environmental microbiology reports

Environmental Microbiology Reports (2018) **10**(3), 378–382 doi:10.1111/1758-2229.12648

PhyMet^2: a database and toolkit for phylogenetic and metabolic analyses of methanogens

Burdukiewicz Michał,[1] Przemysław Gagat,[1] Sławomir Jabłoński,[2] Jarosław Chilimoniuk,[1] Michał Gaworski,[2] Paweł Mackiewicz[1] and Łukaszewicz Marcin[2]*

[1]*Department of Genomics, Faculty of Biotechnology, University of Wrocław, Wrocław, Poland.*

[2]*Department of Biotransformation, Faculty of Biotechnology, University of Wrocław, Wrocław, Poland.*

Summary

The vast biodiversity of the microbial world and how little is known about it, has already been revealed by extensive metagenomics analyses. Our rudimentary knowledge of microbes stems from difficulties concerning their isolation and culture in laboratory conditions, which is necessary for describing their phenotype, among other things, for biotechnological purposes. An important component of the understudied ecosystems is methanogens, archaea producing a potent greenhouse-effect gas methane. Therefore, we created PhyMet^2, the first database that combines descriptions of methanogens and their culturing conditions with genetic information. The database contains a set of utilities that facilitate interactive data browsing, data comparison, phylogeny exploration and searching for sequence homologues. The most unique feature of the database is the web server MethanoGram, which can be used to significantly reduce the time and cost of searching for the optimal culturing conditions of methanogens by predicting them based on 16S RNA sequences. The database will aid many researchers in exploring the world of methanogens and their applications in biotechnological processes. PhyMet^2 with the MethanoGram predictor is available at http://metanogen.biotech.uni.wroc.pl

Introduction

Innovations in DNA sequencing technologies allowed for the rapid development of metagenomics, DNA sequencing of environmental samples, and consequently, identification of a plethora of new uncultivated microorganisms (Meyerdierks *et al*., 2005; Chojnacka *et al*. 2015; Emerson *et al*., 2016). Many of the microorganisms are of great importance as they might play a significant role in climate change, for example, the methanogen, that is, methane-producing archaea, Candidatus '*Methanoflorentaceae stordalenmirensis*' (Mondav *et al*., 2014) or in the understanding of Eukaryote evolution, for example, the methanogens from the Asgard lineage (Zaremba-Niedzwiedzka *et al*., 2017). The former is a substantial contributor to methane-based positive feedback in global warming due to its prevalence in thawing permafrost, the latter represent an archaeal clade that closely affiliates with eukaryotes in phylogenomic analyses. At present, there are many more microorganisms identified by metagenomic high-throughput methods than by isolation. However, in order to describe the phenotype of microorganisms, that is, to gather data on their physiology, morphology and biochemistry, the metagenome analyses need to be supplemented with studies of microorganisms isolated in pure culture. Unfortunately, searching for the optimal culturing conditions is expensive, time-consuming and technically difficult.

As in the case of reverse genetics, which largely dominated the classical genetics approach, it could be hypothesized that *in silico* prediction of culturing conditions for newly discovered microorganisms would be possible based on DNA sequences. Indeed, the first approaches have already shown that the phylogenetic signal from DNA may be used to predict the phenotype of microorganisms, especially the conserved traits, which are encoded by numerous genes (Martiny *et al*., 2013; Goberna and Verdú, 2016; Martínez-García *et al*., 2016). Since the prediction relies on the amount of genetic and phenotypic information available, a comprehensive database is a vital preliminary in the construction of a prediction algorithm (Martiny *et al*., 2013; Goberna and Verdú, 2016; Martínez-García *et al*., 2016). Having a comprehensive database for methanogens (Jabłoński *et al*., 2015) at our disposal, we decided to focus on these particular microorganisms in the search for the best algorithm that would predict

Received 22 December, 2017; revised 6 March, 2018; accepted 27 March, 2018. *For correspondence. E-mail marcin.lukaszewicz@uwr.edu.pl; Tel. 0048 71 3756 250.

culturing conditions based on genetic markers. A robust prediction algorithm would be a milestone in research on uncultivated microorganisms.

We built the methanogen database because we have had experience with these microorganisms from previous studies and consider them of great global importance (Jabłoński *et al*., 2015). They are widespread in nature but restricted to anaerobic habitats, for example wetlands, landfill sites, and digestive tracts of animals. They are often found in extreme environments and might even be capable of growing on Mars (Conrad, 2007; Thauer *et al*., 2008; Wagner and Liebner, 2009; Mickol and Kral, 2016). Methanogens 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 (Houweling *et al*., 2008). They also show growing potential for many biotechnological uses (Goyal *et al*., 2016).

In order to create the predictor of culturing conditions for newly discovered methanogens, the first version of our methanogens database, accessible since January 2015, was updated (Jabłoński *et al*., 2015). The second version of the database transformed it from a simple information repository to an advanced platform for data analysis. PhyMet^2 (Phylogeny and Metabolism of Methanogens) is the largest database 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, searching, sorting, comparing and downloading (Fig. 1). It is the first database that combines species descriptions and culturing conditions with genetic information, thereby setting standards for other biological databases.

The data contained in PhyMet^2 was used to develop a web server, MethanoGram, that quickly and accurately predicts conditions for optimal growth of methanogens: temperature, pH and NaCl concentration, that is, the key factors that shape the composition of methanogenic communities (Wen *et al*., 2017). Using this tool, researchers could reduce the number of experiments and thus the cost of searching for the optimal culturing conditions of newly discovered methanogens. The predictions are based on a standard phylogenetic marker 16S rRNA.

Results and discussion

PhyMet^2 database

PhyMet^2 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 (see Supporting Information).

The simplest access to the data in PhyMet^2 is available via the customizable table on the main page, which allows, for example, to select the species most suitable for cost-effective methanogenesis, or to design optimal operating conditions in bioreactors for a particular methanogen. To make the search effective and specific, the users can apply multiple filters at the same time or explore the database using the 'Advanced' search tab, which provides more filtering options for numerical and character data. Users may also use the 'Taxonomy' tab, where the methanogens are grouped into classes, orders, families and genera. The search results can be easily downloaded into a CSV file and all data in XML format. The user can also create charts visualizing at the same time three selected methanogen features, and infer from their distribution how, for example, key culturing conditions characterize the methanogens and identify interesting outliers.

PhyMet^2 contains a set of bioinformatics utilities, which may be especially helpful in the characterization of new methanogens. They enable the user to: (i) search for potential nucleotide or protein sequence homologues, (ii) interactively analyze the phylogeny of methanogens and (iii) predict key optimal culturing conditions for newly discovered methanogens (Fig. 1). In order to find similarity between a query sequence and manually curated high-quality sequences deposited in PhyMet^2, we set up a standalone Blast algorithm (Boratyn *et al*., 2012). The interactive dot plots and 16S rRNA phylogenetic tree were implemented in Plotly and phylotree.js, respectively (see Supporting Information). The prediction of culturing conditions is performed by the web server MethanoGram based on 16S rRNA.

MethanoGram predictor

The unique feature of PhyMet^2 is a web server, MethanoGram, that predicts conditions for the optimal growth of methanogens: temperature, pH, NaCl concentration and growth doubling time. It makes the prediction using a random forest algorithm (Breiman, 2001) trained on n-grams (k-mers, subsequences of length *n*) extracted from 16S rRNA. The random forests and n-grams were chosen because of their speed and good performance (McNair *et al*., 2012; Burdukiewicz *et al*., 2017). Moreover, with the advent of alignment-free phylogenetics, n-grams are becoming more commonly regarded as the carriers of evolutionary information (Bonham-Carter *et al*., 2014). In order to avoid overfitting and to find the best-performing predictors, we conducted a nested cross-validation over 20,000 different models. The details of the training procedure as well as the results are described in the Supporting Information.

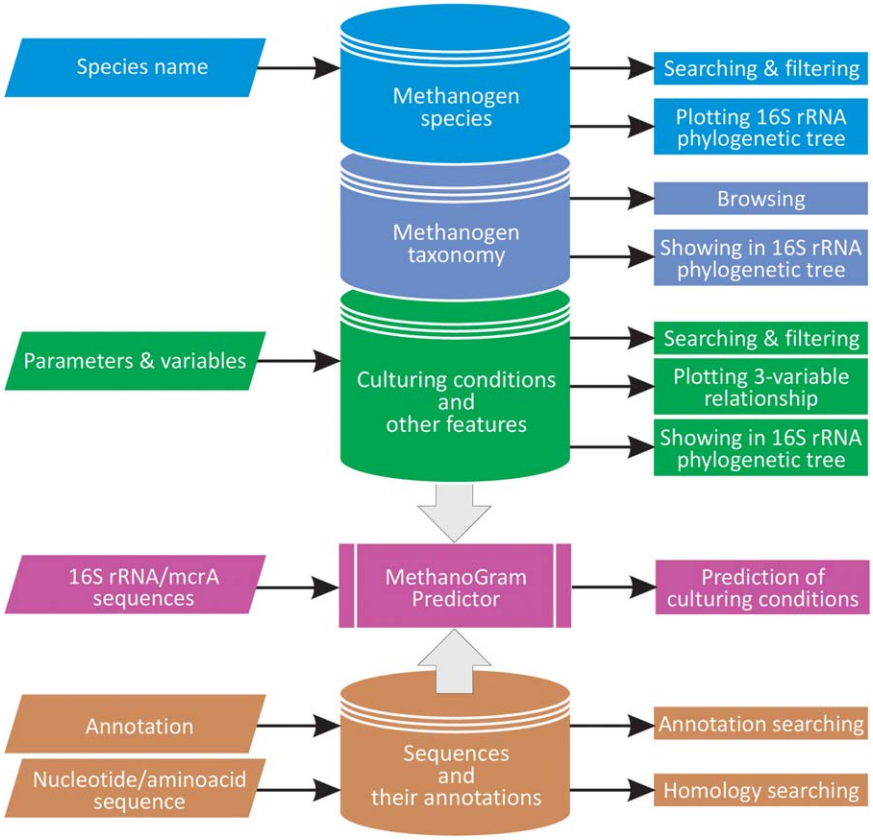

Fig. 1. PhyMet2 is a multi-functional platform that allows for various analyses using methanogens' species names, taxonomy, culturing conditions, environmental/phenotype features and nucleotide/protein sequences. The analyses include advanced data browsing, exploring phylogeny, plotting selected features, searching for potential sequence homologues and predicting key culturing conditions for newly discovered methanogens based on 16S rRNA. The database comprises 153 methanogens characterized by ~ 50 features organized into 13 categories, 88 complete genomes, ~ 200,000 protein coding nucleotide/amino acid sequences, and ~ 1200 rRNA and ~ 4100 tRNA sequences.

While searching for the most accurate predictor, we evaluated combinations of n-grams of different lengths. We also tested our algorithms trained on: (i) the information from a single type of molecular marker, 16S rRNA or *mcrA* and (ii) from both markers at the same time. Comparing the prediction results from the two molecular markers, we discovered that in the majority of cases, the most relevant information contained in n-grams was derived from 16S rRNA. Only for growth doubling time and optimal growth temperature, the addition of n-grams extracted from *mcrA* allowed training of slightly more accurate predictors (see Supporting Information). As iterations of MethanoGram based solely on 16S rRNA are as accurate or almost as accurate as iterations trained on 16S rRNA and *mcrA*, we decided to employ only the rRNA sequences. Therefore, potential users of MethanoGram have to upload only the 16S rRNA sequence to predict the culturing conditions. The mean error, for example, of the predicted optimal growth pH is 0.45, which means that the user can expect the optimal pH in the range of this deviation (Table 1).

We also compared MethanoGram to a null model which predicts a given culturing condition for a single strain as the median value of culturing conditions for all other strains (Fig. 2). Since the null model does not incorporate any sequence-based information, we expected it to show significantly higher mean error than the well-optimized MethanoGram. Although MethanoGram always outperforms the null model for all culturing conditions, the differences range from marginal (0.01) for the optimal growth pH to very drastic (6.13°C) for the optimal temperature (Fig. 2). These differences can

Table 1. Results of jackknife test of MethanoGram.

Culturing condition	Mean error	Normalized mean error
Growth doubling time (h)	16.3	0.45
Optimal growth temp. (°C)	6.67	0.53
Optimal growth pH	0.45	0.73
Optimal growth NaCl (mol dm^{-3})	0.13	0.38

The normalized mean error is the mean error divided by standard deviation of respective culturing condition.

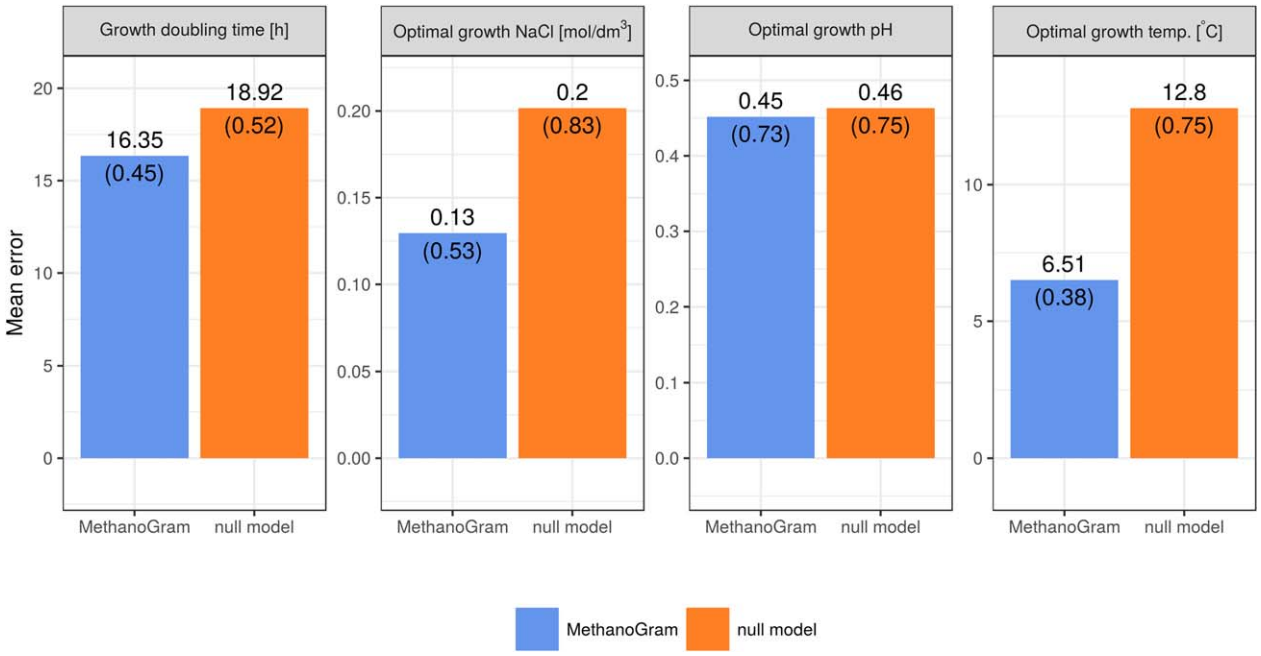

Fig. 2. Comparison of the MethanoGram algorithm with the null model to validate its usefulness in predicting optimal culturing conditions for methanogens. The normalized mean error divided by standard deviation of respective culturing condition is shown in parentheses.

reflect (i) specificity of culturing traits in relation to methanogenic species, (ii) uncertainties concerning their measurements during taxa description, for example, low density of measurements and (iii) weak correlation between culturing traits and n-gram distribution. Therefore, the future versions of MethanoGram have to be trained on n-grams derived from a much wider array of genes and culturing data.

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.

The exact details on training of MethanoGram are accessible in Supporting Information. The code necessary to reproduce the analysis is hosted online: https://github.com/michbur/PhyMet2_supplements.

Acknowledgements

The authors alone are responsible for the content and writing of the article. We are grateful to Michael Nikiel for his language editing. This work was supported by the Leading National Research Center (KNOW) and the National Science Centre grant no. 2015/17/N/NZ2/01845, 2017/24/T/NZ2/00003, 2017/26/D/NZ8/00444.

References

Bonham-Carter, O., Steele, J., and Bastola, D. (2014) Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Brief Bioinform* **15**: 890–905.

Boone, D.R., and Whitman, W.B. (1988) Proposal of minimal standards for describing new taxa of methanogenic bacteria. *Int J Syst Evol Microbiol* **38**: 212–219.

Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., and Madden, T.L. (2012) Domain enhanced lookup time accelerated BLAST. *Biol Direct* **7**: 12.

Breiman, L. (2001) Random forests. *Mach Learn* **45**: 5–32.

Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017) Amyloidogenic motifs revealed by n-gram analysis. *Sci Rep* **7**: 12961.

Chojnacka, A., Szczęsny, P., Błaszczyk, M.K., Zielenkiewicz, U., Detman, A., Salamon, A., and Sikora, A. (2015) Noteworthy facts about a methane-producing microbial community processing acidic effluent from sugar beet molasses fermentation. *PloS One* **10**: e0128008.

Conrad, R. (2007) Microbial ecology of methanogens and methanotrophs. In *Advances in Agronomy*. New York, NY: Academic Press, pp. 1–63.

Emerson, J.B., Thomas, B.C., Alvarez, W., and Banfield, J.F. (2016) Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. *Environ Microbiol* **18**: 1686–1703.

Goberna, M., and Verdú, M. (2016) Predicting microbial traits with phylogenies. *ISME J* **10**: 959–967.

Goyal, N., Zhou, Z., and Karimi, I.A. (2016) Metabolic processes of *Methanococcus maripaludis* and potential applications. *Microb Cell Fact* **15**: 107.

Houweling, S., Van der Werf, G., Klein Goldewijk, K., Röckmann, T., and Aben, I. (2008) Early anthropogenic CH_4 emissions and the variation of CH_4 and $^{13}CH_4$ over

the last millennium. *Global Biogeochem Cycles* **22**: GB1022.
Jabłoński, S., Rodowicz, P., and Łukaszewicz, M. (2015) Methanogenic archaea database containing physiological and biochemical characteristics. *Int J Syst Evol Microbiol* **65**: 1360–1368.
Martínez-García, P.M., López-Solanilla, E., Ramos, C., and Rodríguez-Palenzuela, P. (2016) Prediction of bacterial associations with plants using a supervised machine-learning approach. *Environ Microbiol* **18**: 4847–4861.
Martiny, A.C., Treseder, K., and Pusch, G. (2013) Phylogenetic conservatism of functional traits in microorganisms. *ISME J* **7**: 830–838.
McNair, K., Bailey, B.A., and Edwards, R.A. (2012) PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* **28**: 614–618.
Meyerdierks, A., Kube, M., Lombardot, T., Knittel, K., Bauer, M., Glöckner, F.O., *et al*. (2005) Insights into the genomes of archaea mediating the anaerobic oxidation of methane. *Environ Microbiol* **7**: 1937–1951.
Mickol, R.L.., and Kral, T.A. (2016) Low pressure tolerance by methanogens in an aqueous environment: implications for subsurface life on mars. *Orig Life Evol Biosph* **47**: 511–532.
Mondav, R., Woodcroft, B.J., Kim, E.-H., McCalley, C.K., Hodgkins, S.B., Crill, P.M., *et al*. (2014) Discovery of a novel methanogen prevalent in thawing permafrost. *Nat Commun* **5**: 3212.
Thauer, R.K., Kaster, A.-K., Seedorf, H., Buckel, W., and Hedderich, R. (2008) Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat Rev Microbiol* **6**: 579.
Wagner, D., and Liebner, S. (2009) Global warming and carbon dynamics in permafrost soils: methane production and oxidation. In *Permafrost Soils*. Varma, A. (ed). New York, NY: Springer, pp. 219–236.
Wen, X., Yang, S., Horn, F., Winkel, M., Wagner, D., and Liebner, S. (2017) Global biogeographic analysis of methanogenic archaea identifies community-shaping environmental factors of natural environments. *Front Microbiol* **8**: 1339.
Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., *et al*. (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**: 353–358.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article.

Supporting Information S1 Fig. 1. The Venn diagram of methanogenic species used in the analysis.
Supporting Information S1 Fig. 2. The scheme of tuning procedure which was aimed at receiving the optimal values and combination of random forest parameters and the informative n-grams. The tuning of parameters in the random forest algorithm was subjected to fivefold cross-validation and the final step to threefold cross-validation.
Supporting Information S1 Table 1. Mean errors of the best predictors found in the nested cross-validation for three possible data sources.
Supporting Information S1 Table 2. The hyperparameters found in the nested cross-validation.
Supporting Information S2.
Supporting Information S3 Fig. 1. The phylogenetic tree obtained in MrBayes for 16S rRNA sequences of methanogens. Particular taxonomic groups were indicated in different colours. Values at nodes indicate posterior probabilities. The values smaller than 0.9 were omitted.
Supporting Information S3 Fig. 2. The phylogenetic tree obtained in MrBayes for 16S rRNA sequences of methanogens, an-notated with a matrix of culturing conditions and other features. Values at nodes indicate posterior probabilities. The values less than 0.9 were omitted.