Universidade Federal de Minas Gerais
Programa de Doutorado em Bioinformática

Microarray-based breast cancer classification using logistic regression and beyond

A thesis submitted to UFMG
for the attainment of the degree of
Doctor of Philosophy in Bioinformatics

Francielly Morais Rodrigues da Costa

Supervisor: Prof. Dr. Eng. Marcos Augusto dos Santos

Co-supervisor: Dra. Rita Silvério de Magalhães Machado

Belo Horizonte, July 2017

Com um toque para nascer uma folha começa branquiar-se até pingos começarem a colorir os caminhos em sua fibra frágeis, criando formas apessoadas e exultantes. Mas folhas se tornam feias quando borradas e tristes quando pisoteadas. Eu mergulhava em folhas em mim, recolhia os pequenos pedaços, costurava os rasgos por dentro, desamassava as dobras e misturava os restos das cores que sobraram, quando as folhas foram roubadas ao calar do dia e jogadas ao nada com o cantar da noite.

Tentaram destruir minha aquarela para que eu não existisse mais, o que me restava era recomeçar no silêncio das minhas folhas, com o único pensamento: quem sabe colorir, transformar restos de pingos em litros de cores para se reconstruir, com a maior sutiliza de uma águia, que se recolhe em um silêncio ecoante quando velha e ao despencar de suas asas. Em seu recanto ela arranca o bico vetusto, e com o expctar do seu crescimento, seu corpo é regenerado, suas garras são afiadas, e assim eu fiz…

No silêncio ecoante arranquei todas as folhas borradas, rasgadas e pisoteadas, regenerei minha paleta de cores e afiei meus pincéis. Recomecei a me colorir, e aos poucos os meus olhos, a minha voz, a minha pele voltaram a respirar em cores tão vívidas quanto o primeiro respiro de um recém-nascido. As folhas em mim foram tomando formas com paciência e resiliência de uma águia, até o dia em que pude voar novamente…

Hoje mergulho em folhas com fibras de aço que podem cortar quem tentar roubá-las, rasgá-las, dobrá-las ou pisoteá-las. Novos capítulos de mim virão com maior quantidade de cores e tentarão borrá-las, e por quantas vezes tentarem, será a quantidade de vezes que terão que suportar eu aquarelar esses borrões.

Francielly Morais Rodrigues da Costa

Belo Horizonte, May 2017

Acknowledgments

Primeiramente, agradeço a Deus e a minha Nossa Senhora de Fátima por terem me dado saúde, paciência e sabedoria para cumprir essa jornada de 4 anos. Obrigada Senhor, o maior mestre doutor programador que criou o principal programa, os organismos!

Durante meus 34 anos os livros mais preciosos que pude ter a honra de aprender foram meus pais, Délia e Adelson, meus irmãos Davidson, meu segundo pai, e Itamar Jr que se tornou meu amigo irmão quando eu ainda estava no útero da minha mãe. Nessas páginas aprendi a ser uma pessoa respeitosa, ética, honesta, justa, verdadeira, corajosa e o mais importante, que não é vencer, e sim, nunca desistir. Obrigada pela vida, por terem me proporcionado um lar amoroso e feliz, por me apoiarem nos estudos e por isso consegui chegar neste dia tão almejado na minha vida. Como diria Isaac Newton: "Se vi mais longe, foi por estar de pé sobre ombros de gigantes." Ombros esses que pertencem a vocês.

Com o passar dos anos muitas páginas foram inseridas nesse livros preciosos. Relembrando o meu passado, ao assistir o filme "Ebola" em 1995 quando eu tinha doze anos, decide que seria uma cientista. Com o passar dos anos me espelhei em meu irmão Davidson, e me apaixonei por computação. Foi quando descobri que eu poderia ser uma cientista usando um computador. Resolvi seguir os passos do meu irmão e cursei ciência da computação na faculdade UNI-BH, onde pude conhecer o professor Bráulio. E foi através dele que conheci a pós-graduação multidisciplinar, pois o mesmo estava terminando o doutorado em bioinformática na UFMG, e sendo orientado pelo professor Marcos Augusto. Durante os 5 anos que estive na UNI-BH conheci os professores, Ana Paula Ladeira que me ensinou inteligência artificial e me orientou durante o trabalho de final de curso, Guilherme que me ensinou algoritmos e programação, Evandrino Barros que me ensinou Java, Cayley Guimaraes que escrevemos um artigo juntos com outros alunos sobre usabilidade no dia a dia, e muitos outros que me ensinaram todo conhecimento que pude utilizar durante o mestrado e doutorado. Após minha formatura, me espelhando no meu primo e pesquisador Adilson, comecei a caminhada para me tornar uma cientista. Cursei mestrado em modelagem matemática e computacional no CEFET-MG, que abriram meus caminhos para pesquisas computacionais. Mas meu objetivo ainda era chegar ao doutorado em bioinformática, e

consegui, fui aceita no curso em 2013. Foi um dos melhores momentos da minha vida! Agradeço aos professores Bráulio e Adilson pelas cartas de recomendação.

Alguns capítulos são necessários para o crescimento e entendimento de nós mesmos. Durante os 4 anos que estive na UFMG me deparei com muitos momentos decisivos para a minha jornada. Às vezes o caminho fica ingrime e árduo, o chão se torna grandes dunas de um deserto, e nessas horas que Jesus nos carrega no colo e nos guia. À sua frente vão anjos para nos proteger contra os vendavais e iluminar o caminho, e assim foi comigo, e esses eu chamo de anjos gladiadores, Sheila, Natália e Fernanda, amigas que levarei por toda vida. Vocês seguraram minha mão e oraram comigo me dando o melhor conforto e melhor remédio, as palavras de Jesus, que são fundamentadas no respeito, no amor e no perdão para com nossos irmãos de mundo.

Outras páginas se juntam para somar no caminho abençoado, é quando Jesus vê que já podemos voltar a caminhar, Ele nos punha no chão e entrega nossas mãos para anjos que nos devolve a vontade de sorrir, de levantar cedo para estudar e virar noites com maior satisfação, e esses chamo de anjos mestres, Rita que se tornou minha irmã de coração e Marcos Augusto que se tornou um amigo precioso, que me receberam de braços abertos com um carinho que não pode ser medido. Vocês acrescentaram mais aos valores que aprendi com minha família, respeito, ética, honestidade e justiça, e por isso me senti tão acolhida. Vocês foram os melhores orientadores que eu podia ter. Cada palavra que eu anotava, cada reunião, cada material que eu precisei estudar para chegar até o dia de hoje, foram os maiores tesouros que vocês poderiam presentear o meu crescimento.

Mas alguns ensinamentos para o crescimento do ser advém de pessoas que por algum motivo estiveram na hora certa e no momento certo, como meus tios Dorino, Renê, Dália e Inês, meus amigos Edson, Jessica Bley, Gil, minha cunhada Fátima, Manoel e Maria de Lourdes pais da Rita e seu esposo Bruno, em especial Alberto (que me guiou e me ajudou em momentos precisos), mostrando que todos somos iguais e com palavras e atos enriqueceram minhas forças para continuar a caminhada no doutorado.

E por fim, gostaria de agradecer aos professores Vasco, Miguel e Glória que com suas inteligências aliadas as suas sensibilidades trabalham firmemente para desenvolvimento do

curso. Meus agradecimentos a estrutura oferecida pelo pela UFMG e ao apoio financeiro da CAPES.

Eu dedico este trabalho aos meus pais Adelson e Délia, aos meus irmãos Davidson, Itamar Jr e Rita, aos meus amigos que a vida me presenteou Marcos Augusto, Sheila, Natália, Fernanda e à memória dos meus avós, Wanda, Yeder, Antônio e Manuel e à memória do meu amigo Edson.

A todos citados aqui muito obrigada por colorirem as folhas desses capítulos da minha vida. Vocês todos são livros preciosos para minha existências que eu pude ter o honra de ler. Novos capítulos de mim virão com maior quantidade de cores e tentarão borrá-las, e por quantas vezes tentarem, será a quantidade de vezes que terão que suportar eu aquarelar esses borrões.

Francielly Morais Rodrigues da Costa

Table of Contents

Abstract

Cancer is a major global health problem with millions of new cancer cases emerging each year and millions of cancer-related deaths occurring per year. Breast cancer ranks as the first to affect women with the most disease-related cases being reported in developed countries but with the majority of deaths occurring in developing countries.

In this PhD project, a novel and innovative genome-wide model was developed to classify breast cancer samples. This new logistic regression-based model that we propose uses a stabilizing term in that allows the assignment of values to parameters α, a distinguishing feature among other methods which circumvents the need for variable pruning. Applying this methodology to classify samples found in NCBI's Gene Expression Omnibus (GEO) GSE65194, GSE20711 and GSE25055 data sets we obtained a minimum performance of 80% (both sensitivity and specificity). Genes associated with parameters αi^* holding extreme values were searched in the literature for a relation with breast cancer. Some hold no evidence in the literature of association with breast cancer but based on the rational followed during this PhD project, they were flagged to be investigated as yet-undiscovered candidates with potential diagnostic and/or therapeutic utilities in breast cancer.

We examined the pattern and feature of a GRNs composed of TFs in MCF-7 breast cancer cell lines to provide valuable information relating breast cancer with some particular genes whose αi^* associated parameter values reveal extreme positive values and as such identify breast cancer prediction genes. The topological analysis of these networks, the direct correlation observed between some of the flagged genes with relevant TFs in the context of breast cancer and using the S-score system that has been used by many to confirm the tumour suppressor/oncogenic profile of genes in specific cancer types, allowed us to reveal some potential breast cancer prediction genes that are suggested to be be prioritized for further breast cancer clinical studies. These results establish the proof of concept for the proposed novel and innovative model to classify breast cancer samples that we propose here.

A large number of oncolytic viruses have been proposed for cancer therapy, which includes Seneca Valley Virus. SEMA6A is a gene flagged by application of the new logistic

regression model detailed in this PhD thesis, which produces a cell receptor. Keeping in mind that SVV-001 cancer cell tropism might be governed by binding to specific receptors on the surface of cancer cells, we hypothesize that this specific protein could be the door for Seneca Valley Virus V001 entrance in breast cancer cells. The results obtained make probable the creation of the complex Semaphorin-6A – V001, indicating the oncolytic virus Seneca Valley Virus as a new therapeutic option to be considered and further studied for breast cancer treatment.

Thesis Outline

This PhD thesis consists of three chapters, a section with final conclusions, a section with references and two annexes: the first with the supplementary material of the top publication produced during the course of this PhD Project and submitted to publication to the international peer-reviewed scientific journal BMC Bioinformatics (BioMed Central) and second with other publications produced during a side Project carried out during these 4 years of intensive work and schematized in Table 1.

Table 1: PhD thesis outline

Chapter 1 ***Using a new logistic regression-based model for breast cancer classification***
Potential breast cancer prediction genes
Chapter 3 ***Exploring breast cancer potential therapeutics***
Conclusions and Final Reflections
References
Annex 1 ***Using a new logistic regression-based model for breast cancer classification – supplementary material***
Annex 2 ***Side Project – publications***

Objectives of the Project

Main Objective:

The main objective of this PhD project was to develop a genome-wide new regression-based model for breast cancer classification without reducing the number of features and with good classification performance.

The specific objectives were:

1. To point out an ingenious way to compute the logit function*;*
2. Classify GSE65194, GSE20711 and GSE25055 microarray samples with all *features* included;
3. Flag new potential breast cancer biomarkers;
4. Apply the new model here proposed to classify breast cancer samples, but using only the genes whose α_i* associated parameters that are topologically located in the extremes of the α plots;
5. Explore GRNs to establish the proof of concept for the proposed novel and innovative model to classify breast cancer samples that we propose here;
6. Propose a new Oncolytic Virotherapy for breast cancer using Seneca Valley Virus V001
7. Explore the hypothesis established in the former point using the in silico method molecular docking.

Index of Tables

Index of Figures

List of Abbreviations

BGRMI	Bayesian Gene Regulation Model Inference
BMA	Bayesian Model Averaging
ChIP-seq	Chromatin IP Sequencing
DE	Desolvation
DNA	Deoxyribonucleic acid
EGF	Epidermal Growth Factors
ELEC	Electrostatics
ER	Estrogen receptors
FFT	Fast Fourier Transform
GEO	Gene Expression Omnibus
GRN	Gene regulatory networks
HER2	Human Epidermal growth factor Receptor 2
HR	Hormone receptors
HRG	Heregulin
IARC	International Agency for Research on Cancer
INCA	National Cancer Institute José Alencar Gomes da Silva
LumA	Luminal A
LumB	Luminal B
MCF-7	Michigan Cancer Foundation-7
NCBI	National Centre for Biotechnology Information
PDB	Protein Data Bank
PSC	Pairwise shape complementarity function
PR	Progesterone receptors
RNA-seq	RNA Sequencing
SVD	Singular Value Decomposition
SVV-001	Seneca Valley Virus-001
TNBC	Triple Negative Breast Cancer
WHO	World Health Organization

CHAPTER 1

Using a new logistic regression-based model for breast cancer classification

"If you want to have good ideas you must have many ideas. Most of them will be wrong, and what you have to learn is which ones to throw away."

Linus C. Pauling
chemist

1. Introduction

Cancer is a major global health problem with the World Health Organization (WHO) projecting that by 2035 the world could see 24 million new cancer cases and 14.5 million cancer-related deaths per year (*INCA*, 2016). Based on based in the World Cancer Report 2014 from the International Agency for Research on Cancer (IARC) that is part of the WHO, the National Cancer Institute José Alencar Gomes da Silva (*INCA*) has estimated in the beginning of the year 2016 that for that year and the subsequent (2016 and 2017), about 596.070 new cancer-related cases were expected to be counted in Brazil (Figure 1) (*INCA*, 2016). Breast cancer ranks as the first to affect women with the most disease-related cases being reported in developed countries but with the majority of deaths occurring in developing countries (Siegel, Miller and Jemal, 2015; UK, 2017). In wealthier countries, death rates have become stable since the 90s mainly due to early detection and increased efficacy of the applied treatments, which still isn't a reality in poor countries (Hu et al., 2016).

Genetic changes that promote cancer development occur mainly in genes that regulate cell growth in normal cells: proto-oncogenes and tumour-suppressor genes. The former promote cell growth and when they are mutated or when many copied if it exists, they stay

permanently activated making the cell grow without control which can lead to cancer. The later are genes that slowdown cell division, DNA repair or programmed cell death (apoptosis) and when mutated, cells grow without control, which can also lead to carcinoma. A meaningful difference between oncogene and tumour-suppressor genes is that the former causes cancer when it's activated and the later does the same but when deactivated (American Cancer Society, 2015; Schatten, 2013; Rivenbark, O'Connor and Coleman, 2013).

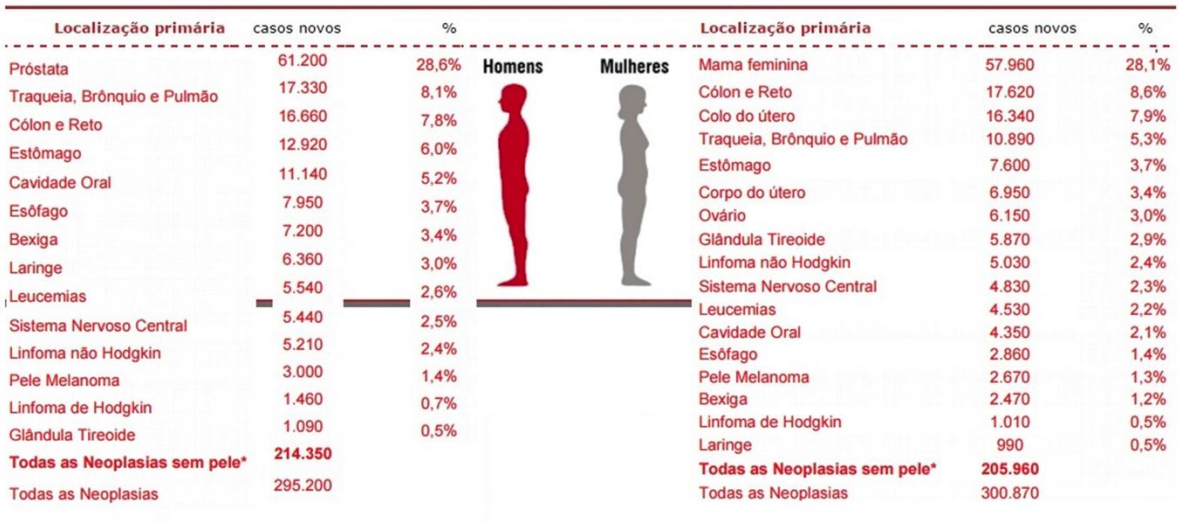

Figure 1: Estimated cancer-related cases in Brazil for years 2016 and 2017. Adapted from (*INCA*, 2016).

Cancer holds clinical, morphological and biological heterogeneity that has implications for cancer therapeutics. The biological characteristics of a tumour are determined by the patterns of changes that cells experience during the disease. Based on similar patterns, morphological and immunophenotypically tumours are grouped in genetically homogeneous tumours (Hu et al., 2016; da Cunha et.al., 2013; Zhao et.al., 2009; Polyak, 2011; Rivenbark, O'Connor and Coleman, 2013; Brooks, Burness and Wicha, 2015). Malignat breast tumours hold different shapes and structures, with ductal carcinoma, which starts in a milk duct of the breast (the passages that drain milk from the lobules to the nipple), as the most common. These show a very slow growth rate and may or may not progress to invasive breast cancer. Invasive ductal carcinoma is the most common type of breast cancer; it breaks through the wall of the duct, and grows into the fatty tissue of the breast. The other common types is lobular carcinoma that start in the milk-producing glands (lobules) and in many cases it spreads to other parts of the body. Less common types of breast cancer are inflammatory breast cancer, Paget disease of the nipple, Phyllodes tumour and angiosarcoma. Inflammatory breast cancer is

an uncommon type of invasive breast cancer, very aggressive as it progresses rapidly in a matter of weeks or months, in which cancer cells block lymph vessels in the skin of the breast. The "inflammatory" designation came from the swollen and red, or inflamed appearance of the breast. Paget disease of the nipple starts in the breast ducts and it spreads to the skin of the nipple and, usually, the darker circle of skin around it (areola). Phyllodes tumour are rare breast tumours that develop in the connective tissue (stroma) of the breast, are usually benign but some are malignant. Angiosarcomas of the breast are very rare cancers that start in the cells that make up the walls of blood vessels or lymphatic vessels (Malhotra et.al., 2010; Badve et.al., 2011; American Cancer Society, 2015; Schatten, 2013).

There are four main molecular subtypes of breast cancer that are based on the genes a cancer expresses: Human Epidermal growth factor Receptor 2 (HER2), Luminal A (LumA), Luminal B (LumB), Triple-Negative Breast Cancer (TNBC). LumA tumours occur mainly in developed countries, represent about 74% of all breast carcinomas and are characterized for expressing estrogen receptors (ER+) and/or progesterone receptors (PR+), but not HER2 (Hu et al., 2016). This subtypes shows a slower growth and less aggressive profile than LumB, which has a more proliferative capacity and as such a higher hostility and less favourable prognostics. It is ER+ and/or PR+, but either HER2+ or HER2-negative. Both these subtypes are sensitive to anti-hormonal therapies with better prognostics than HER2 and TNBC subtypes (American Cancer Society, 2015; Schatten, 2013; Park et al., 2016). Her2 tumour subtypes are much more aggressive than the Lums, don't express hormone receptors (HR-negative) but on the other hand highly express HER2 gene. Traditional treatments for this tumour subtype usually turn off signal channels (American Cancer Society, 2015; Schatten, 2013). TNBC subtype is both HR-negative and lack of HER2 overexpression. This tumour subtype is known being very aggressive, bad prognosis, with fewer treatment options. TNBC is often used as a surrogate for identifying the aggressive basal-like breast cancer subtype, as both are defined by negative immunohistochemical staining for estrogen receptor (ER) and progesterone receptor (PR) and lack of Her2 overexpression. Both basal-like and triple negative breast cancers are associated with poor clinical outcomes and although they share many similarities but they are not synonymous (American Cancer Society, 2015; Schatten, 2013).

Developing countries have limited healthcare resources and use different strategies to diagnose breast cancer that many times aren't accessible to all the population, with

approximately 60% of deaths due to breast cancer occurring in developing countries. In contrast, in developed countries there has been some debate about breast cancer treatment being overrated and women being over-diagnosed. It has been reported that screening healthy women with mammography to find breast cancers before they could be felt as a lump in the breast did not lead to lower death rates for average-risk women in their 40s and 50s. Cancer organizations continue to spread that "early detection saves lives" but their mantra has not changed after being proved that such claims are inflated and imbalanced. Many times, breast cancers found through mammography screening lead to unnecessary surgery, radiation and chemotherapy for non-life threatening cancers. This is a very controversial discussion, but the bottom line here is that we must ensure that we all have access to unbiased information, free from conflict of interest and without the heavy thumb of vested interests tipping the balance (Hu et al., 2016).

2. SVD for Breast Cancer Classification

Microarray is the technology of choice since the 90s for global analysis of gene expression that allows simultaneous investigation of hundreds or thousands of genes in a sample (Brentani et al., 2005). Although this genomic tool is not new (Schena et al.,1995), it has matured in the last fifteen years, with the emergence of high quality arrays due to standardized hybridization protocols, accurate scanning technologies, and robust computational methods (Powell et al., 2015). Still this technology has several limitations and a new powerful technology named RNA-seq is predicted to replace microarrays for transcriptome profiling by avoiding some technical issues in microarray studies related to probe performance such as limited detection range of individual probes, cross-hybridization and non-specific hybridization (Zhao et al., 2014). However, RNA-seq is still facing some challenges that are currently limiting its potential utilization: higher cost that makes its use almost impractical for large studies, high data storage requirements as data produced by an RNA-seq experiment is orders of magnitude greater than microarrays data, and the analysis is quite complex for example, a significant number of sequence reads in RNA-seq are multireads (reads that have high-scoring alignments to multiple positions in a reference genome or transcript set) and the way to assign multireads to genes is still a problem in reads mapping. Therefore microarrays are still the

more common choice of researchers gene expression analysis (Pont et al., 2016; Schulten et al.,2016).

Microarray-based gene expression profiling is used to classify a multitude of tumour types (Kumar, Sharma, and Tiwari 2012; Weigelt, Baehner and Reis-Filho, 2010), that, as explained before, will determine which treatment methods will most likely yield beneficial results for particular cancer patients (Ringnér et.al., 2011; Brentani et.al., 2005; Barnett et al., 2014) and to predict cancer-specific biomarkers in large patient cohorts (Han and Li, 2011). Microarray studies are characterized by a low sample number and a large feature (gene/ probes/ attributes) number, which adversely affect similarity measurements and classification performance, since many of these features are irrelevant to specific traits of interest, and therefore contain no discrimination power. If we would project our samples in the features' space, we would have a thousand-dimensional space and we could talk about the `curse of dimensionality', coined by Richard E. Bellman (Bellman, 2015) and that in general terms is the widely observed phenomenon that data analysis techniques frequently perform poorly as the dimensionality of the analysed 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 tumour type. Many algorithms have been developed to deal with the high-dimensionality problem in microarray studies (Wilcox, 1961; Fort and Lambert-Lacroix, 2005; Giancarlo, Bosco and Pinello, 2010; Zhao et.al., 2013). Some use classical classification tools, but feature selection must occur *a priori* (McKinney et.al.,2007; Saeys, Inza and Larranaga, 2007; Beniwal and Arora, 2012). Dimensionality reduction is another approach taken using linear algebra methods (Zhao et.al., 2013; Kossenkov and Ochs, 2010; Thomas et.al., 2014; Tomfohr, Lu and Thomas Kepler, 2005).

We attempted to use Singular Value Decomposition (SVD) to predict breast cancer in samples from a GSE65194, GSE20711 and GSE25055 data set downloaded from NCBI Gene Expression Omnibus (GEO).

2.1. Method

2.1.1. 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 test the applicability of the proposed methodology. Data sets with the identifiers GSE65194, GSE20711 and GSE25055 were downloaded from GEO and the former two acquired using Affymetrix Human Genome U133 Plus 2.0 arrays and the last acquired using Human Genome HG U133A Affymetrix arrays. GSE65194 data set consists of 178 measurements of gene expression profilings from 153 breast cancer samples, grouped into 4 major subtypes (55 TNBC; 39 HER2; 29 LumA and 30 LumB), 11 non-tumour breast tissue samples obtained from mammoplasty and 14 TNBC cell lines. GSE20711 data set consists of measurements of gene expression profilings from 90 breast cancer samples grouped into 4 major subtypes (27 Basal-like; 26 HER2; 13 LumA and 22 LumB) and 2 non-tumour breast tissue samples. GSE25055 data set consists of measurements of gene expression profilings from 310 samples grouped into 4 major subtypes (122 Basal-like; 20 HER2; 99 LumA and 44 LumB) and 25 non-tumour breast tissue samples.

2.1.2. SVD

The mathematical technique of linear algebra SVD can be applied to a term-document matrix to find relevant documents from query words using a search engine in the context of informational retrieval, enabling the analysis of latent (i.e. hidden) semantics in a document containing words (Deerwester, et al., 1990). As previously mentioned, a typical term-document matrix is very large and quite often very sparce and SVD acts as a method to reduce the dimensionality of this original space and construct a subspace without great loss of descriptiveness. With SVD the less frequently co-occurring features occurring in a given document are excluded from the subspace and as such the “noise” of the original matrix is reduced. This perspective has pushed us to apply SVD in breast cancer classification of samples from GSE65194, GSE20711 and GSE25055 data sets.

2.2. Results and Discussion

Applying SVD to the matrices built using the samples in the three data sets considered produced quite mixed results. For GSE65194, and after reducing to 4 the matrix dimensionality (the same reduction was applied to all data sets), it is clearly observed a topological separation between non-tumour samples and breast cancer samples, as illustrated in Figure 2. For all other breast samples there was no clear separation between breast cancer subtypes. Likewise, for both GSE20711 and GSE25055 data sets there was no clear separation between breast cancer subtypes and not even between breast samples and non-tumour samples, as illustrated in Figures 3 and 4.

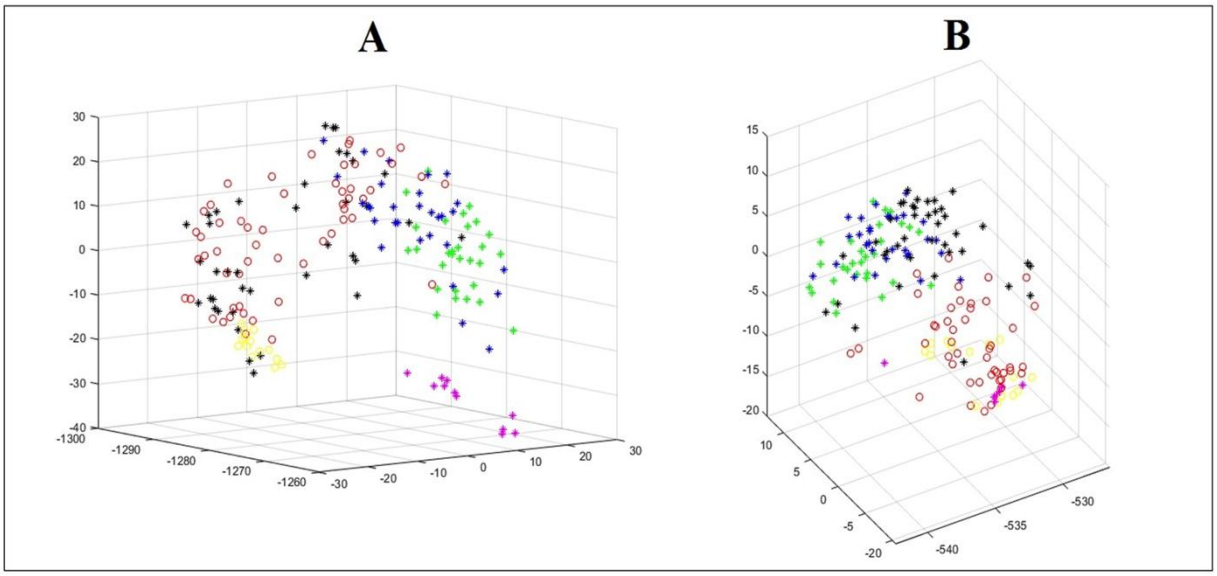

Figure 2: Visualization of breast cancer samples from GSE65194 data set, after application of SVD technique for dimensionality reduction with (A) 462 probes and (B) all genes. Red represents TNBC, black HER2, green LumA, blue LumB, yellow TNBC cell lines and purple normal breast samples. Vectors were projected in space IR^3 using the method described by Marcolino, Couto and Santos (2010).

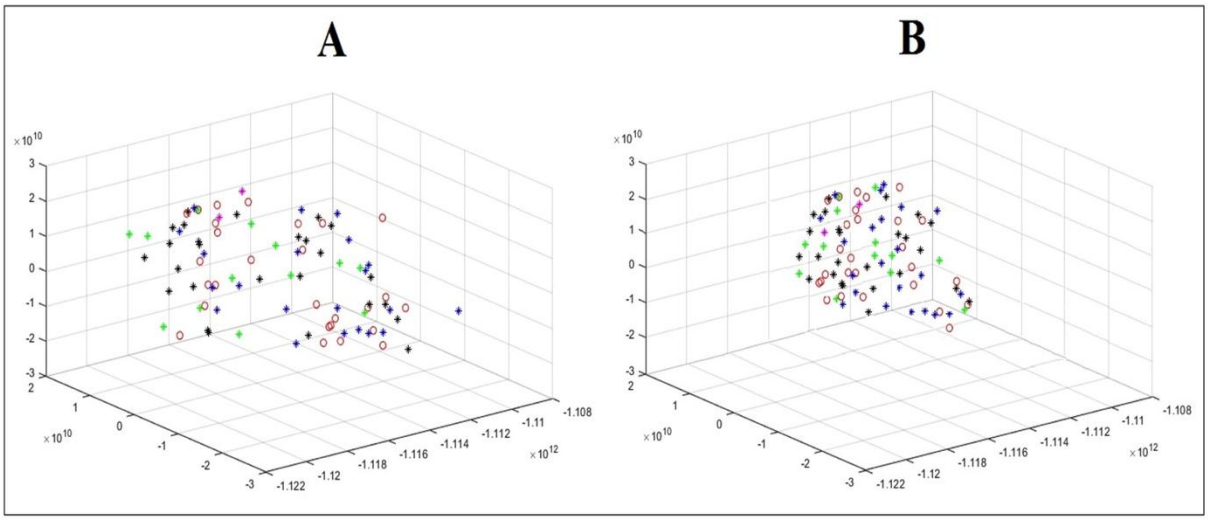

Figure 3: Visualization of breast cancer samples from GSE20711 data set, after application of SVD technique for dimensionality reduction with (A) 320 probes and (B) all genes. Red represents Basal-like, black HER2, green LumA, blue LumB, and purple normal breast samples. Vectors were projected in space IR^3 using the method described by Marcolino, Couto and Santos (2010).

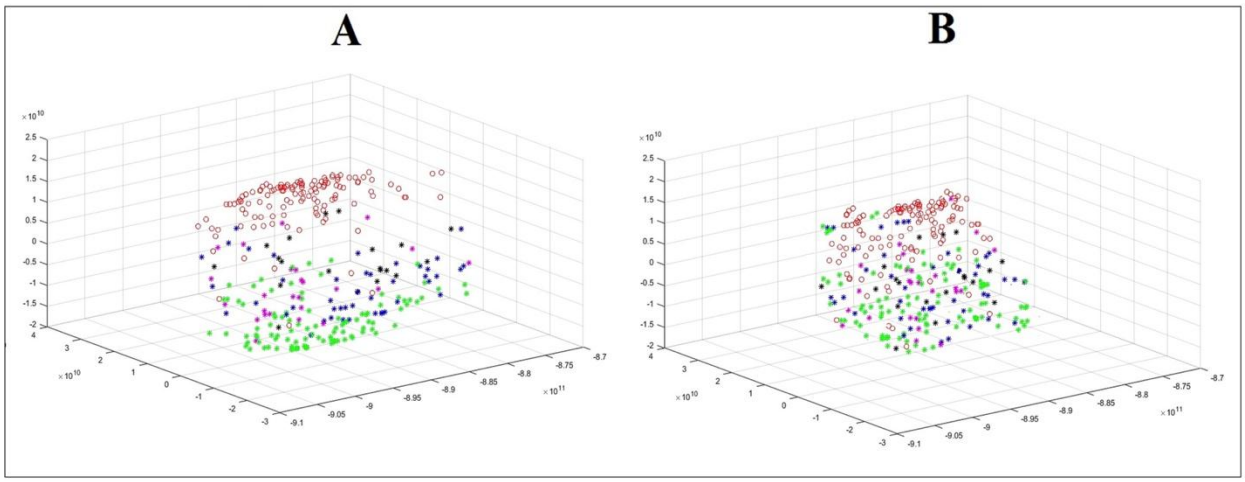

Figure 4: Visualization of breast cancer samples from GSE25055 data set, after application of SVD technique for dimensionality reduction with (A) 319 probes and (B) all genes. Red represents Basal-like, black HER2, green LumA, blue LumB, and purple normal breast samples. Vectors were projected in space IR^3 using the method described by Marcolino, Couto and Santos (2010).

Application of this methodology for breast cancer classification didn't produce the results we were expecting. We suspect that the poor classification performance is due to the enormous resemblance between several probes (some have redundant information). When the cosine of the angle formed between each vector representing an individual and any other vector in the data set is computed, the minimum value observed is always higher than 0.8 (Figure 5), pointing to very similar vectors positioned almost alongside (Marcolino, Couto and Santos, 2010; Xu et al., 2011 and Wu, et al., 2015).

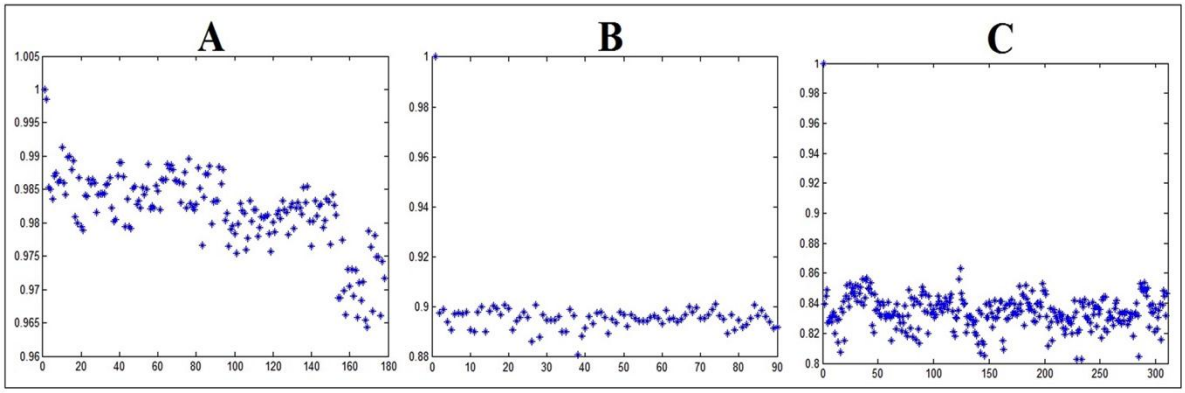

Figure 5: Minimum cosine values for each sample against all samples for each (A) GSE65194, (B) GSE20711 and (C) GSE25055 data set.

3. The New Logistic Regression-based Model

Here we propose a new logistic regression-based model that we developed to classify breast cancer tumour samples based on microarray expression data with all features included and no need for reduction of microarray data matrix. This model uses the logit function for classification of breast cancer subtypes with some particularities that will be detailed further ahead in a scientific paper embedded in this thesis and submitted to publication to the international peer-reviewed scientific journal BMC Bioinformatics (BioMed Central). Logit-based methods have been successfully applied to cancer classification but always involving gene selection for classification to be possible. Certain variable selection schemes for the logistic regression models that exist before the one that we propose here are not suitable for microarray-like problems having large numbers of variables and small sample sizes (James et.al., 2013; Hyeoun-Ae, 2013). The model that we propose here circumvents the need for variable pruning by aggregating the quadratic term to the solution of a system of equations to determine the value of αi^* associated parameters. These parameters are related with the expression of a gene. The variables that are associated with gene expression and do not have a discriminatory 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. Though the text there is the symbols αi^* and αi are presented: the former refers to specific values obtained after application of the new logistic-based regression model proposed and the later is used before model application.

The key point for the development of this model is a stabilizing term that allows the assignment of values to parameters αi^*, allowing the system to have a unique solution. The parameters with some of the extreme values are associated with known breast cancer related genes and other topologically related genes with no reference in the literature as being related with breast cancer. These are fagged here to be investigated as yet-undiscovered candidates with potential diagnostic and/or therapeutic utilities in breast cancer, which is explored in Chapter 2 of this PhD thesis.

3.1. Using a new logistic regression-based model for breast cancer classification submitted paper.

RESEARCH

Using a new logistic regression-based model for breast cancer classification

Francielly Morais-Rodrigues[1,*◕], Rita Silvério-Machado [1◕], J Miguel Ortega[2], Frederico F Campos[3], Sandro J de Souza[4], Marcos A dos Santos[3]

1 Institute of Biological Sciences, Federal University of Minas Gerais, Brazil. Av. Antônio Carlos, 6627, Belo Horizonte, MG 31270-901, Brazil.
2 Department of Biochemistry and Immunology, Federal University of Minas Gerais, Brazil
3 Department of Computer Science, Federal University of Minas Gerais, Brazil
4 Brain Institute, Federal University of Rio Grande do Norte, Brazil

◕These authors contributed equally to this work.
1* corresponding author franrodriguesdacosta@gmail.com; 1 rita_silverio@hotmail.com; 2 miguel@icb.ufmg.br; 3 ffcampos@dcc.ufmg.br; 4 sandro@neuro.ufrn.br; 3 marcos@dcc.ufmg.br

Full list of author information is available at the end of the article

Abstract

Background: More and more statistics and linear algebra methods are used to address questions that emerge in microarray literature. Microarray technology is a long-used tool for global analysis of gene expression that allows simultaneous investigation of hundreds or thousands of genes in a sample, and is characterized by a low sample size and a large feature (gene) number that adversely affect similarity measurements and classification performance. To avoid the problem of the 'curse of dimensionality' many authors have performed feature selection or reduced the size of data matrix. 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.

Results: This methodology allowed the correct classification of breast cancer samples from GEO data series GSE65194, GSE20711, and GSE25055 data sets that contain microarray data of breast cancer samples, with a minimum performance of 80% (sensitivity and specificity) and exploring all possible combinations of data that included breast cancer subtypes.

Conclusions: This new model allows the assignment of values to parameters $\alpha_i{}^*$ that are associated with the expression of a gene. Scrutinizing these parameters $\alpha_i{}^*$ unveiled that some of the topologically extreme parameters are associated with known biomarker in breast cancer and flagged a set of other genes with no identified relation to breast cancer, to be investigated as as-yet-undiscovered biomarker candidates with potential diagnostic and therapeutic utilities in breast cancer.

Keywords: Breast cancer classification, new logistic regression-based model, gene expression, microarrays.

Background

In the past few years, there has been a growing interest in the application of methods of linear algebra and statistics in data mining, social networks, machine learning, bioinformatics, information retrieval, plus others [1–5]. Among these methods, logistic regression approach draw some special 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 [6–11].

Microarray is the technology of choice since the 90s for global analysis of gene expression that allows simultaneous investigation of hundreds or thousands of genes in a sample [12]. Although this genomic tool is not new [13], it has matured in the last fifteen years, with the emergence of high quality arrays due to standardized hybridization protocols, accurate scanning technologies, and robust computational methods [14]. Still this technology has several limitations and a new powerful technology named RNA sequencing (RNA-seq) is predicted to replace microarrays for transcriptome profiling by avoiding some technical issues in microarray studies related to probe performance such as limited detection range of individual probes, cross-hybridization and non-specific hybridization [15]. However, RNA-seq is still facing some challenges that are currently limiting its potential utilization: higher cost that makes its use almost impractical for large studies, high data-storage requirements as data produced by an RNA-seq experiment is orders of magnitude greater than microarrays data, and the analysis is quite complex for example, a significant number of sequence reads in RNA-seq are multireads (reads that have high-scoring alignments to multiple positions in a reference genome or transcript set) and the way to assign multireads to genes is still a problem in reads mapping. Therefore microarrays are still the more common choice of researchers gene expression analysis [16, 17]. Microarray-based gene expression profiling is used to classify a multitude of tumor types [18, 19], to determine which treatment methods will most likely yield beneficial results for particular cancer patients [20] and to predict cancer-specific biomarkers in large patient cohorts [21]. Microarray studies are characterized by a low sample number and a large feature (gene) number, which adversely affect similarity measurements and classification performance, since many of these features are irrelevant to specific traits of interest, and therefore contain no discrimination power. If we would project our samples in the features' space, we would have a thousand-dimensional space and we could talk about the 'curse of dimensionality', coined by Richard E. Bellman [22] and that in general terms is the widely observed phenomenon that data analysis techniques frequently perform poorly as the dimensionality of the analysed 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 [7, 23].

Cancer incidence and mortality statistics reported by the American Cancer

Society [24, 25] and by the United Kingdom Office for National Statistics [26] indicate breast cancer as one of the four most common cancer types, along with lung, colorectal, and prostate. Breast cancer alone is expected to account for 30% of all new cancer diagnoses in women in 2017, being the most frequently diagnosed cancer in women [25]. Breast cancer is a very heterogeneous disease [27, 28] with significant variability between patients. Breast tumors can be grouped in four molecular subtypes, which have major implications for determining treatment (Luminal A, Luminal B, Triple negative/Basal-like, HER2 type) [29, 30].

In this paper we analyse 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 parameters $\alpha_i{}^*$ that are associated with the expression of a certain gene. Scrutinizing these parameters $\alpha_i{}^*$ 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.

Methods

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) [31] 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. GSE65194 dataset consists of 178 measurements of gene expression profilings from 153 breast cancer samples, grouped into 4 major subtypes (55 Triple Negative Breast Cancer - TNBC; 39 Human Epidermal Growth Factor Receptor 2- HER2; 29 Luminal A and 30 Luminal B), 11 non-tumor breast tissue samples obtained from mammoplasty and 14 TNBC cell lines. GSE20711 dataset consists of measurements of gene expression profilings from 90 breast cancer samples grouped into 4 major subtypes (27 Basal-like; 26 HER2; 13 Luminal A and 22 Luminal B) and 2 non-tumor breast tissue samples. GSE25055 dataset consists of measurements of gene expression profilings from 310 samples grouped into 4 major subtypes (122 Basal-like; 20 HER2; 99 Luminal A and 44 Luminal B) and 25 non-tumor breast tissue samples.

For GSE65194 dataset, a set of six systems were created: system 1 distinguishes between TNBC and the other breast cancer subtypes, TNBC cell line samples and non-tumor breast tissue samples; system 2 discriminates HER2 against other breast cancer subtypes, TNBC cell line samples and non-tumor breast tissue samples; system 3 discriminates Luminal A against other breast cancer subtypes, TNBC cell line samples and non-tumor breast tissue samples; system 4 distinguishes between Luminal B and other breast cancer subtypes, TNBC cell line samples and non-tumor breast tissue samples; system 5 distinguishes between TNBC cell line samples and all breast cancer subtypes and non-tumor breast tissue samples; and system 6 distinguishes between presence or absence of breast cancer. For

GSE20711 dataset, a set of five systems were created: system 1 discriminates HER2 against other breast cancer subtypes and non-tumor breast tissue samples; system 2 discriminates Basal-like against other breast cancer subtypes and non-tumor breast tissue samples; system 3 distinguishes between Luminal A and other breast cancer subtypes and non-tumor breast tissue samples; system 4 distinguishes between Luminal B and other breast cancer subtypes and non-tumor breast tissue samples; system 5 distinguishes between presence or absence of breast cancer. Finally for GSE25055 dataset another set of five systems were created: system 1 discriminates HER2 against other breast cancer subtypes and non-tumor breast tissue samples; system 2 distinguishes between Luminal A and other breast cancer subtypes and non-tumor breast tissue samples; system 3 distinguishes between Luminal B and other breast cancer subtypes and non-tumor breast tissue samples; 4 discriminates Basal-like against other breast cancer subtypes and non-tumor breast tissue samples; and system 5 distinguishes between presence or absence of cancer in the breast tissue samples.

Modified Logistic Regression Model

The data obtained from microarray experiments is represented by matrix $A = \{x_{ij}\}$ with m rows and n columns, with rows representing patients and columns representing genes. The value of each position x_{ij} 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 $x = \{x_i, x_2, \ldots, x_n\}$ 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:

$$P_i(x) = g_i(x)/(1 + g_i(x)), \qquad (1)$$

where

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

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 one. Viewing $P_i(x)$ as the probability, the odds $C_i(x)$ are given by:

$$C_i(x) = P_i(x)/(1 - P_i(x)). \qquad (2)$$

Expressing equation (2) using (1), one obtains:

$$C_i(x) = exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \alpha_{n+1}). \qquad (3)$$

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 (3), a linear algebraic model is created to determine α:

$$b_i = (\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \alpha_{n+1}) \qquad (4)$$

for $i = 1, 2, \ldots, m$.

Let $\bar{e} = [1, \ldots, 1]^T$ be a vector of m ones and $b = (b_1, b_2, \ldots, b_m)^T$. The system of linear equations (4) may be represented

by:

$$B\alpha=b,$$

with

$$B=[A\ \bar{e}\ ]. \tag{5}$$

In (5) 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 be hold 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 [32], Golub [33] and Menard [34], that allows the assignment of values to parameters α by minimizing the square sum of the residuals ($B\alpha$ - b), summed to the squares of α . So to assign a solution to (5), we solve an unconstrained quadratic optimization problem given by:

$$\begin{aligned} &Minimize && (6)\\ &f(\alpha)=\alpha^{\ T}\ \alpha + (\mathrm{B}\alpha - \mathrm{b})^T\ (\mathrm{B}\alpha - \mathrm{b}). \end{aligned}$$

As $f(\alpha)$ is convex, the argument that minimizes (6) is given by differentiating $f(\alpha)$ w.r.t α and setting the result equal to zero that yields:

$$(I + B^TB)\alpha=B^Tb, \tag{7}$$

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 (6) is obtained by the solution to (7) and it is unique. So, given a query q=[q_1, $q_2 \ldots$,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:

$$P(q)=g(q)/(1 + g(q)), \tag{8}$$

where

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

Receiver operating characteristic curve (ROC)

ROC curves have been used to determine the goodness of fitting for a logistic regression model to predict binary outcomes, and have also been used with microarray data to predict the performance of a classifier for the presence or absence of cancer [35]. Given some explanatory variables, a patient belongs to the coded classes, for an established cut-off value defined in the test $P_i(x)$ (supplementary material). ROC curve is a graphic presentation of the relationship between both sensitivity and specificity [36–38]. The sensitivity and specificity were calculated as in equations (9) and (10).

$$sensitivity=tp/(tp+fn) \tag{9}$$

where tp represents true positives or the positive instances classified as positive and fn represents false negatives or the positive instances classified as negative.

$$specificity=tn/(fp+tn) \tag{10}$$

where tn represents true negatives or the negative instances classified as negative and fp represents false positives or the negative instances classified as positive.

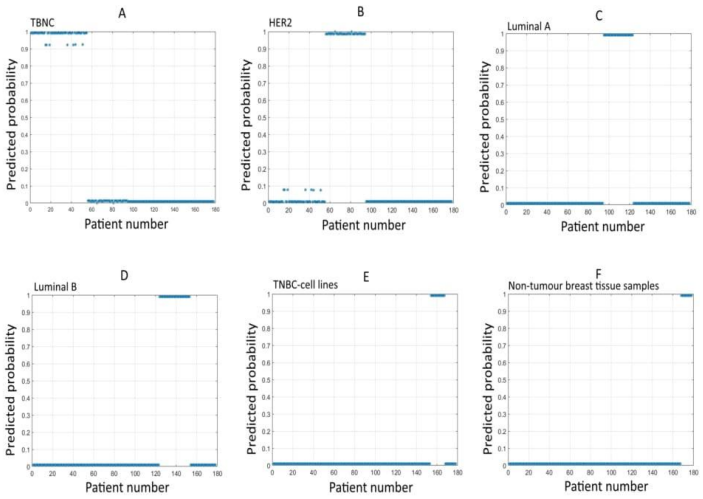

Figure 1 Probability results for samples from patients being classified into breast cancer subtypes TNBC, HER2, Luminal A, or Luminal B, non-tumor tissue samples, or of being TNBC cell lines' samples, as compared to the others, for all systems created from GSE65194 dataset. (A) System 1 discriminates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

Results

Modified Logistic Regression Model: classification using whole features

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 data sets, aiming to perform binary classification discriminating between the subtypes of breast cancer and between each subtype and non-tumor breast tissue samples. The methodology exhibited good performance in the classification task. Figure 1 illustrates these results for all subsystems created from GSE65194 dataset. 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 [7, 39, 40]. The model that we propose here circumvents the need for variable pruning by aggregating the quadratic term to the solution of a system of equations to determine the value of $\alpha_i{}^*$ associated parameters. The variables that are associated with gene expression and do not have a discriminatory role in any of the classification models are indirectly removed from the model as their $\alpha_i{}^*$ associated parameters are either zero or close to zero (Figure 2).

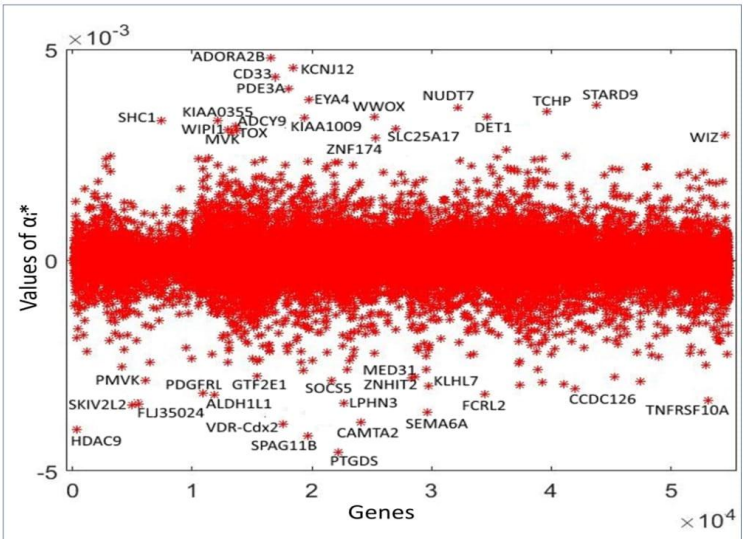

Figure 2 Genes associated with features that represent the 20 most positive and 20 most negative parameters $\alpha_i{}^*$ from one of the 6 systems created from GSE65194 dataset.

Assessment of the Probability of Misclassification

The probability of misclassification is the most important property of a classifier because it quantifies the predictive capability of the classifier [41]. 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 misclassification rate we took a random subset of 15% of samples from different data sets (27 patients from GSE65194 dataset, 13 from GSE20711 dataset and 46 from GSE25055 data set) and applied the new proposed logistic regression model with the stabilizing term to the shortened data sets and to the subsets created, to evaluate the classification performance. 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 data sets, 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 didn't disrupt the matrix organization structure.

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

Discussion

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 don't work for microarray data because generally there will be far more variables (the measured expression levels) than observations [7]. 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 parameters α 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.

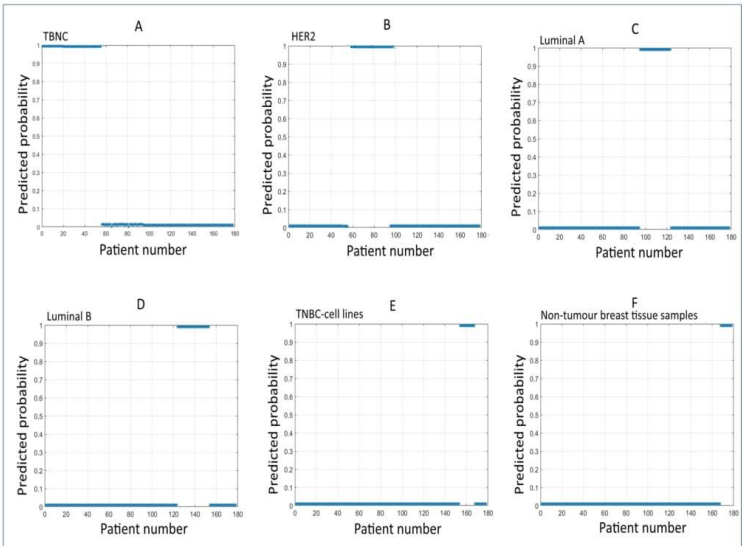

Figure 3 Probability results for samples from patients being classified into breast cancer subtypes TNBC, HER2, Luminal A, or Luminal B, non-tumor tissue samples, or of being TNBC cell lines samples, as compared to the others, for all systems created from GSE65194 dataset, using the 250 most positive and the 250 most negative parameters $\alpha_i{}^*$ of each of the systems created. (A) System 1 discriminates TNBC against the other breast cancer subtypes, TNBC cell line samples and non-tumor breast tissue samples, (B) system 2 discriminates Her2 against the other breast cancer subtypes, TNBC cell lines samples and non-tumor breast tissue samples, (C) system 3 discriminates Luminal A against the other breast cancer subtypes, TNBC cell lines samples and non-tumor breast tissue samples, (D) system 4 discriminates Luminal B against the other breast cancer subtypes, TNBC cell lines samples and non-tumor breast tissue samples, (E) system 5 discriminates TNBC cell lines samples against all breast cancer subtypes and non-tumor breast tissue samples, and (F) system 6 distinguishes between presence or absence of breast cancer.

When plotting the parameters $\alpha_i{}^*$ 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 $\alpha_i{}^*$ 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 parameters $\alpha_i{}^*$ 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 parameters $\alpha_i{}^*$. The genes on the list were matched with the genes in GSE20711 and GSE25055 data sets, and new systems were created considering just these genes, but with the same framework as used in the previously created systems for these data sets. The new proposed logistic regression model was applied to the newly created systems using the parameters $\alpha_i{}^*$ that represent the features in the data sets associated with the 462 genes selected and determined upon classification of samples in GSE65194 dataset, with a minimum performance of 80% from a cross validation procedure (see in supplementary material).These results reveal that the parameters $\alpha_i{}^*$ topologically located at

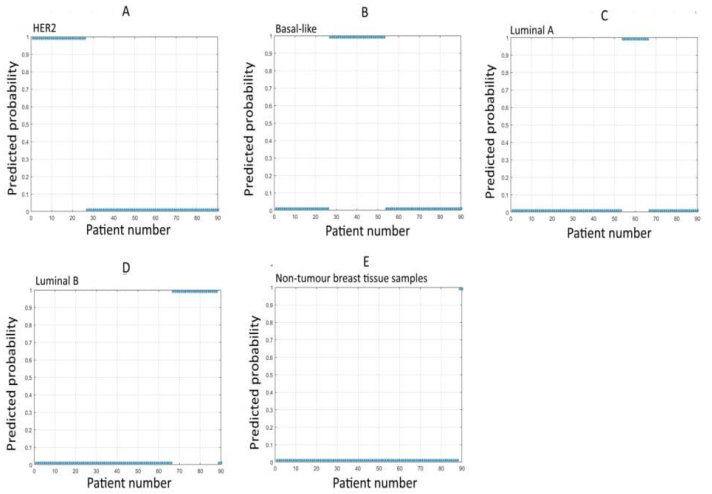

Figure 4 Probability results for samples from patients being classified into breast cancer subtypes HER2, Basal-like, Luminal A, Luminal B, or non-tumour tissue samples, as compared to the others, for all systems created from GSE20711 dataset, using the 250 most positive and the 250 most negative parameters $\alpha_i{}^*$ of each of the 6 systems created from GSE65194 data set. (A) System 1 discriminates HER2 against other breast cancer subtypes and non-tumour breast tissue samples; (B) system 2 discriminates Basal-like against other breast cancer subtypes and non-tumour breast tissue samples; (C) system 3 distinguishes between Luminal A and other breast cancer subtypes and non-tumour breast tissue samples; (D) system 4 distinguishes between Luminal B and other breast cancer subtypes and non-tumour breast tissue samples; (E) system 5 distinguishes between presence or absence of breast cancer.

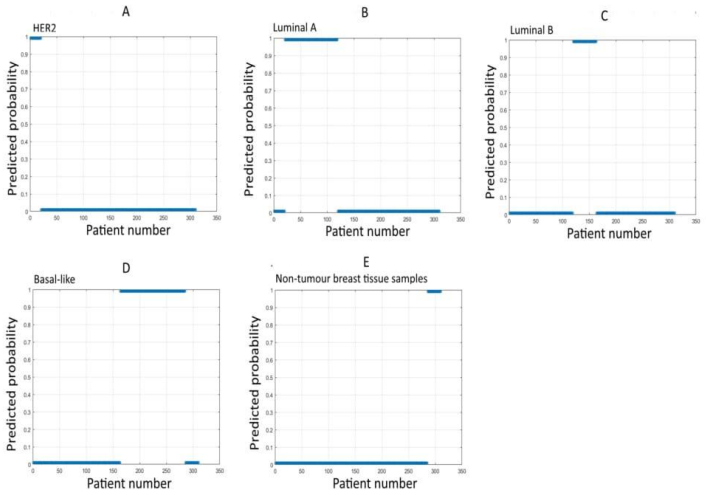

Figure 5 Probability results for samples from patients being classified into breast cancer subtypes HER2, Luminal A, Luminal B, Basal-like, or non-tumour tissue samples, as compared to the others, for all systems created from GSE25055 dataset, using the 250 most positive and the 250 most negative parameters $\alpha_i{}^*$ of each of the 6 systems created from GSE65194 data set. (A) System 1 discriminates HER2 against other breast cancer subtypes and non-tumour breast tissue samples; (B) system 2 distinguishes between Luminal A and other breast cancer subtypes and non-tumour breast tissue samples; (C) system 3 distinguishes between Luminal B and other breast cancer subtypes and non-tumour breast tissue samples; (D) 4 discriminates Basal-like against other breast cancer subtypes and non-tumour breast tissue samples; and (E) system 5 distinguishes between presence or absence of cancer in the breast tissue samples.

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 3, 4 and 5 illustrate the classification of the samples present in all the systems created from GSE65194, GSE20711 and GSE25055 data sets, respectively. To scrutinize this prospect the 40 genes that matched the features represented by the most positive and most negative parameters $\alpha_i{}^*$ were flagged to search the literature. In figures 2 the genes associated with the most negative (PTGDS, HDAC9, VDR-Cdx2, ALDH1L1, SOCS5, SPAG11B, CAMTA2, SEMA6A, TNFRSF10A, ZNHIT2, MED31, PDGFRL, FCRL2, LPHN3, KLHL7, FLJ35024, PMVK, SKIV2L2, CCDC126 and GTF2E1), and the most positive parameters $\alpha_i{}^*$ (ADORA 2B, KCNJ12, CD33, PDE3A, WWOX, ADCY9, WIPI1, SHC1, EYA4, STARD9, NUDT7, SLC25A17, WIZ, TCHP, KIAA 0355, DET1, ZNF174, KIAA1009, TOX and MVK) are identified.

Several of these genes are undoubtedly associated with breast cancer in the literature. WWOX and PDGFRL are known tumor suppressor genes in breast cancer [42, 43]. HDAC9 gene has an important role in the regulation of breast cancer cell proliferation and survival of patients [44]. ADORA2B gene is highly expressed in ER-negative breast cancer cell lines and antagonists of ADORA2B protein have shown to be toxic to breast cancer cells [45, 46].

FLJ35024 gene encodes very-low density lipoprotein receptor (VLDLR) that constitutes an apolipoprotein E-VLDLR ligand receptor system is overexpressed in human Triple-Negative Breast Cancer *in vitro* [47]. In a study of association of the polymorphisms with breast cancer risk, single nucleotide polymorphism (SNP) Met300Val in SHC1 gene shows a protective effect in breast cancer, while the non mutated form of the protein is associated with vascularization and metastatic spread of breast tumors [48, 49]. TOX genes are aberrantly expressed or mutated in several different kinds of malignancies such as breast cancer and proved to be potential diagnostic or prognostic markers in this type of cancer [50]. VDR gene encodes a protein that is a vitamin D receptor, which is expressed in most body tissues as well as on cancer cells [51]. Over 470 SNPs have been discovered in the VDR gene in different individuals [52–55], and recently the correlation between the effect of vitamin D on VDR gene regulation focusing on Cdx2 polymorphism has been established: there is a close association between specific VDR-Cdx2 polymorphism and breast cancer showing a more aggressive phenotype [56]. Expression of SOCS5 gene in breast cancer tissue was found to significantly reduce tumor growth [57]. ALDH1L1 gene codes for a protein that has tumour suppressor-like properties and is down regulated in many human cancers with at least one study correlating ALDH1L1 high expression with good overall survival for breast cancer patients [58–60]. Transcripts of LPHN3 genes are associated with axillary node status in breast cancer, and considered a potential clinical marker for predicting axillary node status accurately and tumor aggressiveness, as the common first route of spread for breast carcinoma is through the axillary lymph nodes [61]. In myeloid cells from human breast cancer CD33 gene is substantially expressed with high levels of CD33 being associated with reduced overall patient survival and accelerated tumour progression [62, 63]. Targeting CD33 is still not a therapeutic reality for breast cancer patients but is being considered as such by researchers [63].

Genes ZNHIT2, STARD9, PDE3A and WIPI1 are in a way related to breast cancer in the literature, although this relation is not very clear. In a study that aimed to characterize the mutational pattern of several Basal- like breast cancer model cell lines to improved the understanding of Basal-like breast cancer biology and for the development of drug targets for this aggressive subtype, the authors reported that ZNHIT2 was mutated in two of them [64]. STARD9 gene product is associated with mitotic microtubule formation and cell division and was associated with breast cancer in an *in vitro* study performed by Torres and collaborators [65] where depletion of STARD9 in MCF-7 (breast adenocarcinoma) cell lines caused the pericentriolar material to fragment. PDE3A gene is expressed in hundreds of cancer cell lines, including a breast cancer cell lines and is associated with cancer maintenance [66]. WIPI1 gene is expressed as two isoforms, designated α and β, and its expression is upregulated in a variety of tumors like breast cancer having gathered attention in the context of breast cancer metastasis understanding [67, 68].

CAMTA2 gene, also known as KIAA 0909, may act as tumor suppressor and when mutated allows malignant cell growth (http://www.uniprot.org/uniprot/O94983). Despite there is no direct reference in the literature to this gene

and breast cancer, CAMTA2 gene maps to human chromosome 17 to which two key tumor suppressor genes are associated, namely, p53 and BRCA1 (a genetic determinant of early onset breast cancer). Like CAMTA2, also MED31 and DET1 genes are not described as being explicitly related to breast cancer, but MED31 gene was validated as a target of microRNA-1 (miR-1) in osteosarcoma [69, 70] and DET1 was identified as one of the 54 miR-155-specific target genes in B-cell lymphoma [71], with miR-1 working as a critical regulator in both osteosarcoma and breast cancer and miR-155 having a known oncogene role in breast cancer [72, 73].

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, SPAG11 B, SEMA6A, FCRL2, KLHL7, KCNJ12, CCDC126, GTF2E1, SKIV2L2, TNFRSF10A and PTGDS with breast cancer.

Conclusions

We introduce here a new logistic regressi on-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 parameters $\alpha_i{}^*$, allowing the system to have a unique solution. These parameters are related with the expression of a gene, with some of the extreme values being associated with known breast cancer related genes and other topologically related genes with no reference in the literature as being related with breast cancer, flagged here to be investigated as-yet-undiscovered candidates with potential diagnostic and/or therapeutic utilities in breast cancer.

Abbreviations

GEO: Gene Expression Omnibus; GSE: GEO data series; HER2: Human Epidermal Growth Factor Receptor 2; miR-1: microRNA-1; NCBI: National Center for Biotechnology Information; RNA-Seq: RNA sequencing; ROC: Receiver Operating Characteristic Curve; TNBC: Triple Negative Breast Cancer.

Acknowledgements

Not applicable.

Funding

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Availability of data and materials

Test data sets analyzed during the current study are available at the NCBI's Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/

Author's contributions

FMR, RSM, and MAS conceived and designed the study. FMR and RSM analyzed the data. All authors interpreted the data, and drafted the manuscript. All authors have read and approved the final version of this manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent to publish

Not Applicable.

Ethics approval and consent to participate

Not Applicable.

Author details

[1]Institute of Biological Sciences, Federal University of Minas Gerais, Brazil. [2]Department of Biochemistry and Immunology,Federal University of Minas Gerais. [3]Department of Computer Science, Federal University of Minas Gerais., Av. Antônio Carlos, 6627, Pampulha, 31270-901 Belo Horizonte, Brazil. [4] Brain Institute, Federal University of Rio Grande do Norte, Av. Nascimento de Castro, 2155, 59056-450 Natal, Brazil.

References

1. Berry, M.W., Dumais, S.T., OBrien, G.W.: Using linear algebra for intelligent information retrieval. SIAM Review **37**, 573–595 (1995)
2. Eldén, L.: Numerical linear algebra in data mining. Acta Numerica **15**, 327–384 (2006)
3. Horn, D., Axel, I.: Novel clustering algorithm for microarray expression data in a truncated svd space. Bioinformatics **19**, 1110–1115 (2003)

4. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**, 30–37 (2009)
5. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular value decomposition and principal component analysis. In: Berrar, D.P., Dubitzky, W., Granzow, e. M (eds.) A Practical Approach to Microarray Data Analysis vol. 2, 4001 edn., pp. 91–109. Kluwer, Norwell, MA (2003)
6. Eilers, P.H., Boer, J.M., Ommen, G.J.B.V., Houwelingen, J.H.C.V.: Classification of microarray data with penalized logistic regression. Proceedings of SPIE - The International Society for Optical Engineering **4266**, 187–198 (2011)
7. Fort, G., Lambert-Lacroix, S.: Classification using partial least squares with penalized logistic regression. Bioinformatics **21**, 1104–1111 (2005)
8. Nguyen, D.V., Rocke, D.M.: Tumor classification by partial least squares using microarray gene expression data. Bioinformatics **18**, 39–50 (2002)
9. Shen, L., Tan, E.C.: Dimension reduction-based penalized logistic regression for cancer classification using microarray data. IEEE/ACM transactions on computational biology and bioinformatics **2**, 166–175 (2005)
10. Zhou, X., Liu, K.Y., Wong, S.T.: Cancer classification and prediction using logistic regression with bayesian gene selection. Journal of biomedical informatics **37**, 249–259 (2004)
11. Zhu, J., Hastie, T.: Classification of gene microarrays by penalized logistic regression. Biostatistics **5**, 427–443 (2004)
12. Brentani, R.R., et al: Gene expression arrays in cancer research: methods and applications. Critical reviews in oncology/hematology **54**, 95–105 (2005)
13. Schena, M., et al: Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science **270**, 467–70 (1995)
14. Powell, J., et al: 3d-dip-chip: a microarray-based method to measure genomic dna damage. Scientific Reports **5** (2015)
15. Zhao, S., et al: Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. Plos ONE **9** (2014)
16. Pont, M.J., et al: Microarray gene expression analysis to evaluate cell type specific expression of targets relevant for immunotherapy of hematological malignancies. Plos ONE **11** (2016)
17. Schulten, H.J., et al: Microarray expression data identify dcc as a candidate gene for early meningioma progression. Plos One **11** (2016)
18. Kumar, R., Sharma, A., Tiwari, R.K.: Application of microarray in breast cancer: An overview. J Pharm Bioallied Sci **4**, 21–26 (2012)
19. Weigelt, B., Baehner, F.L., Reis-Filho, J.S.: The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. The Journal of pathology **220**, 263–280 (2010)
20. Barnett, C.M., et al: Genetic profiling to determine risk of relapse-free survival in high-risk localized prostate cancer. Clin Cancer Res **20**, 1306–1312 (2014)
21. Han, H., Li, X.L.: Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery. BMC Bioinformatics **12**, 1–14 (2011)
22. Bellman, R.E.: Adaptive Control Processes - A Guided Tour. Princeton Legacy Library, New Jersey (2015)
23. Giancarlo, R., Lo Bosco, G., Pinello, L.: Distance functions, clustering algorithms and microarray data analysis. In: Blum, C., Battiti, R.e. (eds.) Learning and Intelligent Optimization vol. 6073, pp. 125–138. Springer, Berlin (2010)
24. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2015. CA: A Cancer Journal for Clinicians **65**, 5–29 (2015)
25. Cancer Facts e Figures. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf
26. United Kingdom Office for National Statistics. Cancer Registration Statistics, England, 2017. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/previousReleases
27. da Cunha, J.P., et al: The human cell surfaceome of breast tumors. BioMed research international **2013** (2013)
28. Zhao, Q., et al: Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proceedings of the National Academy of Sciences of the United States of America **106**, 1886–1891 (2009)
29. Irvin, W.J.J., Carey, L.A.: What is triple-negative breast cancer? Eur. J. Cancer **44**, 2799–2805 (2008)
30. New Analysis of Breast Cancer Subtypes Could Lead to Better Risk Stratification. http://www.cdc.gov/media/releases/2015/p0330-breast-cancer.html
31. NCBI's Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/
32. Linnik, I.V.: Method of Least Squares and Principles of the Theory of Observations. Pergamon Press, Russian (1961)
33. Golub, G.: Numerical methods for solving linear least squares problems. Numerische Mathematik **7**, 206–216 (1965)
34. Menard, S.: Logistic Regression. From Introductory to Advanced Concepts and Applications. SAGE Publications, Colorado, USA (2010)
35. Ma, S., Huang, J.: Regularized roc method for disease classification and biomarker selection with microarray data. Bioinformatics **21**, 4356–4362 (2005)
36. Fawcett, T.: An introduction to roc analysis. Pattern Recogn **27**, 861–874 (2006)
37. Hanczar, B., et al: Small-sample precision of roc-related estimates. Bioinformatics **26**, 822–830 (2010)
38. Lalkhen, A.G., McCluskey, A.: Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care e Pain **8**, 221–223 (2008)
39. Antoniadis, A., Lambert-Lacroix, S., Leblanc, F.: Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics **19**, 563–570 (2003)
40. Liao, J.G., Chin, K.V.: Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics **23**, 1945–1951 (2007)
41. Braga-Neto, U.M., Zollanvari, A., Dougherty, E.R.: Cross-validation under separate sampling: strong bias and how to correct it. Bioinformatics **30**, 3349–3355 (2014)
42. Ge, F., et al: Wwox suppresses klf5 expression and

breast cancer cell growth. Chinese Journal of Cancer Research **26**, 511–516 (2014)
43. Xu, M., et al: An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. BMC Genomics **9** (2008)
44. Marks, P.A.: The clinical development of histone deacetylase inhibitors as targeted anticancer drugs. Expert Opin Investig Drugs **19**, 1049–1066 (2010)
45. Desmeta, C.J., et al: Identification of a pharmacologically tractable fra-1/adora2b axis promoting breast cancer metastasis. PNAS **110**, 5139–5144 (2013)
46. McCarthy, N.: Metastasis: Adora(2b)tion. Nature Reviews Cancer **13**, 294–295 (2013)
47. Shiang, C.Y., Qi, Y., Wang, B., Pusztai, L.: P3-17-01: Apoe and its receptors (lrp8, vldlr) function as growth signals for triple-negative breast cancer and represent a novel therapeutic target. Cancer Research **71** (2012)
48. Wagner, K., et al: The insulin-like growth factor-1 pathway mediator genes: Shc1 met300val shows a protective effect in breast cancer. Carcinogenesis **25**, 2473–2478 (2004)
49. Northey, J.J., et al: Distinct phosphotyrosine-dependent functions of the shca adaptor protein are required for transforming growth factor beta (tgfbeta)-induced breast cancer cell migration, invasion, and metastasis. The Journal of biological chemistry **288**, 5210–5222 (2013)
50. Yu, X., Li, Z.: Tox gene: a novel target for human cancer gene therapy. Am J Cancer Res **15**, 3516–3524 (2015)
51. Pulito, C., et al: Cdx2 polymorphism affects the activities of vitamin d receptor in human breast cancer cell lines and human breast carcinomas. Plos ONE **10** (2015)
52. Engel, L.S., et al: Vitamin d receptor gene haplotypes and polymorphisms and risk of breast cancer: a nested case–control study. Cancer Epidemiol. Biomarkers Prev **21**, 1856–1867 (2012)
53. Anderson, L.N., Cotterchio, M., Cole, D.E., Knight, J.A.: Vitamin d related genetic variants, interactions with vitamin d exposure, and breast cancer risk among caucasian women in ontario. Cancer Epidemiol Biomarkers Prev **20**, 1708–1717 (2011)
54. Dalessandri, K.M., et al: Vitamin d receptor polymorphisms and breast cancer risk in a high-incidence population: a pilot study. J. Am. Coll. Surg. **215**, 652–657 (2012)
55. Rollison, D.E., et al: Vitamin d intake, vitamin d receptor polymorphisms, and breast cancer risk among women living in the southwestern u.s. Breast Cancer Res Treat **132**, 683–691 (2012)
56. Rose, A.A.N., Elser, C., Ennis, M., Goodwin, P.J.: Blood levels of vitamin d and early stage breast cancer prognosis: a systematic review and meta-analysis. Breast Cancer Res Treat **141**, 331–339 (2013)
57. Sasi, W., Sharma, A.K., Mokbel, K.: The role of suppressors of cytokine signalling in human neoplasms. Molecular Biology International **2014** (2014)
58. Oleinik, N.V., Krupenko, N.I., Krupenko, S.A.: Aldh1l1 inhibits cell motility via dephosphorylation of cofilin by pp1 and pp2a. Oncogene **29**, 6233–6244 (2010)
59. Oleinik, N.V., Krupenko, N.I., Krupenko, S.A.: Epigenetic silencing of aldh1l1, a metabolic regulator of cellular proliferation, in cancers. Genes e cancer **2**, 130–139 (2011)
60. Wu, S., et al: Distinct prognostic values of aldh1 isoenzymes in breast cancer. Tumor biology : the journal of the International Society for Oncodevelopmental Biology and Medicine **36**, 2421–2426 (2015)
61. Kotepui, M., et al: Quantitative real-time rt-pcr of itga7, svep1, tns1,lphn3, sema3g, klb and mmp13 mrna expression in breast cancer. Asian Pacific J Cancer Prev, **2012**, 5879–5882 (13)
62. Walter, R.B., Appelbaum, F.R., Estey, E.H., Bernstein, I.D.: Acute myeloid leukemia stem cells and cd33-targeted immunotherapy. Blood **119**, 6198–6208 (2012)
63. Laszlo, G.S., Estey, E.H., Walter, R.B.: The past and future of cd33 as therapeutic target in acute myeloid leukemia. Blood **28**, 143–153 (2014)
64. Olsson, E., et al: Mutation screening of 1,237 cancer genes across six model cell lines of basal-like breast cancer. Plos ONE **10** (2015)
65. Torres, J.Z., et al: The stard9/kif16a kinesin associates with mitotic microtubules and regulates spindle pole assembly. Cell **147**, 1309–23 (2011)
66. de Waal, L., et al: Identification of cancercytotoxic modulators of pde3a by predictive chemogenomics. Nature Chemical Biology **12**, 102–108 (2015)
67. Proikas-Cezanne, T., et al: Wipi-1α (wipi49), a member of the novel 7-bladed wipi protein family, is aberrantly expressed in human cancer and is linked to starvation-induced autophagy. Nature Oncogene **23**, 9314–9325 (2004)
68. Lee, M., Cheung, G., Done, S.J., Nair, R.: Abstract 93: Defining the roles of coil and wipi1 in breast cancer metastasis. Cancer research **2012**, 93–93 (72)
69. JIANG, C., CHEN, H., SHAO, L., WANG, Q.: Microrna-1 functions as a potential tumor suppressor in osteosarcoma by targeting med1 and med31. ONCOLOGY REPORTS **32**, 1249–1256 (2014)
70. Liu, R., et al: Hsa-mir-1 suppresses breast cancer development by down-regulating k-ras and long non-coding rna malat1. International Journal of Biological Macromolecules **81**, 491–7 (2015)
71. Slezak-Prochazka, I., et al: Inhibition of the mir-155 target niam phenocopies the growth promoting effect of mir-155 in b-cell lymphoma. Oncotarget **7** (2015)
72. Mattiske, S., et al: The oncogenic role of mir-155 in breast cancer. Cancer Epidemiol Biomarkers Prev **21** (2012)
73. Bacci, M., et al: mir-155 drives metabolic reprogramming of er+ breast cancer cells following long-term estrogendeprivation and predicts clinical response to aromatase inhibitors. Cancer Research **76**, 1615–26 (2016)

Additional Files

Supplementary material
File format: .pdf. Title of data: Supplementary material:Using a new logistic regression-based model for breast cancer classification. Description of data: Results for application of the new proposed logistic regression model to all systems created from GSE65194, GSE20711 and GSE25055 data sets are presented here. A random subset of 15% of samples from different data sets was removed from each dataset and to reduce variability, five rounds of cross-validation were performed. Probability results for 85% of the samples of patients being classified into the several breast cancer subtypes are shown, as well as sensitivity and specificity values when the new model proposed is applied to all

systems created with 85% the samples of patients, with all possible combinations of data, in five rounds.

CHAPTER 2

Potential breast cancer prediction genes

"Great discoveries and improvements invariably involve the cooperation of many minds."

Alexander Graham Bell
inventor of the telephone

1. Intrinsic Genetic Networks in Cancer Systems Biology

DNA microarrays, among other techniques, are used to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome and have revealed the intrinsic regulatory dynamics functions that remodel gene expression programs within a cell, even under the subtlest perturbations. In systems biology field of study, the chase continues for understanding the cell functions that depend upon precise regulation of thousands of genes that are turned on or off. Gene regulation can occur at any point during gene expression, but most commonly occurs at the transcription level by means of signals from the environment or from other cells that activate proteins called transcription factors (TF). These TFs bind to regulatory regions of a gene and increase or decrease the level of transcription so, by controlling the level of transcription they can determine the amount of protein product that is made by a gene at any given time. As such, as time passes by, methods have been developed to be able to determine complete reading of transcripts, including differentially expressed genes for which there is little or absolutely no information relating them with the system under study (Parikh et al., 2014; Grechkin et al., 2016; Iglesias-Martinez, et al., 2016). 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 cancer maintenance; therefore, the network approach to cancer systems biology, 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 (de la Fuente, 2010). Gene regulation takes place in various stages with many participants among which, TFs are the ones most readily analysed and easy to quantify.

GRNs operate as a "map" or a "blue print" of molecular interactions, helping to solve a number of different biological and biomedical problems (Emmert-Streib, et al. 2014). Molecular networks in mammal cells control cell proliferation and differentiation. Recently, 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 (Li and Wang, 2014; Li and Wang, 2015; Yu and Wang, 2016). As such, pathological cells manifesting tumours have their own characteristic networks; which drove us to cherry-pick the GRNs that were reconstructed using expression profiles of MCF-7 cells after artificially inducing proliferation and differentiation, to look for the topological location of some particular genes whose αi^* associated parameters values reveal extreme positive values, known crucial genes associated with breast cancer, transcription factors identified as the busiest junctions in these GRNs and also some other transcription factors already reported as having an important role on breast cancer development. The purpose is to explore the correlation of the former with breast cancer (Emmert-Streib, et al., 2014; Iglesias-Martinez, et al., 2016; Morais-Rodrigues, et al., 2017; Yu and Wang, 2016).

The GRNs used here were inferred from time-course gene expression data using the model-based method Bayesian Gene Regulation Model Inference (BGRMI) that relies on the principles of Bayesian Model Averaging (BMA) and uses discretized ordinary differential equation based mathematical models to frame the interactions between each gene and its regulators (Iglesias-Martinez, et al., 2016). This model takes into account basal expression and self-regulation to formulate the rate of change in a gene's expression as a function of the expression of its regulators. Existing ChIP-seq data and know protein-protein interactions between TFs were incorporated in BGRMI to reconstruct GRNs of proliferating and differentiating BC cells from time-course gene expression data (Iglesias-Martinez, et al., 2016;

Li et al., 2014).

Many subtypes of breast cancers are formed when breast tissue cells stop differentiating and keep proliferating (Mueller, et al., 1998). Given a certain stimulus or under specific conditions, the relative abundance of a great number of mRNA species may vary due to changes resulting from the activation of a particular gene expression program. As such, the molecular mechanisms that govern proliferation and differentiation in breast cancer cells can be studied by measuring the time course gene expression profile of MCF-7 cells stimulated with heregulin (HRG) and epidermal growth factor (EGF), to 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 (Saeki, et al., 2009). BGRMI found 22692 genes and 19016 interactions for the MCF-7 HRG and EGF stimulated cells (Iglesias-Martinez, et al., 2016). The human breast carcinoma cell line MCF-7 constitutes a powerful system for breast cancer study as in the passed information derived from these powerful experimental tool has translated into clinical benefit (Holliday and Speirs, 2011). MCF-7 is the most studied human breast cancer cell line in the world, and results from this cell line have had a fundamental impact upon breast cancer research and patient outcomes (Lee, Oesterreich and Davidson, 2015).

The genes that were flagged (as detailed in Chapter 1) to be investigated as yet-undiscovered candidates with potential diagnostic and/or therapeutic utilities in breast cancer and that are associated with the 20 αi^* parameters holding the most positive values (it is our premise that these extreme αi^* parameters are associated with genes that are important for breast cancer classification) and selected for the system that discriminates breast cancer against non-cancer samples, were searched for their involvement in the GRN reconstructed for both the EGF and HGR stimulated cells. Founded on the fact that identifying the specific breast cancer subtype that a patient holds is fundamental for the choice of the most efficacious treatment to be applied for better prognosis, we also determined the 20 αi^* parameters holding the most positive values for each breast cancer subtype calculated for the samples in GSE65194 data set and that are associated with genes exclusively associated with each subtype. All these were also searched for their involvement in the GRNs.

The topological location of these genes, of known crucial genes associated with breast cancer, of transcription factors identified as the busiest junctions in these GRNs and also of some other transcription factors already reported as having an important role on breast cancer development were considered to explore the correlation of the former with breast cancer. The flagged genes were used as input data for the prediction of their roles as oncogenes or tumour suppressor genes in breast cancer or in a specific breast cancer subtype, using the S-score system that integrates genome-wide data (de Souza, et al., 2014). The following Table 2 resumes the S-score determined for each flagged gene, as well as the "important" TF in the context of breast cancer that is related with the gene that was flagged using the methodology detailed in the next section, where we present a draft paper to be submitted to a reference paper in the cancer field of interest.

Table 2: S-score value for the genes associated with features that represent the αi* parameter values of breast cancer/ breast cancer subtype sample

Subtype	EGF induced GRN			HRG induced GRN		
	Gene	S-Score	Transcription Factor	Gene	S-Score	Transcription Factor
TNBC System 1	MED23	-3.58	SIX5	FAM102B	1.07	NFE2
	CPEB3	-1.54	CASP7	ZNF514	0.45	MXI1
	USP42	0.90	SIX5			
HER2 System 2	ZKSCAN4	-0.19	SIX5	ME2	1.15	RXRA-VDR
	CNN2	-1.58	SIX5	LRRC8E	-1.64	MXI1
	SEC61A1	1.32	CHD2			
LumA System 3	INTS4	1.82	SIX5	HECTD1	-1.31	RFX5
	SCN1B	1.10	FOXA1	CCDC92	-1.34	RFX5
				CUL5	-2.10	MXI1 and RFX5
				CHD9	-0.82	MXI1
				HDAC11	-1.22	RXRA-NR1H3
				HPS1	-0.97	NFE2
				KLHL20	3.02	RXRA-NR1H3
				VIPR2	-0.39	RXRA-VDR

LumB System 4	TSEN2 0.83 FOXA1 GSTM3 -1.01 GATA3 ZNF516 -1.98 ZEB1	PI4KB 3.92 CTBP2 GSTM3 -1.01 RAD21 GLI3 -3.43 MXI1 MACROD1 1.24 RAD21 SPRED2 -0.62 RXRA-NR1H3 FANCA -1.23 NFE2 FOLR1 2.20 NFE2 and RXRA-VDR
Breast Cancer System 5	ADORA2B 1.54 ESRRA MVK 0.61 GATA2 WWOX 0.76 GATA 3 WIPI1 2.06 FOXA1	CD33 0.90 RXRA-VDR SHC1 3.44 RXRA-NR1H3

2. Potential Breast Cancer Prediction Genes

OPEN ACCESS Freely available online

PLOS ONE

Potential breast cancer prediction genes

Francielly Morais-Rodrigues[1☯], Rita Silverio-Machado[1☯*], J Miguel Ortega[2], Sandro J Sousa[3], Marcos A dos Santos[4]

1 Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Brazil 2 Department of Biochemistry and Immunology, Federal University of Minas Gerais, Brazil 3 Brain Institute, Federal University of Rio Grande do Norte, Natal, Brazil, 4 Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil

Abstract

Among all types of cancer, breast cancer is the main cancer in women both in the developed and the developing countries, with the number of deaths from cancer being projected to continue rising. In this work we examine the pattern and feature of a GRN composed of TFs in MCF-7 breast cancer cell lines to provide valuable information relating breast cancer with some particular genes whose α_i* associated parameters values reveal extreme positive values and as such identify breast cancer prediction genes. We reveal PKN2, MKL1, MED23, CUL5 and GLI genes that demonstrate a tumor suppressor profile and MTR, ITGA2B, TELO2, MRPL9, MTTL1, WIPI1, KLHL20, PI4KB, FOLR1 and SHC1 genes that demonstrate an oncogenic profile and propose these as potential breast cancer prediction genes and that they should be prioritized for further breast cancer clinical studies.

Citation: Morais-Rodrigues F, Silvério-Machado R, dos Santos MA (2017) Potential Biomarkers in Breast Cancer. **PLoS ONE x(x): xxxxxx. doi:10.1371/journal.pone.xxxxxxx**

Editor: xxxxxxxxxxxxxxxx

Received June 30, 2017; Accepted xxxxx, 2017; Published xxxxx, 2017

Funding: This study was supported by xxxxxx. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: franrodriguesdacosta@gmail.com

☯. These authors contributed equally to this work.

Introduction

Cancer, a class of diseases characterized by uncontrolled cell growth, is one of the leading causes of human death worldwide, accounting for 8.2 million deaths in 2012 (WHO). The deaths from cancer are projected to continue rising, with annual cancer cases expected to rise from 14 million in 2012 to 22 within the next twenty years (Ferlay, et al., 2015; WHO). Among all types of cancer, breast cancer is the main cancer in women both in the developed and the developing countries, representing 25% of all cancer diagnoses and 15% of all cancer deaths among females (WHO).

Breast cancer (Purrington, et al.) is a very heterogeneous disease with several histological and molecular manifestations within tumors. (da Cunha, et al., 2013). Breast tumors can be grouped in four molecular subtypes, which have major implications for determining treatment (Luminal A, Luminal B, Triple negative/Basal-like, HER2 type) (CDC; Irvin and Carey, 2008). Recently we developed a modified logistic regression model for breast cancer classification where we present the αi associated parameters whose values are related with gene expression (Morais-Rodrigues, et al., 2017).

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 most subtle 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. 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 cancer maintenance; therefore, the network approach to cancer systems biology, 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 (de la Fuente, 2010). 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.

In this work we investigate the pattern and feature of a GRN composed of TFs in MCF-7 breast cancer cell lines (Iglesias-Martinez, et al., 2016) to provide valuable information relating breast cancer with some particular genes whose α_i* associated parameters values reveal extreme positive values and as such identify prediction breast cancer genes. Furthermore we also prognosticated their roles as oncogenes or tumour suppressor genes in breast cancer, using the S-score system that integrates genome-wide data (de Souza, et al., 2014). By flagging

prediction cancer genes here, we expect to provide new breast cancer biomarkers s, which could make a beneficial contribution to minimizing high mortality rates by providing a better prognosis.

Material and Methods

Data source

The dataset with the identifier GSE65194 was obtained from NCBI Gene Expression Omnibus (Fay, et al.) (NCBI's), containing microarray data of breast cancer samples, with no missing data. This dataset consists of 178 measurements of gene expression profilings from 153 breast cancer samples, grouped into 4 major subtypes (55 Triple Negative Breast Cancer - TNBC; 39 HER2; 29 Luminal A and 30 Luminal B), 11 non-tumor breast tissue samples obtained from mammoplasty and 14 TNBC cell lines that were disconsidered for the purpose of this study. A set of five systems were created: system 1 distinguishes between TNBC subtype samples against all other samples; system 2 discriminates HER2 subtype samples against all other samples; system 3 discriminates Luminal A subtype samples against all other samples; system 4 distinguishes between Luminal B subtype samples against all other samples; and system 5 distinguishes between presence or absence of breast cancer.

The αi* associated parameters values

The logistic regression model developed by Morais-Rodrigues and collaborators (Morais-Rodrigues, et al., 2017) was applied to all systems created to determine the α_i^* associated parameters values. The 20 α_i^* parameters with the most positive values were selected for each system created and their associated gene i were collected for each breast cancer sample classified into a particular subtype or breast cancer existence.

Gene regulatory network

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). This network also integrates ChIP-seq, that provides quantitative measurements of bindings between TFs and DNA molecules, and PPI data between TFs to increase the accuracy of the reconstructed GRN.

Assessment of diagnostic gene candidates

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 (www.bioinformatics-brazil.org/S-score/) (de Souza, et al., 2014).

Results and Discussion

Many subtypes of breast cancers are formed when breast tissue cells stop differentiating and keep proliferating (Mueller, et al., 1998). Given a certain stimulus or under specific conditions, the relative abundance of a great number of mRNA species may vary due to changes resulting from the activation of a particular gene expression program. As such, the molecular mechanisms that govern proliferation and differentiation in breast cancer cells can be studied by measuring the time course gene expression profile of MCF-7 cells stimulated with heregulin (HRG) and epidermal growth factor (EGF), to 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 (Saeki, et al., 2009). The human breast carcinoma cell line MCF-7 constitutes a powerful system for breast cancer study as in the passed information derived from these powerful experimental tool has translated into clinical benefit (Holliday and Speirs, 2011).

GRNs operate as a "map" or a "blue print" of molecular interactions, helping to solve a number of different biological and biomedical problems (Emmert-Streib, et al., 2014). Molecular networks in mammal cells control cell proliferation and differentiation. Recently, 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 (Li and Wang, 2014; Li and Wang, 2015; Yu and Wang, 2016). As such, pathological cells manifesting tumors have their own characteristic networks; which dove us to cherry-pick the GRNs that were reconstructed using expression profiles of MCF-7 cells after artificially inducing proliferation and differentiation, to look for the topological location of some particular genes whose α_i^* associated parameters values reveal extreme positive values, known crucial genes associated with breast cancer, transcription factors identified as the busiest junctions in these GRNs and also some other transcription factors already reported as having an important role on breast cancer development. The purpose is to explore the correlation of the former with breast cancer (Emmert-Streib, et al., 2014; Iglesias-Martinez, et al., 2016; Morais-Rodrigues, et al., 2017; Yu and Wang, 2016).

Each gene associated with the 20 α_i^* parameters holding the most positive values and selected for each system, were searched for its involvement in the GRN reconstructed for the EGF stimulated cells. For system 5 that distinguishes between presence and absence of breast cancer, MVK was found to be regulated by GATA2, a TF known to play a crucial role in BC proliferation (Li, et al., 2014). Also for this system, WIPI1 was found to be regulated by FOXA1. Moreover, WIPI1 gene is upregulated in a variety of tumors like breast cancer, having gathered attention in the context of breast cancer metastasis understanding (Lee, et al., 2014). ADORA2B appears in the network regulated by ESRRA that was implicated in breast cancer progression in two independent clinical studies (Ariazi, et al., 2002; Suzuki, et al., 2004). WWOX is a known tumor suppressor gene regulated in this EGF induced GRN by GATA3, a gene that is referred in the literature as being particularly useful as a marker for metastatic breast carcinoma, especially in TNBC subtype. MED23, USP42, ITGA2B, CPEB3, MTR and TELO2 genes are underscored in system 1 that distinguishes between TNBC subtype samples against all others. The first two were found to be regulated by SIX5 that in a recent study was shown to be correlated with clinical-pathological parameters of BC patients and pointed by Iglesias-Martinez and collaborators as having a potential major role in breast cancer (Iglesias-Martinez, et al., 2016). MED23 is a player recently implicated in breast cancer (Lin, et al., 2017). The ubiquitin proteasome system (UPS) is a key regulator of fundamental cellular processes of which cancer cells depend and USP42 particularly was described as being involved in gastric cancer (D'Arcy, et al., 2015). Likewise ITGA2B gene, CPEB3 gene is regulated by BHLHE40 TF, which is believed to be involved in cell differentiation and positively associated with the malignant phenotype of

invasive breast cancers (Liu, et al., 2013), which in turn regulates CASP7, an apoptosis-related cysteine peptidase that was recently show that is aberrantly expressed in breast cancer and contributes to cell growth and proliferation (Chaudhary, et al., 2016). MTR is regulated by ELF-1 that is though to modulate breast cancer progression to some extent without having an impact on survival of breast cancer patients (Gerloff, et al., 2011) and also by NF-YA, among others. NF-Y was shown to be a heterotrimer composed of NF-YA, NF-YB and NF-YC that turns on cancer related genes and that specific cancer-driving nodes are generally under NF-YA/B control (Benatti, et al., 2016). NF-YA is detected in this GRN as a regulator of IGFBP-3 apoptosis gene. It appears that in some tissues IGFBP-3 functions as a tumor promoter as it is associated with prognosis, particularly in TNBC (Marzec, et al., 2015). TELO2 gene is regulated by BDP1 that is co-expressed with HER2 (overexpressed in TNBC) and regulates its activity too (Gensler, et al., 2004). BDP1 also regulates IGFBP-3 in this EGF induced GRN. For the system that discriminates HER2 subtype samples against all other samples, genes ZKSCAN4, CNN2 are regulated by SIX5, SEC61A1 gene is regulated by CHD2, MRPL9 gene is regulated by NFE2 and METTL1 by STAT2-STAT6. CNN2 has been previously associated with breast tumorigenesis or metastasis (Ren, et al., 2011). Although CHD2 was not previously studied in the context of BC, some have mentioned that CHD2 may play a crucial role not only in the proliferation of BC cells but has having a potential clinical relevance in designing new BC treatments. MRPL9 gene encodes the mitochondrial ribosomal protein L9. Mitochondria powers breast cancer metabolism as mitochondrial respiration was proved to be required for tumor growth and development (Sotgia, et al., 2012; Villanueva, 2015). NFE2 is a TF that plays important roles in the proliferation and/or progression of breast carcinoma, is associated with poor prognosis in several different cancers and regulates MRPL9 (Sporn and Liby, 2012). MERRL1 gene is regulated by STAT2-STAT6: the STAT family of proteins is frequently implicated in breast tumorigenesis and STAT6 is important for the regulation of mammary cell differentiation with a substantial body of evidence indicating the involvement of STAT6 and other STATs in breast cancer formation, progression, prognosis and prediction, although STAT2 has not yet been associated with breast cancer (Haricharan and Li, 2014). For system 3, genes INTS4, and SCN1B are featured as being regulated by SIX5 and FOXA1, respectively. High SCN1B expression is associated with increased tumor growth and metastasis in breast cancer (Nelson, et al., 2014). To what respects system 4, TSEN2, GSTM3 (GST genotypes contribute to the individual breast cancer risk) and ZNF516 are revealed as being regulated by FOXA1, GATA3, ZEB1. This last TF is reported as ZEB1 as required for breast tumor initiation and maintenance (Mitrunen, et al., 2001; Zhou, et al., 2017). Figure 1 presents some BC potentially important genes that are part of GRN reconstructed after artificially inducing proliferation by stimulating MCF-7 cells EGF.

Stimulation of BC cells with HRG leads to variations of the landscape topography in the GRN reconstructed using this data. As such, each gene associated with the 20 α_i^* parameters holding the most positive values and selected for each system, were also searched for its involvement in the HGF induced GRN. For system 1 FAM102B and ZNF514 genes show in this GRN as being regulated by NFE2 and MXI1, respectively, which are TFs with a known role in cell differentiation and found to be two of the largest junctions in this network (Iglesias-Martinez, et al., 2016). MKL1 and PKN2 also show for this system regulated CHD2 for the former and by this TF and SIX5 for the latter. Like ZNF514, also LRRC8E revealed in system 2 is regulated by MXI1 and still in this system, ME2 gene is found to be regulated by RXRA-VDR complex, found to be another largest junction in this network and also with a known role in cell differentiation. HECTD1, CCDC92, CUL5, CHD9, HPS1, VIPR2, HDAC11 and KLHL20 genes are underscored in system 3. The first three are shown to be regulated by RFX5 whose expression was found to be predictive of BC patient survival in addition to SIX and CHD2 (Iglesias-Martinez, et al., 2016). CUL5 has been suggested to be a tumor suppressor in breast tissue, and is also regulated by MXI1, similarly to CHD9 gene (Fay, et al., 2003). HPS1 is regulated by NFE2. VIPR2, HDAC11 and KLHL20 genes are regulated by RXRA complexes: RXRA-VDR for the first and RXRA-NR1H3 for the last two. RXRA-NR1H3 complex has been previously described as master regulator of lipid synthesis in mammary epithelial cells and one of the largest transcriptional hubs in this network (Iglesias-Martinez, et al., 2016). These RXRA complexes regulate also SPRED2 and FOLR1 revealed in system 4, with the later being regulated also by NFE2, along with FANCA. GSTM3 and MACROD1 are regulated by RAD21, that likewise RXRA-VDR complex, was found to be another largest junction in this network and also with a known role in cell differentiation. PI4KB is regulated by CTBP2, whose overexpression is noted to correlate with cancer metastasis in several human cancers including breast cancer (Yang, et al., 2017). Finally, GLI3 is featured in this system and regulated by MXI1. For system 5 that distinguishes between presence and absence of breast cancer, CD33 and SHC1 genes were found to be regulated by RXRA complexes RXRA-VDR and RXRA-NR1H3, respectively.

TABLE 1: S-score value for the genes associated with features that represent the α_i^* parameters of each of the 5 systems created from GSE65194 data set.

	EGF induced GRN		HRG induced GRN	
System	α_i^* associated gene	S-score	α_i^* associated gene	S-score
1	MED23	-3.58	FAM102B	1.07
	CPEB3	-1.54	ZNF514	0.45
	MTR	3.39		
	USP42	0.90	PKN2	-2.40
	ITGA2B	2.12		
	TELO2	2.51	MKL1	-2.30
2	ZKSCAN4	-0.19	ME2	1.15
	CNN2	-1.58		
	MRPL9	4.49		
	SEC61A1	1.32	LRRC8E	-1.64
	MERRL1	2.07		
3	INTS4	1.82	HECTD1	-1.31
			CCDC92	-1.34
			CUL5	-2.10
			CHD9	-0.82
			HDAC11	-1.22
	SCN1B	1.10	HPS1	-0.97
			KLHL20	3.02
			VIPR2	-0.39
4	TSEN2	0.83	PI4KB	3.92
			GSTM3	-1.01
			GLI3	-3.43
	GSTM3	-1.01	MACROD1	1.24
			SPRED2	-0.62
	ZNF516	-1.98	FANCA	-1.23
			FOLR1	2.20
5	ADORA2B	1.54	CD33	0.90
	MVK	0.61		
	WWOX	0.76	SHC1	3.44
	WIPI1	2.06		

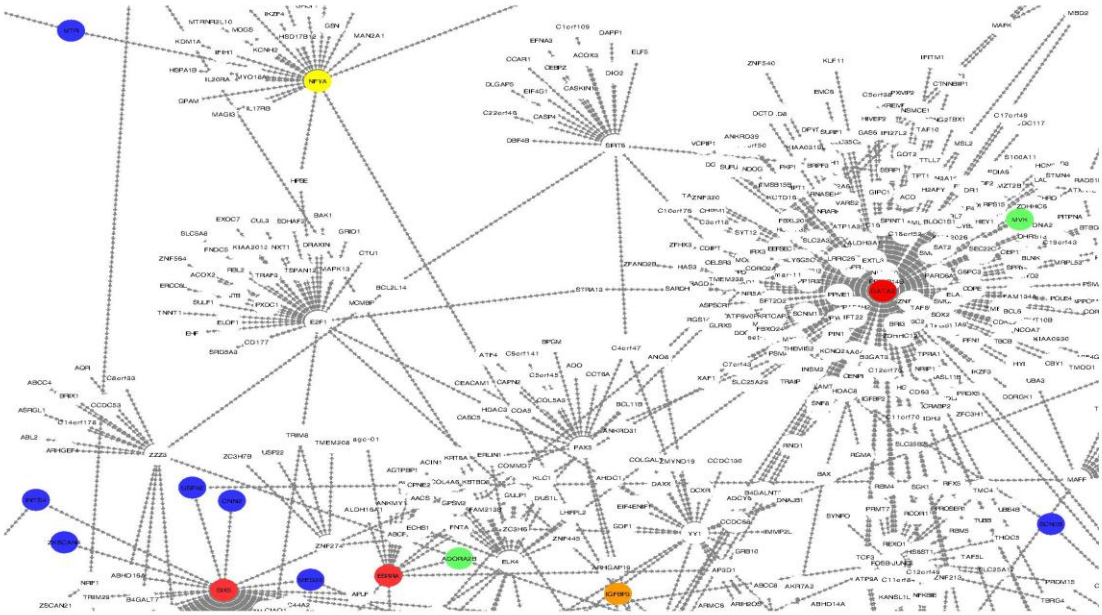

Figure 1. Part of Gene regulatory network reconstructed after artificially inducing proliferation by stimulating MCF-7 cells EGF. Blue nodes indicate potentially important cancer genes associated with breast cancer subtypes, green nodes indicate potentially important cancer genes associated with breast cancer of any subtype; red nodes indicate indicate key transcriptional regulators which play essential roles in BC cell proliferation; yellow nodes indicate proven important genes in BC and orange nodes indicate apoptosis gene.

Cancer continues to be a major health problem in developed, as well as developing countries. Late-stage presentation and inaccessible diagnosis and treatment are common. More than 90% of high-income countries reported treatment services are available compared to less than 30% of low-income countries. In 2015, only 35% of low-income countries reported having pathology services generally available in the public sector (WHO). In developing countries, there is no effective screening tool for diagnosing breast cancer. Identification of this illness can increase the chances of long-term survival of cancerous patients. Identification of the breast cancer subtype is also of paramount importance in cases where there is only one chance of treatment. The genes flagged here as potential breast cancer related genes were used as input data for the prediction of their roles as oncogenes or tumor suppressor genes in breast cancer using the S-score system (www.bioinformatics-brazil.org/S-score/). This scoring system integrates genome-wide data (copy-number variation, expression, methylation and mutations) from a set of tumor samples to generate a gene-specific score that indicates whether that specific gene is a tumor suppressor (negative *S*-score) or an oncogene (positive *S*-score) (de Souza, et al., 2014). The threshold defined was *S*-score <−2 and >2 for breast tumors. The PKN2, MKL1, MED23, CUL5 and GLI3 genes have negative *S*-scores in breast cancer (−2.40, −2.30, −3.58, −2.10, and −3.43, respectively), demonstrating a tumor suppressor profile and the MTR, ITGA2B, TELO2, MRPL9, MTTL1, WIPI1, KLHL20, PI4KB, FOLR1 and SHC1 genes have positive S-scores (3.39, 2.12, 2.52, 4.49, 2.07, 2.06, 3.02, 3.92, 2.20 and 3.44), demonstrating an oncogenic profile. These genes were underscored in different systems, and as such, can potentially be used as breast cancer biomarkers and as such they should be prioritized for further breast cancer clinical studies. On top of the rational that was followed to identify the α_i^* parameters with the most positive values, the TFs that regulate the other genes flagged here suggest that these are potentially important breast cancer genes as the TFs have a very important role in breast cancer and/or have a critical topological location in the GNR .

By flagging prediction cancer genes here, we expect to provide new breast cancer biomarkers identifiable of preference in ambulatory care settings, which could make a beneficial contribution to minimizing high mortality rates by providing a better prognosis.

References

Ariazi, E.A., Clark, G.M. and Mertz, J.E. (2002) Estrogen-related receptor alpha and estrogen-related receptor gamma associate with unfavorable and favorable biomarkers, respectively, in human breast cancer, Cancer research, 62, 6510-6518.

Benatti, P., et al. (2016) NF-Y activates genes of metabolic pathways altered in cancer cells, Oncotarget, 7, 1633-1650.

CDC New Analysis of Breast Cancer Subtypes Could Lead to Better Risk Stratification.
http://www.cdc.gov/media/releases/2015/p0330-breast-cancer.html

Chaudhary, S., et al. (2016) Overexpression of caspase 7 is ERalpha dependent to affect proliferation and cell growth in breast cancer cells by targeting p21(Cip), Oncogenesis, 5, e219.

D'Arcy, P., Wang, X. and Linder, S. (2015) Deubiquitinase inhibition as a cancer therapeutic strategy, Pharmacology & Therapeutics, 147, 32-54.

da Cunha, J.P., et al. (2013) The human cell surfaceome of breast tumors, BioMed research international, 2013, 976816.

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

Emmert-Streib, F., et al. (2014) The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks, Frontiers in genetics, 5, 15.

Fay, M.J., et al. (2003) Analysis of CUL-5 expression in breast epithelial cells, breast cancer cell lines, normal tissues and tumor tissues, Molecular cancer, 2, 40-40.

Ferlay, J., et al. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012, International journal of cancer, 136, E359-386.

Gensler, M., Buschbeck, M. and Ullrich, A. (2004) Negative regulation of HER2 signaling by the PEST-type protein-tyrosine phosphatase BDP1, The Journal of biological chemistry, 279, 12110-12116.

Gerloff, A., et al. (2011) Protein expression of the Ets transcription factor Elf-1 in breast cancer cells is negatively correlated with histological grading, but not with clinical outcome, Oncology reports, 26, 1121-1125.

Holliday, D.L. and Speirs, V. (2011) Choosing the right cell line for breast cancer research, Breast Cancer Research : BCR, 13, 215-215.

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

Irvin, W.J., Jr. and Carey, L.A. (2008) What is triple-negative breast cancer?, European journal of cancer, 44, 2799-2805.

Lee, M., et al. (2014) Abstract 93: Defining the roles of COIL and WIPI1 in breast cancer metastasis, Cancer research, 72, 93.

Li, C. and Wang, J. (2014) Quantifying the underlying landscape and paths of cancer, Journal of The Royal Society Interface, 11.

Li, C. and Wang, J. (2015) Quantifying the Landscape for Development and Cancer from a Core Cancer Stem Cell Circuit, Cancer research, 75, 2607.

Li, Y.W., et al. (2014) Decreased expression of GATA2 promoted proliferation, migration and invasion of HepG2 in vitro and correlated with poor prognosis of hepatocellular carcinoma, PloS one, 9, e87505.

Lin, B., et al. (2017) MED23 in endocrinotherapy for breast cancer, Oncology letters, 13, 4679-4684.

Liu, Y., et al. (2013) DEC1 is positively associated with the malignant phenotype of invasive breast cancers and negatively correlated with the expression of claudin-1, International journal of molecular medicine, 31, 855-860.

Marzec, K.A., Baxter, R.C. and Martin, J.L. (2015) Targeting Insulin-Like Growth Factor Binding Protein-3 Signaling in Triple-Negative Breast Cancer, BioMed research international, 2015, 638526.

Mitrunen, K., et al. (2001) Glutathione S-transferase M1, M3, P1, and T1 genetic polymorphisms and susceptibility to breast cancer, Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 10, 229-236.

Morais-Rodrigues, F., et al. (2017) Using a new logistic regression-based model for breast cancer classification, PloS one.

Mueller, E., et al. (1998) Terminal differentiation of human breast cancer through PPAR gamma, Molecular cell, 1, 465-470.

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

Nelson, M., et al. (2014) The sodium channel beta1 subunit mediates outgrowth of neurite-like processes on breast cancer cells and promotes tumour growth and metastasis, International journal of cancer, 135, 2338-2351.

Purrington, K.S., et al. (2014) Genetic variation in mitotic regulatory pathway genes is associated with breast tumor grade, Human molecular genetics, 23, 6034-6046.

Ren, S., et al. (2011) D2-40: an additional marker for myoepithelial cells of breast and the precaution in interpreting tumor lymphovascular invasion, International journal of clinical and experimental pathology, 4, 175-182.

Saeki, Y., et al. (2009) Ligand-specific sequential regulation of transcription factors for differentiation of MCF-7 cells, BMC Genomics, 10, 545.

Sotgia, F., et al. (2012) Mitochondria "fuel" breast cancer metabolism: Fifteen markers of mitochondrial biogenesis label epithelial cancer cells, but are excluded from adjacent stromal cells, Cell Cycle, 11, 4390-4401.

Sporn, M.B. and Liby, K.T. (2012) NRF2 and cancer: the good, the bad and the importance of context, Nature reviews. Cancer, 12, 564-571.

Suzuki, T., et al. (2004) Estrogen-related receptor alpha in human breast carcinoma as a potent prognostic factor, Cancer research, 64, 4670-4676.

Villanueva, M.T. (2015) Metabolism: the mitochondria thief, Nature reviews. Cancer, 15, 70.

WHO World Health Organization. Cancer. Fact sheet N. 297. http://www.who.int/mediacentre/factsheets/fs297/en/.

Yang, X., et al. (2017) C-terminal binding protein-2 promotes cell proliferation and migration in breast cancer via suppression of p16INK4A, Oncotarget, 8, 26154-26168.

Yu, C. and Wang, J. (2016) A Physical Mechanism and Global Quantification of Breast Cancer, PloS one, 11, e0157422.

Zhou, C., et al. (2017) ZEB1 confers stem cell-like properties in breast cancer

CHAPTER 3

Exploring breast cancer potential therapeutics

"It's a prototype – not the Mona Lisa."

Todd Zaki Warfel
product designer

1. Oncolytic Virotherapy

Oncolytic viruses are the major therapeutic breakthrough in the treatment of cancer, opening a new era in cancer treatment. The strategy is to use "a killer is used to kill a killer" (CBSnews, 2015), meaning using viruses to kill cancer cells. Going back in time, over a century ago the first evidence of the ability of oncolytic viruses to kill cancer cells was documented with the case of a tumour regression that has been observed for a woman diagnosed with uterine cancer and after being given the rabies vaccine (Ferhat, 2017). Still it was only recently that clinical trials demonstrated the effectiveness of this therapeutic in humans, and numerous oncolytic viruses are under clinical development today (Ferhat, 2017).

Oncolytic viruses are naturally occurring or genetically engineered viruses that have gained the oncolytic attribute for their ability to selectively infect, replicate and kill cancer cells while not affecting normal tissue (Figure 6). Oncolytic viruses are thought to mediate antitumour activity through either selective replication within cancer cells resulting in a direct lysis of tumour cells or an induction of systemic antitumour immunity.

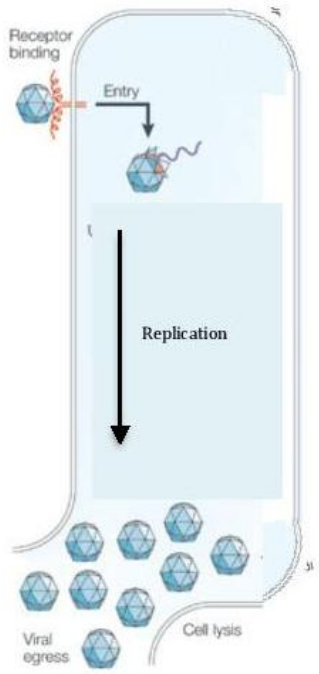

Figure 6: Illustration of an oncolytic virus invasion, replication and consequent tumour cell lysis as the cell bursts due to the number of virus that are created within the cell.

2. Seneca Valley Virus

A large number of oncolytic viruses have been proposed for cancer therapy. This includes Seneca Valley Virus (Rudin et. al., 2011). Seneca Valley Virus isolate 001 (SVV-001) is an oncolytic RNA virus that belongs to the *Picornaviridae* family originally discovered in 2002 as a contaminant in cell culture of human fetal retinoblasts and believed to be introduced through the bovine serum or porcine trypsin that was in the culture media. It is nonpathogenic to both humans and animal species and it is known to replicates through an RNA intermediate, lacking the ability to integrate into the host genome (Burke, 2016). SVV-001 is a very attractive oncolytic virus as it is a self-replicating virus that rapidly enters tumour cells through intravenous delivery, namely, its ability to target and penetrate solid tumours via intravenous administration, inability for insertional mutagenesis, and being a self-replicating RNA virus with selective tropism for cancer cells. Results from the first-in-human and first-in-children Phase I clinical trials with SVV-01 indicate safety and some clinical efficacy, albeit primarily in adult tumours (Burke, 2016; Friedman et al., 2012). SVV-001 has been shown to induce

cytotoxicity in tumours expressing neuroendocrine features, in several in vitro and in vivo models, as well as in small-cell lung carcinoma.

There is no full understanding of what underlies the specific cell tropism of Seneca virus to cancer cells, though it is known that this tropism conditions the ability of the virus to replicate in certain cell types. SVV-001, like other members of the *Picornaviridae* family, kills cells through intracellular viral replication resulting in cell lysis and autophagy. Although it has not been validated using the structure of SVV-001, it is though that, because SVV-001 has the ability to target and kill cells with neuroendocrine features, it is possible that this cell tropism is guided by the binding of receptors expressed on these tumour cells. However, a variety of motifs on the surface or near the surface in depressions or canyons of SVV-001 have been identified, that may bind to specific integrins that are present on tumour cells (Reddy et.al., 2007; Wadhwa, et.al., 2007; Poirier, et.al., 2013).

3. Exploring Breast Cancer Virotherapy using Seneca Valley Virus

The virus encodes one polyprotein that is posttranslationally processed by virus-encoded proteases into 4 structural (VP1 to VP4) and some other non-structural proteins (Hales et al., 2008; Burke, 2016 and Venkataraman et al., 2008). The crystal structure of SVV-001 was obtained at 2.3 Å resolution and stored in RSCB PDB data bank with code 3CJI. Out of these four different subunits, VP1, VP2, VP3 and VP4, of lengths 265, 286, 241 and 74 residues, the surface loops of VP1 and VP2 are predicted to mediate cell tropism of SVV-001. Since SVV-001 is known to target cells with neuroendocrine tumour features, it is possible that the cell tropism of SVV-001 might be governed by binding to receptors NCAM2 (Neural cell adhesion molecule2) and ITGA5 (Integrin alpha-5) expressed on such tumour cells.

SEMA6A is a gene flagged by application of the new logistic regression model detailed in Chapter 1 and that, in accordance with the rational proposed and presented in that chapter, is a potentially important breast cancer gene. SEMA6A codes for Semaphorin-6A protein that is a cell surface receptor. Keeping in mind that SVV-001 cancer cell tropism might be governed by binding to specific receptors on the surface of cancer cells, we hypothesized that this specific protein could be the door for Seneca Valley Virus V001 entrance in breast cancer cells. We used the in silico methodology molecular docking to prove this thesis.

3.1. Molecular Docking

Molecular docking, aims to predict the preferred orientation of one molecule to a second or more, when bound to each other to form a stable intermolecular complex (Sousa, et al., 2013). In the heart of the docking methodology is the notion of steric and physicochemical complementarity at the protein-protein interface (Teodoro, JR, and Kavraki, 2001; Yuriev and Ramsland, 2013; Sousa, et al., 2013). Modelling the interaction of two molecules is a complex problem as many forces are involved in the intermolecular association and there are many degrees of freedom, as well as insufficient knowledge of the effect of solvent on the binding association. Despite having different goals and requirements, all docking algorithms build on two basic components: sampling and scoring. Sampling consists of exploring (some of) the putative ligand protein conformations and orientations (the pose) of the ligand protein docked into the binding cavity of the receptor protein and predicting its various potential binding modes. Scoring consists of estimating the interaction energy (strength) of the binding (binding energy or binding affinity) associated with each of the predicted binding modes, using a specific scoring function. Molecular docking algorithms execute quantitative predictions of binding energetics, providing rankings of docked ligand protein conformations based on the binding affinity of protein-protein complexes. Molecular docking programs perform these tasks through a cyclical process, in which the ligand protein conformation is evaluated by specific scoring functions. This process is carried out recursively until converging to a solution of minimum energy. A scoring function should be able not only to rank the poses, but also to represent the thermodynamics of interaction of the protein-protein system (Yuriev and Ramsland, 2013; Sousa, et al., 2013). Different search algorithms have been developed to generate different poses, based on quite different approaches and at different levels of sophistication. The two critical elements in a search algorithm are speed and effectiveness in covering the relevant conformational space, which is intrinsically related with dealing with the flexibility of a molecule, as the computational time associated scales with the number of degrees of freedom included in the conformational search. Several scoring functions are developed for protein-protein interactions with different accuracies and computational efficiencies, with the availability of some being restricted to specific software packages

(Huang, et al., 2006). Scoring functions should have good accuracy and be fast enough to allow their application to a large number of potential solutions: as speed implies a number of simplifications that tend to reduce the complexity and computational cost of the scoring functions the price to pay is less accuracy.

3.1.1. ZDOC and ClusPro Web Serves

The ZDOCK server (http://zdock.umassmed.edu/) developed by Program in Bioinformatics and Integrative Biology da University of Massachusetts Medical School and Bioinformatics Program de Boston University (Pierce, Tong and Weng, 2005; Pierce et al., 2014) and ClusPro server (https://cluspro.org) developed by Structural Bioinformatics Lab Boston University and Stony Brook University (Kozakov et al., 2017), are widely used tools for protein–protein docking. Both sample the entire 6D conformational space in an initial stage, exploring only the six degrees of translational and rotational freedom for possible relative orientations of the two proteins that are considered rigid. Rigid-body methods, perform exhaustive sampling of the conformational space on a dense grid as illustrated in Figure 7.

The two servers follow a different rational. ZDOCK uses the Fast Fourier Transform algorithm to enable an efficient global docking search on a 3D grid to efficiently explore the rigid-body search space of docking positions, and utilizes a combination of desolvation (DE), shape complementarity (PSC), electrostatics (ELEC) and statistical potential terms for scoring (Chen, Li and Weng, 2003; Pierce, Hourai and Weng, 2011; Pierce, Tong and Weng, 2005). In this process, ZDock score is calculated using the following equation (Wisitponchai et al., 2017):

$$ZDock\ score = \alpha PSC + DE + \beta ELEC$$

where α and β have standard values 0.01 and 0.06, respectively.

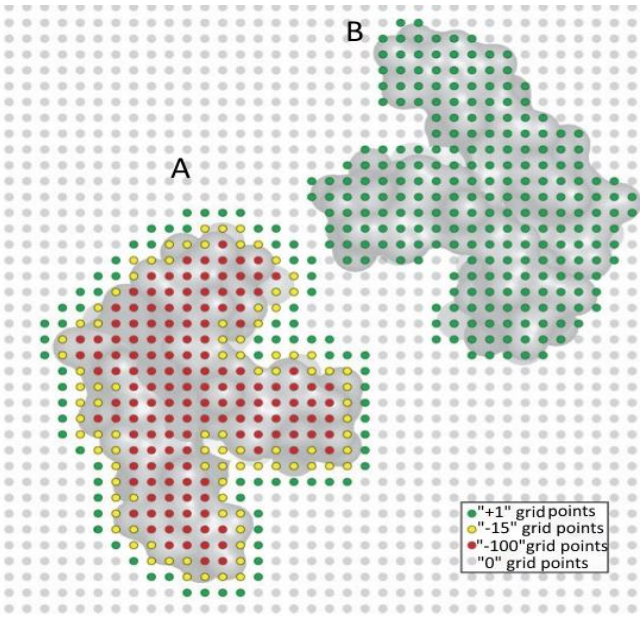

Figure 7: Fixed Grid for receptor protein (A) and mobile Grid for ligand protein (B). Adapted from: Huang, Love and Mayo, (2005).

The ClusPro server performs first rigid-body docking by sampling billions of conformations using a docking program based on the Fast Fourier Transform (FFT) correlation approach that represents the interaction energy between two proteins using an expression of the form (Kozakov et al., 2017):

$$E = w_1 E_{rep} + w_2 E_{attr} + w_3 E_{elec} + w_4 E_{DARS}$$

where E_{rep} and E_{attr} represent the repulsive and attractive contributions to the van der Waals interaction energy, E_{elec} is an electrostatic energy term, E_{DARS} is a pairwise structure-based potential and it primarily represents desolvation contributions. The coefficients $w1$, $w2$, $w3$ and $w4$ define the weights of the corresponding terms, and are optimally selected for different types of docking problems. The following computational steps are root-mean-square deviation (RMSD)-based clustering of the 1,000 lowest-energy structures generated, to find the largest clusters that will represent the most likely models of the complex; and refinement of selected structures using energy minimization.

The rigid-body docking programs based on the Fast Fourier Transform correlation approach are very efficient: in this method, the receptor protein is placed at the origin of the coordinate system on a fixed grid, the ligand protein is placed on a movable grid; and the interaction energy is written in the form of a correlation function that can be efficiently calculated using Fast Fourier Transforms and this results in the ability to exhaustively sample billions of conformations of the two interacting proteins, evaluating the energies at each grid point. A key to the success of rigid-body methods is that the shape complementarity term allows for some overlaps, and hence the methods are able to tolerate moderate differences between bound and unbound (Kozakov et al., 2017).

3.1.2. Preparation of 3D Structures for Molecular Docking

The crystal structure of Seneca Valley Virus-001 with PDB code 3CJI, Neural cell adhesion molecule2 with PDB code 2kBG), Integrin alpha-5 with PDB code 4WJK, and Semaphorin-6A with PDB code 3OKW were checked for the quality of the structures using PROCHECK (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/) which certified the stereochemical quality of all the protein structures. Using PyMol program (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC) structures were cleaned and stripped from ligands, metal ions and water molecules, and protonated

3.2. Results and Discussion

ZDock and CluPro servers were challenged to identify poses through scoring of each predicted pose of Neural cell adhesion molecule2 (2kBG) and Integrin alpha-5 (4WJK) when complexing with V001 (3CJI) in the region set for the Grid (Figure 8), that includes VP1 and VP2 loops identified in the literature as the most probable anchor points for protein interactions with cancer cell surface receptor proteins. Both algorithms were able to predict the interaction of the former mentioned proteins with V001 established at the level of the two loops of VP1 subunit and the loop of VP2 subunit. Figures 9, 10 and 11 illustrate these results for ZDock server.

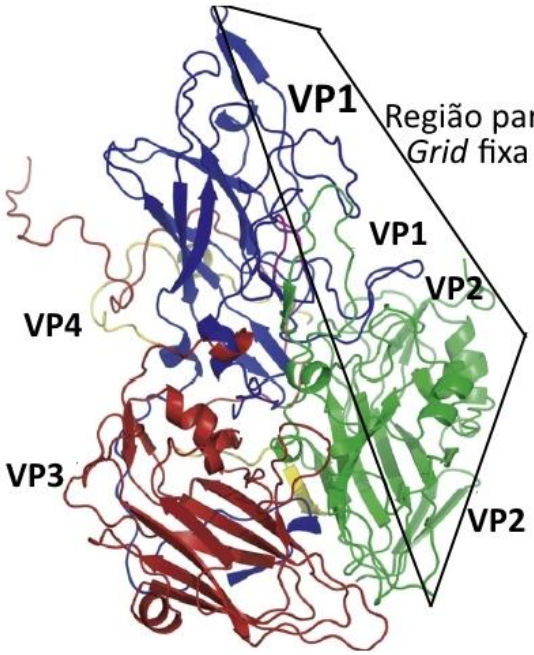

Figure 8: Fixed Grid created for molecular docking (protein-protein) at Seneca Valley Virus capside (PDB entry 3CJI). Grid includes VP1 and VP2 loops. Adapted from: Venkataraman et al., 2008.

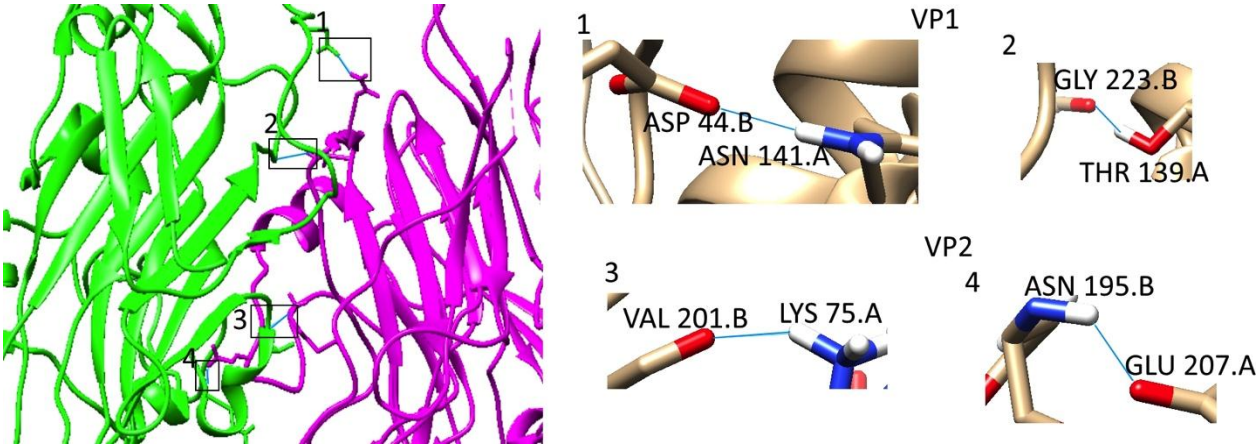

Figure 9: Interactions established between V001 (3CJI) in green and Integrin alpha-5 (4WJK) in magenta as determined by ZDock web server. Images produced using UCSF CHIMERA.

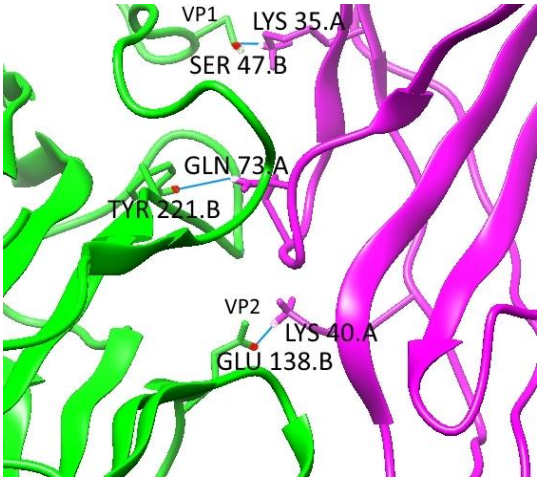

Figure 10: Interactions established between V001 (3CJI) in green and Neural cell adhesion molecule 2 (2kBG) in magenta as determined by ZDock web server. Images produced using UCSF CHIMERA.

These results confirm that both servers and the docking parameters specified in the input files for the docking method are reasonable to be applied for molecular docking of Semaphorin-6A (3OKW) and V001 (3CJI). There are many docking programs and scoring functions created for different molecular docking proteins, so validations like the one that we performed are crucial to guarantee a good performance of the docking program selected (Hevener, 2009).

Table 3 summarizes the interactions determined by web servers ClusPro and ZDOCK upon docking of Semaphorin-6A (3OKW) and V001 (3CJI) in the region of the established GRID, as well as upon molecular docking of Neural cell adhesion molecule2 (2kBG) and Integrin alpha-5 (4WJK) with the same receptor region. Interactions (mainly establishment of hydrogen bonds) were analysed using UCSF Chimera. Interaction values flanked by the * symbol were exclusive for ClusPro web server calculations and as such were not generated by ZDOCK web server, otherwise the generated results are common to both web servers. (Whitten et al., 2010).

Table 3: Interactions determined by web servers ClusPro and ZDOCK upon docking of ligand proteins with V001

Ligand Protein	Free Energy	Interactions established with 3CIJ-VP1 (chain B) – ligand P	Interactions established with 3CJI-VP2 (chain B) – ligand P
Semaphorin-6A Chain A	-39,88 (*Cluspro*) -39,80 (*ZDOCK*)	TYR 319.A O PHE 20.B H ASP 288.B O ASN 320.A H GLU 22.B O ARG 417.A H PHE 20.B O ARG 417.A H *PRO 21.B O ARG 417.A H GLU 22.B O ARG 417.A H PHE 20.A O ARG 417.B H *GLU 22.A O ARG 417.B H *PRO 21.A O ARG 417.B H *THR 414.A O ARG 417.B H	TYR 319.B O PHE 20.A H ASP 350.A O THR 352.B H ASP 350.B O THR 352.A H LYS 248.B O SER 289.A H SER 289.A O LYS 248.B H ASP 288.A O LYS 248.B H *PRO 286.A O ASN 320.B H *SER 289.A O ASN 320.B H
Neural cell adhesion molecule 2 Chain A	-39,90 (*Cluspro*) -38,60 (*ZDOCK*)	SER 47.B O LYS 35.A H TYR 221.B O GLN 73.A H *GLY 223.B O THR 74.A H *GLU 28.A O SER 47.B H	GLU 138.B O LYS 40.A H TYR 36.B O ARG 68.A H HIS 50.A O TYR 36.B H TYR 10.A O TYR 197.B H LEU 11.A O GLN 198.B H
Integrin alpha-5 Chain A	-42,35 (*Cluspro*) -39,70 (*ZDOCK*)	GLY 223.B O THR 139.A H ASP 44.B O ASN 141.A H GLU 138.B O ARG 144.A H	VAL 201.B O LYS 75.A H GLU 207.A O ASN 195.B H *LEU 79.A O GLN 198.B H

Just like for Integrin alpha-5 and Neural cell adhesion molecule 2, Semaphorin-6A established strong interactions between the two loops in VP1 structure and the one loop in VP2 structure, identified in the literature as the most likely V001 structure features responsible for cell tropism to cancer cells. The here designated strong interactions are not more than hydrogen bonds that are established when hydrogen atoms attached to nitrogen or oxygen atoms interact with highly electronegative nitrogen, oxygen or sulphur atoms, which provoke a slight polarization. This causes these atoms to be attracted or repulsed by a dipole-dipole interaction (non-covalent), stabilizing the protein-protein complex. These results make probable the creation of the complex Semaphorin-6A – V001, indicating the oncolytic virus Seneca Valley Virus and a new therapeutic option to be considered and further studied for breast cancer

treatment.

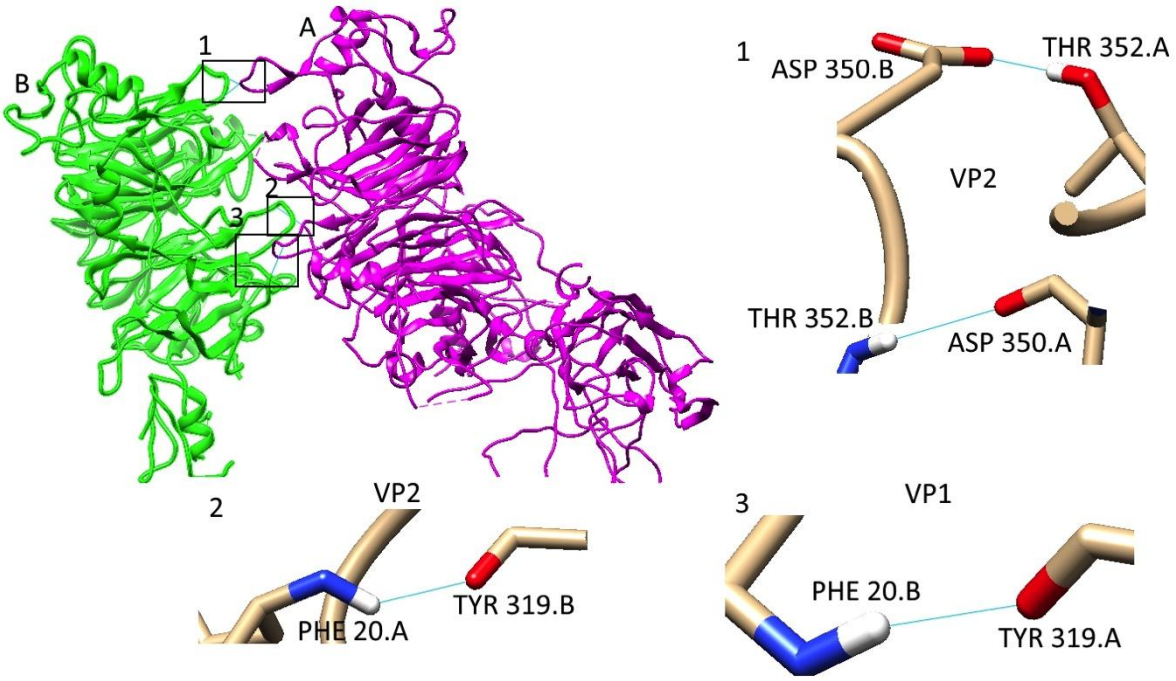

Figure 11: Interactions established between V001 (3CJI) in green and *Semaphorin-6A* (3OKW) in magenta as determined by ZDock web server. Images produced using UCSF CHIMERA.

CONCLUSIONS AND FINAL REFLEXIONS

"Tweedledum and Tweedledee decided to have a battle"

Lewis Carroll
mathematician and English writer

More and more statistics and linear algebra methods are used to address questions that emerge in microarray literature. Microarray technology is a long-used tool for global analysis of gene expression that allows simultaneous investigation of hundreds or thousands of genes in a sample, and is characterized by a low sample size and a large feature (gene) number that adversely affect similarity measurements and classification performance. To avoid the problem of the `curse of dimensionality' many authors have performed feature selection or reduced the size of data matrix. In Chapter 1 of this PhD thesis we introduce a new logistic regression-based model developed to classify breast cancer tumour samples based on microarray expression data with all features included and no reduction of microarray data matrix. This methodology allowed the correct classification of breast cancer samples from GSE65194, GSE20711, and GSE25055 data sets that contain microarray data of breast cancer samples, with a minimum performance of 80% (sensitivity and specificity) and exploring all possible combinations of data that included breast cancer subtypes. Conclusions: This new model allows the assignment of values to parameters $\alpha i*$ that are associated with the expression of a gene. Scrutinizing these parameters $\alpha i*$ unveiled that some of the topologically extreme parameters are associated with known biomarker in breast cancer and flagged a set of other genes with no identified relation to breast cancer, to be investigated as as-yet-undiscovered biomarker candidates with potential diagnostic and therapeutic utilities in breast cancer.

In Chapter 2 we examine the pattern and feature of a GRNs composed of TFs in MCF-7 breast cancer cell lines to provide valuable information relating breast cancer with some particular genes whose $\alpha i*$ associated parameter values reveal extreme positive values and as such identify breast cancer prediction genes. The topological analysis of these networks, the direct correlation observed between some of the flagged genes with relevant TFs in the context of breast cancer and using the S-score system that has been used by many to confirm the

tumour suppressor/oncogenic profile of genes in specific cancer types, we reveal PKN2, MKL1, MED23, CUL5 and GLI genes that demonstrate a tumour suppressor profile and MTR, ITGA2B, TELO2, MRPL9, MTTL1, WIPI1, KLHL20, PI4KB, FOLR1 and SHC1 genes that demonstrate an oncogenic profile and propose these as potential breast cancer prediction genes and that they should be prioritized for further breast cancer clinical studies.

A large number of oncolytic viruses have been proposed for cancer therapy, which includes Seneca Valley Virus. This is a very attractive virus for cancer therapy as it's nonpathogenic to both humans and animal species but is a self-replicating virus that rapidly penetrates solid tumours via intravenous administration and destroys it through direct lysis of the host cancer cells that bursts when too many virus replicate. SEMA6A is a gene flagged by application of the new logistic regression model detailed in Chapter 1 and that, in accordance with the rational proposed and presented in that chapter, is a potentially important breast cancer gene. SEMA6A codes for Semaphorin-6A protein that is a cell surface receptor. Keeping in mind that SVV-001 cancer cell tropism might be governed by binding to specific receptors on the surface of cancer cells, we hypothesized that this specific protein could be the door for Seneca Valley Virus V001 entrance in breast cancer cells. We used the in silico methodology molecular docking to prove this thesis. These results obtained make probable the creation of the complex Semaphorin-6A – V001, indicating the oncolytic virus Seneca Valley Virus as a new therapeutic option to be considered and further studied for breast cancer treatment.

REFERENCES

"The human brain is built to compare; it's Darwinian to consider an alternative when one presents itself"

Hellen E. Fisher
anthropologist

American Cancer Society. (2015) Cancer Facts & Figures. Available from: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2015-2016.pdf.

Badve, S., et.al. (2011) Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. Modern Pathology 24, 157–167.

Barnett, C.M., et al, (2014) Genetic profiling to determine risk of relapse-free survival in high-risk localized prostate cancer. Clin Cancer Res 20, 1306-1312.

Bellman, R.E. (2015) Adaptive Control Processes – A Guided Tour. Princeton Legacy Library, New Jersey.

Beniwal, S. and Arora, J. (2012) Classification and feature selection techniques in data mining, International Journal of Engineering Research & Technology (IJERT), 1(6).

Brentani, R.R., et al. (2005) Gene expression arrays in cancer research: methods and applications, Critical reviews in oncology/hematology,54: 95-105.

Brooks, M.D., Burness, M.L. and Wicha, M.S. (2015) Therapeutic Implications of Cellular Heterogeneity and Plasticity in Breast Cancer, Cell Stem Cell, 17.

Burke, M.J. (2016) Oncolytic Seneca Valley Virus: past perspectives and future directions. Oncolytic Virotherapy, 5 81–89

CBSnews (2015). Using polio to kill cancer: A producers' notebook. CBS News, USA.

Chen, R., Li,L. and Weng, Z. (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins, volume 52, Issue 1, pages 80–87.

da Cunha, J.P., et al. (2013) The human cell surfaceome of breast tumours, BioMed research international, 976816.

de la Fuente, A. (2010) From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends in genetics : TIG.26(7):326-33.

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

Deerwester, S., et al. (1990) Indexing by latent semantic analysis, Journal of the American society for information science, 41, 391.

Emmert-Streib, F., et. al. (2014) The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. Frontiers in GENETICS, 2014.00015.

Ferhat, M. (2017) Oncolytic Viruses: The Next Major Breakthrough in Cancer Treatment. J Hum Virol Retrovirol 5(1): 00141. DOI: 10.15406/jhvrv.2017.05.00141.

Fort, G. and Lambert-Lacroix, S. (2005) Classification using partial least squares with penalized logistic regression, Bioinformatics, 21(7):1104-1111.

Friedman, G.K., et al. (2012) Targeting pediatric cancer stem cells with oncolytic virotherapy. Nature Volume 71 | Number 4.

Giancarlo, R., Bosco,G.L. and Pinello,L. (2010) Distance Functions, Clustering Algorithms and Microarray Data Analysis. In: Blum C, Battiti R. (eds). Learning and Intelligent Optimization: 4th International Conference, LION 4, Venice, Italy, January, p. 18-22. Selected Papers. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 125-138.

Grechkin, M., et al. (2016) Identifying Network Perturbation in Cancer. PlosOne 12(5).

Hales, L.M., et al. (2008) Complete genome sequence analysis of Seneca Valley virus-001, a novel oncolytic picornavirus. Journal of General Virology, 89, 1265–1275

Han, H., Li, X.L. (2011) Multi-resolution independent component analysis for high-performance tumour classification and biomarker discovery. BMC Bioinformatics 12, 1-14.

Hevener, K.E. (2009) Validation of Molecular Docking Programs for Virtual Screening against Dihydropteroate Synthase. J Chem Inf Model. 49(2): 444–460.

Holliday, D.L. and Speirs, V. (2011) Choosing the right cell line for breast cancer research, Breast Cancer Research : BCR, 13, 215-215.

Hu, K., et al. (2016) The Outcome of Breast Cancer Is Associated with National Human Development Index and Health System Attainment. PlosOne 11(7).

Huang, N., et al. (2006) Physics-based scoring of protein-ligand complexes: enrichment of known inhibitors in large-scale virtual screening, *Journal of chemical information and modeling*, **46**, 243-253.

HUANG, P.S., LOVE, J.J, MAYO, S.L. (2005) Adaptation of a Fast Fourier Transform-Based Docking Algorithm for Protein Design. J Comput Chem 26: 1222–1232, 2005

Hyeoun-Ae, P. (2013) An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain, J Korean Acad Nurs, 43(2).

Iglesias-Martinez, 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 | DOI: 10.1038/srep37140.

INCA National Cancer Institute José Alencar Gomes da Silva (2016). Available from: http://www.inca.gov.br/estimativa/2016/index.asp?ID=2.

James, G., et.al. (2013) An Introduction to Statistical Learning with Applications in R, ISBN 978-1-4614-7138-7 (eBook).

Kossenkov, A.V. and Ochs , M.F. (2010) Matrix Factorization Methods Applied in Microarray Data Analysi, Int J Data Min Bioinform, 4(1): 72–90.

Kozakov, D., et al. (2017) The ClusPro web server for protein–protein docking. Nature Protocols 12, 255–278

Kumar, R., Sharma, A., Tiwari, R.K. (2012) Application of microarray in breast cancer: An overview. JPharm Bioallied Sci 4, 21-26.

Lee, A.V, Oesterreich, S. and Davidson, N.E. (2015). MCF-7 Cells—Changing the Course of Breast Cancer Research and Care for 45 Years. J Natl Cancer Inst (2015) 107 (7)

Li, C. and Wang, J. (2014) Quantifying the underlying landscapeand paths of cancer, Journal of The Royal Society Interface, 11.

Li, C. and Wang, J. (2015) Quantifying the Landscape for Development and Cancer from a Core Cancer Stem Cell Circuit, Cancer research, 75, 2607.

Li, P. et.al. (2014) Gene regulatory network inference and validation using relative change ratio analysis and time-delayed dynamic Bayesian network. Journal on Bioinformatics and Systems Biology, 2014:12.

Malhotra, et.al. (2010) Histological, molecular and functional subtypes of breast cancers. Cancer Biology & Therapy 10:10, 955-960

Marcolino, L.S., Couto, B.R.G.M. and Santos, M.A.D. (2010) Genome Visualization in Space, Proceedings of IWPACBB, 225-232.

McKinney, B.A., et al. (2007) Evaporative cooling feature selection for genotypic data involving interactions, Bioinformatics, 2, no. 16 2007, pages 2113–2120.

Morais-Rodrigues, et al. (2017) Using a new logistic regression-based model for breast cancer classification.

Mueller, E., et al. (1998) Terminal differentiation of human breast cancer through PPAR gamma, Molecular cell, 1, 465-470.

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

Parikh, A.P., et al. (2014) Network Analysis of Breast Cancer Progression and Reversal Using a Tree-Evolving Network Algorithm. PlosOne, 10(7), e1003713.

Park, C., et al. (2016). Prognostic values of negative estrogen or progesterone receptor expression in patients with luminal B HER2-negative breast cancer. World Journal of Surgical Oncology 14:244
Pierce, B., Tong, W. and Weng, Z. (2005) M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. Bioinformatics, Vol. 21 no.8, pages 1472–1478.

Pierce, B.G., et. al., (2014) ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. Bioinformatics, Vol. 30 no.12, pages 1771–1773

Pierce, B.G., Hourai,Y. and Weng, Z. (2011) Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library. PlosOne, 6(9), e24657.

Poirier, J.T., et.al. (2013) Selective Tropism of Seneca Valley Virus for Variant Subtype Small Cell Lung Cancer. J Natl Cancer Inst;2013;105:1059–1065.
Polyak, K. (2011) Heterogeneity in breast cancer. The Journal of Clinical Investigation, 121(10).

Pont, M.J., et al. (2016) Microarray gene expression analysis to evaluate cell type specific expression of targets relevant for immunotherapy of hematological malignancies. Plos ONE 11.

Powell, J., et al. (2015) 3d-dip-chip: a microarray-based method to measure genomic dna damage. Scientific Reports 5.

Reddy, P.S., et.al. (2007) Seneca Valley Virus, a Systemically Deliverable Oncolytic icornavirus, and the Treatment of Neuroendocrine Cancers. J Natl Cancer Inst 2007;99: 1623-33.

Ringnér, M., et.al. (2011) GOBO: Gene Expression-Based Outcome for Breast Cancer Online. PlosOne, 6(3), e17911.

Rivenbark, A.G., O'Connor, S.M. and Coleman, W.B. (2013) Molecular and Cellular Heterogeneity in Breast Cancer Challenges for Personalized Medicine, The American Journal of Pathology, 183(4).

Rudin, C.M., et al. (2011) Phase I clinical study of Seneca Valley Virus (SVV‐001), a replication-competent picornavirus, in advanced solid tumors with neuroendocrine features. Clin. Cancer Res. 17, 888–895.

Saeki, Y., et al. (2009) Ligand-specific sequential regulation of transcriptionfactors for differentiation of MCF-7 cells, BMC Genomics, 10, 545.

Saeys,Y., Inza I., and Larranaga, P. (2007) A review of feature selection techniques in bioinformatics, Bioinformatics, 23, no. 19 2007, pages 2507–2517.

Schatten, H. Cell and molecular biology of breast cancer. New York : Springer, 2013. ISBN: 978-1-62703-633-7 (Print) 978-1-62703-634-4 (eBook).

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

Schulten, H.J., et al. (2016) Microarray expression data identify dcc as a candidate gene for early meningioma progression. Plos One 11.

Siegel, R.L., Miller, K.D. and Jemal, A. (2015) Cancer statistics, 2015, CA: A Cancer Journal for Clinicians, 65(1):5-29.

Sousa, S.F., et al. (2013) Protein-ligand docking in the new millennium--a retrospective of 10 years in the field, Current medicinal chemistry, 20, 2296-2314.

Teodoro, M.L., JR,G.N.P. and Kavraki, L.E., (2001) Molecular Docking: A Problem With Thousands Of Degrees Of Freedom. IEEE International Conference on Robotics and Automation (ICRA 2001), Seoul, Korea, 2001, pp. 960–966.

Thomas, M., et al. (2014) Predicting breast cancer using an expression values weighted clinical classifier, BMC Bioinformatics, 15:411.

Tomfohr, J., Lu, J. and Thomas Kepler, T.B. (2005) Pathway level analysis of gene expression using singular value decomposition, BMC Bioinformatics, 6:225.

United Kingdom O_ce for National Statistics. Cancer Registration Statistics, England,2017. Available from:
https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/previousReleases.

Venkataraman, S., et al. (2008) Structure of Seneca Valley Virus-001, An oncolytic picornavirus representing a new genus. Structure, 16(10): 1555–1561.

Wadhwa, L., et.al. (2007) Treatment of Invasive Retinoblastoma in a Murine Model Using an Oncolytic Picornavirus. American Association for Cancer Research. CAN-07-2352.

Weigelt, B., Baehner, F.L., Reis-Filho, J.S. (2010) The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. The Journal of pathology 220, 263-280.

Whitten,K.W., et al. Chemistry. EUA: Brooks Cole, 2010. 10° edition.

Wilcox, R.H. (1961) Adaptive control processes - A guided tour, by Richard Bellman, Princeton

University Press, Princeton, New Jersey. Naval Research Logistics Quarterly,8,315-316.

Wisitponchai, T., et al. (2017) AnkPlex: algorithmic structure for refinement of near-native ankyrin-protein docking. BMC Bioinformatics 18:220.

Wu, M., et al. (2015) Prioritization Of Nonsynonymous Single Nucleotide Variants For Exome Sequencing Studies Via Integrative Learning On Multiple Genomic Data. Nature Scientific Reports | 5:14955 | DOI: 10.1038/srep14955.

Xu, L., et al. (2011) Functional Cohesion of Gene Sets Determined by Latent Semantic Indexing of PubMed Abstracts. PlosOne, 6(4), e18851.

Yu, C. and Wang, J. (2016) A Physical Mechanism and Global Quantification of Breast Cancer, PloS one, 11, e0157422.

Yuriev, E. and Ramsland, P.A. (2013) Latest developments in molecular docking: 2010-2011 in review, Journal of molecular recognition : JMR, 26, 215-239.

Zhao, H., et al. (2013) Modified Logistic Regression Models Using Gene Coexpression and Clinical Features to Predict Prostate Cancer Progression, Hindawi Publishing CorporationComputational and Mathematical Methods in Medicine, 917502.

Zhao, Q., et al. (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line, Proceedings of the National Academy of Sciences of the United States of America, 106(6):1886-1891.

Zhao, S., et al. (2014) Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. Plos ONE 9.

ANNEX 1

Using a new logistic regression-based model for breast cancer classification – supplementary material

"The writer writes only half the book; the other half is with the reader."

Joseph Conrad
novelist

Supplementary Material

Using a new logistic regression-based model for breast cancer classification

Francielly Morais-Rodrigues, Rita Silvério-Machado, J Miguel Ortega, Frederico F Campos Filho, Sandro J de Souza and Marcos A dos Santos.

Results for application of the new proposed logistic regression model to all systems created from GSE65194, GSE20711 and GSE25055 data sets are presented here. A random subset of 15% of samples from different data sets was removed from each dataset and to reduce variability, five rounds of cross-validation were performed. Probability results for 85% of the samples of patients being classified into the several breast cancer subtypes are shown in Figures 1 to 3, for all systems created from GSE65194, GSE20711 and GSE25055 data sets, respectively. Sensitivity and specificity of the new model proposed was determined when applied to all systems created with 85% the samples of patients, with all possible combinations of data, in 5 rounds. Values are calculated for all genes included and with the number of genes that arise as the most frequently occurring genes that matched the 500 parameters α_i^* topologically located on the extremes of each of the 6 systems created from GSE65194 dataset, which corresponds to 462 genes for GSE65194 dataset, 320 genes for GSE20711 dataset and 319 genes for GSE25055 dataset. Sensitivity and specificity values are presented in Tables I to III.

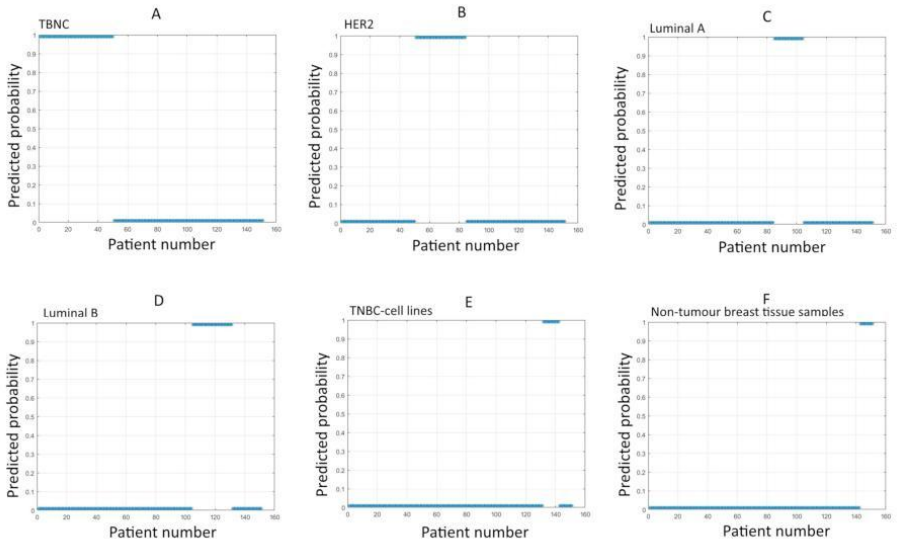

Figure 1: Probability results for samples from patients being classified into breast cancer subtypes TNBC, HER2, Luminal A, Luminal B, TNBC cell line, or non-tumor tissue samples, as compared to the others, for all systems created from GSE65194 dataset. (A) System 1 discriminates TNBC against the other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

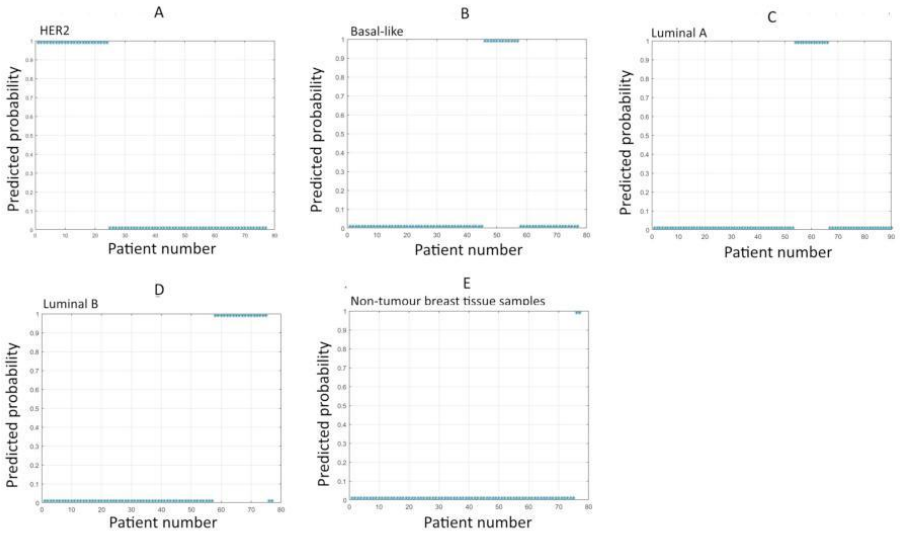

Figure 2: Probability results for samples from patients being classified into breast cancer subtypes HER2, Basal-like, Luminal A, Luminal B, or non-tumor tissue samples, as compared to the others, for all systems created from GSE20711 dataset. (A) System 1 discriminates HER2 against other breast cancer subtypes and non-tumor breast tissue samples; (B) system 2 discriminates Basal-like against other breast cancer subtypes and non-tumor breast tissue samples; (C) system 3 distinguishes between Luminal A and other breast cancer subtypes and non-tumor breast tissue samples; (D) system 4 distinguishes between Luminal B and other breast cancer subtypes and non-tumor breast tissue samples; (E) system 5 distinguishes between presence or absence of breast cancer.

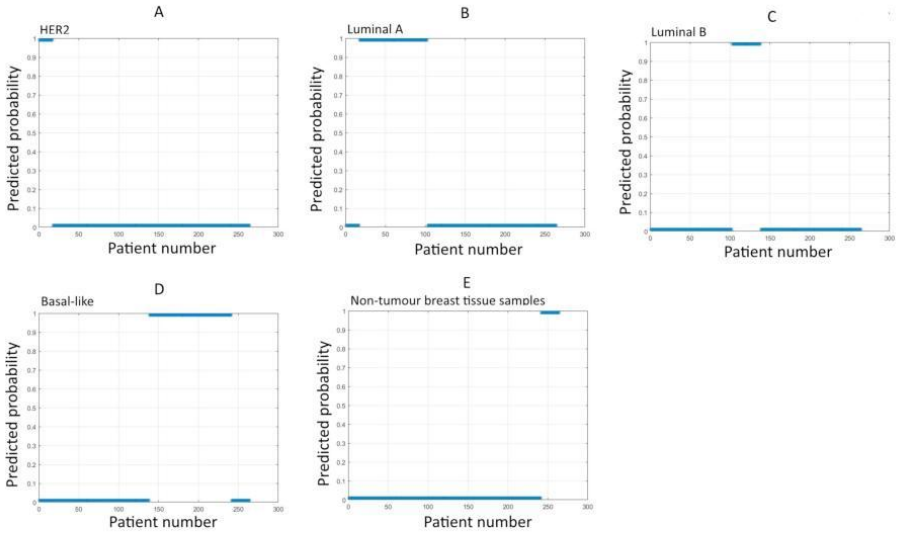

Figure 3: Probability results for samples from patients being classified into breast cancer subtypes HER2, Luminal A, Luminal B, Basal-like, or non-tumor tissue samples, as compared to the others, for all systems created from GSE25055 dataset. (A) System 1 discriminates HER2 against other breast cancer subtypes and non-tumor breast tissue samples; (B) system 2 distinguishes between Luminal A and other breast cancer subtypes and non-tumor breast tissue samples; (C) system 3 distinguishes between Luminal B and other breast cancer subtypes and non-tumor breast tissue samples; (D) 4 discriminates Basal-like against other breast cancer subtypes and non-tumor breast tissue samples; and (E) system 5 distinguishes between presence or absence of cancer in the breast tissue samples.

Table I. Sensitivity and specificity values for the new model applied to all systems (GSE65194 dataset) .

Round #	Breast cancer subtype	# patients to classify	All genes included			462 genes included		
			# patients classified using P(x)	Sensitivity	Specificity	# patients classified using P(x)	Sensitivity	Specificity
1	TNBC	5	4	0.80	1	5	1	1
	HER2	5	4	0.80	1	3	0.60	1
	Luminal A	9	6	0.67	1	7	0.78	1
	Luminal B	3	2	0.67	1	3	1	1
	TNBC cell line	3	2	0.67	1	2	0.67	1
	non-tumor	2	2	1	1	2	1	1
2	TNBC	10	9	0.90	1	10	1	1
	HER2	7	4	0.57	1	5	0.71	1
	Luminal A	4	4	1	1	3	0.75	1
	Luminal B	2	3	1	0.96	3	1	0.96
	TNBC cell line	2	2	1	1	1	0.50	1
	non-tumor	2	2	1	1	2	1	1
3	TNBC	5	5	0.83	1	5	1	1
	HER2	6	4	0.67	1	6	1	1
	Luminal A	5	4	0.80	1	4	0.80	1
	Luminal B	4	5	1	0.96	3	0.75	1
	TNBC cell line	3	3	1	1	3	1	1
	non-tumor	4	4	1	1	3	0.75	1
4	TNBC	11	9	0.82	1	10	0.91	1
	HER2	5	5	1	1	5	1	1
	Luminal A	6	5	0.83	1	6	1	1
	Luminal B	3	3	1	1	3	1	1
	TNBC cell line	2	2	1	1	1	0.50	1
	non-tumor	0	0	0	0	0	0	0
5	TNBC	5	4	0.80	1	5	1	1
	HER2	9	6	0.67	1	8	0.89	1
	Luminal A	5	5	1	1	5	1	1
	Luminal B	6	6	1	1	5	0.83	1
	TNBC cell line	1	4	1	0.90	2	1	0.96
	non-tumor	1	1	1	1	1	1	1

Table II. Sensitivity and specificity values for the new model applied to all systems (GSE20711 dataset) .

Round #	Breast cancer subtype	# patients to classify	All genes included			320 genes included		
			# patients classified using P(x)	Sensitivity	Specificity	# patients classified using P(x)	Sensitivity	Specificity
1	HER2	2	2	1	1	2	1	1
	Basal-like	6	5	0.83	1	6	1	1
	Luminal A	1	3	1	0.85	1	1	1
	Luminal B	4	3	0.75	1	3	0.75	1
	non-tumor	0	0	0	0	0	0	0
2	HER2	3	3	1	1	3	1	1
	Basal-like	4	3	0.75	1	4	1	1
	Luminal A	1	1	1	1	2	1	0.92
	Luminal B	3	3	1	1	1	0.33	1
	non-tumor	2	2	1	1	1	1	1
3	HER2	4	4	1	1	4	1	1
	Basal-like	3	2	0.67	1	3	1	1
	Luminal A	2	2	1	1	2	1	1
	Luminal B	2	2	1	1	2	1	1
	non-tumor	2	2	1	1	2	1	1
4	HER2	4	3	0.75	1	4	1	1
	Basal-like	4	2	0.50	1	3	0.75	1
	Luminal A	2	2	1	1	2	1	1
	Luminal B	3	3	1	1	3	1	1
	non-tumor	0	0	0	0	0	0	0
5	HER2	4	3	0.75	1	4	1	1
	Basal-like	2	2	1	1	2	1	1
	Luminal A	2	1	0.50	1	1	0.50	1
	Luminal B	3	3	1	1	3	1	1
	non-tumor	2	2	1	1	2	1	1

Table III. Sensitivity and specificity values for the new model applied to all systems (GSE25055 dataset) .

Round #	Breast cancer subtype	# patients to classify	All genes included			319 genes included		
			# patients classified using P(x)	Sensitivity	Specificity	# patients classified using P(x)	Sensitivity	Specificity
1	HER2	3	3	1	1	4	1	0.98
	Luminal A	14	12	0.85	1	14	1	1
	Luminal B	8	9	1	0.98	7	0.87	1
	Basal-like	19	19	1	1	19	1	1
	non-tumor	2	3	1	0.98	1	0.50	1
2	HER2	2	2	1	1	2	1	1
	Luminal A	12	10	0.83	1	12	1	1
	Luminal B	10	12	1	0.94	9	0.90	1
	Basal-like	18	18	1	1	18	1	1
	non-tumor	4	3	0.75	1	3	0.75	1
3	HER2	3	2	0.67	1	3	1	1
	Luminal A	12	12	1	1	10	0.83	1
	Luminal B	7	8	1	0.97	7	1	1
	Basal-like	20	20	1	1	20	1	1
	non-tumor	4	3	0.75	1	4	1	1
4	HER2	0	0	0	0	0	0	0
	Luminal A	18	18	1	1	18	1	1
	Luminal B	8	8	1	1	7	0.87	1
	Basal-like	15	10	0.67	1	13	0.86	1
	non-tumor	5	4	0.80	1	4	0.80	1
5	HER2	4	4	1	1	3	0.75	1
	Luminal A	14	13	0.92	1	13	0.92	1
	Luminal B	6	6	1	1	6	1	1
	Basal-like	17	15	0.88	1	16	0.94	1
	non-tumor	5	4	0.80	1	5	1	1

ANNEX 2

Side Project

"How does a project get to be a year late? One day at a time."

Frederick P. Brooks Jr.
computer scientist

These following references concern the comprehensive work accomplished between August 2013 and June 2015 that was published in a high-ranking international peer-reviewed scientific journal as well as conference abstracts.

Van Voorhis, W.C., ... , **Morais Rodrigues da Costa, F.**, ... , et al. (2016) Open Source Drug Discovery with the Malaria Box Compound Collection for Neglected Diseases and Beyond, PLoS pathogens, 12, e1005763.

Silva, E.B., Cardoso, M.V.O., Siqueira, L.R.P., **Costa, F.M.R.**, Villela, F.S., Ferreira, R.S., Leite, A.C.L. Novel Thiophenol-Thiosemicarbazones Derivatives as Cruzain Inhibitors. 23rd International Congress of the IUBMB and 44th Annual Meeting of the Brazilian Society for Biochemistry and Molecular Biology - Foz do Iguaçu, Paraná, Brazil. August 2015.

Costa, F.M.R., Villela, F.S., Pereira, G.A.N., Ferreira, R.S. Discovery And Characterization of Cruzain and Rhodesain Small Molecule Inhibitors through Experimental Screening. X European Workshop in Drug Design - Siena, Itally. May 2015.

Santos, L.H., Silva, V.S., **Costa, F.M.R.**, Caffarena, E.R., Ferreira, R.S. Recognition of a suitable receptor structure for the virtual screening of novel HIV-1 reverse transcriptase inhibitors. 7th Brazilian Symposium on Medicinal Chemistry - Campos do Jordão, São Paulo, Brazil. November 2014.

Costa, F.M.R., Oliveira, A., Villela, F.S., Azevedo, V.A.C., Ferreira, R.S. Discovery of Cruzain and Rhodesain Small Molecule Inhibitors through Experimental Screening and Molecular Docking. ISCB-Latin America x-Meeting on Bioinformatics with BSB and SoiBio - Belo Horizonte, Minas Gerais, Brazil. October 2014.

Oliveira, A., **Costa, F.M.R.**, Ferreira, R.S., Azevedo, V.A.C. Molecular Modeling of Phospolipase D (Sphngomyelinase) of Corynebacterium pseudotuberculosis (CpEMaseD) in the development of Potential Microbicides Molecules. ISCB-Latin America x-Meeting on Bioinformatics with BSB and SoiBio - Belo Horizonte, Minas Gerais, Brazil. October 2014.

Costa, F.M.R., Villela, F.S., Ferreira, R.S. Discovery of Cruzain and Rhodesain Small Molecule Inhibitors through Experimental Screening. Brazilian Symposium on Chemistry and Physiology of Proteases and their Inhibitors - São Carlos, São Paulo, Brazil. September 2014.

Costa, F.M.R., Oliveira, A., Villela, F.S., Azevedo, V.A.C., Ferreira, R. S. Discovery of Cruzain and Rhodesain Small Molecule Inhibitors through Experimental Screening and Molecular Docking. Annual Meeting Program Casadinho/PROCAD between the Program in Bioinformatics UFMG and the Program on Computational and Systems Biology (BCS) - Institute Oswaldo Cruz / FIOCRUZ, 2014 - - Belo Horizonte, Minas Gerais, Brazil. 2014.

Costa, F.M.R. Presentation of all versions of the DOCK program. Brazilian Workshop in Structural biology - ICB/UFMG, Belo Horizonte, Minas Gerais, Brazil. 2014.