# Computing Molecular Signatures as Optima of a Bi-Objective Function: Method and Application to Prediction in Oncogenomics

Vincent Gardeux[1,2], Rachid Chelouah[1], Maria F. Barbosa Wanderley[3], Patrick Siarry[2], Antônio P. Braga[3], Fabien Reyal[4], Roman Rouzier[5], Lajos Pusztai[6] and René Natowicz[7,*]

[1]EISTI engineering school, Department of Computer Science, Cergy, France. [2]LISSI laboratory, University of Paris-Est, Créteil, France. [3]Federal University of Minas Gerais, Laboratório de Inteligência Computacional, Belo Horizonte, Brazil. [4]Curie Institute, Department of Translational Research, Paris, France. [5]Curie Institute, Department of Surgery, Paris, France. [6]Breast Medical Oncology, Yale School of Medicine, New Haven, CT, USA. [7]ESIEE-Paris, University of Paris-Est, Noisy-le-Grand, France.

**ABSTRACT**

**BACKGROUND:** Filter feature selection methods compute molecular signatures by selecting subsets of genes in the ranking of a valuation function. The motivations of the valuation functions choice are almost always clearly stated, but those for selecting the genes according to their ranking are hardly ever explicit.

**METHOD:** We addressed the computation of molecular signatures by searching the optima of a bi-objective function whose solution space was the set of all possible molecular signatures, ie, the set of subsets of genes. The two objectives were the size of the signature–to be minimized–and the interclass distance induced by the signature–to be maximized–.

**RESULTS:** We showed that: 1) the convex combination of the two objectives had exactly $n$ optimal non empty signatures where $n$ was the number of genes, 2) the $n$ optimal signatures were nested, and 3) the optimal signature of size $k$ was the subset of $k$ top ranked genes that contributed the most to the interclass distance. We applied our feature selection method on five public datasets in oncology, and assessed the prediction performances of the optimal signatures as input to the diagonal linear discriminant analysis (DLDA) classifier. They were at the same level or better than the best-reported ones. The predictions were robust, and the signatures were almost always significantly smaller. We studied in more details the performances of our predictive modeling on two breast cancer datasets to predict the response to a preoperative chemotherapy: the performances were higher than the previously reported ones, the signatures were three times smaller (11 versus 30 gene signatures), and the genes member of the signature were known to be involved in the response to chemotherapy.

**CONCLUSIONS:** Defining molecular signatures as the optima of a bi-objective function that combined the signature size and the interclass distance was well founded and efficient for prediction in oncogenomics. The complexity of the computation was very low because the optimal signatures were the sets of genes in the ranking of their valuation. Software can be freely downloaded from http://gardeux-vincent.eu/DeltaRanking.php

**KEYWORDS:** molecular signatures, bi-objective optimization, filter method, feature selection, breast cancer

**CITATION:** Gardeux et al. Computing Molecular Signatures as Optima of a Bi-Objective Function: Method and Application to Prediction in Oncogenomics. *Cancer Informatics* 2015:14 33–45 doi: 10.4137/CIN.S21111.

**RECEIVED:** October 23, 2014. **RESUBMITTED:** December 14, 2014. **ACCEPTED FOR PUBLICATION:** December 17, 2014.

**ACADEMIC EDITOR:** J.T Efird, Editor in Chief

**TYPE:** Methodology

**FUNDING:** This research was supported by the French-Brazilian program CAPES-COFECUB. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

**COMPETING INTERESTS:** Authors have disclosed no potential conflicts of interest.

***CORRESPONDENCE:** r.natowicz@esiee.fr



## Introduction

The important amount of data that are generated by high-throughput technologies increased the need for computational analysis.[1] Among these technologies, microarray chips are used to measure the expressions of tens of thousands mRNAs simultaneously. Supervised and unsupervised classifications are the two broad methods for analyzing the gene expressions measured by these devices. Supervised classification methods are the major tool for extracting quantitative information from the datasets: the analyst selects representative variables from a training dataset (*feature selection*), designs a classifier relying on these variables (*model design*), then assesses the predictor performances on independent datasets from similar platforms (*statistical validation*) or on the same dataset by cross-validation procedures. The whole process is termed as *predictive modeling*.[2]

In transcriptomic studies, a *molecular signature* is a set of genes whose expressions are predictive of a molecular class or a phenotype. From the machine learning perspective, classifier models are built from a particular signature to robustly assign the patients to their proper classes or phenotypes across a wide range of settings (eg, the patients' responses to specific chemotherapy treatments). Numerous methods for gene selection in genomics have been proposed (see[3] for an exhaustive survey).

These methods belong to three broad families of models: *filter methods* (either univariate or multivariate), *wrapper methods* and *embedded methods*.[1] Filter methods rank the features by a valuation criterion and retain the features whose values are above a fixed threshold. These methods are independent of the classification models used afterwards for delivering the predictions. Conversely, wrapper methods[4] search the set of features for optimal subsets of a given specific classifier. An objective function value is computed for each subset according to the performances of this particular classifier assessed on the learning dataset. Thus, the classifier is part of the valuation criterion, if not the valuation function itself. Finally, embedded methods perform the feature selection while designing the classification model. Typical examples are decision trees which have built-in mechanisms for performing feature selection while constructing the tree.

Regarding the possible use of these classifiers in clinical routine, it is of upmost importance to minimize the number of genes in the molecular signatures because predictors relying on small sized signatures would be easier to implement into cheap low-throughput technologies. Moreover, small signatures could be of higher value for biological investigations when gene interaction networks are the main concerns, because the complexity of the gene interaction networks is exponential in the networks' sizes.

In the present work we addressed the computing of molecular signatures as optimizing an explicit bi-objective function.[5] Because we were seeking for robust classifications and small molecular signatures, we aimed at both maximizing the interclass distance relative to the signature and minimizing the signature's size. Because the molecular signatures were sets of genes, the total number of potential molecular signatures was $2^n$ where $n$ was the number of genes. This astronomic number suggested to conduct the optimization process in the general frame of metaheuristic methods applied to gene selection.[5–7] A central point of our study was to show that computing the optima of the bi-objective function did not require searching the whole set of signatures. Rather, we showed that our bi-objective function had exactly $n + 1$ non-dominated optima (Pareto sets) that we could compute by ranking the probesets according to their contributions to the interclass distance. Therefore, the optimization problem could be solved by a non-combinatorial approach where both the gene valuation function and the gene ranking process were direct consequences of the bi-objective optimization. The second step was to compute, among $n + 1$ optimal signatures, the one whose size was optimal regarding the classifier model. We achieved this computation with a wrapper approach.[1,4] This computation too was non combinatorial because conducted in the very small set of $n + 1$ optimal signature (not in the huge set of all the signatures).

We assessed our predictive modeling on seven datasets in oncology. Five of them were used for benchmarking: we assessed the predictive performances of the optimal signatures and compared them to those of previous methods. We analyzed the performances on the two last datasets more thoroughly, paying attention to the biological relevancy of the optimal signatures, and comparing the signatures and prediction performances to previously reported methods and results. The two datasets came from a clinical trial of preoperative chemotherapy in breast cancer.

## Methods

**Definition of the objective function.** Let $R$ and $R'$ be the two classes distinguishing the samples (eg, the patient cases who were responder to a treatment and those who were not). For any signature $S$ (subset of variables, ie, subset of probesets) of size $|S| = m$, each data $p$ (a patient case) was represented by the real $m$-component vector whose values were the expression levels of the respective probesets in $S$. Now, let $c = c(S)$ and $c' = c'(S)$ be the centroids of the two classes $R$ and $R'$ for the signature $S$. We defined the interclass distance between $R$ and $R'$ as the Euclidean distance $d(c, c')$ between their centroids.

We aimed at maximizing the interclass distance and minimizing the size of the signature $S$, which are two conflicting objectives. Hence, we proposed to combine them into a single bi-objective function $F_w(S)$. We defined the bi-objective function $F_w(S)$ as a convex linear combination of the interclass distance and of the signature's size. The respective weights of the convex linear combination were $w$ and $(1 - w)$, $w \in [0, 1]$:

$$F_w(S) = w \times d^2(c(S), c'(S)) + (1-w) \times (1-|S|) \quad (1)$$

For any fixed weight $w$, the optimal solution was a signature $S^*(w)$ that maximized the function $F_w(S)$, ie:

$$S^*(w) = \arg \max_{S \in \mathcal{P}(\mathbb{S})} F_w(S) \quad (2)$$

where $\mathcal{P}(\mathbb{S})$ is the set of all the possible signatures.

Regarding the weight parameter $w$, the two limit cases were $w = 0$ at which $F_w(S) = 1 - |S|$ whose maximum was reached at $S^*(w) = \emptyset$, and $w = 1$ where $F_w(S) = d^2(c(S), c'(S))$, whose maximum was $S^*(w) = \mathbb{S}$ (the whole set of probesets). The function $F_w$ was bi-objective, except at each of these two limit cases, where it was mono-objective. The optimal solution $S^*(w)$ was the one that, given the weight parameter $w$, made the best balance between the two conflicting objectives. A large value of the parameter $w$ emphasized the interclass distance at the detriment of the signature's size, and conversely for a small value.

**Computing the optimal signature.** Let $S$ be any signature, $m$ its size, and let $S' = S \cup \{\sigma\}$ where $\sigma$ is a mRNA probeset not yet member of the signature $S$. Letting $\delta(\sigma)$ be the contribution of probeset $\sigma$ to the interclass distance:

$$\delta(\sigma) = \sqrt{(\bar{\sigma} - \bar{\sigma}')^2} \quad (3)$$

where $\bar{\sigma}$ and $\bar{\sigma}'$ are the mean values of the probeset's expressions on the two classes, we considered the difference between the values of the bi-objective function at $S'$ and $S$:

$$\begin{aligned} F_w(S') - F_w(S) &= w \times \left( \sum_{S \in S'} \delta^2(s) - \sum_{S \in S} \delta^2(s) \right) \\ &+ (1-w) \times \left( (1-(m+1)) - (1-m) \right) \\ &= w \times \delta^2(\sigma) - (1-w) \end{aligned} \quad (4)$$

The difference was positive if and only if $\delta^2(\sigma) > \frac{1-w}{w}$, a condition that did not depend on the signature $S$ the probeset $\sigma$ was added to. Hence, for any signature $S$ and any probeset $\sigma$ such that $\delta^2(\sigma) > \frac{1-w}{w}$, one had $F_w(S') > F_w(S)$. Conversely, for any probeset $\sigma$ such that $\delta^2(\sigma) \leq \frac{1-w}{w}$, one had $F_w(S') \leq F_w(S)$. Therefore, the optimal signature $S^*(w)$ was

$$S^*(w) = \left\{ \sigma \in \mathbb{S}; \delta^2(\sigma) > \frac{1-w}{w} \right\} \quad (5)$$

Because the weight parameter $w$ belonged to the real interval [0, 1], any augmenting probeset of the function $F_w$ was an augmenting probeset of all function $F_{w'}$ s.t. $w' > w$ since:

$$\delta^2(\sigma) > \frac{1-w}{w} \text{ and } w' > w \Rightarrow \delta^2(\sigma) > \frac{1-w'}{w'} \quad (6)$$

Hence, the optimal signatures followed an inclusion property:

$$w' > w \Rightarrow S^*(w') \subseteq S^*(w) \quad (7)$$

Now let the set $\mathbb{S}$ of all the probesets be ranked in decreasing order according to their contributions $\delta$ to the interclass distance:

$$\mathbb{S} = \{s_1, s_2, \ldots\} \text{ s.t. } i < j \Rightarrow \delta(s_i) \geq \delta(s_j) \quad (8)$$

and, for each probeset $s_k$, let us define the value $w_k = \frac{1}{\delta^2(s_k)+1}$, equation $\delta^2(s_k) = \frac{1-w_k}{w_k}$. The inclusion property of the signatures $S^*(w)$ led to the following property: "*The set $S(k)$ of the $k$ top probesets in the $\delta$ ranking is an optimal solution of any function $F_w$ $w \in [w_k, w_{k+1}]$*". Hence, the set $W$ of weight values $w$ leading to different optimal signatures was finite:

$$W = \{w_1, w_2, \ldots, w_n\} \quad (9)$$

Otherwise stated, among the weight values of the real unit interval [0, 1] the only values of interest were those of the finite set $W$.

This result holds for our bi-objective function which is the convex combination of the signature size and interclass distance. Other combinations of objectives would make sense. Which of them share the optimal signature inclusion property is an open question that we do not address in the present article.

**Automatic selection of the predictor's optimal size.** The above analysis shows that the optimal solutions of $F_w(S)$ (eq. (1)) are the $n + 1$ subsets of probesets in the ranking of the contributions to the interclass distance δ. This first stage computed the $n + 1$ optimal subsets of probesets. This computation was independent of the classifier model chosen for the prediction. By contrast, finding the signature whose size was optimal among the $n + 1$ optimal signatures required a wrapper approach. This second stage relied on the classifier model.

An in-depth study of the prediction of the response to preoperative chemotherapy in breast cancer[8] evaluated a total of 780 distinct classifiers (sets of genes and classifier models). The conclusions were that the Diagonal Linear Discriminant Analysis (DLDA) classifier had better prediction performances for gene expression data[9] and higher sensitivities.[10] For these reasons we decided to use DLDA for prediction modelling, taking as input the expressions of the genes belonging to the signatures that were the optima of the bi-objective function.

The classifier model being chosen, we selected the optimal-sized signature by a wrapper approach, searching for the signature whose accuracy (see *Terminology and Abbreviations*) was maximal on the learning set of samples. Because the set of optimal signatures was of size $n + 1$, the wrapping process assessed the performances of the DLDA classifier of $n + 1$ signatures, which in terms of computational complexity was far less than searching the whole set of signatures (of size $2^n$). The wrapping procedure can be summarized as follows: for each of the $n + 1$ optimal signatures, we computed the DLDA model and its accuracy on the learning set. We defined the optimal predictor as the smallest non-singular signature of highest accuracy. Qualitatively, this non-singular predictor was the smallest predictor whose accuracy was maximal and which was not an outlier, ie, had comparable performances with predictors using similar signatures. By considering nonsingular predictors of highest accuracy one wanted to avoid selecting a predictor whose high accuracy was the consequence of an overfitting of the data, which would have led to non-robust performances. More precisely, we defined a quasi-plateau as a set of predictors of size $k$, $k \in [l, u]$, such that the accuracy of each of them was not lower than a fixed threshold value ($l$ and $u$ were the respective lower and upper bound of the quasi-plateau). The quasi-plateau had exactly $u - l + 1$ different predictors whose accuracies were all above the fixed threshold value. In all the experiments reported in the present article, the threshold value was set to 95 percent of the maximum accuracy value measured on the learning set of cases. On a quasi-plateau of length greater than or equal to 2

(ie, $u - l + 1 \geq 2$), we define the size of the locally optimal non-singular predictor as $k^*_{[l,u]} = l$. Then, the optimal predictor was the smallest locally optimal non-singular predictor of highest accuracy. Its size was the minimum of the $k^*_{[l,u]}$ values.

**Datasets.** We compared the performances of our predictors to those of the major published ones on seven different datasets in oncology. The purpose of this benchmarking was to assess the relevancy and predicting power of the optimal signatures our method unveiled. To this end, we computed the signatures on a significant number of different prediction problems in oncogenomics and compared their performances (using a DLDA classifier) to the state-of-the-art published predictive modelings. Each of them was a two-class dataset, whose characteristics are summarized in Table 1. First, we used five datasets of tumors in different tissues (Datasets I-V) to evaluate the performance of our predictors. We invite the reader to refer to the cited articles for further explanations of the different methods and protocols that were used in the reported studies. In a second stage, we conducted a detailed study on two breast cancer datasets (Datasets VI-VII).

**Cross validation.** Three different cross validation techniques were used in this study: 1) $k$–FOLD CROSS-VALIDATION ($k$-fold CV) which consists of splitting the dataset into $k$ sets of equal length, choosing $k - 1$ sets as training sets for building the model, and testing the performances on the remaining set. This training/testing procedure was performed times with the $k$ unique different combinations, and the average performances were computed. 2) REPEATED RANDOM SUB-SAMPLING is a cross-validation procedure very similar to the $k$-fold CV. The main difference is that instead of splitting the dataset only once then assessing the performances on the randomly created subsets, it splits the dataset randomly several times. Therefore the average performances can be computed on a broader number of simulated subsets, enhancing the robustness of the results. 3) LEAVE-ONE-OUT CROSS-VALIDATION is a specific case of the $k$-fold CV, taking $k$ as the total number of samples, which implies that the model is built on $k - 1$ samples and tested on the last one, this operation being repeated times. Since there are only partitions of $k - 1$ samples, this method is run only once, and the total number of FP, TP, FN, and TN (whose sum is $k$) gives the performance results.

**Experimental protocol (non-biased).** In all of the aforementioned experiments, the signatures, their sizes and the parameters of the classification model were computed on the training set only, without any reference to the test sets.[11,12] Then the performances of the model designed in the training phase were computed on the external test sets. The same protocol was used in the cross-validation procedures: at each run of the cross-validation, the whole predictive modeling was repeated on the new independent training set of samples.

## Results

We have applied our predictive modeling (Methods: Experimental Protocol (Non- biased)) to seven different microarray datasets in oncology. Five of them were used for benchmarking, and two of them for a thorough analysis of breast cancer signatures.

**Evaluation of the performance of the predictors unveiled by the bi-objective optimization.** We first performed a benchmarking on different datasets to objectively assess the performances of the signatures predicted by the bi-objective optimization. To this end, we have applied our predictive modeling to Datasets I-V (Methods: Datasets) and conducted two kinds of cross-validation procedures (Methods: Cross-validation): 1) a three-fold cross-validation whose results are reported in Table 2, and 2) a leave-one-out cross-validation procedure whose results are in Table 3. The results outlined three interesting features: 1) the signature sizes (between 3 and 10 probesets in average), were rather small compared to what was commonly reported in the literature (most of the time containing between 20 and 50 genes); this property being exemplified in Table 4 where we compared different published classifiers to our model, 2) the performances were high and concordant (high accuracy, with both high specificity and sensitivity, see Terminology and Abbreviations). Only the brain cancer dataset seemed to harbor relatively weaker performances, but one still had ~68% accuracy with signatures of ~10 probesets. 3) The two different cross validation protocols had very similar results, which strengthen the confidence in the robustness of the predictions.

**Table 1.** Description of the seven publicly available cancer datasets used in this study.

| DATASET | DATA TYPE | ARTICLE | # CASES | # PROBESETS | SOURCE |
|---|---|---|---|---|---|
| Dataset I | Colon | [29] | 62 | 2000 | (1) |
| Dataset II | Lymphoma | [30] | 77 | 5469 | (2) |
| Dataset III | Leukemia | [31] | 72 | 7129 | (3) |
| Dataset IV | Prostate | [32] | 102 | 10509 | (2) |
| Dataset V | Brain | [33] | 60 | 7129 | (3) |
| Dataset VI | Breast | [8] | 133 | 22283 | (4) |
| Dataset VII | Breast | [16] | 91 | 22283 | GSE20271 |

**Notes:** (1) http://genomics-pubs.princeton.edu/oncology/ (2) http://www.gems-system.org/ (3) http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi (4) http://bioinformatics.mdanderson.org/main/Public_Datasets

**Table 2.** Average performances and size of the signatures predicted by bi-objective function optimization; three-fold cross-validation. We computed the $\delta$-filtered signatures 100 times for different random training/testing subsets on Datasets I-V (Methods: 3-fold CV). Average performances across the runs are reported along with their standard deviation.

| | COLON | LYMPHOMA | LEUKEMIA | PROSTATE | BRAIN |
|---|---|---|---|---|---|
| #Probesets | 8.420 ± 5.459 | 4.700 ± 2.816 | 3.100 ± 1.187 | 3.340 ± 0.764 | 10.420 ± 5.668 |
| Accuracy | 0.825 ± 0.018 | 0.872 ± 0.017 | 0.961 ± 0.008 | 0.909 ± 0.009 | 0.672 ± 0.038 |
| Sensitivity | 0.867 ± 0.029 | 0.872 ± 0.036 | 0.937 ± 0.012 | 0.904 ± 0.008 | 0.584 ± 0.084 |
| Specificity | 0.747 ± 0.013 | 0.889 ± 0.015 | 0.974 ± 0.010 | 0.913 ± 0.015 | 0.718 ± 0.022 |
| PPV | 0.865 ± 0.006 | 0.702 ± 0.035 | 0.950 ± 0.020 | 0.524 ± 0.040 | 0.650 ± 0.040 |
| NPV | 0.752 ± 0.043 | 0.939 ± 0.012 | 0.967 ± 0.006 | 0.910 ± 0.007 | 0.766 ± 0.045 |

**Table 3.** Average performances and size of the signatures predicted by bi-objective function optimization; leave-one-out cross-validation. We computed the $\delta$-filtered signatures for different random training/testing subsets on Datasets I-V (Methods: Leave- one-out CV). Performances are computed from the summary of the runs.

| | COLON | LYMPHOMA | LEUKEMIA | PROSTATE | BRAIN |
|---|---|---|---|---|---|
| #Probesets | 9.048 | 4.156 | 2.972 | 4.000 | 8.267 |
| Accuracy | 0.855 | 0.883 | 0.986 | 0.941 | 0.683 |
| Sensitivity | 0.925 | 1.000 | 0.960 | 0.960 | 0.619 |
| Specificity | 0.727 | 0.845 | 1.000 | 0.923 | 0.718 |
| PPV | 0.860 | 0.679 | 1.000 | 0.923 | 0.542 |
| NPV | 0.842 | 1.000 | 0.979 | 0.960 | 0.778 |

**Table 4.** Comparison of average performances and size of signatures reported in the literature. In this table are reported the mean accuracies (Ac., in percent) and mean number of probesets ($\bar{p}$) of non-biased results (Methods: Experimental Protocol (Non-biased)) published for the five benchmarking datasets (Datasets I-V). Results of our predictive modeling are reported on the last line.

| ARTICLES | COLON | | LYMPHOMA | | LEUKEMIA | | PROSTATE | | BRAIN | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Ac.* | $\bar{p}$ | *Ac.* | $\bar{p}$ | *Ac.* | $\bar{p}$ | *Ac.* | $\bar{p}$ | *Ac.* | $\bar{p}$ |
| [32] | – | – | – | – | – | – | 86.00 | 29 | – | – |
| [34] | – | – | – | – | – | – | – | – | 60.00 | 21 |
| [35] | 85.83 | 20 | – | – | – | – | – | – | – | – |
| [36] | – | – | 83.33 | 6 | – | – | – | – | – | – |
| [37] | 82.33 | 20 | – | – | – | – | – | – | – | – |
| [38] | 82.03 | (*) | – | – | 94.40 | (*) | 91.22 | (*) | – | – |
| [39] | – | – | – | – | – | – | 94.12 | 22 | – | – |
| [40] | 85.71 | 30 | – | – | – | – | 94.11 | 20 | – | – |
| [13] | | | | | | | | | | |
| F-test | 84.05 | 15.1 | – | – | – | – | 91.18 | 126.4 | – | – |
| *∂W* | 76.70 | 35.1 | – | – | – | – | 94.60 | 756.6 | – | – |
| *∂RW* | 78.60 | 43.3 | – | – | – | – | 94.70 | 573.3 | – | – |
| *∂Spb* | 80.30 | 31.8 | – | – | – | – | 94.80 | 95.5 | – | – |
| SVM-RFE | 85.48 | 26.4 | – | – | – | – | 94.18 | 43.2 | – | – |
| GLMPath | 81.91 | 1.3 | – | – | – | – | 94.09 | 1.6 | – | – |
| Random Forest | 89.40 | 49.8 | – | – | – | – | 94.10 | 81 | – | – |
| **δ-DLDA** | **85.50** | **9.05** | **88.30** | **4.16** | **98.60** | **2.97** | **94.10** | **4.00** | **68.30** | **8.27** |

**Note:** (*) Not in the article.

Next, we compared our results to previously published ones that we have selected because they were non-biased (Methods: Experimental Protocol (Non-biased)). The comparative results are reported in Table 4. For the colon cancer dataset, our predictive modeling was only outperformed by the random forest (accuracy ~89% vs. ~85%) but the sizes of these predictors were five times larger (~50 vs. ~10 probesets). For the leukemia dataset, our predictive modeling was not outperformed (accuracy ~98% with ~3 variables). Regarding the prostate cancer dataset, our modeling was not outperformed. GLMPath method had the same performances (accuracy ~94%) and was twice smaller (~2 vs. ~4 variables). Finally, our model was not outperformed in both the brain cancer (accuracy ~68% with ~9 variables) and the lymphoma datasets (~88%, ~5 variables).

Hence, the performances of our predictors were at the level of, or higher than those of the previous studies. Moreover, our signatures were most of the time significantly smaller. A noticeable exception is GLMPath[13] whose performances, even though slightly lower than ours, were obtained with remarkably small signature sizes. Our predictive modeling appeared to be among the two best methods, if not the best one. Beside this result, one should stress that our predictive modeling is very simple, quick to compute on commonly available personal computers, and fully automated (no parameter to tune).

**Application in breast cancer: molecular signature designed for prediction of preoperative chemotherapy treatments.** After having assessed the performances of the predictive modeling we propose, we applied it on two breast cancer datasets: Datasets VI & VII (Methods: Datasets).

Dataset VI comes from a clinical trial conducted at the MD Anderson Cancer Center (Houston, Texas, USA). One of the purposes was to provide data for designing predictors of the response to preoperative chemotherapy treatments in breast cancer. We used the same protocol as in[8]: the set of 133 patient cases was split into a fixed training set of 82 patient samples and a test set of 51 patient samples, each one showing the same ratio of responder (Pathologic Complete Response, PCR) and partially responder patient cases (NoPCR or Residual Disease, RD), respectively 1/3 and 2/3.

According to our predictive modeling, the probesets were ranked by decreasing contributions to the interclass distance measured on the learning set of 82 patient cases[8] and the signatures were computed in this ranking. In Table 5, we compared the performances of our predictive modeling to two other recently published classifiers. We computed two signatures containing respectively 30 ($\delta$-DLDA-30) and 11 ($\delta$-DLDA-11) probesets. The $\delta$-DLDA-30 signature was created by taking the 30 top ranked genes according to their $\delta$ score, and was built for comparing its performance to the two published classifiers of same size. The $\delta$-DLDA-11 signature was built by the automated procedure which chooses the optimal size of the signature on the training dataset without human supervision (see Methods). The results of our predictive modeling were compared to two published studies on the same datasets: 1) a predictor using the same DLDA classifier (DLDA-30) designed on the 30 probesets of smallest *P*-value to a *t*-test,[8] and 2) a signature unveiled by a majority voting predictor (bi-majority-30) designed on the 30 probesets of highest bi-informative values[14] (Terminology and Abbreviations). The $\delta$-DLDA-11 signature had the best performances on the test dataset, and they were even slightly higher than the best performances ever published for this dataset. The main point was that these performances were achieved with three times less probesets (11 vs. 30).

**Table 5.** Comparison of the performances of signatures predicted on breast cancer dataset VI (training set=82 samples). The first column of this table ($\delta$-DLDA-30) corresponds to the result of our predictive modeling with a fixed size of 30 probesets (for direct comparison with other methods). The second column ($\delta$-DLDA-11) contains the results of our predictive modeling obtained without fixing the number of probesets of the signature. The optimal non-singular predictor found by our method contained 11 probesets. The third (DLDA-30[8]) and fourth (Bi-Majority-30[14]) columns report results found in the literature with the same data/protocol.

| | $\delta$-DLDA-30 | $\delta$-DLDA-11 | DLDA-30 | Bi-Majority-30 |
|---|---|---|---|---|
| Accuracy | 0.863 | 0.882 | 0.765 | 0.863 |
| Sensitivity | 0.846 | 0.923 | 0.923 | 0.923 |
| Specificity | 0.868 | 0.868 | 0.711 | 0.842 |
| PPV | 0.688 | 0.706 | 0.522 | 0.667 |
| NPV | 0.943 | 0.971 | 0.964 | 0.970 |

Figure 1 shows heatmaps for each of the four signatures cited above, applied on the testing Dataset VI (51 test samples). We can observe that the signatures unveiled by our method (Panels A&B) clearly dichotomize the responder patients. The up-regulated and down-regulated genes are clearly visible, while this seems less apparent for the two other signatures found in the literature (Panels C&D): the classifier rules would probably be more complicated to analyze/interpret. Interestingly, we found that patients PERU11, PERU14, M120, M353 and M503 were misclassified by all the methods. This can suggest that these five patients form one or several sub-clusters regarding the response to therapy.

In order to put more focus on the biological relevancy of the genes unveiled in the $\delta$-DLDA-11 signature, we detailed the characteristics of the 11 selected probesets in Table 6 and plotted the boxplots of their expression levels in Figure 2. Table 6 unveils a relatively high overlap between our signature and the two other published ones. Such compliance across studies is rarely observed in the literature and is worth mentioning. The highest ranked gene of our signatures was ESR1 (estrogen receptor protein coding gene 1). Although this gene is known to be one of the most determinant marker of the response to the chemotherapy (eg,[15]), it had not been selected in neither[8] nor.[14] Beside ESR1 three other genes of the $\delta$-DLDA-11 signature were neither in

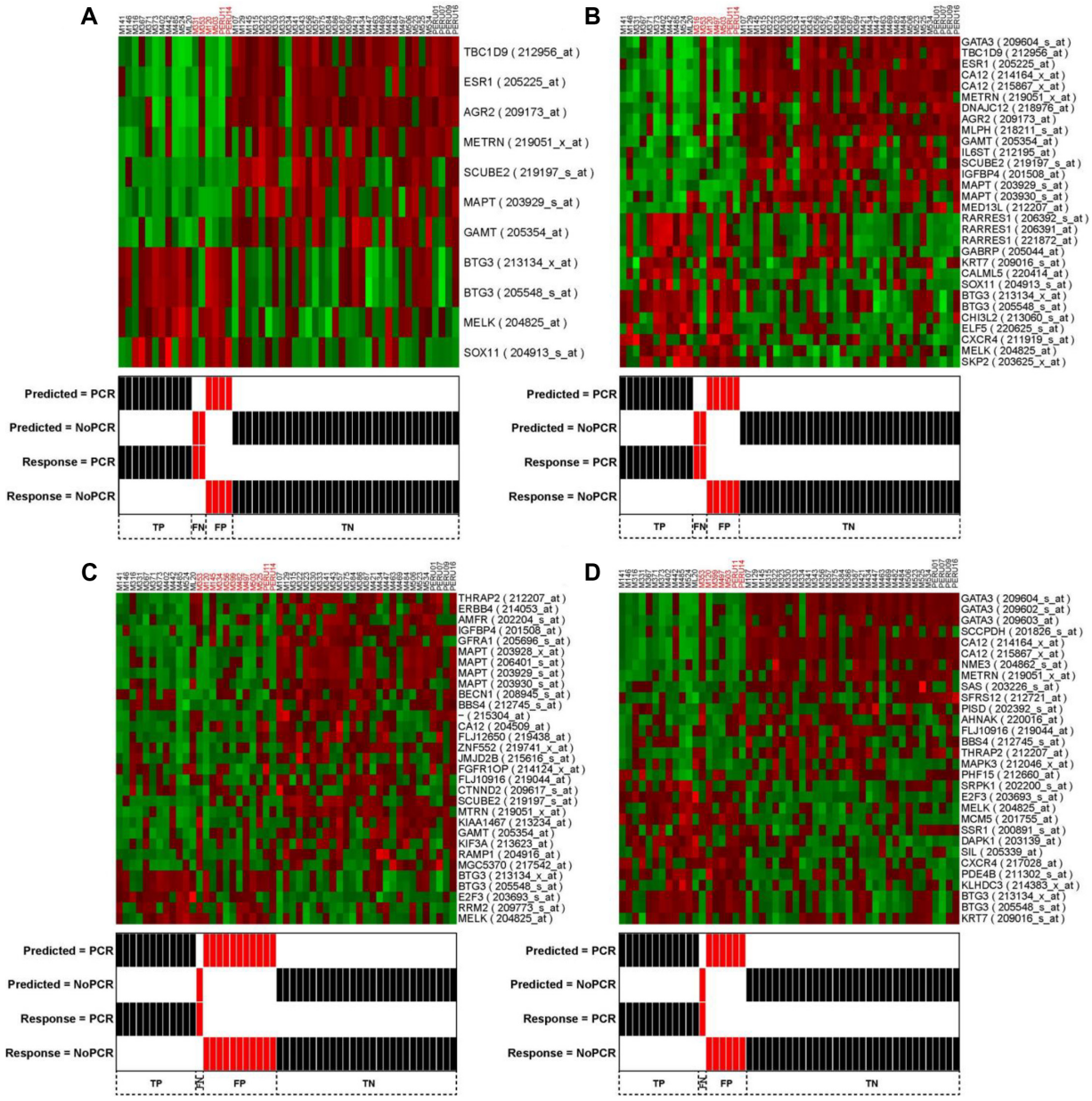


**Figure 1.** Heatmaps of the four signatures on testing data (51 patients).

**Notes:** The four heatmaps represent, for each of the 51 patients of the breast cancer testing set (in columns), the different genes (in rows) of each of the four molecular signatures detailed in this paper. Green colors represent down-regulated genes, and red colors represent up-regulated genes. In parenthesis are the names of the corresponding probesets in the Affymetrix microarray. Panel A corresponds to the 11 genes of the $\delta$-DLDA-11 signature. Panel B corresponds to the 30 genes of the $\delta$-DLDA-30 signature. Panel C corresponds to the 30 genes of the DLDA-30 signature.[8] Panel D corresponds to the 30 genes of the Bi-Majority-30 signature.[14] At the bottom of each heatmap two variables are represented: "Predicted" is the predicted response for each patient using the DLDA classifier, and "Response" is the true response class (PCR: Pathologic Complete Response, or NoPCR: residual disease). In the subpanel, the red bars represent the misclassified patient (False Positives or False Negatives).

the DLDA-30 signature nor in the bi-majority-30 one: AGR2, SOX11, and TBC1D9 (see Discussion section).

Dataset VII is a second independent dataset of 91 patient cases[16] coming from a cohort subject to the same clinical trial and protocol as Dataset VI.[8] Measurements of the expression levels were conducted with the same microarray (Affymetrix U133A). With this dataset we aimed at validating the performances of the predictors we computed on Dataset VI (82 training data samples). Hence, we only used this new dataset as a test dataset. The performances of our predictors are reported in Table 7 together with those of the DLDA-30[8] and bi-majority-30[14] signatures. The performances of the $\delta$-DLDA-predictors were close to those of the two other predictors. More specifically, the $\delta$-DLDA-11 predictor's specificity was slightly lower than that of the DLDA-30 but the two predictors had the same sensitivity value, and almost the same high negative predictive values, ie, the probability of a partial response given that the prediction is a partial response to the treatment.

**Table 6.** Detailed caracteristics of the $\delta$-DLDA-11 signature. This table contains descriptions of the 11 probesets of highest contributions to the interclass distance unveiled by our predictive modeling on Dataset VI (rank of the probesets following our prioritization by $\delta$ scores, names of the targeted genes, Affymetrix references of the probesets, values of their contributions to the interclass distance, and *P*-values to a *t*-test). In bold are the genes of the 11 $\delta$-signature that were neither member of DLDA-30 nor bi-majority-30 signatures.

| $\delta$-RANK | GENE | PROBESET | $\delta(\sigma)$ | *P*-VALUE |
|---|---|---|---|---|
| 1 | **ESR1** | 205225 at | 0.102 | 5.261E-6 |
| 2 | BTG3 | 213134_x_at | 0.090 | 2.956E-5 |
| 3 | BTG3 | 205548_s_at | 0.088 | 3.307E-5 |
| 4 | MELK | 204825_at | 0.083 | 1.224E-4 |
| 5 | METRN | 219051_x_at | 0.076 | 1.705E-6 |
| 6 | GAMT | 205354_at | 0.075 | 2.768E-7 |
| 7 | MAPT | 203929_s_at | 0.074 | 2.312E-8 |
| 8 | **AGR2** | 209173_at | 0.073 | 5.451E-7 |
| 9 | **SOX11** | 204913_s_at | 0.073 | 8.000E-3 |
| 10 | **TBC1D9** | 212956_at | 0.072 | 2.037E-5 |
| 11 | SCUBE2 | 219197_s_at | 0.071 | 4.736E-5 |

Finally, we used the two test datasets to assess more deeply the robustness of the $\delta$-DLDA-11 predictor. In this last experiment, the signature was kept constant while the parameters of the DLDA classifier were tuned at each run of the 3-fold cross-validation procedure (Methods: Cross-Validation). The results of these two distinct cross-validations are reported in Table 8. Neither DLDA-30[8] nor bi-majority-30[14] signatures reported cross-validation assessment. Both results found on Dataset VI (51 test samples) and Dataset VII (91 test samples) are concordant with (and thus reinforce) the ones found without cross-validation.

**Global assessment of the robustness of the method.** We statistically assessed the robustness of our predictive modeling on the 133 cases dataset[8] by means of permutation resampling procedures. We first assessed that our predictor was better than a random predictor. To this end, we performed 1,000 random subsampling cross-validations (Methods: Cross-Validation), and conducted our predictive modeling twice at each run: once on the learning set of samples, leading to a first predictor, and once on the same learning set where the samples were randomly assigned to the classes. The null hypothesis was *"The performances of the two predictors are equal"*. It was strongly rejected in each of the measures of performance (sensitivity, specificity, accuracy, PPV, NPV; see Terminology and Abbreviations; data not shown). This result demonstrated that the predictive modeling comprising both the signature selection and the classifier model (DLDA), was not random guesses.

Next, we assessed that this result was not the consequence of the classifier model alone or, otherwise stated, that the signatures were relevant for the predictions. To this end, we have assessed the optimal signatures by designing another random subsampling cross-validation procedure (1000 runs; Methods: Cross-Validation) conducted on the same dataset. At each run the predictive modeling was conducted, together with the design of a predictor with a random signature of same size. The null hypothesis was stated as *"The performances of the predictors with optimal and random signatures of same sizes are equal."* This null hypothesis was strongly rejected as well (data not shown). This last result demonstrated that the performances of the predictive modeling were not the consequence of the DLDA alone, ie, the optimal signatures were relevant for the data at hand.

## Discussion

In this study we proposed a feature selection method by which the molecular signatures were the optima of a bi-objective function expressing the tradeoff between the class discrimination (interclass distance) and the size of the signature. We demonstrated that this optimization was reducible to ranking the probesets by their contributions to the distance between the class centroids. This method of signature selection is non parametric because the contribution to the interclass distance makes no assumption on the distribution of the expression levels. Moreover, we have proposed a fully automated method for computing the optimal sized signature on a training dataset without supervision.

In this study we used the DLDA classifier method, following the conclusions of an in-depth study that compared several predictive modelings.[8] Other classifier methods could have been used such as SVM, Naïve Bayes, Random Forest, etc. Investigating the effects or performances of each classifier model on the molecular signatures unveiled by our feature selection method would be interesting. However, since we used the classifier model after the feature selection step, and according to the referenced study,[8] we may reasonably assume that the difference should be small or in favor of the DLDA classifier results.

In this study we defined the interclass distance as the Euclidean distance between the centroids of the two classes. This definition implies that we used two specific parameters for defining our objective function: 1) the "representative" element of each class, that we here defined as the centroid of all the samples pertaining to this class, and 2) the choice of the Euclidean distance to measure the difference between the two classes "representative" elements. Arguably, other representative elements can be used, such as the vector of medians, the medoids,[17] etc. However, since the distance between the representative elements will always increase with the dimensionality (ie, the number of probesets included in the signature), this choice will not change the conclusions of our study. This choice have the potential to increase (or decrease) the performances of the predictors,[18] but the global methodology will remain the same.

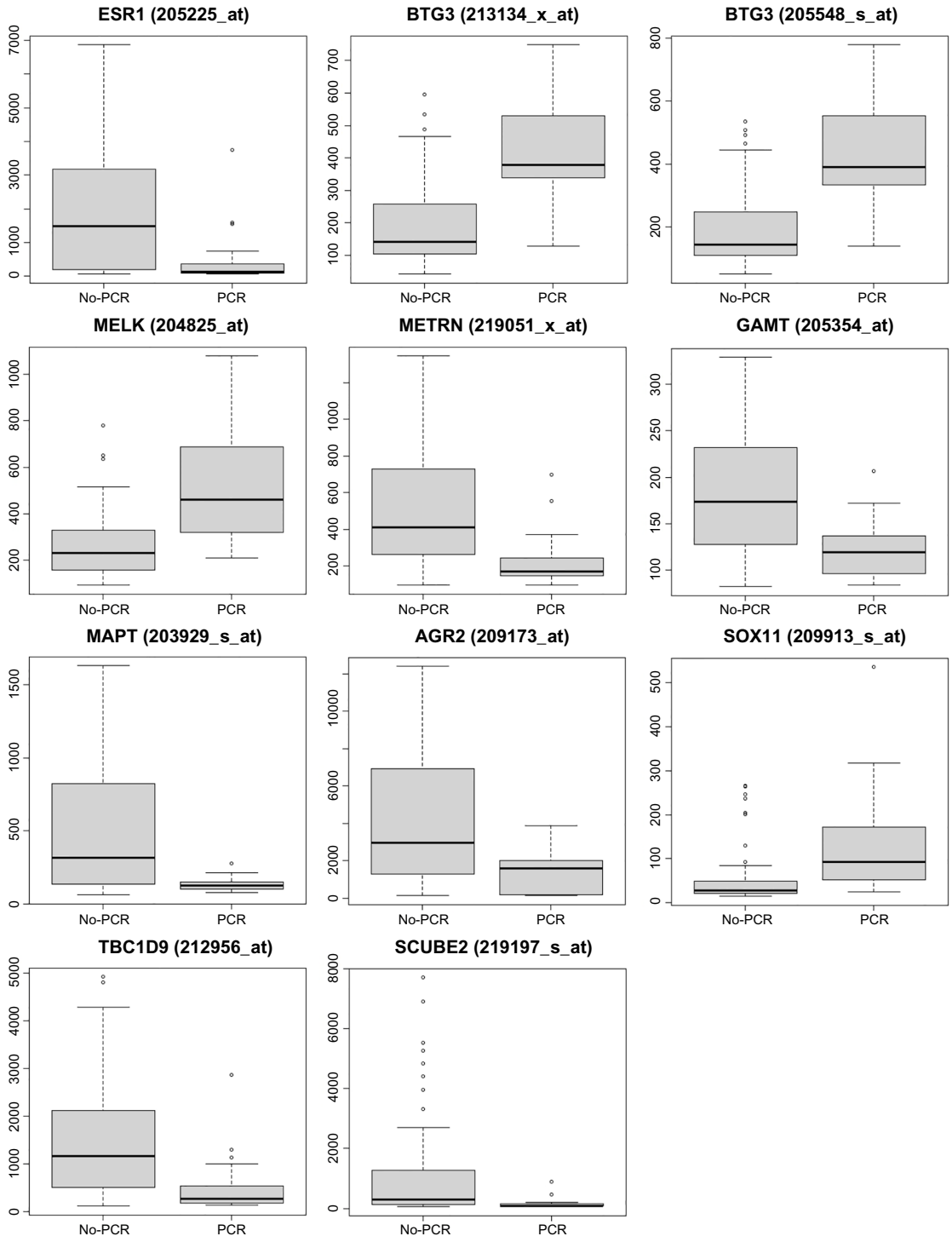


**Figure 2.** Boxplots of the expressions of the 11 probesets of the $\delta$-DLDA-11 signature.
**Notes:** The boxplots represent the expression values of the 11 probesets selected by our predictive modeling for breast cancer prediction of the response to preoperative chemotherapy.
**Abbreviations:** PCR, pathologic complete response; No-PCR, residual disease.

Concerning the bi-objective function choice, we can also stress out the fact that this choice could have been different. Indeed, concerning the second objective, the minimization of the size is produced by the right part of equation (1): $1 - |S|$, and could have been chosen as any other decreasing function of $|S|$. The probesets being indexed by decreasing contribution to the interclass distance, the scores of the $n + 1$ optima would be different, but the probesets within those $n + 1$ optima would remain the same. Therefore, since the optimum size determination is conducted on the set of probesets, by decreasing $\delta$ cutoffs, it would not change the accuracy of the classifier either. Furthermore, the first objective of the bi-objective

**Table 7.** Performances of the predictors trained on breast cancer Dataset VI (82 training samples) and applied on Dataset VII (91 test samples). In the first column ($\delta$-DLDA-30) are the performances of our predictive modeling with the 30 probesets signature predicted on Dataset VI (for direct comparison with other methods). In the second column ($\delta$-DLDA-11) are the results of our predictive modeling obtained with the optimal non-singular DLDA predictor (11 probesets signature) predicted on Dataset VI. The third (DLDA-30[8]) and fourth (Bi-Majority-30[14]) columns report the performances of two predictors whose signatures were the 30 probesets of smallest *P*-values to the Student *t*-test (DLDA-30) and the 30 probesets of highest *bi-informativeness* (Bi-Majority–30).

| | $\delta$-DLDA-30 | $\delta$-DLDA-11 | DLDA-30 | Bi-Majority-30 |
|---|---|---|---|---|
| Accuracy | 0.670 | 0.659 | 0.725 | 0.681 |
| Sensitivity | 0.632 | 0.632 | 0.632 | 0.579 |
| Specificity | 0.681 | 0.667 | 0.750 | 0.708 |
| PPV | 0.343 | 0.333 | 0.400 | 0.343 |
| NPV | 0.875 | 0.873 | 0.885 | 0.864 |

**Table 8.** Three-fold cross-validation average performances of the $\delta$-DLDA-11 signature, applied on the two external test datasets. The table contains the results of our predictive modeling obtained with the 11 probesets signature ($\delta$-DLDA-11) predicted on Dataset VI (82 training samples). The predictor was applied on both test datasets following a 3-fold cross-validation experimental protocol (Methods: 3-fold cross-validation). Average performances and standard deviations are reported.

| | DATASET VI (51 TEST SAMPLES) | DATASET VII (91 TEST SAMPLES) |
|---|---|---|
| Accuracy | 0.84 ± 0.12 | 0.61 ± 0.15 |
| Sensitivity | 0.92 ± 0.18 | 0.72 ± 0.30 |
| Specificity | 0.81 ± 0.16 | 0.58 ± 0.17 |
| PPV | 0.68 ± 0.22 | 0.31 ± 0.15 |
| NPV | 0.97 ± 0.06 | 0.90 ± 0.11 |

function does not require being a distance metric. "In regards to the analysis that we have conducted, the first objective only needs to fulfill the two following requirements: its value must be non-decreasing in the size of the signature and the contribution of a variable must not depend on the set of variables it is added to. To illustrate this second point by a counter-example, would the first objective take into account the correlations of the variables, the contribution of a variable would depend on the set it is added to. In such situations our method would not be applicable. Such problems are to be solved in the general frame of metaheuristic optimization procedures.[5]

We applied our method on microarray data, but it could as well be applied to RNA-Seq, methylation or other type of data. The complexity of the signatures computation is that of assigning each gene its δ value (this step is in linear time) plus the complexity of sorting the genes according to this score (this step is in log-linear time). This first step of finding the n + 1 optimal signatures is performed very fast. It drastically reduces the dimension of the search space, so that the computation of the optimal-sized signature is very efficient. Hence, the final wrapper step for selecting the best signature is performed among this small subset of n + 1 optimal signatures. It implies that the complexity of the whole process is low (close to log-linear). This efficiency makes the method a good candidate for predictor design on Next Generation Sequencing data.

Our results are supported by a study concerning expression data of breast cancer tumors.[19] The authors have assessed three factors underlying the ability of predicting phenotypes in breast cancer: 1) the presence of genes whose expressions had high fold changes between the different classes, 2) the amount of such genes in the microarray and 3) the number of learning cases at which these genes had strong different mean expressions.

The conclusion of this study was that the fold-change had the greatest influence on the success of model building, followed by the size of the gene signature then by the number of informative cases. Our molecular signatures are designed by selecting the genes in the ranking of their contributions to the interclass distance. Because this ranking and that of the fold-change are equivalent, our method and result are coherent with the study.[19]

In,[20] the authors compared and assessed a large panel of feature selection methods that are available for DNA microarray studies. The conclusion of the study was: "*Surprisingly, complex wrapper and embedded methods generally do not outperform simple univariate feature selection methods, and ensemble feature selection has generally no positive effect. Overall a simple Student's t-test seems to provide the best results.*" This statement supports our predictive modeling design. It is a simple filter methods based on the ranking of the probesets by their contributions to the interclass distance, ie, by the difference of the means between the two classes. It has similarities with a *t*-test, but does not take into account the standard deviations or the distribution of the data. Figure 3 highlights this result, plotting the contributions to the interclass distances ($\delta$-values on the X axis) versus the *P*-values to the Student's *t*-test (Y axis). One can see that the probesets of highest $\delta$-values are precisely those of smallest *P*-values to the Student's *t*-test. This result is general to microarrays because the variance of the expression levels is independent of the fold-change between the two classes. Hence, selecting the genes of highest contribution to the interclass distance implies selecting the genes of smallest *P*-values. The reverse is not true: on this figure one can see that probesets of low $\delta$ rankings may be of small *P*-values to the Student *t*-test (eg, at $\delta = 0.10$ where the *P*-values of some probesets are not higher than those at $\delta = 0.30$ and over).

Regarding molecular signatures, the motivation for selecting subsets of genes whose contributions to the interclass distance are the highest (implying that they are of small

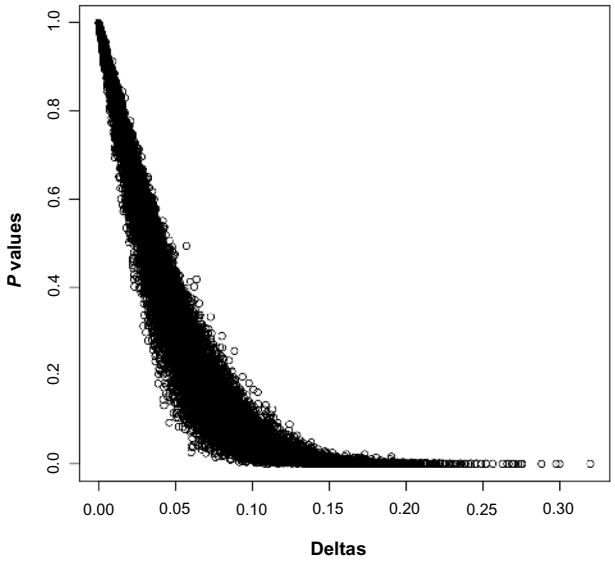


**Figure 3.** Scatterplot of the probesets.
**Notes:** X-axis: probesets' contributions ($\delta$ values) to the interclass distance (PCR and no-PCR classes). Y-axis: probesets' *P*-values to the Student *t*-test. The $\delta$ contributions to the interclass and the *P*-values to the *t*-test were computed on the 133 tumor samples of Dataset VI.

*P*-values) is now established: they are the optima of the bi-objective function. To our best knowledge no reason was ever given for selecting subsets of genes of smallest *P*-values with no regard to their contribution the interclass distance.

Concerning the breast cancer datasets and the issue of predicting the responses to preoperative chemotherapy treatments, our predictive modeling was at the level of, or outperformed the previous studies, with three times smaller signatures. The highest ranked gene of our signatures was ESR1 (estrogen receptor protein coding gene 1, cf. Table 6). Although this gene is known to be one of the most determinant marker of the response to the chemotherapy (eg,[15]), it was neither selected in[8] nor.[14] In the latter study the probesets were ranked according to their "*bi-informativeness*". For instance, the gene BTG3 in Figure 2 is bi-informative: a low expression value is a clue of a partial responder case and a high expression value is that of a responder case. In contrast, the gene ESR1 in the same figure is mono-informative: when its expression level is high the probability of the patient case to be a partial responder is high, but when the expression level is low the two classes are of equal probabilities. Because the valuation function of[14] gave high rankings to bi-informative genes, ESR1 received a low ranking that prevented it to be in the signatures. In the former study the probesets were ranked according to the *P*-value to a *t*-test and the genes were selected in this ranking. The highest ranked gene was MAPT (microtubule-associated protein tau.) Although the boxplots of ESR1 and MAPT are very similar (cf. Figure 2), the ranking of ESR1 was too low for it to enter the DLDA-30 predictor. Its low ranking, despite the similar boxplots, may be the consequence of the parametric assumption underlying the *P*-value to the *t*-test, to which the distribution of ESR1 may be less fitted than that of MAPT. Beside ESR1 three other genes of the 11 $\delta$-signature were neither in the DLDA-30 signature nor in the bi-majority-30 one. These genes are AGR2, SOX11, and TBC1D9. According to some recent studies, these three genes may make sense in a predictor of the response to chemotherapy treatments: gene AGR2 (anterior gradient-2) is a potential novel oncogene overexpressed in estrogen receptor positive tumors[21]); gene SOX11's protein (SRY-related HMG-box 11) is a tumor suppressor[22]; and the expression of the gene TBC1D9, closely adjacent to ESR1 on its chromosome, was shown to be positively correlated with that of ESR1 in breast cancer.[23] Moreover, both TBC1D9 and ESR1 were found to be among the seven most important predictors of breast cancer, for both disease mortality and recurrence.[24]

It should be noted that two of the probesets belonging to the 30 $\delta$-signature (but not to the 11 $\delta$-signature) had *P*-values to a *t*-test that were slightly higher than the 5% threshold, above which a null hypothesis is usually considered not rejected. These two probesets (221872_at and 206391_at, *P*-values 0.053 and 0.072) are of the same gene, RARRES1 (retinoic acid receptor responder 1). Its selection comes from our valuation of the probesets, the distance between the mean expressions on the two learning classes, which does not take into account the normal distribution of the values. We refer the reader to[1] for a thorough discussion of the issue of individually assessing the features of multivariate predictors. Regarding RARRES1, this gene is known to be a tumor suppressor (eg,[25–27]).

We saw in Table 4 that only one method, GLMPath, had prediction performances that were very close to ours, and signature sizes equal or smaller than ours. The GLMPath method is a generalized LASSO method (Least Absolute Shrinkage and Selection Operator.) The LASSO algorithm is based on a regression method estimating the best parameters of a Gaussian multi-variable model and simultaneously discarding the non-explicative variables. An analysis of LASSO conducted by R. Tibshirani,[28] the author of the method, showed that LASSO is a greedy algorithm, solving a linear regression model by adding at each step the variable that is the most correlated to the current residual. Hence LASSO method computes a greedy solution to the problem of molecular signature selection. This greedy solution is not the optimal one in the general case. In contrast we demonstrated that the greedy solutions of our bi-objective function were the global optima of the bi-objective function. They were not greedy approximations of the optimal signatures: they were the optimal signatures. We also demonstrated that our bi-objective function had a very small number of optima, precisely $n$ optimal non empty signatures.

Multi-objective functions whose optima can be found by a greedy optimization are very rare. The reason why this optimization problem could be solved by a greedy approach is that

the two objectives of the function are *separable*: when a gene is added to the signature, the new value of the objective function only depends on this gene, not on the genes that were already member of the signature. In its whole generality the problem of feature/gene selection is the optimization of multi-objective functions whose objectives are non-separable. An example of such a multi-objective function is the following three objective function: minimize the signature size, maximize the interclass distance, and minimize the signature's gene correlations.

Because of the third criterion this three-objective function is not *separable*. Addressing the issue of searching optimal molecular signatures in non-separable multi-objective functions might bring significant improvements in oncology prediction.

**Programs availability:** A Web application and scripts written in R language that implement the feature selection procedure can be freely accessed or downloaded from http://gardeux-vincent.eu/DeltaRanking.php.

## Terminology and Abbreviations

The performances of a two-class predictor (binary classification) can be measured according to the following criteria, where TP, TN, FP and FN are respectively the numbers of True Positive, True Negative, False Positive and False Negative of the predictor.

- Accuracy: true prediction rate. Accuracy = (TP + TN) / (TP + FP + FN + TN)
- Sensitivity (or Recall): true positive prediction rate. Sensitivity = TP / (TP + FN)
- Specificity: true negative prediction rate. Specificity = TN / (TN + FP)
- PPV (Positive Predictive Value, or Precision): probability that a case belongs to the positive class when the positive class is predicted. PPV = TP / (TP + FP)
- NPV (Negative Predictive Value): probability that a case belongs to the negative class when the negative class is predicted. NPV = TN / (FN + TN)

The following abbreviations are used for specific signatures and classification methods:

- DLDA: Diagonal Linear Discriminant Analysis, which is a classification method.
- $\delta$-DLDA-30: The signature found by our method, using a fixed number of 30 features (the 30 top-ranked genes).
- $\delta$-DLDA-11: The signature found by our method, using the automatic process selecting an optimal number of features of 11 (the 11 top-ranked genes).
- DLDA-30: The 30-genes signature found in[8] by *t*-test *P*-value ranking and using DLDA as classifier.
- Bi-majority-30: The 30-genes signature found in[14] corresponding to the highest bi-informative values and using majority voting as classifier.

## Conclusions

Defining molecular signatures as optima of a bi-objective function as a tradeoff between the interclass distance and the signature size, is relevant. Computing the signatures in the ranking of genes' contributions to the interclass distance is well founded and effective. Our benchmarking showed that the performances of the linear discriminant predictors designed on these optimal signatures were at the level of the best reported studies and that the signatures were significantly smaller than almost all of them. Applying the method on two breast cancer datasets brought very small and biologically relevant molecular signatures. They further helped designing robust and efficient predictors for the response to the chemotherapy.

## Author Contributions

Defined the method: VG, RN, PS, RC. Performed the experiments: VG. Provided and analyzed the data and results: LP, RR, FR. Wrote the paper: VG, RN, PS, RC, LP, APB, MFBW. All authors reviewed and approved of the final manuscript.

## REFERENCES


1. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–82.
2. Simon R, Lam A, Li M.C, et al. Analysis of Gene Expression Data Using BRB-Array Tools. *Cancer Inform*. 2007;3:11–17.
3. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23:2507–17.
4. Kohavi R, John G.H. Wrappers for feature subset selection. *Artificial Intelligence*. 1997;97:273–324.
5. Duval B, Hao J.K. Advances in metaheuristics for gene selection and classification of microarray data. *Briefings in bioinformatics*. 2010;11:127–41.
6. Gardeux V, Chelouah R, Siarry P. et al. in: Unidimensional Search for Solving Continuous High-Dimensional Optimization Problems, (Ed.)^(Eds.) Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference on, 2009;1096–101.
7. Gardeux V, Chelouah R, Siarry P, et al. EM323: a line search based algorithm for solving high-dimensional continuous non-linear optimization problems. *Soft Comput*. 2011;15:2275–85.
8. Hess K.R, Anderson K, Symmans W.F. et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer, Journal of clinical oncology : official journal of the *American Society of Clinical Oncology*. 2006;24:4236–44.
9. Dudoit S, Fridlyand J, Speed T.P. Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*. 2002;97:77–87.
10. 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. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102:13550–5.
11. Simon R, Radmacher M.D, Dobbin K. et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*. 2003;95:pp. 14–8.
12. Dupuy A, Simon R.M. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute*. 2007;99:147–57.
13. Ghattas B, Ishak B. Sélection de variables pour la classification binaire en grande dimension : comparaisons et application aux données de biopuces. *Journal de la société française de statistique*. 2008;149:43–66.
14. Natowicz R, Incitti R, Horta EG, et al. Prediction of the outcome of preoperative chemotherapy in breast cancer using DNA probes that provide information on both complete and incomplete responses. *BMC Bioinformatics*. 2008;9:149.
15. Kim C, Tang G, Pogue-Geile K.L. et al. Estrogen receptor (ESR1) mRNA expression and benefit from tamoxifen in the treatment and prevention of estrogen receptor- positive breast cancer. *Journal of clinical oncology*: official journal of the *American Society of Clinical Oncology*. 2011;29:4160–7.

16. Tabchy A, Valero V, Vidaurre T, et al. Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer, Clinical cancer research : an official journal of the *American Association for Cancer Research*. 2010;16:5351–61.
17. Reynolds A.P, Richards G, de la Iglesia B, et al. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *J Math Model Algor*. 2006;5:475–504.
18. Szabo A, Boucher K, Carroll W.L, et al. Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Mathematical Biosciences*. 2002;176:71–98.
19. Hess K.R, Wei C.M.A, Qi Y, et al. Lack of sufficiently strong informative features limits the potential of gene expression analysis as predictive tool for many clinical classification problems. *BMC Bioinformatics*. 2011;12:463.
20. Haury A.C, Gestraud P. Vert J.P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One*. 2011;6:e28210.
21. Hrstka R, Nenutil R, Fourtouna A, et al. The pro-metastatic protein anterior gradient-2 predicts poor prognosis in tamoxifen-treated breast cancers. *Oncogene*. 2010;29:4838–47.
22. Sernbo S, Gustavsson E, Brennan D.J, et al. The tumour suppressor SOX11 is associated with improved survival among high grade epithelial ovarian cancers and is regulated by reversible promoter methylation. *Bmc Cancer*. 2011;11:405.
23. Dunbier A.K, Anderson H, Ghazoui Z, et al. ESR1 is co-expressed with closely adjacent uncharacterised genes spanning a breast cancer susceptibility locus at 6q25.1. *PLoS Genet*. 2011;7:e1001382.
24. Andres S.A, Brock G.N, Wittliff J.L. Interrogating differences in expression of targeted gene sets to predict breast cancer outcome. *BMC Cancer*. 2013;13:326.
25. Jing C, El-Ghany M.A, Beesley C, et al. Tazarotene-induced gene 1 (TIG1) expression in prostate carcinomas and its relationship to tumorigenicity. *J Natl Cancer Inst*. 2002;94:482–90.
26. Sahab Z.J, Hall M.D, Me Sung Y, et al. Tumor suppressor RARRES1 interacts with cytoplasmic carboxypeptidase AGBL2 to regulate the alpha-tubulin tyrosination cycle. *Cancer research*. 2011;71:1219–28.
27. Wilson C.L, Sims A.H, Howell A, et al. Effects of oestrogen on gene expression in epithelium and stroma of normal human breast tissue. *Endocrine-related cancer*. 2006;13:617–28.
28. Tibshirani R, in: A simple explanation of the Lasso and Least Angle Regression., (Ed.)^(Eds.), Stanford, CA, USA, http://statweb.stanford.edu/~tibs/lasso/simple.html.
29. Alon U, Barkai N, Notterman D.A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*. 1999;96:6745–50.
30. Shipp M.A, Ross K.N, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*. 2002;8:68–74.
31. Golub T.R, Slonim D.K., Tamayo P. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science (New York, N.Y.), 1999;286:531–7.
32. Singh D, Febbo P.G, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*. 2002;1:203–9.
33. Pomeroy S.L, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*. 2002;415:436–42.
34. Ramaswamy S, Ross K.N, Lander E.S, et al. A molecular signature of metastasis in primary solid tumors. *Nature genetics*. 2003;33:49–54.
35. Weston J, Elisseeff A, Schölkopf B, et al. Use of the zero norm with linear models and kernel methods. *J Mach Learn Res*. 2003;3:1439–61.
36. Deutsch J.M. Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics*. 2003;19:45–52.
37. Rakotomamonjy A. Variable selection using SVM based criteria. *J Mach Learn Res*. 2003;3:1357–70.
38. Pochet N, De Smet F, Suykens J.A, et al. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*. 2004;20:3185–95.
39. Shah S, Kusiak A. Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine*. 2007;37:251–61.
40. Orsenigo C, Gene Selection and Cancer Microarray Data Classification Via Mixed- Integer Optimization, in: E. Marchiori, J. Moore (Eds.) Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. *Springer Berlin Heidelberg*. 2008:141–52.