MethanoGram

Michał Burdukiewicz, Przemysław Gagat, Michał Gaworski, Sławomir Jabłoński, Paweł Mackiewicz, Marcin Łukaszewicz

Contents

Introduction

MethanoGram is a predictor of culture conditions of methanogenes. Using random forests trained on n-gram encoded 16 rRNA and mcrA sequences, MethanoGram is able to estimate:

- growth rate,
- growth doubling time [h],
- optimal growth temperature,
- optimal growth pH,
- optimal growth NaCl.

Here we document the process of tuning and evaluation of set of classifiers constituting MethanoGram.

Tuning and evaluation of MethanoGram

Datasets

To train MethanoGram we used n-grams (subsequences of length n) extracted from mcrA and 16 rRNA sequences found in the PhyMet^2 database. We chose only sequences for which we were able to identify all culture conditions described in the database (both optimal and non-optimal). Thus, we chose only records that have known 16 rRNA sequence, mcrA sequence and all culture conditions (growth rate, growth doubling time, optimal growth temperature, growth temperature, optimal growth pH, growth pH, optimal growth NaCl, growth NaCl).

We considered two different sets of 16 rRNA sequences and three different sets of mcrA sequences. We removed all sequences containing atypical or unknown nucleotides (b, d, k, m, n, r, s, v, w, y). After purification steps described above we ended with 60 methanogenes (Fig. 1).

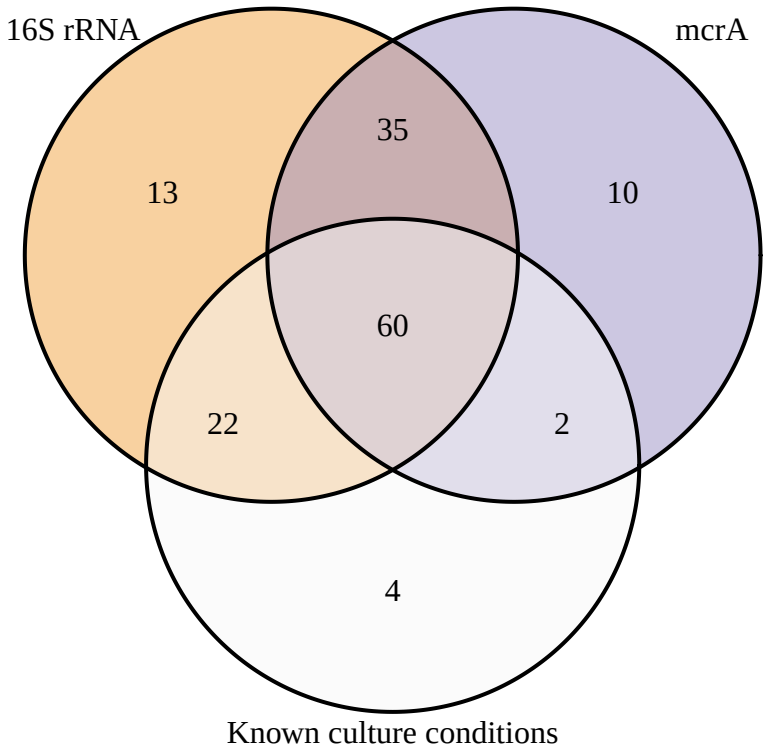


Figure 1: The Venn diagram of methanogens species in the analysis.

Tuning procedure

We chose the random forests implemented in the **ranger** R package for estimation of culture conditions, because of its speed and high accuracy. To find optimal values of hyperparameters, we performed a nested cross-validation of random forest classifiers. The inner loop was a 5-fold cross-validation and the outer loop was more demanding 3-fold cross-validation.

We have optimized three hyperparameters related to the random forest algorithm: a number of variables to possibly split at in each node, a number of trees in the forest and a minimal node size. In the tuning procedure we also incorporated different data sources, levels of feature selection and n-gram lengths.

n-gram length

We considered continuous 2-, 3-, 4- and 5-grams. The number of possible n-grams for a nucleotide sequence is equal to 4^n, so the number of feature ranges between 16 (for 2-grams) to 1024 (for 5-grams). Since further increases in the n-gram size were not providing the algorithm with satisfying decrease in the error, we did not considered longer n-grams.

n-gram source

The algorithm was trained on n-grams extracted from:

- 16 rRNA,
- mcrA,
- 16 rRNA and mcrA.

In the third case, n-grams were annotated by their source. For example, in the case of bigrams, GA_RNA and GA_mcrA (bigram GA coming from RNA and mcrA) were treated as two different features.

Feature selection

To select the most informative n-grams, we used Pearson's correlation between the feature and the target as implemented in the Rfast package. We retained a fraction of features instead of an absolute number of features to keep the feature selection consistent between datasets with varying number of features. We considered following fractions of features: 0.25 and 0.5. In all cases, the strictest feature selection (0.25) proved to create the most efficient classifiers.

We also considered a predictor without any feature selection, but it consistently had the worst performance (results not shown).

Number of variables to possibly split at in each node (mtry)

In addition to the standard number of variables to possibly split at in each node for regression tasks ($\frac{1}{3}$ of all considered features), we have also examined $\frac{1}{2}$ and $\frac{1}{4}$. In most cases, $\frac{1}{4}$ was the most optimal.

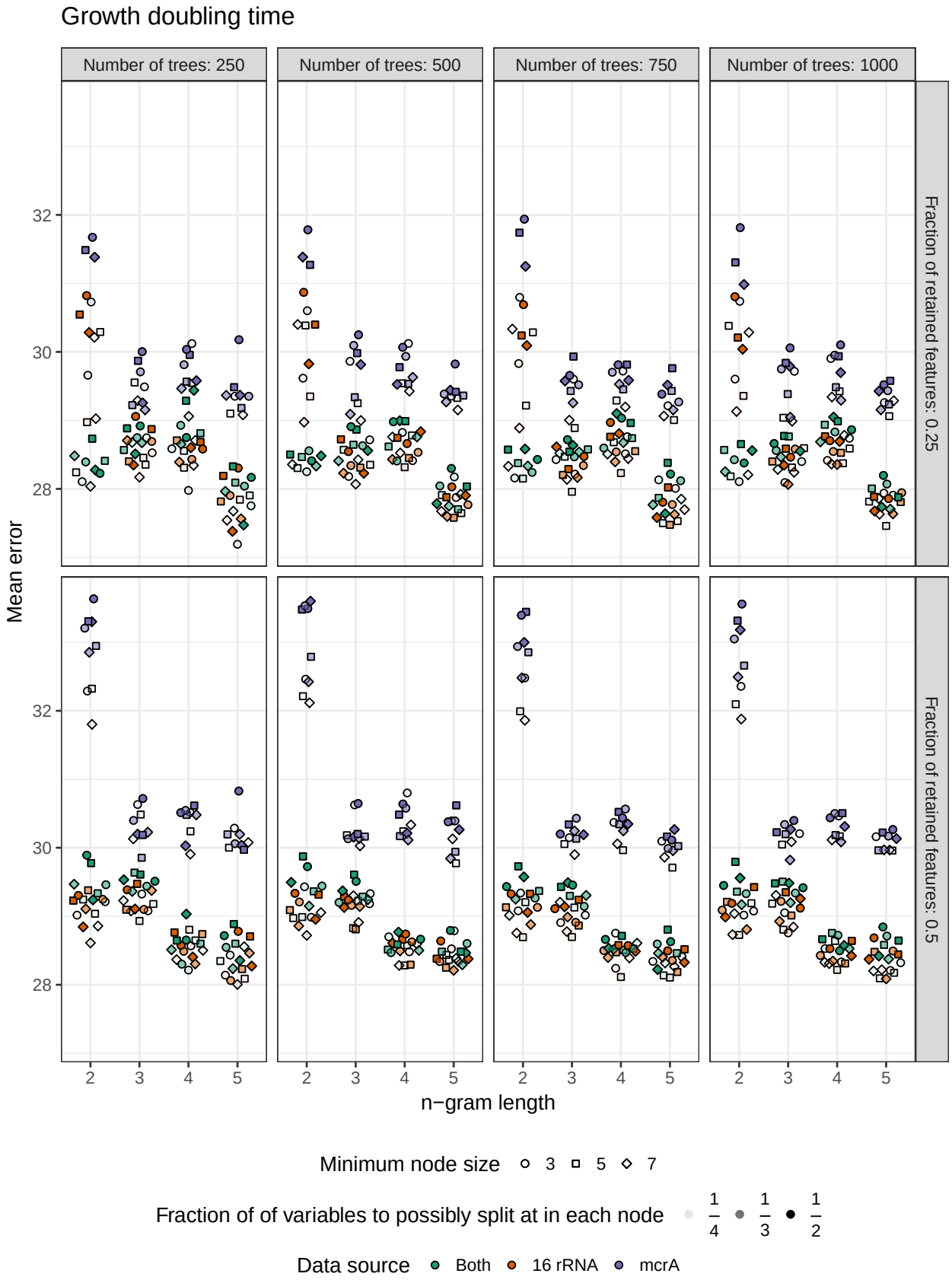
Minimal node size

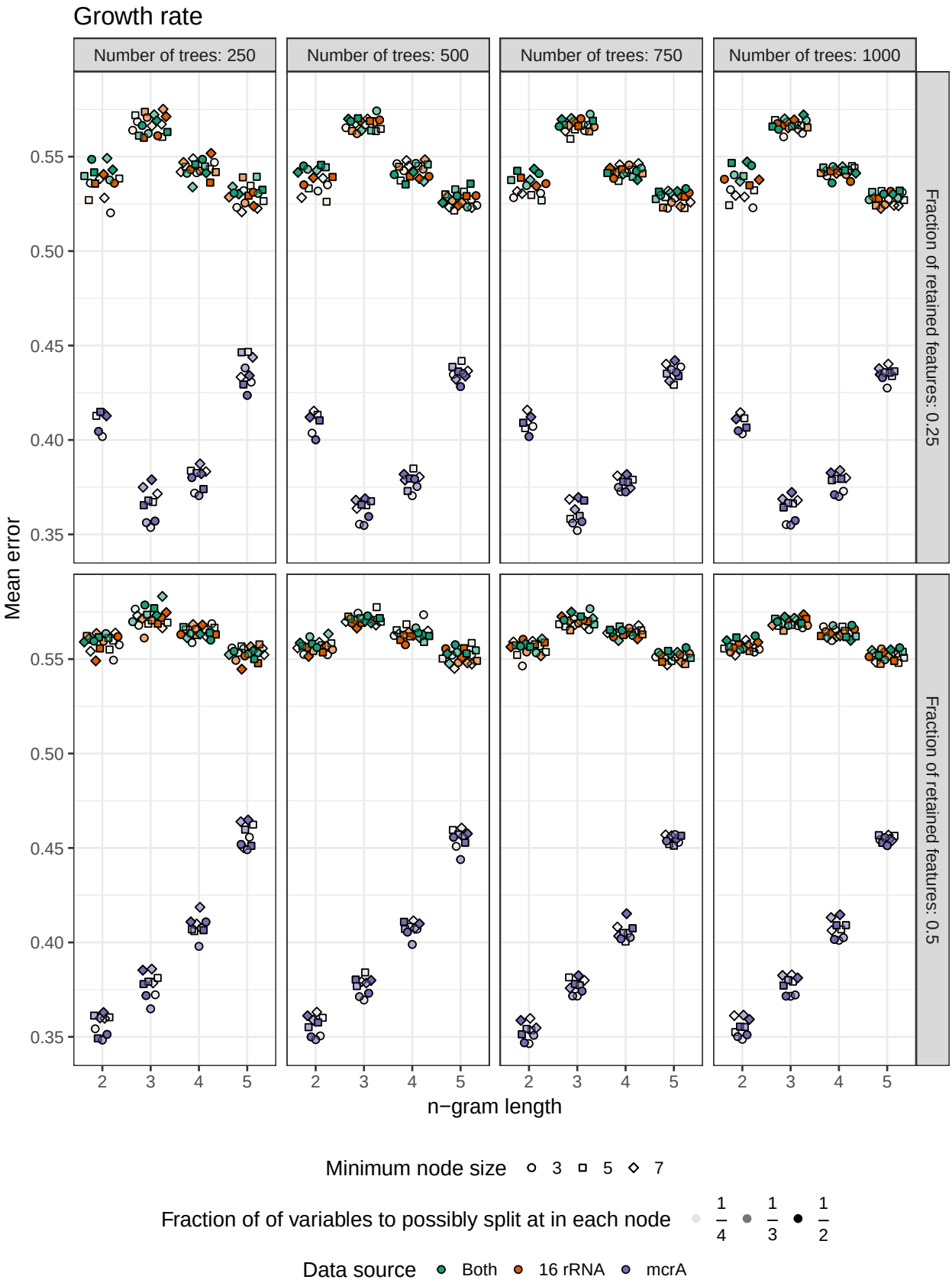
Aside from the optimal number of variables to possibly split advised by literature (5), we have also considered 3 and 7 variables. There were no visible patterns in the optimal value of the minimum node size, aside from the fact that the value advised by the literature was never producing the best-performing predictors.

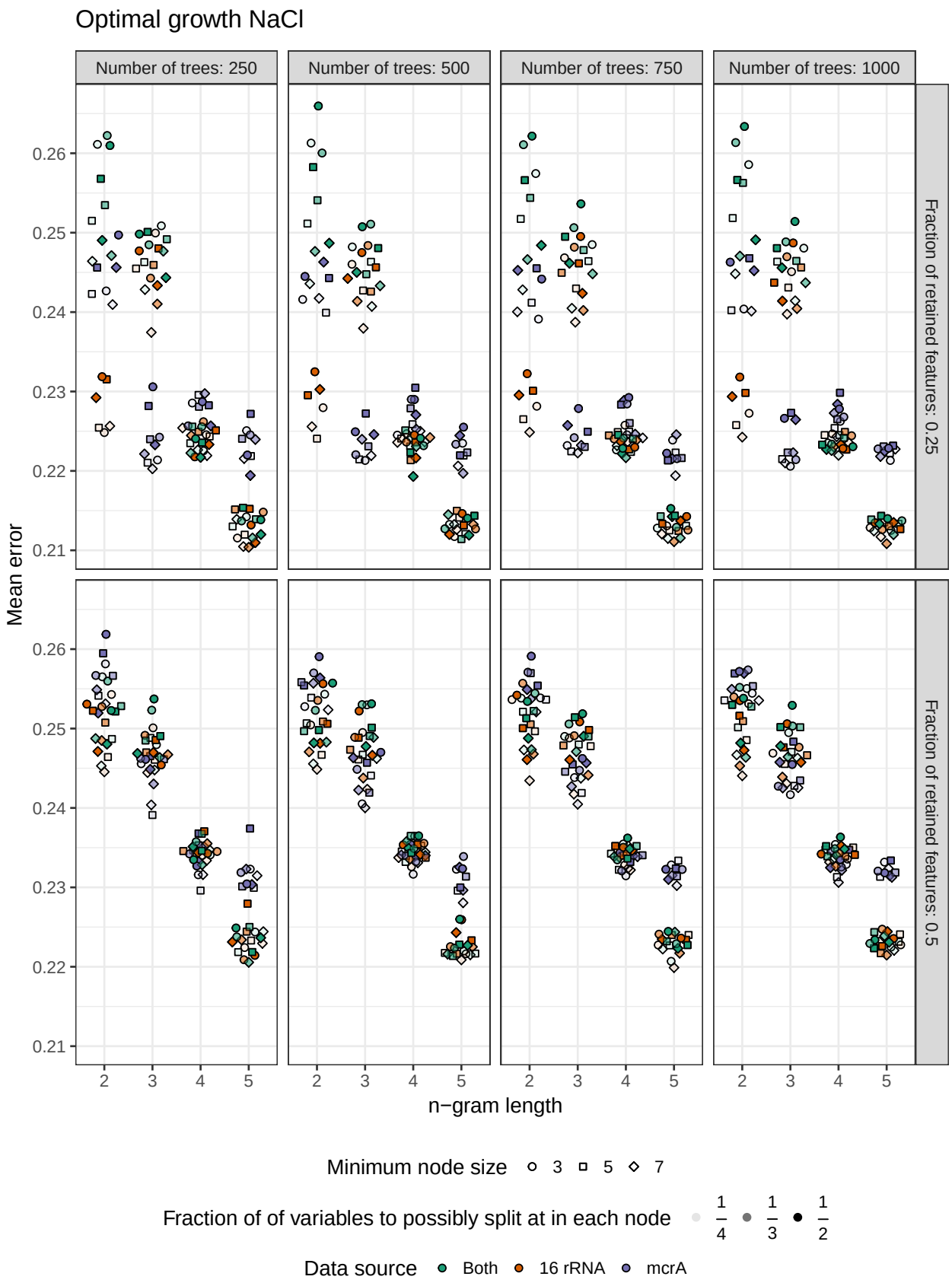
Results of tuning

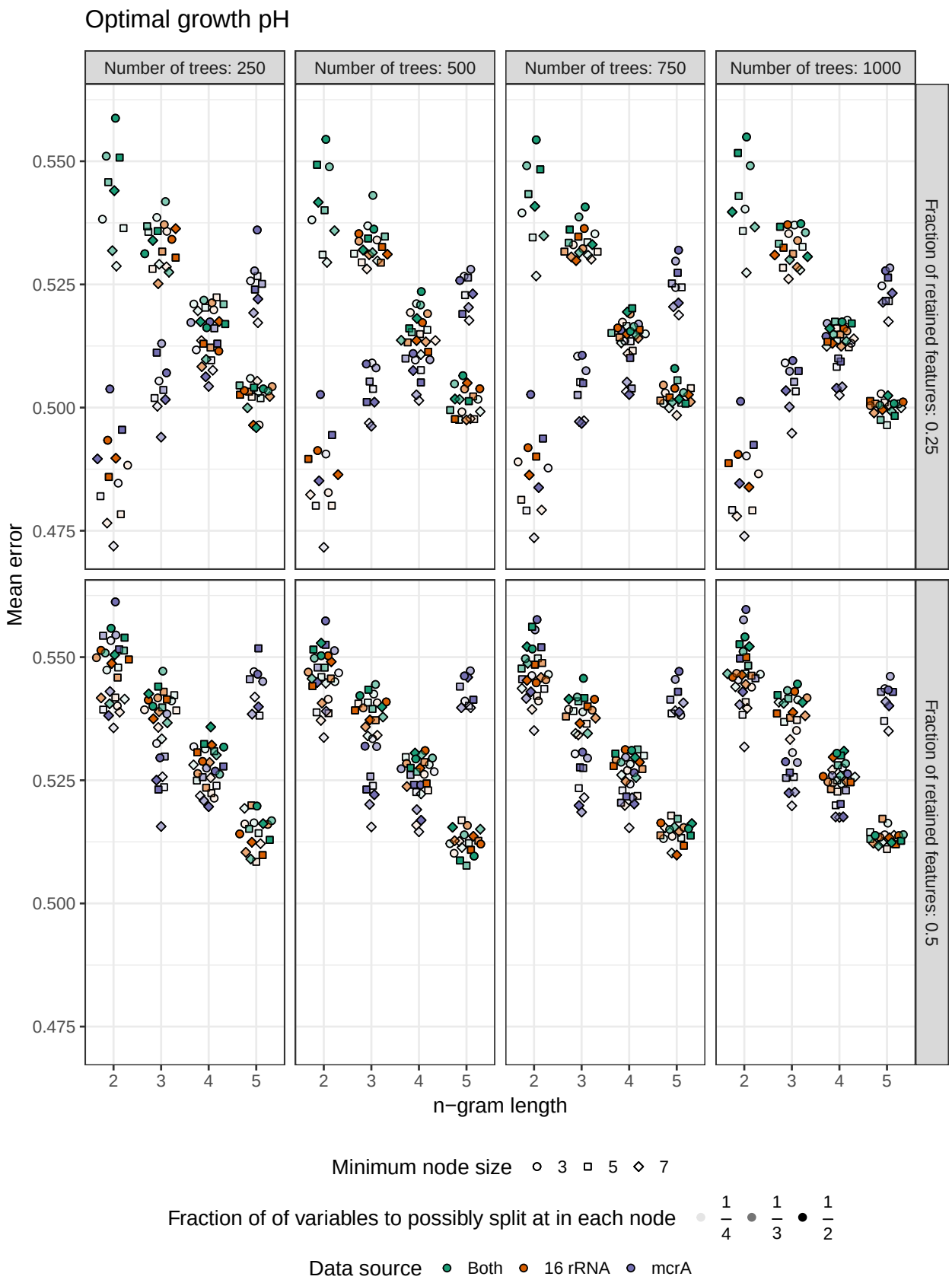
Below are included bee swarm plots for all parameters tuned for MethanoGram. The shape of points represents minimal node size and transparency distinguish between different number of variables to possibly split at in each node (mtry).

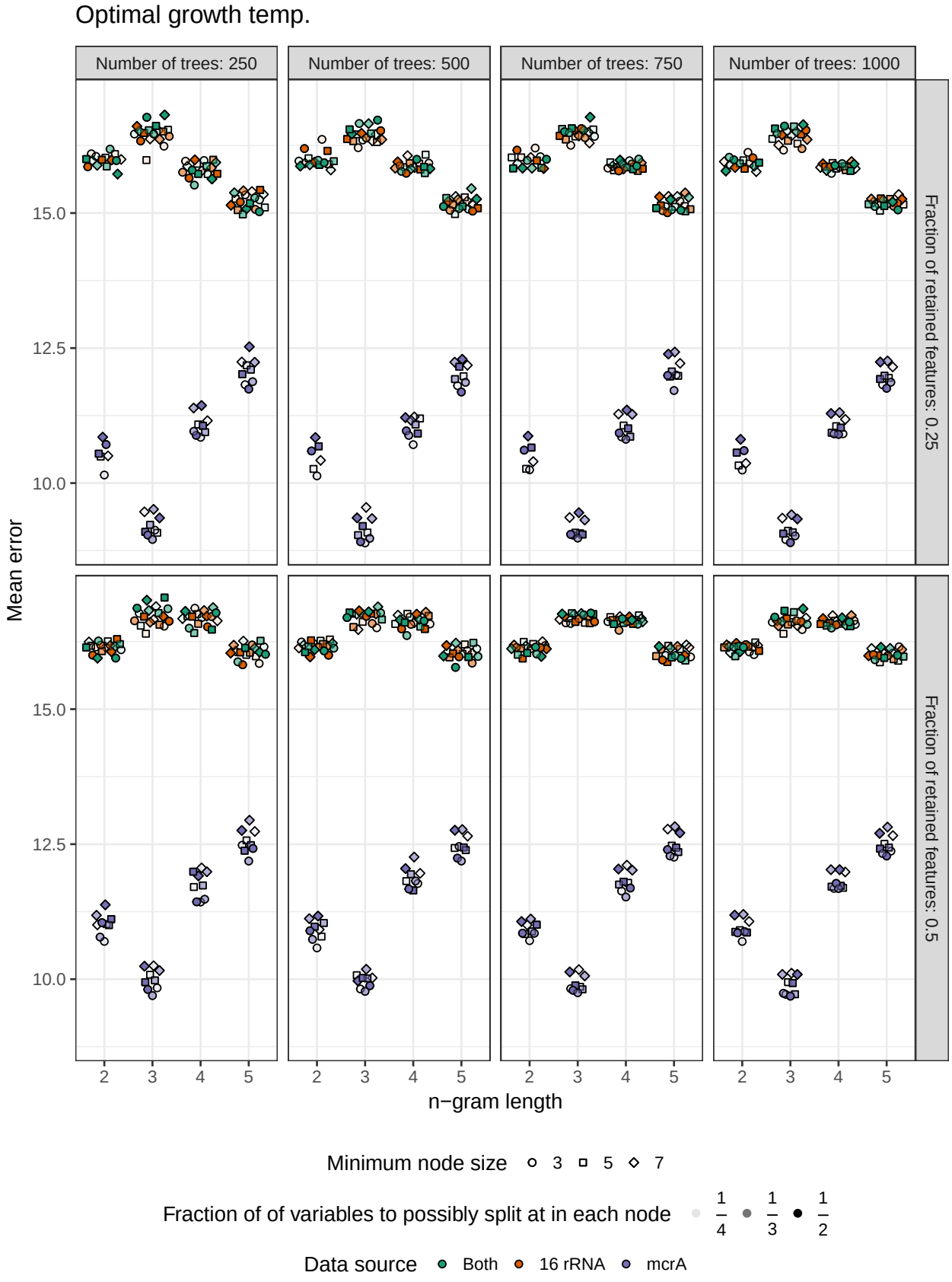
The best combinations of parameters for each condition can be also accessed in the table 1.


Growth doubling time
Number of trees: 250
Number of trees: 500
Number of trees: 750
Number of trees: 1000
Fraction of retained features: 0.25
Fraction of retained features: 0.5
Mean error
32
30
28
2
3
4
5
n−gram length
Minimum node size 3 5 7
Fraction of of variables to possibly split at in each node 1/4 1/3 1/2
Data source Both 16 rRNA mcrA


Growth rate
Number of trees: 250
Number of trees: 500
Number of trees: 750
Number of trees: 1000
Fraction of retained features: 0.25
Fraction of retained features: 0.5
Mean error
0.55
0.50
0.45
0.40
0.35
2
3
4
5
n−gram length
Minimum node size 3 5 7
Fraction of of variables to possibly split at in each node 1/4 1/3 1/2
Data source Both 16 rRNA mcrA


Optimal growth NaCl
Number of trees: 250
Number of trees: 500
Number of trees: 750
Number of trees: 1000
Fraction of retained features: 0.25
Fraction of retained features: 0.5
Mean error
0.26
0.25
0.24
0.23
0.22
0.21
2
3
4
5
n−gram length
Minimum node size 3 5 7
Fraction of of variables to possibly split at in each node 1/4 1/3 1/2
Data source Both 16 rRNA mcrA


Optimal growth pH
Number of trees: 250
Number of trees: 500
Number of trees: 750
Number of trees: 1000
Fraction of retained features: 0.25
Fraction of retained features: 0.5
Mean error
0.550
0.525
0.500
0.475
2
3
4
5
n−gram length
Minimum node size 3 5 7
Fraction of of variables to possibly split at in each node 1/4 1/3 1/2
Data source Both 16 rRNA mcrA


Optimal growth temp.
Number of trees: 250
Number of trees: 500
Number of trees: 750
Number of trees: 1000
Fraction of retained features: 0.25
Fraction of retained features: 0.5
Mean error
15.0
12.5
10.0
2
3
4
5
n−gram length
Minimum node size 3 5 7
Fraction of of variables to possibly split at in each node 1/4 1/3 1/2
Data source Both 16 rRNA mcrA

Condition	n-gram length	Source	Feature selection	mtry	Number of trees	Minimal node size	Mean error
Growth rate	3	mcrA	0.25	0.25	750	3	0.35
Growth doubling time [h]	5	16 rRNA	0.25	0.25	250	3	27.19
Optimal growth temp.	3	mcrA	0.25	0.25	500	3	8.89
Optimal growth pH	2	mcrA	0.25	0.25	500	7	0.47
Optimal growth NaCl	5	16 rRNA	0.25	0.33	250	7	0.21

Table 1: Results of nested cross-validation. mtry denotes fraction of variables to possibly split at in each node.