# Robust biomarker identification for cancer diagnosis using ensemble feature selection methods

Thomas Abeel [1,2], Thibault Helleputte [3], Yves Van de Peer [1], Pierre Dupont [3] and Yvan Saeys [1]

1) Department of Plant Systems Biology, VIB, Gent, Belgium and Department of Molecular Genetics, Ghent University, Gent, Belgium – http://bioinformatics.psb.ugent.be
2) Broad Institute of MIT and Harvard, Cambridge, USA
3) Department of Computing Science & Engineering INGI – Université catholique de Louvain , UCL Machine Learning Group - http://www.ucl.ac.be/mlg

**Contact: yvan.saeys@ugent.be**

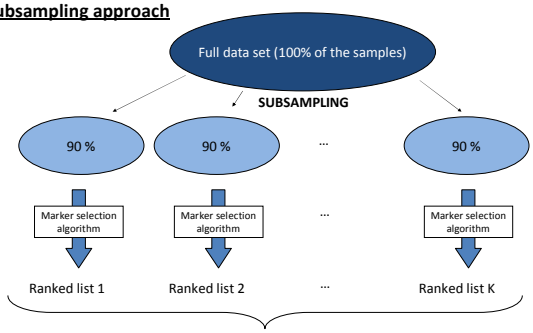
## The Idea

Biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high dimensional data [1] . Surprisingly, the stability with respect to sampling variation or robustness of such selection processes has received attention only recently. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. In addition, a more robust set of markers may strengthen the confidence of an expert in the results of a selection method.

Our first contribution is a general framework for the analysis of the robustness of a biomarker selection algorithm. Secondly, we conducted a large scale analysis of the recently introduced concept of ensemble feature selection [2], where multiple feature selections are combined in order to increase the robustness of the final set of selected features. We focus on selection methods that are embedded in the estimation of support vector machines (SVMs).

## Measuring robustness

**Subsampling approach**




Calculate robustness as average pair wise similarities between all possible signature pairs ($\mathbf{f}_i$,$\mathbf{f}_j$) :

$$\text{Robustness} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k}\mathrm{KI}(\mathbf{f}_i,\mathbf{f}_j)}{k(k-1)} \quad \text{with} \quad \mathrm{KI}(\mathbf{f}_i,\mathbf{f}_j) = \frac{r\cdot N - s^2}{s\cdot(N-s)} \quad [3]$$

where

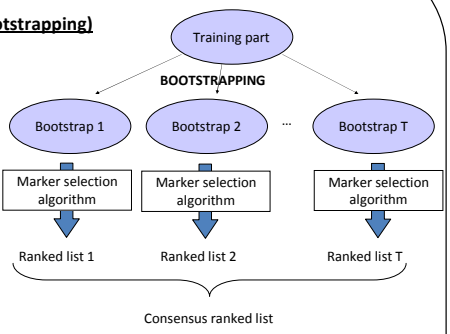
- $s = |\mathbf{f}_i| = |\mathbf{f}_j|$ denotes the signature size
- $r = |\mathbf{f}_i \cap \mathbf{f}_j|$ is the number of common elements in both signatures
- $N$ denotes the number of features (dataset dimensionality)

## Ensemble feature selection

**Resampling approach (bootstrapping)**

*Based on the idea of ensemble classifiers, **ensemble feature selection techniques** combine different feature selection techniques.*

*Two components are essential to each feature selection ensemble: a) a method to obtain **diverse** feature selection algorithms, and b) a way to **aggregate** the single feature selection algorithms into an overall ensemble/ consensus result.*

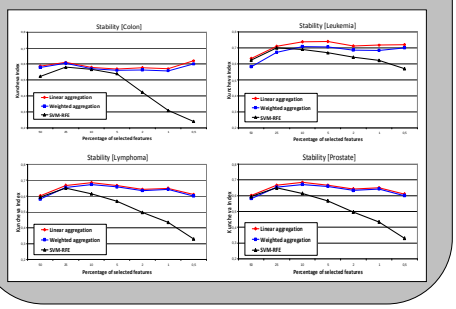



**Aggregation operators to create the consensus:**

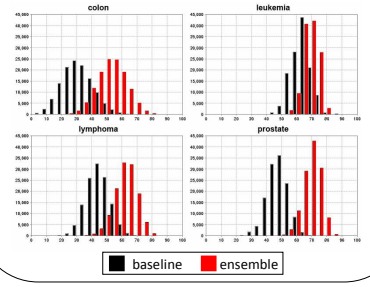
- Linear aggregation: consensus ranking is obtained by summing the ranks over all bootstrap samples
- Weighted aggregation: consensus ranking is obtained by a *weigthed* sum of the ranks over all bootstrap samples. The weights are obtained by evaluating the resulting on the out-of-bag samples.

## Experiments and Results

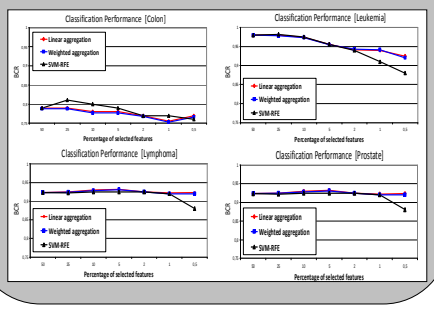
***Stability*** of the baseline method (SVM-RFE [4]) compared to two different ensemble methods, using the ***Kuncheva Index***. Default: 40 bootstraps in the ensemble, 20% of features eliminated at each iteration of SVM-RFE.




Stability distributions of the baseline method (SVM-RFE) compared to the linear aggregation ensemble.




***Classification performance*** of the baseline method (SVM-RFE [4]) compared to two different ensemble methods, using the ***Balanced Classification Rate*** (BCR). Default: 40 bootstraps in the ensemble, 20% of features eliminated at each iteration of SVM-RFE.




## Conclusions

**Stability**

- Ensemble feature selection (EFS) techniques outperform the baseline largely in terms of feature selection stability
- Especially for smaller signature sizes, the gain in stability of EFS methods increases
- More bootstraps in the ensemble increase the method's stability

**Classification performance**

- For the balanced classification rate, ensemble methods perform equally well, compared to the baseline
- For very small signature sizes, ensemble methods outperform the baseline in terms of balanced classification rate

## References/Acknowledgements

All implemented algorithms are available in our open source package **JAVA-ML**, available at http://java-ml.sf.net

⇒ see our other poster *"Java-ML a Java Library for Data Mining"*

[1] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.

[2] Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Proceedings of the 25th ECML/PKDD*, Part II, pages 313–325.

[3] Kuncheva, L. (2007). A stability index for feature selection. In Proceedings of the 25th International Multi-Conference on Artificial Intelligence and Applications, pages 309–395.

[4] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), 389–422.