A miRNA expression based diagnostic tool for breast cancer using Random Forests

Stephane Wenric[1,2*], Pierre Freres[2], Claire Josse[1,2], Vincent Bours[1], Guy Jerusalem[2]

University of Liege, GIGA-Research, Human Genetics Unit[1] ; University of Liege Hospital (ULg CHU), Medical Oncology Laboratory[2]. *s.wenric@ulg.ac.be

Introduction

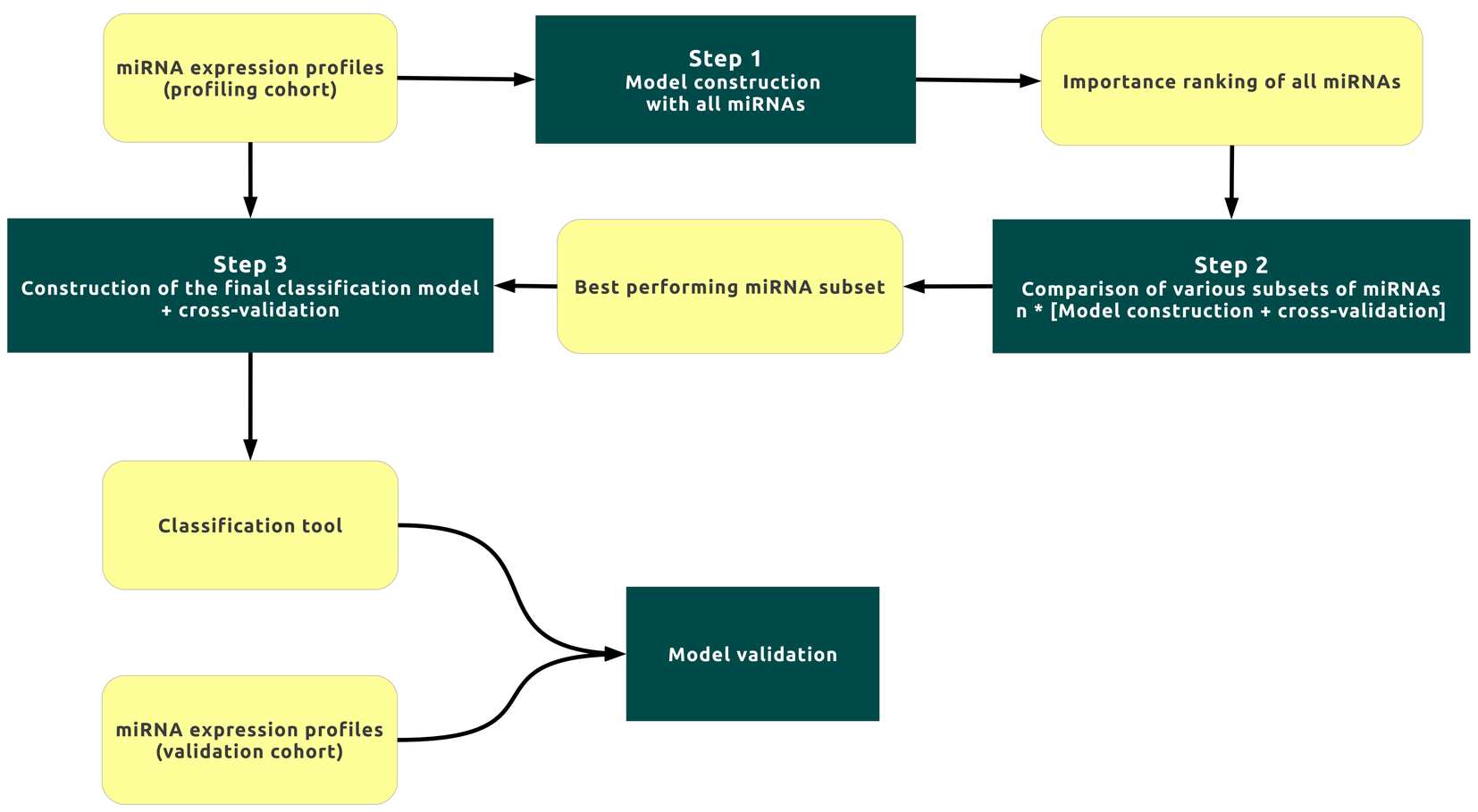
Breast cancer is the leading cause of death by cancer among women and there is a need to improve diagnosis methods. MicroRNAs (miRNAs) are noncondig RNAs that regulate gene expression and many have been implicated in breast cancer. Here, we show that **an accurate diagnostic tool for breast cancer can be built based on the expression levels of 8 circulating miRNAs (out of 188 probed miRNAs) and the use of the Random forests classification algorithm.**

Methods

The expression levels of **188 circulating miRNAs** was determined for **101 patients** with breast cancer and **20 controls.**
The individuals were randomly separated into **2 independent cohorts** with the same patients/controls ratio:

- **profiling cohort, *n = 85***
- **validation cohort, *n = 36***

A **Random forests model** using all 188 miRNAs has been built on the profiling cohort, yielding two variable importance metrics (*mean decrease in accuracy* and *mean decrease in Gini*)
Based on these metrics, the miRNAs are ranked, and a selection of 15 miRNAs (which were all ranked among the 20 first miRNAs of both rankings) has been performed.



From these 15 miRNAs, 32767 combinations of 2 to 15 miRNAs have been generated.
A Random forests model was then built for each of these combinations, and its classification performance was assessed by carrying **ten-fold cross-validation** and comparing the resulting AUC.
Finally, the model built with the profiling cohort and **the best performing combination of miRNAs has been validated by predicting the classes of each individual of the validation cohort.**

Results & Discussion

The best performing combination of miRNAS among the 32767 combinations was composed of 8 miRNAs and yielded an AUC of 0.9625 when using ten-fold cross-validation on the profiling cohort.

The model built with the profiling cohort and said combination of miRNAs has been validated by predicting classes for the independent validation cohort, and gave an **AUC of 0.9522.**

To our knowledge, this is the first time the Random forests method is used to perform classification using circulating miRNAs expression levels as features.

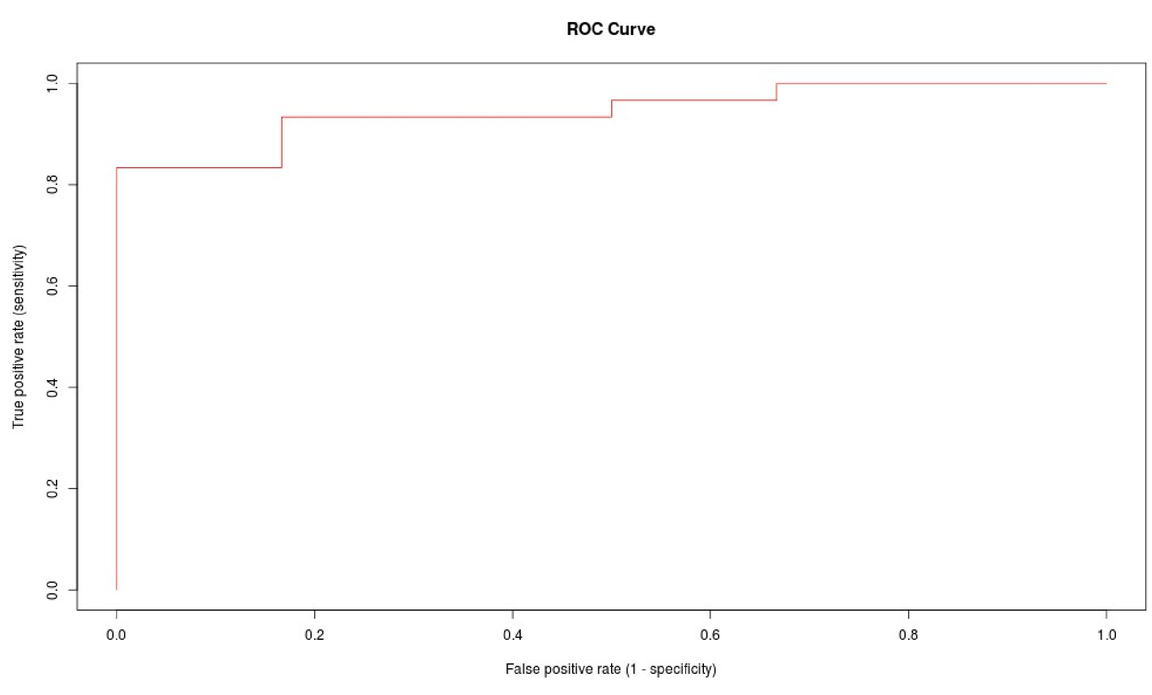


***Figure 1**: ROC curve obtained through validation of the final model (built on the profiling cohort, with 8 miRNAs) on the independent validation cohort. AUC = 0.9522*