 runs. The first design is an orthogonal array consisting on 47 runs with between 10 to 18 genes each; the second design has 47 runs

with between 1 to 26 genes generated randomly; and the third design consisted of 28 runs, each with only one gene. Fig. 3 shows the composite experimental design.

Figure 3. Composite Experimental Design. 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. Depending on the absence and presence of certain genes in each run a range of different classification performances will result. It is expected that if the biomarkers are efficient and the more genes present in a specific run, the better the model will classify the tissue (healthy or cancerous). Using an analogy, given the most characteristics possible of a person the better to identify the person will be. However each single gene is also being tested for this analysis and it should be noted that their contribution to the classification performance was more than sufficient (See DOE 3 for detailed information).

DOE 1		DOE 2		DOE 3	
No.	CP	No.	CP	No.	CP
1	0.37	48	0.37	36	0.303
2	1	49	1	37	0.873
3	1	50	1	38	0.873
4	1	51	1	39	0.303
5	1	52	0.303	100	0.873
6	1	53	0.333	101	0.303
7	1	54	0.333	102	0.818
8	1	55	0.333	103	0.873
9	1	56	1	104	0.818
10	1	57	0.333	105	0.873
11	1	58	1	106	0.873
12	1	59	1	107	0.818
13	1	60	1	108	0.848
14	1	61	1	109	0.873
15	1	62	0.37	110	0.873
16	1	63	0.303	111	0.848
17	1	64	1	112	0.333
18	1	65	1	113	0.873
19	1	66	1	114	0.848
20	1	67	1	115	0.788
21	1	68	1	116	0.873
22	1	69	1	117	0.303
23	1	70	1	118	0.758
24	1	71	1	119	0.818
25	1	72	1	120	0.873
26	1	73	1	121	0.818
27	1	74	1	122	0.818
28	1	75	1	123	0.873
29	1	76	1		
30	1	77	1		
31	1	78	1		
32	1	79	1		
33	1	80	1		
34	1	81	1		
35	1	82	1		
36	1	83	1		
37	1	84	1		
38	1	85	1		
39	1	86	1		
40	1	87	1		
41	1	88	1		
42	1	89	1		
43	1	90	1		
44	1	91	1		
45	1	92	1		
46	1	93	1		
47	1	94	1		
		95	1		

Tables 1-3. Classification performance for each experimental design.

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.

Noticing that most of the Classification Performances give a 100% classification, two different regression models were made (from two different samples). The first included all of the 122 original runs from the experimental designs and by consequent the 122 classification performances. For the second model, the sample was reduced to only 38 runs, using the 28 runs from DOE 3 and 10 random ones from DOE 2. By doing this a hierarchy of the genes can be extracted. With the regression, the relationship between each “important” gene and the classification performance can be seen in a linear and

$$\hat{CP} = \hat{\beta}_0 + \hat{\beta}_1 g_1 + \hat{\beta}_2 g_2 + \dots + \hat{\beta}_{28} g_{28} \quad (1)$$

$$g_1, g_2, \dots, g_{28} \in \{0,1\} \quad (2)$$

Variable	Coefficient Symbol	Regression Coefficient
	β_0	0.8868
g_1	β_1	0.0152
g_2	β_2	0.0027
g_3	β_3	0.0097
g_4	β_4	0.0146
g_5	β_5	0.003
g_6	β_6	0.0083
g_7	β_7	-0.0034
g_8	β_8	0.0051

g_9	β_9	0.0001
g_{10}	β_{10}	0.0054
g_{11}	β_{11}	0.0008
g_{12}	β_{12}	-0.002
g_{13}	β_{13}	0.012
g_{14}	β_{14}	-0.0027
g_{15}	β_{15}	0.0138
g_{16}	β_{16}	0.0089
g_{17}	β_{17}	0.0166
g_{18}	β_{18}	0.0145
g_{19}	β_{19}	0.0089
g_{20}	β_{20}	0.012
g_{21}	β_{21}	0.0137
g_{22}	β_{22}	0.0105
g_{23}	β_{23}	-0.0068
g_{24}	β_{24}	-0.0025
g_{25}	β_{25}	0.0093
g_{26}	β_{26}	0.005
g_{27}	β_{27}	0.0079
g_{28}	β_{28}	0.0158

Table 4. Linear regression model using 122 experimental designs.

Table 4 represents the first regression model. It shows that in order to obtain a classification performance of a 100% at least 23 different specific genes are needed. This behavior comes from the small variability in the classification performances for the 122 runs. To obtain a yet better and simpler model we reduced the dataset used for the regression as explained before.

Variable	Coefficient Symbol	Regression Coefficient
1	β_0	0.854962
g_1	β_1	0.02688
g_2	β_2	0.01723
g_3	β_3	0.04222
g_4	β_4	0.02565
g_5	β_5	0.02926
g_6	β_6	0.05826
g_7	β_7	-0.0345
g_8	β_8	-0.00439
g_9	β_9	-0.03323
g_{10}	β_{10}	0.0265
g_{11}	β_{11}	0.0189
g_{12}	β_{12}	-0.01878
g_{13}	β_{13}	0.03906
g_{14}	β_{14}	0.02826
g_{15}	β_{15}	0.04174
g_{16}	β_{16}	-0.02205
g_{17}	β_{17}	0.04799
g_{18}	β_{18}	0.00947
g_{19}	β_{19}	-0.00714
g_{20}	β_{20}	-0.01286
g_{21}	β_{21}	0.01577

g_{22}	β_{22}	0.02561
g_{23}	β_{23}	-0.08053
g_{24}	β_{24}	-0.02284
g_{25}	β_{25}	0.00324
g_{26}	β_{26}	-0.03274
g_{27}	β_{27}	-0.02898
g_{28}	β_{28}	0.0265

Table 5. Linear regression model using 38 experimental designs.

After reducing the dataset to only 38 runs the model becomes more targeting, for it now has added variability to it. The model represented in Table 5, shows that in order to obtain a 10% classification performance only 17 genes are needed. It also shows like the first model the specific genes in order to obtain it.

Step 8. Using the two linear regression models described before, the optimization model is to find the combination of genes (through the use of binary variables) to maximize the predicted classification performance using Linear Integer Programming. Equation 3 shows the objective function and its constraints.

$$\begin{aligned} \text{Max_CP} &= f(g_1, g_2, \dots, g_{28}) \\ \text{Subject to,} \\ g_1, g_2, \dots, g_{28} &\in \{0,1\} \end{aligned} \quad (3)$$

Index	Frontier	Accession	Optimization
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, 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 6. Optimization models.

Such optimization resulted in the identification of 17 important genes, that is, potential cervix cancer biomarkers. Also, a hierarchy of genes can be found using this procedure, which helps visualize exactly what genes and in which order are most necessary and significant when it comes to detect Cervix Cancer.

Shown in Table 7 are the 17 potential biomarkers drawn from the second model. Also the hierarchy and the classification performance are shown, after adding the subsequent gene to the model. CP_Predicted come from the Classification Performance obtained using the Linear Regression from model 2. CP_Real is the Classification Performance obtained by applying LDA method.

Genes	Hierarchy	Classification Performance Predicted	Classification Performance Real
1	6	0.91	0.91
2	17	0.96	0.97
3	3	1.00	0.94
4	15	1.00	0.97
5	13	1.00	0.97
6	5	1.00	1.00
7	14	1.00	1.00
8	1	1.00	1.00
9	10	1.00	1.00
10	28	1.00	1.00
11	4	1.00	1.00
12	22	1.00	1.00
13	11	1.00	1.00
14	2	1.00	1.00
15	21	1.00	1.00
16	18	1.00	1.00
17	25	1.00	1.00

Table 7. Hierarchy of genes to obtain a maximized classification performance.

Validation on Colon Cancer

As part of our study and to reinforce the performance of the proposed method we proceeded to validate the steps illustrated before with another microarray database.

In this validation part the microarray database under analysis was related to colon cancer. A different cancer type

database was chosen to prove the effectiveness of the method among different types of cancer. This database consisted of 22 healthy tissues and 40 colon cancer tissues, all of them with expression level readings of 2000 genes. We kept only 27 genes of the 2000 original genes prevenient of the first ten frontiers founded through DEA. Table 8 shows the resulting efficient genes prevenient from the DEA analysis.

Genes	Frontier	Accession
g ₁	1	M22382
g ₂	1	R87126
g ₃	2	H08393
g ₄	2	R36977
g ₅	3	J05032
g ₆	3	M26383
g ₇	4	X63629
g ₈	4	H40095
g ₉	4	Z50753
g ₁₀	4	M63391
g ₁₁	5	J02854
g ₁₂	5	X12671
g ₁₃	6	U09564
g ₁₄	6	H43887
g ₁₅	6	M76378
g ₁₆	7	M36634
g ₁₇	7	T86473
g ₁₈	8	H06524
g ₁₉	8	R84411
g ₂₀	8	X14958
g ₂₁	9	T92451
g ₂₂	9	M26697
g ₂₃	9	T71025
g ₂₄	10	X86693
g ₂₅	10	T47377
g ₂₆	10	U30825
g ₂₇	10	D31885

Table 8. Efficient genes prevenient of DEA analysis.

After the identification of these 27 efficient genes, a composite experimental design was created and these 27 efficient genes were used to vary their efficiency. The experimental design involved 27 binary variables and formed 37 runs. The first design consists of 10 runs with between 18 to 24 genes generated randomly; and the second design consisted of 27 runs, each with only one gene activated. Fig. 4 shows the resulting designs.

Figure 4. Composite Experimental Design (10 runs with between 18 to 24 genes generated randomly and 27 runs with only one gene activated). Shaded in gray are the values of 1.

Using the composite experimental design we measured and recorded the classification performance through linear discriminant analysis using the different combination of genes given by each run. Results are shown in Table 9.

Run	Classification Performance
1	0.968
2	0.952
3	0.952
4	0.919
5	0.919
6	0.919
7	0.952
8	0.887
9	0.919
10	0.935
11	0.694
12	0.823
13	0.71
14	0.726
15	0.629
16	0.597
17	0.71
18	0.661
19	0.758
20	0.855
21	0.79
22	0.71
23	0.694
24	0.758
25	0.823
26	0.774
27	0.661
28	0.726
29	0.677
30	0.694

31	0.79
32	0.677
33	0.774
34	0.774
35	0.742
36	0.677
37	0.645

Table 9. Classification performances for experimental designs.

After the completion of the experimental designs a linear regression model was used to relate the classification performance with the absence or presence of each of the 27 efficient genes. The results gave us the identification of 15 potential colon cancer biomarkers. This means that to obtain a classification performance of 100% at least these 15 genes are needed. Table 10 shows a list of these potential colon cancer biomarkers.

Variable	Coefficient Symbol	Regression Coefficient
1	β_0	0.716
g_1	β_1	-0.00379
g_2	β_2	0.081
g_3	β_3	0.01
g_4	β_4	-0.02
g_5	β_5	-0.1
g_6	β_6	-0.08
g_7	β_7	-0.01
g_8	β_8	-0.03
g_9	β_9	0.027
g_{10}	β_{10}	0.134
g_{11}	β_{11}	0.068
g_{12}	β_{12}	0.01
g_{13}	β_{13}	-0.05
g_{14}	β_{14}	0.077
g_{15}	β_{15}	0.074
g_{16}	β_{16}	0.036
g_{17}	β_{17}	-0.05
g_{18}	β_{18}	0.01
g_{19}	β_{19}	-0.05
g_{20}	β_{20}	0
g_{21}	β_{21}	0.016
g_{22}	β_{22}	-0.01
g_{23}	β_{23}	0.035
g_{24}	β_{24}	0.01
g_{25}	β_{25}	0.028
g_{26}	β_{26}	0

Table 10. Genes needed to obtain a 100% classification performance.

Once completed the linear regression analysis we used this equation to find the combination of genes that maximize

the classification performance applying integer linear programming. The use of this tool allowed us to obtain a hierarchy of genes that permit us to know the significance of each potential colon cancer biomarker in the detection of cancer. The hierarchy obtained in the colon cancer study is shown in Table 11.

Genes	Hierarchy	Prediction	Classification Performance-Real
1	10	0.85	0.855
2	2	0.931	0.855
3	14	1	0.855
4	15	1	0.855
5	11	1	0.855
6	16	1	0.855
7	23	1	0.839
8	25	1	0.871
9	9	1	0.903
10	21	1	0.903
11	3	1	0.919
12	12	1	0.935
13	24	1	0.935
14	18	1	0.935
15	27	1	0.935

Table 11. Hierarchy of genes to obtain a maximized classification performance.

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 and validation study on colon cancer were presented to demonstrate the capabilities of the strategy. Indeed, in the studies 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.

Acknowledgements

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.

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
- [5] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*. 1999;96(12):6745-6750.

Biographical Sketch

Eulisa Rivera, is an undergraduate student in the Industrial Engineering Department at the University of Puerto Rico at Mayagüez and a member of the Bio IE lab. Her interests include operations research and statistics applied to healthcare and biological systems. She is currently a member of the Institute of Industrial Engineers and the Society for Healthcare Systems and plans to pursue doctoral studies in the near future. http://ininweb.uprm.edu/mcabrera/eng_home.html