Model Selection in Meta-Analytical Framework for Prototype-Based Clustering

B. Haibe-Kains[1,2], C. Desmedt[2], C. Sotiriou[2], and G. Bontempi[1]

[1]Functional Genomic Unit, Institut Jules Bordet, Brussels, Belgium
[2]Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium

1. Introduction

The microarray technology allows biologists and doctors to measure the expression of several thousand of genes in parallel. Each microarray chip uses up to 50,000 probes to measure these expressions, generating a huge amount of data to analyze. The high cost of microarray experiments and the coexistence of different microarray technologies make difficult the generation of datasets with large number of samples. In addition to the high feature-to-sample ratio, these data are characterized by the high correlation of coexpressed genes and the high level of noise due to complex technology, making the analysis of microarray data a complex task.

The dimension reduction is widely used in microarray analysis. The inclusion of *a priori* biological knowledge could both reduce the complexity and improve the performance of this important step. This *a priori* knowledge could be a list of key biological processes in the biological problem under study. The role of dimension reduction in this case would be to compute from the original microarray data, new features which quantify as well as possible each of these biological processes.

Clustering analysis is widely used to perform dimension reduction, keeping the new features interpretable. This method consists in replacing a cluster of correlated genes by a cluster centroid (called feature). We aimed at efficiently using *a priori* biological knowledge to improve clustering methodology for dimension reduction.

2. Materials

Gene expression data were measured from breast cancer (BC) tissues and were retrieved from public databases or authors' website. We used normalized data (log2 intensity in single-channel platforms or log2 ratio in dual-channel platforms) as published by the original studies. The Table 1 gives the characteristics of all the datasets used for the analysis.

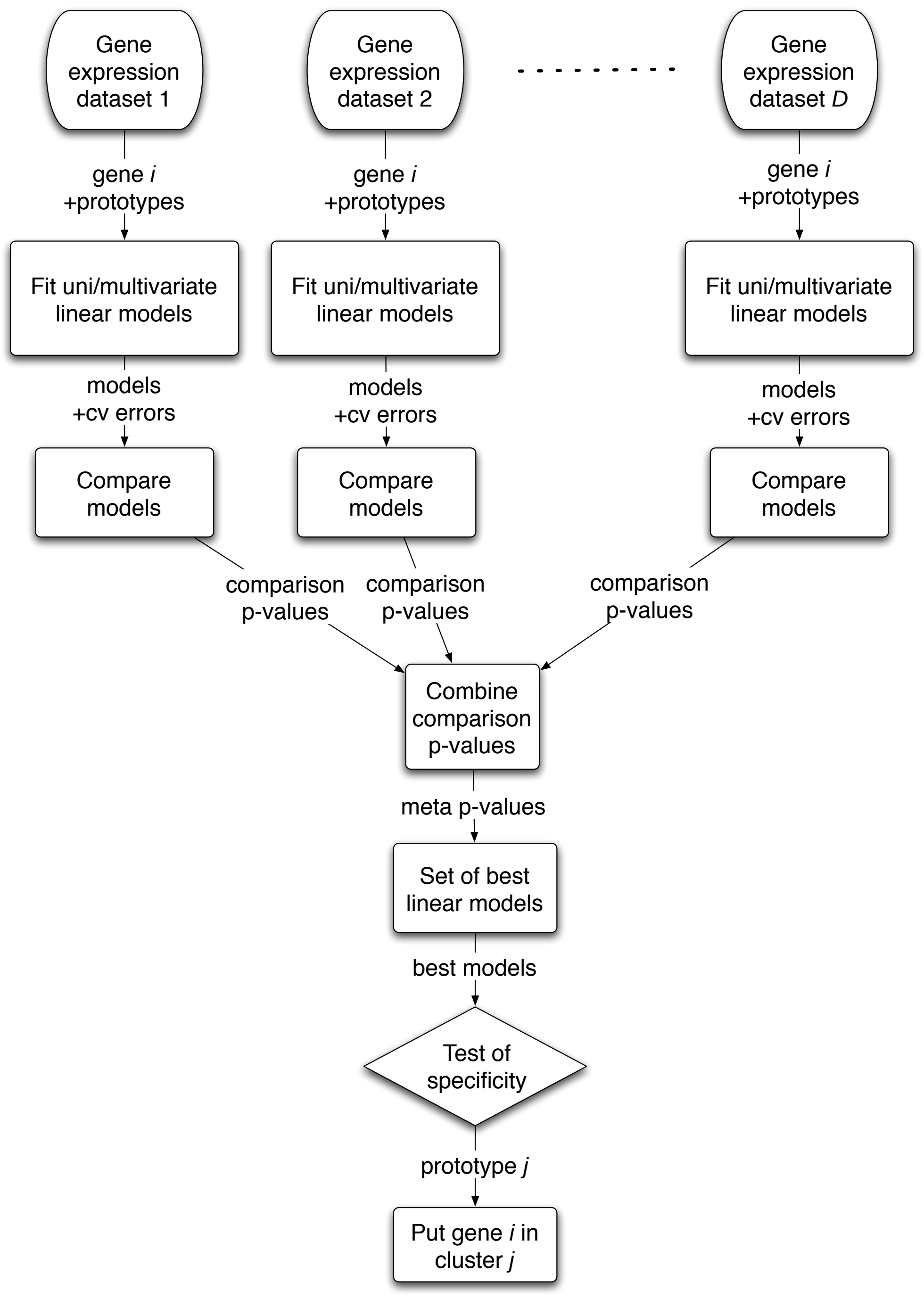
Article	Dataset	Size
van't Veer *et al*, 2002	NKI	117
van de Vijver *et al*, 2002	NKI2	295
Sorlie *et al*, 2003	STNO2	122
Sotiriou *et al*, 2003	NCI	99
Ma *et al*, 2004	MGH	60
Miller *et al*, 2005	UPP	251
Pawitan *et al*, 2005	STK	159
Wang *et al*, 2005	VDX	286
Foekens *et al*, 2006	VDX2	180
Sotiriou *et al*, 2006	UNT	137
Oh *et al*, 2006	UNC	153
Buyse *et al*, 2006	TRANSBIG	307
Desmedt *et al*, 2007	TRANSBIG	198
Naderi *et al*, 2007	NCH	135
Loi *et al*, 2007	TAM	255

Table 1: *Gene expression datasets of breast cancer patients.*

3. Methods

In order to represent the biological processes of interest, we selected one gene (called *prototype*) per process. These genes were selected from literature and biological databases. We introduced a new method called *prototype-based clustering* to identify genes that are specifically coexpressed with one prototype, i.e. genes that can be predicted using only one prototype. For each gene to cluster, we fitted univariate and multivariate linear models with the prototypes which play the role of explanatory variables. We compared these models based on their leave-one-out cross-validation (cv) error computed by the PRESS statistic. Using Friedman's test, the models exhibiting the lowest cv errors were selected to test the specificity of the gene. If only one univariate model (with prototype *j*) remains in the set of best model, this gene is put in cluster *j*. Otherwise, the gene is discarded from analysis because it can be predicted similarly by several univariate models or a multivariate model. The Figure 1 sketches the different steps of the method.

This method was used in a meta-analytical framework in order to combine model selection from different datasets.



Gene expression dataset 1
Gene expression dataset 2
Gene expression dataset *D*
gene *i* +prototypes
gene *i* +prototypes
gene *i* +prototypes
Fit uni/multivariate linear models
Fit uni/multivariate linear models
Fit uni/multivariate linear models
models +cv errors
models +cv errors
models +cv errors
Compare models
Compare models
Compare models
comparison p-values
comparison p-values
comparison p-values
Combine comparison p-values
meta p-values
Set of best linear models
best models
Test of specificity
prototype *j*
Put gene *i* in cluster *j*

Figure 1: *Prototype-based clustering method in a meta-analytical framework.*

Once the clusters were identified, we computed the cluster centroids (signed average of expressions of genes in the cluster) for each sample. In order to assess the relevance of these new features, we used them 1) to define robustly BC molecular subtypes and 2) to investigate the impact of the new features on clinical outcome. These two questions were addressed in using public microarray datasets given in Table 1 (≈ 2100 patients).

4. Results

We applied our method to two large public microarray datasets of BC patients (NKI2 and VDX). These datasets came from two different microarray technologies. We used hallmarks of BC involving various biological processes such as estrogen receptor (ESR1), her2/neu signaling (ERBB2), proliferation (AURKA), tumor invasion (PLAU), immune response (STAT1), angiogenesis (VEGF), and apoptosis (CASP3) as prototype genes. We reduced the number of variables from ≈ 20,000 to seven in keeping valuable information for BC subtyping and prognostication.

4.1 Breast Cancer Subtypes

Several publications reported that BC are molecularly heterogeneous [1, 4], mainly in terms of estrogen receptor and her2/neu signaling. Using the features associated to these two biological processes (ESR1 and ERBB2 respectively), we identified BC subtypes in all the datasets (see Table 1) using a simple model-based clustering. The BIC criterion was used to estimate the most likely number of clusters and we consistently found three well conserved patterns (called subtype 1, 2 and 3), confirming the past studies (see Figure 2).

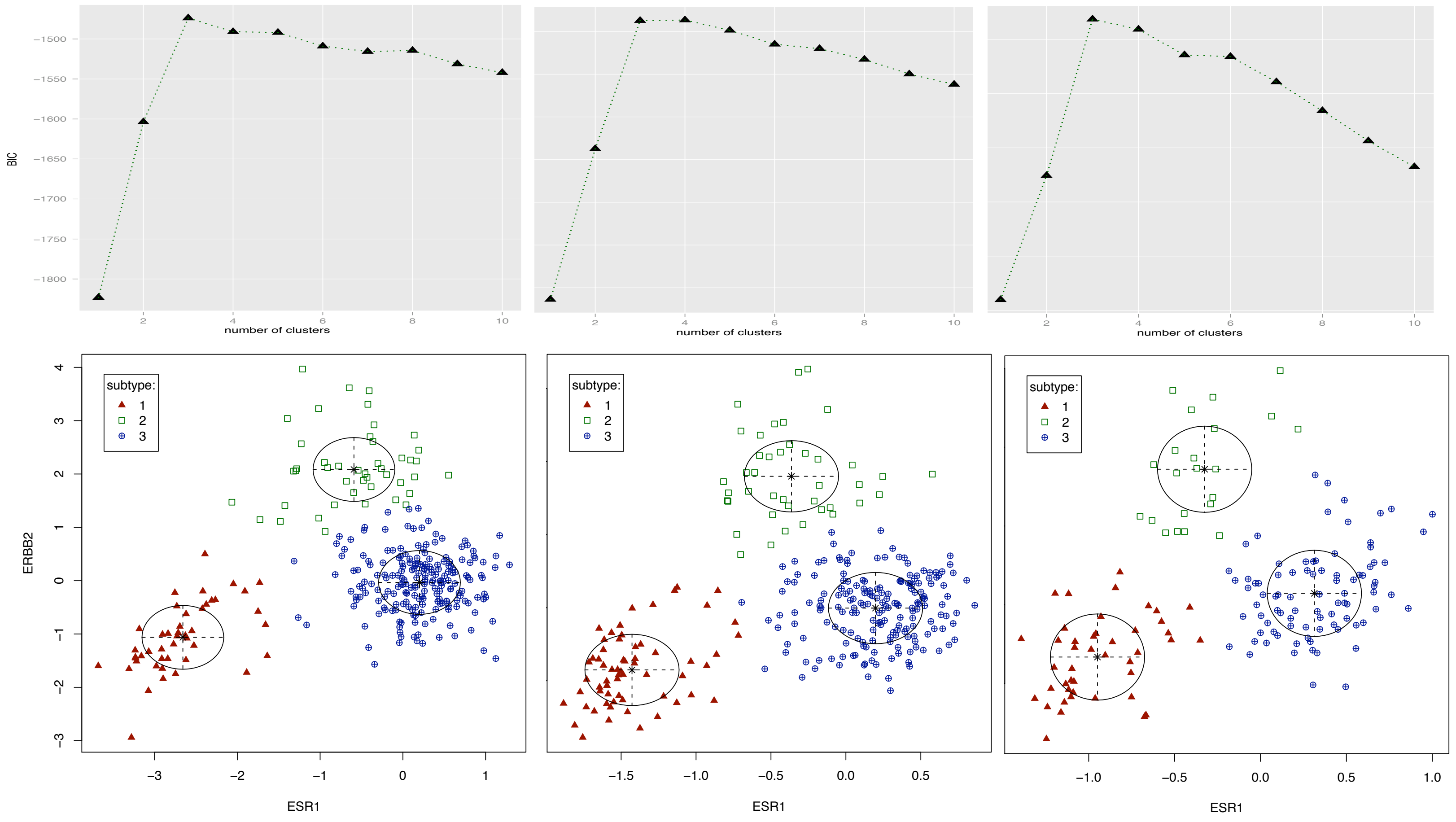


Figure 2: *BC subtypes in NKI2, VDX and UNC datasets respectively.*

4.2 Breast Cancer Prognostication

We performed a meta-analysis on all the datasets (see Table 1), considering only untreated patients. The most relevant features depend on the subtype, AURKA being highly prognostic in subtype 3 and STAT1 being moderately prognostic in subtype 1 and 2 (see Figure 3).

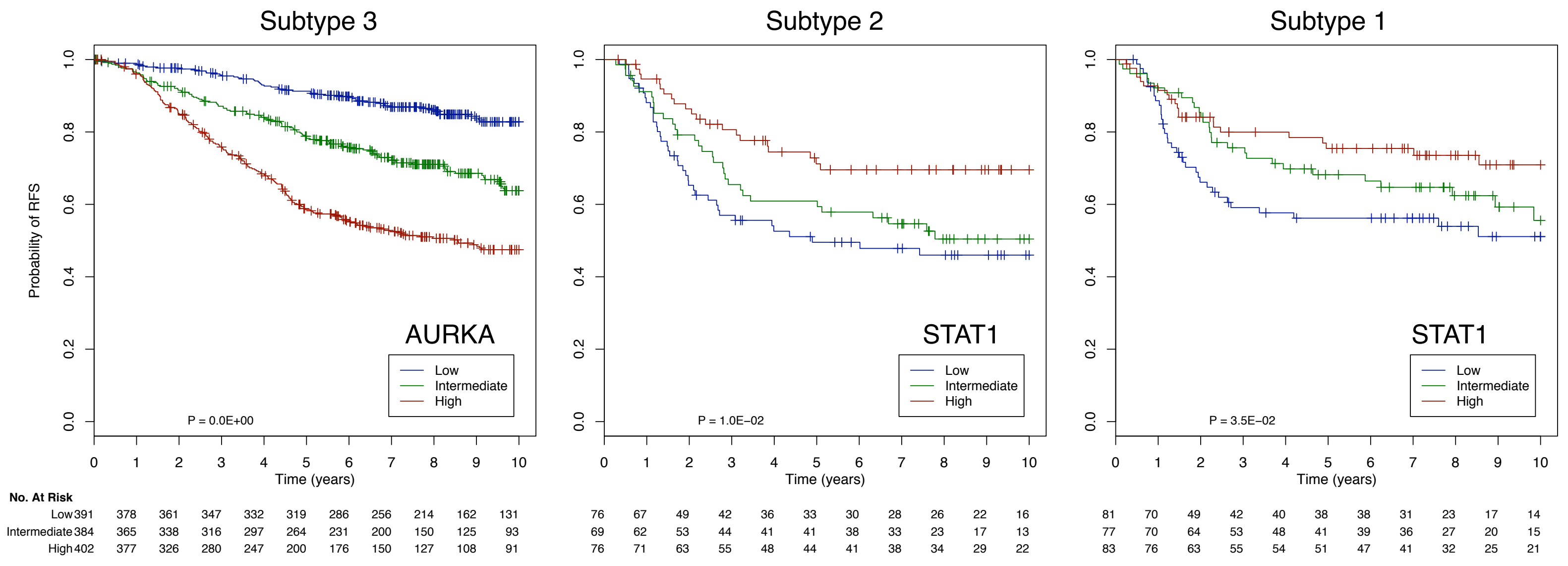


Figure 3: *Kaplan-Meier curves for most relevant features in each BC subtype.*

5. Conclusion

The use of prototype-based clustering allowed for efficient reduction of the dimensionality of microarray data in focusing on target biologically processes. We successfully applied this method to BC samples in order to gain new insights into BC biology.

This work was presented in [2, 3].

References

[1] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geister, J. Demeter, C. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in indepedent gene expression data sets. *Proc Natl Acad Sci USA*, 1(14):8418–8423, 2003.

[2] C. Sotiriou, C. Desmedt, B. Haibe-Kains, A. Harris, D. Larsimont, M. Buyse , P. Wirapati, M. Delorenzi, G. Bontempi, and M. J. Piccart. Biological mechanisms that trigger breast cancer (bc) tumor progresion are molecular subtype dependent. In *American Society of Clinical Oncology*, number 10581, 2007.

[3] C. Sotiriou, B. Haibe-Kains, C. Desmedt, P. Wirapati, V. Durbecq, A. Harris, D. Larsimont, G. Bontempi, M. Buyse, M. Delorenzi, and M. Piccart. Comprehensive molecular analysis of several prognostic signatures using molecular indices related to hallmarks of breast cancer: proliferation index appears to be the most significant component of all signatures. In Springer, editor, *Breast Cancer Research and Treatment*, volume 100, page S86, 2006.

[4] C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S.B. Fox, A. L. Harris, and E. T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci.*, 100(18):10393–10398, 2003.

Emerging Design and Analysis Issues in Genomics Studies in Population Sciences