33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

A disease annotation study of gene signatures in a breast cancer microarray dataset

Foivos Gypas, Ekaterini S Bei, Michalis Zervakis, *Member, IEEE,* Stelios Sfakianakis

***Abstract*— Breast cancer is a complex disease with heterogeneity between patients regarding prognosis and treatment response. Recent progress in advanced molecular biology techniques and the development of efficient methods for database mining lead to the discovery of promising novel biomarkers for prognosis and prediction of breast cancer. In this paper, we applied three computational algorithms (RFE-LNW, Lasso and FSMLP) to one microarray dataset for breast cancer and compared the obtained gene signatures with a recently described disease-agnostic tool, the Genotator. We identified a panel of 152 genes as a potential prognostic signature and the ERRFI1 gene as possible biomarker of breast cancer disease.**

I. Introduction

Breast cancer, a leading cause of cancer death in women, is characterized by its molecular and clinical heterogeneity. Breast, cervical, endometrial and ovarian cancer contribute to 45% of total female malignancies and approximately 880000 cancer related deaths annually [1]. Our understanding of the biology and molecular basis for this common disease, as well as the factors that contribute to breast cancer risk, has greatly increased over the past few decades. Markers such as estrogen receptor (ER), progesterone receptor (PR) and epidermal growth factor receptor family member (ERBB2/HER2) are used for prognostication and multiple gene profiling studies have been conducted, searching for genomic measurements with predictive power for breast cancer prognosis [2]–[4]. One challenge for bio-informatists is to tease out useful information from massive data sets for further analysis.

II. Computational methods

In this work we have used multiple machine learning and statistical methods for the elimination of non informative genes and the identification of possible biomarkers.

A. Support Vector Machines

Support vector machines [5] map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes (see Fig. 1). An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalisation error of the classifier will be.

E. S. Bei, F. Gypas and M. Zervakis are with the Department of Electronic and Computer Engineering, Technical University of Crete, Chania 73100, Greece, beieka@yahoo.gr, foivos@gypas.com, michalis@display.tuc.gr

S. Sfakianakis is with the Institute of Computer Science, Foundation for Research and Technology, Heraklion 71110, Greece ssfak@ics.forth.gr

In cases where the data are not linearly separated, the training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function φ. Then SVM finds a linear separating hyperplane $\boldsymbol{w}^T\varphi(\boldsymbol{x}) + b$ with the maximal margin in this higher dimensional space by solving the following optimization problem [6]

$$\begin{array}{ll} \min_{\boldsymbol{w},b,\xi} & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} & y_i(\boldsymbol{w}^T\varphi(\boldsymbol{x}) + b) \geq 1 - \xi_i, i = 1, \ldots, N \\ & \xi_i \geq 0, i = 1, \ldots, N \end{array}$$

The introduction of the "slack" variables ξ_i permits the calculation of the *soft margin* in non-separable cases where some misclassifications are inevitable while the positive C is the penalty parameter of the error term. Furthermore, for dealing with the higher dimensionality introduced by the φ transformation, a *kernel function* $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^T\varphi(\boldsymbol{x}_j)$ suffices to bypass the tranformation and solve the optimization problem in the original, finite dimensional space.

B. RFE-LVM

The output of the linear and non-linear parts of a single neuron (Fig. 2) in a neural network are given as [7]

$$\begin{array}{rcl} u & = & \sum_{i=1}^{m} w_i g_i \\ f(u) & = & \frac{1}{1+e^u} = y \\ f'(u) & = & y(1-y) \end{array}$$

where w_i is the weight associated to gene g_i. The error function that is pursued for minimization is

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{j=1}^{n}(d_j - y_j)^2 \qquad (1)$$

where n corresponds to the number of samples, d_j represents the desirable neuron output associated with sample j and y_j is the actual output produced by this neuron for the

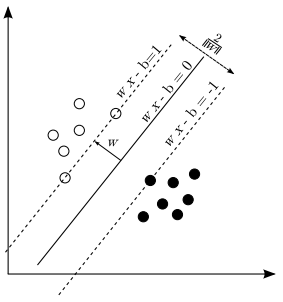

Fig. 1. A separating hyperplane between two classes

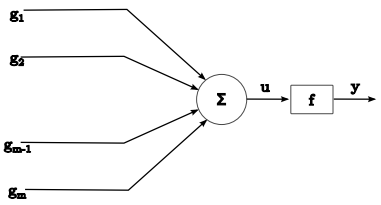

Fig. 2. A single linear neuron

given sample. Using the gradient descent method for the minimization of Eq. (1), we update the weight w_i associated to gene g_i as follows

$$w_i \leftarrow w_i - \eta \frac{\partial E(\boldsymbol{w})}{\partial w_i} = w_i + \eta \sum_{j=1}^{n} (d_j - y_j) y_j (1 - y_j) g_{ij} \quad (2)$$

The Recursive Feature Elimination based on Linear Neuron Weights (RFE-LNW) algorithm [8] is similar to the RFE-SVM [9] approach which uses a linear kernel to assess the weights of the direction vector w which in turn is used as a gene ranking criterion. It introduces a Fisher's metric in the above update equation:

$$w_i \leftarrow w_i + \eta \sum_{j=1}^{n} (d_j - y_j) y_j (1 - y_j) g_{ij} \frac{|g_{ij} - \bar{g_i}|}{\sigma_+(g_i) + \sigma_-(g_i)} \quad (3)$$

where $\bar{g_i}$ is the mean expression of the gene g_i, and $\sigma_+(g_i)$, $\sigma_-(g_i)$ are the standard deviation of the expression of g_i in the positive and the negative class respectively.

C. LASSO

The LASSO technique (Least Absolute Shrinkage and Selection Operator [10]) can be used in cases where we expect a response variable to be determined by a linear combination of a subset of potential covariates. It minimizes the following problem

$$\begin{array}{rl} \min_{\boldsymbol{\beta}} & \sum_{k=1}^{m} (y_k - \boldsymbol{\beta} \cdot \boldsymbol{x}_k)^2 \\ \text{subject to} & \sum_{i=1}^{n} |\beta_i| \leq t \end{array}$$

The t is the parameter that performs the "shrinkage" of the β parameters closer to zero and thus producing a sparse model. In this sparse model covariates x_i that are associated with $\beta_i = 0$ have been effectively eliminated doing some automatic feature selection.

D. Feature Selection MLP

In this multilayer perceptron (MLP) network genes are selected in a supervised way using "gate opening" [11]. The gates are located as nodes in the input layer of the MLP and implement parameterized functions that when given the expression of a gene as input, either close, when evaluate to 0, or partially open, when return a value in the $(0, 1]$ interval. At the beginning of the learning all gates are almost closed as if no gene is important in the classification task. During the learning phase each input node computes the product of the expression of a gene and the gate function value as its output. This output is passed on to the next layers of the network and the gate function produces high values for marker genes and low values for non important ones. The learning algorithm, called the Feature Selection MLP (FSMLP), adapts the parameters of the gate functions and the weights of the network so that genes that can reduce the error faster are "opened" faster.

III. Dataset and Methodology

A. Dataset

Our analysis was performed on the breast cancer dataset from vant Veer et al. [12] which is divided into train and test set. Primary gene expression data files of clinical samples as well as information on the associated standardisation of the data and system used can be found online and in the original publication [12]. The train set consists of 78 sporadic lymph node patients; 44 of them remained free of disease after the initial diagnosis, and 34 developed distant metastasis within five years. The test set comprises 19 lymph node negative breast cancer patients. 7 patients remained metastasis free for at least five years, whereas the other 12 developed distant metastasis within five years.

B. Methodology

Based on the computational methods (RFE-LNW, Lasso and FSMLP) described in the above section we obtain three gene signatures (see supplementary information). In combination with these methods, we employ an SVM in order to check the classification accuracy whenever we need it. As classification accuracy we denote the correct classification rate as calculated using SVM.

The first gene signature (190 genes) is the output of the RFE-LNW algorithm. At each step of the RFE-LNW we estimate the classification accuracy with SVM, keeping the minimum number of genes with the best classification accuracy (82-83%), which in our case is 190 genes. (Note that SVM correct rate for the whole dataset is 78-79%).

The second gene signature is the outcome of the Lasso algorithm; we fine-tuned the algorithm's parameter t via manual trials; with the right choice of parameter, a small subset of the genes is selected - 82 in our case- which results in good classification accuracy (circa 74%).

The third gene signature is derived from a combination of RFE-LNW and FSMLP; one should note that other feature selection techniques may be used as well. Using RFE-LNW we keep 1500 genes (using 10-fold cross validation in the training set). Those 1500 genes are fed into Multilayer Perceptron Network with gate opening, which is described in paragraph (II-D) above. Our Neural Network consists of 3 layers: the first layer contains 1500 nodes, the middle (hidden) layer contains 150 nodes, and the output layer comprises a single node. The Neural Network is trained using the aforementioned training set. After the training process, 200 genes corresponding to the highest gate opening values are kept (accuracy around 80-85%). Due to randomization, we have to iterate through many realizations of the NN; from those realizations, we select those genes that appear in 90% of the times. Consequently, this third gene signature consists of 152 genes.

TABLE I
THE THREE GENE SIGNATURES

a/a	Number of genes	Method used	Training set Accuracy
1	190	RFE-LNW	82-83%
2	82	Lasso	74%
3	152	RFE-LNW, FSMLP	80-85%

Table I summarizes these three signatures.

IV. BIOLOGICAL EVALUATION OF THE SIGNATURES

First, to illustrate their potential we compared these three gene signatures with the Genotator, a disease-agnostic tool for genetic annotation of disease [13]. Genotator is a software tool, developed to facilitate multi-database searching and to provide a more complete picture of advances in genetic research of human diseases. Genotator generates a comprehensive set of results for breast cancer disease and for any disease by integrating gene and annotation data from 11 externally accessible and best-of-breed genetic resources. More interestingly, the results from Genotator are ranked using a scoring system that integrates bibliomic and genomic data and provides a preliminary likelihood of strength of association for use in future thesis testing [13].

Of note, in the publication of vant Veer et al. [12] many cDNA sequences had no gene symbol, gene name or information associated with them. Given this fact, we have updated and examined ontology information for all genes included in all three gene signatures and their encoded proteins to examine their significance in Genotator database. 35.79%, 46.34% and 32.9% of genes from the first, second and third gene signature respectively had no yet gene names and could not associated with any information (see supplementary information). Second, we address a comparison of the three gene signatures to identify the common genes. Here, we report the findings of our analyses.

A. Results - Discussion

An important result to come from our analyses addresses Genotator specialization. For Breast Cancer, Genotator database yielded 29 genes (23.2%) from the first gene signature (125 known genes in 190 gene set), 12 genes (27.27%) from the second gene signature (44 known genes in 82 gene set) and 39 (38.24%) from the third gene signature (102 known genes in 152 gene set), suggesting that their inclusion within the Genotator workflow provided new, and potentially valuable information about the genes involved in breast cancer disease (see supplementary information). In view of the above, we focus on the third signature. Finally, we also cannot rule out the possibilities that the first and second gene signatures may have additional characteristics that differ from those of the more promising third gene signature or that applying the same method to a larger dataset may result in different signatures. We argue that the data produced so far may be preliminary to launch large-scale study.

We linked the 102 known gene products (third signature) to their Biological Process Gene Ontology (GO) annotations, a procedure relying upon a controlled vocabulary for describing proteins with respect to their biological processes [14]. All biological processes identified through this procedure were mapped to the corresponding proteins. As shown in Fig. 3, comparison appears to give clearer insights into this gene signature suggesting at least five principal processes; metabolic processes, response to stimulus, signal transduction, gene expression, and protein modification processes, each associated with breast cancer pathology.

Comparison analysis on the three gene signatures identified 19 common genes between RFE and LASSO, 5 common genes between RFE and MLP, and 5 common genes between MLP and LASSO. Interestingly, a single gene, the ERRFI1 gene is a particularly distinct marker among all three gene signatures. As a meta-query engine, Genotator has provided that the ERRFI1 gene is associated with breast cancer. Therefore, we focused our attention on the ERFFI1.

ERBB receptor feedback inhibitor 1 or mitogen-inducible gene 6 protein (ERRFI1 or MIG-6 also known as RALT or Gene 33) is a multiadaptor protein thought to be involved in the regulation of receptor tyrosine kinase (RTK) and stress signalling [15].

Epidermal growth factor receptor (EGFR) is a membrane tyrosine kinase that is implicated in the regulation of a wide variety of biological processes [16]. Members of the epidermal growth factor receptor family (EGFR/ERBB1, ERBB2/HER2, ERBB3/HER3 and ERBB4/HER4) are key targets for inhibition in cancer therapy. The cytoplasmic protein ERRFI1 interacts with and inhibits the kinase domains of EGFR and ERBB2, which are critical for the activation by the formation of an asymmetric dimmer [17]. Recent data, using gene expression analysis showed that ERRFI1 expression is correlated with the phosphorylated active state of EGFR and that ERRFI1 expression is associated with basal EGFR kinase activity in the absence of ligand [18].

Animal studies demonstrate that ERRFI1 is a specific negative regulator of EGFR signaling in skin morphogenesis, and a novel tumor suppressor of Egfr-dependent carcinogenesis [15], and also ERRFI1 is a crucial regulator of pulmonary development and vascularization [19] and in the tumorigenesis of endometrial cancer [20].

Duncan et al. [21] using a multifaceted genome-wide analysis in glioblastomas, indicate that ERRFI1 is a potential glioblastoma-targeted tumor suppressor gene and a key component in the EGFR signaling pathway involved in glioblastoma development. Also, they demonstrate that restoring ERRFI1 expression in an ERRFI1-deficient glioblastoma cell line decreases glioblastoma cell migration.

In vitro study has demonstrated that the ERRFI1 gene does not undergo mutational inactivation in breast cancer and suggested a role for loss of ERRFI1 signalling in the pathogenesis of ERBB2-amplified breast carcinomas [22], while Xu et al. have shown that ERRFI1 gene promotes breast cancer cell growth by an anti-apoptotic rather than a mitogenic effect, possibly involving up-regulation of Poly(ADP-

ribose) Polymerase (PARP-1) protein in multiple human breast cancer cell lines [23]. Interestingly, in a previous clinical study, using a tissue-wide expression profile analysis, ERFFI1 was identified as a down-regulated gene in tumors of breast cancer patients with a poor prognosis [24]. The above studies strongly support the significance of ERRFI1 as a crucial regulator in intracellular signalling and give rise to our notion that ERRFI1 could be a possible biomarker for breast cancer disease.

V. Conclusions

The present study provides a combined analysis of three computational algorithms with a disease-agnostic tool, the Genotator for the identification of prognostic gene signature for breast cancer disease and the selection of candidate biomarkers.

In agreement with the analysis using the Genotator, there appears to be an important role played by the third gene signature and suggests that the third gene signature is a potential prognostic signature and the ERRFI1 gene could be a promising biomarker for breast cancer disease. Molecular studies are necessary to delineate the role of ERRFI1 signaling in breast cancer.

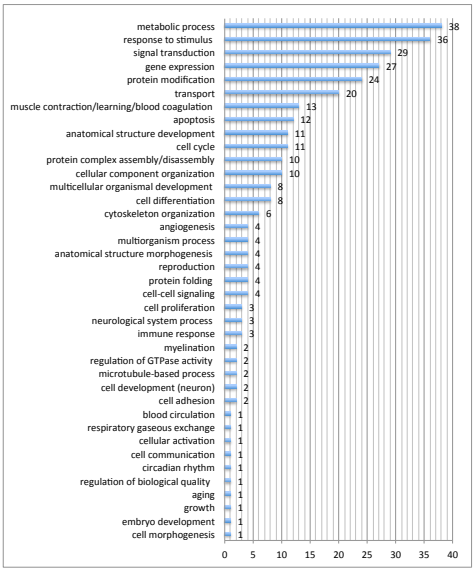

Fig. 3. Distribution of all biological processes from the third gene signature. The number indices the number of gene products that participate to the corresponding biological processes.

Supplementary information accompanies the paper on lab's website (http://www.display.tuc.gr/bcannotstudy/)

Acknowledgement

Present work was supported by "OASYS" project funded by the National Strategic Reference Framework 2007-13 of the Greek Ministry of Development.

References

[1] R. Doe, "Cancer incidence, mortality and prevalence worldwide." 2002. [Online]. Available: http://www-dep.iarc.fr

[2] D. Subramaniam and C. Isaacs, "Utilizing prognostic and predictive factors in breast cancer," *Current Treatment Options in Oncology*, vol. 6, no. 2, pp. 147–159, 2005-04-01.

[3] M. C. Cheang, M. van de Rijn, and T. O. Nielsen, "Gene expression profiling of breast cancer," *Annual Review of Pathology: Mechanisms of Disease*, vol. 3, no. 1, pp. 67–97, 2008.

[4] S. Knudsen, *Cancer diagnostics with DNA microarrays*. Wiley, 2006.

[5] B. E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Computational Learning Theory*, 1992, pp. 144–152.

[6] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[7] B. Ripley, *Pattern recognition and neural networks*. Cambridge Univ Pr, 2008.

[8] M. Blazadonakis and M. Zervakis, "Wrapper filtering criteria via linear neuron and kernel approaches," *Computers in Biology and Medicine*, vol. 38, no. 8, pp. 894–912, 2008.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[11] N. Pal, K. Aguan, A. Sharma, and S. Amari, "Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering," *BMC bioinformatics*, vol. 8, no. 1, p. 5, 2007.

[12] L. van't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart, M. Mao *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer." *Nature*, vol. 415, no. 6871, p. 530, 2002. [Online]. Available: http://www.nature.com/nature/journal/v415/n6871/full/415530a.html

[13] D. Wall, R. Pivovarov, M. Tong, J.-Y. Jung, V. Fusaro, T. DeLuca *et al.*, "Genotator: A disease-agnostic tool for genetic annotation of disease," *BMC Medical Genomics*, vol. 3, no. 1, p. 50, 2010.

[14] S. Carbon, A. Ireland, C. Mungall, S. Shu, B. Marshall, and S. Lewis, "AmiGO: online access to ontology and annotation data," *Bioinformatics*, vol. 25, no. 2, p. 288, 2009.

[15] I. Ferby, M. Reschke, O. Kudlacek, P. Knyazev, G. Pante, K. Amann *et al.*, "Mig6 is a negative regulator of EGF receptor-mediated skin morphogenesis and tumor formation," *Nat. Med.*, vol. 12, pp. 568–573, May 2006.

[16] Y. Yarden and M. X. Sliwkowski, "Untangling the ErbB signalling network," *Nat. Rev. Mol. Cell Biol.*, vol. 2, pp. 127–137, Feb 2001.

[17] X. Zhang, K. A. Pickin, R. Bose, N. Jura, P. A. Cole, and J. Kuriyan, "Inhibition of the EGF receptor by binding of MIG6 to an activating kinase domain interface," *Nature*, vol. 450, pp. 741–744, Nov 2007.

[18] T. Nagashima, R. Ushikoshi-Nakayama, A. Suenaga, K. Ide, N. Yumoto, Y. Naruo *et al.*, "Mutation of epidermal growth factor receptor is associated with MIG6 expression," *FEBS J.*, vol. 276, pp. 5239–5251, Sep 2009.

[19] N. Jin, S. N. Cho, M. G. Raso, I. Wistuba, Y. Smith, Y. Yang *et al.*, "Mig-6 is required for appropriate lung development and to ensure normal adult lung homeostasis," *Development*, vol. 136, pp. 3347–3356, Oct 2009.

[20] T. H. Kim, D. K. Lee, H. L. Franco, J. P. Lydon, and J. W. Jeong, "ERBB receptor feedback inhibitor 1 regulation of estrogen receptor activity is critical for uterine implantation in mice," *Biol. Reprod.*, vol. 82, pp. 706–713, Apr 2010.

[21] C. G. Duncan, P. J. Killela, C. A. Payne, B. Lampson, W. C. Chen, J. Liu *et al.*, "Integrated genomic analyses identify ERRFI1 and TACC3 as glioblastoma-targeted genes," *Oncotarget*, vol. 1, pp. 265–277, Aug 2010.

[22] S. Anastasi, G. Sala, C. Huiping, E. Caprini, G. Russo, S. Iacovelli *et al.*, "Loss of RALT/MIG-6 expression in ERBB2-amplified breast carcinomas enhances ErbB-2 oncogenic potency and favors resistance to Herceptin," *Oncogene*, vol. 24, pp. 4540–4548, Jun 2005.

[23] J. Xu, A. B. Keeton, L. Wu, J. L. Franklin, X. Cao, and J. L. Messina, "Gene 33 inhibits apoptosis of breast cancer cells and increases poly(ADP-ribose) polymerase expression," *Breast Cancer Res. Treat.*, vol. 91, pp. 207–215, Jun 2005.

[24] S. Amatschek, U. Koenig, H. Auer, P. Steinlein, M. Pacher, A. Gruenfelder *et al.*, "Tissue-wide expression profiling using cDNA subtraction and microarrays to identify tumor-specific genes," *Cancer research*, vol. 64, no. 3, p. 844, 2004.