PhyMet^2: complex database containing records on methanogens with unique feature (MethanoGram) allowing prediction of culture conditions based on 16S rRNA

Michał Burdukiewicz[1], Przemysław Gagat[1], Sławomir Jabłoski[2], Jarosław Chilimoniuk[1], Michał Gaworski[2], Paweł Mackiewicz[1] and Marcin Łukaszewicz[2]*
*marcin.lukaszewicz@uwr.edu.pl

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

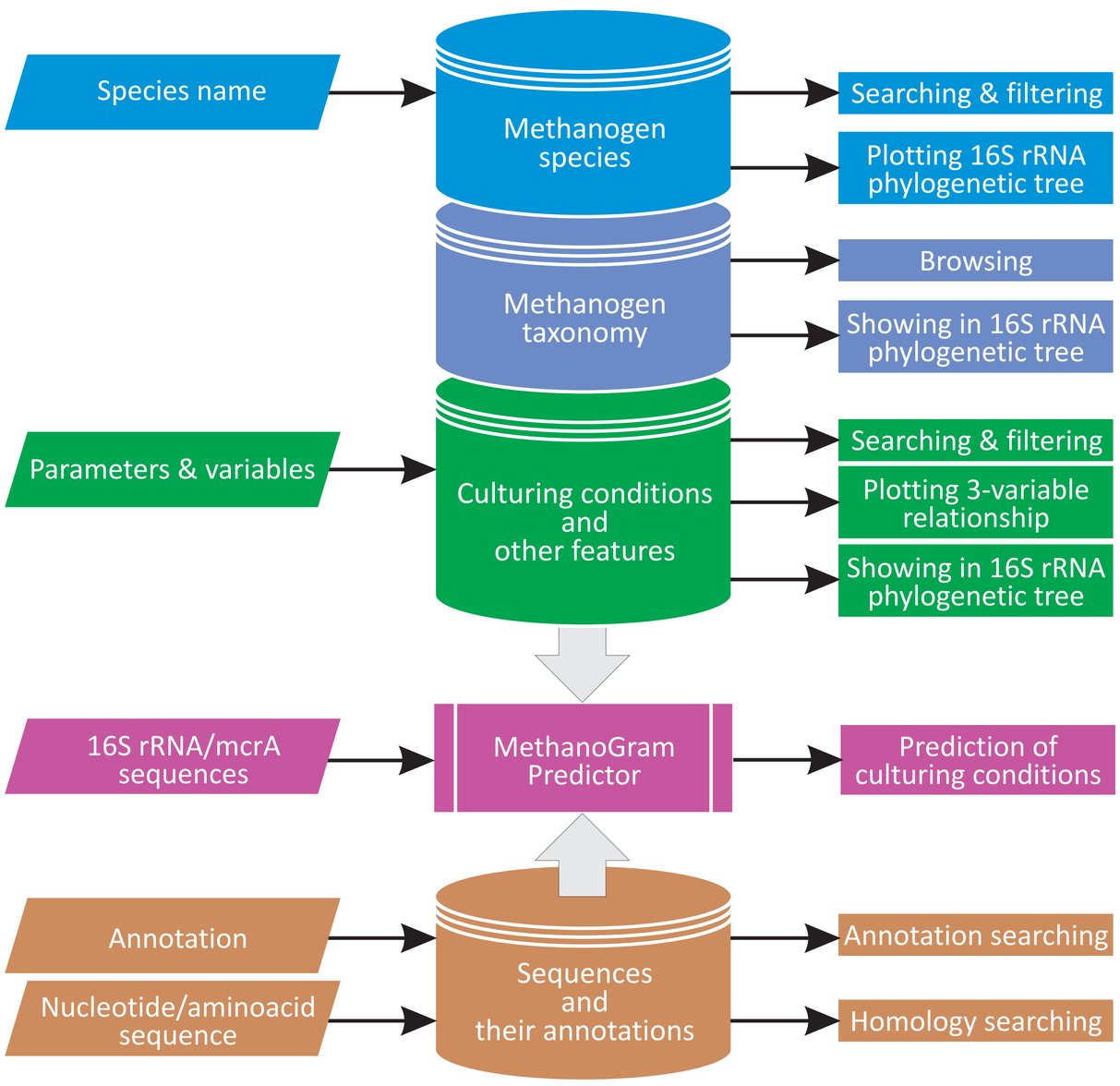
Introduction

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 PhyMet^2, 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, was obtained by thorough manual search of literature.

PhyMet^2 as multifunctional platform



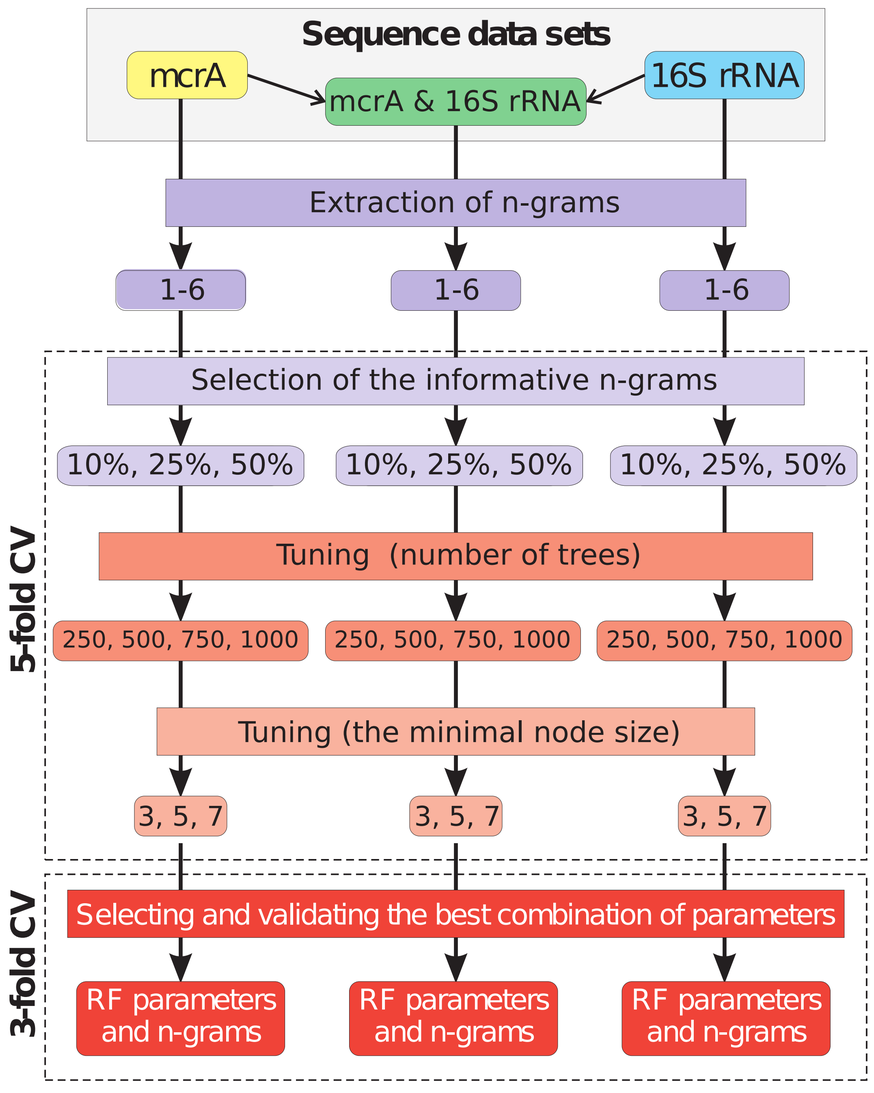
PhyMet^2 (Phylogeny and Metabolism of Methanogens) is the largest data analysis platform that provides information on culturing conditions and sequence data for methanogenic archaea with a user-friendly interface. The analyses include advanced data browsing, exploring phylogeny, plotting selected features, searching for potential sequence homologues and predicting key culturing conditions for newly discovered methanogens based on 16S rRNA sequences.

MethanoGram

The unique feature of PhyMet2 is a web server, MethanoGram, that predicts conditions for the optimal growth of methanogens: temperature, pH, NaCl concentration, and growth doubling time.
MethanoGram is one of the first approaches aiming at predicting the phenotype of microorganisms based on molecular markers, and hopefully will boost further research in the field. We would also like to apply our algorithm to prediction of culturing conditions for other microorganisms.

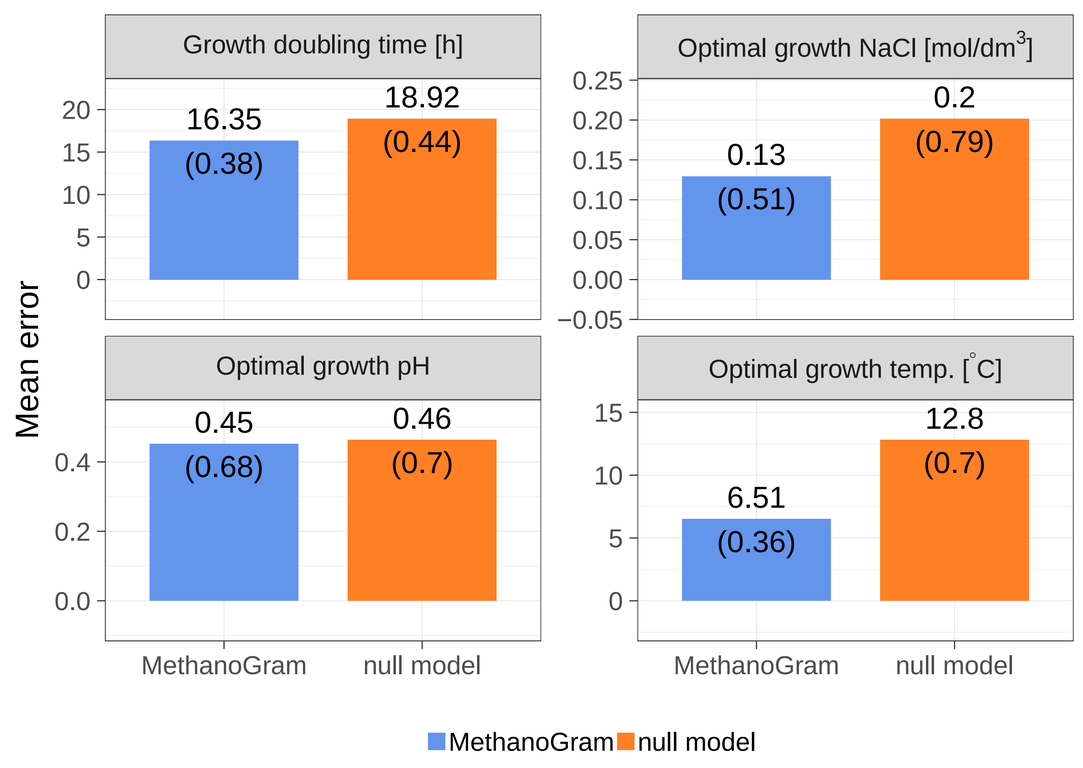
Tuning of MethanoGram

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.



To estimate the culturing conditions we implemented random forests algorithm and performed a nested cross-validation.

Mean error



Null model does not incorporate any sequence-based information.
Values above columns represent mean error of algorithm, below show mean errors of the best predictors found in the nested cross-validation.

Funding and aviability

PhyMet^2 is avaible at: `http://metanogen.biotech.uni.wroc.pl/`.

MethanoGram is avaible as a web-server: `http://www.smorfland.uni.wroc.pl/shiny/mgp/`.

The exact details on training of MethanoGram are accessible at: `https://github.com/michbur/PhyMet2_supplements`.

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.