Biomarker signatures discovery to support cancer diagnosis
Towards an accurate and robust machine learning strategy

Ahmed Debit, PhD Student (*)
adebit@uliege.be

(*) GIGA-R, Medical Genomics, University of Liège, Belgium
Context

- Diagnosis of Breast Cancer
- Breast Cancer treatment response
- Design of short biomarker signatures
From Omics to Clinics

Omics
~20k genes (RNA-seq)

Clinics
Dozen of genes (biomarkers)
From Omics to Clinics

Omics
~20k genes (RNA-seq)

Clinics
Dozen of genes (biomarkers)

Signatures of biomarkers

Cancer Diagnosis
Cancer Prognosis
From Omics to Clinics

Omics
~20k genes (RNA-seq)

Clinics
Dozen of genes (biomarkers)
Signatures of biomarkers
Robust and precise
From Omics to Clinics

Omics

~20k genes (RNA-seq)

Clinics

Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise Models and signatures

Deep Learning

Genetic Algorithm

Boosting

Unsupervised Clustering

Neural Networks

Random Forest

SVM
From Omics to Clinics

Omics
~20k genes (RNA-seq)

Clinics
Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise Models and signatures

Random Forest

R implementations of Random Forest algorithm proposed by [Breiman et al. 2001]
From Omics to Clinics

Omics
~20k genes (RNA-seq)

Clinics
Dozen of genes (biomarkers)

Signatures of biomarkers

Robust and precise
Models and signatures

Random Forest

R implementations of
Random Forest algorithm
proposed by [Breiman et al. 2001]
Objectives

Toward a robust RF method for the Biological question asked

Which method is suitable for which dataset (platform/technology)?
Objectives

• Empirical comparison of random forest based methods

• Differences/Similarities of RF methods → groups of methods

• Designing a high stability score to rank RF methods

Toward a robust RF method for the Biological question asked

Which method is suitable for which dataset (platform/technology)?
Materials and Methods

- Datasets (Perfectly balanced)
  - TCGA-BRCA (RPKM): 182 samples x 9560 genes
  - TCGA-LUSC (RPKM): 96 samples x 9262 genes

- Main classification question

  The difference between paired Tumor / Normal samples will be used as a strong classification parameter, allowing for strong modeling only
Overview of the method

Step I
Stable RF-based FS
- Number of more stable & important variables indexVar**
- Number of trees nTrees*

Step II
Pre-comparison
- List of signatures
- List of Partitions
- Number of trees nTrees**

Step III
Comparison
- Construction of q ∈ [5, 15, 25] classification models for each training partition

Extreme dimensionality reduction

Dataset

Orthogonal
- randomForest
- randomForestSRC
- Ranger
- rf
- mRanger
- extraTrees
- randomUniformForest
- RRF
- WSRF
- IterativeForest

Oblique
- canonicalForest
- PForest
- obliqueRF
- rotationForest
- randomForest
Overview of the method

Extreme dimensionality reduction

Generation of random combinations
Generation of random training partitions
Assessment of RF parameters \( ntrees \)

Step I
Stable RF-based FS
- Number of more stable & important variables \( nbVar*\)
- Number of trees \( ntrees*\)

Step II
Pre-comparison
- List of signatures
- List of Partitions
- Number of trees \( ntrees**\)

Step III
Comparison
- Construction of \( q \in \{5, 15, 25\} \) classification models for each training partition

Dataset

Orthogonal
- randomForest
- randomForestSRC
- Ranger
- xgboost
- extraTrees
- randomUniformForest
- RRF
- WRF
- HerdingForest

Oblique
- canonicalForest
- PPRforest
- obliqueRF
- rotationForest
- randomForest
Overview of the method

1. **Step I**
   - Stable RF-based FS
   - Number of more stable & important variables \( nbVar** \)
   - Number of trees \( ntrees^* \)

2. **Step II**
   - Pre-comparison
   - List of signatures
   - List of Partitions
   - Number of trees \( ntrees^{**} \)

3. **Step III**
   - Comparison
   - Metrics
     - AUC
     - Runtime

---

**Extreme dimensionality reduction**

**Generation of random combinations**

**Generation of random training partitions**

**Assessment of RF parameters (ntrees)**

**RF stability assessment**
Stable Feature Selection
Extreme dimensionality reduction
First pass Feature Selection results

TCGA-BRCA dataset
First pass Feature Selection results

Number of trees ($n_{trees}$) = 2000

Number of variables ($n_{Var}$) = 200

TCGA-BRCA dataset
First pass Feature Selection results

\[ n_{trees}^* = 2000 \]
\[ n_{Var}^* = 200 \]

~9000 to 200 variables (Genes)

TCGA-BRCA dataset
Second pass Feature Selection results

TCGA-BRCA dataset
Second pass Feature Selection results

TCGA-BRCA dataset

\[ nVar^{**} = 30 \]
Second pass Feature Selection results

\[ n_{Var}^{**} = 30 \]

TCGA-BRCA dataset
Second pass Feature Selection results

\[ n\text{Var}^{**} = 30 \]

TCGA-BRCA dataset

~200 to 30 variables (Genes)
Pre-Comparison
Assessment of RF parameters & Generation of random combinations
Pre-comparison

I- Generation of random combinations (Cancer signatures)

• Multiple predictive models using combinations of different lengths

\[ (2^{nVar^*} - 1) \] combinations

Random selection of 3 signatures per length

II- Generation of random training partitions

• 50 random training partitions

training partition = a set of samples used to construct a model
Pre-comparison

III- Tuning the parameter $ntrees$ for each RF method
Pre-comparison

III- Tuning the parameter \textit{ntrees} for each RF method

TCGA-BRCA dataset
## Summary of step I + step II

<table>
<thead>
<tr>
<th>Dataset</th>
<th>nVar</th>
<th>nVar*</th>
<th>nVar**</th>
<th>ntrees*</th>
<th>ntrees**</th>
<th>Nbre combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA-BRCA</td>
<td>9560</td>
<td>200</td>
<td>30</td>
<td>2000</td>
<td>500</td>
<td>78</td>
</tr>
<tr>
<td>TCGA-LUSC</td>
<td>9262</td>
<td>200</td>
<td>10</td>
<td>2000</td>
<td>500</td>
<td>21</td>
</tr>
</tbody>
</table>
Comparison

Random Forest stability assessment
Random Forest Method Comparison

- Comparison of RF methods under **same** conditions
- Using **same** random training partitions
- Assessing the **same** signatures
- On computational cores of **same** characteristics
For each signature, we’ll focus on:

- 50 resampling to build the Training and the Validation set.
- 25 modeling and validations.

Analysis of:

- Coefficient of Variation of 1,250 models & AUCs
Random Forest Method Comparison

• For each signature, we’ll focus on:
  • 50 resampling to build the Training and the Validation set.
  • 25 modeling and validations.

• Analysis of:
  • Coefficient of Variation of 1,250 models & AUCs

Clinics ➔ Hyper Stability : CV == 0
Hyper Stability discriminates RF methods

TCGA-LUSC dataset

Training partitions

Signatures

Coefficient of variation of pd_tidy_tumor_AUC over signature and resampling

Resample01
Resample03
Resample05
Resample07
Resample09
Resample11
Resample13
Resample15
Resample17
Resample19
Re...
Hyper Stability discriminates RF methods

Best Methods

Worst Methods

TCGA-LUSC dataset
Hyper Stability Score helps finding the best method(s)

TCGA-BRCA

- Hyper Stability Score / Average AUC
- Average modeling Time (ms)

RF methods

- cc
- cForest
- extraTrees
- iforest
- obliqueRF
- PpForest
- randomForest
- ranger
- randomUniformForest
- Rforest
- Rq
- ssr
Hyper Stability Score is dataset dependent

TCGA-BRCA

Hyper Stability Score / Average AUC

TCGA-LUSC

Hyper Stability Score / Average AUC
Conclusions

• The AUC precision is dataset dependent
  • The Methods are dataset dependent.

• Trade-off:
  • AUC precision (hyper-stability)
  • Average AUC value
  • Modeling Time

Classification of classification methods
Towards robust signatures and predictions
Perspectives

- Datasets of same dimensions
- Same datasets coming from other platforms: micro-array, ...
- Other machine learning methods
Acknowledgment

**Human Genetics (GIGA)**
- Prof. Vincent Bours, MD, PhD
- Christophe Poulet, PhD
- Corinne Fasquelle, Ir

**BIO3 Unit (GIGA)**
- Prof. Kristel Van Steen

**Oncology (CHU-Liege)**
- Prof. Guy Jerusalem, MD
- Dr. Claire Josse
- Jerome Thiry

**CBIO, Mines ParisTech, Institut Curie**
- Chloe-Agathe Azencott, PhD