Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression

Francielly Morais-Rodrigues, Rita Silv´erio-Machado, Rodrigo Bentes Kato, Diego Lucas Neres Rodrigues, Juan Valdez-Baez, Vagner Fonseca, Emmanuel James San, Lucas Gabriel Rodrigues Gomes, Roselane Gonçalves dos Santos, Marcus Vinicius Canário Viana, Joyce da Cruz Ferraz Dutra, Mariana Teixeira Dornelles Parise, Doglas Parise, Frederico F Campos, Sandro J de Souza, José Miguel Ortega, Debmalya Barh, Preetam Ghosh, Vasco A.C. Azevedo, Marcos A dos Santos

PII: S0378-1119(19)30827-3
DOI: https://doi.org/10.1016/j.gene.2019.144168
Reference: GENE 144168

To appear in: *Gene Gene*

Received Date: 5 July 2019
Revised Date: 21 September 2019
Accepted Date: 11 October 2019

Please cite this article as: F. Morais-Rodrigues, R. Silv´erio-Machado, R.B. Kato, D.L.N. Rodrigues, J. Valdez-Baez, V. Fonseca, E.J. San, L.G.R. Gomes, R.G. dos Santos, M. Vinicius Canário Viana, J. da Cruz Ferraz Dutra, M. Teixeira Dornelles Parise, D. Parise, F.F. Campos, S.J. de Souza, J.M. Ortega, D. Barh, P. Ghosh, V.A.C. Azevedo, M.A. dos Santos, Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression, *Gene Gene* (2019), doi: https://doi.org/10.1016/j.gene.2019.144168

Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression

Francielly Morais-Rodrigues[a,*,1,franrodriguesdacosta@gmail.com], Rita Silvério-Machado[a], Rodrigo Bentes Kato[a], Diego Lucas Neres Rodrigues[a], JuanValdez-Baez[a], Vagner Fonseca[a,b], Emmanuel James San[b], Lucas Gabriel Rodrigues Gomes[a], Roselane Gonçalves dos Santos[a], Marcus Vinicius Canário Viana[a,c], Joyce da Cruz Ferraz Dutra[a], Mariana Teixeira Dornelles Parise[a], Doglas Parise[a], Frederico F Campos[d], Sandro J de Souza[e], José Miguel Ortega[a], Debmalya Barh[f], Preetam Ghosh[g], Vasco A. C. Azevedo[a], Marcos A dos Santos[d]

[a]Institute of Biological Sciences, Federal University of Minas Gerais, Brazil. Av. Antônio Carlos, 6627, Belo Horizonte, MG 31270-901, Brazil

[b]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZulu-Natal, Durban 4001, South Africa

[c]Federal University of Pará, UFPA, Brazil

[d]Department of Computer Science, Federal University of Minas Gerais, Brazil Av Antônio Carlos, 6627 Belo Horizonte, MG 31270-901, Brazil

[e]Brain Institute, Federal University of Rio Grande d oNorte, Brazil

[f]Centre for GenomicsandApplied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal 721172, India

[g]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

[*]Corresponding author.

[1]First author.

Abstract: Methods based around statistics and linear algebra have been increasingly used in attempts to address emerging questions in microarray literature. Microarray technology is a long-used tool in the global analysis of gene expression, allowing for the simultaneous investigation of hundreds or thousands of genes in a sample. It is characterized by a low sample size and a large feature number created a non-square matrix, and by the incomplete rank, that can generate countless more solution in classifiers. To avoid the problem of the 'curse of dimensionality' many authors have performed feature selection or reduced the size of data matrix. In this work, we introduce a new logistic regression-based model to classify breast cancer tumor samples based on microarray expression data, including all features of gene expression and without reducing the microarray data matrix. If the user still deems it necessary to perform feature reduction, it can be done after the application of the methodology, still maintaining a good classification. This methodology allowed the correct classification of breast cancer sample data sets from Gene Expression Omnibus (GEO) data series GSE65194, GSE20711, and GSE25055, which contain the microarray data of said breast cancer samples. Classification had a minimum performance of 80% (sensitivity and specificity), and explored all possible data combinations, including breast cancer subtypes.This methodology highlighted genes not yet studied in breast cancer, some of which have been observed in Gene Regulatory Networks (GRNs). In this work we examine the patterns and features of a GRN composed of transcription factors (TFs) in MCF-7 breast cancer cell lines, providing valuable information regarding breast cancer. In particular, some genes whose αi∗associated parameter values revealed extreme positive and negative values, and, as such, can be identified as breast cancer prediction genes. We indicate that the PKN2, MKL1, MED23, CUL5 and GLI genes demonstrate a tumor suppressor profile, and that the MTR, ITGA2B, TELO2, MRPL9, MTTL1, WIPI1, KLHL20, PI4KB, FOLR1 and SHC1 genes demonstrate an oncogenic profile. We propose that these may serve as potential breast cancer prediction genes, and should be prioritized for further clinical studies on breast cancer. This new model allows for the assignment of values to the αi∗ parameters associated with gene expression. It was noted that some αi∗ parameters are associated with genes previously described as breast cancer biomarkers, as well as other genes not yet studied in relation to this disease.

Keywords: Tumor classification, samples, new logistic regression-based model, GRN, TFs, MCF-7, oncogenic.

Abbreviations: BC, Breast cancer; BGRMI, Bayesian Gene Regulatory Model Inference; EGF, Epidermal growth factor; GEO, Gene Expression Omnibus; GRNs, Gene Regulatory Networks; HER2, Human Epidermal Growth Factor Receptor 2; HRG, Cells stimulated with heregulin; LumA, Luminal A; LumB, Luminal B; NCBI, National Center for Biotechnology Information; RNA-Seq, RNA sequencing; TFs, Transcription factors; TNBC, Triple Negative Breast Cancer;

1. Introduction

In the past few years, there has been a growing interest in the application of methods of linear algebra and statistics in data mining, machine learning, bioinformatics, and in other areas (Asnaoui *et al*., 2016; Ding *et al*., 2018; Enshaeifar *et al*., 2019; Piwowar *et al*., 2018). Among these methods, logistic regression approach draw some particular interest as it is a standard method for data classification using gene expression data and is the most frequently used method for disease prediction, with good results for cancer classification as shown by many (Bazzoli and Lambert-Lacroix, 2018; Li *et al*., 2018).

The classifiers, which use several methodologies as logistic regression, are the core component of microarray data analysis. With the emergence of the DNA microarray technology it has enabled the generation of massive microarray data of gene expression and used for the discovery and classification of diseases. Over the years the characteristics of these data have remained virtually unchanged, among them, which allows simultaneous investigation of hundreds of thousands of genes in a sample, small sample size and a greater number of features. Because of this dimension one can generate 'curse of dimensionality' a problem that needs to be treated (Li *et al*., 2018; Shao *et al*., 2019), coined by Bellman, 2015, and that in general terms is the widely observed phenomenon that data analysis techniques frequently perform poorly as the dimensionality of the analyzed data increases. Conceptually, the samples are lost in the features space as the dimensionality increases and we would need an enormous number of samples to obtain a satisfactory estimate of, for example, which genes have altered expression patterns in a specific tumor type. Many algorithms have been developed to deal with the high-dimensionality problem in microarray studies including the ones that are based on distance functions, clustering or dimensionality reduction (de Meulder *et al*., 2018; Pereda *et al*., 2018; Toledano *et al*., 2018) Although this genomic tool is not new (Schena *et al*., 1995), and despite this technology has several limitations and a powerful technology named RNA sequencing (RNA-seq) is predicted to replace it for transcriptome profiling, the microarray has been matured in the last years, with the emergence of high quality arrays due to standardized hybridization protocols, accurate scanning technologies, and robust computational methods. Therefore their became use has been extended to a broad spectrum of biologically relevant studies and researches, like as the clinical researches the microarray together with advanced transcription kits designed for low input derived from microdissected tissue generating better results. For this favorable information about the microarrays, they are still the one common choice of researches gene expression analysis. (Shao *et al*., 2019; Wimmer *et al*.,2018; Zhao *et al*.,2018).

Microarray-based gene expression profiling is used to classify a multitude of tumor types, to determine which treatment methods will most likely yield beneficial results for particular cancer patients and to predict cancer-specific biomarkers in this disease and this technology is applied in breast cancer (BC) classifiers. (Duan *et al*., 2018; Gálvez *et al*., 2018; Gong, *et al*., 2018; Kagaris *et al*., 2018).

The breast cancer is a very heterogeneous disease with significant variability between patients. Breast tumors can be grouped in four molecular subtypes, which have major implications for determining treatment (Luminal A-LumA, Luminal B-LumB, Triple negative-TNBC/Basal-like, Human Epidermal Growth Factor Receptor 2-HER2 type) (Chiu *et al*., 2018; Gálvez *et al*., 2018), and in 2016 and 2019 cancer incidence and mortality statistics reported by the American Cancer Society and by the United Kingdom Office for National Statistics indicate breast cancer as one of the four most common cancer types, along with lung, colorectal, and prostate. Breast cancer together with lung, and colorectal are expected to account for 30% of all new cancer diagnoses in women in in those years, being the most frequently diagnosed cancer in women.

According by American Cancer Society in relation as breast cancer death rates were registered in the women a declined 40% from 1989 to 2016 because of early detection, but coined by Bellanger *et al*., 2018, the poorest countries have higher burden of breast cancer mortality particularly women younger than age 50 years. Because of these statistics and the importance of the fastest possible detection of breast cancer, in this paper we analyzed three microarray data sets of patients with breast cancer distributed in several cancer subtypes and we introduce a new logistic regression-based model to classify breast cancer tumor samples based on microarray expression data and with no initial reduction of features' dimensionality. This new model allows the assignment of values to intercept *αi* parameters*,* which are associated with the expression of a certain gene. Scrutinizing these *αi*∗ parameters unveiled that some of the parameters topologically located further away from the majority of the parameters are associated with known breast cancer related genes and flagged others for further investigation inside gene regulatory networks from search of Iglesias-Martinez, *et al*., 2016.

The advent of high-throughput genomics that started with DNA microarrays has been revealing the intricate regulatory dynamics that reshape the gene expression programs of a cell even, under the mostsubtle perturbations. Considerable effort has been directed towards methods for the efficient analysis and interpretation of whole transcriptome read-outs, including differentially expressed genes for which there is little if any knowledge related to the system under study (Li *et al*., 2018; Shao *et al*., 2019; Wimmer *et al*.,2018; Zhao *et al*.,2018). Genome-wide data available has allowed the development of methods to infer the gene regulatory program responsible for an observed expression profile. Regulatory mechanisms foster proper genetic interactions that maintain health and perturbations of gene regulatory networks (GRNs) are essentially responsible for both oncogenesis and breast cancer maintenance. Therefore, the use of a network approach to the study of cancer systems is critical to overcoming cancer. In a GRN, collections of interacting DNA elements (indirectly through their RNA and protein expression products) in a cell are represented, thereby indicating the influence a gene product has on the expression rate of gene i. Gene regulation takes place in various stages with many participants among which, transcription factors (TFs) are the ones most readily analyzed and easy to quantify (Gao, *et al*., 2018; Iglesias-Martinez, *et al*., 2016; Joo, *et al*., 2018).

In this work we investigate the patterns and features of a GRN composed of TFs in MCF-7 breast cancer cell lines (Iglesias-Martinez, et al., 2016), to provide valuable information relating to breast cancer . Some particular genes of these networks had α_i^* associated parameters values with extreme positive and negatives values from all of the 6 systems created from GSE65194 dataset. Next, we try to predict breast cancer associated genes (Figure 2). Finally, we prognosticate their roles as oncogenes or tumour suppressor genes in breast cancer, using the S-score system, which integrates genome-wide data (de Souza, et al., 2014). By flagging prediction cancer genes here, we expect to provide new breast cancer biomarkers, which could make a beneficial contribution to minimizing mortality rates by providing a better prognosis.

2 Materials and Methods

2.1 Modified Logistic Regression

2.1.1 Materials: Data Collection and Generation

A collection of three available data sets containing microarray data of breast cancer samples, with no missing data, was used to demonstrate the usefulness of the proposed methodology. Data sets were downloaded from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) with the identifiers GSE65194 and GSE 20711, acquired using Affymetrix Human Genome U133 Plus 2.0 arrays, and GSE25055, acquired using Human Genome HG U133A Affymetrix arrays. The table 1 with the formation and with the Modified Logistic Regression Models were created for each datasets.

2.1.2 Methods: Modified Logistic Regression Model

The data obtained from microarray experiments is represented by matrix $\boldsymbol{A} = \{x_{i,j}\}$ with m rows and n columns, with rows representing patients and columns representing genes. The value of each position $x_{i,j}$ represents the expression levels of a certain gene j for a patient i. We will omit the indication of row i in the elements of vector x. That is $x = \{x_1, x_2, \ldots, x_n\}$ every time row i to which x refers to is clear in the context. Associated with each row i is $P_i(x) = 0/1$ that informs the origin of the gene profile (no membership or membership of breast cancer/breast cancer subtype). The logit function expressed for each patient is given by:

$$P_i(x) = g_i(x) / \left(1 + g_i(x)\right) \quad (1)$$

where

$$g_i(x) = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \alpha_{n+1}) \quad (2)$$

for $i = 1, 2, \ldots, m$ and exp is the exponential function ($exp(x) = e^x$).

The logistic regression consists on finding a vector $\alpha = (\alpha_1, \ldots, \alpha_n, \alpha_{n+1})^T$ to fit the set of equations (1). We observe that when $g_i(x)$ drops to zero, $P_i(x)$ also drops to zero. On the other hand, if $g_i(x)$ tends to infinity, $P_i(x)$ approximates to one. Viewing $P_i(x)$ as the probability, the odds $C_i(x)$ are given by:

$$C_i(x) = P_i(x) / \left(1 - P_i(x)\right) \quad (3)$$

Expressing equation (3) using (1), one obtains:

$$C_i(x) = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \alpha_{n+1}) \quad (4)$$

To implement the method, one uses $\hat{C}_i(x) \approx C_i(x) = (0.99 / (1 - 0.99))$ instead of $C_i(x)$, when the odds are related to $P_i(x) = 1$. When $P_i(x) = 0$, one considers $\hat{C}_i(x) \approx C_i(x) = (0.01 / (1 - 0.01))$. Letting $b_i = log\,(\hat{C}_i(x))$ and taking the logarithm on both sides of (4), a linear algebraic model is created to determine α:

$b_i = (\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \alpha_{n+1})$ (5)

for i=1, 2, . . . , m.

Let $\bar{e}$=[1, . . . , 1]^T be a vector of m ones and b=(b_1, b_2, . . . , b_m)^T. The system of linear equations (5) may be represented by:

$B\alpha = b$ (6)

with

$B = [A\,\bar{e}]$ (7)

In (7) there are fewer equations than unknowns, and the system is undetermined, with an infinite number of solutions. The classical approach in linear algebra minimizes α subject to $B\alpha = b$, which requires full rank of B^TB a property not expected to behold by matrix B. It is usual to circumvent this difficulty by pruning the model and keeping only a small subset of the n genes. This procedure resembles the feature selection in data mining an open research area.

We propose the usage of a stabilizing term in the logistic regression model found in the works of Linnik 1961, Golub 1965 and Menard 2010, that allows the assignment of values to α parameters by minimizing the square sum of the residuals ($B\alpha$ - b), summed to the squares of α. So to assign a solution to (7), we solve an unconstrained quadratic optimization problem given by:

$Minimize: \ f(\alpha) = \alpha^T\alpha + (B\alpha - b)^T(B\alpha - b)$ (8)

Note that B^TB is a positive semi-definite matrix, so it has at least one negative self-determinant of zero and so is a singular matrix, has no inverse, and the rank is incomplete. Unlike this situation, it is known that a square matrix is full-rank if, and only if, it is non-singular, for this was added identity matrix (I) with B^TB as can be seen in equation (9). The identity matrix is positively defined so it is not singular, which implies having a complete rankand allowing the system to have a unique solution, and it happens and is presented in equation 9 (Bapat, 2012; Boldrini et al., 1980; Harville, 2011).

$(I + B^TB)\alpha = B^T b$ (9)

where I is an identity matrix of dimension n.

One should note that the identity matrix does not allow the rank to become deficient. The optimal solution α^* to (8) is obtained by the solution to (9) and it is unique. So, given a query q=[q_1, q_2. . . ,q_n] with the levels of expression of n genes, the probability of q to be associated with a breast cancer subtype is given by:

$P(q) = g(q) / (1 + g(q))$, (10)

where:

$g(q) = \exp([q\ 1]\alpha)$. (11)

2.2 Gene regulatory network

2.2.1 Materials: Data Collection

The Gene Regulatory Networks (GRNs) of breast cancer cells inferred from time-course gene expression data and reconstructed using the algorithm Bayesian Gene Regulatory Model Inference (BGRMI) was used in the analysis presented here (Iglesias-Martinez, *et al*., 2016).The flagged genes were used as input data for the prediction of their roles as oncogenes or tumor suppressor genes in breast cancer or in a specific breast cancer subtype, using the S-score system with negative value to indicate of tumor suppressing or reduced gene activity and positive value to indicate of oncogene or increased gene activity (de Souza, *et al*., 2014).

2.2.2 Methods: Gene regulatory network and S-score

To scrutinize this prospect the genes that matched the features represented by the most positive and negative αi∗ parameters were flagged to search the literature. The 20 most positive values and 20 most negative values αi∗ parameters were selected from all of the 6 systems created from GSE65194 dataset, were searched at Gene Regulatory Networks (GRNs) presented in the paper by Iglesias-Martinez, et al., 2016 and received S-score values from Souza, et al., 2014.

3 Results and Discussion

3.1 Results and Discussion: Modified Logistic Regression Model

For logistic regression method application, two classes coded as 1 and 0 are attributed to each patient sample and one computes the probabilities that given some explanatory variables a patient belongs to the coded classes. We applied the proposed classification methodology to all established systems created from the three mentioned breast cancer datasets, aiming to perform binary classification discriminating between the subtypes of breast cancer and between each subtype and non-tumor breast tissue samples. Frequently when a model is presented, the principle that less is always more is followed and so the possibility of variable reduction is frequently explored (Bazzoli and Lambert-Lacroix, 2018; Li *et al*., 2018; Lien *et al*., 2018; Shao *et al*., 2019).The model that we propose here circumvents the need for initial variable pruning by aggregating the quadratic term to the solution of a system of equations to determine the value of α_{i^*} associated parameters. It was observed that the reduction of features can be applied after applying the methodology and still maintain a good classification of the classifier. (Figure 1 inside article and Figures 1 and 2 inside supplementary material illustrates these results). The αi∗ parameters that are associated with gene expression and do not have a important role in any of the classification models are indirectly removed from the model as their αi∗ associated parameters are either zero or close to zero. The parameters associated with gene expression, that to have important role in breast cancer progression or do not have, are located on the extremes (Figure 2).

Logistic regression provides a good method for classification by modelling the probability of membership of a class based on linear combinations of exploratory variables. Classical logistic regression models do not work for microarray data because generally there will be far more variables (the measured expression levels) than observations (Lien *et al.*, 2018; Xu *et al.*, 2018).

One particular problem is multicollinearity: the estimated equations have no unique solution. The modified logistic regression model proposed in this work provides a solution to this problem, with no need for previous feature selection or matrix dimensionality reduction. The key point for the development of this model is the inclusion of a stabilizing term that allows the assignment of values to α parameters by minimizing the square sum of the residuals ($B\alpha$ - b) summed to the squares of α, allowing the system to have a unique solution . Applying the concepts of logistic regression and with all samples and all features included, we were able to correctly classify all samples from all data sets used, with a minimum performance of 80% from a cross-validation procedure (see in supplementary material) and exploring all possible combinations of data, establishing a good model for breast cancer classification.

When plotting the α_i^* parameters calculated upon classification of samples in all systems, it has not escaped our notice the intriguing distribution of the elements. At this point we hypothesized that the model created could suggest a framework to study gene expression in the context of a feature selection procedure, and also that α_i^* parameter value is close to zero every time the expression of gene i is irrelevant to the computation of the function logit. Keeping this in mind we selected the 500 α_i^* parameters topologically located on the extremes (250 most positive and 250 most negative) of each of the 6 systems created from GSE65194 dataset, matched the corresponding genes and subsequently built a list with 462 genes which arise as the most frequently occurring genes corresponding to the selected α_i^* parameters. Of these genes from the list, 320 were finded in GSE20711 and 319 were finded in GSE25055 datasets, then new systems were created considering just these genes, but with the same framework as used in the previously created systems for these datasets, small sample size and a greater number of features. The new proposed logistic regression model was applied to the newly created systems with five rounds of misclassification cross-validation. Assessment of the Probability of Misclassification was applied too and were obtained with a minimum performance of 80% from a cross validation procedure (see in supplementary material). These results reveal that the α_i^* parameters topologically located at the extremes are relevant for classification and we prognosticate that the genes that matched these features are associated with biomarkers with potential diagnostic and/or therapeutic utilities in breast cancer. Figures 1 and 2 inside supplementary material illustrate the classification of the samples present in all the systems created from GSE20711 and GSE25055 datasets, respectively.

The probability of misclassification is the most important property of a classifier because it quantifies the predictive capability of the classifier (Bazzoli and Lambert-Lacroix, 2018; Li *et al*., 2018; Ding *et al*., 2018; Gálvez *et al*., 2018; Pereda *et al*., 2018) The feature label distribution is known for the data set used and so the true error can be exactly found. For one round of cross-validation of the misclassification rate, a random subset of 15% of samples from different data sets (27 patients from GSE65194 dataset, 13 from GSE20711 dataset and 46 from GSE25055 dataset). The new proposed logistic regression model with the stabilizing term was applied to the subsampled datasets and to the created subsets, in order to evaluate the performance of the classification. To reduce variability, five rounds of cross-validation were performed. Cut-off values (values are shown in the supplementary material) were defined and for those, the modified logistic regression model correctly classified 88%, 89% and 94% (average values) of the patients for the five rounds of patients extracted from GSE65194, GSE20711 and GSE25055 datasets, respectively.

All possible combinations of data (provided in supplementary material) were explored and the new proposed model was able to classify all samples in all data sets, considering all possible combinations of data, with good performance. This also discloses that the removal of subsets did not disrupt the matrix organization structure.

To assess the discriminatory power of the proposed method, we also performed relationship between both sensitivity and specificity (representing values for Roc curves). For all possible combinations of data explored, the sensitivity and specificity values range is 0.8 - 1 (supplementary material).

3.2. Results and Discussion: Genes associated to the 20 most positive values and 20 most negative values αi∗ parameters, from all of the 6 systems created from GSE65194 dataset, inside Gene Regulatory Networks (GRNs) from search Iglesias-Martinez, et al., (2016).

GRNs operate as a “map” or a “blue print” of molecular interactions, helping to solve a number of different biological and biomedical problems (Gao, 2018). Molecular networks in mammal cells control cell proliferation and differentiation(Liang *et al*., 2018; Wang *et al*., 2018; Yuan *et al*., 2018). Some researchers propose that cancer is a particular cell state associated with complex molecular networks therefore, the transformation from “normal cells” to cancer cells is governed by network landscape changes, which contribute to cancer cell autonomy (Wang, 2018; Zhang *et al.*, 2018). Given a specific stimulus or under specific conditions, the relative abundance of a high number of mRNA species may vary due to changes resulting from the activation of a particular gene expression program in the Gene Regulatory Networks. As such, the molecular mechanisms that govern proliferation and differentiation in breast cancer cells can be studied by measuring the gene expression profile of MCF-7 cells stimulated with heregulin(HRG) and epidermal growth factor (EGF) over time. These stimuli artificially induce differentiation and proliferation, respectively: HRG induces a sustained signal activity in MCF-7 breast cancer cells which triggers an irreversible cell phenotype change toward differentiation (accumulation of lipid droplets within the cells) and EGF only elicits a transient signal activity in these cells that drives them toward proliferation. The human breast carcinoma cell line MCF-7 constitutes a powerful system for the study of breast cancer, as in the past information derived from this powerful experimental tool has translated into clinical benefit (Iglesias-Martinez, *et al.*, 2016; Majumder, *et al.*, 2018; Roncato, *et al.*, 2018; Takagi, *et al.*, 2018).

Since pathological cells manifesting tumors have their own characteristic networks; which dove us to cherry-pick the GRNs that were reconstructed, in the research of Iglesias-Martinez and collaborators, using expression profiles of MCF-7 cells after artificially inducing proliferation and differentiation. Looking at the results after the application of the new proposed logistic regression model, the topological location of some particular genes whose α_i^* associated parameters values reveal extreme positive and negative values, we explored the correlation of these genes with breast cancer in literature, and the following are known crucial genes associated with breast cancer, transcription factors identified as the busiest junctions in these GRNs, as well as some other transcription factors already reported as having an important role on breast cancer development.

Each Gene associated to the 20 most positive values and 20 most negative values $\alpha i*$ parameters and selected for all of the 6 systems created from GSE65194 dataset, was searched for its involvement with S-scores in the research of Souza, *et al.* (2014). According to the author the negative value S-score is indicative of tumor suppressing or reduced gene activity and the positive value S-score is indicative of oncogene or increased gene activity. But he noted that even classifying in this way, some genes that had a classification by their S-score in one of the two groups (positive or negative) the literature indicated their existence in the opposite group. The following the tables 2 and 3 with value S-score for the genes associated with features that represent the $\alpha i*$ parameters of each of the systems created from GSE65194 data set and the participation these genes in the Gene Regulatory

Networks (GRNs) from search Iglesias-Martinez, *et al.*, (2016), so were also searched for its involvement in the the EGF stimulated cells and for stimulation of BC cells with HRG. These stimulations leads to variations of the landscape topography in the GRN reconstructed using this data.

The $\alpha i*$ parameters associated with gene expression that to have discriminatory role are topologically located on the extremes, but we still had to find a way to evidence the roles of these genes, and S-score it may be possible. In this work we hypothesized that the majority of the genes associated with the αi * topologically located at the positive side have oncogenic role, and in negative side are found of the genes protection factors as shown by in Figure 2. S-score confirming our hypothesis that genes are suppressive and other oncogenic as shown in Table 2 and Table 3, but other genes were not confirmed by this value.

But some genes are not inside the EGF and HRG networks. The gene Tox (expressed in the subtypes TNBC and Lum B) located at the positive side have oncogenic role (Figure 2) and has S-score 2.77. The genes KCNJ12 (expressed in the subtypes HER2 and LumA) with S-score 1.8, KIAA1009 (expressed in the subtypes TNBC, LumA and LumB) with S-score -1.34, NUDT7 (expressed in the subtypes HER2 and LumA) with S-score 0 and STADARD9 no S-score, are located at the positive side have oncogenic role (Figure 2). And in negative side are found of the genes protection factors as shown by in Figure 2: ZNHIT2 with S-score -0.02 (TNBC, LumA, LumB); the next need to negative but they not FLJ35024 S-score 0.36 (LumA, LumB), ALDH1L1 S-score 1.52 (HER2, LumA, LumB), LPHN3 S-score 1.50 (HER2, LumA), SOCS5 S-score 0.27 (TNBC, LumA), KLHL7 S-score 1.35 (TNBC, HER2, Lum B); and the next genes do not have S-score FCRL2(LumA), VDR-Cdx2 (LumA), SPAG11B (HER2, LumA,LumB).

To the best of our knowledge there are no evidences in the literature of association of genes ADCY9, SLC25A17, KIAA1009, WIZ, KIAA0355, ZNF174, MVK, EYA4, NUDT7, TCHP, SPAG11B, SEMA6A, FCRL2, KLHL7, KCNJ12, CCDC126, GTF2E1, SKIV2L2, TN-FRSF10A and PTGDS with breast cancer.

4 Conclusions

We introduce here a new logistic regression-based model to classify breast cancer tumor samples based on microarray expression data with all features included and no reduction of microarray data matrix and that has also put a light on some genes as breast cancer related. This methodology allowed the correct classification of all samples from all data sets tested, with a minimum performance of 80% and exploring all possible combinations of data, establishing a good model for breast cancer classification. The key point for the development of this model is a stabilizing term that allows the assignment of values to $\alpha_i{}^*$ parameters, allowing the system to have a unique solution. These parameters are related with the expression of the genes that are related to the progression of breast cancer and others genes not been studied yet. The application of the methodology does not generate diagnosis to be used by physicians in consultations with common patients. The present work, being computational only, is an assumption of reality and is being offered as a first step for laboratories to try to perform an analysis of TF role in gene expression need to be validated at the protein expression and interaction levels both in dry lab and wet lab. And this rearch used microarray with gene expression values from healthy and breast cancer patients to create a classifier based on these values. These microarrays were built based on these expressions and do not bring metabolic dysregulations at the protein level.

References

American Cancer Society (2019) Facts & Figures 2019: US Cancer Death Rate has Dropped 27% in 25 Years. Available: https://www.cancer.org/latest-news/facts-and-figures-2019.html.

Asnaoui,K. EL. *et al*. (2016) An Application of Linear Algebra to Image Compression (). In: Badawi,A., Vedadi,M.R., Yassemi,S., Darani,A.Y. (eds.) *Homological and Combinatorial Methods in Algebra*, pp.41-54 Springer, Iran (2016).

Bapat, R.B. Linear Algebra and Linear Models. New York:Srpinger, 2012. 3° ed.

Bazzoli,C. and Lambert-Lacroix,S. (2018) Classification based on extensions of LS-PLS using logistic regression: application to clinical and multiple genomic data. BMC Bioinformatics 19:314.

Bellanger,M et al., (2018) Are Global Breast Cancer Incidence and Mortality Patterns Related to CountrySpecific Economic Development and Prevention Strategies?. Journal of Global Oncology.

Boldrini, J.L., et al. Algebra Linear. São Paulo Brasil:Harper &Row do Brasil, 1984. 3° edição.

Chiu, A.M. *et al*. (2018) Integrative analysis of the intertumoral heterogeneity of triplenegative breast cancer. Nature, 8:11807.

de Meulder,B. *et al*. (2018) A computational framework for complex disease stratification from multiple large-scale datasets. BMC Systems Biology,12:60.

de Souza, J.E. *et al*. (2014) S-score: a scoring system for the identification and prioritization of predicted cancer genes, PLoS ONE, 9(4): e94147.

Ding,Y. *et al*. (2018) Identification of a gene-expression predictor for diagnosis and personalized stratification of lupus patients. PLoS ONE,13(7): e0198325.

Duan,S. *et al*. (2018) Novel prognostic biomarkers of gastric cancer based on gene expression microarray: COL12A1, GSTA3, FGA and FGG. MOLECULAR MEDICINE REPORTS 18: 3727-3736.

Enshaeifar,S. *et al*. (2019) Machine learning methods for detecting urinary tract infection and analysing daily living activities in people with dementia. PLoS ONE, 14(1): e0209909.

Gálvez,J.M. *et al*. (2018) Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series. PLoS ONE 13(5): e0196836.

Gao,L. (2018) Understanding the Integrated Gene Regulatory Networks for Hepatocellular Carcinoma. Scientific Journal of Gastroenterology & Hepatology – SJGH, Volume 1 - Issue 1.

Gao,L. *et al*. (2018) Identifying noncoding risk variants using diseaserelevant gene regulatory networks. Nature, 9:702.

Golub, G. (1965) Numerical methods for solving linear least squares problems. Numerische Mathematik 7, 206–216

Gong,I.Y. *et al*. (2018) Prediction of early breast cancer patient survival using ensembles of hypoxia signatures. PLoS ONE, 13(9): e0204123.

Harville, D.A. Matrix Algebra: Exercises and Solutions. New York: Springer, 2001. 1st Edition, Kindle Edition

Iglesias-Martinez, L.F., *et al*. (2016) BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research, Nature Scientific reports, 6, 37140.

Joo,J.LL. *et al*. (2018) Determining Relative Dynamic Stability of Cell States Using Boolean Network Model. Nature 8:12077.

Kagaris,D. *et al*. (2018) AUCTSP: an improved biomarker gene pair class predictor. BMC Bioinformatics,19:244.

Li,Z. *et al*. (2018) Efficient feature selection and classification for microarray data. PLoS ONE, 13(8): e0202167.

Liang,Y. *et al*. (2018) CD36 plays a critical role in proliferation, migration and tamoxifen-inhibited growth of ER-positive breast cancer cells. Oncogenesis, 7:98.

Lien,T.G. *et al*. (2018) Integrated analysis of DNA-methylation and gene expression using high-dimensional penalized regression: a cohort study on bone mineral density in postmenopausal women. BMC Medical Genomics, 11:24.

Linnik, I.V. (1961) Method of Least Squares and Principles of the Theory of Observations. Pergamon Press, Russian.

NCBI's Gene Expression Omnibus. Available: http://www. ncbi.nlm.nih.gov/geo/

Majumder,A. *et al*. (2018) Epidermal growth factor receptor-mediated regulation of matrix metalloproteinase-2 and matrix metalloproteinase-9 in MCF-7 breast cancer cells. Molecular and Cellular Biochemistry.

Menard, S. (2010) Logistic Regression. From Introductory to Advanced Concepts and Applications. SAGE Publications, Colorado, USA.

Pereda,E. *et al.* (2018) The blessing of Dimensionality: Feature Selection outperforms functional connectivity-based feature transformation to classify ADHD subjects from EEG patterns of phase synchronization. PLoS ONE, 13(8): e0201660.

Piwowar,M. *et al*. (2018) Regularization and grouping-omics data by GCA method: A transcriptomic case. PLoS ONE, 13(11): e0206608.

Roncato,F. *et al*. (2018) Improvement and extension of anti-EGFR targeting in breast cancer therapy by integration with the Avidin-Nucleic-Acid-Nano-Assemblies.Nature Communications, 9:4070.

Schena, M., *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467–70.

Shao,G. *et al.* (2019) Automatic microarray image segmentation with clustering-based algorithms. PLoS ONE, 14(1): e0210075.

Takagi,K. *et al*. (2018) ARHGAP15 in Human Breast Carcinoma: A Potent Tumor Suppressor Regulated by Androgens. International Journal of Molecular Sciences, 19, 804.

Toledano,D.T. *et al*. (2018) Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on TIMITExperiments on TIMIT. PLoS ONE 13(10): e0205355.

United Kingdom Office for National Statistics 2016. Available:https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/final2016.

UniProt. Available in: < https://www.uniprot.org/uniprot/O94983>. Accessed on: April 13, 2019

Wang,G. *et al*. (2018) TFPI-2 suppresses breast cancer cell proliferation and invasion through regulation of ERK signaling and interaction with actinin-4 and myosin-9. Nature, 8:14402.

Wimmer,I. *et al*. (2018) Systematic evaluation of RNA quality, microarray data reliability and pathway analysis in fresh, fresh frozen and formalin-fxed parafnembedded tissue samples. Nature, 8:6351.

Xu, E.L. *et al*. (2018) Feature selection with interactions in logistic regression models using multivariate synergies for a GWAS application. BMC Genomics, 19(Suppl 4):170.

Yuan,J. *et al*. (2018) Identification of protein kinase inhibitors to reprogram breast cancer cells. Cell Death and Disease, 9:915.

Zhang, W. *et al*. (2018) Classifying tumors by supervised network propagation. Bioinformatics, 34, i484–i493.

Zhao,S. *et al*. (2018) Evaluation of two main RNA-seq approaches for gene quantifcation in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Nature, 8:4781.

Fig. 1. Probability results for samples from patients being classified into breast cancer subtypes TNBC, HER2, Luminal A, or Luminal B, non-tumor tissue samples, or of being TNBC cell lines' samples, as compared to the others, for all systems created from GSE65194 dataset, using all genes and 462 genes. (A) Result for model 1 discriminates TNBC against the other breast cancer subtypes, (B) result for model 2 discriminates Her2 against the other breast cancer subtypes, (C) result for model 3 discriminates Luminal A against the other breast cancer subtypes, (D) result for model 4 discriminates Luminal B against the other breast cancer subtypes, (E) result for model 5 discriminates TNBC cell lines' samples against all breast cancer subtypes and (F) result for model 6 distinguishes between presence or absence of breast cancer.

Fig. 2. Genes associated with features that represent the 20 most positive and 20 most negative α_i^* parameters from one of the 6 systems created from GSE65194 dataset. 1-ADORA2B, 2-CD33, 3-PDE2A,4-KCNJ12, 5-SHC1, 6-KIAA0355, 7-WIPI1, 8-MVK, 9-ADCY9, 10-TOX, 11-KIAA1009, 12-ZNF174, 13-WWOX, 14-EYA, 15-SLC25A17, 16-DET1, 17-NUDT7, 18-TCHP, 19-STARD9, 20-WIZ, 21-PMVK, 22-SKIV2L2, 23-HDAC9, 24-PDGFRL, 25-FLJ35024, 26-GTF2E1, 27-ALDH1L1, 28-VDR-Cdx2, 29-SPAG11B, 30-PTGDS, 31-CAMTA2, 32-SEMA6A, 33-LPHN3, 34-SOCS5, 35-ZNHIT2, 36-MED31, 37-KLHL7, 38-FCRL2, 39-CCDC126, 40-TNFRSF10A.

Fig. 3. Genes (Lilac/Blue) with S-score selected with the classifier and identified inside clusters BGRMI network with transcription factors (red) involved in the proliferation of breast cancer. S-score negative indicative of tumor suppressing or reduced gene activity and positive indicative of oncogene or increased gene activity.Figure adapted from BGRMI network from Iglesias-Martinez, *et al*. 2016.

Fig. 4. Genes (Lilac/Blue) with subtype of breast cancer and S-score selected using the classifier and identified inside clusters BGRMI network with transcription factors (Red/Yellow) involved in various processes in breast cancer. S-score negative indicative of tumor suppressing or reduced gene activity and positive indicative of oncogene or increased gene activity. And the CASP7 gene is apoptosis-related cysteine peptidase. Figure adapted from BGRMI network from Iglesias-Martinez, et al., 2016.

Table 1. Modified Logistic Regression Models created for each datasets.

Datasets	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
GSE65194 178 Samples	55 Samples TNBC against other samples	39 Samples HER2 against other samples	29 Samples LumA against other samples	30 Samples LumB against other samples	14 TNBC cell lines against other samples	11 Samples Non-tumor against other samples
GSE20711 90 Samples	26 Samples HER2 against other samples	27 Samples Basal-like against other samples	13 Samples LumA against other samples	22 Samples LumB against other samples	2 Samples Non-tumor against other samples	
GSE25055 310 Samples	20 Samples HER2 against other samples	99 Samples LumA against other samples	44 Samples LumB against other samples	122 Samples Basal-like against other samples	25 Samples Non-tumor against other samples	

Table 2: Value S-score for the genes associated with features that represent the $\alpha i*$ parameters of each of the systems created from GSE65194 data set. S-score negative indicative of tumor suppressing or reduced gene activity and positive indicative of oncogene or increased gene activity.

EGF induced GRN and HRG induced GRN	
Tumor suppressing Genes / S-score negative / BC subtype	Oncogene Genes / S-score positive / BC subtype
PMVK / 4.11 / (HER2, LumA, LumB) (S-score needed to be negative)	ADORA2B / 1.54 / (TBNC, Lum A, Lum B)
SKIV2L2 / -1.54 / (HER2, LumA, LumB)	KIAA0355 / 0.41 / (TNBC, LumB)
PDGFRL / -2.52 / (HER2, LumA)	MVK / 0.60 / (TNBC, LumA)
GTF2E1 / -0.95 / (TNBC, LumA, LumB)	WWOX / 0.75 / (TNBC, HER2, LumA, LumB)
CAMTA2 / -0.99 / (TNBC, LumA)	WIPI1 / 2.06 / (TNBC, LumA)
SEMA6A / 0.73 / (HER2, LumA) (S-score needed to be negative)	ADCY9 / 2.17 / (HER2, LumA)
MED31 / -2.08 / (LumA, LumB)	EYA4 / 0.77 / (HER, LumA)
CCDC126 / 1.31 / (HER2, LumA, LumB) (S-score needed to be negative)	ZNF174 / 2.46 / (TNBC, LumA)
TNFRSF10A / -3.20 / (HER2, LumB)	SLC25A17 / -1.56 / (HER2, LumA) (S-score needed to be positive)
	TCHP / 0.48 / (HER2, LumA)

Table 3: Value S-score for the genes associated with features that represent the $\alpha i*$ parameters of each of the systems created from GSE65194 data set, and they are located only in one of the networks or in the EGF or in the HRG. S-score negative indicative of tumor suppressing or reduced gene activity and positive indicative of oncogene or increased gene activity.

EGF induced GRN	HRG induced GRN	
Oncogene	Tumor suppressing	Oncogene
Genes / S-score positive / BC subtype	Genes / S-score negative/ BC subtype	Genes / S-score positive / BC subtype
WIZ / 0.51 / (HER 2 LumA)	HDAC9 / 1.95 / (HER2, LumA)(S-score needed to be negative) PTGDS / -1.34 / (TNBC, LumA)	CD33 / 0.89 / (TNBC, HER2, LumA) SHC1 / 3.44 / (HER2, LumB) DET1 / -1.52 / (TNBC, LumA, LumB) (S-score needed to be negative)