



HAL
open science

The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures

Anne-Claire Haury, Pierre Gestraud, Jean-Philippe Vert

► **To cite this version:**

Anne-Claire Haury, Pierre Gestraud, Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. 2010. hal-00559580v1

HAL Id: hal-00559580

<https://hal.science/hal-00559580v1>

Preprint submitted on 25 Jan 2011 (v1), last revised 23 Jun 2011 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures

Anne-Claire Haury*

Mines ParisTech, CBIO
Institut Curie, Paris, F-75248
INSERM, U900, Paris, F-75248
anne-claire.haury@ensmp.fr

Pierre Gestraud

Mines ParisTech, CBIO
Institut Curie, Paris, F-75248
INSERM, U900, Paris, F-75248
pierre.gestraud@curie.fr

Jean-Philippe Vert

Mines ParisTech, CBIO
Institut Curie, Paris, F-75248
INSERM, U900, Paris, F-75248
jean-philippe.vert@mines-paristech.fr

January 25, 2011

Abstract

Motivation Biomarker discovery from high-dimensional data is a crucial problem with enormous applications in biology and medicine, such as, e.g. breast cancer prognosis. While it is now common belief that feature selection methods should output both accurate and stable signatures - at least at the functional level -, no study has focused on comparing algorithms in these regards.

Methods Borrowing 4 public datasets, we define systematic procedures to compare a representative panel of 8 feature selection algorithms, among which filters, wrappers and embedded methods in light of predictive performance, stability and functional interpretability of the signatures that they output. We also implemented Ensemble methods and propose to estimate the advantages of using them.

Results We observe that ensemble feature selection techniques have generally no substantial impact on accuracy or stability. Additionally we notice a possible trade-off between stability and accuracy as some methods produce predictive but unstable signatures while others behave in the opposite way. Filter feature selection by a Student's t-test seems to provide a good accuracy/stability trade-off overall.

Availability Our Matlab code is available upon demand.

1 Introduction

Biomarker discovery from high-dimensional data, such as transcriptomic or SNP profiles, is a crucial problem with enormous applications in biology and medicine, such as diagnosis, prognosis, patient stratification in clinical trials or prediction of the response to a treatment. Numerous studies have for example investigated molecular signatures, i.e., predictive models based on the expression of a small number of genes, for the classification of cancer patients into groups with different prognosis (e.g. metastasis or relapse potential). In breast cancer, for

*To whom correspondance should be addressed: 35, rue Saint Honoré, F-77300 Fontainebleau, France.

instance, several prognostic signatures have been proposed to help in the therapeutic choice (Sotiriou and Pusztai, 2009), including two prognostic signatures in clinical trials: the 70-gene *MammaPrint* signature of van 't Veer *et al.* (2002), and the 76-gene *Rotterdam* signature of Wang *et al.* (2005).

While classifiers could be based on the expression of more than a few tens of genes, several reasons have motivated the search for short lists of predictive genes. First, from a statistical and machine learning perspective, restricting the number of variables is often a way to reduce overfitting to learn in high dimension from few samples, and can thus lead to better predictions on new samples. Second, inspecting the genes selected in the signature may shed light on biological processes involved in the disease, and suggest novel targets. Third, and to a lesser extent, a small list of predictive genes allows the design of cheap dedicated prognostic chips.

The comparison of existing signatures shows, however, that they have very few genes in common, if any, raising questions about their biological significance (Ioannidis, 2005). Independently of differences in cohorts or technologies, Ein-Dor *et al.* (2005) and Michiels *et al.* (2005) showed that a major cause for the lack of overlap between signatures is that many different signatures show similar accuracies, and that the process of estimating a signature is very sensitive to the samples used in the phase of gene selection. In particular, Ein-Dor *et al.* (2006) suggested that many more samples than currently available would be required to reach a descent level of signature stability, meaning in particular that no biological insight should be expected from the analysis of current signatures. Alternatively, some authors noticed that the lack of stability in terms of genes did not prevent different signatures from sharing functional interpretation (Shen *et al.*, 2008; Reyat *et al.*, 2008; Wirapati *et al.*, 2008).

From a machine learning point of view, estimating a signature from a set of expression data is a problem of *feature selection*, an active field of research in particular in the high-dimensional setting (Guyon and Elisseeff, 2003). While the limits of some basic methods for feature selection have been highlighted in the context of molecular signatures, such as gene selection by Pearson correlation with the output (Ein-Dor *et al.*, 2006), there are surprisingly very few and only partial investigations that focus on the *influence of the feature selection method* on the performance and stability of the signature. Relevant work include Lai *et al.* (2006), who compared various feature selection methods in terms of predictive performance only, or Abeel *et al.* (2010), who suggest that ensemble feature selection improves both stability and accuracy of SVM recursive feature elimination (RFE), without comparing it with other methods. However, it remains largely unclear how "modern" feature selection methods such as the elastic net (Zou and Hastie, 2005) or stability selection (Meinshausen and Bühlmann, 2010) behave in these regards, and how they compare to more basic univariate techniques.

Here we propose for the first time to the best of our knowledge an empirical comparison of a panel feature selection techniques in terms of accuracy and stability both at the gene level and at the functional level. Using four breast cancer datasets, we observe significant differences in accuracy and stability, both at the gene level and at the functional level, showing that the method used to select genes has a strong influence on these properties. Among our findings, we observe that, surprisingly, ensemble feature selection by combining multiple signatures estimated on random subsamples has generally no clear impact on accuracy and stability; that some methods have higher stability than others at the cost of producing less accurate models, suggesting a possible trade-off between stability and accuracy; and that, globally, feature selection by a simple Student t-test provides a good accuracy/stability trade-off.

2 Methods

2.1 Feature selection methods

We compare eight common feature selection methods to estimate molecular signatures. All methods take as input a matrix of gene expression data for a set of samples from two categories (good and bad prognosis in our case), and return a set of genes of a user-defined size s . These genes can then be used to estimate a classifier to predict the class of any sample from the expression values of these genes only. Feature selection methods are usually classified in three categories (Kohavi and John, 1997; Guyon and Elisseeff, 2003): *filter methods*, which select subsets of variables as a pre-processing step, independently of the chosen predictor; *wrapper methods*, which utilize the learning machine of interest as a black box to score subsets of variable according to their predictive power; and *embedded methods*, which perform variable selection in the process of training and are usually specific to given learning machines. We have selected popular methods representing these different classes, as described below.

2.1.1 Filter methods

Univariate filter methods rank all variables in terms of relevance, as measured by a score which depends on the method. They are simple to implement and fast to run. To obtain a signature of size s , one simply takes the top s genes according to the score. We consider the following four scoring functions to rank the genes: the *Student's t-test* and *Wilcoxon sum-rank test*, which evaluate if each feature is differentially expressed between the two classes; and the *Bhattacharyya distance* and *relative entropy*, to compute a distance between the distributions of the two groups. We used the MATLAB Bioinformatics toolbox to compute these scoring functions.

2.1.2 Wrapper methods

Wrapper methods attempt to select jointly sets of variables with good predictive power for a predictor. Since testing all combinations or variables is computationally impossible, wrapper methods usually perform a greedy search in the space of sets of features. We test *SVM recursive feature elimination (RFE)* (Guyon *et al.*, 2002), which starts with all variables and iteratively remove the variables which contribute the less to a linear SVM classifier trained on the current set of variables. We remove 20% of features at each iteration until s remain, and then remove them one by one until the desired number of variables is reached. Following (Abeel *et al.*, 2010), we set the SVM parameter C to 1, and checked afterwards that other values of C did not have a significant influence on the results. Alternatively, we test a *Greedy Forward Selection (GFS)* strategy combined with a nearest centroid classifier, where we start from no variable and add them one by one by selecting each time the one which maximizes the AUC of the nearest centroids classifier, in a 3-fold internal cross-validation setting.

2.1.3 Embedded methods

Embedded methods are learning algorithm which perform feature selection in the process of training. We test the popular *Lasso* regression (Tibshirani, 1996), where a sparse linear predictor $\beta \in \mathbb{R}^p$ is estimated by minimizing the objective function $R(\beta) + \lambda \|\beta\|_1$, where $R(\beta)$ is the mean square error on the training set (considering the two categories as ± 1 values) and $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$. λ controls the degree of sparsity of the solution, i.e., the number of features selected. We fix λ as the smallest value which gives a signature of the desired size s . Alternatively, we tested the elastic net (Zou and Hastie, 2005), which is similar to the Lasso but where we replace the ℓ_1 norm of β by a combination of the ℓ_1 and ℓ_2 norms, i.e., we minimize $R(\beta) + \lambda \|\beta\|_1 +$

$\lambda/2\|\beta\|_2^2$ and $\|\beta\|_2^2 = \sum_{i=1}^p \beta_i^2$. By allowing the selection of correlated predictive variables, the elastic net is supposed to be more robust than the Lasso while still selecting predictive variables. Again, we tune λ to achieve a user-defined level of sparsity.

2.2 Ensemble feature selection

Many feature selection methods are known to be sensitive to small perturbations of the training data, resulting in unstable signatures. In order to "stabilize" variable selection, several authors have proposed to use ensemble feature selection on bootstrap samples: the variable selection method is run on several random subsamples of the training data, and the different lists of variables selected are merged into a hopefully more stable subset (Bi *et al.*, 2003; Bach, 2009; Meinshausen and Bühlmann, 2010; Abeel *et al.*, 2010).

For each feature selection method described above, we tested the following three aggregation strategies for ensemble feature selection. We first bootstrap, i.e. draw a sample of size n from the data with replacement, the training samples $B = 50$ times and thus get B rankings ($r^1 \dots r^B$) of all features. For filter methods, the ranking of features is naturally obtained by decreasing score. For RFE and GFS, the ranking is the order in which the features are added or removed in the iterative process. For Lasso and elastic net, the ranking is the order in which the variables become selected when C increases. We then aggregate the B lists by computing a score $S_j = 1/B \sum_{b=1}^B f(r_j^b)$ for each gene j as an average function of its rank r_j^b in the b -th bootstrap experiment. We tested the following functions of the rank for aggregation:

- *Ensemble-mean*: we just average the rank of a gene over the bootstrap experiments, i.e., we take $f(r) = r$ (Abeel *et al.*, 2010).
- *Ensemble-stability selection*: we measure the percentage of bootstrap samples where the gene is in the top s genes, i.e., we take $f(r) = 1$ if $r \leq s$, 0 otherwise. This is similar to the stability selection strategy investigated by Meinshausen and Bühlmann (2010).
- *Ensemble-exponential*: we propose a soft version of stability selection, where we average an exponentially decreasing function of the rank, namely $f(r) = \exp -r/s$.

Finally, for each rank aggregation strategy, the aggregated list is the set of s genes with the largest score.

2.3 Accuracy of a signature

In order to assess the predictive accuracy of a feature selection method, we assess the performance of various supervised classification algorithms trained on data restricted to the selected signature. More precisely, we tested 5 classification algorithms: nearest centroids (NC), k-nearest neighbors (KNN) with $k = 9$, linear SVM with $C = 1$, linear discriminant analysis (LDA) and naive Bayes (BAYES). The parameters of the KNN and SVM methods were fixed to arbitrary default values, and we checked that no significantly better results could be obtained with other parameters by testing a few other parameters. We assess the performance of a classifier by the area under the ROC curve (AUC), in two different settings. First, on each dataset, we perform a 10-fold cross-validation (CV) experiment, where on each iteration feature selection is performed and the classifier is trained on 90% of the data, and the AUC is computed on the remaining 10% of the data. This is a classical way to assess the relevance of feature selection of a given dataset. Second, to assess the performance of the signature across dataset, we estimate a signature on one dataset, and assess its accuracy on other datasets by again running a 10-fold CV experiment where only the classifier (restricted to the genes in the signature) is retrained on each training set.

2.4 Stability of a signature

To assess the stability of feature selection methods, we compare the signatures estimated on different samples in various settings. First, to evaluate stability with respect to small perturbation of the training set, we randomly subsample each dataset into pairs of subset with 80% of sample overlap, estimate a signature on each subset, and compute the overlap between two signatures in a pair as the fraction of shared genes, i.e., $|S_1 \cap S_2|/s$. Note that this corresponds to the figure of merit defined by Ein-Dor *et al.* (2006). The random sampling of subsets is repeated 20 times on each dataset, and the stability values are averaged over all samples. We call this procedure the *soft-perturbation* setting below. Second, to assess stability with respect to strong perturbation within a dataset, we repeat the same procedure but this time with no overlap between two subsamples whose signatures are compared. In practice, we can only sample subsets of size $N/2$, where N is the number of samples in a dataset, to ensure that they have no overlap. Again, we measure the overlap between the signatures estimated on training set with no sample in common. We call this procedure the *hard-perturbation* setting below. Finally, to assess the stability across datasets, we estimate signatures on each dataset independently, using all samples on each dataset, and measure their overlap. We call this procedure the *between-datasets setting* below.

2.5 Functional analysis and stability of a signature

To interpret a signature in terms of biological function, we perform functional enrichment analysis by inspecting the signature for over-represented Gene Ontology (GO) terms. This may hint at biological hypothesis underlying the classification (Shen *et al.*, 2008; Reyal *et al.*, 2008). We performed a hypergeometrical test on each of the 5830 GO biological process (BP) terms that were associated to at least one gene in our dataset, and corrected the resulting p-values for multiple testing through the procedure of Benjamini and Hochberg (1995). To assess the *interpretability* of a signature, i.e. how easily one can extract a biological interpretation, we computed the number of GO terms over-represented at 5% FDR. To compare two signatures in functional terms, we first extracted from each signature the list of 10 GO terms with the smallest p-values, and compared the two lists of GO terms by the similarity measure of Wang *et al.* (2007) which takes into account not only the overlap between the lists but also the relationships between GO BP. Finally, to assess the functional stability of a selection method, we followed a procedure similar to the one presented in Section 2.4 and measured the mean functional similarity of signatures in the soft-perturbation, hard-perturbation and between-datasets settings.

3 Data

We collected 4 breast cancer datasets with relapse information from Gene Expression Omnibus (Barrett *et al.*, 2009), as described in Table 1. The four datasets were obtained through the same technology, namely Affymetrix HG-U133A, and all normalized using Robust Multi-Array (RMA) (Irizarry *et al.*, 2003). Furthermore, we used a custom CDF file with EntrezGene ids as identifiers (see Dai *et al.*, 2005), resulting in expression levels for 12,065 genes. All datasets address the problem of predicting metastatic relapse in breast cancer, although on different cohorts.

Dataset name	# examples	# positives	source
GSE1456	159	40	Pawitan <i>et al.</i> (2005)
GSE2034	286	107	Wang <i>et al.</i> (2005)
GSE2990	125	49	Sotiriou <i>et al.</i> (2006)
GSE4922	249	89	Ivshina <i>et al.</i> (2006)

Table 1: The four breast cancer datasets used in this study.

4 Results

4.1 Accuracy

We first assess the accuracy of signatures obtained by different feature selection method. Intuitively, the accuracy refers to the performance that a classifier trained on the genes in the signature can reach in prediction. Although some feature selection methods (wrapper and embedded) jointly estimate a predictor, we dissociate here the process of selecting a set of genes and training a predictor on these genes, in order to perform a fair comparison common to all feature selection methods. We tested the accuracy of signatures of 100 genes obtained by each feature selection method, combined with five classification algorithms to make a predictor, as explained in Section 2.3. Table 2 shows the accuracies, measured in AUC, reached by the different combinations. Each value is the average over the different datasets of the 10-fold cross-validation AUC. Note that on each training set (corresponding to 90% of a given dataset), a signature of 100 genes is learned and a predictor based on these genes is trained, without using the 10% of left-out samples in order to prevent any selection bias ([Ambroise and McLachlan, 2002](#)).

Globally, we observe limited differences between the feature selection methods, for a given classification method. In particular the selection of a random signature reaches a baseline AUC comparable to that of other methods, confirming results already observed by [Ein-Dor *et al.* \(2005\)](#). Second, we observe that, among all classification algorithms, the simple nearest centroid classifier consistently gives good results compared to other classifiers. We therefore choose it as a default classification algorithm for further assessment of the performance of the signatures below. Figure 1 depicts graphically the AUC reached by each feature selection method with nearest centroids as a classifier, reproducing the first three lines of Table 2. In the single-run framework, the t-test performs significantly better than most methods ($p < 0.05$ against random, entropy, Bhattacharyya, Wilcoxon and GFS and $p < 0.1$ against RFE). Lasso and Elastic Net perform similarly and show an AUC significantly higher than entropy and GFS ($p < 0.01$). Lasso also significantly overperforms random, Bhattacharyya and RFE at 10%. Except for the t-test ($p = 0.0065$), random feature selection is not significantly worse than any other algorithm at 5%. Somehow surprisingly, some method perform even worse than random: the relative entropy ($p = 0.052$) and GFS ($p = 0.015$). Finally, we observe that ensemble methods for feature selection do not bring any improvement in accuracy in general since only Bhattacharyya and GFS benefit from Ensemble-mean (resp. $p = 0.004$ and $p = 0.02$) and no significant improvement is obtained from the use of the two remaining Ensemble aggregation methods.

In order to check how these results depend on the size of the signature, we plot in Figure 2 the AUC of the 9 feature selection methods, with or without ensemble averaging, combined with a nearest centroid classifier, as a function of the size of the signature. Interestingly, we observe that in some cases the AUC seems to increase early, implying that a few genes may be sufficient to obtain the maximal performance. For instance, a 10-gene signature obtained

Class.	Type	Random	t-test	Entropy	Bhatt.	Wilcoxon	SVM RFE	GFS	Lasso	Elastic Net
NC	S	0.64(0.14)	0.66(0.15)	0.60(0.14)	0.62(0.14)	0.62(0.17)	0.63(0.16)	0.53(0.18)	0.64(0.15)	0.65(0.15)
	E-M	0.62(0.16)	0.66(0.15)	0.61(0.15)	0.63(0.15)	0.63(0.18)	0.62(0.16)	0.60(0.17)	0.63(0.17)	0.63(0.16)
	E-E	0.62(0.15)	0.65(0.15)	0.59(0.15)	0.61(0.14)	0.63(0.17)	0.63(0.16)	0.54(0.15)	0.65(0.15)	0.65(0.16)
	E-S	0.62(0.15)	0.65(0.15)	0.59(0.15)	0.59(0.14)	0.63(0.17)	0.63(0.16)	0.53(0.16)	0.64(0.15)	0.64(0.15)
SVM	S	0.56(0.12)	0.57(0.14)	0.56(0.13)	0.54(0.16)	0.54(0.14)	0.62(0.19)	0.53(0.15)	0.64(0.16)	0.62(0.18)
	E-M	0.50(0.16)	0.57(0.12)	0.59(0.14)	0.58(0.12)	0.57(0.12)	0.63(0.16)	0.51(0.16)	0.62(0.18)	0.62(0.16)
	E-E	0.53(0.15)	0.55(0.13)	0.55(0.15)	0.51(0.15)	0.56(0.13)	0.65(0.16)	0.54(0.13)	0.63(0.16)	0.63(0.17)
	E-S	0.62(0.14)	0.65(0.14)	0.59(0.14)	0.59(0.15)	0.63(0.15)	0.63(0.14)	0.53(0.15)	0.64(0.18)	0.64(0.18)
KNN	S	0.57(0.13)	0.63(0.14)	0.58(0.13)	0.59(0.13)	0.59(0.17)	0.63(0.14)	0.52(0.16)	0.62(0.15)	0.61(0.15)
	E-M	0.60(0.13)	0.62(0.15)	0.58(0.13)	0.60(0.15)	0.61(0.17)	0.63(0.15)	0.53(0.18)	0.60(0.16)	0.62(0.15)
	E-E	0.58(0.18)	0.62(0.15)	0.57(0.15)	0.58(0.14)	0.62(0.19)	0.60(0.16)	0.54(0.17)	0.62(0.14)	0.61(0.17)
	E-S	0.54(0.17)	0.60(0.15)	0.57(0.15)	0.58(0.13)	0.62(0.17)	0.62(0.15)	0.58(0.13)	0.61(0.16)	0.60(0.16)
LDA	S	0.54(0.12)	0.52(0.10)	0.55(0.16)	0.53(0.14)	0.52(0.14)	0.56(0.15)	0.53(0.13)	0.58(0.12)	0.59(0.14)
	E-M	0.50(0.11)	0.55(0.12)	0.53(0.11)	0.51(0.12)	0.53(0.11)	0.57(0.12)	0.53(0.12)	0.55(0.14)	0.57(0.12)
	E-E	0.55(0.11)	0.51(0.12)	0.54(0.12)	0.50(0.14)	0.55(0.12)	0.60(0.15)	0.54(0.12)	0.59(0.12)	0.58(0.14)
	E-S	0.53(0.14)	0.53(0.12)	0.54(0.13)	0.49(0.13)	0.53(0.10)	0.59(0.13)	0.50(0.12)	0.56(0.13)	0.58(0.15)
BAYES	S	0.60(0.14)	0.62(0.13)	0.57(0.13)	0.57(0.12)	0.58(0.14)	0.57(0.13)	0.52(0.12)	0.59(0.13)	0.58(0.14)
	E-M	0.58(0.13)	0.61(0.15)	0.58(0.13)	0.60(0.14)	0.59(0.16)	0.59(0.13)	0.56(0.15)	0.58(0.14)	0.60(0.13)
	E-E	0.59(0.14)	0.61(0.14)	0.55(0.13)	0.57(0.13)	0.59(0.15)	0.58(0.14)	0.52(0.14)	0.60(0.12)	0.59(0.14)
	E-S	0.58(0.12)	0.61(0.14)	0.55(0.13)	0.57(0.11)	0.59(0.14)	0.58(0.14)	0.54(0.14)	0.58(0.12)	0.59(0.12)

Table 2: AUC obtained for each combination of feature selection and classification method, in 10-fold cross validation and averaged over the datasets. Standard error is shown within parentheses. For each selection algorithm, we highlighted the setting in which it obtained the best performance. The *Type* column refers to the use of feature selection run a single time (S) or through ensemble feature selection, either with the mean (E-M), exponential (E-E) or stability selection (E-S) procedure to aggregate lists.

with t-test selection achieves the same average performance as a random 100-gene signature. However, it is worth noting that some algorithms have an increasing AUC curve in this range of sizes, and we observe no overfitting that may lead to decreasing AUC when the number of features increases. Random selection was previously shown to give an AUC equivalent to other methods for a large signature, but as we observe on this picture, the less genes the larger the gap in AUC.

Finally, we aim to estimate the predictive performance of a signature across datasets. For that purpose we use each dataset in turn to learn a list of 100 genes; restraining the three other datasets to these genes, we estimate the AUC of a nearest centroid classifier by 10-fold cross-validation on each dataset. We report the results in Table 3 as averages over the three test datasets. For each training dataset, we highlighted the method with the best results. Single-run t-test performs significantly better than every other filter and wrapper method at 5%. Elastic Net and Lasso perform significantly better than every method except for the t-test at 10%. In this between-datasets setting, we also note that both entropy and Bhattacharyya benefit from Ensemble-mean in terms of AUC ($p < 0.05$). However, no method performs better when used with Ensemble-exponential or Ensemble-stability selection.

4.2 Stability of gene lists

We now assess the relative stability of signatures created by different feature selection methods in terms of genes they contain. While the lack of stability of particular methods due to the small number of samples used to estimate the signatures has already been noticed and investigated by several authors (Ein-Dor *et al.*, 2005; Michiels *et al.*, 2005; Ein-Dor *et al.*, 2006), we wish here to investigate the influence of the feature selection method on signature stability.

The lack of stability observed between different signatures can be attributed to different factors, including (i) differences in cohorts that may differ in potentially relevant factors, (ii) differences in microarray technologies, (iii) differences in experimental protocols and (iv) random instability due to small sample size. Ein-Dor *et al.* (2006) has highlighted the importance of the last factor, the small size effect, by testing the stability of signatures estimated

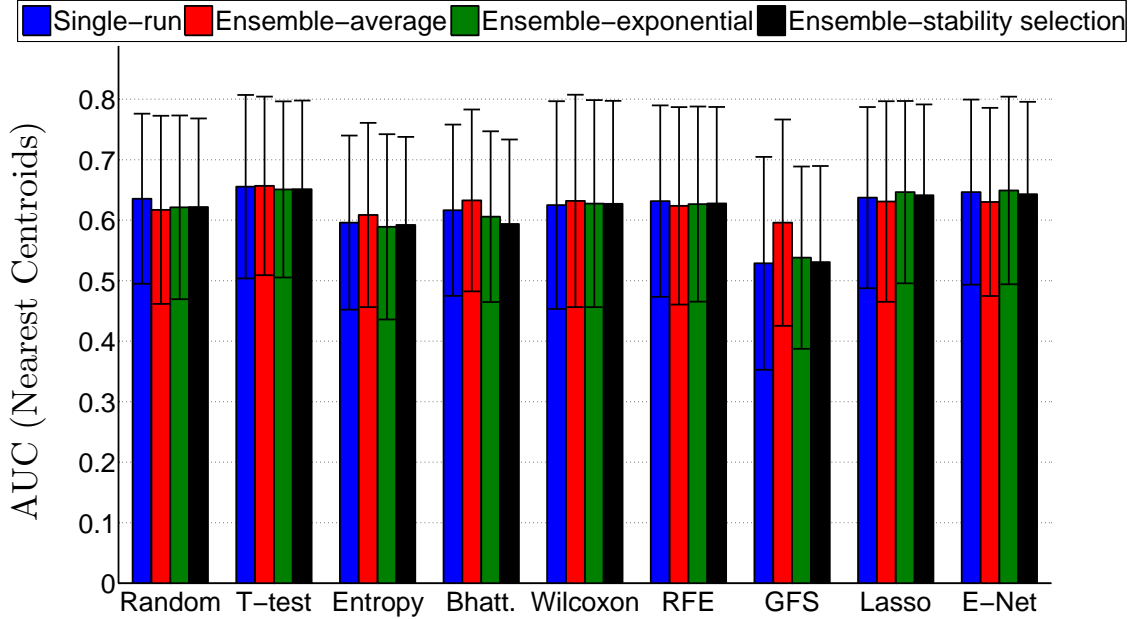


Figure 1: Area under the ROC curve for a signature of size 100 in a 10-fold CV setting and averaged over the four datasets

Training data	Type	Random	t-test	Entropy	Bhattacharyya	Wilcoxon	SVM RFE	GFS	Lasso	Elastic Net
GSE1456	S	0.61(0.17)	0.62(0.11)	0.59(0.16)	0.63(0.13)	0.60(0.12)	0.62(0.16)	0.58(0.17)	0.62(0.14)	0.62(0.13)
	E-M	0.60(0.16)	0.62(0.12)	0.61(0.12)	0.62(0.11)	0.60(0.12)	0.62(0.15)	0.58(0.15)	0.62(0.15)	0.64(0.15)
	E-E	0.59(0.14)	0.62(0.11)	0.60(0.16)	0.62(0.14)	0.61(0.12)	0.62(0.15)	0.58(0.17)	0.64(0.14)	0.64(0.14)
	E-S	0.60(0.15)	0.62(0.11)	0.58(0.16)	0.62(0.14)	0.61(0.12)	0.62(0.14)	0.56(0.16)	0.61(0.14)	0.65(0.14)
GSE2034	S	0.64(0.17)	0.64(0.16)	0.58(0.18)	0.62(0.16)	0.58(0.16)	0.60(0.19)	0.61(0.16)	0.64(0.14)	0.64(0.15)
	E-M	0.61(0.16)	0.64(0.16)	0.63(0.16)	0.64(0.15)	0.60(0.16)	0.63(0.16)	0.63(0.16)	0.64(0.15)	0.64(0.16)
	E-E	0.63(0.17)	0.64(0.16)	0.56(0.17)	0.62(0.17)	0.61(0.17)	0.62(0.17)	0.61(0.16)	0.65(0.15)	0.64(0.14)
	E-S	0.63(0.18)	0.64(0.17)	0.56(0.18)	0.60(0.17)	0.62(0.17)	0.63(0.16)	0.62(0.15)	0.66(0.15)	0.65(0.14)
GSE2990	S	0.65(0.14)	0.66(0.15)	0.55(0.14)	0.60(0.14)	0.61(0.16)	0.62(0.11)	0.62(0.14)	0.65(0.14)	0.65(0.14)
	E-M	0.62(0.13)	0.66(0.16)	0.61(0.14)	0.65(0.15)	0.58(0.15)	0.64(0.12)	0.63(0.16)	0.64(0.11)	0.64(0.14)
	E-E	0.62(0.14)	0.68(0.15)	0.53(0.13)	0.57(0.13)	0.60(0.17)	0.64(0.13)	0.57(0.15)	0.66(0.14)	0.65(0.15)
	E-S	0.61(0.12)	0.68(0.15)	0.52(0.14)	0.57(0.12)	0.60(0.16)	0.64(0.13)	0.59(0.17)	0.66(0.13)	0.66(0.14)
GSE4922	S	0.64(0.16)	0.67(0.13)	0.57(0.17)	0.63(0.14)	0.68(0.17)	0.64(0.16)	0.61(0.16)	0.66(0.15)	0.68(0.15)
	E-M	0.67(0.14)	0.67(0.13)	0.64(0.12)	0.67(0.13)	0.67(0.17)	0.65(0.15)	0.66(0.15)	0.65(0.16)	0.66(0.16)
	E-E	0.66(0.16)	0.68(0.14)	0.58(0.16)	0.61(0.15)	0.68(0.16)	0.64(0.15)	0.59(0.15)	0.66(0.14)	0.67(0.15)
	E-S	0.65(0.18)	0.69(0.14)	0.57(0.16)	0.59(0.15)	0.68(0.16)	0.64(0.15)	0.61(0.16)	0.67(0.14)	0.66(0.15)

Table 3: AUC obtained with Nearest Centroids when a signature is learnt from one dataset and tested by 10-fold cross-validation on the three remaining datasets. Standard error is shown within parentheses. For each training dataset, we highlighted the best performance. The *Type* column refers to the use of feature selection run a single time (S) or through ensemble feature selection, either with the mean (E-M), exponential (E-E) or stability selection (E-S) procedure to aggregate lists.

on non-overlapping bootstrap samples of a given dataset were all other factors are constant. Comparing the stability of signature in this hard-perturbation setting with the stability in the between-datasets setting (see definitions in Section 2.4) offers the opportunity to investigate the instability due to the first and third factor: how less stable are signatures estimated on data from two independent cohorts, than signatures estimated on data from the same cohort? Figure 3 illustrates this difference for one feature selection method. It shows the stability of the t-test in both settings with respect to the number of samples used to estimate signatures. While both curves remain low, we observe like [Ein-Dor et al. \(2006\)](#) a very strong effect of the number of samples. Interestingly, we observe that for very small sample sizes the stability in

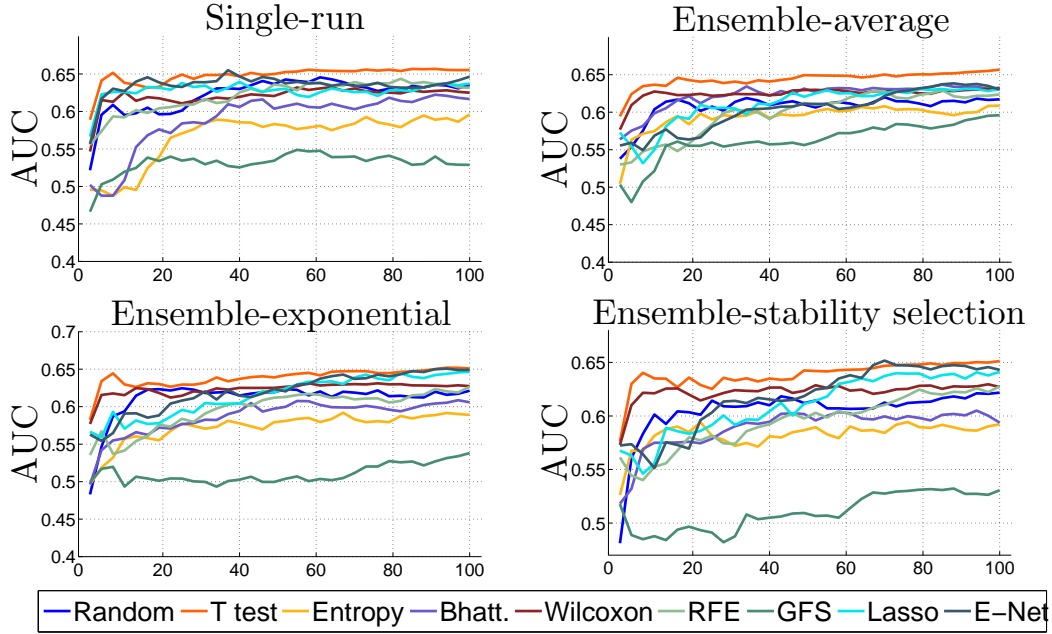


Figure 2: AUC of a nearest centroid classifier trained as a function of the size of the signature, for different feature selection methods, in a 10-fold CV setting averaged over the four datasets

the hard-perturbation setting is a good proxy for the stability in the between-dataset setting. However, the slope of the hard-perturbation setting stability seems sharper, suggesting that the gap would stretch for larger sample sizes, should the blue curve be extrapolated. These results suggest that the main reason for signature instability for a given microarray technology is really the sample size issue.

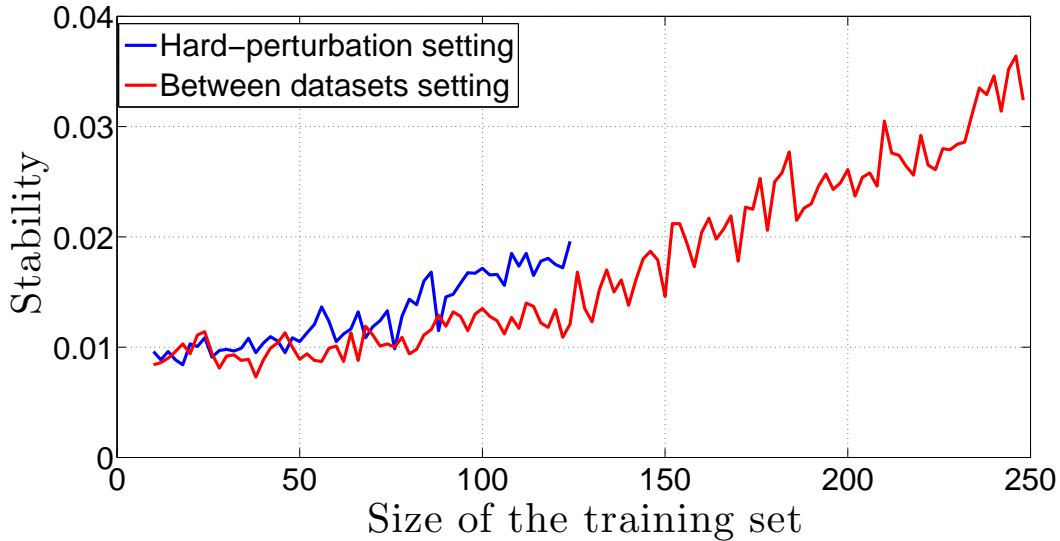


Figure 3: Evolution of stability of t-test signatures with respect to the size of the training set in the hard-perturbation and the between datasets settings from GSE2034 and GSE4922.

Figure 4 compares the stability of 100-gene signatures estimated by all feature selection

methods tested in this benchmark, in the three experimental settings: soft-perturbation (80% of samples in common), hard-perturbation (no overlap) and between-datasets settings. The results are averaged over the bootstrap replicates and the four datasets. Significance analysis in the hard-perturbation setting reveals that every single-run algorithm outputs more stable signatures than random selection ($p < 10^{-3}$). Moreover, filter methods return significantly more stable signatures than wrappers and embedded methods. This is especially true for entropy and Bhattacharyya ($p < 10^{-20}$). Except for GFS, the Ensemble-mean setting does not significantly improve the stability of the signatures. However, Ensemble-exponential and Ensemble-stability selection do for entropy ($p < 10^{-25}$), Bhattacharyya ($p < 10^{-23}$) and GFS ($p = 10^{-20}$). We also observe a small gain in the Ensemble-exponential setting for the Lasso and RFE ($p < 0.1$).

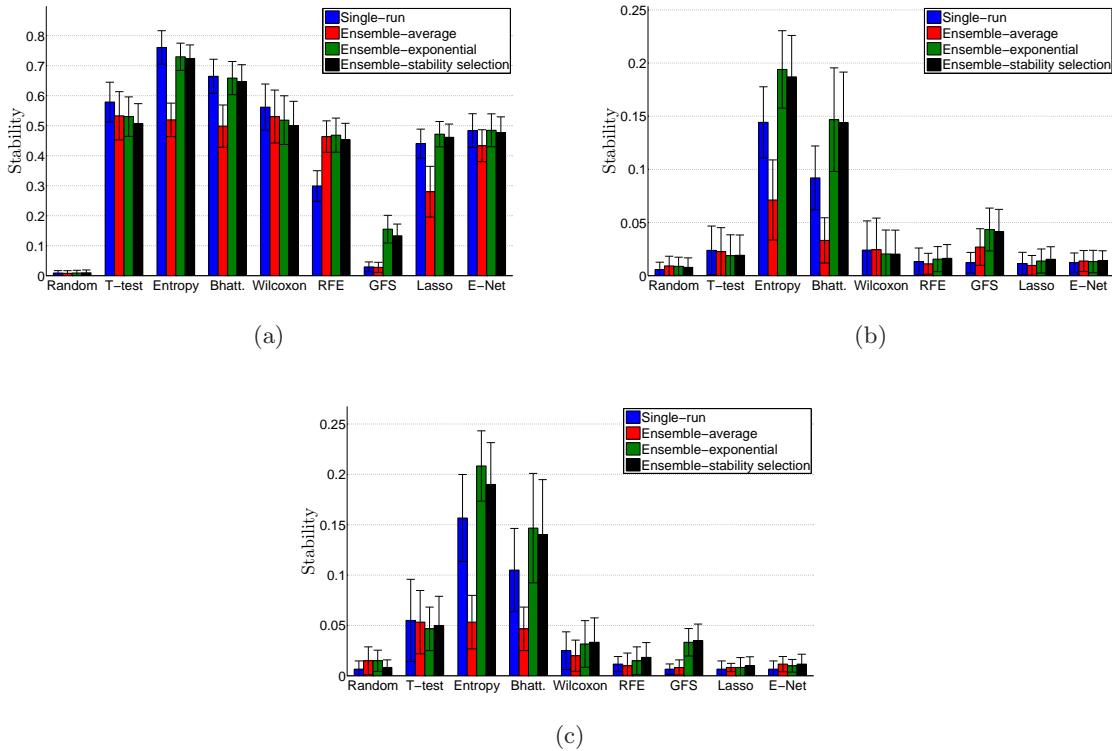


Figure 4: Stability for a signature of size 100. Average and standard errors are obtained over the four datasets. a) Soft-perturbation setting. b) Hard-perturbation setting. c) Between-datasets setting.

Obviously, Figures 4b and 4c are very much alike while Figure 4a stands aside. They confirm that the hard-perturbation setting is the best way to estimate the behavior of the algorithms between different studies. The larger stability observed in the between-datasets setting compared to the hard perturbation setting for some methods (e.g., t-test) are essentially due to the fact that signatures are trained on more samples in the between-dataset setting, since no split is required within a dataset.

It appears very clearly and significantly that univariate filter methods provide more stable lists than wrappers and embedded methods. It also seems that Ensemble-exponential and Ensemble-stability selection yield much more stable signatures than Ensemble-average. It is also worth noting that a significant gain in robustness through bootstrap is only observable for entropy and Bhattacharyya distance. Interestingly, SVM-RFE seems to benefit from Ensemble aggregation in the soft-perturbation setting, as observed by [Abeel et al. \(2010\)](#), but the

beneficial effect seems to vanish in the more relevant hard-perturbation and between-dataset settings.

Figure 5 shows how stability of the different methods varies with the size of the signature, in the between-dataset setting. We observe that the relative stability of the different methods does not depend on the size of the signature over a wide range of values, confirming that the differences observed for signatures of size 100 reveal robust differences between the methods.

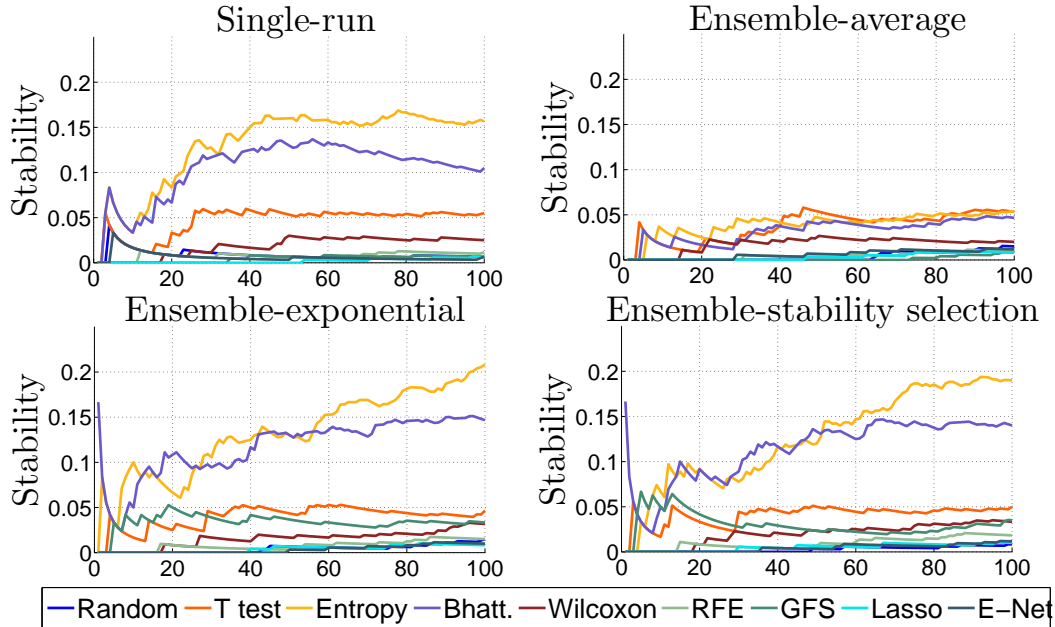


Figure 5: Stability of different methods in the between-dataset setting, as a function of the size of the signature.

4.3 Functional stability

Even when different lists of genes share no or little overlap in terms of genes, it is possible that they encode the same biological process and be useful if we can extract information about these processes from the signatures in a robust manner. In the case of breast cancer prognostic signatures, for example, several recent studies have shown that functional analysis of the signatures can highlight coherent biological processes (Fan *et al.*, 2006; Reyat *et al.*, 2008; Shen *et al.*, 2008; Abraham *et al.*, 2010; Shi *et al.*, 2010). Just like gene stability, it is therefore important to assess the stability of biological interpretation that one can extract from signatures.

First, we evaluate the *interpretability* of signatures of size 100, i.e., the ability of functional analysis to extract a biological interpretation from a signature. Figure 6 shows the average number of Gene Ontology biological processes (GO BP) significantly over-represented at 5% FDR in the signatures returned by each algorithm. Strikingly, the four filter methods appear to be much more interpretable than wrappers/embedded methods. However, it should be pointed out that, regardless of the algorithm, the number of significant terms is often zero, leading to large error bars. Ensemble methods do not seem to change the interpretability of signatures.

Second, we assess the *stability* of interpretation. For that purpose we represent each signature by the 10 most significant GO term they suggest, and assess the functional similarity between two signatures by comparing these GO terms as explained in Section 2.5. Figure 7 compares the algorithms in terms of the functional stability of the signatures. The results are

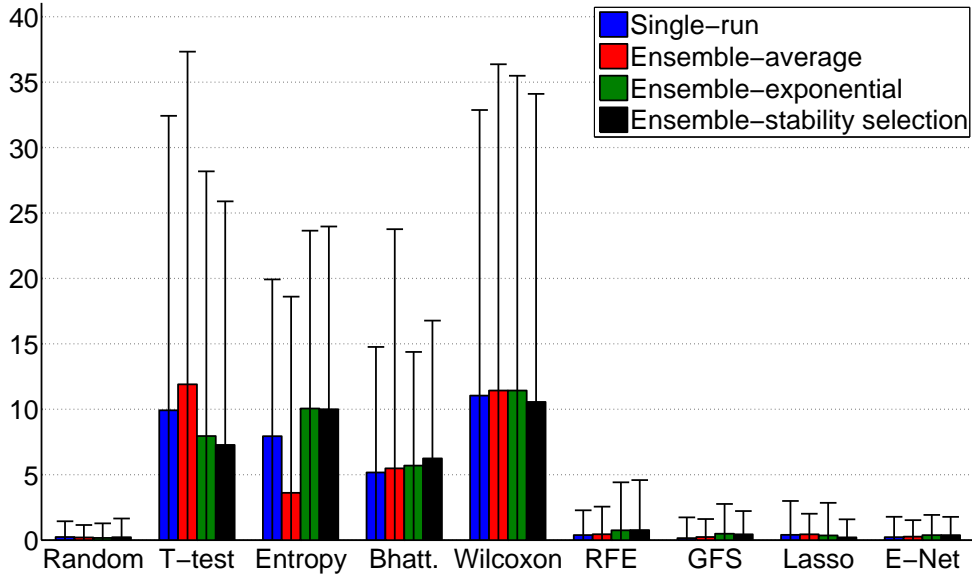


Figure 6: Interpretability of the signatures : average number of GO BP significantly over-represented. Mean and standard errors are obtained over 20 replications on each dataset.

overall very similar to the results at the gene level, namely, we observe that univariate filters are overall the most stable methods, and that the hard-perturbation setting returns a trustworthy estimate of the inter-datasets stability. In particular, we note that in the hard-perturbation and single-run settings, only signatures obtained from filters are significantly more stable than random at ($p < 0.05$). We also note that Ensemble-mean never improves the functional stability and that Ensemble-exponential/Ensemble stability selection return more stable signatures than single-run for Entropy and Bhattacharyya ($p < 10^{-22}$) as well as for GFS ($p < 10^{-6}$) and Lasso ($p = 0.029$) although less significantly.

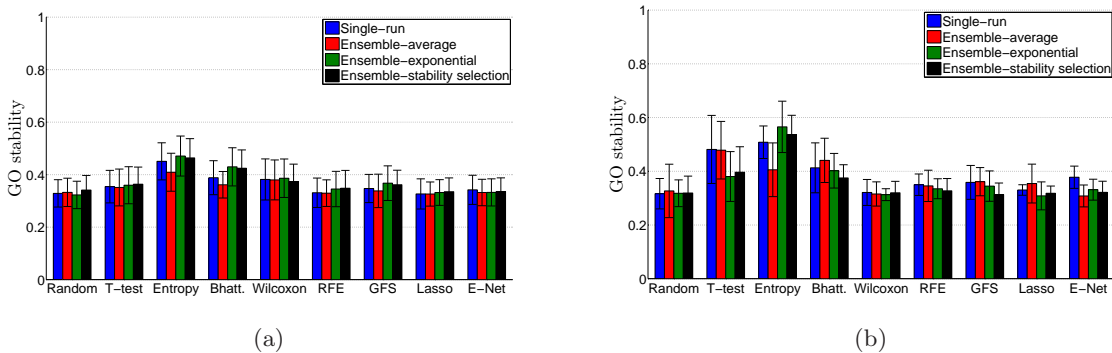


Figure 7: GO Stability for a signature of size 100. Average and standard errors are obtained over the four datasets. a) Hard-perturbation setting. c) Between-datasets setting.

4.4 The trade-off between stability and accuracy

Combining the results of the previous sections, Figure 8 shows both AUC and stability (at the gene level) for each method, in the between-dataset setting. Surprisingly, we observe a negative correlation between stability and AUC, in particular for filter methods: Entropy is the most stable but less accurate method, followed by Bhattacharya, while t-test is less stable but very accurate. Being more stable but less accurate than random, which is the case of Entropy and Bhattacharyya, suggests a bias in the method which may preferably and consistently select particular genes, not necessarily very predictive. To elucidate this behavior, we investigated the genes selected by these two methods, and noticed that they tend to be systematically expressed at low levels, as shown in Figure 9. It is therefore likely that, although stable and interpretable, the molecular signatures generated by these methods may lead to erroneous interpretation. Overall, the t-test seems to give the best stability among methods with good accuracy.

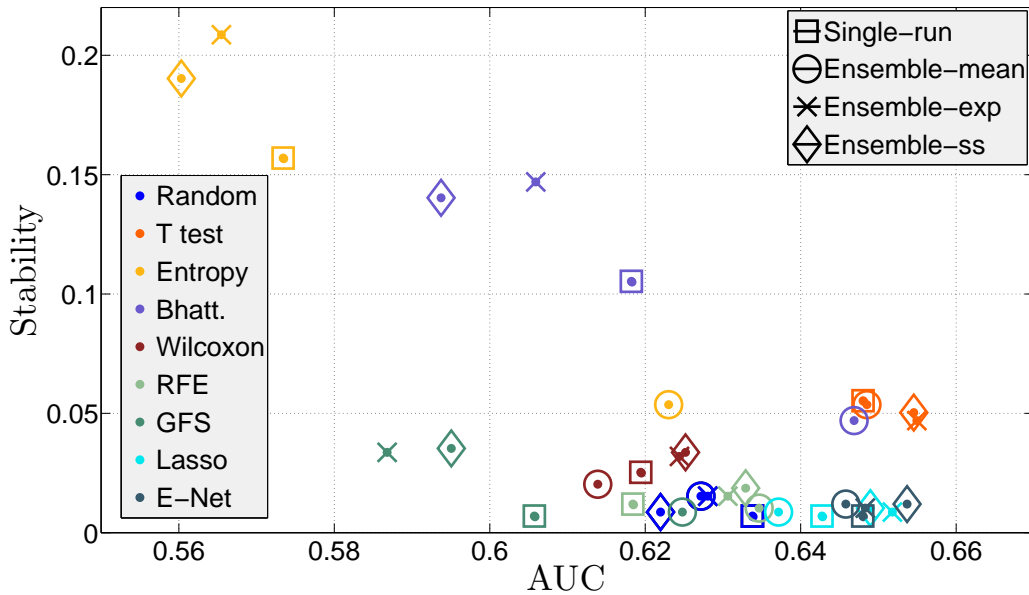


Figure 8: Accuracy versus stability for each method in the between-datasets setting. We show here the average results over the four datasets.

5 Discussion

We compared a panel of 9 feature selection methods in light of criteria that are relevant to bioinformatics analysis: we assessed each method’s accuracy and stability, both at the gene level and at the functional level. The results we observe deserve several comments.

First, we noticed the overall good performance of the t-test, that gave both the best performance among all methods, and the best stability among accurate methods. More generally, all univariate filter methods are more stable at the gene and functional level than multivariate wrapper and embedded methods. This is in fact not surprising because multivariate methods typically try to avoid redundant genes, which certainly impact their stability in data where typically many genes encode for functionally related proteins. While the elastic net was designed exactly to fight this detrimental property of Lasso by allowing the selection of groups of correlated genes (Tibshirani, 1996), we did not observe any improvement in stability between

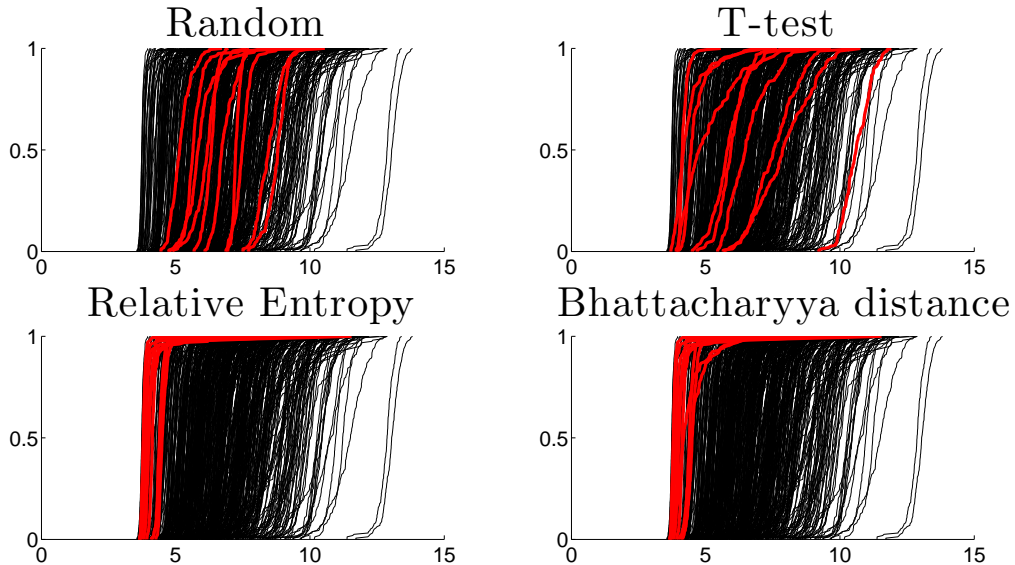


Figure 9: Estimated cumulative distribution functions (ECDF) of the first ten genes selected by four methods on GSE1456. They are compared to the ECDF of 500 background genes.

Lasso and elastic net in our experiments, although both methods had a good accuracy. The behavior of wrapper methods was overall disappointing. SVM RFE and Greedy Forward Selection are neither more accurate, nor more stable or interpretable than other methods, while their computational cost is much higher. On the other hand, it is easy to see that lists obtained from univariate methods are often composed of genes that are very much alike. Indeed, if two genes are very correlated, which is often the case in this kind of dataset, their involvements with the response will probably be similar. This partly explains why univariate methods are overall more stable in terms of genes. They also produce the most stable and interpretable lists in terms of pathways, for two main reasons. First, stability in genes obviously encourages stability in pathways. Second, pathways like GO BP contain similar genes by construction, leaving no chance of interpretability to a heterogeneous signature, as obtained by multivariate methods.

Second, we observe that ensemble method which select features by aggregating signatures estimated on different bootstrap samples increased the stability of some methods in some cases, but did not bring huge advantages to the best methods. Regarding the aggregation step itself, we advise against the use of Ensemble-average, i.e. averaging the ranks of each gene over the bootstrapped lists, regardless of the selection method. Ensemble-stability selection or Ensemble-exponential gave consistently better results. The superiority of the latter two can be explained by the high instability of the rankings. It is indeed very likely that the same gene will be ranked very differently in two lists, implying that the average rank will not be informative as to the relevancy of a gene compared to another. Therefore, considering a - strictly or not - decreasing function of the ranks is a better choice. The instability of the ranks is discussed in e.g. [Iwamoto and Pusztai, 2010](#).

Regarding the choice of method to train a classifier once feature are selected, we observed that the best accuracy was achieved by the simplest one, namely the *nearest centroids* classifier, used e.g. by [Lai et al. \(2006\)](#); [Abraham et al. \(2010\)](#). An advantage of this classifier is that it does not require any parameter tuning, making the computations fast and less prone to overfitting.

It appears from our experiments that evaluating the stability and the interpretability in a

soft-perturbation setting may lead to untrustworthy results. The best estimation seems to be obtained in the hard-perturbation setting experiments. The lack of stability between datasets has been explained by three arguments. First the data come from different technologic platforms. In this work, we only considered Affymetrix U133A data, thus getting rid of this bias. Second and third, the differences in experimentation and in patient cohorts. We compared datasets from different studies, i.e. meeting the two latter sources of bias. However, when dividing one set into two non-overlapping parts and repeating this operation several times on each dataset, we obtain stability results that are extremely similar to the *inter-datasets* results. This suggests that the main source of instability is the data itself: the higher stability observed in the soft-perturbation setting may only be explained by the overlap of the samples.

References

- Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, **26**(3), 392–398.
- Abraham, G., Kowalczyk, A., Loi, S., Haviv, I., and Zobel, J. (2010). Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC Bioinformatics*, **11**(1), 277.
- Ambroise, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA*, **99**(10), 6562–6566.
- Bach, F. (2009). Model-consistent sparse estimation through the bootstrap. *Arxiv preprint arXiv:0901.3202*.
- Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I., Soboleva, A., Tomashevsky, M., Marshall, K., *et al.* (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic acids research*, **37**(Database issue), D885.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.
- Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. (2003). Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.*, **3**, 1229–1243.
- Dai, M., Wang, P., Boyd, A., Kostov, G., Athey, B., Jones, E., Bunney, W., Myers, R., Speed, T., Akil, H., *et al.* (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic acids research*, **33**(20), e175.
- Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**(2), 171–178.
- Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA*, **103**(15), 5923–5928.
- Fan, C., Oh, D., Wessels, L., Weigelt, B., Nuyten, D., Nobel, A., van’t Veer, L., and Perou, C. (2006). Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.*, **355**(6), 560.
- Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**(1/3), 389–422.
- Ioannidis, J. P. A. (2005). Microarrays and molecular research: noise discovery? *Lancet*, **365**(9458), 454.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.
- Ivshina, A., George, J., Senko, O., Mow, B., Putti, T., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., *et al.* (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, **66**(21), 10292.
- Iwamoto, T. and Pusztai, L. (2010). Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data? *Genome Medicine*, **2**(11), 81.
- Kohavi, R. and John, G. (1997). Wrappers for feature selection. *Artificial Intelligence*, **97**(1-2), 273–324.
- Lai, C., Reinders, M., Van’t Veer, L., Wessels, L., *et al.* (2006). A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC bioinformatics*, **7**(1), 235.
- Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B*, **72**(4), 417–473.

- Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**(9458), 488–492.
- Pawitan, Y., Bjoehle, J., Amler, L., Borg, A., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., *et al.* (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, **7**(6), R953–R964.
- Reyal, F., Van Vliet, M., Armstrong, N., Horlings, H., De Visser, K., Kok, M., Teschendorff, A., Mook, S., Van't Veer, L., Caldas, C., *et al.* (2008). A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Res*, **10**(6), R93.
- Shen, R., Chinnaiyan, A., and Ghosh, D. (2008). Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC Medical Genomics*, **1**(1), 28.
- Shi, W., Bessarabova, M., Dosymbekov, D., Dezso, Z., Nikolskaya, T., Dudoladova, M., Serebryiskaya, T., Bugrim, A., Guryanov, A., Brennan, R. J., Shah, R., Dopazo, J., Chen, M., Deng, Y., Shi, T., Jurman, G., Furlanello, C., Thomas, R. S., Corton, J. C., Tong, W., Shi, L., and Nikolsky, Y. (2010). Functional analysis of multiple genomic signatures demonstrates that classification algorithms choose phenotype-related genes. *The pharmacogenomics journal*, **10**, 310–23.
- Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer. *N. Engl. J. Med.*, **360**(8), 790–800.
- Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., *et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *JNCI Cancer Spectrum*, **98**(4), 262.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**(1), 267–288.
- van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerckhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancers. *Nature*, **415**(6871), 530–536.
- Wang, J., Du, Z., Payattakool, R., Yu, P., and Chen, C. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**(10), 1274.
- Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M., Yu, J., Jatkoe, T., Berns, E., Atkins, D., and Foekens, J. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancers. *Lancet*, **365**(9460), 671–679.
- Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schutz, F., *et al.* (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res*, **10**(4), R65.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.