MethanoGram

Michal Burdukiewicz

Contents

Introduction **1**

Tuning and evaluation of MethanoGram **1**
Datasets 1
Random forest tuning 1

Introduction

MethanoGram is a predictor of culturing 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 evalutation 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 PhyMet^2 database. We chose only sequences for which we were able to identify all culturing 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 culturing 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).

Random forest tuning

We chose the random forests implementated in the **ranger** R package for estimation of culturing 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 lenghts.

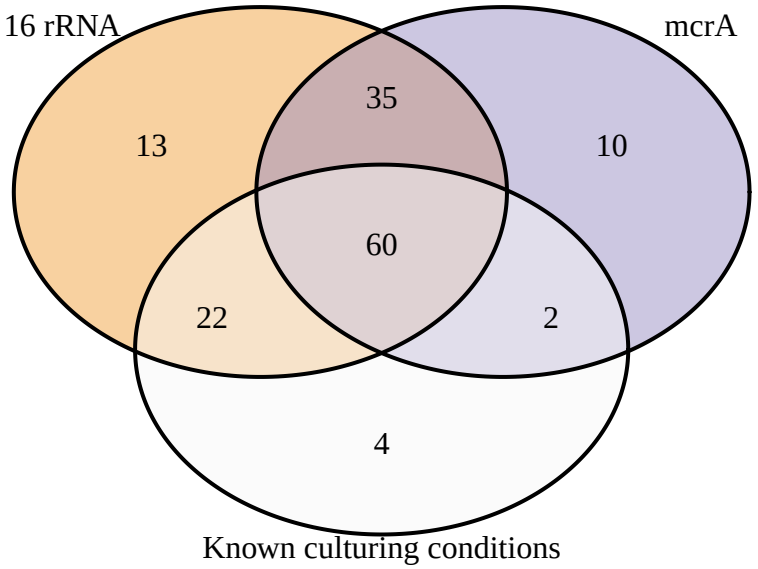


Figure 1: The Venn diagram of methanogens species in the analysis.

n-gram length

We considered continous 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 descrese 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

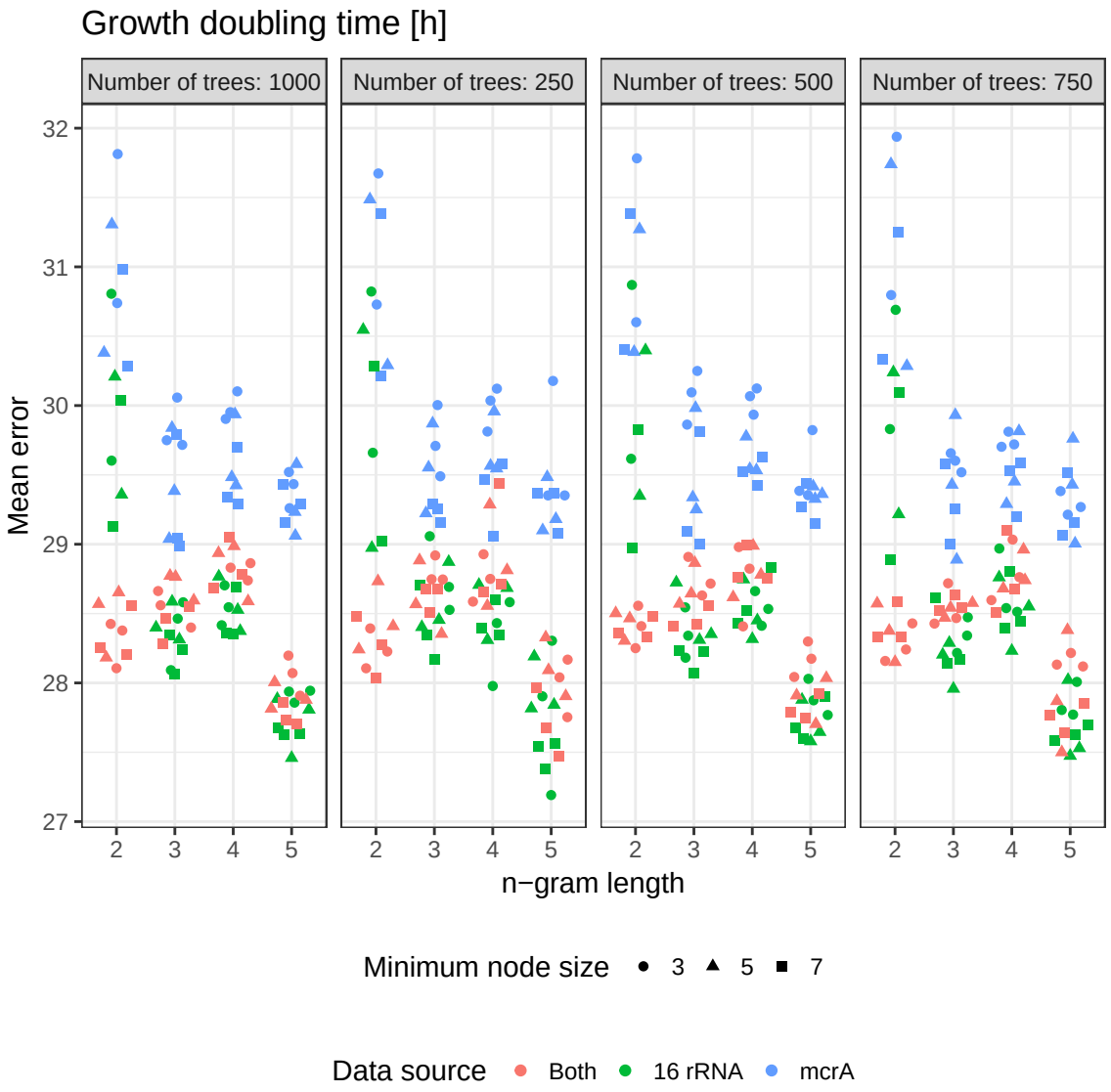
To select the most informative n-grams, we used Pearson's correlation between the feature and the target as implemented in the Rfast package. We retained a fraction of features instead of an absolute number of features to keep the feature selection consistent between datasets with varying number of features. We considered following fractions of features: 0.25, 0.5 and 1 (when we do not select any features). In all cases, the strictest feature selection (0.25) proved to create the most efficient classifiers.

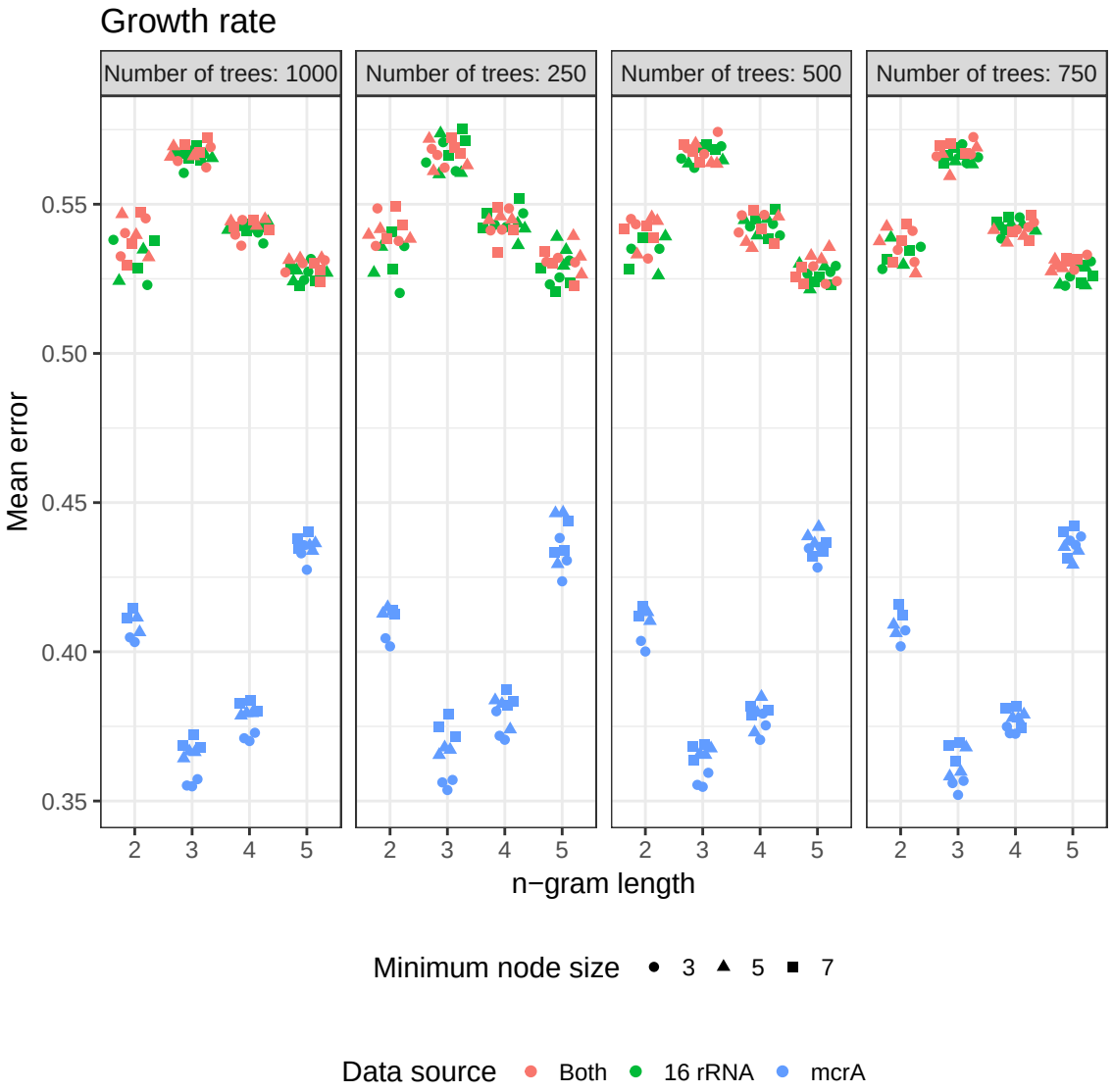
Number of variables to possibly split at in each node

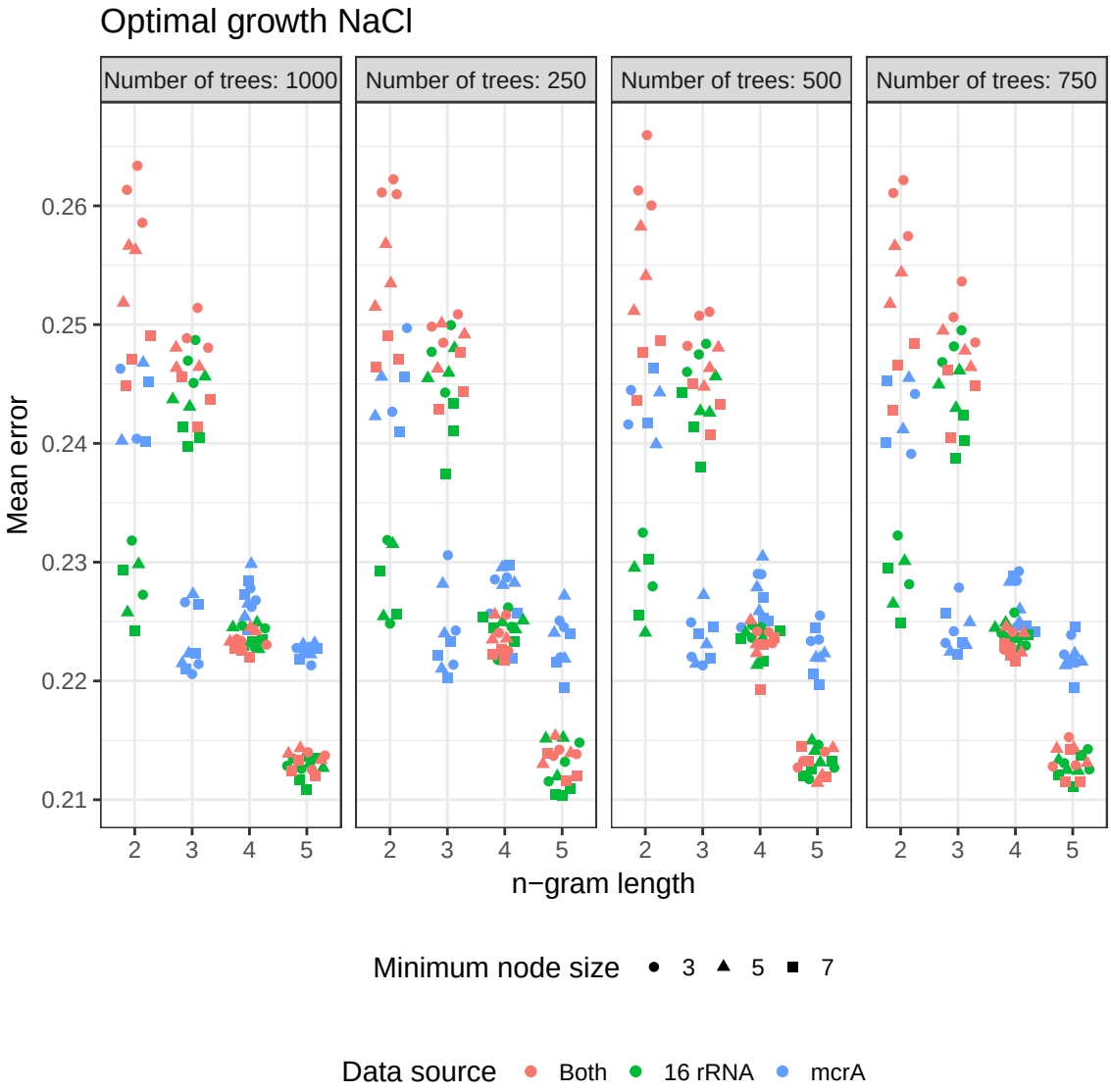
Aside from the optimum number of variables to possibly split (5), we have also considered 3 and 7 variables.

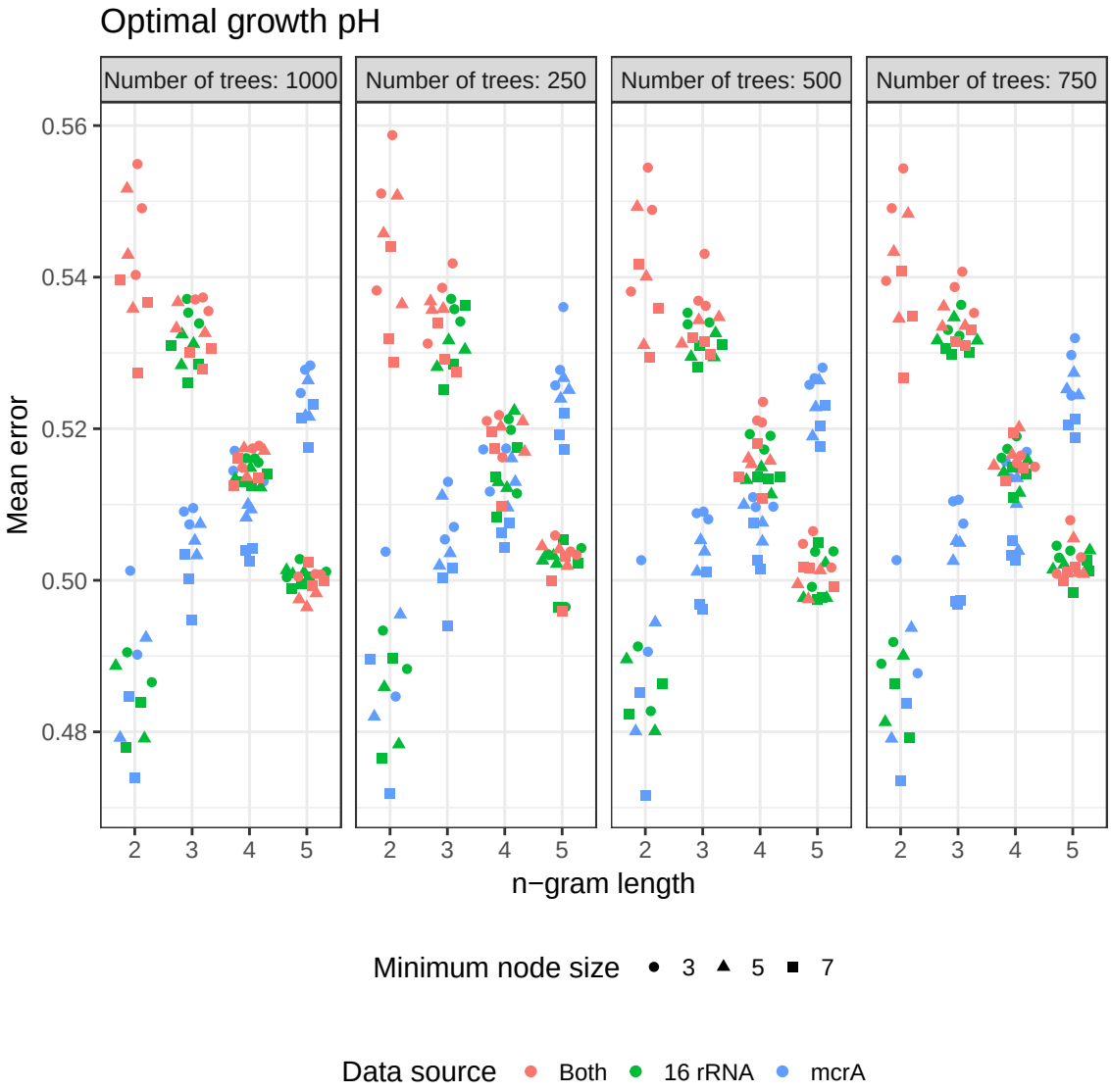
Minimal node size

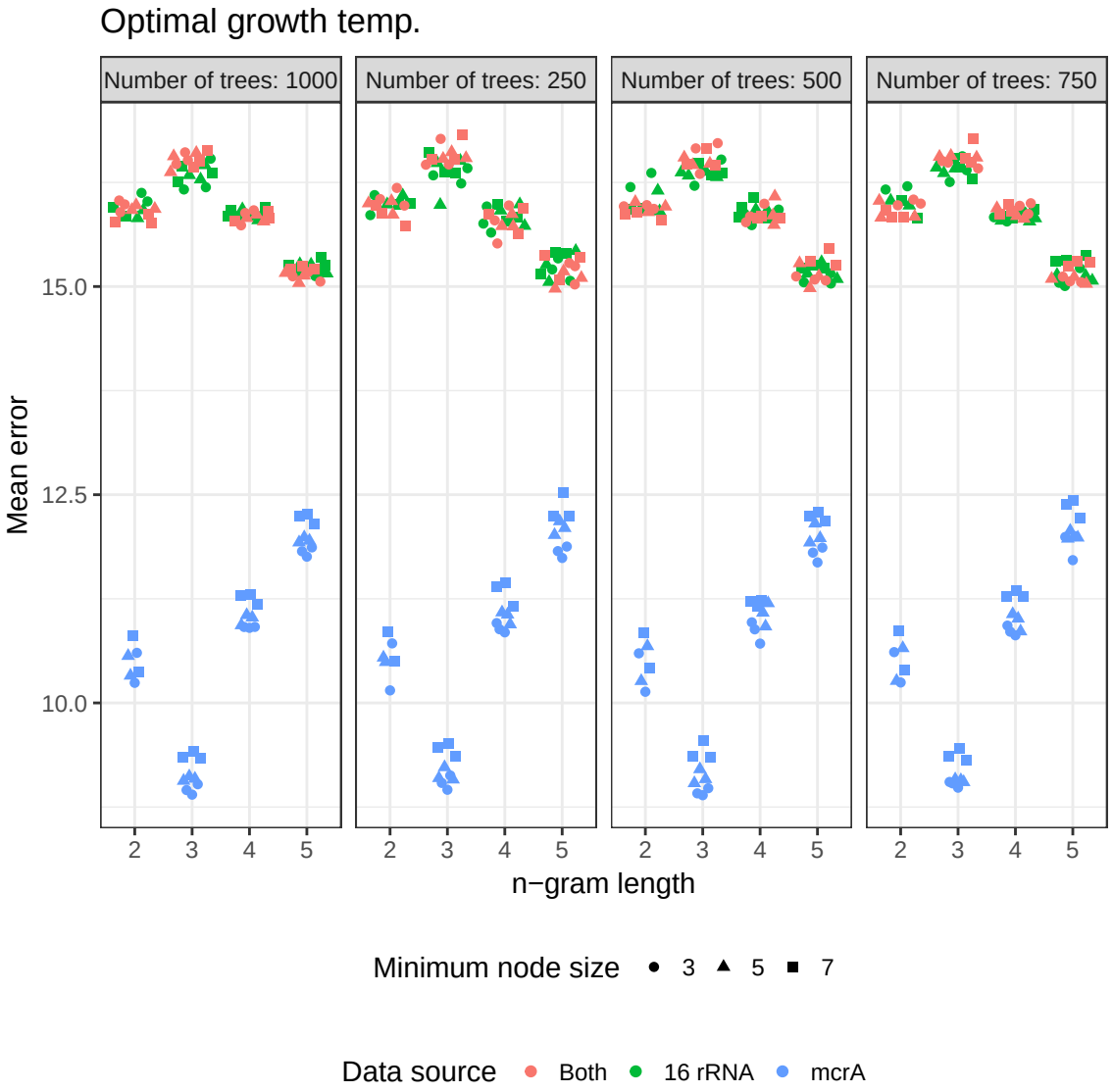
```
## Warning: Column `task.id` joining factors with different levels, coercing
## to character vector
```




Growth rate
Number of trees: 1000
Number of trees: 250
Number of trees: 500
Number of trees: 750
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
Data source
Both
16 rRNA
mcrA


Optimal growth NaCl
Number of trees: 1000
Number of trees: 250
Number of trees: 500
Number of trees: 750
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
Data source
Both
16 rRNA
mcrA


Optimal growth pH
Number of trees: 1000
Number of trees: 250
Number of trees: 500
Number of trees: 750
Mean error
0.56
0.54
0.52
0.50
0.48
2
3
4
5
n−gram length
Minimum node size
3
5
7
Data source
Both
16 rRNA
mcrA


Optimal growth temp.
Number of trees: 1000
Number of trees: 250
Number of trees: 500
Number of trees: 750
Mean error
15.0
12.5
10.0
2
3
4
5
n−gram length
Minimum node size
3
5
7
Data source
Both
16 rRNA
mcrA