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

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

1 Accuracy across datasets

To estimate the predictive performance of a signature across datasets, we used 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 centroids (NC) classifier by 10-fold cross-validation on each dataset. We report the results in Table 1 as averages over the three test datasets. For each training dataset, we highlighted the method with the best results. In the last row, we report average results over the $4 \times 3 \times 10 = 120$ folds.

Training data	Type	Random	T-test	Entropy	Bhatt.	Wilcoxon	SVM RFE	GFS	Lasso	Elastic Net
GSE1456	S	0.59(0.10)	0.63(0.13)	0.60(0.10)	0.63(0.13)	0.61(0.14)	0.61(0.13)	0.61(0.11)	0.62(0.11)	0.62(0.11)
	E-M	0.60(0.12)	0.63(0.14)	0.60(0.12)	0.61(0.14)	0.61(0.14)	0.61(0.11)	0.60(0.12)	0.63(0.11)	0.60(0.12)
	E-E	0.60(0.13)	0.63(0.13)	0.58(0.10)	0.63(0.12)	0.61(0.13)	0.61(0.11)	0.62(0.12)	0.63(0.11)	0.62(0.11)
	E-S	0.60(0.14)	0.63(0.14)	0.59(0.10)	0.63(0.11)	0.61(0.13)	0.61(0.13)	0.62(0.13)	0.63(0.12)	0.63(0.09)
GSE2034	S	0.62(0.15)	0.62(0.15)	0.57(0.20)	0.59(0.19)	0.58(0.19)	0.60(0.18)	0.62(0.15)	0.63(0.16)	0.63(0.16)
	E-M	0.63(0.17)	0.63(0.15)	0.60(0.15)	0.64(0.16)	0.58(0.19)	0.63(0.17)	0.62(0.16)	0.62(0.16)	0.62(0.16)
	E-E	0.64(0.14)	0.63(0.15)	0.56(0.19)	0.58(0.19)	0.59(0.19)	0.63(0.16)	0.60(0.18)	0.61(0.16)	0.61(0.16)
	E-S	0.61(0.17)	0.63(0.16)	0.56(0.17)	0.57(0.19)	0.59(0.19)	0.63(0.15)	0.62(0.17)	0.62(0.16)	0.63(0.16)
GSE2990	S	0.64(0.14)	0.64(0.15)	0.56(0.14)	0.60(0.16)	0.60(0.16)	0.62(0.16)	0.64(0.15)	0.66(0.13)	0.65(0.13)
	E-M	0.61(0.15)	0.66(0.16)	0.59(0.17)	0.65(0.13)	0.58(0.16)	0.65(0.15)	0.62(0.14)	0.64(0.15)	0.64(0.15)
	E-E	0.61(0.14)	0.66(0.15)	0.54(0.14)	0.57(0.19)	0.59(0.15)	0.62(0.15)	0.63(0.15)	0.65(0.14)	0.66(0.14)
	E-S	0.62(0.15)	0.66(0.14)	0.55(0.14)	0.57(0.18)	0.60(0.16)	0.64(0.15)	0.63(0.14)	0.65(0.14)	0.65(0.14)
GSE4922	S	0.65(0.15)	0.66(0.15)	0.59(0.16)	0.63(0.14)	0.64(0.16)	0.64(0.14)	0.62(0.12)	0.65(0.14)	0.65(0.14)
	E-M	0.65(0.12)	0.67(0.15)	0.64(0.13)	0.66(0.16)	0.65(0.15)	0.64(0.13)	0.65(0.15)	0.66(0.14)	0.64(0.13)
	E-E	0.65(0.15)	0.66(0.15)	0.57(0.16)	0.63(0.15)	0.66(0.15)	0.64(0.12)	0.65(0.13)	0.67(0.13)	0.66(0.14)
	E-S	0.65(0.15)	0.65(0.15)	0.60(0.16)	0.62(0.16)	0.66(0.16)	0.63(0.12)	0.63(0.10)	0.66(0.13)	0.65(0.13)
Average	S	0.62(0.14)	0.64(0.15)	0.58(0.15)	0.61(0.15)	0.61(0.16)	0.62(0.15)	0.62(0.13)	0.64(0.13)	0.64(0.14)
	E-M	0.62(0.14)	0.65(0.15)	0.61(0.15)	0.64(0.15)	0.61(0.16)	0.63(0.14)	0.62(0.14)	0.64(0.14)	0.62(0.14)
	E-E	0.62(0.14)	0.64(0.15)	0.56(0.15)	0.60(0.17)	0.61(0.16)	0.63(0.13)	0.62(0.14)	0.64(0.14)	0.64(0.14)
	E-S	0.62(0.15)	0.64(0.15)	0.58(0.15)	0.60(0.16)	0.61(0.16)	0.63(0.14)	0.62(0.14)	0.64(0.14)	0.64(0.13)

Table 1: 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.

2 Stability as a function of signature size

We observe in Figure 1 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

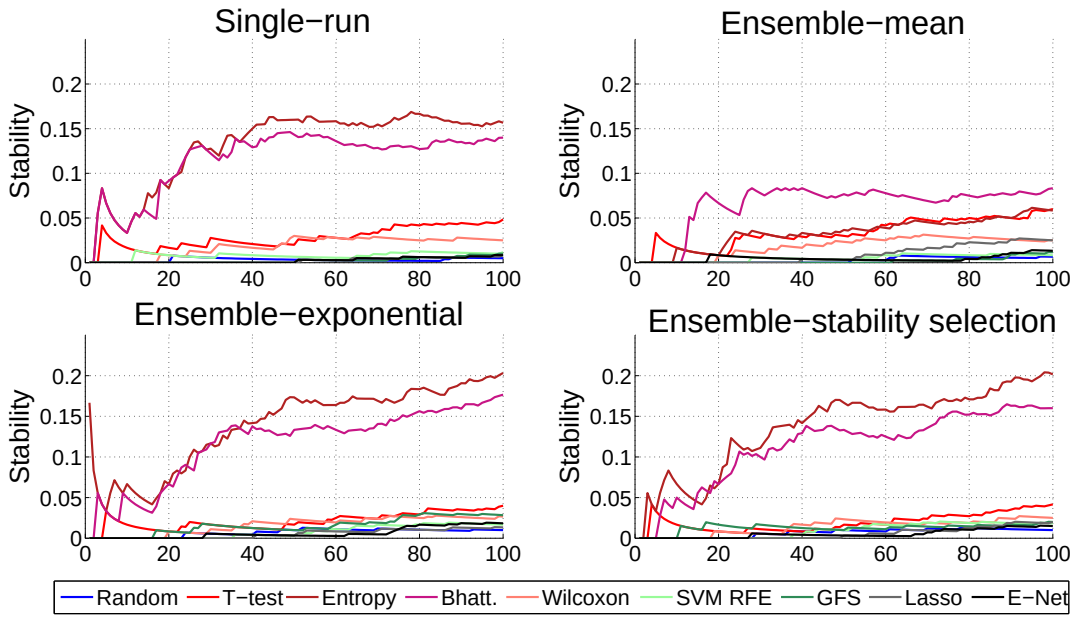


Figure 1: Stability of different methods in the between-dataset setting, as a function of the size of the signature.

3 Stability in the hard-perturbation and between-dataset settings

The lack of stability observed between different signatures can be attributed to different factors, including (i) diferences 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 small size effect by testing the stability of signatures estimated on non-overlapping bootstrap samples of a given dataset where 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 2 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 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.

4 Functional stability in the soft- and hard-perturbation settings

Figures 3a, 3b and 3c show the functional stability for all methods in the three settings. While the baseline stability, as obtained by random signatures, is approximatively the same regardless of

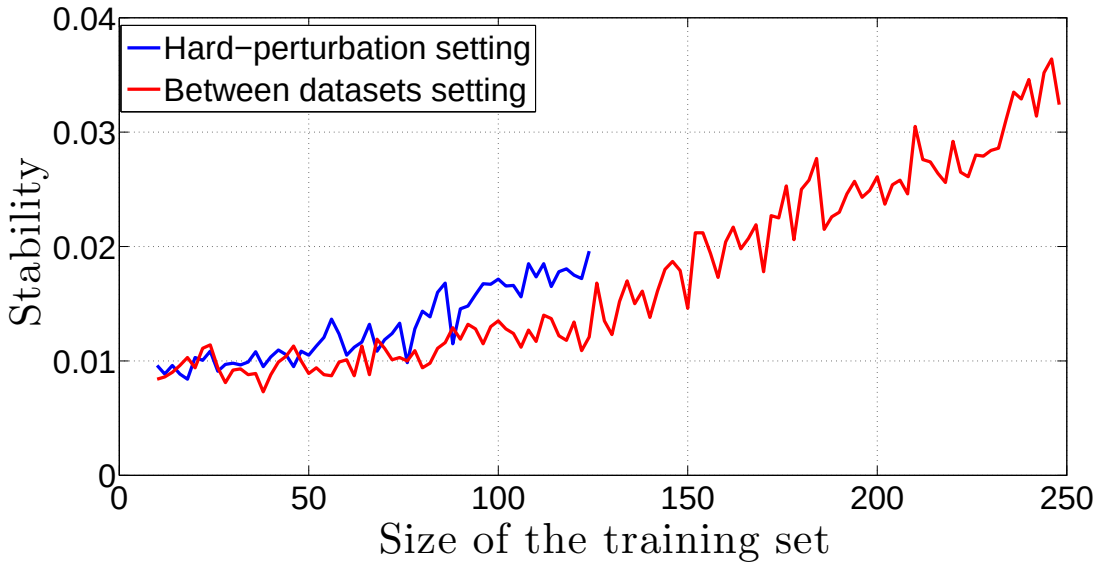


Figure 2: Evolution of stability of t-test signatures with respect to the size of the training set in the hard-perturbation and the between datasets settings from GSE2034 and GSE4922.

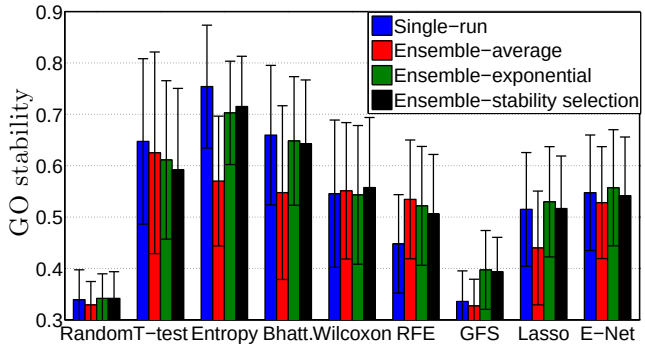
the setting, we observe that, like stability at the gene level, soft- and hard-perturbation can lead to very different interpretations. This suggests again that the high functional stability obtained by several methods in the soft-perturbation setting is mainly due to the overlap in samples. Hence the hard-perturbation setting seems to be a much better proxy for the between-datasets framework.

5 Issues with selection by relative entropy and Bhattacharyya distance

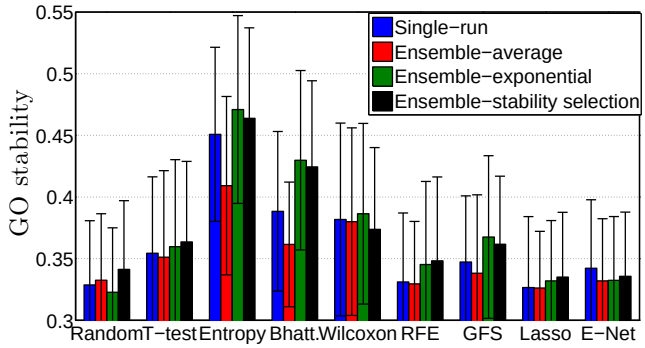
Gene selection by relative entropy and Bhattacharyya distance is more stable but less accurate than random selection, which suggests a bias in the method which may preferably and consistently select particular genes, not necessarily very predictive. To elucidate this behavior, we investigated the genes selected by these two methods. We noticed that they tend to be systematically expressed at low levels, as shown in Figure 4, and that they barely depend on the labels, which explains the high stability but small accuracy. In fact the frequently selected genes systematically show a multimodal yet imbalanced distribution due to the presence of outliers, as illustrated on figure 5. As soon as, by chance, one class of samples contains one or more outliers when the other class doesn't, this type of distribution is responsible for a very high variance ratio between the two classes, thus leading to a very high value of the entropy and Bhattacharyya statistics. It is therefore likely that, although stable and interpretable, the molecular signatures generated by these two methods lead to erroneous interpretation.

References

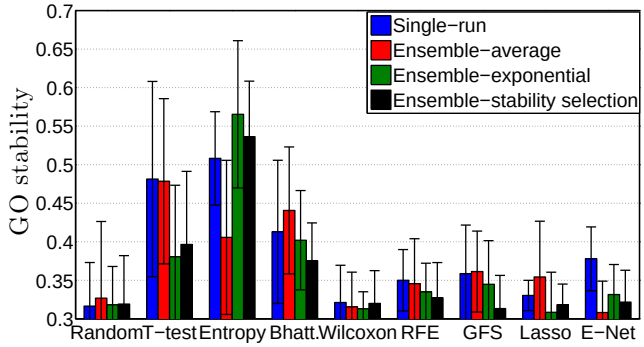
Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA*, **103**(15), 5923–5928.



(a)



(b)



(c)

Figure 3: GO 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.

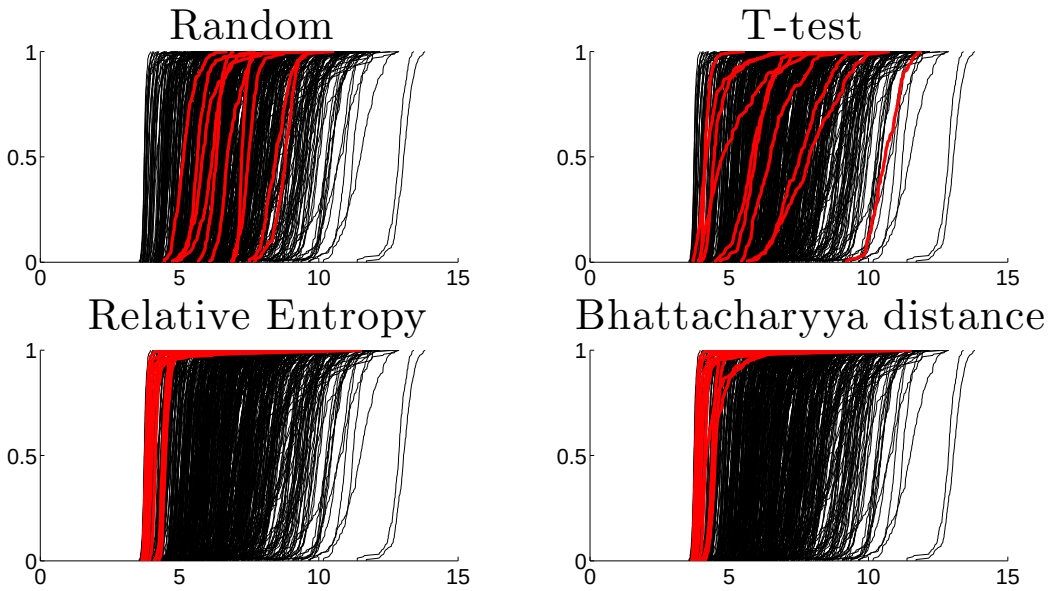


Figure 4: Estimated cumulative distribution functions (ECDF) of the first ten genes selected by four methods on GSE1456. They are compared to the ECDF of 500 randomly chosen background genes.

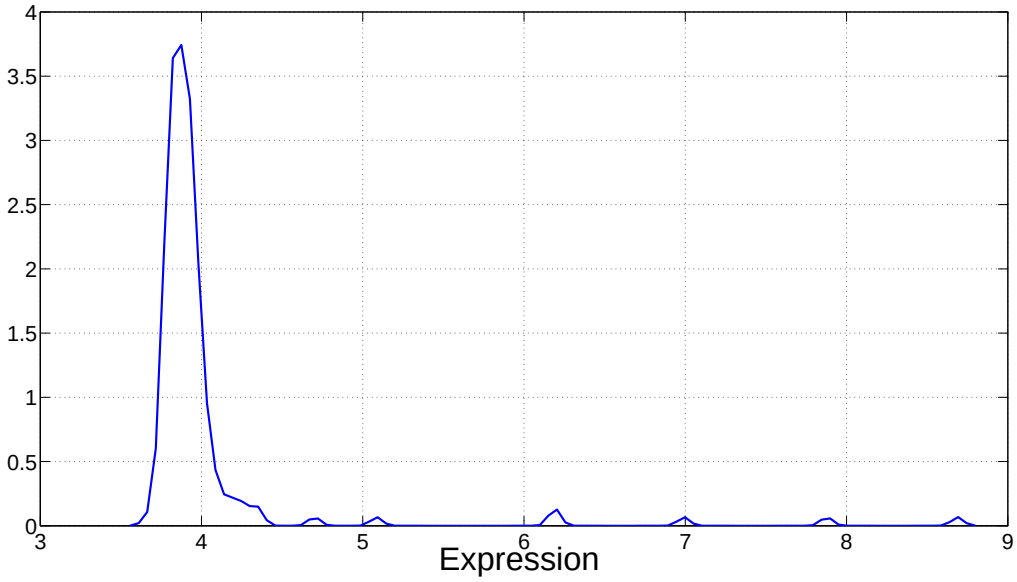


Figure 5: Estimated distribution of the first gene selected by entropy and Bhattacharyya distance.