31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Comparison and Unification of Genomic Signatures in Breast Cancer

Michalis E. Blazadonakis and Michalis E. Zervakis, *Member, IEEE*

Abstract— **The concept of deriving a gene signature in breast cancer has been addressed by different research groups, each one proposing a different solution with minor overlap among them. There is still an open issue of unifying results among different research groups. In this study we evaluate two published signatures, namely the 70 gene signature of Netherlands group and a 57 gene signature published in our previous study and propose an evaluation platform under which the underlined signatures could be compared effectively. After such an evaluation, we proceed with a unified signature and assess its performance with improved efficiency over the initial signatures.**

I. Introduction

THE release of the human genome working draft [1] along with the development of DNA microarray technology has opened a new era in the battle against cancer. Using such a technology scientists can derive genomic markers (gene signatures) highly associated with various types of cancer. One widely discussed and accepted [2] [3] signature has been published by the Amsterdam group [4] for breast cancer. Several other signatures have been obtained on the basis of the same dataset, which show little overlap in terms of selected genes [5]. This fact has been previously addressed and explained on the basis of data exploratory issues and algorithmic limitations in view of the large domain of the problem and the small number of samples in the dataset (curse of dimensionality) [6]. In fact, there might be a number of gene sets that result in similar predictive accuracy [5]. Furthermore, testing a signature derived from one dataset on a different dataset reflects significantly decreased prognostic value.

In this study we explore the hypothesis that different signatures, even with little overlap, might carry complementary information that can aid in improving the performance of gene signatures. Towards this direction, we use and evaluate two signatures as presented in section II. We further merge the two signatures and evaluate the prognostic power of the new set of genes. Besides accuracy performance evaluation, we assess a qualitative comparison between all signatures addressing their significance in terms of qualitative characteristics that deal: **a)** with the ability of the expression profile to differentiate its behavior between the two prognostic groups [10] **b)** the statistical significance of the expression profile of a signature under the hypothesis that genes with similar expression pattern are associated with similar outcome and **c)** the survival prediction ability of the underlined signatures. Our aim is to perform an objective and fair comparison that would not necessarily result in a strict ranking of signatures but rather propose additional evaluation criteria under which genomic signatures could be compared in an effective, reliable and objective manner.

M.E.B. and M.E.Z. Authors are with Technical University of Crete, Department of Electronic and Computer Engineering (e-mail: mblazad@ier.forthnet.gr, michalis@display.tuc.gr).

II. Background Knowledge

In this section we provide the background information on the various tools that will be used to assess the comparison between signatures. The quality of signatures is evaluated through the Global test, whereas their predictive ability is assessed through the nearest centroid classifier.

A. Tested Signatures

The Amsterdam signature (denoted by S1) was derived using a training set of 78 patients, 44 belonging to the good prognosis group and 34 to the poor prognosis category (relapse before five years). A three step procedure was applied for the derivation of the final signature. Initially, those genes that had a two fold difference and a p-value less than 0.01 in more than five tumors were selected as more significant, while the ones not satisfying the condition were discarded. In turn, correlation between the prognosis category and the expression value of each gene across samples for all remaining 5000 genes was calculated using Pearson correlation. Among those 5000 genes 231 were found to be significantly regulated with the prognosis groups either in a positive or negative way i.e. genes that gave a Pearson correlation value of less than -0.3 or greater than 0.3. Finally, those 231 genes were ranked in order according to a variation of Fisher's ratio and groups of five genes were repeatedly extracted from this ranked list and were added to the feature list for designing a classifier. The leave one out cross validation method was implemented and classification was based on the correlation of the expression profile of the left out sample with the mean expression profiles of the two prognostic groups formed by the remaining samples. Performance was increasing up to 70 genes, which constitute the final gene signature (S1), giving an 89.47% success rate on the independent test set of 19 samples [4]. Even though the group has received criticism for not cross validating their result [10], the signature was further evaluated on 234 new cases [7] outperforming the performance of the NIH and St. Gallen criteria on Breast

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

Cancer. Besides, the signature has been accepted by FDA [2], while the European Organization for Research and Treatment of Cancer has initiated randomized trials to further strengthen the value of the signature [8].

As a second signature (S2) we consider the 57-gene set published in our previous study [9] that was derived using an external leave one out cross validation procedure using the same set of 78 patients for training purposes as the Amsterdam group. Along each of the 78 runs, we applied the recursive feature elimination procedure based on linear neuron weights (RFE-LNW) [9], [11], while genes were eliminated in a recursive manner, so that surviving ones form the closest power of two. Maximum average performance was measured at the 64 gene cut off point, so that the entire iterative process derives 78 different sets of 64 genes as possible candidates for a final gene signature. Taking their union, an ensemble of 200 different genes is formed Reapplying RFE-LNW and eliminating one gene per iteration we derive a 57-gene signature (denoted by S2) with 89.47% success rate on the independent test set of 19 samples. Notice, that the derived result is directly comparable with that of Amsterdam's group in terms of prediction accuracy while the latter signature consists of a smaller number of genes.

We can point out some important differences related to the philosophies of these two signatures. The Amsterdam's groups applied a drastic preprocessing step that resulted in a 231 set of possible candidate genes, while we derived the candidate set of approximately same size (200 genes) through the use of an external evaluation process [13] with no preprocessing on the initial set of genes. Both methods used Fisher's ratio as the gene ranking criterion. However, in the case of signature S1 it was applied in a static manner, while in S2 was applied in a dynamic approach through the RFE-LNW methodology [9], [11].

B. The Global Test

The global test [14] elaborates on the connection between gene expressions and clinical outcome. If a group of genes can be used to predict the clinical outcome, the gene expression patterns must express different behaviour for different clinical outcomes. Defining now $X=\left[x_{ij}\right]$ as the $n\times m$ data matrix containing the m genes of interest for n samples and Y as the $n\times 1$ clinical outcome vector, we can model the dependence of Y on X. The model in [15] defines an intercept α, a length-m vector of regression coefficients β and a function h such that:

$$E\left(Y_i|\beta\right)=h^{-1}\left(\alpha+\sum_{j=1}^{m}x_{ij}\beta_j\right) \tag{1}$$

Testing whether there is a predictive effect of the gene expression on the clinical outcome is equivalent to testing the hypothesis:

$$H_0:\beta_1=\beta_2=\cdots=\beta_m=0 \tag{2}$$

It can be shown that if H_0 is true the test statistic Q is derived as:

$$Q=\frac{(Y-\mu)'R(Y-\mu)}{\mu_2} \tag{3}$$

where $R=(1/m)XX'$, $\mu=h^{-1}(\alpha)$ is the expectation of Y under H_0 and μ_2 is the second moment of Y under H_0. Hence, Q is a score test which can be interpreted in two alternative ways as follows:

$$Q=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\mu_2}\left[X_i'(Y-\mu)\right]^2 \tag{4}$$

or as:

$$Q=\frac{1}{\mu_2}\sum_{i=1}^{n}\sum_{j=1}^{n}R_{ij}(Y_i-\mu)(Y_j-\mu) \tag{5}$$

Equation (4) indicates that genes with large variance have much more influence on the outcome of the test than genes with low variance. Additionally, equation (5) concentrates on samples and checks whether samples with similar outcomes share also similar gene expression patterns.

C. The Nearest Centroid Classifier (NCC)

For classification purposes we use the nearest centroid prediction rule [16]. Each patient is classified according to the distance between his/her signature and the two average profiles; the predicted class is the one closer to examined profile, by means of the Euclidean distance. Such a classifier can be formulated as follows:

$$f(x)=sign\left((\boldsymbol{x}-\boldsymbol{c})\cdot\boldsymbol{w}\right) \tag{6}$$

where

$$\boldsymbol{c}=\frac{\boldsymbol{c}_+ +\boldsymbol{c}_-}{2} \tag{7}$$

$$\boldsymbol{w}=\boldsymbol{c}_+ -\boldsymbol{c}_- \tag{8}$$

and $\boldsymbol{c}_+$, $\boldsymbol{c}_-$ are the centroids of the positive and negative classes respectively.

III. CLASS COMPARISON AND SURVIVAL PREDICTION ANALYSIS

We assess the classification performance for each signature in the 234 cases published in [7], while the training of NCC was performed on the 78 patient set published in [4]; for a more objective evaluation we removed the 61 patients that were commonly included in the two sets and were considered in the derivation of the two signatures. Classification results are presented in Table 1 where we notice an advantage of S1 signature by three units on the AUC measure. Furthermore, we assess the statistical significance of the two signatures with respect to gene expression value and the prognosis outcome; We use the Q-Test (section II.B) to test whether genes in the two signatures with similar expression patterns also point to same clinical outcome. This result is directly associated to the quality characteristic (**b**) that was stressed out earlier in the introduction and is depicted in Fig. 1 (left part). We notice that signature S1 achieves a higher score on the Q-

Test than S2, and hence it shows better correlation between gene expression and clinical outcome. Signature S2, even though it achieves a smaller test value, it still achieves a high score, demonstrating also significant correlation between gene expression and clinical outcome.

TABLE 1: ACCURACY PERFORMANCE OF THE TWO SIGNATURES ON 234 NEW CASES.

Signature	AUC	SEN	SPE
S1	0.69	0.76	0.62
S2	0.66	0.74	0.59
S1-S2	0.70	**0.80**	0.59

We formulate next, a unified signature composed of the gene union between S1 and S2. This resulted in a new S1-S2 signature of 122 genes (5 genes are common) achieving an AUC performance of 0.7, but also achieving an increased sensitivity (true positive) rate of 80% (Table 1). The specificity decrease of the unified signature is considered minor with respect to the gain in sensitivity.

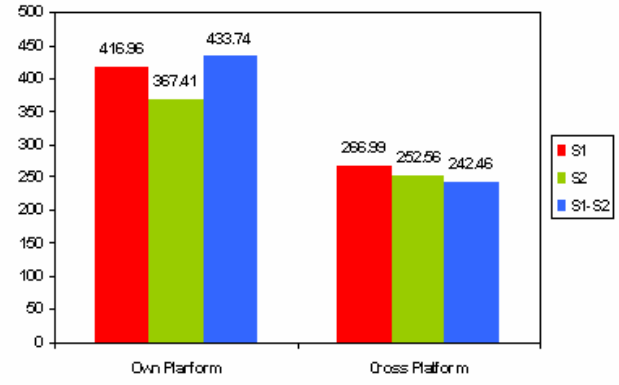

Fig. 1. Significance of the expression profile with respect to the prognosis group.

Analyzing further the implications of the achieved classification measures, the 59% specificity reflects a ratio of 107 recognized cases out of 180 true negatives, while the 80% sensitivity rate maps to 43 correctly categorized out of the 54 true positive cases (11 classified as false negatives). From the set of 118 cases categorized as negatives, 107 are indeed true negatives. In a clinical setup, the implication of this score is that for cases not suggesting chemotherapy the decision is correct with 91% (107/118) probability.

This unified signature also reflects an improvement on the correlation between gene expression and clinical outcome, by achieving a 433.74 Q-Score, higher than its predecessor signatures S1 and S2 (Fig. 1, left part). We also demonstrate the expression profile of the unified 122-gene signature in Fig. 2 (part A). We took the classification result and performed a labeled hierarchical clustering procedure, where the predicted classification result was given as an attribute to the clustering algorithm. Intermingling of samples (columns) is then avoided when clustering is performed. As we observe in Fig. 2 (part A) we notice a substantial difference in the expression level between the green and red group of the clustering result. This finding is directly associated to the quality characteristic (**a**) that was pointed out in the introduction. In addition we used the survival times available with the data set and performed a Kaplan-Meier survival analysis on the classification result. We derived the graph depicted in Fig. 2 (part B), where we observe that there is a substantial gap between the two curves, meaning that the unified signature can effectively discriminate the two prognostic groups. The good prognosis group, green line, which is derived using the survival times corresponding to the green sub-tree patients of the clustering result (Fig. 2 (Part A)), approaches a 12-year survival with a probability of approximately 0.9. On the other hand, the poor prognosis group (red-line) approaches the same survival period but with a probability of less then 0.5. These results demonstrate that the unified gene signature can indeed effectively discriminate between the two prognostic groups, addressing the quality characteristic (**c**) that was pointed out in the introduction.

We proceed one step further by validating the performance of the unified signature on yet another data set derived from a different microarray platform, but also a different experimental design [17]. None of the signatures can be considered as a clinical predictor on this data set, while their significance on the correlation between gene expression and clinical outcome is also significantly decreased (Fig. 1, right part). One possible way to overcome such limitations in cross-platform evaluation studies could be to search for biological knowledge hidden behind the signatures, in terms of biological processes and pathways. Then, instead of combining gene sets, integration should be attempted at the level of biological processes involved, so as to combine knowledge from different sources towards a more global and biologically meaningful solution.

IV. CONCLUSION

In this work we evaluated two gene signatures using various metrics. We used standard success-rate criteria but in addition we assessed the significance of the two signatures based on 'quality' measures such as expression profile analysis through Q-Test and class comparison through the labeled clustering on the classification result; both signatures were derived on the same data set. Even though Amsterdam's signature appears to have a slight advantage in a direct comparison, the unification of these two signatures improves significantly all assessed measures, indicating that both signatures bear complementary information regarding the outcome. Moving one step further towards a cross platform evaluation, the performance of signatures considered drops significantly. One factor that may play a catalytic role on this reduced performance is related to incompatibilities encountered between the different microarray platforms. Different microarray platforms may examine different set of genes belonging to different biological processes or pathways. Research groups then start from different basis which may lead to diverse or

.

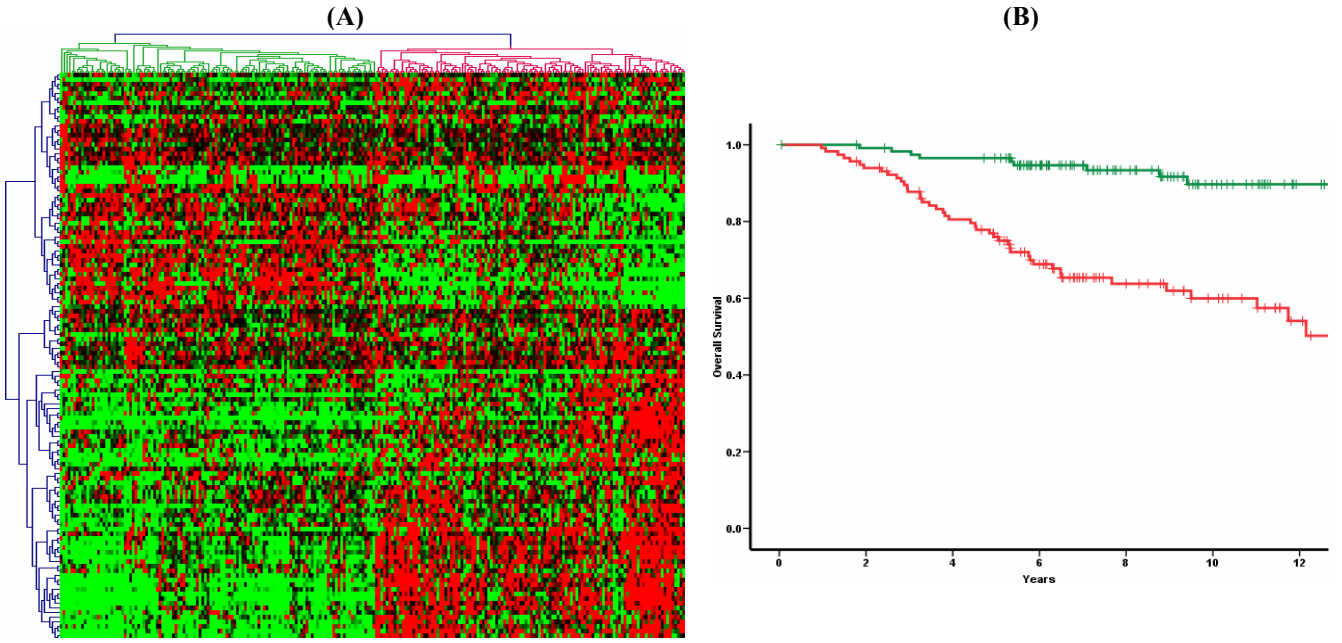

Fig. 2. Class comparison and Kaplan-Meier analysis of the unified signature of 122 genes. In part A, rows correspond to genes columns to patients.

complementary solutions. Another aspect is the difference in the experimental protocols that are used by research groups, while the 'philosophy' of the methodological procedure that is applied to derive a gene signature also plays its role. Normalization of datasets to the same reference, as well as biological knowledge integration, is expected to play catalytic role in improving cross-platform validation results.

ACKNOWLEDGMENT

Research work was supported by EU funded project CRH-BME, no: 144537-TEMPUS-2008-GR-JPCR and the Greek Ministry of Education

REFERENCES

[1] University of California Santa Cruz Genome Bioinformatics, Human Genome Working Draft, http://genome.ucsc.edu, (2001)
[2] http://medgadget.com/archives/2007/02/mammaprint_a_br.html.
[3] http://www.eortc.be/services/unit/mindact/MINDACT_websiteii.asp
[4] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, Y. D. He, et al., "Gene expression profiling predicts clinical outcome of breast cancer", *Letters to Nature* vol. 415 2002, pp. 530-536.
[5] M. Gormley, W. Dampier, A. Ertel, B. Karacali and A. Tozeren, "Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets", BMC Bioinformatics (2007) , 8:415.
[6] L. Ein-Dor, I. Kela, G. Getz, D. Givol and Eytan Domany, "Outcome signature genes in breast cancer: is there a unique set?", Bioinformatics 21 (2005) , pp 171-178.
[7] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer et al., "A gene expression signature as a predictor of survival in breast cancer", *The New England Journal of Medicine,* vol. 347, 2002 pp. 1999-2009.
[8] http://www.eortc.be/services/unit/mindact/MINDACT_websiteii.asp
[9] M. Blazadonakis and M. Zervakis, "The Linear Neuron as Marker Selector and Clinical Predictor in Cancer Gene Analysis", *Computer Methods and Programs in Biomedicine*, vol 91, 2008, pp 22-35.
[10] R. Simon, M. D. Radmacher, K. Dobbin, L. M. McShane, "Pitfalls in the use of DNA Microarray Data for Diagnostic and Prognostic Classification", *Journal of the National Cancer Institute* Vol. 95, 2003 pp.14–18.
[11] M. E. Blazadonakis and M.Zervakis, "Wrapper Filtering Criteria via Linear Neuron and Kernel Approaches", *Computers in Biology and Medicin* , to appear.
[12] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using Support vector machines, *machine learning*, vol. 36, 2002, pp. 389-422.
[13] C. Ambroise, G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data", *Proc. Natl. Acad. Sci. USA,* Vol 99, 2002, pp. 6562-6566.
[14] J. J. Goeman, S. A. Van de Geer, F. de Kort, H.C. Van Houwelingen. "A global test for groups of genes: testing association with a clinical outcome", *Bionformatics*, Vol. 20, 2004, pp. 93-99.
[15] P. McCullagh and J. A. Nelder, "Generalized Linear Models", *Chapman and Hall*, Boca Raton, USA, 1989.
[16] R. Simon, "Diagnostic and prognostic prediction rule using gene expression profiles in high dimensional microarray data", *British Journal of Cancer* Vol. 89, 2003, pp. 1599-1604.
[17] Wang Y., Klijn J. G. M., Zhang Y., Sieuerts A. M. et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer", *the lancet* Vol. 365, 2005, pp. 671-679.