

MethanoGram

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

Contents

Introduction	1
Tuning and evaluation of MethanoGram	1
Datasets	1
Tuned parameters	1
Results of tuning	3

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) extrated from mcrA and 16 rRNA sequences found in the PhyMet² 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).

Tuned parameters

We chose the random forests implemented in the **ranger** R package for estimation of culture conditions, because of its speed and high accuracy. We have optimized three parameters: 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 levels of feature selection and n-gram lengths.

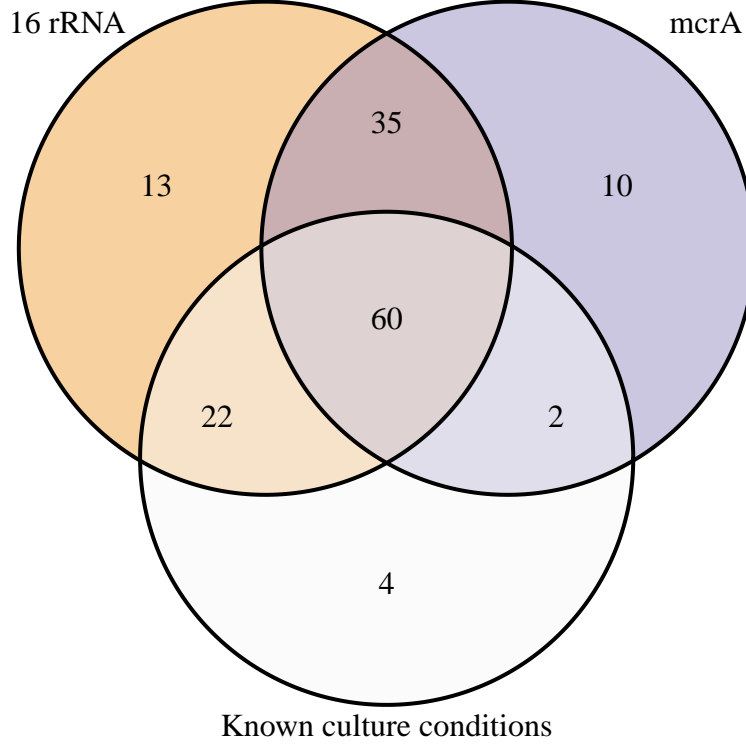


Figure 1: The Venn diagram of methanogens species in the analysis.

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 consider 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.

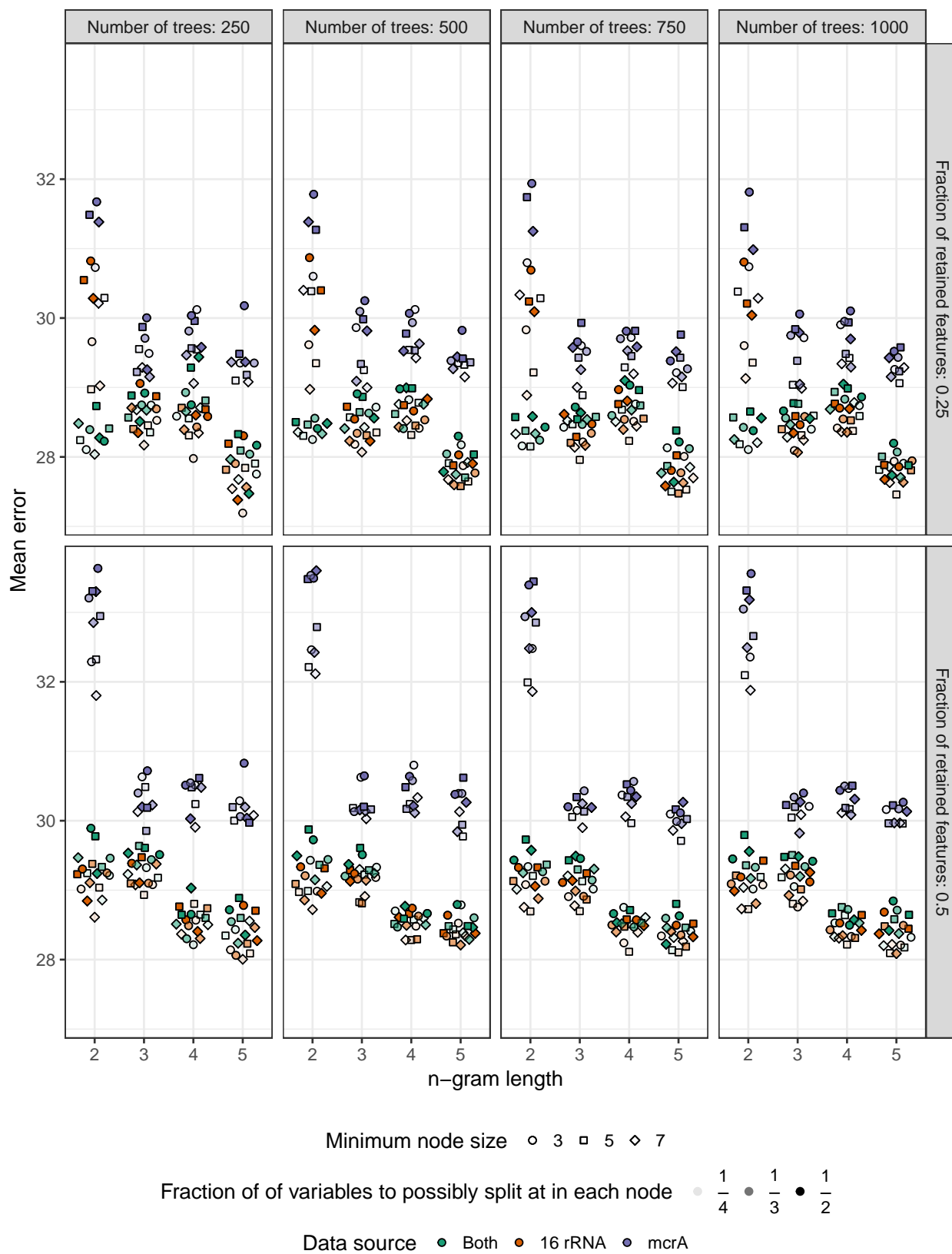
Minimal node size

Aside from the optimum number of variables to possibly split (5), we have also considered 3 and 7 variables. There were no visible patterns in the optimal value of the minimum node size.

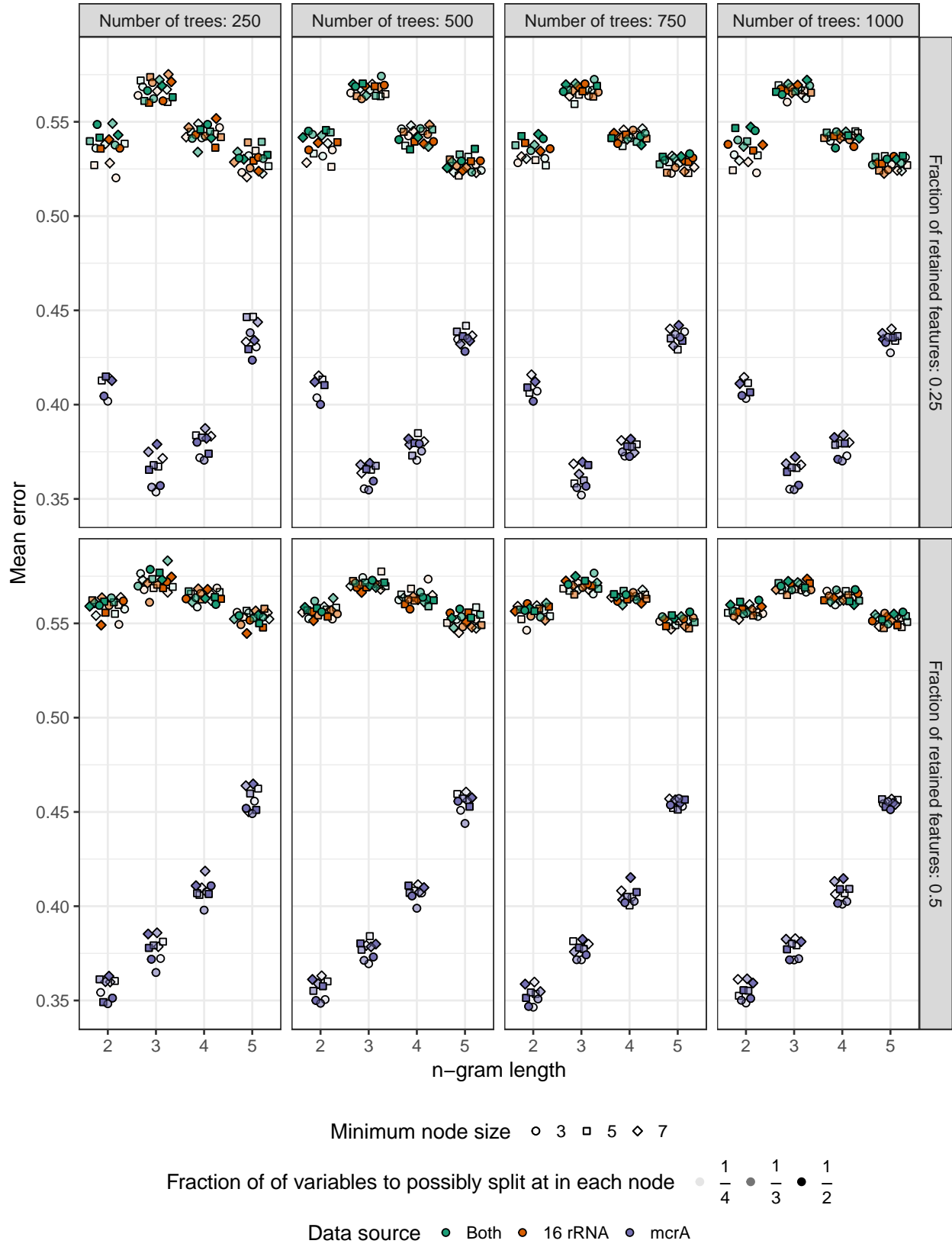
Results of tuning

Below are included beeswarm 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).

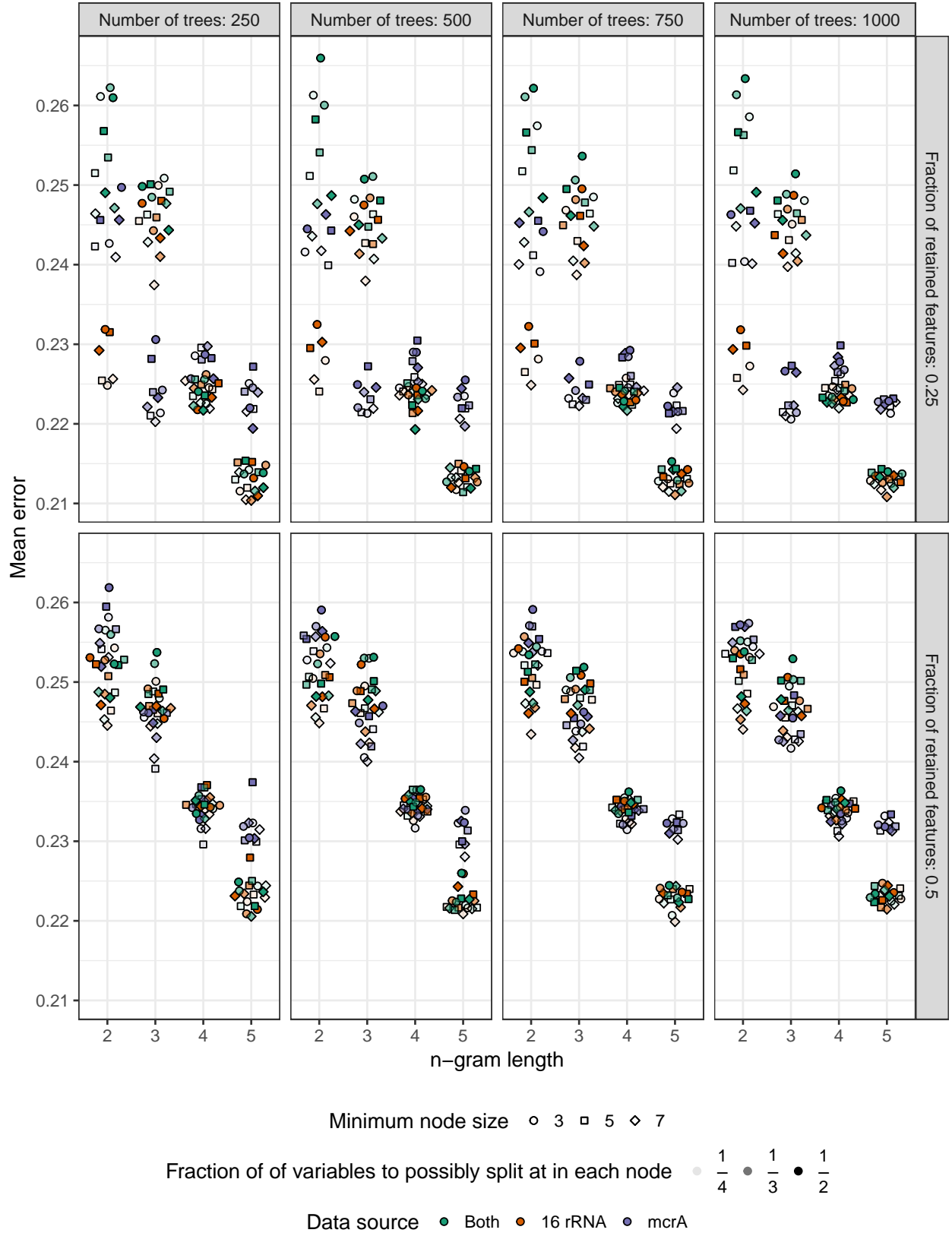
Growth doubling time [h]



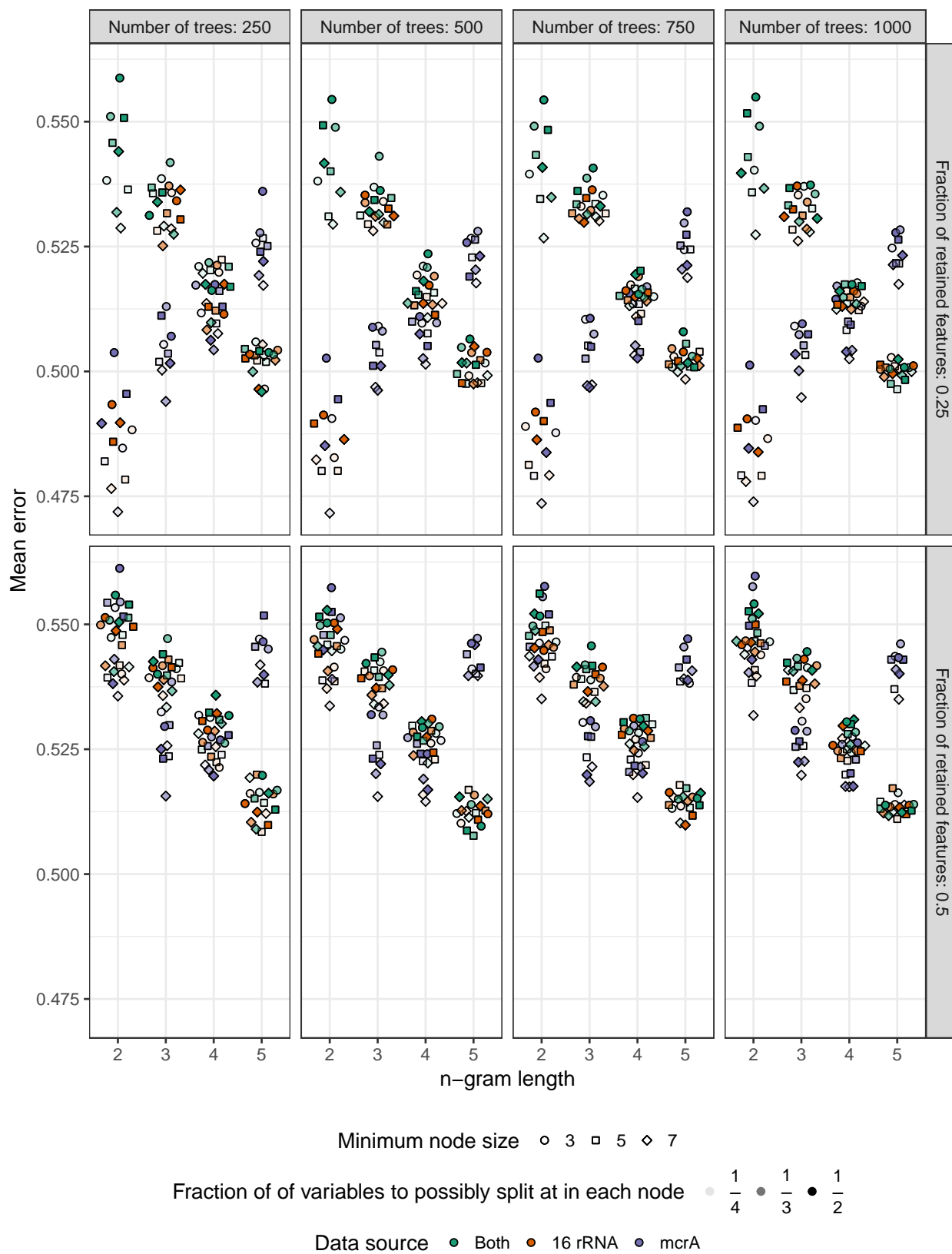
Growth rate



Optimal growth NaCl



Optimal growth pH



Optimal growth temp.

