A Predictive Based Regression Algorithm for Gene Network Selection

Stéphane Guerrier[1] ⋆, Nabil Mili[2] ⋆, Roberto Molinari[2],
Samuel Orso[2], Marco Avella-Medina[2] & Yanyuan Ma[3]

[1]Department of Statistics
University of Illinois at Urbana-Champaign, USA
Email: stephane@illinois.edu

[2]Research Center for Statistics
Geneva School of Economics and Management
University of Geneva, Switzerland

[3]Department of Statistics
University of South Carolina, USA

Abstract

Gene selection has become a common task in most gene expression studies. The objective of such research is often to identify the smallest possible set of genes that can still achieve good predictive performance. To do so, many of the recently proposed classification methods require some form of dimension-reduction of the problem which finally provide a single model as an output and, in most cases, rely on the likelihood function in order to achieve variable selection. We propose a new prediction-based objective function that can be tailored to the requirements of practitioners and can be used to assess and interpret a given problem. Based on cross-validation techniques and the idea of importance sampling, our proposal scans low-dimensional models under the assumption of sparsity and, for each of them, estimates their objective function to assess their predictive power in order to select. Two applications on cancer data sets and a simulation study show that the proposal compares favorably with competing alternatives such as, for example, Elastic Net and Support Vector Machine. Indeed, the proposed method not only selects smaller models for better, or at least comparable, classification errors but also provides a set of selected models

[0⋆] The first two authors are Joint First Authors.

instead of a single one, allowing to construct a network of possible models for a target prediction accuracy level.

Keywords: Biomarker selection, Genomic networks, Disease classification, Breast cancer, Acute leukemia, Model averaging

1 Introduction

Gene selection has become a common task in most gene expression studies. The problem of assigning tumours to a known class is an example that is of particular importance and has received considerable attention in the last ten years. Conventional class prediction methods of leukemia or other cancers are in general based on microscopical examination of stained tissue specimens. However, such methods require highly trained specialists and are subjective (Tibshirani *et al.*, 2002).

To avoid these drawbacks, many automatic selection methods have been proposed recently. The goal of these methods is often to identify the smallest possible set of genes that can still achieve good predictive performance (Díaz-Uriarte and De Andres, 2006), although this is not necessarily the only criterion based on which model (gene) selection is carried out (see, for example Leng *et al.*, 2006). However, these methods have the advantage of being objective and have improved the correct classification rate in various cases. Among the different methodologies brought forward in this context we can find those proposed by Tibshirani *et al.* (2002), Dudoit *et al.* (2002), Zhu and Hastie (2004), Zou and Hastie (2005). See also Díaz-Uriarte and De Andres (2006) and the references therein for other approaches.

Nonetheless, many of these methods do not necessarily respond to the needs of practitioners and researchers when they approach the gene selection process. First of all, many of them have to rely on some form of size reduction and often require a subjective input to determine the dimension of the problem. Also, many of these methods often provide a single model as an output whereas genes interact inside biological systems and can be interchangeable in explaining a specific response. The idea of interchangeability of genes in explaining responses appears for instance in Kristensen *et al.* (2012). These authors use the PARADIGM algorithm of Vaske *et al.* (2010) to combine mRNA expression and DNA copy number in order to construct clusters of patients that provide the best predictive value. The resulting clusters can be seen as being characterized by different significantly expressed genes and we can refer to their interactive structure as *paradigmatic* networks.

Another issue of most existing gene selection methods is their reliance on the likelihood function, or a penalized version of it, as a means to develop a selection criterion. However, the likelihood function may not necessarily be the quantity that users are interested in as they may want to target some other kind of loss function such as, for example, the classification error. Of course, maximizing

the likelihood function is not typically the same as minimizing a particular loss function. Moreover, adapting these methods to handle missing or contaminated data is not straightforward. This has limited the applicability and reliability of these methods in many practical cases.

To eliminate the limitations of the gene selection procedures described above, this paper proposes an objective function for out-of-sample predictions that can be tailored to the requirements of practitioners and researchers. This is achieved by enabling them to select a criterion according to which they would like to assess and/or interpret a given problem. However, the optimization of such a criterion function is typically not an easy task since the function can be discontinuous, non-convex and would require computationally intensive techniques. To tackle this issue, we propose a solution using a different approach based on a procedure that resembles *importance sampling*. This new approach provides a general and flexible framework for gene selection as well as for other model selection problems.

The advantages of this proposal are multiple:

- **Flexibility**: It allows the users to specify a criterion that can be tailored to the specific problem setting. It is able to handle different kinds of responses, problems of missing and contaminated data, multicollinearity, etc.
- **Prediction Power**: The result of the procedure is a set of models with high predictive power with respect to the specified criterion. It is especially suitable in selecting genes and models to achieve accurate predictions.
- **Dimension-reduction**: It can provide an assessment of the dimension of the problem because it greatly reduces the number of necessary covariates and eases the interpretation without requiring any preliminary size reduction.
- **Network-building**: With the reduced model size, it preserves the capacity to build gene-networks to provide a more general view of the potential paradigmatic structures of the genetic information.

This last aspect is of great interest for gene selection since this list can provide insight into the complex mechanisms behind different biological phenomena. Different cases, some of which can be found in Section 4, indicate that this method appears to outperform other methods in terms of criteria minimization while, at the same time, selects models of considerably smaller dimension which allow improved interpretation of the results. The set of selected models can naturally be viewed as a network of possible structures of genetic information. We call

this a paradigmatic network. In Section 4 we give an example of a graphical representation of such networks based on the analysis of one of two cancer data sets which are discussed therein.

In this paper we first describe and formalize the proposed approach within the model selection statistical framework in Section 2. In Section 3 we illustrate the techniques and algorithms used to address the criterion minimization problem highlighted in Section 2. The performance of our approach is then illustrated on two data sets concerning leukemia classification (Golub *et al.*, 1999) and breast cancer classification (Chin *et al.*, 2006) in Section 4. We conclude the paper in Section 6 by summarizing the benefits of the new approach and providing an outlook on other potential applications that can benefit from this methodology.

2 Approach

To introduce the proposed method, let us first define some notation which will be used throughout this paper:

1. Let $\mathcal{J}_f = \{1, 2, ..., p\}$ be the set of indices for p potential covariates included in the $n \times p$ matrix $\mathbf{X}$. We allow $\mathbf{X}$ to include a vector of 1s.
2. Let $\mathcal{J} = \mathcal{P}(\mathcal{J}_f) \setminus \emptyset$, $|\mathcal{J}| = 2^p - 1$, be the power set including all possible models that can be constructed with the p covariates excluding the empty set.
3. Let $\jmath \in \mathcal{J}$ be a model belonging to the above mentioned power set.
4. Let $\boldsymbol{\beta}^{\jmath} \in \mathbb{R}^p$ be the parameter vector for model $\jmath$, i.e.

$$\boldsymbol{\beta}^{\jmath}_k = \begin{cases} \boldsymbol{\beta}_k & \text{if } \; k \in \jmath \\ 0 & \text{if } \; k \notin \jmath \end{cases}$$

   where $\boldsymbol{\beta}^{\jmath}_k$, $\boldsymbol{\beta}_k$ are respectively the kth element of $\boldsymbol{\beta}^{\jmath}$ and $\boldsymbol{\beta}$, with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T \in B \subseteq \mathbb{R}^p$.

Keeping this notation in mind, for a given model $\jmath \in \mathcal{J}$ we have that

$$\mathbb{E}[Y|\mathbf{X}] = g(\mathbf{X}, \boldsymbol{\beta}^{\jmath}), \tag{1}$$

where $\mathbb{E}[\cdot]$ is the expectation operator and $g(\cdot,\cdot)$ is a link function known up to the parameter vector $\boldsymbol{\beta}^{\jmath} \in \mathbb{R}^p$. Models of the form (1) are very general and include all parametric models and a large class of semiparametric models when $g(\cdot,\cdot)$ is not completely known.

We assume that for a fixed $\jmath$, based on a specific choice for model (1) with corresponding parameter vector $\boldsymbol{\beta}^{\jmath}$ and given a new covariate vector $\mathbf{X}_0$, the user can construct a prediction $\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath})$. To assess the quality of this prediction we assume that we have a divergence measure available which we denote as $D\{\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}), Y_0\}$. The only requirement imposed on the divergence measure is that it satisfies the property of positiveness, i.e.

$$\begin{aligned} D(u,v) &> 0 \ \text{ for } \ u \neq v \\ D(u,v) &= 0 \ \text{ for } \ u = v. \end{aligned}$$

With this property being respected, the divergence measure can arbitrarily be specified by the user according to the interest in the problem. Examples of such divergence measures include the L_1 loss function

$$D\{\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}), Y_0\} = |\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}) - Y_0|$$

or an asymmetric classification error

$$\begin{aligned} D\{\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}), Y_0\} =& I\{\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}) = 1, Y_0 = 0\} w_1 \\ &+ I\{\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}) = 0, Y_0 = 1\} w_2. \end{aligned}$$

where $w_1, w_2 \geq 0$. The latter is for a Bernoulli response and is typically an interesting divergence measure when asymmetric classification errors have to be considered. Indeed, in most clinical situations, the consequences of classification errors are not equivalent with respect to the direction of the misclassification. For instance, the prognosis and the treatment of Estrogen Receptor (ER) positive Breast Cancers (BC) are quite different from those of ER negative ones. Indeed, if a patient with ER negative is treated with therapies designed for patients with ER positive, the consequence is much more severe than if this were done the other way round because of the excessive toxicities and potentially severe side effects. It therefore makes sense to give different values to w_1 and w_2. By defining $w_1 > w_2$ we would take these risks into account, where w_1 would be the weight for a misclassification from ER negative to ER positive BC and w_2 for the opposite direction. Weight values can be modulated according to the current medical knowledge and the clinical intuition of the physicians.

Considering this divergence measure $D(\cdot,\cdot)$, we are consequently interested in finding the best models within the general class given in (1). To do so, we would ideally aim at solving the following risk minimization problem :

$$\widehat{\boldsymbol{\beta}}^{\jmath} \in \mathcal{B} \equiv \underset{\jmath \in \mathcal{J}}{\operatorname{argmin}} \underset{\boldsymbol{\beta}^{\jmath}}{\operatorname{argmin}} \mathbb{E}_0 \left[D \left\{ \widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}), Y_0 \right\} \right], \tag{2}$$

where $\mathbb{E}_0$ denotes the expectation on the new observation $(Y_0, \mathbf{X}_0)$. Let $\jmath_0$ denote the models with the smallest cardinality among all $\widehat{\boldsymbol{\beta}}^{\jmath} \in \mathcal{B}$. Note that there could be more than one model with the same prediction property and of the same size, hence $\jmath_0$ could contain more than one model. Let us define the models corresponding to $\jmath_0$ as the "true" models. Thus, our "true" models are essentially the most parsimonious models that minimize the expected prediction error.

The optimization problem in (2) is typically very difficult to solve. First of all, supposing we do not consider interaction terms, the outer minimization would require to compare a total of $2^p - 1$ results, each a result of the inner minimization problem. In addition, each of the $2^p - 1$ inner minimization problems is also very hard to solve, even if the risk $\mathbb{E}_0[D\{\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}), Y_0\}]$ were a known function of $\boldsymbol{\beta}^{\jmath}$. Indeed, the inner minimization problem is in general non-convex and could be combinatorial, implying that the minimizer might not be unique. For example, when $D(\cdot,\cdot)$ is the classification error, this problem is combinatorial by nature. In practice, the computational challenge is even greater because the risk function $\mathbb{E}_0[D\{\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}), Y_0\}]$ is a function of $\boldsymbol{\beta}^{J}$ without explicit form and needs to be approximated.

We propose to estimate $\mathbb{E}_0[D\{\widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}), Y_0\}]$ via an m-fold cross-validation (typically $m = 10$) repeated K times. More specifically, for a sample of size n, we repeat the following procedure K times. At the kth repetition, we randomly permute the sets $(\mathbf{X}_r, Y_r)$, with r indexing a row of the data (i.e. $r = 1, \ldots, n$), and then select $\lfloor n/m \rfloor$ observations from the permuted data to form "test" data sets, subindexed (i, l) (with $i = 1, \ldots, \lfloor n/m \rfloor$ and $l = 1, \ldots, m$) and superindexed $k = 1, \ldots, K$, i.e. $(\mathbf{X}^k_{(i,l)}, Y^k_{(i,l)})$. Therefore l indicates the test set, i indicates the observation within this test set and k represents the repetition (associated with a certain permutation of the data). The classical 10-fold cross-validation, for example, is obtained by defining $K = 1$ and $m = 10$. Given this, the estimated risk is

$$\widehat{\mathbb{E}}_0 \left[D \left\{ \widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^{\jmath}), Y_0 \right\} \right] = \frac{1}{\lfloor n/m \rfloor m K} \sum_{k=1}^{K} \sum_{l=1}^{m} \sum_{i=1}^{\lfloor n/m \rfloor} D\{\widehat{Y}(\mathbf{X}^k_{(i,l)}, \boldsymbol{\beta}^{\jmath}), Y^k_{(i,l)}\}. \tag{3}$$

Having approximated the expectation $\mathbb{E}_0$, the minimization problem in (2) becomes

$$\underset{\jmath\in\mathcal{J}}{\operatorname{argmin}} \underset{\boldsymbol{\beta}^\jmath}{\operatorname{argmin}} \widehat{\mathbb{E}}_0 \left[D\left\{ \widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^\jmath), Y_0 \right\} \right]. \tag{4}$$

Despite the above approximation, the minimization problem remains complicated for the reasons mentioned earlier. Thus, we further eliminate the inner minimization problem in (4) by inserting an estimator $\widehat{\boldsymbol{\beta}}^\jmath$ obtained independently from the minimization procedure. More specifically, we assume that an estimator of $\boldsymbol{\beta}^\jmath$, say $\widehat{\boldsymbol{\beta}}^{\jmath,k}$, is available based on model (1) and "training" observations $(\mathbf{X}^k_{\lfloor n/m \rfloor+1}, Y^k_{\lfloor n/m \rfloor+1}), \ldots, (\mathbf{X}^k_n, Y^k_n)$ (i.e. those observations excluded from the above mentioned "test" data sets). This estimator can be any available estimator, for example, the maximum likelihood estimator (MLE), a moment based estimator, or a quantile regression based estimator, etc. (see, for example, Azzalini, 1996; Hall *et al.*, 2005; Koenker, 2005). We then replace the inner minimization in (4) directly with the approximate expectation evaluated at $\widehat{\boldsymbol{\beta}}^{\jmath,k}$'s and simplify (4) to

$$\underset{\jmath\in\mathcal{J}}{\operatorname{argmin}} \frac{1}{\lfloor n/m \rfloor m K} \sum_{k=1}^{K} \sum_{l=1}^{m} \sum_{i=1}^{\lfloor n/m \rfloor} D\{\widehat{Y}(\mathbf{X}^k_{(i,l)}, \widehat{\boldsymbol{\beta}}^{\jmath,k}), Y^k_{(i,l)}\}. \tag{5}$$

The intuition of replacing the inner minimization in (4) with a sample average evaluated at an arbitrary estimator is due to the fact that this estimator, under a fixed "true" model and regardless of whether this estimator is a standard MLE or a minimizer of the divergence measure $D(\cdot,\cdot)$, is an approximation to the "true" parameter. This means that, consequently, different estimators are "close" to each other. As a consequence, the minimization problem in (5) can be considered to be a close approximation to $\min_{\boldsymbol{\beta}^\jmath} \mathbb{E}_0 \left[D\left\{ \widehat{Y}(\mathbf{X}_0, \boldsymbol{\beta}^\jmath), Y_0 \right\} \right]$. In fact, using an informal law of large numbers argument, as $n/m \to \infty$, then we have that $\widehat{\boldsymbol{\beta}}^\jmath \xrightarrow{p} \boldsymbol{\beta}^\jmath$. If in addition $m \to \infty$ then, under some regularity conditions on $D(\cdot,\cdot)$, the averages tend to the desired expectation. On the other hand, if instead we consider m as fixed, we would have an unbiased estimator of the expected risk.

We now have an optimization problem in (5) which requires a comparison of $2^p - 1$ values and is much easier to solve. To further reduce the number of comparisons, the following section describes some procedures and algorithms allowing to solve this problem in a more efficient manner.

3 Heuristic procedure

To solve the optimization problem in (5), we propose an approach designed to have the following three features:

1. Identify a **set of models** that carry large predictive power instead of a single "best" model;
2. Find this set of models within a **reasonable time**, without having to explore all possible models;
3. This set achieves **sparsity**, i.e. most of the parameters in $\boldsymbol{\beta}$ will be fixed at zero in each of the models in the set.

Note that the last feature above reflects the belief that most of the covariates are irrelevant for the problem under consideration and should be excluded. Indeed, our method is designed to work effectively if such a sparsity assumption holds, putting it on the same level of almost all variable selection procedures in the literature. Moreover, we require the method to have the first feature in order to increase flexibility in terms of interpretation. Indeed, in many domains such as gene selection, for example, the aim may not be to find a single model but a set of variables (genes) that can be inserted in a paradigmatic structure to better understand the contribution of each of them via their interactions.

Given this goal, assume that we have at our disposal an estimate of the measure of interest $D(\cdot, \cdot)$ for all possible $2^p - 1$ models. In this case, our interest would be to select a set of "best" models by simply keeping the set of models that have a low discrepancy measure $D(\cdot, \cdot)$. It is of course unrealistic to obtain a discrepancy measure for all models in most practical cases because this would require a considerable amount of time for computation. Therefore, in order to achieve the second feature, instead of examining all possible models, we can randomly sample covariates from $\mathcal{J}$. The random sampling needs to be carefully devised because in practice, for example in gene selection problems, the number of covariates p can easily reach thousands or tens of thousands (see examples in Section 4, where $p = 7,129$ and $p = 22,215$ respectively). In such situations, $2^p - 1$ is an extremely large number and the probability of randomly sampling a "good" set of variables from the $2^p - 1$ variables is very small. Using the sparsity property of the problem, we propose to start with the set of variables $\mathcal{M}_0$ (typically an empty set) and increase the model complexity stepwise. Throughout this procedure, we ensure that at step k, the most promising covariates based

on the evaluation at step $k-1$ are given higher probabilities of being randomly drawn. The last idea is in the spirit of "importance sampling" in the sense that covariates with more importance based on the previous step are "encouraged" to be selected in the current step. Note that by construction we achieve sparsity if we stop the stepwise search at models of size $d_{\max} \ll p$.

More formally, let us first define the set of all possible models of size d as

$$\mathcal{S}_d = \{(i_1, \ldots, i_d) \mid i_1, \ldots, i_d \in \mathcal{J}_f;\ i_1 < \ldots < i_d\}.$$

We then define the set of promising models, $\mathcal{S}_d^*$, as the ones with an estimated out-of-sample divergence measure $D(\cdot, \cdot)$ below a certain estimated α-quantile. The value of α is user-defined depending on the problem at hand, and is typically a small value such as $\alpha = 1\%$. The formal definition of this set would then be

$$\mathcal{S}_d^* = \{\jmath \mid \jmath \in \mathcal{S}_d\ ;\ \widehat{D}_\jmath \leq \widehat{q}_d(\alpha)\},$$

where

$$\widehat{D}_\jmath \equiv \frac{1}{\lfloor n/m \rfloor mK} \sum_{k=1}^{K} \sum_{l=1}^{m} \sum_{i=1}^{\lfloor n/m \rfloor} D\{\widehat{Y}(\mathbf{X}_{(i,l)}^k, \widehat{\boldsymbol{\beta}}^{\jmath,k}), Y_{(i,l)}\}, \tag{6}$$

and $\widehat{q}_d(\alpha)$ is the α-quantile of the $\widehat{D}_\jmath$ ($\jmath \in \mathcal{S}_d$) values issued from B randomly selected models. Finally, we define the set of indices of covariates that are in $\mathcal{S}_d^*$ as

$$\mathcal{I}_d^* = \{i \mid i \in \jmath,\ \jmath \in \mathcal{S}_d^*\}$$

whose complement we define as $\mathcal{I}_d^c$ (i.e. all those covariates that are not included in $\mathcal{I}_d^*$).

With this approach in mind and using the above notations, to start the procedure we assume that we have p variables from which to select.

A. *Initial Step:* We start by adding the number of variables $d = 1$ to our initial variable set $\mathcal{M}_0$ with the goal of finally obtaining the set $\mathcal{I}_1^*$.

   1. Construct the p possible one variable models by augmenting $\mathcal{M}_0$ with each of the p available variables.
   2. Compute $\widehat{D}$ for every model obtained in Step A.1.

3. From Steps A.1 and A.2, construct the set $\mathcal{I}_1^*$ using (3). Go to Step B and let $d = 2$.

B. *General Step:* We define here the general procedure to construct $\mathcal{I}_d^*$ for $2 \leq d \leq d_{\max}$.

1. Augment $\mathcal{M}_0$ with d variables as follows:
   (i) Randomly select a set, either set $\mathcal{I}_{d-1}^*$ with probability π or its complement $\mathcal{I}_{d-1}^c$ with probability $1 - \pi$.
   (ii) Select one variable uniformly at random and without replacement from the set chosen in Step (i) and add this variable to $\mathcal{M}_0$.
   (iii) Repeat Steps (i) and (ii) until d variables are added to $\mathcal{M}_0$.
2. Construct a model of dimension d using the d variables selected in Step B.1. Repeat Step B.1 B times to construct B such models.
3. From Steps B.1 and B.2, construct the set $\mathcal{I}_d^*$ according to (3). If $d < d_{\max}$, go to Step B and let $d = d + 1$, otherwise exit algorithm.

Once the algorithm is implemented, the user obtains an out-of-sample discrepancy measure for all evaluated models. Given that the goal is to obtain a set of models $\mathcal{S}_d^*$ with high predictive power, the discrepancy measure delivers the criterion based on which it is possible to determine the optimal model dimension and the corresponding network structure.

3.1 Practical Considerations

The algorithm described above lays out the basic procedure to solve the problem in (2). However, as many other heuristic selection procedures, there are a series of "hyper-parameters" to be determined and certain aspects to be considered. In the following paragraphs we will discuss some of these issues arising when implementing our algorithm in practice.

3.1.1 Choice of algorithm inputs

The parameters $d_{\max}$, B, α and π of the above algorithm are to be fixed by the user. As mentioned earlier $d_{\max}$ represents a reasonable upper bound for the model dimension which is constrained to $d_{\max} \leq l$, where l depends on the limitations of the estimation method and is commonly the sample size n. As

for the parameter B, a larger value is always preferable to better explore the covariate space. However, a larger B implies heavier computations, hence a rule of thumb that could be used is to choose this parameter such that $p \leq B \leq \binom{p}{2}$. As mentioned earlier, the parameter α should define a small quantile, typically 1%. Finally, π determines to what extent the user assigns importance to the variables selected at the previous step. Given that $d_{\max} \ll p$ and α is small, we will typically have that $|\mathcal{I}^*_{d-1}| < |\mathcal{I}^c_{d-1}|$. In this setting, a choice of $\pi = 0.5$ for example would deliver a higher probability for the variables in $\mathcal{I}^*_{d-1}$ to be included in $\mathcal{I}^*_d$. All other parameters being equal, increasing the value of π would decrease the probability of choosing a variable in $\mathcal{I}^c_{d-1}$ and vice versa. Moreover, we discuss in Appendix A how the proposed algorithm can be adjusted to situations where p is either small or very large.

As a final note, it is also possible for the initial model $\mathcal{M}_0$ to already contain a set of p_0 covariates which the user considers to be essential for the final output. In this case, the procedure described above would remain exactly the same since the procedure would simply select from the p covariates which are not in the user-defined set and the final model dimension would simply be $p_0 + d$.

3.1.2 Model Dimension and Network Building

The final goal of the algorithm is to find a subset of models of dimension d^* that in some way minimize the considered discrepancy. A possible solution would be to select the set of models $\mathcal{S}^*_{d^*}$ such that $d^* = \min_{\in\{1,\dots,d_{\max}\}} q_d(\alpha)$. However, the quantity $q_d(\alpha)$ is unknown and replaced by its estimator $\widehat{q}_d(\alpha)$. Due to this, a solution that might be more appropriate would be to consider a testing procedure to obtain d^* taking into account the variability of $\widehat{q}_d(\alpha)$. For example, we could find the dimension d^* such that we cannot reject the hypothesis that $\widehat{q}_{d^*}(\alpha) = \widehat{q}_{d^*+1}(\alpha)$. Thus we sequentially test whether $\widehat{q}_{j+1}$ is smaller than $\widehat{q}_j$ for $j = 1, \dots, d_{\max}$. As long as the difference is significant we increment j by one unit, otherwise the minimum is reached and $d^* = j$.

The type of test and its corresponding rejection level are determined by the user based on the nature of the divergence measure. For example, if we take the L_1 loss function as a divergence, one could opt for the Mann-Whitney test or if the loss function is a classification error (as in the applications in Section 4), one could choose the binomial test or other tests for proportions. The rejection level will depend, among others, on the number of tests that need to be run, typically less than $d_{\max} - 1$, and need to be adjusted using, for example, the Bonferroni correction. Finally, once the set $\mathcal{S}^*_{d^*}$ is obtained, the user may still want to "filter"

the resulting models. Indeed, the number of models in the solution $\mathcal{S}^*_{d^*}$ may be large and the corresponding divergence estimates may vary considerably from model to model. Since these divergence measures are estimators, we again propose a multiple testing procedure to reduce the number of models in $\mathcal{S}^*_{d^*}$. Before doing so, we eliminate redundant models, thereby making sure that every model is included only once. Then, we start the testing procedure with an empty set $\mathcal{S}^0_{d^*} = \emptyset$ to which we add the model (or one of the models) that has the minimum divergence measure estimate, denoted $\widehat{D}_{\jmath_{\min}}$, where $\jmath_{\min} \in \mathcal{S}^*_{d^*}$ denotes this model. Then for every model $\jmath \in \mathcal{S}^*_{d^*} \setminus \jmath_{\min}$, we test whether $\widehat{D}_\jmath$ is greater than $\widehat{D}_{\jmath_{\min}}$. We add the model to $\mathcal{S}^0_{d^*}$ if the difference is not significant and stop adding models as soon as the test deems that the divergence of the next model is indeed larger. By doing so we finally obtain $\mathcal{S}^0_{d^*} \subseteq \mathcal{S}^*_{d^*}$ which is the set containing the models (and hence covariates) which can be interpreted in a paradigmatic network. Generally speaking, this network can be built starting from the most frequent covariate(s) present in $\mathcal{S}^0_{d^*}$ (we call these "hubs") and, subsequently, connecting these with the most frequent covariates included in the models with the previous hubs. This can be continued until the number of connected hubs is equal to d^*.

3.2 Related literature

Some of the ideas put forth in this work have also been considered in the literature. An extensive survey of the related works goes beyond the scope of this paper. Here we briefly describe some of the connections to three main ideas that have been explored to this point.

The first one is recognizing that practitioners might aim to minimize some criterion that differs from likelihood-type losses. An interesting paper illustrating this point is Juang *et al.* (1997) in the context of speech recognition. For their classification problem, these authors propose to minimize a "smoothed" version of the decision rule used for classification. The advantage of this procedure is that it yields better misclassification errors than using pure likelihood based criteria which intrinsically fit a distribution to the data. In the approach presented in this work we also deliver an approximate solution but, as opposed to approximating the problem and solving the latter in an exact manner as in Juang *et al.* (1997), we define the exact problem and try to approximately minimize the misclassification error through our algorithm.

Secondly, there is a large literature that uses stochastic search procedures to explore the space of candidate models. Influential work in this direction includes

George and McCulloch (1993) and George and McCulloch (1997) who postulate hierarchical Bayesian models. In their set-up, subsets of promising predictors form models with higher posterior probabilities. An interesting application of this framework for disease classification using gene expression data is the work of Yang and Song (2010). Cantoni *et al.* (2007) also consider a random exploration of the space of possible models, but avoiding the Bayesian formulation of George and McCulloch (1993). Their approach defines a probability distribution for the various candidate models based on a cross-validated prediction error criterion and then uses a Markov Chain Monte-Carlo method to generate a sample from this probability distribution. An important feature of the stochastic search implied by our algorithm is that it is a greedy method, while the aforementioned methods are not. The typical forward/backward greedy algorithms proposed in the literature are not random, while existing stochastic procedures are not greedy. Thus, the combination of greedy approach and random search approach seems to be new (see for instance Zhang, 2011, for some theory on greedy algorithms in sparse scenarios).

Finally, other authors have also considered providing a set of interesting models as opposed to a single "best" model. The stochastic search procedures mentioned in the above paragraph can naturally be used to obtain a group of interesting models. For example, Cantoni *et al.* (2007) consider a set of best indistinguishable models in terms of prediction. Random forests can be used to select variables and account for the stability of the chosen model as in Díaz-Uriarte and De Andres (2006). These methods can also be used to construct a set of interesting models.

4 Case Studies

In this section we provide an example of how the methodology proposed in this paper selects and groups genes to explain, describe and predict specific outcomes. We focus on the data-set (hereinafter *leukemia*) which collects information on Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) and is frequently used as an example for gene selection procedures. Indeed, Golub *et al.* (1999) were among the first to use this data to propose a gene selection procedure which was then followed up by other proposals that used the same data to compare their performance. We will use this data-set to underline the features and advantages of the proposed method. A second data-set concerning the research on breast cancer (presented in Chin *et al.* (2006)) is analysed in

Appendix C to show the outputs of the proposed method from another example[1].

The analysis of these data-sets focuses both on the advantages of the proposed methodology and the biological interpretation of the outcomes. One of the goals of our method is to help decipher the complexity of biological systems. We will take on an overly simplified view of the cellular processes in which we will assume that one biomarker maps to only one gene that in turn has only one function. Although this assumption is not realistic, it allows us to give a straightforward interpretation of the selected models or "networks" which can therefore provide an approximate first insight into the relationships between variables and biomarkers (as well as between the biomarkers themselves). We clarify that we do not claim any causal nature in the conclusions we present in these analyses but we believe that the selected covariates can eventually be strongly linked to other covariates that may have a more obvious and direct interpretation for the problem at hand. Finally, the data-set has binary outcomes (as does the data-set in Appendix C), hence we will make use of the Classification Error (CE) as a measure of prediction performance and we will not assign weights to a given prediction error. This means that misclassification errors are given the same weight, in the sense that a false positive prediction (e.g. predicted "presence" when the truth is "absence") is considered as undesirable as a false negative prediction. However, our method can consider also divergence measures based on unequal weights as highlighted in Section 2.

4.1 Acute Leukemia

Golub *et al.* (1999) were among the first to propose an automatic selection method for cancer classification and demonstrated the advantages of using such a method. One of the main applications of their method was on the *leukemia* data-set in which information regarding 72 patients is included, namely their type of leukemia (25 patients with AML and 47 patients with ALL) and 7,129 gene expressions used as explanatory variables to distinguish between two types of leukemia. As explained in Golub *et al.* (1999) this distinction is critical for successful treatment which substantially differs between classes. In fact, although remissions can be achieved using any of these therapies, cure rates are markedly increased and unwarranted toxicities are avoided when targeting the specific type of leukemia with the right therapy.

[1]The *Acute Leukemia* (Section 4.1) and the *Breast Cancer* (Appendix C) data-sets are made available in the R package "`datamicroarray`".

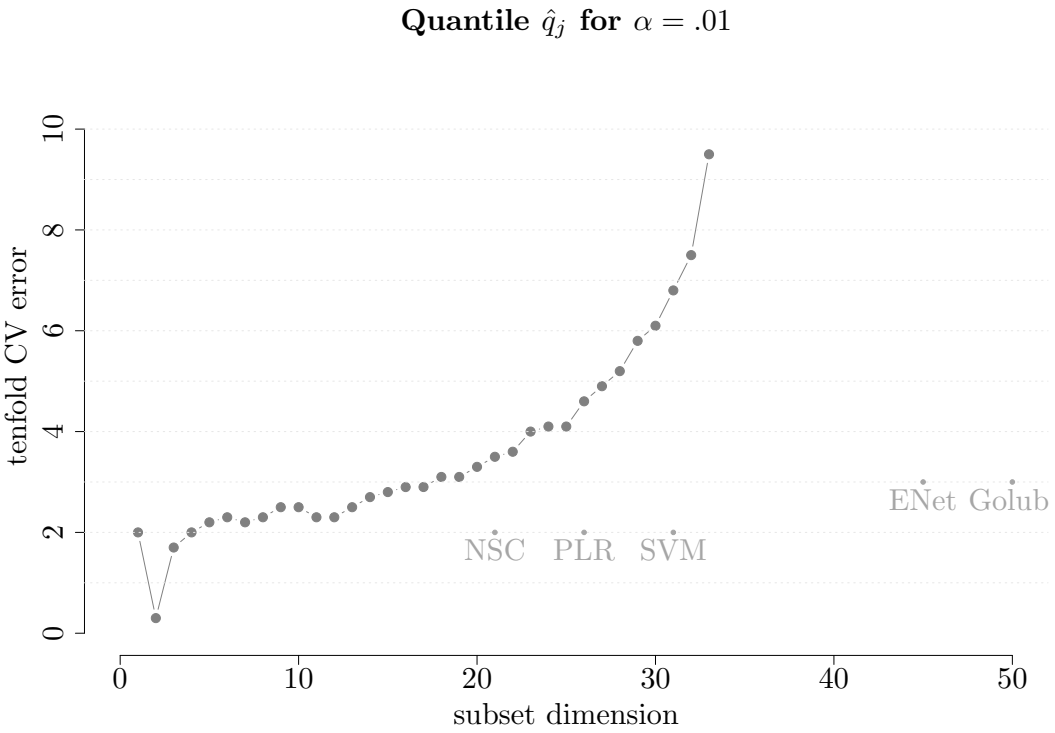

FIGURE 1: *Number of covariates vs.* $\widehat{D}$ *on leukemia cancer classification training set. The names are abbreviations for other selection method referred to in Table 1.*

4.1.1 Statistical analysis

In order to understand how our proposed methodology performs compared to existing ones, we split the *leukemia* data into the same training set (38 patients) and test set (34 patients) as in the original work by Golub *et al.* (1999). We employ our method on the training set to understand the dimension of the model and to select the most relevant genes. Setting $\alpha = 0.01$, the corresponding observed quantile of the 10-fold cross-validation CE ($\widehat{D}$) is shown in Figure 1. It can be seen that the error immediately decreases to almost zero when using two covariates instead of one, after which it roughly monotonically increases, suggesting that the optimal model dimension is two.

In Figure 1 we also plotted the performance of the other selection methods used on this training data which are represented by labelled dots reporting the acronyms of these methods that are listed in Table 1. These cross-validation errors are taken from Zou and Hastie (2005) in the same setting in which we ran

the proposed method. However, another table in which the competing methods were ran using currently available software is presented in Appendix B where the conclusions in terms of comparison do not differ from those presented in Table 1[2]. Indeed, the approach proposed in this work compares favourably to all other methods in terms of prediction power since they lie under the curve to the right of its minimum indicating that, compared to our method, they select models of considerably higher dimensions without achieving the same degree of performance in terms of CE. Therefore, for this particular case, our method outperforms the other methods. The sparsity and tenfold CV error are further illustrated in Table 1, where we also present the average prediction error on the test data. Considering the latter, it can be seen how the performance of the different methods are similar but the proposed method (which we refer to as *Panning*) is able to achieve the same performance by selecting models of a considerably lower dimension. As a final note to the table, the last line reports the performance of model averaging. Indeed, if the interest lies in predicting, the algorithm of Section 3 provides a set of models whose CE is below a given quantile α. The predictions of these models can be used in the spirit of model averaging where a general prediction can be obtained by taking the average of predictions of the selected set of models. The proposed methodology can therefore be potentially seen as a bridge between model selection and model averaging.

Once this procedure is completed, we can create a gene network to facilitate interpretation. This is a direct benefit of our method which does not deliver a single model after the selection process but provides a series of models that can be linked to each other and interpreted jointly. Indeed, the existence of a single model that links the covariates to the explained variable is probably not realistic in many settings, especially for gene classification. For this reason, the frequency with which each gene is included within the selected models and with which these genes are coupled with other genes provides the building block to create an easy-to-interpret gene network with powerful explanatory and predictive capacities. A graphical representation of this gene network can be found in Figure 2 where the size of a disk represents the frequency with which a particular biomarker is included in the selected models, and the line connecting the disks indicates the biomarkers that are included in the same model. Since the model dimension in this case is two, each biomarker is connected with only one other biomarker and, as can be observed, the proposed method identifies three main

[2]The use of the software making available the competing methods is described in Section 5.

Method	**Tenfold CV error**	**Test error**	**Number of genes**
Golub	3/38	4/34	50
Support vector machine (with recursive feature elimination)	2/38	1/34	31
Penalised logistic regression (with recursive feature elimination)	2/38	1/34	26
Nearest shrunken centroids	2/38	2/34	21
Elastic net	3/38	0/34	45
Panning Algorithm (107)			
Model a	0/38	2/34	2
Model b	0/38	2/34	2
Model c	0/38	2/34	2
[...]			
Model averaging		2/34	2

TABLE 1: *Summary of Leukemia classification results. The table is taken from Zou and Hastie (2005) where we added the Panning Algorithm. We obtained a total of 107 models of size 2 (109 different biomarkers) using a probability* $\alpha = 0.01$, $B = 20'000$ *bootstrap replicates, a selection probability* $\pi = 0.5$ *with* $D(\cdot, \cdot)$ *estimated through tenfold-CV repeated* $K = 10$ *times. Models "a" to "c" are three examples out of the 107 models. All 107 models have a tenfold-CV error of 0. The best test error is 2 and the worst is 12. For model averaging all models are equally weighted.*

"hubs" for the networks (green disks) generating three networks. Appendix B also reports a related table where the biomarkers are listed according to their position in the model. These positions represent families of biomarkers (or genes) whose members are interchangeable. By the latter we mean that, given the presence of biomarkers from other families, specific biomarkers can be replaced by another biomarker from within the same family without losing predictive power. This is the idea behind finding a paradigmatic network for gene selection purposes. In the following paragraph we provide a summary biological interpretation of the the three main biomarkers (i.e. the most frequent in the selected models) which we call "hubs" from which the networks start.

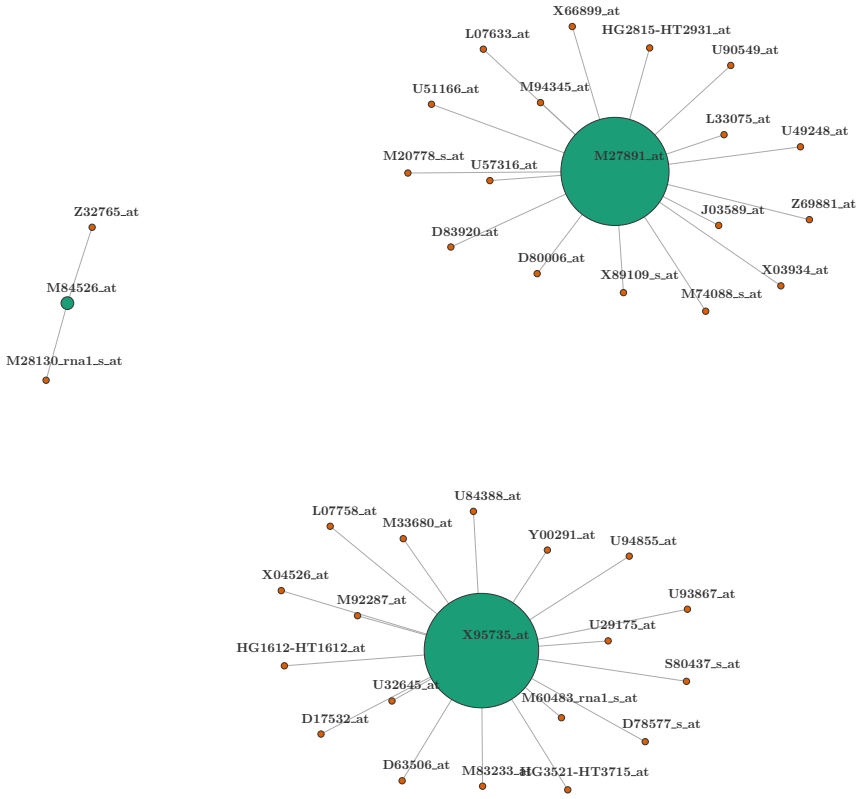

FIGURE 2: *Network representation of biomarkers selected from the leukemia data-set. Colors represent the position of covariates within the model: green for first position (hub) and orange for second. The width of the connecting lines is proportional to the frequency with which two biomarkers appear in the same model. The size of the disk is proportional to the frequency with which a biomarker is present within the selected set of models.*

4.1.2 Biological interpretation

The three hubs that were identified are the following:

1. Cystatin C: a secreted cysteine protease inhibitor abundantly expressed in body fluids (see Xu *et al.*, 2015);
2. Zyxin: a zinc-binding phosphoprotein that concentrates at focal adhesions and along the actin cytoskeleton;
3. Complement factor D: a rate-limiting enzyme in the alternative pathway of complement activation (see White *et al.*, 1992).

In the current state of knowledge about acute leukemia, these three hubs appear to make sense from a biological viewpoint. Cystatin C is directly linked to many pathologic processes through various mechanisms and recent studies indicate that the roles of Cystatin C in neuronal cell apoptosis induction include decreasing B-cell leukemia-2 (BCL-2) whose deregulation is known to be implicated in resistant AML (see Sakamoto *et al.*, 2015). Zyxin is a protein that interacts with Vasodilator-stimulated phosphoprotein (VASP) with both being involved in cellular adhesion and motility. VASP interacts with ABL (breakpoint cluster region-abelson) and is a substrate of the BcrAbl oncoprotein which drives oncogenesis in patients with chronic myeloid leukemia (CML) due to a constitutive activation of tyrosine kinase activity (see Bernusso *et al.*, 2015). Further results suggest that the phosphorylation and dephosphorylation cycle of VASP by the Abi-1-bridged mechanism regulates association of VASP with focal adhesions, which may regulate adhesion of Bcr-Abl-transformed leukaemic cells (see Masahiro *et al.*, 2012). Finally, Complement factor D, together with several other components of both the classical and alternative complement cascade, is primarily expressed through both adipocytes and monocytes-macrophages in human subjects (see White *et al.*, 1992; Gabrielsson *et al.*, 2003). A recent review in Ratajczak (2014) has stressed the role of the complement cascade as a trigger for hematopoietic stem cells from bone marrow into blood.

The interpretation of the network can be carried out through plots or tables such as those presented in Appendix B where the biomarkers can be grouped together into clusters having the same biological traits, e.g. transcription/translation factor activity, DNA repair and catabolism, apoptotic activity. This grouping allows a more straightforward interpretation of the links between the different families thereby providing a more general overview of how the elements of the identified network interact.

5 Simulation study

In this section we present a simulation study whose goal is to highlight the practical benefits of the proposed method over competing methods frequently used in genomics. Considering the complexity of simulating from a gene network, in this setting we limit ourselves to considering the existence of a unique true model which therefore does not allow to assess one of the features of the proposed approach which is its network building capacities. Hence, this section specifically focuses on the prediction power and dimension-reduction ability of the method and, for the comparison with alternative methods to be fair, we only keep one model for each simulation replicate. This means that, once the dimension of the model has been identified, the model with the lowest estimated prediction error is kept (thereby discarding the other potential candidates).

In this optic, for the simulation study we mimicked the acute leukemia dataset seen in Section 4.1 where we set the true model to be generated by a combination of two gene expressions: Cystatin C (X_1) and Thymine-DNA Glycosylase (X_2) (see Section 4.1.2). Hence the response $y^\star$ in the simulations is a realization of a Bernoulli random variable with probability parameter γ which is obtained through a logit-link function applied to a linear combination of the two above-mentioned variables plus an intercept (with all β coefficients equaling one) i.e.:

$$\gamma = \frac{1}{1 + \exp^{(1+X_1+X_2)}}.$$

Once the binary response variable $y^\star$ is generated, this is then separated into a training and a test set of the same size as that in the original data-set (i.e. 38 and 34 respectively).

Using the implementation of the proposed algorithm available at the corresponding GitHub repository[3], the results of the simulations based on 100 replications can be found in Table 2 where the median performances are reported. The proposed algorithm's hyper-parameters are $\alpha = 0.01$, $B = 20'000$, $\pi = 0.5$ and $D(\cdot, \cdot)$ based on the classical tenfold-CV ($K = 1$). To select the dimension d^*, we ran the testing procedure described in Section 3.1.2 based on a p-value of 0.1. As mentioned earlier, unlike Table 1, we only kept one model of dimension d^* instead of a set of models. This model was chosen such that it had the minimum training error and, if this minimum was not unique, then the model was randomly chosen among those achieving this minimum.

[3] https://github.com/SMAC-Group/panning

Concerning the competing methods, these were implemented using existing R functions with default values. For the Elastic Net we used the R package "`glmnet`", that implements the coordinate descent algorithm described in Friedman *et al.* (2010), using the `cv.glmnet()` function to select the lasso parameter. We performed a grid search over the values $\{0.2, 0.4, 0.6, 0.8, 1\}$ for the parameter α of the Elastic Net and kept the value yielding the best deviance[4]. As for the Nearest Shrunken Centroids method of Tibshirani *et al.* (2002) we considered the R package "`pamr`". We applied the function `pamr.train()` on the training data and took the value of the tuning parameter (threshold) yielding the best classification. The Support Vector Machines approach with recursive feature elimination was obtained through the function `fit.rfe()` in the "`pathClass`" R package . We used the function `crossval()` to select the soft-margin tuning parameter discussed in Chapelle *et al.* (2002). Finally, the penalized $L2$ logistic regression with greedy forward selection and backward deletion was implemented with the function `step.plr()` of the "`stepPlr`" R package. Note that this function also considers all possible interactions among the active variables and it is an implementation of the methodology proposed by Park and Hastie (2008). Finally, we used our own implementation for the logistic regression with greedy forward selection, selecting the model with the minimum BIC.

Method	**Tenfold CV error**	**Test error**	**Number of genes**
Panning Algorithm	0/38 (all)	1/34 (min: 0/34; max: 12/34)	2/7129 (all)
Elastic net	10/38 (min: 9/38; max: 12/38)	0/34 (all)	81/7129 (min: 1; max: 104)
Support vector machine	0/38 (all)	15/34 (all)	4/7129 (min: 4; max: 6)
Penalised logistic regression		12/34 (min: 8/34; max: 12/34)	5/7129 (all)
Nearest shrunken centroids	12/38 (min: 7/38; max: 18/38)	5/34 (min: 0/34; max: 5/34)	30/7129 (min: 3; max: 30)

TABLE 2: *Median performances of selection methods on 100 simulations based on a dataset of 7129 genes where only two are relevant.*

[4]Note that the special cases $\alpha = 0$ and $\alpha = 1$ correspond respectively to ridge regression and lasso.

Table 2 shows how the proposed method compares favorably in terms of median performance with the respect to the competing methods. Indeed, it is the best approach (or it is among the best) both in terms of cross-validation error as in terms test error. Even considering its maximum test error it is comparable to the other methods, keeping in mind that it selects models of extremely low (and above all correct) dimensions. For example, the Elastic Net is the without doubt the best in terms of test error but it selects a unique model of size 81 (in median) making its genetic interpretation much more complex. On the other hand, the proposed algorithm selects the correct dimension and, if considering the set of best models, would deliver a network which is more straightforward to interpret.

6 Conclusions

This paper has proposed a new model selection method with various advantages compared to existing approaches. Firstly, it allows the user to specify the criterion according to which they would like to assess the predictive quality of a model. In this setting, it gives an estimate of the dimension of the problem, allowing the user to understand how many gene expressions are needed in a model to well describe and predict the response of interest. Building on this, it provides a paradigmatic structure of the selected models where the selected covariates are considered as elements in an interconnected biological network. The approach can handle more variables than observations without going through dimension-reduction techniques such as pre-screening or penalization.

The problem definition of this method and the algorithmic structure used to solve it deliver further advantages such as the ability to cope with noisy inputs, missing data, multicollinearity and the capacity to deal with outliers within the response and the explanatory variables (robustness).

Some issues which must be taken into account concerning the proposed method are (i) its computational demand and (ii) its need for an external validation. As far as the first aspect goes, this can be considered indeed negligible compared to the time often required to collect the data it should analyse and can be greatly reduced according to the needs and requirements of the user. Concerning the second aspect, external validation is a crucial point which is often overlooked and is required for any model selection procedure. In this sense, the proposed method does not differ from any other existing approach in terms of additional requirements.

Having proposed a method with considerable advantages for gene selection using statistical ideas in model selection and machine learning, future research aims at studying the statistical properties of this approach to understand its asymptotic behaviour and develop the related inference tools.

Acknowledgements

We are very thankful to John Ramey (http://ramhiser.com/) for having processed the breast cancer and leukemia data set in Github and for having kindly answered our requests.
We thank Maria-Pia Victoria-Feser (Research Center for Statistics, University of Geneva, Switzerland) for her valuable comments and inputs as well as her institutional support.

Funding and Conflict of interest

No conflict of interest can be declared.

References

Andres, S. A. and Wittliff, J. L. (2012). Co-expression of genes with estrogen receptor-α and progesterone receptor in human breast carcinoma tissue. *Hormone molecular biology and clinical investigation*, **12**(1), 377–390.

Azzalini, A. (1996). *Statistical inference based on the likelihood*, volume 68. CRC Press.

Bernusso, V. A., Machado-Neto, J. A., Pericole, F. V., Vieira, K. P., Duarte, A. S., Traina, F., Hansen, M. D., Saad, S. T. O., and Barcellos, K. S. (2015). Imatinib restores vasp activity and its interaction with zyxin in bcr–abl leukemic cells. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, **1853**(2), 388–395.

Bohrer, L. R., Chuntova, P., Bade, L. K., Beadnell, T. C., Leon, R. P., Brady, N. J., Ryu, Y., Goldberg, J. E., Schmechel, S. C., Koopmeiners, J. S., *et al.* (2014). Activation of the fgfr–stat3 pathway in breast cancer cells induces a hyaluronan-rich microenvironment that licenses tumor formation. *Cancer research*, **74**(1), 374–386.

Cantoni, E., Field, C., Mills Flemming, J., and Ronchetti, E. (2007). Longitudinal variable selection by cross-validation in the case of many covariates. *Statistics in medicine*, **26**(4), 919–930.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine learning*, **46**(1-3), 131–159.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., *et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, **10**(6), 529–541.

Chou, J., Provot, S., and Werb, Z. (2010). Gata3 in development and cancer differentiation: cells gata have it! *Journal of cellular physiology*, **222**(1), 42–49.

Christer, H., Peter, K., Margaret, L. A., Stephen, H., and Kathryn, M. T. (2013). A mechanism for epithelial-mesenchymal transition and anoikis resistance in breast cancer triggered by zinc channel zip6 and stat3 (signal transducer and activator of transcription 3). *Biochemical Journal*, **455**(2), 229–237.

Chung, S. S., Giehl, N., Wu, Y., and Vadgama, J. V. (2014). Stat3 activation in her2-overexpressing breast cancer promotes epithelial-mesenchymal transition and cancer stem cell traits. *International journal of oncology*, **44**(2), 403–411.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene Selection and Classification of Microarray Data using Random Forest. *BMC Bioinformatics*, **7**(1), 3.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. *Journal of the American statistical association*, **97**(457), 77–87.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**(1), 1.

Gabrielsson, B. G., Johansson, J. M., Lönn, M., Jernås, M., Olbers, T., Peltonen, M., Larsson, I., Lönn, L., Sjöström, L., Carlsson, B., *et al.* (2003). High expression of complement components in omental adipose tissue in obese men. *Obesity research*, **11**(6), 699–708.

George, E. and McCulloch, R. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, **88**(423), 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, **7**(2), 339–373.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., and Caligiuri, M. A. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**(5439), 531–537.

Hall, A. R. *et al.* (2005). *Generalized method of moments*. Oxford University Press Oxford.

Juang, B., Hou, W., and Lee, C. (1997). Minimum classification error rate methods for speech recognition. *Speech and Audio Processing, IEEE Transactions on*, **5**(3), 257–265.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Kouros-Mehr, H., Slorach, E. M., Sternlicht, M. D., and Werb, Z. (2006). Gata-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell*, **127**(5), 1041–1055.

Kristensen, V. N., Vaske, C. J., Ursini-Siegel, J., Van Loo, P., Nordgard, S. H., Sachidanandam, R., Sørlie, T., Wärnberg, F., Haakensen, V. D., Helland, Å., *et al.* (2012). Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (dcis) reveal differential vascular and interleukin signaling. *Proceedings of the National Academy of Sciences*, **109**(8), 2802–2807.

Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284.

Masahiro, M., Mizuho, S., Yunfeng, Y., Masayoshi, I., Ryosuke, F., Takuya, O., Norihiro, I.-K., Tatsuo, T., and Naoki, W. (2012). Abi-1-bridged tyrosine phosphorylation of vasp by abelson kinase impairs association of vasp to focal adhesions and regulates leukaemic cell adhesion. *Biochemical Journal*, **441**(3), 889–899.

Park, M. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**(1), 30–50.

Ratajczak, M. (2014). A novel view of the adult bone marrow stem cell hierarchy and stem cell trafficking. *Leukemia*.

Sakamoto, K. M., Grant, S., Saleiro, D., Crispino, J. D., Hijiya, N., Giles, F., Platanias, L., and Eklund, E. A. (2015). Targeting novel signaling pathways for resistant acute myeloid leukemia. *Molecular genetics and metabolism*, **114**(3), 397–402.

Taniguchi, K. and Karin, M. (2014). Il-6 and related cytokines as the critical lynchpins between inflammation and cancer. In *Seminars in immunology*, volume 26, pages 54–74. Elsevier.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. *Proceedings of the National Academy of Sciences*, **99**(10), 6567–6572.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**(12), i237–i245.

White, R. T., Damm, D., Hancock, N., Rosen, B., Lowell, B., Usher, P., Flier, J., and Spiegelman, B. (1992). Human adipsin is identical to complement factor d and is expressed at high levels in adipose tissue. *Journal of Biological Chemistry*, **267**(13), 9210–9213.

690 Xu, Y., Ding, Y., Li, X., and Wu, X. (2015). Cystatin c is a disease-associated
691 protein subject to multiple regulation. *Immunology and cell biology*.

692 Yang, A.-J. and Song, X.-Y. (2010). Bayesian variable selection for disease
693 classification using gene expression data. *Bioinformatics*, **26**(2), 215–222.

694 Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse
695 representations. *Information Theory, IEEE Transactions on*, **57**(7), 4689–4708.

696 Zhu, J. and Hastie, T. (2004). Classification of Gene Microarrays by Penalized
697 Logistic Regression. *Biostatistics*, **5**(3), 427–443.

698 Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the
699 Elastic Net. *Journal of the Royal Statistical Society: Series B*, **67**(2), 301–320.

A Adapting the algorithm to p

In this subsection we provide two variants of the algorithm proposed in Section 3 in order to adapt it to situations where p is either small or large.

A.1 Adapting the algorithm to very large p

In situations where p is extremely large and the initial step of the algorithm is not computationally feasible, this step can, for example, be replaced by the following modified initial step:

A′. *Large p Modified Initial Step:* We start by augmenting our initial variable set $\mathcal{M}_0$ with $d = 1$ variable in order to construct the set $\mathcal{I}_1^*$.

1. Augment $\mathcal{M}_0$ with $d = 1$ variable selected uniformly at random in $\mathcal{J}_f$.
2. Construct B models of dimension 1 by repeating Step A′.1 B times.
3. From Steps A′.1 and A′.2, construct the set $\mathcal{I}_1^*$ using (3). Go to Step B and let $d = 2$.

A.2 Adapting the algorithm to small p

On the other hand, when p is of reasonable size it may be possible to compute and evaluate all the $\binom{p}{d'}$ models of dimension $2 \leq d' \leq d_{\max}$. In such cases, it may be feasible to also modify the initial step of the proposed algorithm to a different modified initial step. A possible modification is the following:

A″. *Small p Modified Initial Step:* We start by augmenting our initial varable set $\mathcal{M}_0$ with d $(1 \leq d \leq d')$ variables in order to construct the sets $\mathcal{I}_1^*, ..., \mathcal{I}_{d'}^*$.

1. We augment our initial variable set $\mathcal{M}_0$ with 1 variable in order to construct the set $\mathcal{I}_1^*$.
   (i) Construct the p possible models obtained by augmenting $\mathcal{M}_0$ with each of the p available variables.
   (ii) Compute $\widehat{D}(\cdot, \cdot)$ for every model obtained in Step (i).
   (iii) From Steps (i) and (ii), construct the set $\mathcal{I}_1^*$ using (3). Go to Step A″.2 and let $d = 2$.

2. We augment our initial model $\mathcal{M}_0$ set by d variables in order to construct the set $\mathcal{I}_d^*$.
   (i) Construct the $\binom{p}{d}$ possible models and augment $\mathcal{M}_0$ with all variables of these constructed models.
   (ii) Compute $\widehat{D}$ for every model obtained in Step (i).
   (iii) From Steps (i) and (ii), construct the set $\mathcal{I}_d^*$ using (3) and let $d = d+1$. Go to Step A$''$.2 (if $d < d'$) or Step B.1 (if $d \geq d'$), with model dimension starting value d.

B Complementary results on Acute Leukemia

Table 3 reports the main biomarker hubs and related biomarker networks for the *leukemia* data set analysed in Section 4.1.

Table 4 reports the performances of our implementation of the competing methods as described in Section 5. Unlike reported in Table 1, here the proposed method uses the classical tenfold-CV for $D(\cdot,\cdot)$ ($K = 1$). The other hyper-parameters are kept the same (i.e. $\alpha = 0.01$, $B = 20'000$ and $\pi = 0.5$).

C Breast Cancer

The second data-set we analyzed is the *breast cancer* data presented in Chin *et al.* (2006). The main goal behind analyzing this data is to identify the estrogen receptor expression on tumor cells which is a crucial step for the correct management of breast cancer. Similarly to Table 4 in Appendix B, Table 5 reports the performances of our implementation of the competing methods and the proposed approach on the *breast cancer* data. For the sake of this comparison, the data-set was randomly split into training (60) and test (58) sets. The hyper-parameters of the proposed method are $\alpha = 0.01$, $B = 30'000$, $\pi = 0.5$ and $D(\cdot,\cdot)$ is the classical tenfold-CV ($K = 1$).

	Affy ID	Gene ID	Gene Function	Biological Process
NETWORK 1				
Position 1	M27891_at	ENSG00000101439	Cystatin C	AA
Position 2	D80006_at	ENSG00000114978	MOB kinase activator 1A	AA
	M20778_s_at	ENSG00000163359	Collagen, type VI, alpha 3	AA
	U57316_at	ENSG00000108773	K(lysine) acetyltransferase 2A	TF
	U90549_at	ENSG00000182952	High mobility group nucleosomal binding domain 4	TF
	X66899_at	ENSG00000182944	Ewing Sarcoma region 1; RNA binding protein	TF
	M74088_s_at	ENSG00000134982	Adenomatous polyposis coli, DP2, DP3, PPP1R46	TF
	U51166_at	ENSG00000139372	thymine-DNA glycosylase	TF
	Z69881_at	ENSG00000074370	ATPase, Ca++ transporting, ubiquitous	IPT
	U49248_at	ENSG00000023839	ATP-binding cassette, sub-family C (CFTR/MRP), member 2	IPT
	X89109_s_at	ENSG00000102879	Coronin, actin binding protein, 1A	IPT
	HG2815-HT2931_at	ENSG00000092841	Myosin, Light Chain, Alkali, Smooth Muscle (Gb:U02629)	ACC
	M94345_at	ENSG00000042493	Capping protein (actin filament), gelsolin-like	ACC
	L33075_at	ENSG00000140575	IQ motif containing GTPase activating protein 1	ACC
	L07633_at	ENSG00000092010	Proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)	APC
	J03589_at	ENSG00000102178	Ubiquitin-like 4A	APC
	D83920_at	ENSG00000085265	FCN1, Ficolin-1	IR
	X03934_at	ENSG00000167286	CD3d molecule, delta (CD3-TCR complex)	IR
NETWORK 2				
Position 1	X95735_at	ENSG00000159840	Zyxin	ACC
Position 2	X04526_at	ENSG00000185838	Guanine nucleotide binding protein (G protein), beta polypeptide 1	ST
	D78577_s_at	ENSG00000128245	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta	ST
	U32645_at	ENSG00000102034	E74-like factor 4 (ets domain transcription factor)	TF
	U93867_at	ENSG00000186141	Polymerase (RNA) III (DNA directed) polypeptide C (62kD)	TF
	U29175_at	ENSG00000127616	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily A, member 4	TF
	Y00291_at	ENSG00000077092	Retinoic acid receptor, beta	TF
	D17532_at	ENSG00000110367	DEAD (Asp-Glu-Ala-Asp) Box Helicase 6	TF
	HG3521-HT3715_at	ENSG00000127314	Ras-Related Protein Rap1b	TF
	M83233_at	ENSG00000140262	Transcription factor 12	TF
	U94855_at	ENSG00000175390	Eukaryotic translation initiation factor 3, subunit F	TF
	L07758_at	ENSG00000136045	PWP1 homolog	TF
	D63506_at	ENSG00000116266	Syntaxin binding protein 3	IR
	M33680_at	ENSG00000110651	CD81 molecule	IR
	HG1612-HT1612_at	ENSG00000175130	Macmarcks	CG
	M92287_at	ENSG00000112576	Cyclin D3	CG
	M60483_rna1_s_at	ENSG00000113575	Protein Phosphatase 2 (formerly 2A), catalytic subunit, alpha isoform	CG
	U84388_at	ENSG00000169372	CASP2 and RIPK1 domain containing adaptor with death domain	AA
	S80437_s_at	ENSG00000169710	Fatty acid synthase	
NETWORK 3				
Position 1	M84526_at	ENSG00000197766	Complement factor D (adipsin)	IR
Position 2	M28130_rna1_s_at	ENSG00000169429	Interleukine-8	IR
	Z32765_at	ENSG00000135218	CD36 - Thrombospondin receptor	IR

TABLE 3: *Biomarker network organisation* - leukemia *data set - Lymphoblastic / Myeloblastic leukemia. TF = Transcription/translation factor activity, DNA repair and catabolism - AA = apoptotic activity - IR = immunity, inflammatory response (blood coagulation, antigen presentation and complement activation) - IPT = intracellular protein trafficking, transmembrane transport - ACC = actin activity, cytoskeleton organisation - APC = protein catabolism - ST = intracelular signal transduction - CG = cell growth, proliferation and division. Source:* www.ensembl.org ; www.uniprot.org

Method	**Tenfold CV error**	**Test error**	**Number of genes**
Support vector machine (with recursive feature elimination)	0/38	5/34	2/7129
Penalised logistic regression (with forward selection followed by backward deletion)	0/38*	4/34	3/7129
Nearest shrunken centroids	3/38	1/34	372/7129
Elastic net	3/38	2/34	74/7129
Panning Algorithm (131)			
Model a	0/38	1/34	2/7129
Model b	0/38	2/34	2/7129
Model c	0/38	2/34	2/7129
[...]			2/7129

TABLE 4: *Performances of our implementation of the competing methods on the leukemia data-set. For the Penalised logistic regression(*), the in-sample error is reported instead of the tenfold-CV error. For the Panning Algorithm, models "a" to "c" are three examples out of the 131 models. All the 131 models have a tenfold-CV error of 0. The best test error is 1 and the worst is 20.*

Method	**Tenfold CV error**	**Test error**	**Number of genes**
Support vector machine (with recursive feature elimination)	0/60	10/58	3/22215
Penalised logistic regression (with forward selection followed by backward deletion)	0/60*	12/58	15/22215
Nearest shrunken centroids	2/60	11/58	5/22215
Elastic net	3/60	11/58	196/22215
Panning Algorithm (241)			
Model a	0/60	9/58	3/22215
Model b	0/60	9/58	3/22215
Model c	0/60	10/58	3/22215
[...]			3/22215

TABLE 5: *Performances of our implementation of the methods on the breast cancer data-set. For the Penalised logistic regression (*), the in-sample error is reported instead of the tenfold-CV error. For the proposed method, models "a" to "c" are three examples out of 241 models. All the 241 models have a tenfold-CV error of 0. The best test error is 9 and the worst is 28.*

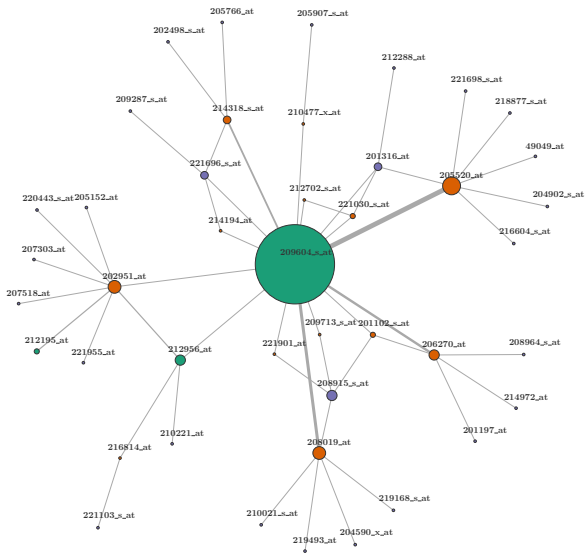

FIGURE 3: *Network representation of biomarkers selected from breast cancer data-set. Colors represent the position of covariates within the model: green for first position (hub), orange for second and purple for third. The width of the connecting lines is proportional to the frequency with which two biomarkers appear in the same model. The size of the circles is proportional to the frequency with which a biomarker is present within the selected set of models. (Note: biomarker "209602_s_at" is merged with biomarker "209604_s_at").*

Figure 3 shows the paradigmatic network identified by our method for the *breast cancer* data for which the selected model dimension is three (i.e. only three biomarkers are needed in a model to well classify the breast cancer). We used the hyper-parameters $\alpha = 0.01$, $B = 22'215$, $\pi = 0.05$ and for $D(\cdot, \cdot)$ the tenfold-CV repeated $K = 10$ times was used. Table 6 provides the details of the networks based on the three main hubs and is to be interpreted as described in Section 4.1.

This figure is a clear example of the advantages of the proposed method since, it not only selects a set of low-dimensional models with a high predictive power, but also provides the basis for a more general biological interpretation which takes into account interactions between different biomarkers as opposed to one single

model. The three main hubs identified through the proposed algorithm are:

1. GATA binding protein 3 (GATA3): a transcription factor regulating the differentiation of breast luminal epithelial cells;
2. IL6 Signal Transducer (IL6 ST): a pro-inflammatory cytokine signal transducer;
3. TBC1 domain family, member 9 (TBC1D9): a GTPase-activating protein for Rab family protein involved in the expression of the ER in breast tumors.

GATA3 is known to regulate the differentiation of epithelial cells in mammary glands (see Kouros-Mehr *et al.*, 2006) and is required for luminal epithelial cell differentiation. Its expression is progressively lost during luminal breast cancer progression as cancer cells acquire a stem cell-like phenotype (see Chou *et al.*, 2010). IL6 ST has been linked to breast cancer epithelial-mesenchymal transition and cancer stem cell traits (see Chung *et al.*, 2014), cancer-promoting microenvironment (see Bohrer *et al.*, 2014) and resistance (see Christer *et al.*, 2013). Moreover, this result supports the assertion by Taniguchi and Karin (2014) that IL6 ST and related cytokines are the critical lynchpins between inflammation and cancer. Finally, concerning the third biomarker, a recent publication by Andres and Wittliff (2012) has shown that the expression of the ER on the surface of breast tumor cells is highly correlated with the coordinate expression of different genes among which we can find TBC1D9 and GATA3. These two genes are not only considered as relevant genes according to the proposed method but as actual hubs of the "best" models which define the structure of the identified network. Instead of selecting a single model with many biomarkers whose interactions may be difficult to interpret, the proposed method selects a set of models with few biomarkers that allow them to be individually easy to interpret without losing the possibility of interpreting them within the larger network. This is what this paper intends with the expression "paradigmatic network" since by taking this approach it is possible to identify a set of biomarker families within which each biomarker is interchangeable with the others.

	Affy ID	Gene ID	Gene Function	Biological Process
NETWORK 1				
Position 1	209604_s_at	ENSG00000107485	GATA binding protein 3	TF
Position 2	205520_at	ENSG00000115808	Striatin, calmodulin binding protein	ER
Position 3	204902_s_at	ENSG00000168397	Autophagy related 4B, cysteine peptidase (APG4B, AUTL1, DKFZp586D1822, KIAA0943)	APC
	221698_s_at	ENSG00000172243	C-type lectin domain family 7, member A	IR
	49049_at	ENSG00000178498	Deltex 3, E3 ubiquitin ligase	APC
	209602_s_at	ENSG00000107485	GATA binding protein 3	TF
	216604_s_at	ENSG00000003989	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 2	IPT
	218877_s_at	ENSG00000066651	TRNA methyltransferase 11 homolog	TF
	201316_at	ENSG00000106588	Proteasome (prosome, macropain) subunit, alpha type, 2	APC
Position 2	208019_at	ENSG00000147117	Zinc finger protein 157	TF
Position 3	219168_s_at	ENSG00000186654	PRR5 (Proline rich 5 (renal))	CG
	219493_at	ENSG00000171241	SHC SH2-domain binding protein 1	CG
	204590_x_at	ENSG00000139719	Vacuolar protein sorting 33 homolog A	APC
	210021_s_at	ENSG00000152669	Cyclin O	CG
	208915_s_at	ENSG00000103365	Golgi-associated, gamma adaptin ear containing, ARF binding protein 2	IPT
Position 2	214318_s_at	ENSG00000073910	Furry homolog	ACC
Position 3	205766_at	ENSG00000173991	Titin-cap (Telethonin)	ACC
	221696_s_at	ENSG00000060140	Serine/threonine/tyrosine kinase 1	CG
	202498_s_at	ENSG00000059804	Solute carrier family 2 (facilitated glucose transporter), member 3	STM
Position 2	201102_s_at	ENSG00000141959	Phosphofructokinase, liver	STM
Position 3	208915_s_at	ENSG00000103365	Golgi-associated, gamma adaptin ear containing, ARF binding protein 2	IPT
Position 2	201316_at	ENSG00000106588	Proteasome (prosome, macropain) subunit, alpha type, 2	APC
Position 3	212288_at	ENSG00000187239	Formin binding protein 1	ACC
Position 2	209713_s_at	ENSG00000116704	Solute carrier family 35 (UDP-GlcA/UDP-GalNAc transporter), member D1	STM
Position 3	208915 _s_at	ENSG00000103365	Golgi-associated, gamma adaptin ear containing, ARF binding protein 2	IPT
Position 2	212702_s_at	ENSG00000185963	Bicaudal D homolog 2	ACC
Position 3	221030_s_at	ENSG00000138639	Rho GTPase activating protein 24	ACC
Position 2	212956_at	ENSG00000109436	TBC1 domain family, member 9 (with GRAM domain)	IPT
Position 3	210221_at	ENSG00000080644	Cholinergic receptor, nicotinic, alpha 3 (neuronal)	ITT
Position 2	214194_at	ENSG00000083520	DIS3 mitotic control homolog (Ribosomal RNA-processing protein 44)	TF
Position 3	221696_s_at	ENSG00000060140	Serine/threonine/tyrosine kinase 1	CG
Position 2	216814_at	ENSG00000232267	ACTR3 pseudogene 2	PUP
Position 3	221103_s_at	ENSG00000206530	Cilia and flagella associated protein 44	ACC
Position 2	221030_s_at	ENSG00000138639	Rho GTPase activating protein 24	ACC
Position 3	201316_at	ENSG00000106588	Proteasome (prosome, macropain) subunit, alpha type, 2	APC
Position 2	221696_s_at	ENSG00000060140	Serine/threonine/tyrosine kinase 1	CG
Position 3	209287_s_at	ENSG00000070831	Cell division control protein 42 homolog	ACC
Position 2	221901_at	ENSG00000138944	KIAA1644	PUP
Position 3	208915_s_at	ENSG00000103365	Golgi-associated, gamma adaptin ear containing, ARF binding protein 2	IPT
Position 1	209602_s_at	ENSG00000107485	GATA3	TF
Position 2	202951_at	ENSG00000112079	Serine/threonine kinase 38	CG
Position 3	220443_s_at	ENSG00000116035	VAX2 (ventral anterior homeobox 2)	TF
	221955_at	ENSG00000088256	Guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	ITT
	207303_at	ENSG00000154678	Phosphodiesterase 1C, calmodulin-dependent 70kDa	ST

	205152_at	ENSG00000157103	Solute carrier family 6, member 1	ST
	207518_at	ENSG00000153933	Diacylglycerol kinase, epsilon 64kDa	ST
Position 2	206270_at	ENSG00000126583	Protein kinase C, gamma	ST
Position 3	208964_s_at	ENSG00000149485	Fatty acid desaturase 1	FAM
	201197_at	ENSG00000123505	Adenosylmethionine decarboxylase 1	CG
	201102_s_at	ENSG00000141959	ATP-dependent 6-phosphofructokinase, liver type	STM
	214972_at	ENSG00000198408	Protein O-GlcNAcase (Meningioma expressed antigen 5 (hyaluronidase))	ST
Position 2	210477_x_at	ENSG00000107643	Mitogen-activated protein kinase 8	CG
Position 3	205907_s_at	ENSG00000127083	Osteomodulin	STM
NETWORK 2				
Position 1	212195_at	ENSG00000134352	IL6 Signal Transducer	ICT
Position 2	202951_at	ENSG00000112079	Serine/threonine kinase 38	CG
Position 3	221955_at	ENSG00000088256	Guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	ITT
	207303_at	ENSG00000154678	Phosphodiesterase 1C, calmodulin-dependent 70kDa	ICT
NETWORK 3				
Position 1	212956_at	ENSG00000109436	TBC1 domain family, member 9 (with GRAM domain)	IPT
Position 2	202951_at	ENSG00000112079	Serine/threonine kinase 38	CG
Position 3	205152_at	ENSG00000157103	Solute carrier family 6, member 1	ST
	207518_at	ENSG00000153933	Diacylglycerol kinase, epsilon 64kDa	ST
Position 2	216814_at	ENSG00000232267	ACTR3 pseudogene 2	PUP
Position 3	221103_s_at	ENSG00000206530	Cilia and flagella associated protein 44	ACC

TABLE 6: *Biomarker network organisation* - breast cancer *data set - Estrogen Receptor - Breast Cancer.*
TF = Transcription/translation factor activity, DNA/RNA repair and catabolism - ER = estrogen receptor activity - APC = autophagy - protein catabolism - IR = immunity, inflammatory response (blood coagulation, antigen presentation and complement activation) - CC = cell/cell communication - ST = intracellular signal transduction, protein glycosylation - CG = cell growth and division - IPT = intracellular protein trafficking , transmembrane amino-acid transporter - ACC = actin activity, cytoskeleton organisation, cell projection - STM = sugar transport and metabolism - ITT = ion transmembrane transport, transmembrane signaling systems - PUP = pseudogene, uncharacterized protein - FAM = fatty acid metabolism. Source: www.uniprot.org; www.ncbi.nlm.nih.gov/gene