Bioassay Classification Study via LC-MS and Machine Learning in Conjunction with Dimensionality Reduction

Ivan Plyushchenko, Igor Rodin

Lomonosov Moscow State University, Chemistry Department, 119992, GSP-2, Lenin Hills, 1, Moscow, Russia

plyush1993@bk.ru

Computation scheme

- Initiate multicore processing.

packages (“doParallel” , “parallel”)

- Load RAW data.

- Univariate filtering (UVF).

Sequential implementation: Shapiro-Wilk, Wilcox, Bartlett, Student tests (Benjamini-Hochberg correction).
package (“stats”)

- Repeated Machine Learning (RML)

Train 4 models (RF, SVM with RBF, KNN, PLS) with 5 tune length and 3 repeats of 5 fold cross-validation and also repeat all calculations 10 times.
package (“caret”)

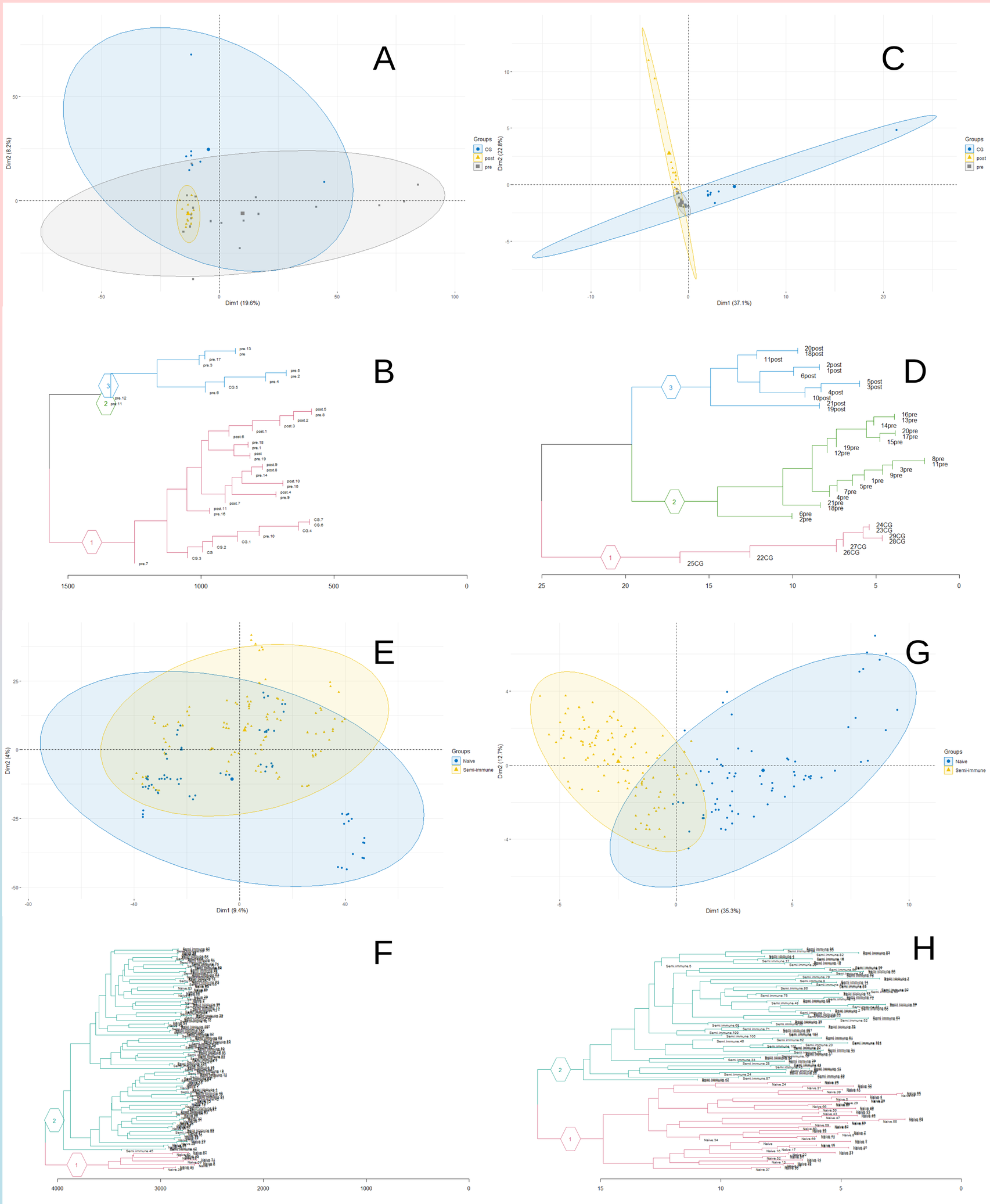
- Recursive Feature Elimination (RFE)

For all models with mean accuracy ≥ 80% generate set of unique features (initial numbers of features: 0.5*√TotalFeatures(TF), √TF, 2* √TF, 3* √TF, 5* √TF) simultaneously with 100% of frequency of features and by top mean rank of variable importance from all 10 outputs of all selected models.
Select best set of features by Naïve Bayes (NB) classification (min number of features with max accuracy), then optimize set by RFE with NB.
package (“caret”, “Deducer”)

- Unsupervised Learning (UL) by HCA (distance = "canberra", aggregation = "average") and PCA (with scaling).

For checking of computation results.
packages (“stats”, “FactoMineR”, “dendextend”)

Figures



Datasets & Results

Project ID	Instrument	LC	MS polarity	Sample	№ of samples	№ of groups	Numbers of features			Ref.
							raw data	after UVF	after UVF+RML+RFE	
Our dataset (OD)	LC-IT-TOF	RP	POS	urine	40	3	3441	1887	31	NA
MTBLS 665	LC-QEx	RP	NEG	plasma	180	2	6671	2807	38	(Gardinassi et al., 2018) 10.1016/j.redox.2018.04.011

A. - PCA for raw data of OD
B. - HCA for raw data of OD
C. - PCA for data after UVF+RML+RFE of OD
D. - HCA for data after UVF+RML+RFE of OD
E. - PCA for raw data of MTBLS665
F. - HCA for raw data of MTBLS665
G. - PCA for data after UVF+RML+RFE of MTBLS665
H. - HCA for data after UVF+RML+RFE of MTBLS665

Conclusion

In all datasets (experimental and from open repository) clinical groups were clearly and properly separated by HCA and PCA. Correct pattern recognition was achieved for highly reduced datasets after feature selection based on combination of machine learning training and results of univariate analysis.
This report slightly demonstrate potential opportunities to creation and validation of some useful approaches for marker research in high dimensional data.

This work was funded by the Russian Foundation for Basic Research (RFBR), according to the research project No. 19-33-90071.