AUCRF: A LIBRARY FOR GENOMIC PROFILING

M.Luz Calle, Víctor Urrea

Bioinformatics and Medical Statistics Group (UVic)

Context

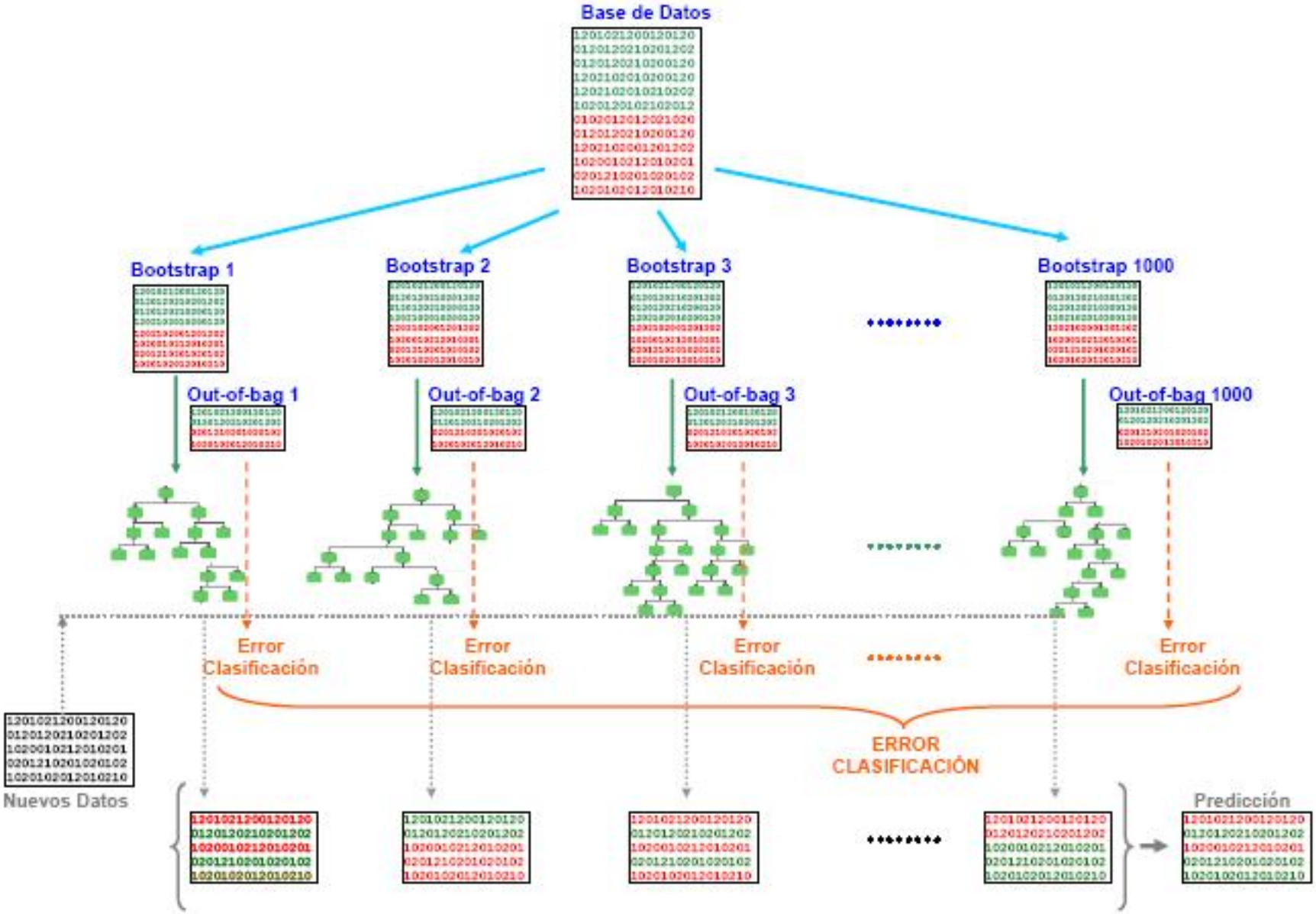
- Genomic profiling is the use of genetic variants at multiple loci simultaneously for prediction of disease risk
- Requires the selection of the set of genetic variants that best predicts disease status
- Focus on variable selection for prediction: selection based on the predictive accuracy of the selected set of variables

Outline

- AUCRF implements an approach for variable selection using Random Forest in case/control studies:

1. Explores the performance of RF through the ROC curve and AUC
2. Variable selection using RF based on optimizing the AUC
3. Predictive accuracy of selection by cross-validation
4. Provides the probability of selection as measure of robustness of the selection

Random Forest



Prediction Based on Votes

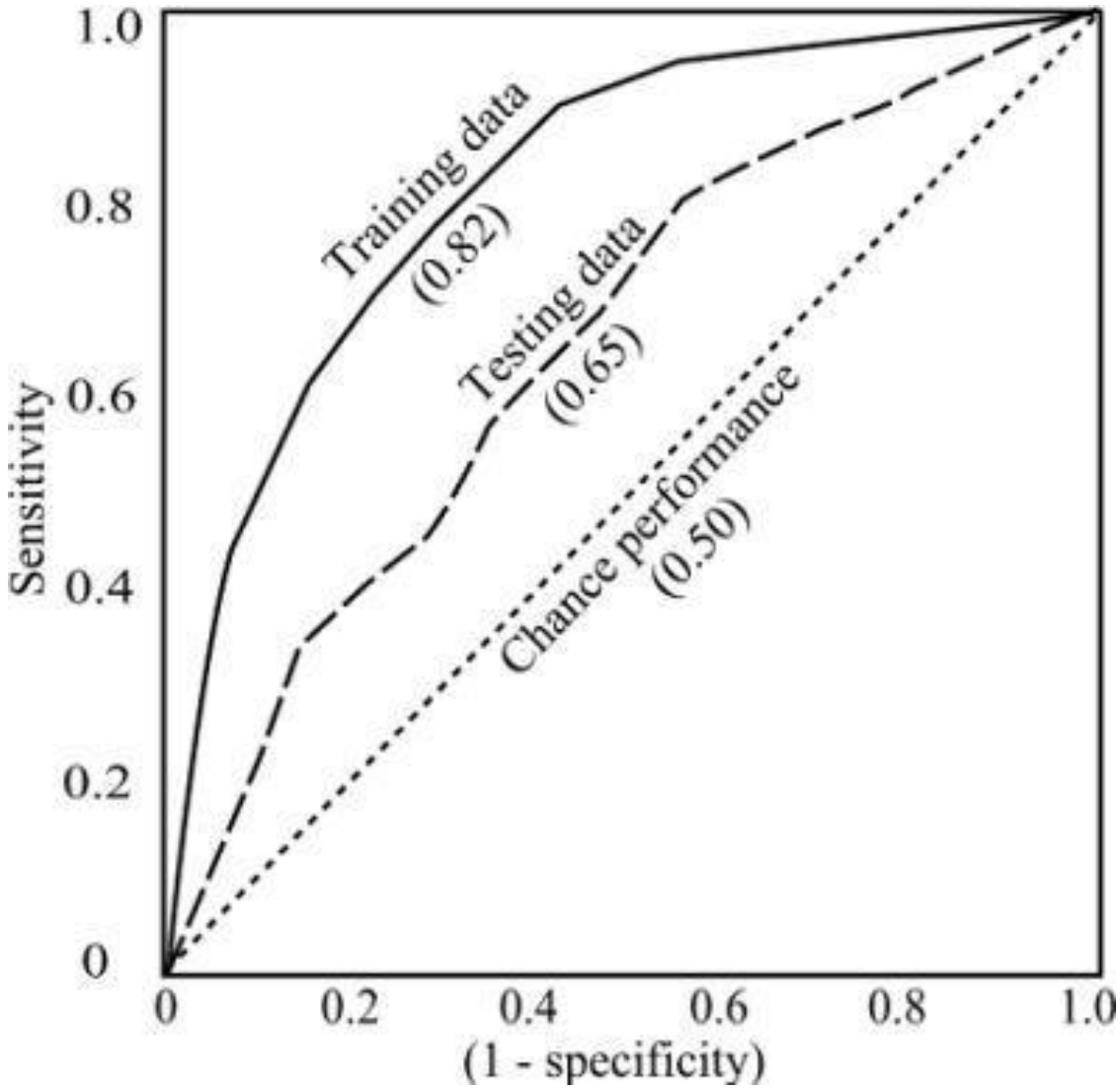
- RF makes class prediction based on votes:

# trees that votes...		Total votes	RF Prediction	Outcome
0 *(control)*	1 *(case)*			
163	198	361	1	1
250	138	388	0	0
372	2	374	0	0
340	4	344	0	0
167	198	365	1	1
180	186	366	1	0
212	127	339	0	1
357	1	358	0	0
297	74	371	0	1
272	87	359	0	0
387	1	388	0	0
157	182	339	1	1
365	5	370	0	0
296	64	360	0	1
215	144	359	0	0
186	194	380	1	0

ROC CURVE AND AUC OF A RF

- The performance of a RF is explored through the ROC curve and its AUC:
  - Varying the probability threshold and explore the proportions of FN and FP
  - This allows to obtain the ROC curve (1-specificity vs sensitivity) of the RF
  - Compute the AUC as the predictive accuracy measure of the RF

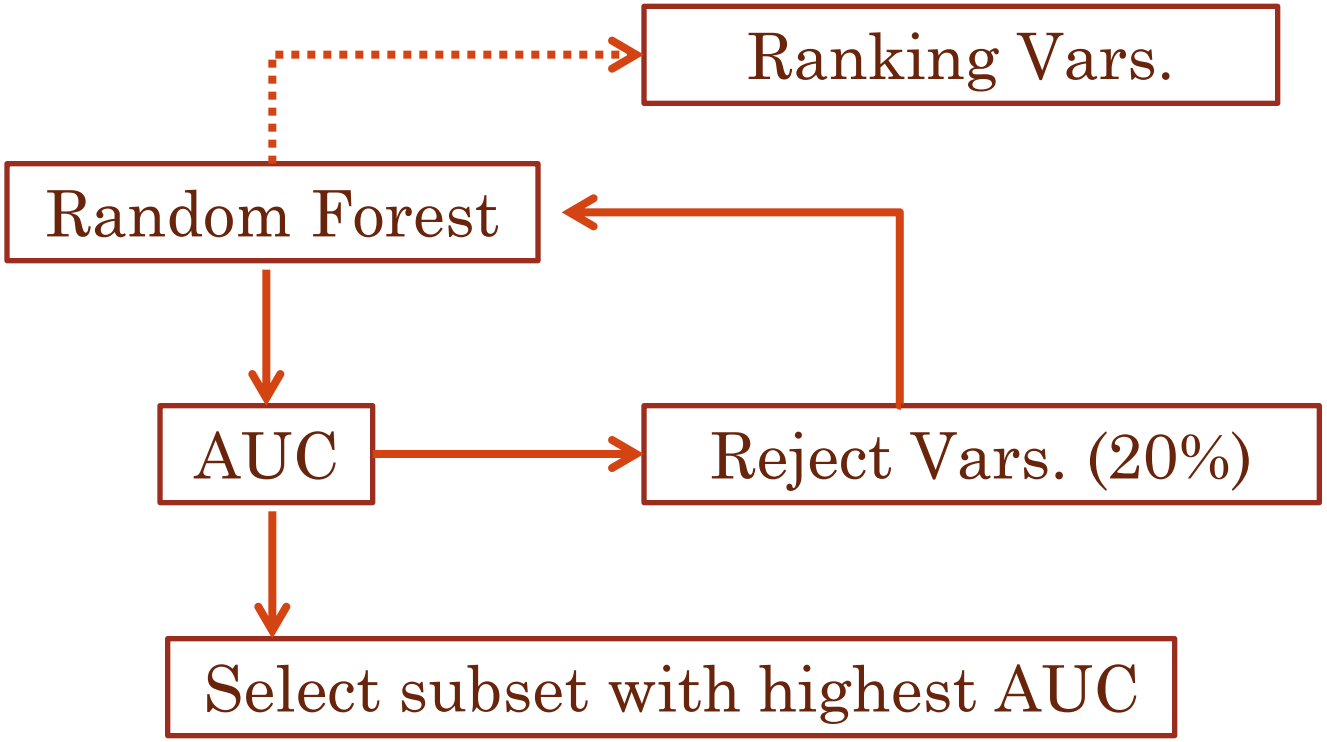
ROC CURVE



TP frac

FP frac

AUCRF WORK FLOW



- Modification of the method proposed in *Gene selection and classification of microarray data using random forest* by Diaz-Uriarte et al. (2006), based on overall prediction accuracy

AUCRF Usage

```
AUCRF ( formula, data,
    k0=1, pdel=0.2, ranking=c("MDG","MDA"), ...)

AUCRFcv ( x, nCV = 5, M = 20)
```

Examples:

```
AUCRF(Y~., data=exampleData, ntree=1000, nodesize=20)

fit   <- AUCRF(Y~., data=exampleData)
fitCV <- AUCRFcv(fit)
```

AUCRF OUTPUT

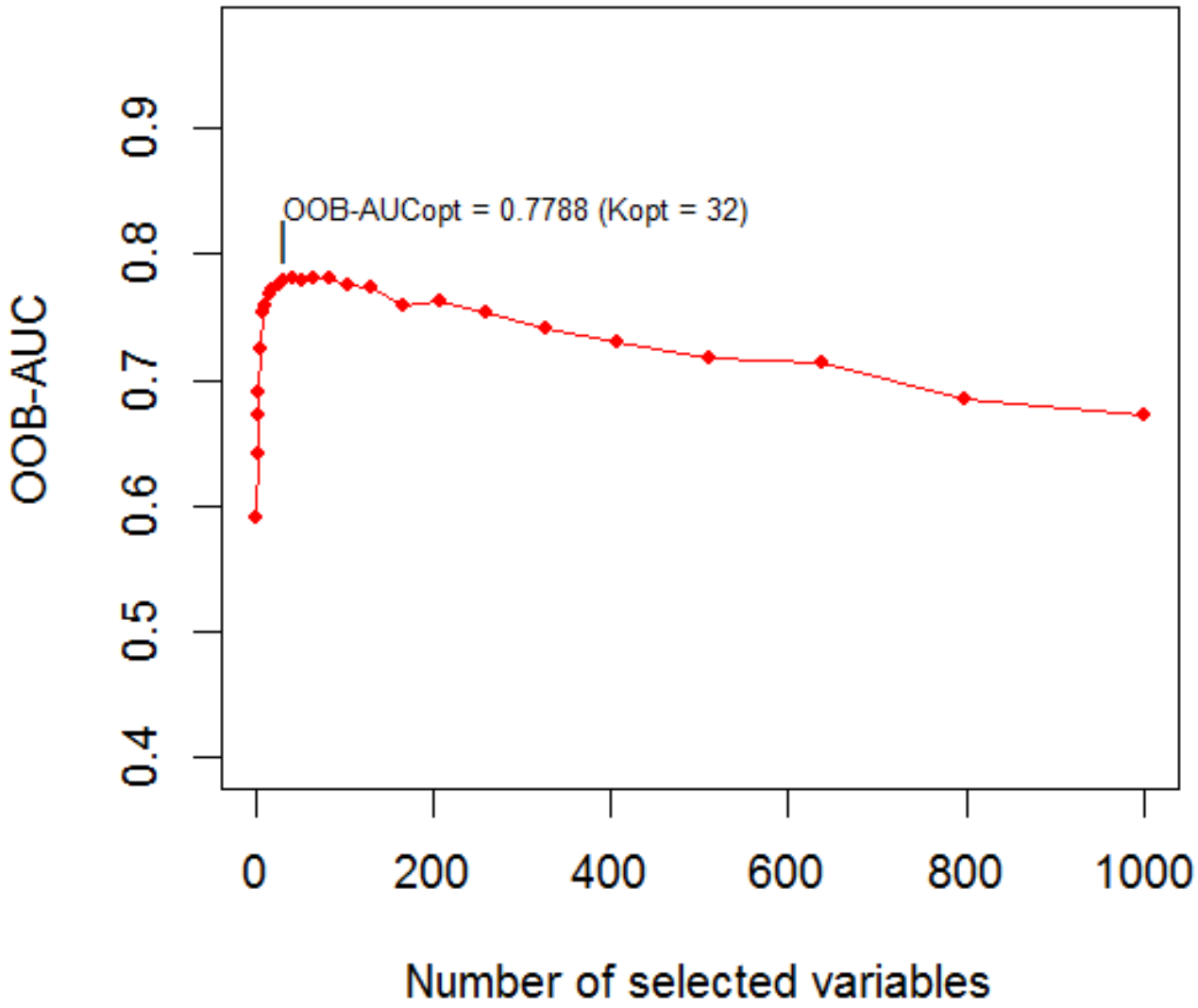
```
AUCRF (Y~. , data, pdel = 0.2, ranking="MDG“, )
AUCRFcv (fit, nCV = 5, M = 20)
```

```
    Number of selected variables: Kopt= 32
    AUC of selected variables: OOB-AUCopt= 0.7787711
    AUC from cross validation: 0.759109
    Importance Measure: MDG
```

	Selected.Variables	Importance	Prob.Select
1	SNP9	15.047305	1.00
2	SNP4	12.912120	1.00
3	SNP3	10.486599	1.00
4	SNP7	9.767075	1.00
5	SNP8	9.283819	1.00
6	SNP2	9.043039	1.00
7	SNP6	8.743129	0.95
8	SNP10	8.465736	0.92
9	SNP5	7.844703	0.80
10	SNP1	7.533021	0.77
11	SNP369	2.677609	0.35

Backward elimination process

```
plot.AUCRF (fit)
```



Human Heredity
International Journal of
Human and Medical Genetics

Vol. 72, No. 2, 2011

Original Paper

AUC-RF: A New Strategy for Genomic Profiling with Random Forest

M. Luz Calle[a], Victor Urrea[a], Anne-Laure Boulesteix[c], Nuria Malats[b]

[a]Systems Biology Department, University of Vic, Vic, and
[b]Centro Nacional de Investigaciones Oncológicas, Madrid, Spain;
[c]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Munich, Germany

Address of Corresponding Author

Hum Hered 2011;72:121-132 (DOI: 10.1159/000330778)