

PhyMet²: complex database containing records on methanogens with unique feature (MethanoGram) allowing prediction of culture conditions based on 16S rRNA.

Michał Burdukiewicz¹, Przemysław Gagat¹, Sławomir Jabłoski², Jarosław Chilimoniuk¹, Michał Gaworski², Paweł Mackiewicz¹ and Marcin Łukaszewicz^{2*}
*marcin.lukaszewicz@uwr.edu.pl

¹University of Wrocław, Department of Genomics, ²University of Wrocław, Department of Biotransformation

Introduction

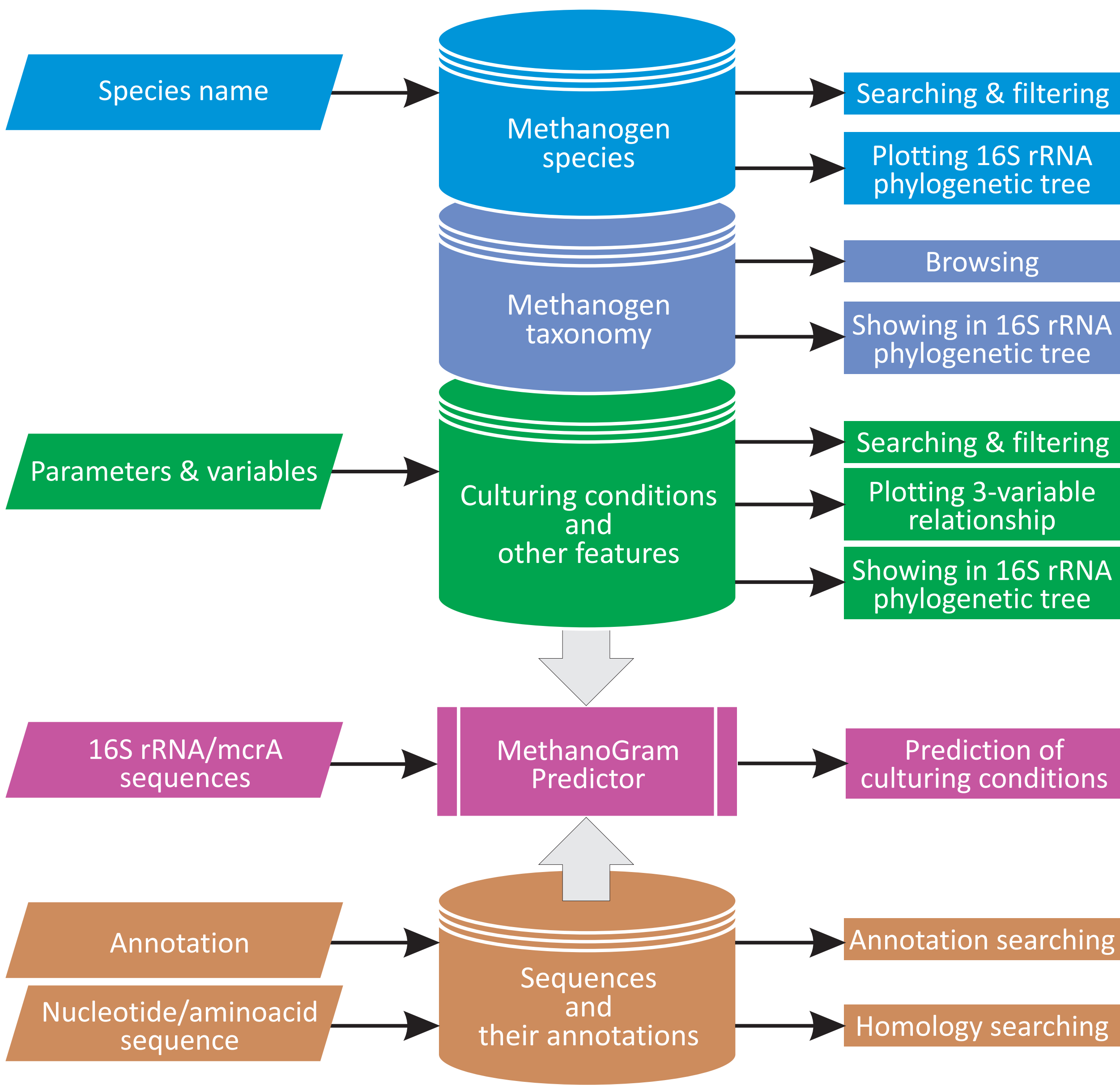
Tuning of MethanoGram

Methanogens are methane-producing anaerobic archaea, that can be found in many anaerobic habitats. They are recognized as the largest biogenic source of methane, which is a potent greenhouse gas, and consequently as an important factor in the global carbon cycle. They also show growing potential for biotechnological uses. Our rudimentary knowledge about them results from difficulties with their isolation and culturing in laboratory conditions, which are necessary to describe their phenotype. Innovations in DNA sequencing technologies allowed for rapid development of metagenomics. DNA sequencing of environmental samples resulted in identifying a plethora of new uncultivated methanogens. Therefore, we created PhyMet², the first database that combines description of methanogens and their culturing conditions with genetic information.

Data collection

PhyMet2 contains 153 manually curated and up-to-date high quality records of methanogenic species. Sequence data was collected from the NCBI (www.ncbi.nlm.nih.gov) and Silva (www.arb-silva.de) databases, and additional information, according to the minimal standards (Boone and Whitman, 1988), was obtained by thorough manual search of literature.

PhyMet² as multifunctional platform

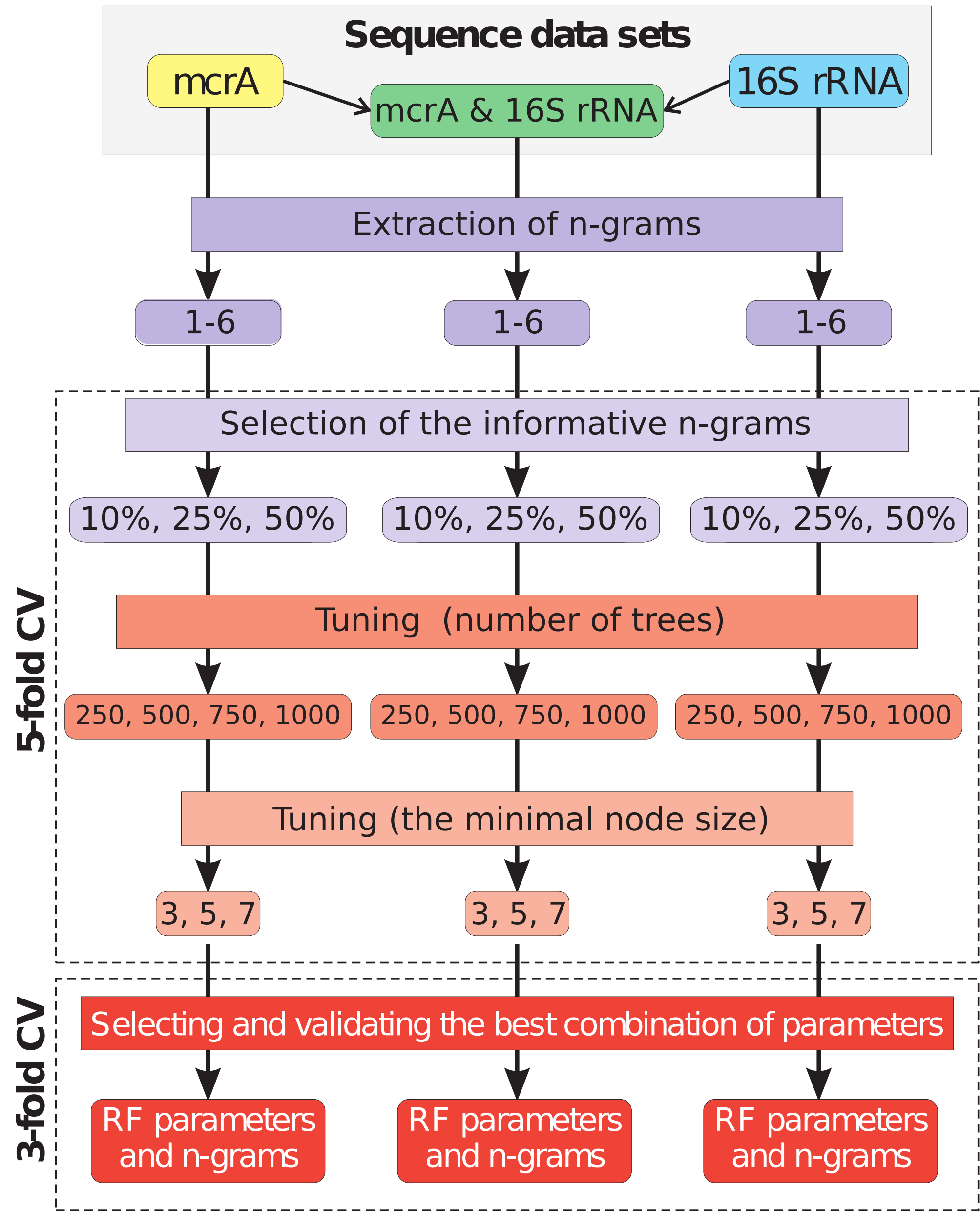


PhyMet² (Phylogeny and Metabolism of Methanogens) is the largest data analysis platform that provides information on culturing conditions and sequence data for methanogenic archaea with a user-friendly interface and a set of tools for interactive data browsing using methanogens species names, taxonomy, culturing conditions, environmental/phenotype features and nucleotide/protein sequences. The analyses include advanced data browsing, exploring phylogeny, plotting selected features, searching for potential sequence homologues and predicting key culturing conditions for newly discovered methanogens based on 16S rRNA sequences. The database contains 153 methanogens characterized by 50 features organized into 13 categories, 88 complete genomes, 200,000 protein coding nucleotide/amino-acid sequences, and 1,200 rRNA and 4,100 tRNA sequences.

MethanoGram

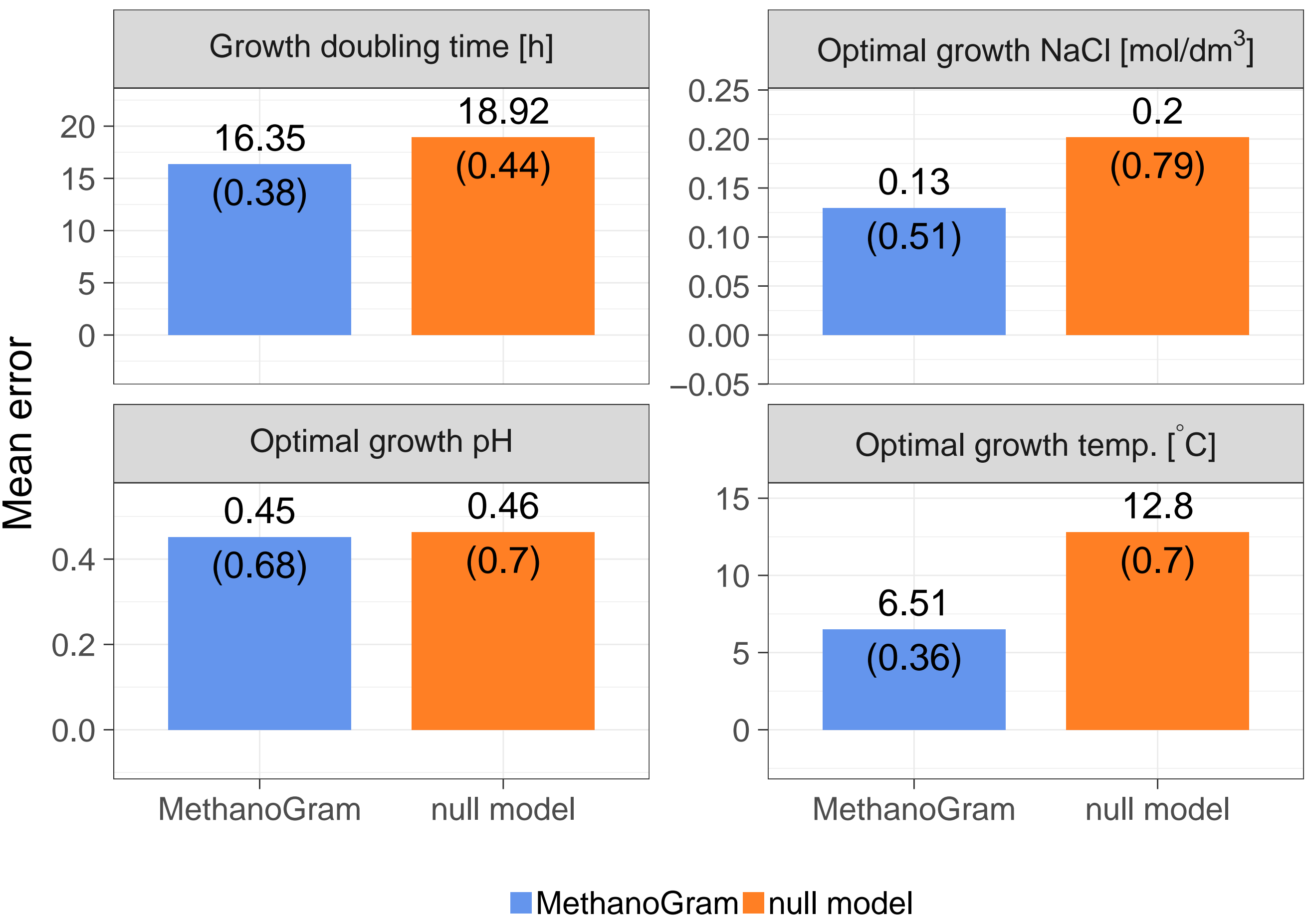
From metagenomics analyses, we know that there is a plethora of new uncultivated microorganisms, including methanogens. Unfortunately, we can often cultivate much less species than actually occur in a given environment. The data contained in PhyMet2 was used to develop a web server, MethanoGram, that quickly and accurately predicts the conditions for optimal growth of methanogens: temperature, pH, and NaCl concentration, i.e. the key factors that shape the composition of methanogenic communities.

In order to train MethanoGram, we used n-grams, i.e. subsequences of the length n that were extracted from 16S rRNA. We chose only those species that have known 16S rRNA as well as all important culturing conditions. The final set included 67 methanogens.



To estimate the culturing conditions we chose the random forests algorithm because of its high performance and resistance to over-fitting(Breiman, 2001). They were implemented in the ranger R package(Wright and Ziegler, 2015). In order to find the optimal values of hyperparameters, we performed a nested cross-validation of random forest classifiers.

Mean error



Funding and aviability

PhyMet² is available at: <http://metanogen.biotech.uni.wroc.pl/>.
MethanoGram is available as a web-server:
<http://www.smorfland.uni.wroc.pl/shiny/mgp/>.

This work was supported by the Leading National Research Center (KNOW) and the National Science Centre grant no. 2015/17/N/NZ2/01845 and 2017/24/T/NZ2/00003.

Bibliography

Boone, D. R. and Whitman, W. B. (1988). Proposal of Minimal Standards for Describing New Taxa of Methanogenic Bacteria†. *International Journal of Systematic and Evolutionary Microbiology*, 38(2):212–219.
Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.
Wright, M. N. and Ziegler, A. (2015). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv:1508.04409 [stat]*. arXiv: 1508.04409.