Extensive benchmark of machine learning methods for quantitative microbiome data

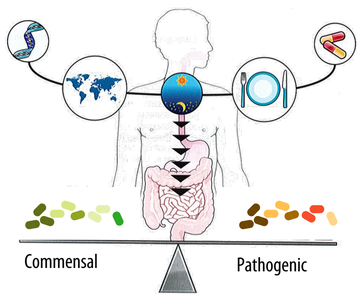
Sébastien Fromentin, Florian Plaza Oñate, Nicolas Maziers, Samar Berreira Ibraim, Guillaume Gautreau, Oscar Gitton-Quent, Manolo Laiola, Soufiane Maski, Raphaëlle Momal, Florence Thirion, et al.

Extensive benchmark of machine learning methods for quantitative microbiome data

Sébastien Fromentin[1], Florian Plaza Oñate[1], Nicolas Maziers[1], Samar Berreira Ibraim[1], Guillaume Gautreau[1], Oscar Gitton-Quent[1], Manolo Laiola[1], Soufiane Maski[1], Raphaëlle Momal[1], Florence Thirion[1], Franck Gautier[1], Nicolas Pons[1] and Magali Berland[1]

Introduction

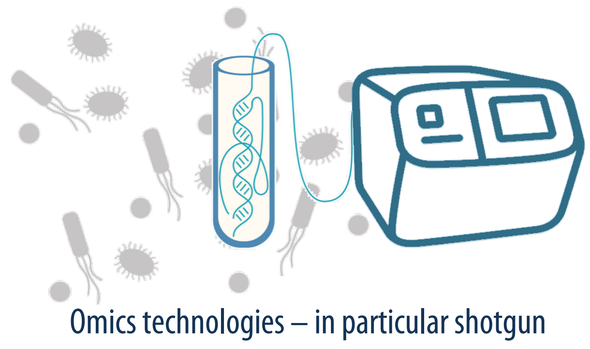
Microbial dysbiosis is associated with many **chronic diseases** and often correlates with **severity**.



Prediction of phenotypic features can help to stratify patients:

- Clinical status (classification)
- Disease severity (regression)

Which machine learning (ML) method should be used?



Omics technologies – in particular shotgun metagenomics – allows a highly precise microbiome profiling

Miaw

Microbiome **a**rtificial intelligence **w**orkflow

The code is based in the R package Caret (topepo.github.io/caret) to train and tune ML models. The Activeeon Proactive solution was used to efficiently distribute the computing load on multiple servers.

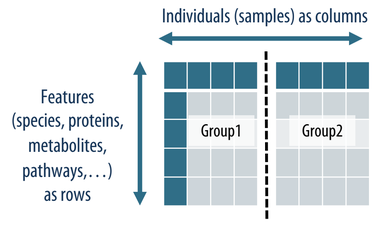
ACTIVEEON SCALE BEYOND LIMITS

LET'S BUILD A ML MODEL WITH CARET

Machine Learning workflow

Input data

Abundance tables and metadata



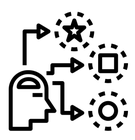
Standard preprocessings

- Downsizing
- Normalization by gene length

Specific preprocessings

- Near zero variance
- Linear combos / correlated predictors

Validation scheme



N = 100

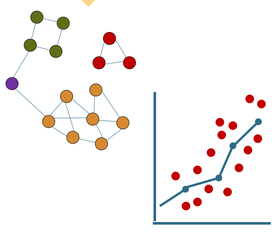
Split Data

Training data – 75%: For each learning approach: Parameters tuning in 10 CV

Test data – 25%: Model performance evaluation

- Tuneable parameters

Benchmarked methods

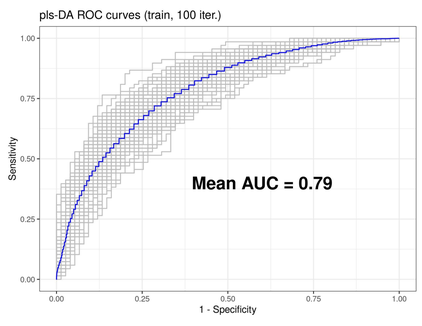


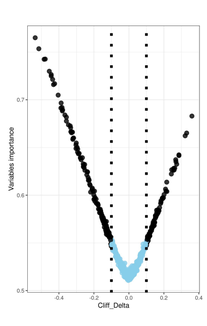
PLS | GLMnet | SVM | RF | LASSO | MLP | KNN | SDA | …

- **69** ML classification methods
- **53** ML regression methods

Outputs

- Indicators of predictive performance
- Variable importance



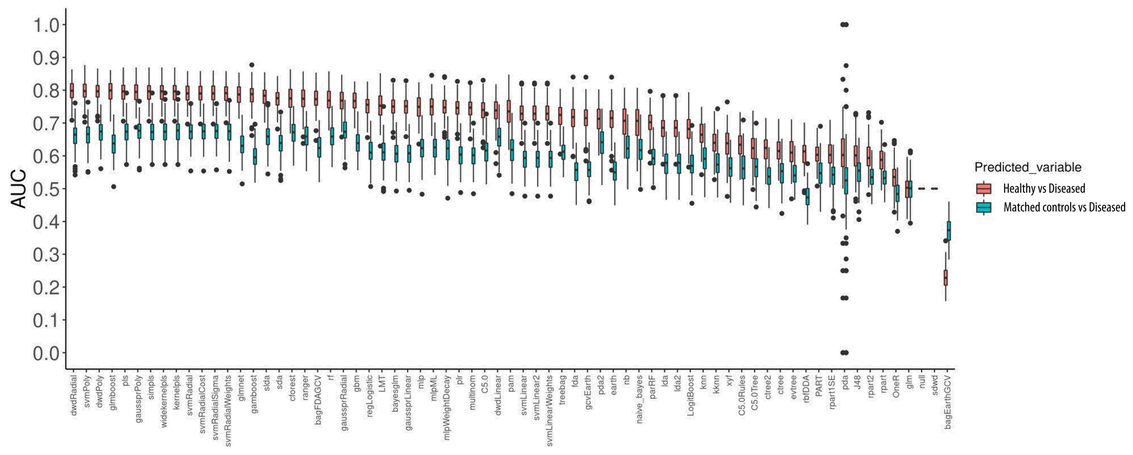


Results

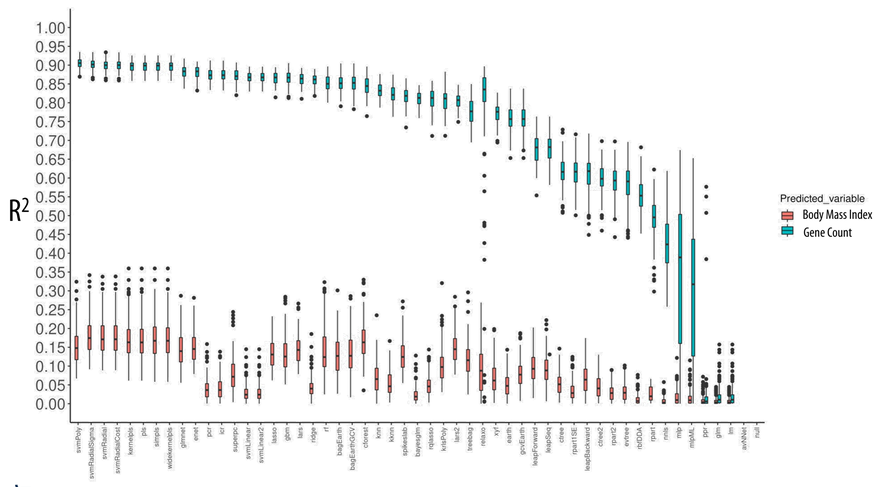
Study Case	Liver cirrhosis	Coronary diseases	Coronary diseases (deconfunded)	Coronary diseases	Vegan	Anorexia	Chronic kidney disease
Sample size	237	578	744	99	150	147	240
Omic type	Metagenomics	Metagenomics	Metagenomics	Metaproteomics	Metagenomics	Metagenomics	Metagenomics
# features	436	729	729	755	619	580	754
Best mean AUC	0.95	0.80	0.68	0.82	0.93	0.89	0.63

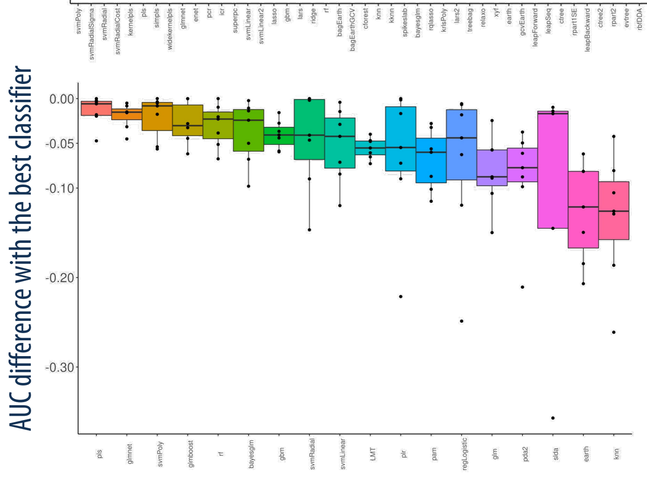
Example of results (coronary diseases - metagenomics):

- Many classification models provide similar performances both for an easy prediction task (Healthy vs Diseased) and a difficult prediction task (individuals matched on the principal confounders [Age, BMI, Diabetes, Gender] vs Diseased)



- Regression model performances are more heterogeneous than those of classification models





- Performances on 7 datasets
- PLS, SVM (poly) and GLMnet were the best predictive approaches to classify microbiome data
- Some methods lack of stability

Conclusion

- **Miaw** allows an easy **benchmark** across methods and datasets
- **New methods** can be **easily added** with the caret wrapping functions
- Depending on the dataset, some methods perform better or worse; so far PLS, GLMnet and SVM performed generally well
- Results will be extended to other public datasets (database under construction)

INRAE science for people, life & earth

Centre de recherche de Jouy-en-Josas
Unité MetaGenoPolis
www.mgps.eu

[1] MetaGenoPolis, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France

www.linkedin.com/company/metagenopolis/

twitter.com/MgpsLab

Contact: magali.berland@inrae.fr