Sequential Application of Feature Selection and Extraction for Predicting Breast Cancer Aggressiveness

Jonatan Taminau[1], Stijn Meganck[1], Cosmin Lazar[1], David Y. Weiss-Solis[2], Alain Coletta[2], Nic Walker[2], Hugues Bersini[2], and Ann Nowé[1]

[1] Computational Modeling Lab, Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium
http://como.vub.ac.be
{jonatan.taminau,stijn.meganck,cosmin.lazar,ann.nowe}@vub.ac.be
[2] IRIDIA, Université Libre de Bruxelles
Avenue Franklin D. Roosevelt 50, 1050 Brussels, Belgium
http://iridia.ulb.ac.be
{david.weiss,alain.coletta,nic.walker,hugues.bersini}@ulb.ac.be

Abstract. Breast cancer is a heterogenous disease with a large variance in prognosis of patients. It is hard to identify patients who would need adjuvant chemotherapy to survive. Using microarray based technology and various feature selection techniques, a number of prognostic gene expression signatures have been proposed recently. It has been shown that these signatures outperform traditional clinical guidelines for estimating prognosis. This paper studies the applicability of state-of-the-art feature extraction methods together with feature selection methods to develop more powerful prognosis estimators. Feature selection is used to remove features not related with the clinical issue investigated. If the resulted dataset is still described by a high number of probes, feature extraction methods can be applied to further reduce the dimension of the data set. In addition we derived six new signatures using three independent data sets, containing in total 610 samples.

Additional information:
http://como.vub.ac.be/~jtaminau/CSBio2010/

Keywords: Breast Cancer Signatures, Feature Selection, Feature Extraction.

1 Introduction

Breast cancer is a very heterogenous disease and it still remains a challenge to distinguish patients who would need adjuvant chemotherapy from those who don't need it. Using microarray based technology, a number of prognostic gene expression signatures have been proposed recently [1,2,3,4] to guide the clinicians with the selection of patients who should receive such treatments. Those signatures,

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 46–57, 2010.

all developed with a different approach and on different data sets, seem to have similar prognostic performance [5,4] despite their limited overlap in genes.

In this paper six new gene signatures are proposed, all based on correlation based analysis with respect to the survival outcome of breast cancer patients. These signatures were derived using several completely independent studies, aiming to increase the generality of the results, and were validated on another independent test set.

A single microarray can contain tens of thousands of probes; the good part in collecting such rich data sets is the fact that one experiment can accomplish many genetic tests in parallel. However, only few probes are relevant for particular clinical issues and their identification is quite difficult even for clinicians or biologists. The analysis of such big data sets is in general subjected to the well known *curse of dimensionality* or *the empty space phenomenon* [6,7]. In order to deal with this challenge, two main strategies can be used: feature selection and feature extraction, each one of them with their specific algorithms. Previous works on the analysis of microarray data have focused mainly on feature selection methods. Basically a number of genes are selected which are meaningful with respect to some clinical features such as disease outcome [1,2] or histological grade [3], but the dimension of the resulting set of genes is still high, typically around 100 probes or genes. On the other hand, feature extraction methods are mainly used to improve the results of the analysis in situations where one deals with high dimensional data sets and no *a priori* information about the data is known [8]. The result of feature extraction methods is a new representation of the original data in a compressed form, by minimizing the loss of information and by preserving the distribution of the original data.

Sequential application of feature selection and feature extraction methods can also be used for dimension reduction especially when the selection of features still results in a high dimensional data set. In a first instance, feature selection is used to remove those features which are not related with the clinical issue investigated but also features carrying low information (features with low variance). If the resulted data set is still described by a high number of features, extraction methods can be applied to further reduce the dimension of the data set. The increase in prognostic accuracy after applying feature extraction methods is shown both in combination with existing signatures and with our six newly proposed ones.

2 Methods

In this section we describe the data, the different gene signatures and the feature extraction methods used in this study.

2.1 Data

We used a collection of three different and independent breast cancer data sets which were retrieved from the InSilico DB[1] in order to consistently pre-process

[1] `http://insilico.ulb.ac.be/` (manuscript in progress)

Table 1. Different microarray data sets used in this study. The first three data sets are the *training* data from which the CoMo signatures are derived. The fourth data set can be seen as an independent *test* set for validation.

GEO Acc. *InSilicoDB Acc.*	*# Samples*	*Author*	*Ref.*
GSE1456	159	Pawitan	[10]
GSE3494	251	Miller	[11]
GSE11121	200	Schmidt	[12]
GSE7390	198	Desmedt	[9]

and annotate them. As our independent validation set we selected the TRANSBIG data set because it was already used before to test and compare different breast cancer gene expression signatures [5]. For consistency we only selected the Affymetrix study (TBVDX serie [9]).

All four studies, listed in Table 2.1, were computed on the Affymetrix HG-U133 research GeneChip™ and we checked for duplicated samples between all studies by looking at the correlation across samples. No sample pairs with a correlation higher than 0.95 were observed, ensuring the independence of all the different data sets.

2.2 Existing Gene Signatures

Several different prognostic gene signatures for breast cancer aggressiveness have been proposed in recent years [1,2,3]. They have been shown to be advantageous compared to standard clinical guidelines and could therefore reduce the number of patients subject to adjuvant chemotherapy.

Gene70. In [1], genes were identified that were differentially expressed between two groups of patients with differing survival. They used a cut-off of 5 years after diagnosis to check whether someone had developed distant metastasis or not, and divided the patients in two groups accordingly. They used microarrays of Agilent technology and identified a set of 70 relevant probes.

Gene76. With a similar method the Erasmus Medical Center, Netherlands and Veridex LLC, USA identified a set of 76 probes [2] using Affymetrix microarrays. These genes were used to build a risk prediction model taking into account the difference between estrogen receptor positive (ER+) and negative (ER-) patients.

GGI. In [3], they proposed a gene signature that is predictive for the histological grade of the tumor. Since the histological grade is highly correlated with predictive outcome, this can be used as a prognostic signature. Probes were selected based on their ranking with respect to their differential expression between histological grade 1 and 3. GGI was also derived using Affymetrix microarrays.

2.3 CoMo-Signatures

The existing gene signatures mentioned above have few overlap between them and the number of selected genes is quite high (tens of probes). Moreover, these signatures have been selected from a single microarray data set and tested in some cases only on few samples which limits their power to generalize. With the public availability of well annotated large breast cancer data sets in the InSilico DB, new strategies for discovering gene signatures can be executed. In this paper we derive new gene signatures using information from three public microarray data sets and validate them on a fourth independent data set. Correlation based techniques were used to develop these signatures as follows.

For each of the three training data sets individually, the correlation of each gene with the survival time was computed. All genes were ranked according to their correlation and six gene signatures were created as follows:

Intersect Low: The intersection of the top 500 negatively correlated probes per individual training data set (8 probes).
Intersect High: The intersection of the top 500 positively correlated probes per individual training data set (13 probes).
Intersect Both: The union of *Intersect low* and *Intersect high*.
Multiple Low: Any probe that appeared in the top 500 negatively correlated probes in at least two training data sets (256 probes).
Multiple High: Any probe that appeared in the top 500 positively correlated probes in at least two training data sets (152 probes).
Multiple Both: The union of *Multiple low* and *Multiple high*.

All details of the six gene signatures can be found in additional information. The use of three completely independent data sets should offer more robust signatures because it decreases the risk of overfitting by combining different sources of the same signals. However, missing or incorrect probes from a defective study will not show up in a strict intersection, regardless their overall importance in the other studies. Therefore, we developed signatures with two different strategies, the probes in the *Intersect* signatures capture relevant information in all data sets and are therefore assumed to be very important for estimating survival. The *Multiple* signatures are less prone to the absence of relevant probes in a specific study but the size of these signatures grows rapidly.

2.4 Feature Extraction

Following, we describe two different simple feature extraction methods we used in this paper. We focused on PCA and ICA since they are both well understood techniques which have proven to be useful in many applications. PCA is perhaps the most popular feature extraction technique used in a wide range of applications such as face recognition [13], multivariate image segmentation or multidimensional data clustering. On the other side, ICA has only been recently used as a feature extraction tool: it has been successfully applied in applications such as target detection from multi/hyperspectral remote sensing images

[14] and [15] and more recently in bioinformatics [16]. In [17] PCA and ICA are investigated as feature extraction tools for brain tumor classification.

Principal Component Analysis (PCA). PCA consists in finding the eigenvectors as well as the eigenvalues of the covariance matrix of the original gene-expression matrix X [18]. Then the principal components are obtained by projecting every sample from X on the eigenvectors with the highest eigenvalues.

$$S = BX \tag{1}$$

where B is the eigenvectors matrix of the covariance matrix of X and S are the principal components. The dimension reduction is performed by choosing those eigenvectors whose corresponding eigenvalues are greater than a fixed threshold. A general and comprehensive description of the method can be found in [18].

Independent Component Analysis (ICA). ICA is similar to PCA. The orthogonality assumption, inherent to the eigenvectors, is replaced with that of independence between the newly derived features and basis vectors [19]. It can also be employed as a dimension reduction method and it sometimes embeds PCA as a preprocessing step. The independence of two random variables is expressed in various ways explaining the big number of algorithms developed for ICA. In our simulations we used the FastICA algorithm [19]. In [16], the use of ICA is motivated by the fact that the expression of genes is the result of a specific combination of cellular variables. ICA has been used to derive a linear model based on hidden variables called expression modes and the expression of each gene is a linear function of those modes. A recent survey of the use of ICA for feature extraction from microarray data can be found in [20].

2.5 Combining Feature Selection and Feature Extraction

Our goal is to look at the combination of feature selection and feature extraction. Although there is a lot of work on using feature extraction methods on entire microarray data sets [16,21], we believe that combining both methods will have a positive impact on the overall results. We motivate our believe by the fact that a microarray data set encodes rich and various information about many aspects of the cell such as functions, states or processes taking place inside it, some of them being more dominant than another (for instance ER status). This is the reason why a first selection of probes, the most relevant with respect to a particular issue in question (in our case the breast cancer aggressiveness) is mandatory.

Extracted features are harder to interpret in the sense that a particular value of a probe is a linear (or non linear) combination of some basis vectors resulted in the feature extraction process. This is the main drawback of these methods, in the sense that a particular value of a probe is a linear (or non linear) combination of some basis vectors resulted in the feature extraction process. By combining both approaches we render interpretation easier since the extracted features depend on a limited number of probes.

In our approach, we start by only keeping probes inside the signature (both the existing as the CoMo ones) and calculate a set of features from this data set by using both PCA and ICA, resulting in two extra transformed data sets per training set and also for the TRANSBIG test data set. On every data set we then perform a k-means clustering. The assumption is that the set of probes will naturally cluster samples based on their expected survival time. Furthermore, we expect that reducing the noise and dimension by feature extraction might help improve these results. We then perform Kaplan-Meier and Cox proportional hazard ratio analysis for the groups that were identified by the clustering algorithm.

3 Results

In this section we describe our experimental setup and the survival analysis results on all training data sets as well as on the independent TRANSBIG test data set.

3.1 Experimental Setup

All code was written in $\boldsymbol{R}$[2]. All data sets were downloaded from the InSilico DB, which automatically retrieved the original CEL files, performed RMA for each individual data set and normalized them between data sets using Batch Mean Centering [22].

Feature selection, which in this case amounts to probe selection, for the CoMo-signatures was done as explained above. For both the *GGI* and *Gene*76 signatures the *genefu* package[3] was used to identify the probes available in $data(sig.ggi[\text{“}probe\text{”}])$ and $data(sig.gene76[\text{“}probe\text{”}])$ respectively. We did not include the *Gene*70 signature in our analysis since this study was not performed on Affymetrix arrays.

The basic *prcomp* function and $fastICA$ package[4] were used to perform PCA and ICA feature extraction respectively. The number of components for PCA was chosen by removing those eigenvectors whose standard deviation was less than a fifth of that of the first eigenvector. The number of components of ICA was chosen to be the same as that of PCA.

For the CoMo-signatures samples were divided into two groups based on k-means clustering with correlation as a distance metric by using the $Kmeans$ function of the *amap* package[5] with five random restarts. For GGI and Gene76 a similar division was done and also one using the risk score as provided in the *genefu* package.

3.2 Analysis

We performed a cox proportional hazard ratio (HR) analysis on all data sets for each of the six new gene signatures. We performed this analysis on all training

[2] www.r-project.org/
[3] http://cran.r-project.org/web/packages/genefu/
[4] http://cran.r-project.org/web/packages/fastICA/index.html
[5] http://cran.r-project.org/web/packages/amap/index.html

data sets as well to see whether the signatures still captured essential information of the data sets that they were derived from. These results however can give an overly positive view as all information of these data sets was used in the derivation of the signatures.

As can be seen in Table 3.2, for the training data sets, in all but one case the best results (higher HR) are obtained after sequentially performing feature selection and feature extraction. Full details, including 95% confidence intervals and p-values can be found in additional material.

Clearly the highest HR is obtained when using the signatures composed of probes which are both negative and positively correlated with survival which shows that these two sets are complementary for the estimation of aggressiveness.

On the TRANSBIG data set, we see similarly that feature extraction aids the division in good and bad prognosis classes, thereby affirming the results on the training data sets. The improvements are however not always as significant. Applying our strategy on both GGI and Gene76, see Table 3.2, shows similar

Table 2. Hazard ratio's using the different gene signatures on both training and test data sets. The bold numbers indicate the best (highest HR) per data set/gene signature combination.

Signature	*Data set*	*Original*	*PCA*	*ICA*
Intersect low	GSE1456	3.498072	**5.025954**	2.953581
	GSE3494	1.483355	**2.548084**	1.303337
	GSE11121	1.717594	**3.265009**	1.281384
Intersect high	GSE1456	2.803051	**5.197641**	3.061206
	GSE3494	2.086769	**3.082308**	2.225119
	GSE11121	1.977458	**2.633595**	1.987610
Intersect both	GSE1456	4.555611	**6.287580**	4.811514
	GSE3494	2.582587	2.981011	**3.006106**
	GSE11121	3.192547	2.072678	**3.742142**
Multiple low	GSE1456	2.884515	**6.421448**	4.398484
	GSE3494	2.450127	**2.565744**	2.373649
	GSE11121	2.113939	**2.863432**	1.943166
Multiple high	GSE1456	3.836090	**7.170804**	3.913145
	GSE3494	**3.597214**	3.435415	3.263882
	GSE11121	2.397085	1.935239	**2.705435**
Multiple both	GSE1456	8.054912	7.564513	**8.314927**
	GSE3494	2.950295	2.628732	**3.089008**
	GSE11121	3.274170	2.870842	**3.327138**
Intersect low	TRANSBIG	2.184360	**2.767724**	1.219542
Intersect high	TRANSBIG	1.774095	**2.065701**	1.610812
Intersect both	TRANSBIG	2.737137	**2.870929**	2.793816
Multiple low	TRANSBIG	4.968810	4.904974	**5.047237**
Multiple high	TRANSBIG	1.037704	**1.910830**	1.432156
Multiple both	TRANSBIG	3.693934	3.041489	**4.947116**

Table 3. Hazard ratio's using GGI and Gene76 on the test data set. The bold numbers indicate the best (highest HR) per data set/gene signature combination using the k-means approach. The first column shows the HRs based on the signature score and risk classification from GGI and Gene76 respectively.

Signature	*Risk Factor*	*Original*	*PCA*	*ICA*
GGI	*5.091227*	**3.816467**	3.414592	1.438335
Gene76	*5.057281*	2.245084	**3.512873**	3.210784

results although the division based on k-means for GGI outperforms those with feature extraction. With several of the CoMo-signatures we are able to obtain higher HRs than with GGI and Gene76 with our approach. However, none of the results can improve the HR of GGI and Gene76 based on their own risk score.

We also performed a Kaplan-Meier (KM) analysis for each data set/gene signature combination. We show the results for the training sets for the *Intersect low* signature in Figure 1, the other figures can be found online as additional material. Results of both *Intersect low* and *Multiple low* on the TRANSBIG data set can be found in Figure 2. These figures map the corresponding results that were discussed previously for the hazard ratios in Table 3.2.

We created a single sample predictor (SSP) for each of the CoMo-signatures. The SSP was created based on the merged data set combining the three test data sets using Batch Mean Centering [22] by taking the mean for each probe/component for samples with metastasis and samples without. We then assigned each sample in the TRANSBIG data set to a group based on the highest correlation with each of the SSPs. The resulting HRs are given in Table 3.2 and corresponding KM plots can be found in additional information.

The SSP based on the *Multiple Low* signature outperforms both the existing GGI and Gene76 signatures and all the divisions based on k-means of all CoMo-signatures on the TRANSBIG data set.

3.3 Discussion on Feature Extraction Methods

PCA tends to work better in most cases than any of the other. ICA seems to perform poorly in general but works good on the largest collection of genes.

Table 4. Hazard ratios for the TRANSBIG data set based on the SSP created on the merged data set of all three training data sets for all CoMo-signatures

Feat. Extr. Method	*Intersect Low*	*Intersect High*	*Intersect Both*	*Multiple Low*	*Multiple High*	*Multiple Both*
None	2.726886	1.719567	2.489592	**8.988187**	2.728809	3.379779
PCA	2.685891	1.898967	2.328386	**6.276682**	2.345202	3.757589
ICA	1.642725	1.841632	2.727805	**7.193519**	2.855434	3.757589

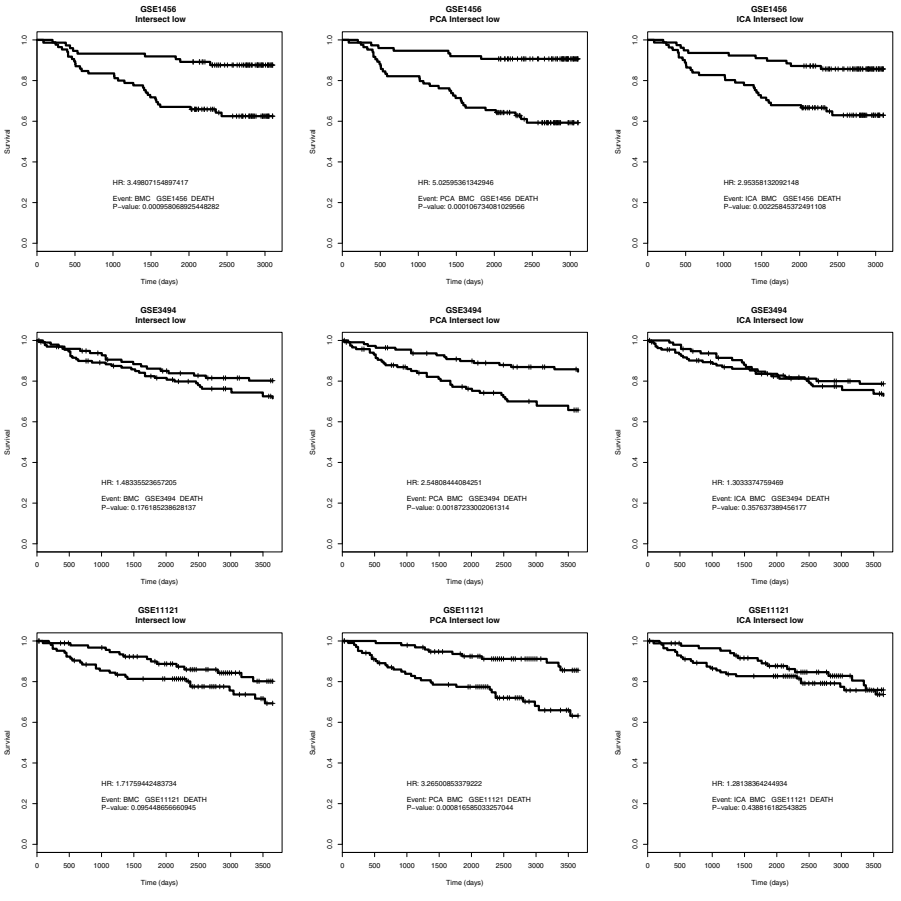

Fig. 1. Comparison of Kaplan-Meier analysis for *Intersect Low* signature. Columns indicate the different feature extraction algorithm used: *none, PCA*, and *ICA* respectively. Rows correspond to the different training data sets: *GSE1456*, *GSE3494* and *GSE11121* respectively.

While a PCA decomposition preserves in the best way possible the original data by implicitly imposing the minimum loss of information, the independence constraints imposed by ICA might give completely misleading results depending on the application. This is why it is not surprising that PCA performs in general better than ICA in this particular application. It is not yet fully understood why feature selection + ICA performs better for gene signatures containing a higher number of probes. One explanation could be the fact that gene signatures containing few probes have relatively independent features and thus the ICA can not improve the results significantly; more than that, removing some features might result in a significant loss of information. On the other hand, the large gene signatures are likely to have several significant correlated features and thus ICA can make a significant improvement.

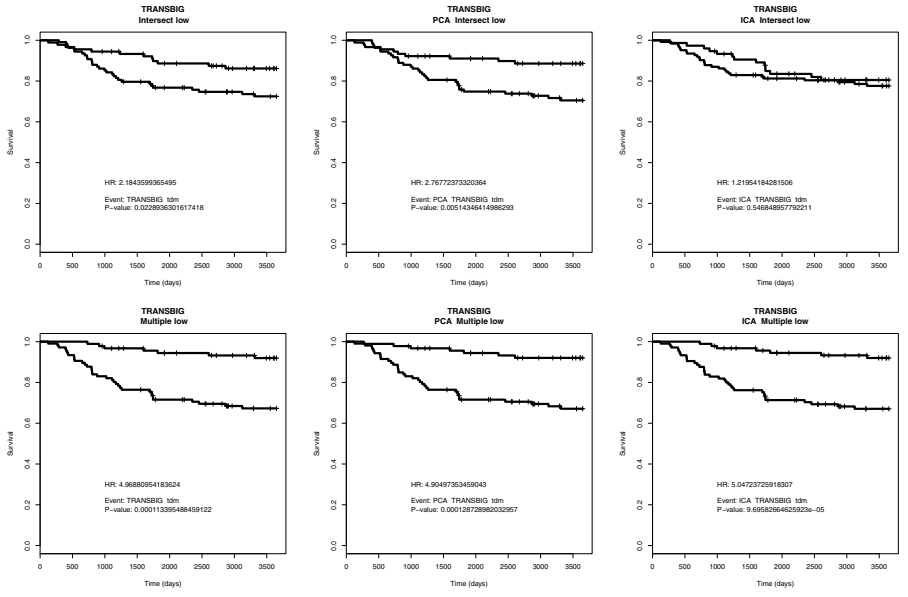

Fig. 2. Comparison of Kaplan-Meier analysis for several CoMo-signatures on the TRANSBIG data set. Columns indicate the different feature extraction algorithm used: *none, PCA*, and *ICA* respectively. Rows correspond to the different signatures: *Intersect Low* and *Multiple Low* respectively.

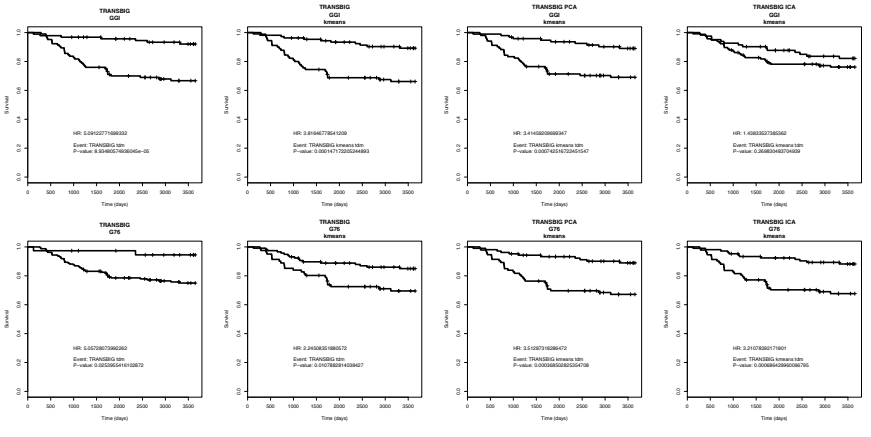

Fig. 3. Comparison of Kaplan-Meier analysis for the TRANSBIG data set for both the GGI and Gene76 signatures. The first column uses the risk factor of the original signals to form risk groups. The following three columns indicate the different feature extraction algorithm used: *none, PCA*, and *ICA* respectively. Rows correspond to the different signatures: *GGI* and *Gene76* respectively.

4 Conclusions

Here we investigate the beneficial aspects of jointly use feature selection and feature extraction methods for breast cancer aggressiveness prediction. For feature selection we have used already existing gene signatures (GGI and Gene76) derived to predict breast cancer aggressiveness but we also derived our own gene signatures from three independent data sets. Further, these signatures have been used as inputs for feature extraction (PCA and ICA) aiming the dimension reduction and prediction improvement. Results show that for this application PCA applied on gene signatures improves the breast cancer aggressiveness prediction in most of the cases.

The promising results of our newly created signatures encourage us to further investigate the use of information of multiple studies in order to derive consistent, robust and predictive signatures.

Acknowledgements

This research is partially funded by the Institute for the encouragement of Scientific Research and Innovation of Brussels (IRSIB).

References

1. van 't Veer, L.J., Dai, H., van de Vijver, M.J., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871), 530–536 (2002)
2. Wang, Y., Klijn, J.G.M., Zhang, Y., et al.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365(9460), 671–679 (2005)
3. Sotiriou, C., Wirapati, P., Loi, S., et al.: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. 98(4), 262–272 (2006)
4. Korkola, J.E., Blaveri, E., DeVries, S., et al.: Identification of a robust gene signature that predicts breast cancer outcome in independent data sets. BMC Cancer 7, 61 (2007)
5. Haibe-Kains, B., Desmedt, C., Piette, F., et al.: Comparison of prognostic gene expression signatures for breast cancer. BMC Genomics 9, 394 (2008)
6. Scott, D., Thompson, J.: Probability density estimation in higher dimensions. In: Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface (1983)
7. Somorjai, R.L., Dolenko, B., Baumgartner, R.: Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 19(12), 1484–1491 (2003)
8. Bild, A.H., Yao, G., Chang, J.T., et al.: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439(7074), 353–357 (2006)
9. Desmedt, C., Piette, F., Loi, S., et al.: Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. Clin. Cancer Res. 13(11), 3207–3214 (2007)

10. Pawitan, Y., Bjöhle, J., Amler, L., et al.: Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. 7(6), R953–R964 (2005)
11. Miller, L.D., Smeds, J., George, J., et al.: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc. Natl. Acad. Sci. USA 102(38), 13550–13555 (2005)
12. Schmidt, M., Böhm, D., von Törne, C., et al.: The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res. 68(13), 5405–5413 (2008)
13. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neurosci. 1(3), 71–86 (1991)
14. Chiang, S.S., Chang, C.I.: Unsupervised hyperspectral image analysis using independent component analysis. In: IEEE International Geoscience and Remote Sensing Symposium, vol. 1(7), pp. 3136–3138 (July 2000)
15. Robila, S.A., Varshney, P.K.: Target detection in hyperspectral images based on independent component analysis. In: SPIE AeroSense, Orlando, Florida, USA, vol. 1(7), pp. 3136–3138 (April 2002)
16. Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. Bioinformatics (January 2002)
17. Luts, J., Poullet, J.B., Garcia-Gomez, J.M., et al.: Effect of feature extraction for brain tumor classification based on short echo time 1h mr spectra. Magn. Reson. Med. 60(2), 288–298 (2008)
18. Jolliffe, I.: Principal component analysis. Springer Series in Statistics (2002)
19. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. 13(4-5), 411–430 (2000)
20. Kong, W., Vanderburg, C.R., Gunshin, H., et al.: A review of independent component analysis application to microarray gene expression data. BioTechniques 45(5), 501–520 (2008)
21. Alter, O., Brown, P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA 97(18), 10101–10106 (2000)
22. Sims, A.H., Smethurst, G.J., Hey, Y., et al.: The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. BMC Medical Genomics 1, 42 (2008)