Detection of unknown genetically modified organisms (GMO) by statistical analysis of high throughput sequencing data

Julie Hurel, Mathieu Roland, Sophie Schbath, Stéphanie Bougeard, Mauro Petrillo, Fabrice Touzain

Detection of unknown genetically modified organisms (GMO) by statistical analysis of high-throughput sequencing data

Julie Hurel[1], Mathieu Roland[2], Sophie Schbath[3], Stéphanie Bougeard[1], Mauro Petrillo[4], Fabrice Touzain[1]

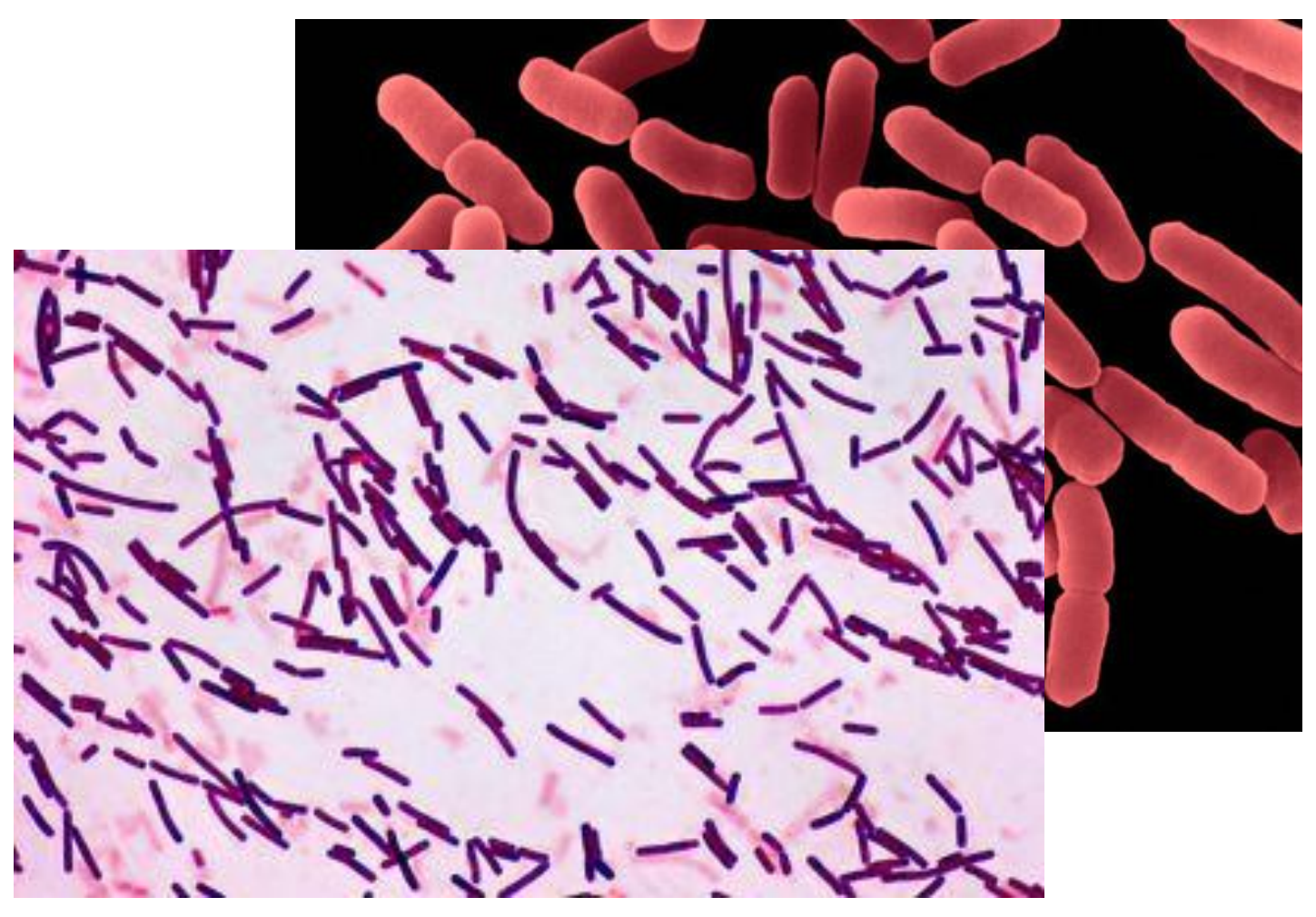
[1] ANSES, Agence Nationale de Sécurité Sanitaire, Laboratoire de Ploufragan, 22440 Ploufragan, France
[2] ANSES, Agence Nationale de Sécurité Sanitaire, Laboratoire de la santé des végétaux, 49000 Angers, France
[3] INRA, Institut National de Recherche Agronomique, Centre de Jouy-en-Josas, 78352 Jouy-en-Josas, France
[4] JRC, Joint Research Centre Ispra, 21027 Ispra, Italy

anses
French agency for food, environmental and occupational health & safety
Investigate, evaluate, protect

UNIVERSITE BRETAGNE LOIRE

INTRODUCTION

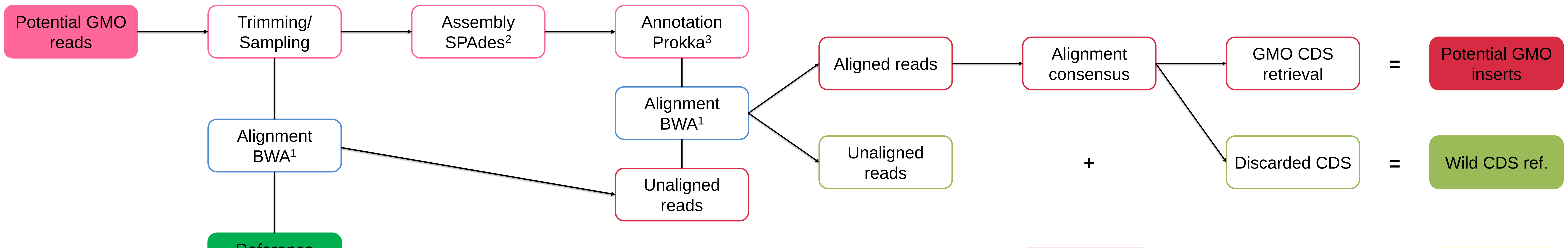
European legislation requires the reporting of **GMOs in food products** when their proportion is **higher than 0.9%**. In recent years, not described GMOs have been produced whose sequence is unknown making them not detectable by current PCR approaches. To date, no detection method have been described for the detection of unknown GMOs. The method was developed using **two sets of raw reads of the bacteria *Bacillus Subtilis*** *:* the first related to a genetically modified genome and the second to a wild bacteria.



Aim : Highlight the sequences of inserts representing a tiny part of the genome if the unknown genome submitted is indeed a GMO.

MATERIAL AND METHODS

Cleaning pipeline



Blast Pangenome

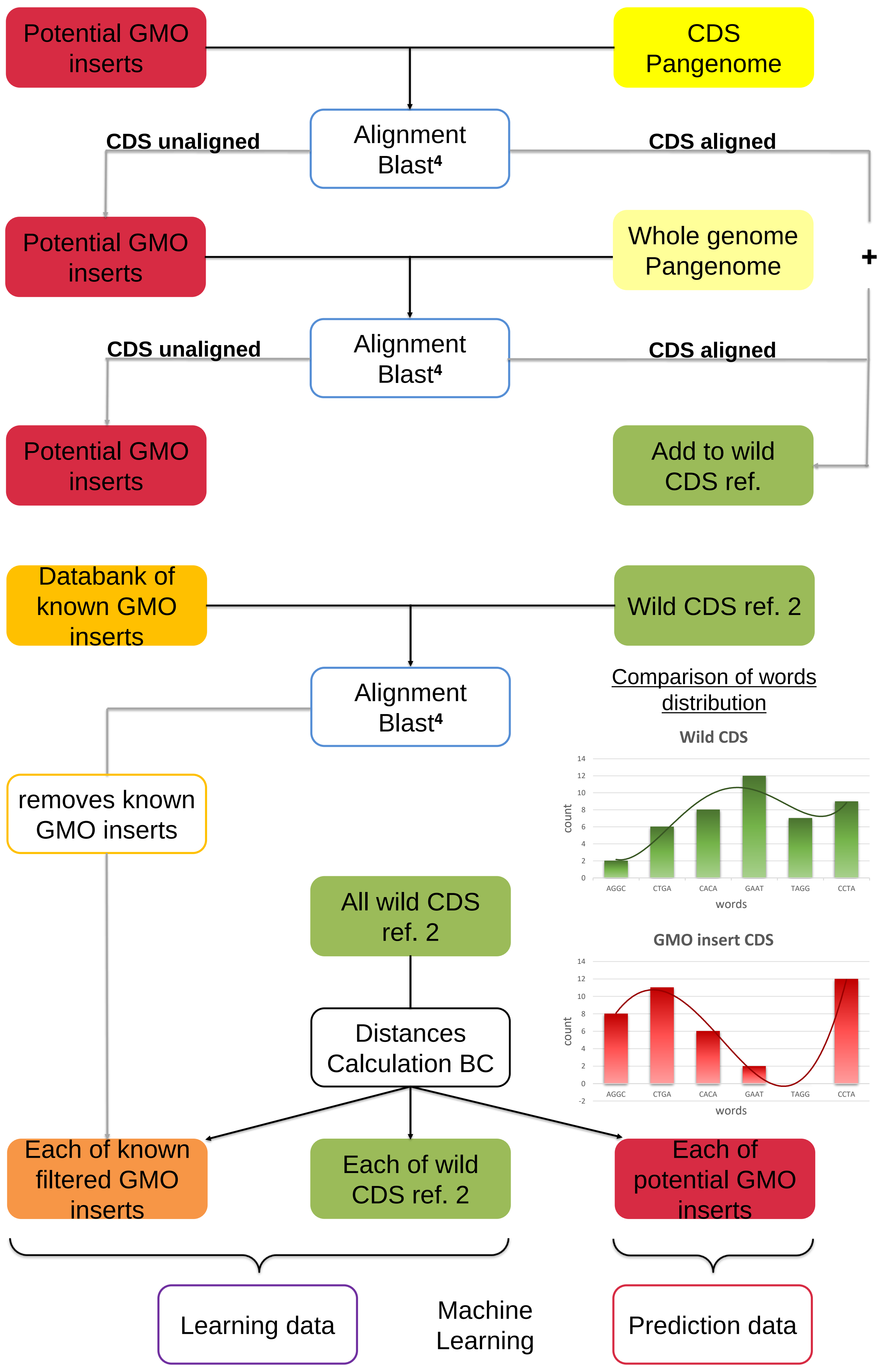
A pangenome groups most of the different genes present in the same species. Two **Blastn**[4] are performed on the pangenome of *Bacillus Subtilis* one on the coding sequences (CDS) and one on the whole genome. This pangenome involves **35 different strains of *Bacillus Subtilis***. The sequences selected by Blast alignment have an **identity percentage greater than 98%** and an alignment **coverage percentage greater than 80%**.

Distance Calculation

The insert sequences of an unknown GMO have a different vocabulary from that of the wild genome from which it is derived. **A genome has its own vocabulary, consisting of words**. Each word is a set of nucleotides with predefined length such as "ATGCCT". We search **over-represented or rare words using R'mes**[5] in the wild genome to define specificities of its vocabulary and highlight the vocabulary of an insert that shows different word proportions. This difference is evaluated through a **distance calculation** between the words vector of all CDS of the wild genome and the words vector of each candidate GMO insert CDS. The distance between a CDS of the wild genome and all the CDS of the wild genome - almost similar - will be close to 0 while the distance between all CDS of the wild genome and an insert CDS will be higher. We use **Bray-Curtis distance**[6] **(BC)** after comparison with alternative distances and employ **two types of calculation** : one based on **proportions** of short words, the other based on **frequencies** of long words over-expressed in wild genome.

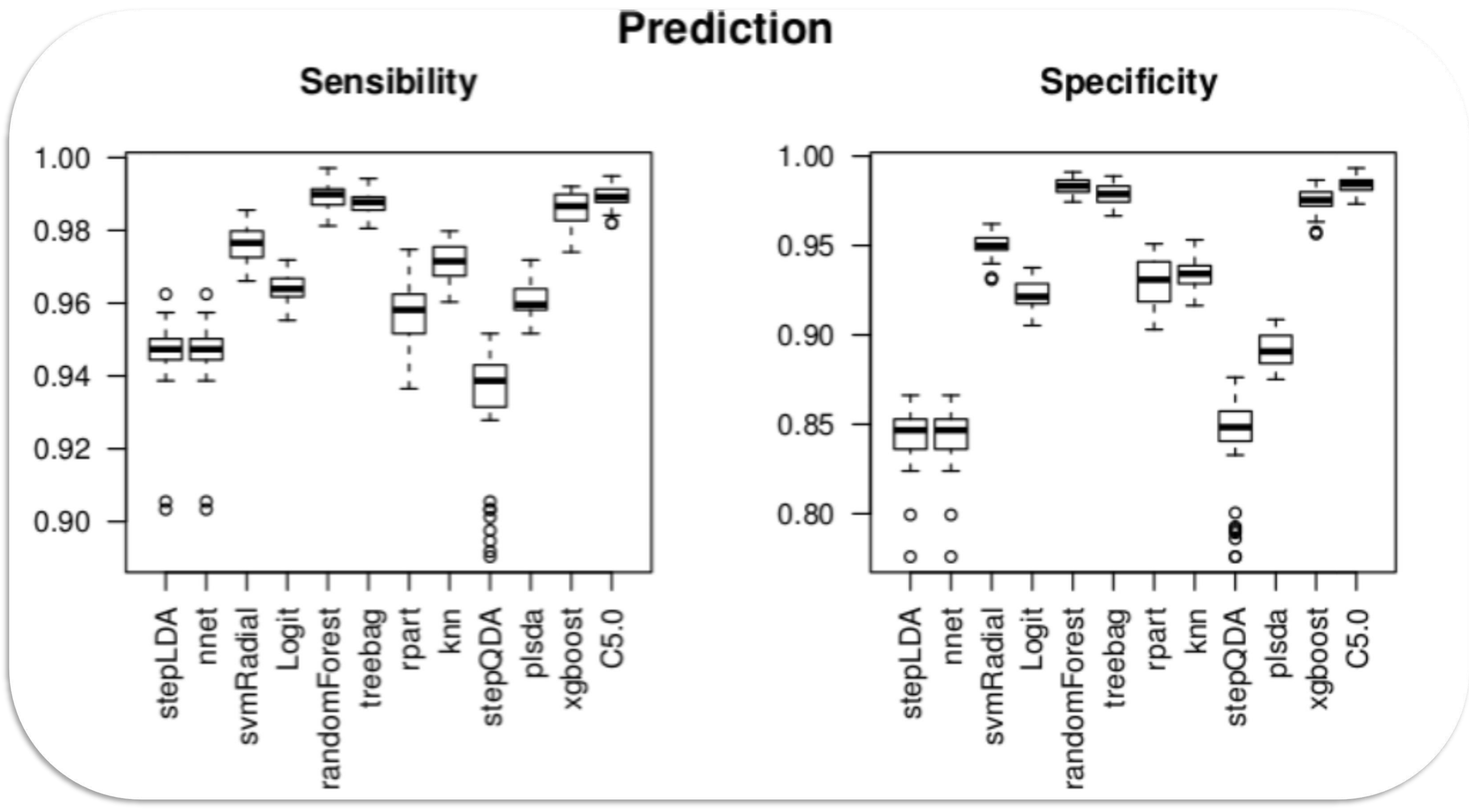
Machine Learning

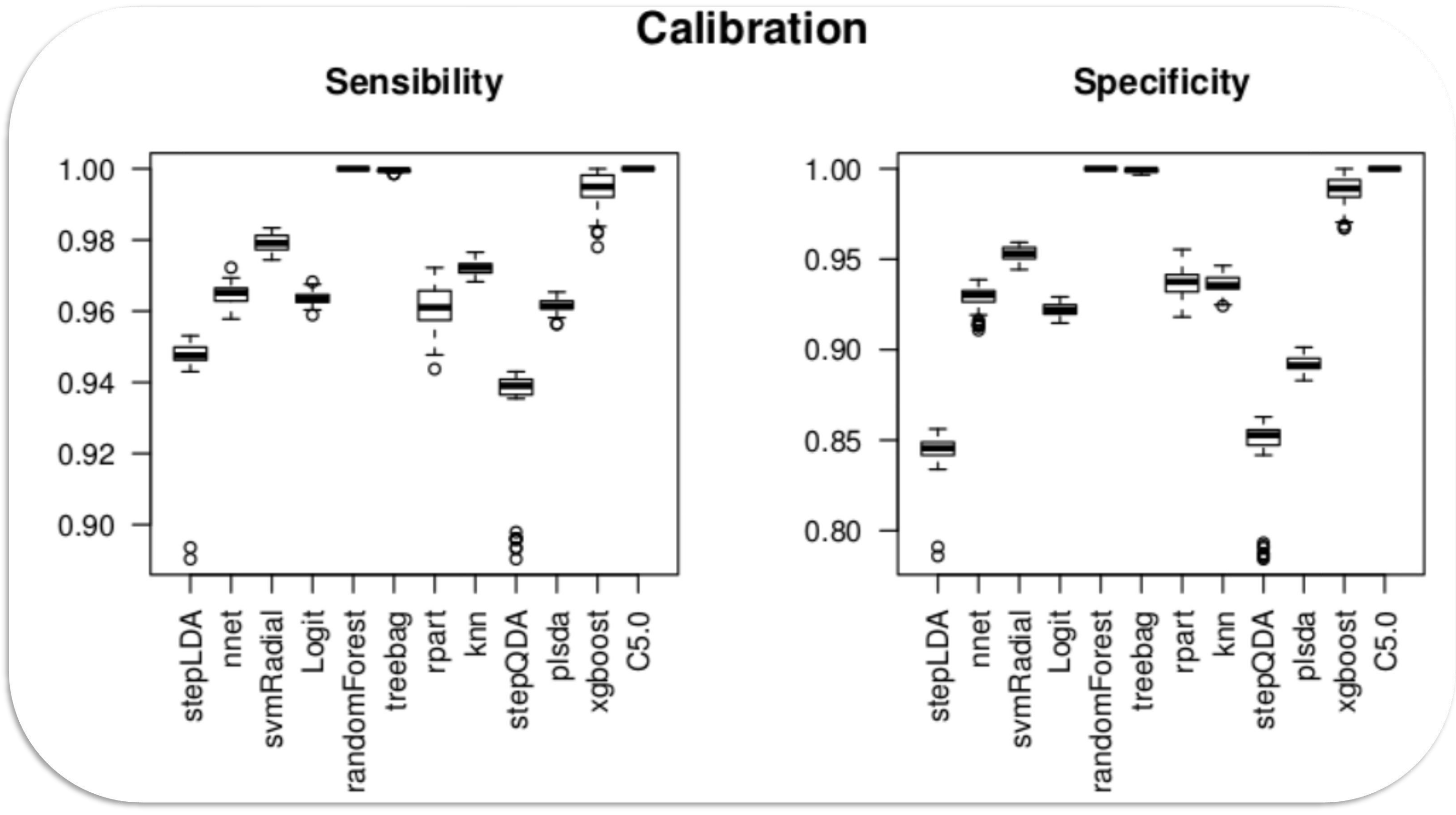
A machine learning step is implemented to **discriminate CDS of GMO inserts**. The **learning datasets** correspond to the wild CDS found in sample (Wild CDS ref. 2) and to a databank of known GMOs (known filtered GMO inserts). This **databank contains only filtered sequences** that don't belong to the species of the wild genome. The sequences of candidates GMO inserts **not matching** pangenome after the two Blasts are used as **prediction data** (Potential GMO inserts). The **variables** retained for this learning are the Bray-Curtis distance, the CDS length, the average of the R'mes words exceptionality scores, the over-represented words density per nucleotide (sum of the counts of over-represented words divided by the length of the CDS), the GC percentage and the codon usage.



RESULTS

12 machine learning methods were tested using the **Caret package**[7]. The parameters of each method were optimized by **cross-validation**. The performance of these methods was then compared on the basis of a **second cross-validation** to obtain the most efficient methods in terms of predicting sequences of GMO inserts (highest sensitivity and specificity). We retained **Random Forest** method for Machine Learning. It preserves most of the GMO inserts. Based on searched properties of the wild genome, **we cannot exclude to find in results genes coming from recent horizontal gene transfer instead of GMO insert**.





Conclusion : On a GM *Bacillus Subtilis*, most of the inserts are found. On a wild Bacillus Subtilis, we obtained only one false positive.

[1] Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60, 2009.
[2] Nurk, Bankevich and al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5): 455-477, 2012.
[3] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14):2068-9, 2014.
[4] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool J. Mol. Biol. 215:403-410, 1990.
[5] Schbath S. and Hoebeke M. R'MES: a tool to find motifs with a significantly unexpected frequency in biological sequences. In Advances in genomic sequence analysis and pattern discovery. Science, Engineering, and Biology Informatics, vol. 7, World Scientific. 2011.
[6] Ricotta C. and Podani J. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. Ecol. Complex, 31:201-205, 2017.
[7] Max Kuhn and al. Caret: Classification and Regression Training. 2017. R package version 6.0-78. https://CRAN.R-project.org/package=caret.