Prokaryote growth temperature prediction with machine learning

Akseli Reunamo

Master's Thesis

University of Turku
Department of Biology

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

UNIVERSITY OF TURKU
Department of Biology
REUNAMO, AKSELI: Prokaryote growth temperature prediction with machine learning
Master's thesis, 44 p., 51 appendix pages
Physiology and Genetics, Master of Science
August 2021
The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service

---

Archaea and bacteria can be divided into four groups based on their growth temperature adaptation: mesophiles, thermophiles, hyperthermophiles, and psychrophiles. The thermostability of proteins is a sum of multiple different physical forces such as van der Waals interactions, chemical polarity, and ionic interactions. Genes causing the adaptation have not been identified and this thesis aims to identify temperature adaptation linked genes and predict temperature adaptation based on the absence or presence of genes. A dataset of 4361 genes from 711 prokaryotes was analyzed with four different machine learning algorithms: neural network, random forest, gradient boosting machine, and logistic regression. Logistic regression was chosen to be an explanatory and predictive model based on micro averaged AUC and Occam's razor principle. Logistic regression was able to predict temperature adaptation with good performance. Machine learning is a powerful predictor for temperature adaptation and less than 200 genes were needed for the prediction of each adaptation. This technique can be used to predict the adaptation of uncultivated prokaryotes. However, the statistical importance of genes connected to temperature adaptation was not verified and this thesis did not provide much additional support for previously proposed temperature adaptation linked genes.

Keywords: Artificial Intelligence, Genome, Adaptation, Computational Biology, Supervised learning (Machine learning), Clusters of Orthologous Groups (COGs)

Table of Contents

1 Introduction

1.1 Prokaryotes

Prokaryotes are one of the keystone species maintaining the flow of nutrients and carbon, and other ecological processes on the Earth (Torsvik et al. 2002). Without prokaryotes, the equilibrium of a stable climate would be disturbed. Carbon found in prokaryotes has been estimated to be as large as total carbon found in plants and they have been found almost everywhere: aquatic environments, soil, subsurface, animals, plants, and even air (Whitman et al. 1998). In addition, the DNA amount of prokaryotes is estimated to be about the same as the total DNA in all eukaryotic groups (Landenmark et al. 2015). The kingdom consists possibly millions of different species and they are capable to live in extreme environments where very few other species are found (Pedrós-Alió & Manrubia 2016).

The prokaryote kingdom is divided into two domains: bacteria and archaea. Archaea were separated from the bacteria kingdom in 1977 (Woese & Fox 1977), but they still share several common features. Archaea and bacteria share a similar prokaryotic cell structure of which three main features are the absence of nuclear membrane, they are, as few exceptions excluded, smaller than eukaryotes, and there are major differences in cytoplasmic membrane compared to eukaryotic one (Whitman 2009). Prokaryotes can be defined shortly as cells that employ co-transcriptional translation on their main chromosomes in which translation occurs same time as the messenger RNA is growing (Martin & Koonin 2006).

Prokaryote cell structure is hard to define because exceptions can be found in all cell attributes. Thus, the following description of prokaryote cell structure by (Bertrand et al. 2018) suits typical prokaryotes. Prokaryote cell size is less than 5μm and prokaryotes' genome usually consists of one circular chromosome and several plasmids. In addition, their genetic material does not contain histones. The prokaryotic cytoplasmic membrane contains different lipids than eukaryotes, for example, the prokaryotic membrane lack sterols. In general, prokaryotes have a wider and more diverse metabolism than eukaryotes.

Regardless of the huge variety of prokaryotes, they share detectable similar genome architecture; both archaea and bacteria domains lack introns and a large fraction of genes are organized as operons (Bertrand et al. 2018). Operons are clusters of co-regulated genes with related functions (Osbourn & Field 2009). Prokaryotes share a small number of conversed operons and a huge number of unique and rare operons (Koonin & Wolf 2008). Albeit a vast number of unique or rare operons exist, corresponding operons can be identified by analyzing clusters of orthologous groups (Galperin et al. 2019). Genome size in bacteria ranges between about 112 Kb (*Nasuia deltocephalinicola*, intracellular endosymbiont) to 16.04 Mb (*Minicystis rosea*, myxobacteria), whereas in archaea ranges between 490 Kb (*Nasuia deltocephalinicola*, ectosymbiont of other archaea) (Huber et al. 2002) to 5.75Mb (*Methanosarcina acetivorans*) (Kellner et al. 2018).

1.1.1 Prokaryote evolution

According to literature, the first living organism was probably some kind of prokaryote (Bertrand et al. 2018). Interactions among species, such as competition for space and resources and cooperation, have been proposed to be the driving force of differentiation and genomic development.

The evolution of prokaryotes can happen through multiple different processes including mutations, rearrangements, and horizontal gene transfer (HGT) within different or closely related taxa (Juhas et al. 2009). HGT occurs through transformation, conjugation, and transduction. It has been stated that HGT is the major evolutionary force of prokaryotic evolution and it helps in adaptation to new environmental conditions because this process expanses the gene content and introduces genes for new metabolic functions (López-García et al. 2015).

HGT has a major effect on adaptation, but gene duplication and *de novo* gene appearance may also lead to new properties, although *de novo* gene formation is a rare and highly unlikeable event in gene gain compared to HGT (Puigbò et al. 2014). Genes can be gained, but adaptation may also be obtained through gene loss. Gene loss is more common than gene gain, leading to the suggestion that in the evolution of prokaryotes genome reduction would be the default evolutional process and the loss of genes is compensated by gene gain via HGT (Puigbò et al. 2014). To understand genome function, genome annotation accuracy depends on the accurate identification of orthologous genes (Makarova et al. 2007).

Orthologous proteins are products of genes that can be found in at least two different species, but the genes are inherited from a single gene of the last common ancestor (Sonnhammer & Koonin 2002). Information about conserved protein groups can be leveraged to understand functional and evolutional perspectives of prokaryotes. Orthologs typically have the same function, thus identification of them gives a framework for functional and evolutionary genome analysis (Tatusov et al. 1997). The Clusters of Orthologous groups database was created in 1997 (Tatusov et al. 1997) and further updated first in 2003 (Tatusov et al. 2003), second in 2014 (Galperin et al. 2015) and, third in 2020 (Galperin et al. 2021). Original COGs were formed in six stage process defined by Tatusov et al. (2000). First, protein sequences were compared all-against-all. Next, paralogs were detected and combined meaning a combination of proteins in the same genome that are more like each other than any protein of other species. Next, triangles of mutually consistent genome-specific best hits were detected taking into account the paralogs detected in the second phase. Next, triangles were combined with a common side to form COGs (Figure 1). Next, each COG was analyzed individually to remove false positives and to identify multidomain proteins. Detected multidomain proteins were divided into single-domain components and treated with four previous steps. Finally, COGs that had multiple members and were found from all or multiple genomes were inspected with phylogenetic trees, cluster analysis and, visual inspection of alignments to define the final set of COGs.

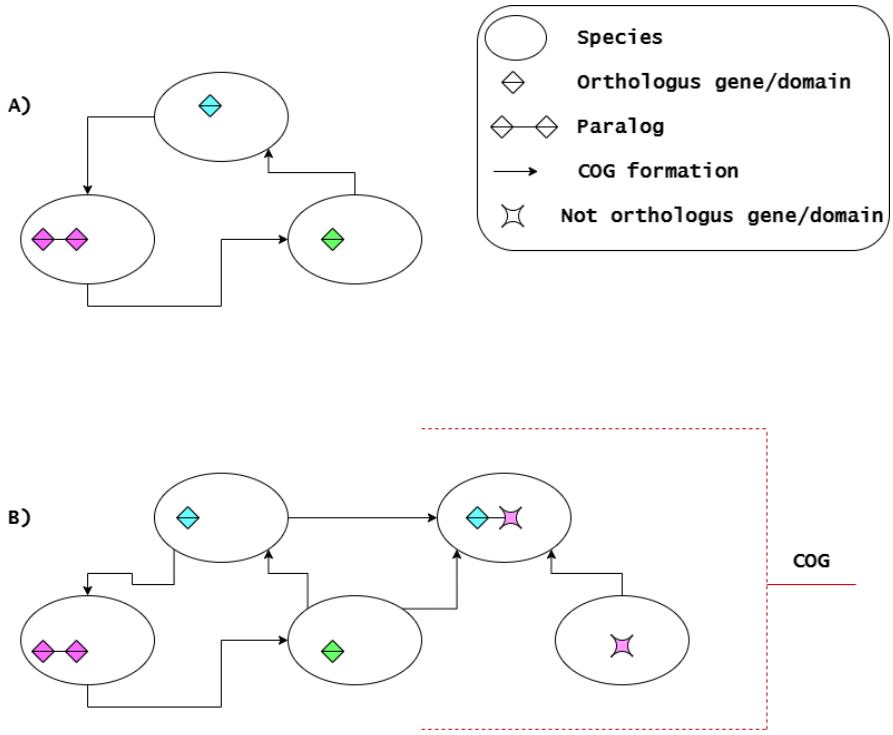


Figure 1. Formation and combination of COG triangles. a) Species that have orthologous genes or domains including paralogs are combined to form a COG triangle. b) Formed COG triangle is combined with common side (Kristensen et al. 2011)

1.1.2 Prokaryote temperature adaptation

An environment where nutrient and habitat space are limited causes competition between species that share the same living conditions. If the environment provides ecological niches, competition may lead to variant selection. Habitat temperature can also be an environmental niche that provides a chance to avoid competition. In the case of most bacterial species, they have a rapid growth rate and large populations that give rise to possibilities of new mutations into the population (Hibbing et al. 2010).

New genes that have arisen through duplication are called paralogs and bacteria genomes consist of a significant number of them, ranging from 7% to 41% in a dataset containing 106 bacterial genomes (Gevers et al. 2004). Adaptation to specific temperatures has been found to be a fluxing feature that can be gained particularly through HGT or lost several times in evolutionary short periods (Puigbò et al. 2008).

Prokaryotes can be classified into four classes by their favored growth temperature. Psychrophiles grow the best below 20°C, mesophiles grow the best at moderate temperatures

between 20 and 50°C, thermophiles grow the best between 50 and 80°C, and hyperthermophiles temperature higher than 80°C (Puigbò et al. 2008). The exact temperature ranges of classes vary slightly among literature (Table 1) (Berezovsky & Shakhnovich 2005, Goldstein 2007, Allaby 2010), thus the above definition has to be taken with reservations.

Article	**Psychrophiles**	**Mesophiles**	**Thermophiles**	**Hyperthermophiles**
Berezovsky and Shakhnovich, 2005	Not defined	G<60°C	G<80°C	G>80°C
Puigbo et al., 2008	G<20°C	20<G<50 °C	50<G<80°C	G>80°C
Goldstein, 2009	G<20°C	20<G<45 °C	45<G<80°C	G>80°C
Allaby, 2010	G<15°C	20<G<45 °C	G<=60°C	G>90°C

Table 1. Psychrophile, mesophile, thermophile, and hyperthermophile growth temperature definitions among literature. G is an abbreviation for growth temperature.

Two major features affecting the physical thermostability of proteins are the amino acid sequence and protein structure. The most reported features that affect protein stability are van der Waals interactions, higher core hydrophobicity, additional networks of hydrogen bonds, secondary structure, ionic interactions, packing and length of surface loops, and proteins with high thermostability possesses various combinations of these forces (Berezovsky & Shakhnovich 2005). The general trend is that the interaction energies of molecules such as interaction energy between hydrophobic residues and aromatic residues increase towards higher growth temperature in mesophiles, thermophiles, and hyperthermophiles, and declines in psychrophiles (Goldstein 2007). In the analysis of mesophile protein and thermophile protein homologs, a general mechanism of thermostability was not found, thus suggesting that a small number of apparently strong molecule interactions cause the thermostability (Berezovsky & Shakhnovich 2005).

The variability of amino acid usage is caused by the guanine-cytosine (G + C) composition of DNA and organisms' optimal growth temperature (OGT) (Pasamontes & Garcia-Vallve 2006), but as stated above specific amino acid properties affecting the OGT has not been determined. However, a study suggests that thermophile specific DNA repair

system might be one factor that allows thermophiles to live in higher temperatures (Makarova et al. 2002).

In previous research, genome derived features, such as G + C composition, genome length and sequence, proteome derived features, and metabolic networks have been used to predict the OGT (Jensen et al. 2012, Sauer & Wang 2019, Weber Zendrera et al. 2019). Yet, none of these studies have been able to predict the OGT of psychrophiles reliably, because of the low number of psychrophiles in available data.

1.1.3 Growth temperature

Bacteria growth is usually modeled as three stage process that contains lag, exponential, and stationary phases and the parameters of the model are the numbers of cells, at the beginning, the maximum specific growth rate, and the beginning of the stationary phase. (Zwietering et al. 1990). Assuming that no other factor is limiting the growth, the OGT can be derived from the maximum specific growth rate. For example, if cells are cultured in different temperatures than their habitat temperature, the growth rate will change according to the number of modifications that cells have to undergo to adapt to the new environment (Zwietering et al. 1990, Buchanan et al. 1997).

The temperature being the only limiting factor is unreal and creating a culture that possesses exactly the same conditions as the environment is impossible, thus defined optical growth temperature is always slightly biased (Musto et al. 2006). If optical growth temperature cannot be defined by culturing, growth is defined by growth temperature minimum to growth temperature maximum or simply growth or no growth (Reimer et al. 2019). The vast majority of thermophiles and hyperthermophiles are not formally described and their characteristics are defined from DNA-sequence and sample collection sites (Hedlund et al. 2015).

1.2 Artificial intelligence: Machine learning and deep learning

1.2.1 Overview

Artificial intelligence (AI) is generally defined as a system that has the ability to understand external data correctly, learn from the data, and change its behavior according to it (Haenlein & Kaplan 2019). The proposed alternative definition to the whole field of AI

is: “the effort to automate intellectual tasks normally performed by humans” (Chollet 2018). Machine learning (ML) is a subset of AI and deep learning (DL) is a subfield of ML (Helm et al. 2020). Traditional programming is based on the concept that humans define rules of the program and the program outputs the answers according to the rules and in AI systems instead of predefined rules, the system interprets the data and given answers to produce rules which leads to given answers (Chollet 2018).

In genomics, and biology in general, data of interest is usually too complex to be investigated with simple statistical methods, hence ML algorithms are well suited for genomics (Eraslan et al. 2019). ML has been used in numerous genomic studies including cell DNA methylation state prediction (Angermueller et al. 2017), schizophrenia detection from mRNA expression levels (Zhu et al. 2021), and microRNA targets detection (Shuang Cheng et al. 2016). Because of the success of various applications in previous studies, it has been stated that ML will become a more important tool for genomics as large datasets become available through international collaborative projects (Libbrecht & Noble 2015).

ML systems are trained with data and they transform the data to a more meaningful representation of which is evaluated by prediction performance (Chollet 2018). ML algorithm can be interpreted as a process of searching a large space of candidate programs led by prediction performance (Jordan & Mitchell 2015). In ML, the learning of the system is defined automatic search process for better representations or finding a combination of model parameters that yields to best possible result (Chollet 2018). In statistics basis of the analysis is modeling, and phenomena are explained by estimating the values of parameters from the data, and the goodness of the model is evaluated usually by R^2-test and residual analysis (Breiman 2001a). If the data is high dimensional, this approach may lead to a large number of models that fit data acceptable, but perform badly in prediction tasks (Breiman 1996). In addition, in traditional statistics features of the data are usually selected or created manually and the outcome is heavily dependent on these features and relevant features may not be generated or selected by this approach, which may affect the overall modeling performance (Eraslan et al. 2019). ML systems and statistics use a lot of similar base models and they both can be evaluated by predictive accuracy, but statistics lack standards for comparison of models that are common in ML (Breiman 2001a).

ML algorithms can further divide into supervised learning, reinforcement learning, and unsupervised learning. In supervised learning, systems are trained to find rules why certain characteristics lead to specific results (Jordan & Mitchell 2015). Supervised systems

can either perform classification tasks or regression tasks in which results are continuous values. These systems use annotated training data which means that the true outcome or result of every sample used in training is known. In reinforcement learning, an algorithm chooses actions based on its environment to maximize a reward, for example, a system analyses customer feedback and outputs answers that maximize the review score (Chollet 2018). Unsupervised learning is used for finding unknown relationships between data points. This is also referred to as clustering where similar instances are grouped (Theodoridis & Koutroumbas 2009).

DL system is a ML algorithm that has multiple simple modules combined into a single model and these modules are often called layers of which purpose is to represent the data in a more meaningful way for the next layer (LeCun et al. 2015). The last layer is called an output layer, which yields the final result and the layers are connected to each other usually in sequence, but other architectures are also possible (Chollet 2018). Each layer has its operations to data and contribution of them to result are stored in layer weights and these are adjusted within training according to the whole performance of the model (Chollet 2018). The performance can be interpreted in many ways, but in the training of the DL system, this means measuring how far the final layer output is from the true value. An objective function is used as the distance score, which tells how the model has succeeded with a specific sample and the score is used as feedback to adjust the weights to the direction where the objective function obtains better results (LeCun et al. 2015). The objective function's value cannot be used directly to adjust the weights. To extract the contribution of each layer's weights, the most common way is to use a backpropagation algorithm of which is based on gradient descent (Zhang 2019). The gradient is a derivative of a multidimensional function, which can be further disassembled to a chain of layer operation derivatives that in DL reveal the contribution of each parameter had in loss (Chollet 2018).

1.2.2 Model selection

To measure how well the model generalizes, it must be tested with unseen data (Géron 2017). Thus, available data need to be divided into training set that is used in training, and test set that is used to evaluate system after training. Usually, some part of the training set is used as a validation set which is used to monitor the performance during the training (Chollet 2018). Assumptions for this procedure are that the training set and test set are reprehensive and large enough and the data is not redundant. A test set large enough is

capable to evaluate the system's true error (Varma & Simon 2006). If a system that appears to predict well on the training data fails to generalize to the test set, this is caused by overfitting. Overfitting indicates that the system rather memorizes the seen patterns than generalizes or the system is fitted to the noise in the data (Dietterich 1995). A model can also be underfitted which means that the used model is too simple to learn underlying complex data structure (Géron 2017).

The selection of the best ML algorithm can be seen as optimization (Cawley & Talbot 2010). The goal of optimization is to achieve better generalization on unseen data by finding optimal parameters of the system dealing with specific learning task (Bottou et al. 2018). Optimization can be divided into convex optimization problems and highly nonlinear and nonconvex problems and in machine learning optimization problems are often highly nonlinear and nonconvex which means that finding the global optimum of function is not guaranteed (Bottou et al. 2018).

Cross-validation (CV) is a common method for both parameter optimization and algorithm selection (Cawley & Talbot 2010). CV is especially recommended to be used if the number of samples is small (Varma & Simon 2006). In this method, data is partitioned randomly into non-overlapping *k*-folds that are as equal sized as possible and the performance of the system is evaluated with fold *i,* and the rest of the folds are used in training (Géron 2017). The system's generalization performance is derived from the mean performance of folds. The system's error can be divided into bias and variance. Bias describes the difference between the estimated value and unknown true generalization error and variance describes the variability of expected value due to the sampling of the data. If algorithm selection and parameter optimization are treated separately in CV, optimistic bias in performance, potentially in high magnitude, can be expected (Cawley & Talbot 2010). To estimate the system's generalization error reliable with small datasets, two nested CV loops are needed. This procedure is called nested CV in which the outer CV estimates the generalization error while the inner CV is used for parameter optimization leading to an almost unbiased estimate of the true error (Varma & Simon 2006).

Two common strategies for the search of system parameters are grid search CV and random grid search CV (Géron 2017). In grid search CV, a multidimensional array of all possible parameter combinations from selected parameters is created. Then all the parameter combinations are evaluated with all folds. In random grid search CV, the same multidimensional array is created, but instead of trying all possible values, only a selected

number of random parameters are tested. Random grid search CV is preferred when the number of possible parameter combinations is large (Géron 2017).

1.3 Classification methods

1.3.1 Overview

There is a huge number of different classification algorithms, but the goal of this chapter is to give an overview of supervised learning classification methods and their parameters that are related to or utilized in this thesis. Supervised classifiers use prior known information about the samples and in this case, it means that all training samples are labeled to some class (Theodoridis & Koutroumbas 2009). The simplest classification task is the binary classification where the classifier needs to distinguish two classes that are usually referred to as positive and negative or 1 and 0. In single label multiclass classification there are more than two possible classes and each sample can be set into one class (Chollet 2018).

The main metric to measure a performance of the classifier is accuracy; the fraction of samples that were correctly classified. Accuracy is a very simple method to evaluate performance. However, it may exaggerate the goodness of the classifier, if the dataset is imbalanced (Géron 2017).

A more informative way to measure the performance of the classifier is to evaluate it with the receiver operating characteristic (ROC) curve and the area under the curve (AUC). ROC curves can only be produced from binary class setting (Berthold et al. 2010). To extend ROC curve and AUC analysis to multiclass problem, each class need to be treated as a binary problem. This can be achieved in two ways. In the one-versus-all approach, a classifier for each class is trained where all classes are separately treated as a positive class and the rest as a negative class (Galar et al. 2011). Another way is to encode the classes to binary format, where true class equals 1 and others 0. This is called one-hot encoding because only one class equals 1 (hot) in the label vector (Chollet 2018). ROC curve expresses how true positive rate changes against false positive rate. True positive rate is the fraction of positive samples, or in other words, samples that actually belong to a particular class are correctly spotted by the classifier, and the false positive rate is the fraction of negative samples, or in other words samples that do not belong to a particular class are incorrectly declared as positive (Géron 2017). AUC can be used to compare different classifiers because larger AUC values indicate better performance of the

classifier on average (Theodoridis & Koutroumbas 2009). AUC is the probability that a randomly chosen positive sample will have a smaller estimated probability of belonging to a negative class than a randomly chosen negative sample (Hand & Till 2001). The value of AUC is between 0.5 and 1 where 0.5 corresponds to random guessing and 1 perfect classifier (Berthold et al. 2010).

Overall performance of the classifier in multiclass classification can be evaluated by micro averaging and macro averaging the AUC. Micro averaged AUC score gives equal weight to sample; it expresses an average over all the sample and class pairs and macro averaged AUC score gives equal weight to every class without taking account of its frequency, thus it is an unweighted mean of each class (Yang 1999).

1.3.2 Logistic Regression

Logistic regression (LG) is a generalized linear model and it is used to examine questions in which the dependent variable is binary or categorical (Tibshirani et al. 2015). LG is a nonparametric technique; it does not require any distributional assumptions (Osborne 2015). Prediction of the model is based on root level to conditional probabilities and odds. The dependent variable is transformed to logit which is the natural logarithm of the odds. This allows regression to use the logit link function, thus the LG equation is $Logit(\hat{Y}) = b_0 + b_1X_1...b_nX_n$. Here b_0 is the constant, b_1 is the coefficient of X_1, and b_n is the coefficient of X_n. When LG has more than two variables, the model estimates the unique effects of individual variables in the whole variable effect space of the equation (Osborne 2015). Typically, logistic models are fitted by maximizing a binomial log-likelihood of the data (Tibshirani et al. 2015).

Regularization is one way to avoid overfitting of logistic regression and it can be carried out with multiple techniques. Common regularization methods are Lasso regression (*l1* regularization) and Ridge regression (*l2* regularization). The *l1* regularization uses a penalty term for coefficients to shrink them or setting some of them to zero (Tibshirani et al. 2015). The *l2* regularization is very similar to *l1*, but the geometry of the *l2* optimization condition region is disk-like which prevents coefficients to be set to zero (Tibshirani et al. 2015). The size of the penalty term determines how much effect coefficients are allowed to have in the model in both *l1* and *l2* regularization.

1.3.3 Random Forests

Random Forest algorithms are ensemble techniques for classification and regression tasks in which a large number of individual decision trees are constructed based on a random sample and feature selection and their results are combined as the prediction (Breiman 2001b). Random Forest belongs to a large class of nonlinear classifiers. The goal of the algorithm is to find boundaries of feature space that separates the samples. The search of these boundaries in trees is performed via a sequence of decisions which are called nodes. The nodes represent a decision based on feature values for example "is the feature value $x > threshold$". The individual node's prediction is called a leaf. To select which features are used as nodes, the order of nodes and the threshold values algorithm need to be trained (Theodoridis & Koutroumbas 2009).

In theory, each node can consist infinite set of questions. If the threshold value is continuous, in practice though, only a finite number of questions can be considered. In order to decide the threshold, the goal is to find the best value that divides samples into homogenous or in decision tree terminology pure subsets compared to starting set of samples (Theodoridis & Koutroumbas 2009). A variety of purity measures has been defined and usage of them depends on the task. A decision tree does not have to use all the available features to declare a subset as the leaf. Usually, the purity of the subset is in a certain threshold is used as the stop splitting rule (Song & Lu 2015). When a node is declared as the leaf, it defines the outcome either as a class or continuous value. When designing a decision tree, it is important to take tree size into account. The tree needs to be large enough, but if it is too large, the tree tends to overfit (Theodoridis & Koutroumbas 2009). For example, in RandomForestClassifier implemented by Scikit-learn, tree size can be controlled with parameters *max_depth*, *min_sample_split*, *min_sample_leaf*, and *max_features* (Pedregosa et al. 2011).

1.3.4 Gradient tree boosting

Boosting is an approach to improve selected algorithm's performance by combining classifiers. However, boosting is conceptually different from ensemble methods. In boosting a series of systems are trained iteratively, that all use the same base system, but using a different subset of training set or different weighting over the samples of the training set (Theodoridis & Koutroumbas 2009). At each iteration, the computed weighting

distribution emphasizes the samples that the system performed poorly. The final system obtained is a weighted average of the previously trained systems.

As said in chapter 1.2.1 gradient is used to find the best set of weights to minimize or maximize the objective function. In traditional gradient optimization parameters of the system are adjusted with small steps, of which size the user determines as learning rate, to minimize or maximize the objective function (Géron 2017). Usually in ML, objective function is defined as loss function of which is tried to minimize. This optimization approach is called gradient descent. Gradient boosted tree algorithms leverage this idea which means that functions performance is evaluated by objective function which measures the difference between prediction and the target (Géron 2017). As traditional decision trees use purity as a measure of tree structure quality, the quality of gradient boosted trees is derived from a wider range of objective functions (Chen & Guestrin 2016).

XGBoost has been one of the most popular and successful gradient boosted tree systems in Kaggle ML competitions (Chen & Guestrin 2016). XGboost's success is based on objective function optimization which takes into account training loss and regularization of the complexity of the model (XGBoost developers 2020a). Overfitting of a model in XGboost can be controlled by controlling the model complexity with parameters *max_depth*, *min_child_weight*, and *gamma* or making the model more robust to noise which is controlled with parameters *lerning_rate* and *reg_lamda* (XGBoost developers 2020b).

1.3.5 Neural networks

Neural networks are multilayer architecture systems. Each layer consists of neurons or depending on terminology nodes. These neurons form a hidden layer. The first layer is called the input layer of which number of neurons defines the dimension of the input space and the last layer is called the output layer which computes as many predictions as desired output space has (Theodoridis & Koutroumbas 2009). For example, in four class classification problem, the output layer has four neurons.

The anatomy of a neural network can be decomposed into layers, objective function, and optimizer which determines how the gradient is used to change the parameters of the model (Chollet 2018). There are a vast number of different types of layers, neurons, objective functions, and optimizers, thus only relevant for this thesis are introduced below.

In fully connected layers each output neuron is connected to all previous and next hidden layer neurons (Liu et al. 2018). Usually, neurons of fully connected layers are perceptrons with a non-linear activation function. The fully connected layer receives a feature vector shaped according to the previous layer and then each perceptron is multiplied with an individual weight of which result bias term is added (Chollet 2018). This result is inputted to the activation function which outputs the result in a differentiable form (Géron 2017). In 2017 the most popular activation function was the rectified linear unit (ReLU) because it is efficient and easy to compute (LeCun et al. 2015).

Neural networks can be regularized with multiple approaches. One efficient technique is a dropout layer, which reduces overfitting and has been part of successful supervised learning tasks such as sequence and structure motif identification (Budach & Marsico 2018) and patient prognosis prediction from genes and pathways (Hao et al. 2018). The dropout layer removes units and their incoming and outgoing connections randomly from the system during training which prevents the units from excessive co-adaption (Srivastava et al. 2014).

Categorical cross entropy is a loss function used in multi-class classification and the loss minimizes the distance between output and true probability distributions (Chollet 2018). In supervised learning, cross entropy is the distance between a predicted distribution and label distribution. Entropy is a measure of uncertainty illustrated as a probability distribution and according to information theory, maximum entropy distribution makes the least assumptions about the data, thus leading to the least biased estimate on a given task (Jaynes 1957). Maximizing entropy of distribution is the same as minimizing cross entropy of distribution (Kern-Isberner 1998).

Momentum optimization is based on gradient descent, but it also takes account of previously computed gradients (Ruder 2016). In momentum optimization, the previous gradient accelerates the optimization by the user defined momentum term. The momentum algorithm is defined as $1.\ B\boldsymbol{m} - \eta \times \Delta_{u} J(u) \rightarrow \boldsymbol{m}$ $2.\ u + \boldsymbol{m} \rightarrow u$ (Géron 2017). Here B is momentum term, $\boldsymbol{m}$ is momentum vector, η is learning rate, u is the weights, $J(u)$ is the objective function, and $\Delta_{u} J(u)$ represents the gradient vector which contains all the partial derivatives of the cost function. The advantage of momentum optimization is that it is faster and escapes from plateaus easier than traditional gradient descent (Géron 2017).

1.4 Related work

The main objectives of previous related studies have been explaining the source of variability in prokaryote growth temperatures (Berezovsky & Shakhnovich 2005, Puigbò et al. 2008). Berezovsky and Shakhnovich suggest that thermostability is possible in prokaryotes living in high temperatures due to more compact and hydrophobic proteins than mesophilic prokaryotes. Only a handful of studies have tried to predict the OGT and they used metabolic network or genome derived features with linear models and regression for the prediction, rather than individual genes (Sauer & Wang 2019, Weber Zendrera et al. 2019). Sauer and Wang used multiple linear regression, thus the prediction was continuous value. Their model was evaluated with root mean squared error RMSE and R^2-test of the test set and the model performed well with unseen validation data sized 528 species (RMSE = 5.18 °C, R^2=0.758). However, a study by Jensen et al. (2012) predicting the OGT of bacteria, that is the most related to the prediction section of this thesis, contains optimistic bias in its performance; classifier was trained with 70 samples and predictive performance was evaluated only single test set that contained 25 samples. Their classifier got 76 % accuracy. In addition, the classification was only done for three temperature adaptation classes. On grounds of the above, additional research is needed to provide more information about adaptation mechanisms and individual genes affecting this phenomenon.

The COGs database has been previously used in an attempt to define thermophilic gene signature (Makarova et al. 2002). Makarova et al. used fully sequenced genomes from 12 hyperthermophile archaea and 2 hyperthermophile bacteria that were used to define conserved gene neighborhood linked to thermophilic adaptation. This gene neighborhood was suggested to be a thermophile specific DNA repair system. There was not a single gene present in all genomes that would explain the temperature adaptation. However, the majority of genomes had a group of five core COGs.

1.5 Research aims

The aims of this thesis are to leverage ML techniques for genomic analysis and use the developed pipeline to predict prokaryote growth temperature based on the presence and absence of individual genes, as well as to identify genes that affect the adaptation to certain environment temperature. This thesis uses NBCI's COGs database (Tatusov et al.

1997) that is a publicly available prokaryote genome dataset. In order to find reliable and unbiased results, I examine several different algorithms with *de facto* ML standards.

Hypotheses of this thesis are that prokaryote growth temperature can be predicted with reasonable reliability only from genetic data and a set of genes affecting temperature adaptation can be identified.

Growth temperature prediction and identification of genes responsible for temperature adaptation can lower the costs of empirical research and give useful insights for future research. Additionally, the techniques used in this thesis are applicable to other genomic datasets.

2 Materials and Methods

2.1 Data

Data used in this thesis was obtained from the COGs database (Galperin et al. 2015). During the data analysis stage of the thesis in summer 2019, the database from 2014 had 711 fully sequenced prokaryote genomes and 4631 COGs in total. These COGs covered 60-80% of species' proteome. The COGs are assigned into 26 functional categories, thus generally individual COG has a similar function among species, but there are instances that the same COG possesses different biological purposes in different species.

2.1.1 Data format

Data is provided in publicly available NBCI's file transfer system (FTP) as comma delimited files, tab delimited files, and FASTA format files. FTP contains necessary information of COGs including COGs functions, COGs functional categories, COG usage of species, protein sequence, protein identifier (accession number), and genome identifier (NBCI TaxId identifier) (Tatusov et al. 2000).

2.1.2 Preliminary data processing

Raw data processing was performed in either Python (version 3.7.3) or VI. Python environment was maintained with open-source package management system Conda (version 4.7.12).

Because of the research aim, obtaining COG usage from the data was the main objective of preliminary processing. COG usage was defined for every species and represented as a presence and absence matrix (Table 2). This was done with the Python package pandas (version 0.24.2) *pivot_table* function. After matrix creation temperature adaptations were annotated to 192 species according to previous research (Puigbò et al. 2008) and BacDive-database (Reimer et al. 2019). In total the annotated data contains 192 species (Figure 2, Figure 3); 87 mesophiles, 66 thermophiles, 26 hyperthermophiles, and 13 psychrophiles.

Species	COG1	COG2	COG3	COG4	COG5	COG6	COG7	COGN	Adaptation
P1	0	1	1	0	1	0	0	1	T
P2	0	0	1	0	1	0	1	1	H
P3	1	1	1	1	1	1	1	0	M
PN	1	0	1	0	0	0	0	1	M

Table 2. Visual representation of COG usage matrix. Rows represent the species, columns 1 to N represent the COG usage as 1 for present and 0 for absence and the last column represents the temperature adaptation class.

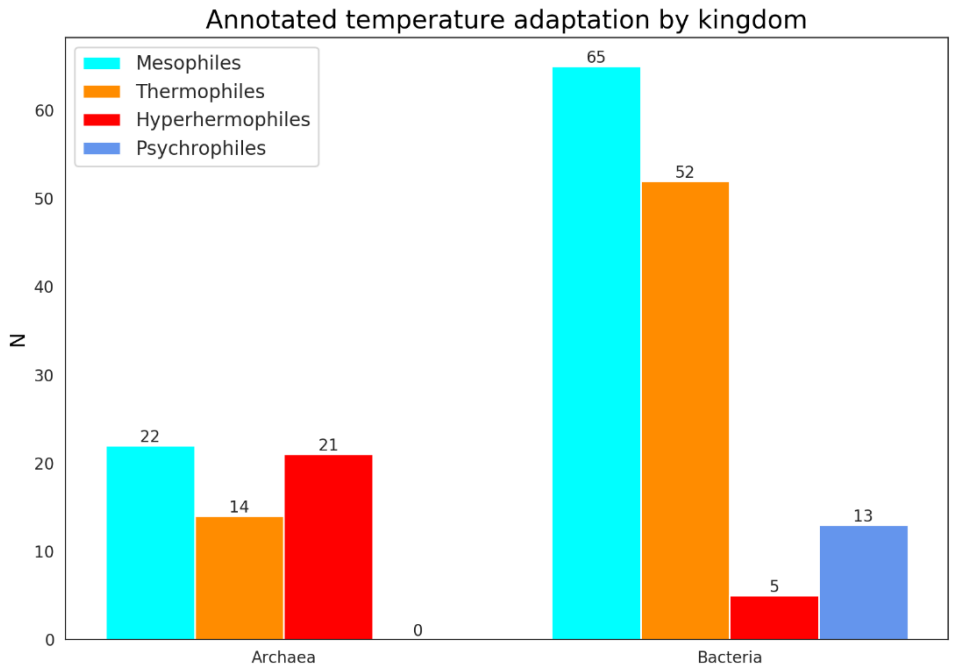


Figure 2. The number of annotated adaptations by the kingdom. Distributions of the adaptations are unbalanced between kingdoms and the adaptation distribution is especially uneven in the bacteria kingdom.

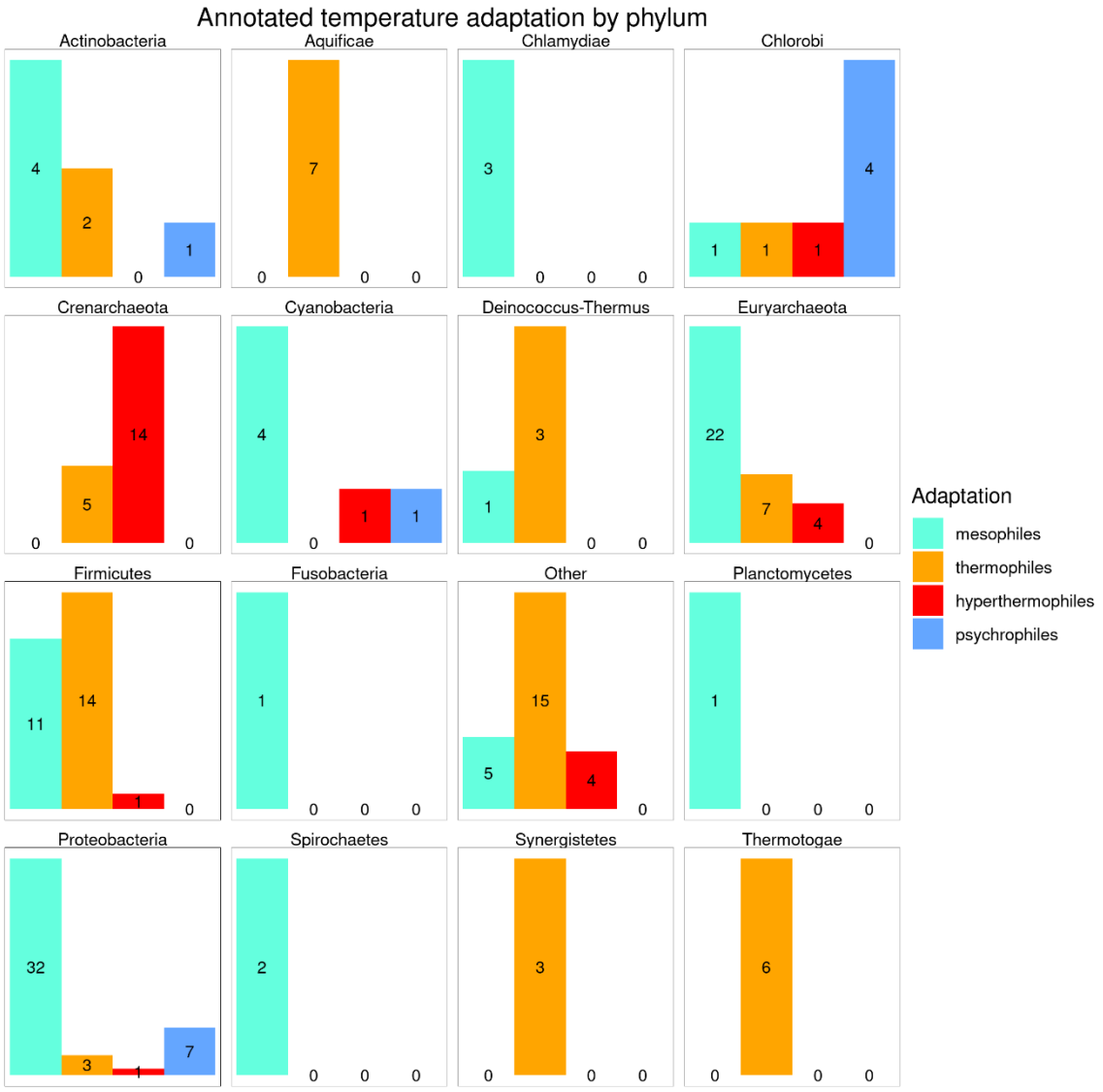


Figure 3. The number of annotated adaptations by phylum. Distributions of adaptations are unbalanced between phyla.

First, the Random Forest model was used to experiment with the data, which gave an insight into possible error sources. In the initial testing phase, the Random Forest classifier gave promising accuracy, but the temperature classification was mostly based on distinguishing archaea and bacteria kingdoms. The majority of the thermophilic prokaryotes in the data are archaea, which misleads the classifier to associate being a member of archaea kingdom also to have thermophilic adaptation, thus COGs in functional category J (Translation, ribosomal structure, and biogenesis) and additional 36 COGs were manually discarded (**Appendix 1**). Additionally, to avoid the overfitting of models, COGs that were present or absent in more than 90% of samples were removed. The final data matrix contained 2653 COGs.

2.2 Model selection

Four different classification algorithms, logistic regression, Random Forest, gradient boosting machine, and neural network performance were evaluated. Logistic regression and Random Forest were implemented with Scikit-learn Python package (version 0.21.3) modules LogisticRegression and RandomForestClassifier (Pedregosa et al. 2011). The gradient boosting machine was implemented with XGBoost's Python package (version 0.81) module XGBClassifier (Chen & Guestrin 2016). The neural network was implemented with the Tensorflow Python package (version 1.13.1) (Abadi et al. 2016).

Logistic regression, Random Forest, gradient boosting machine, and neural network classifiers were evaluated with 5-fold nested CV, which takes account of overfitting the model selection. Hyperparameters of the classifiers in the inner loop of nested CV were chosen by random grid search in Random Forest, gradient boosting machine, and neural network and by grid search in logistic regression. Detailed descriptions of tested hyperparameters and other settings are provided in **Appendix 2**. Nested CV leads to a low bias estimate of classifiers generalization performance. In nested CV training and model selection were done together in a manner that they were never separated.

2.2.1 Hyperparameter optimization and final model

Hyperparameter optimization was made with 5-fold CV and 1000 parameters were tested. The best parameter was chosen based on micro averaged AUC. The final predictive model parameter was set to the same as the best found in 5-fold CV and the final model was trained with all available annotated data. After the training, the model was used to predict the unannotated species.

2.3 Feature selection

The goal of the feature selection was to find models for each temperature adaptation that use as few features as possible and still sustain reasonable performance. Logistic regression classifier was trained with 1000 different regularization strengths and models were evaluated with 5-fold CV. CV results were plotted and regularization strengths for individual models were selected by visual interpretation. After the selection, four different models for each temperature adaptation class were trained with all available annotated data.

2.4 Phylogenetic tree

A phylogenetic tree was constructed from protein sequences of COGs that were present in 90 % of the species and the length of sequences was over 100 amino acids, resulting in 14 COGs in total. Some species had multiple entries of the same COG, thus only the longest sequences were used in the analysis.

First, sequences were aligned with MUSCLE (version 3.8.31), a program for creating multiple alignments of protein sequences (Edgar 2004). MUSCLE algorithm is based on *k*mer distance and Kimura distance metrics which allows the estimation of the evolutional relationships.

Second, the aligned sequences were further trimmed with Gblocks (version 0.91b), program for detecting and eliminating poorly aligned positions, with parameters *-t=p -b1=356 -b2=356 -b3=30 b4=5 -b5=h*. Parameter settings are an imitation of relaxed elimination settings (Talavera & Castresana 2007). The trimming is based on multiple rules: sequence parts selected for inclusion must not contain a large number of contiguous non-conserved positions, the flanks of the parts must be surrounded with highly conserved positions, and the parts need to be at least a certain minimum length (Castresana 2000). Despite information loss due to shortening the alignments, in most alignment conditions trimming the problematic regions leads to better trees (Talavera & Castresana 2007).

Finally, processed sequences were concatenated to a single file of which was used to build an approximately-maximum-likelihood phylogenetic tree with FastTree (Price et al. 2009, 2010) (version 2.1.10), with parameter *-pseudo* of which is recommended for highly gapped sequences (MicrobesOnline 2010). FastTree tree building can be summarized into four major components. First, an initial tree is built based on the neighbor joining heuristic variant of which distance metrics are derived from the sequence position frequency vector. In this phase, the preliminary tree topology is defined. Second, FastTree aims to reduce the length of the tree based on the balanced minimum evolution principle with a mixture of nearest neighbor interchanges and subtree prune regraft moves. The balanced minimum evolution principle relies theoretically on the proven concept of minimum evolution principle that the tree with the smallest sum of branch length estimates is the most likely to be the true tree (Rzhetsky & Nei 1993). In the balanced minimum evolution principle distance between branches are approximations. Third, the tree topology and branch length are optimized based on maximum-likelihood rearrangements.

Finally, the tree quality is estimated by estimating the reliability of each split in the tree with the Shimodaira-Hasegawa test. Visualization of the tree was done with the iTOL online tool (Letunic & Bork 2019).

3 Results

3.1 Model selection and hyperparameter optimization

Models performed with very similar classification accuracy (Figure 4), but logistic regression was chosen as an explanatory model in accordance with Occam's razor principle (Allaby 2010) and this algorithm had the best micro averaged AUC (Figure 5). Detailed figures of ROC curves and AUCs of classifiers are provided in **Appendix 3**.

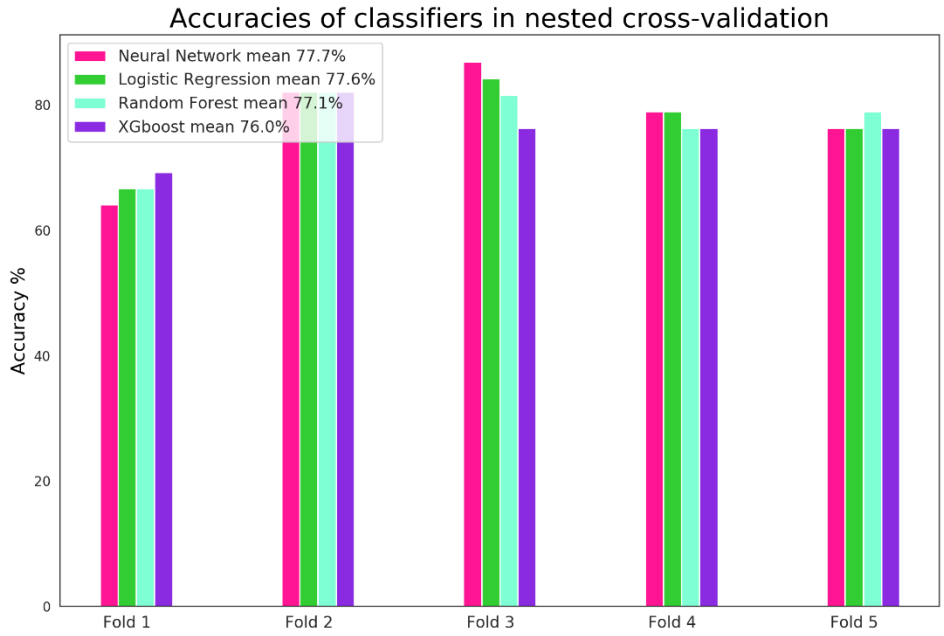


Figure 4. Accuracy of each classifier in different nested CV outer loop folds.

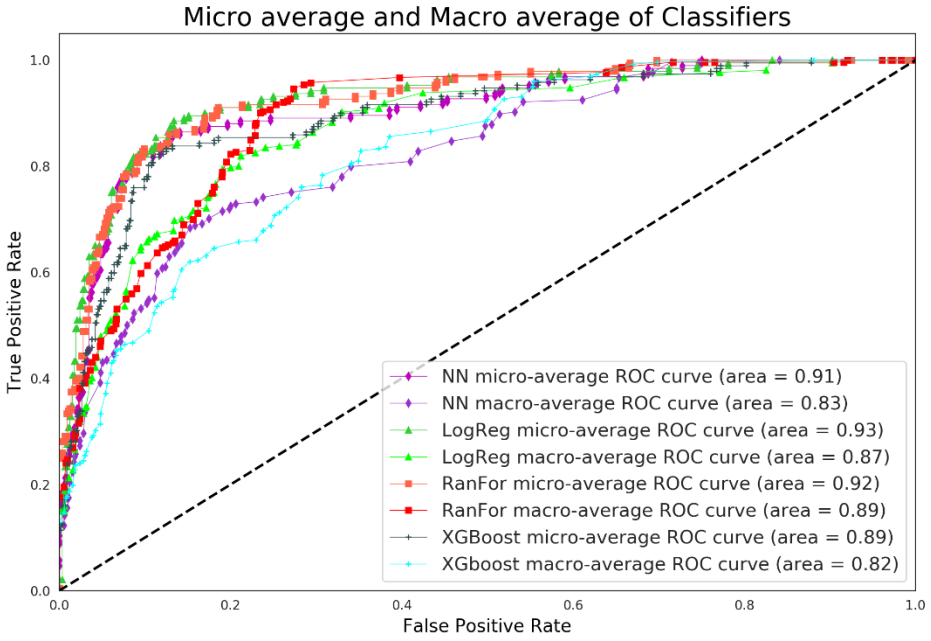


Figure 5. The micro average and the macro average of each classifier in nested CV outer loop folds. Logistic regression (LogReg) had the best micro averaged AUC.

Logistic regression with regularization strength 93.91 was selected as the final predictive model. This model performed the best in 5-fold CV and its micro averaged AUC was 0.936.

3.2 Prediction

The final predictive model used 342 COGs in prediction mesophiles, 342 COGs in prediction thermophiles, 278 COGs in the prediction of hyperthermophiles, and 208 COGs in the prediction of psychrophiles. Evolutionary relationships between annotated species and predicted species can be observed visually in Figure 6.

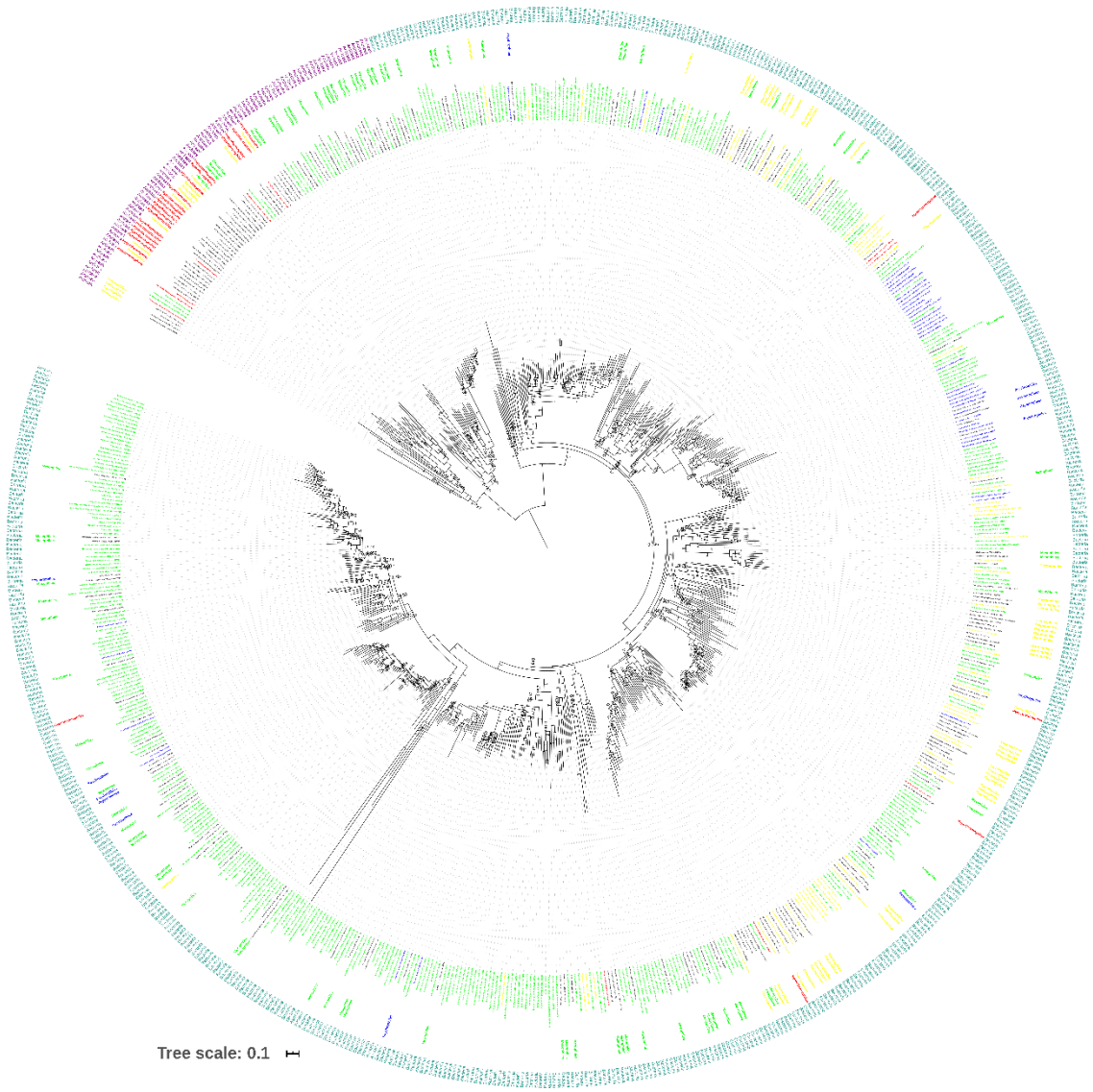


Figure 6. Approximately-maximum-likelihood phylogenetic tree and temperature adaptations. In total tree contains 711 species of which 519 are predicted and 192 annotated. Turquoise samples are bacteria, purple samples are archaea, red samples represent hyperthermophiles, yellow samples represent thermophiles, green samples represent mesophiles and blue samples represent psychrophiles. Annotated adaptations are in the outer layer of the circle and predicted adaptations are in the inner layer. The interactive tree can

be accessed from the ITOL website **http://itol.embl.de/shared/rakseli** under the project title "Master's thesis".

The most predicted adaptation was mesophile in both archaea and bacteria (Figure 7). In the prediction of archaea, none of the species was predicted as psychrophile or thermophile. In the prediction of Bacteria mesophile and thermophile were the most common predictions. Most of the predicted psychrophiles belong to the phylum *Chlorobi* (Figure 8). A table of predictions, kingdom, and phylum is provided in **Appendix 4**

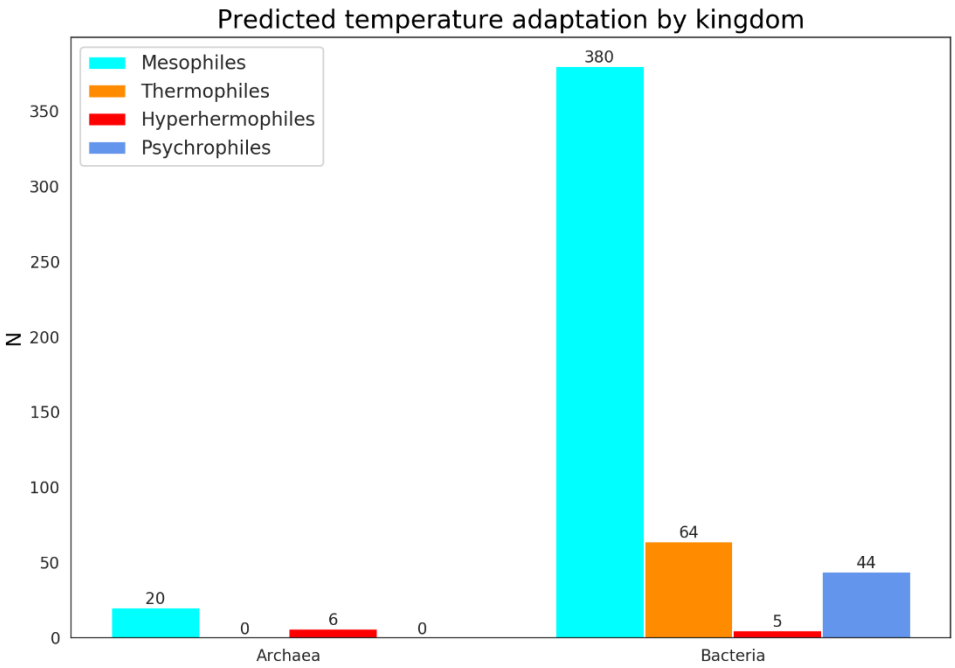


Figure 7. Predicted temperature adaptation count by the kingdom.

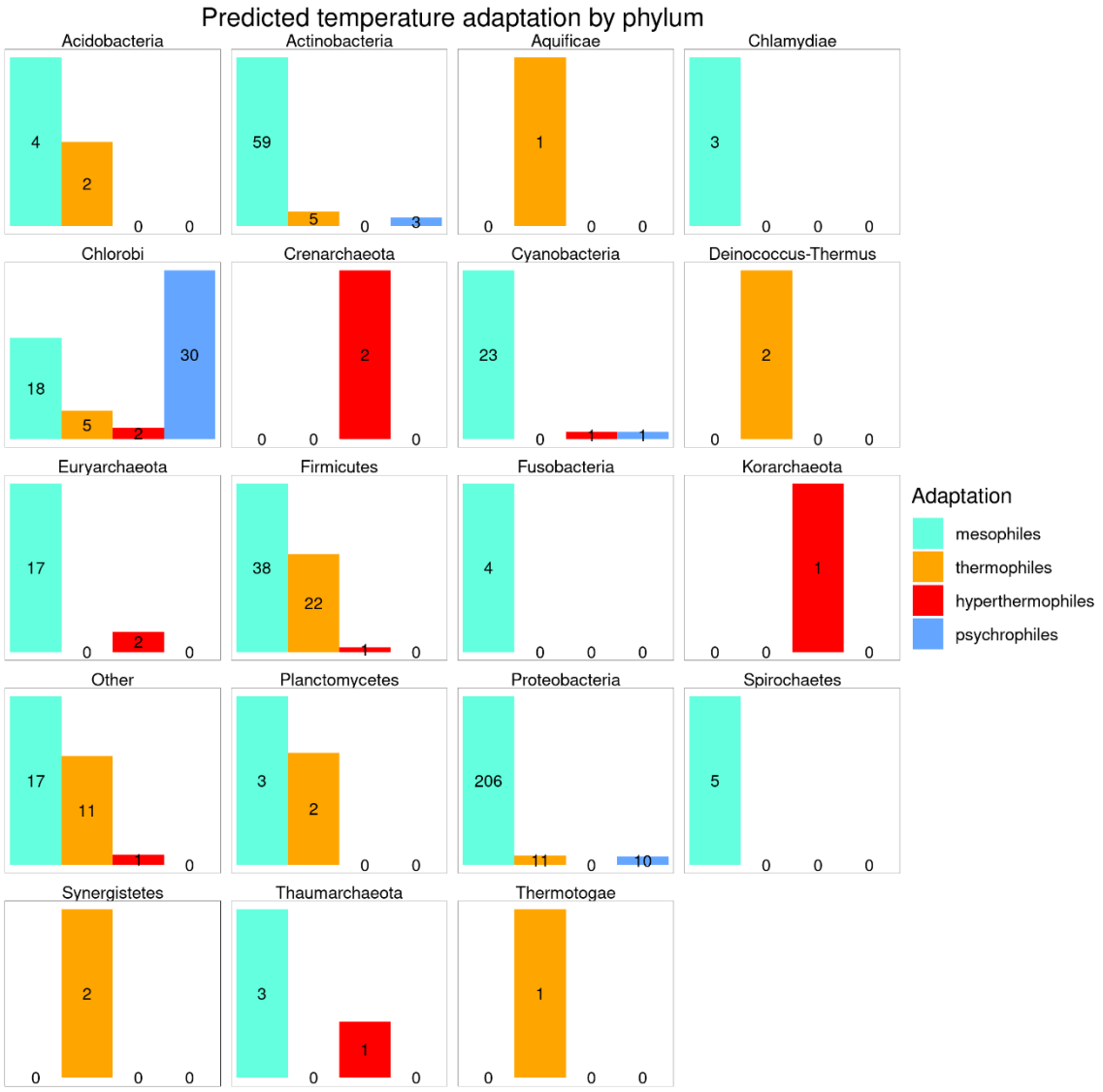


Figure 8. Predicted temperature adaptation count by phylum.

3.3 Feature selection

Chosen regularizations lowered the performance of the models by about 0.01 AUC score compared to best obtained AUCs (Figure 9). Regularization strengths for different adaptations were set for mesophiles to 15.01, for thermophiles to 0.41, for hyperthermophiles to 50, and psychrophiles to 75. Additional COGs increase the performance only a little after a certain threshold (Figure 10). These settings resulted in models where mesophiles can be predicted with 160 COGs, thermophiles can be predicted with 53 COGs, hyperthermophiles can be predicted with 187 COGs and psychrophiles can be predicted with 126 COGs. The number of COGs that models use is connected to the performance. If additional COGs increase the performance of the model, they may be added to the model. In total models had 78 mutual COGs. Usually, they were shared by two models, but in

few cases by three. Names, coefficients, and biological functions of the COGs are provided in **Appendix 5**.

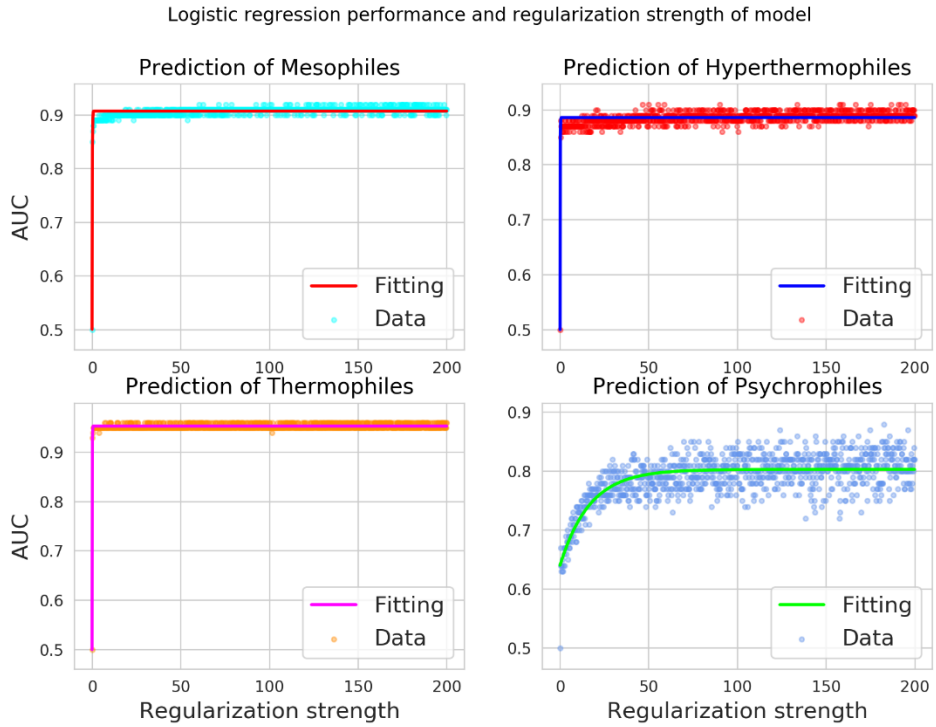


Figure 9. Logistic regression mean AUC in 5-fold CV and the strength of *l1* regularization.

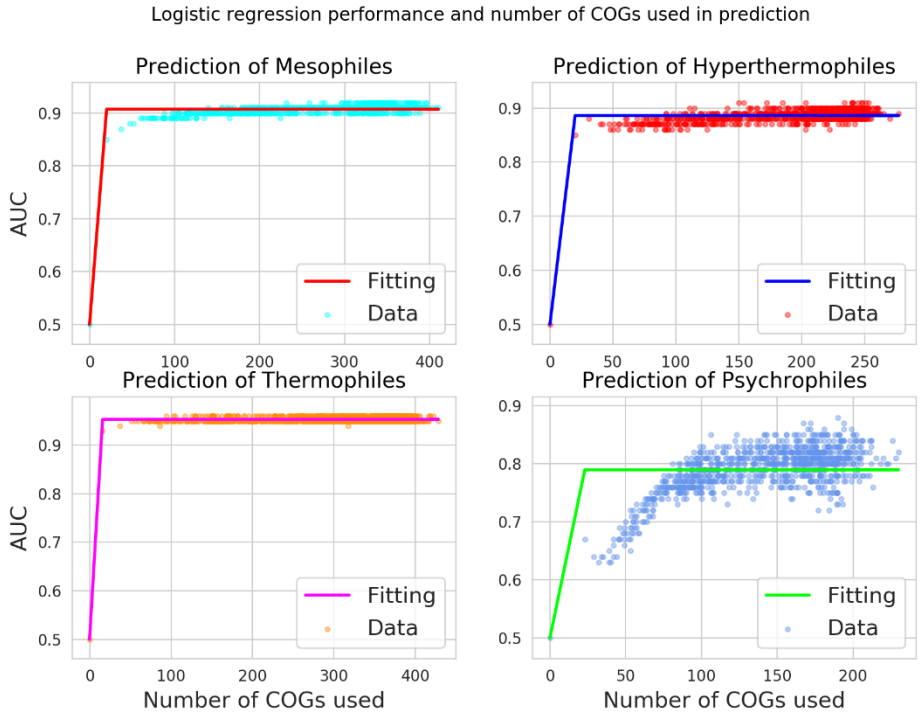


Figure 10. Logistic regression mean AUC in 5-fold CV and the mean number of used COGs.

There was only a little commonality between previously proposed thermophilic adaptation linked COGs (Makarova et al. 2002) and coefficients that degree the logistic regression outcome (Table 3).

COGs	**Mesophile coefficients**	**Thermophile coefficients**	**Hyperthermophile coefficients**	**Psychrophile coefficients**
COG1857	-1.47	0.0	0.0	0.0
COG1688	0.0	0.0	0.0	0.0
COG1203	-0.53	0.0	0.0	0.0
COG1468	0.0	0.0	0.0	0.0
COG1518	-0.22	0.0	0.0	0.0
COG2254	0.0	0.0	0.0	0.0
COG3578	0.0	0.0	0.0	0.0
COG1353	0.0	0.0	0.0	0.0
COG2462	0.0	0.0	0.0	0.0
COG1769	0.0	0.0	0.0	0.0
COG1583	-0.41	1.2	0.0	0.0
COG1567	0.0	0.0	0.0	0.0
COG1336	0.0	0.0	0.0	0.0
COG1367	0.0	0.0	0.0	0.0
COG1604	0.0	0.0	0.0	0.0
COG1337	0.0	0.0	0.0	0.0
COG1332	0.0	0.0	0.0	0.0
COG3337	0.0	0.0	0.0	0.0
COG1517	0.0	0.0	0.0	0.0
COG3574	0.0	0.0	0.0	0.0
COG1343	0.0	0.0	0.0	0.0
COG1421	0.0	0.0	0.0	0.0
COG3649	0.0	0.0	0.0	0.0
COG3512	0.0	0.0	0.0	0.0
COG3513	0.0	0.0	0.0	0.0

Table 3. The COGs suggested by previous research constitutes predicted thermophile-specific DNA repair system and logistic regression coefficients. The proposed core COGs are marked in bold. In modeling of mesophiles, thermophile specific COGs 1857, 1203,

1518, and 1583 are negative which represents an inverse relationship. In modeling thermophiles, the coefficient of thermophile specific COG1583 is positive.

4 Discussion

4.1 Main results

This research shows that machine learning can be leveraged to predict prokaryote temperature adaptation with micro averaged AUC 0.93 and only a small number of COGs are needed for the prediction of each class. These results show that adaptation of uncultivated prokaryotes can be predicted which gives new tools for metagenomic data analysis (Parks et al. 2017). Growth temperature has been predicted successfully previously from genome derived features (Sauer & Wang 2019), thus the COG usage and machine learning provide an additional method for this task. The COGs with the highest coefficients of the logistic regression model may be linked to biological processes that affect temperature adaptation. This research did not provide much additional support for previously proposed thermophile specific COGs (Makarova et al. 2002), but some kind of link between them can be observed; COG1857, COG1203, COG1518, and COG1583 have negative coefficient in the model that predicts mesophiles and COG1583 has positive coefficient in the model that predicts thermophiles. The COG database has not been used much in growth temperature prediction tasks. However, the database has been used in the functional analysis of microbial communities. Tringe et al. (2005) used the database to identify environment specific gene fingerprints of soil, sea surface water, and deep sea. Antunes et al. (2016) used the database to identify the functional profile of high temperature compost microbes by analyzing the relative abundance of CDSs and corresponding COG category.

Prediction of adaptation is quite consistent with the adaptation of closely related species. However, in multiple cases, loss of adaptation to high temperature can be seen in the phylogenetic tree such as found in previous research (Puigbò et al. 2008). *Methylococcus capsulatus* is annotated as hyperthermophile, but *Methylomonas methanica* and *Methylomicrobium alcaliphilum* from the same node are predicted as mesophiles. In addition, variability of predicted adaptation can be observed inside clades.

Adaptation distributions between annotated and predicted species dived by kingdom are visually similar; most of the predicted and annotated species are mesophiles. The majority of bacteria and archaea inhabit deep oceanic subsurface (4×10^{29} of total 1×10^{30} bacterial

and archaeal cell numbers on earth) (Flemming & Wuertz 2019), thus mesophile abundance in data may be caused by easier culturing or abundance in more accessible locations. Prediction distribution of bacteria is more consistent with annotation distribution of bacteria than prediction distribution of archaea compared to annotation distribution of archaea. This may be caused by the small number of archaea in the dataset.

Adaptation distributions between annotated and predicted species divided by phylum provide additional details from the function of the classifier. Phylum *Chlorobi* contains all adaptations in annotations and predictions. In addition, most of the annotated *Chlorobi* are psychrophiles which may be the reason why most of predicted *Chlorobi* are also psychrophiles. Annotated data did not contain any *Acidobacteria*, *Korarchaeota,* or *Thaumarchaeota*, but classifier predicted also other adaptation than mesophile to these phyla which can be seen as successful training of classifier; it is able to predict unseen phyla's adaptation versatilely not just the most common adaptation. In general, distributions are visually alike, suggesting that phylum foretell possible temperature adaptation of species.

4.2 Possible error sources

There are some factors that may be important in the prediction of temperature adaptation that this research and used techniques possibly did not take into account. The main challenge of this research was to cope with a small dataset; it is hard to recognize meaningful patterns from the low number of hyperthermophiles, thus this had to been taken into account by preventing classifiers to predict all archaea to be in this class. The small and unbalanced dataset also may have affected the selection of the classifier, because complex dependencies are hard to find from this kind of data, thus the simplest model had the best performance.

Also, differences in classification performance may have been due to the different number of hyperparameters between classifiers. For example, only one hyperparameter of the logistic regression classifier was tuned which gave a more reliable picture of its performance compared to six hyperparameters of neural network classifier that had 972 different parameter combinations of which only 50 were tested. Unfortunately, exhaustive parameter search is very rarely possible. For example, in this research 5-fold nested CV was used, so the inner loop included training the model 250 times. With 972 parameters, the model would have been trained 4860 times making computation significantly heavier.

Other critical points to consider are growth temperature definition and temperature variability of adaptation classes (Table 1). This may have caused errors in borderline cases, where species could belong to multiple classes, but not have been cultured or were annotated by different standards. This empathizes the importance of global cooperation with researchers and institutes.

4.3 Possible improvements

Machine learning is a fairly young field of which development is rapid. This research could be further improved, but as a master's thesis is defined to be certain extent all possible options could not be fitted in.

The most effective improvement would be to increase the sample size. This would provide more reliable results as distributions of the temperature classes and kingdoms are uneven. Also, additional features such as codon usage, amino acid composition, and environmental factors such as pH and salinity that previous research leveraged for growth temperature related analyses (Puigbò et al. 2008, Lecocq et al. 2021) could have increased the classification performance.

One of the most reliable ways to evaluate the classification performance is to use a leave-one-out CV where each fold consists only from one sample providing an almost unbiased estimate of true generalization performance (Cawley & Talbot 2010). Unfortunately, this was not possible due to computationally heavy models.

Another approach to predict growth temperature could have been a regression model that was used successfully previously (Sauer & Wang 2019). With this technique, the prediction is a continuous value that may have been more suitable for the growth temperature prediction of class borderline species.

4.4 Future studies

Disruption of the gene allows to determine the outcome of loss of gene function (Giaever et al. 2002), thus results of this research can be used as a basis for knockout analysis. Although the statistical significance of the highest coefficients of the explanatory models cannot be tested, they still may indicate actual biological function.

ML is a good technique to find patterns. One possible application for it could be an identification of genes that Last Common Ancestor (LUCA) possessed. LUCA has been

proposed to be a relatively complex organism that was a moderate thermophile or a thermotolerant mesophile (Glansdorff et al. 2008). Findings that mitochondria are bacterial origin and found in eukaryote common ancestor, and the tendency of eukaryotes to branch within archaeal lineages indicates that eukaryotes arose from prokaryotes and genes that trace to the common ancestor of archaea and bacteria trace to LUCA (Weiss et al. 2018). This relationship can be used in supersized learning to predict ancestral genes from ancestor descendent gene pairs.

Natural language or text data shares a lot of commonalities with gene data, thus techniques used in natural language processing (NLP) may provide a new perspective for genetics. One interesting approach could be to utilize word embeddings that are commonly used in NLP tasks. Word embeddings represent words as vectors based on their contexts in a large corpus, thus this technique could be used to represent nucleotide codons as vectors based on their context in the genome. Word embeddings are able to capture semantic and syntactic information of the words (Liu et al. 2015), hence provide more insight into why certain genes function in some way.

Another interesting technique to analyze genetic data could be to leverage long short-term memory (LSTM) networks. LSTM networks are good for sequential data where the order of events matter (Greff et al. 2017). This feature is critical when analyzing raw genetic data; the order of the codons defines the function of the gene and simpler approaches may not catch this.

ML and DL are widely utilized in current bioinformatics and are likely to become dominant in forthcoming research projects. These methods can capture relationships that are impossible to find with other techniques. The complexity of genetics and life is massive, thus more and more comprehensive techniques need to be put into operation. However, the interpretability of ML and DL models is low which makes analysis of model function hard (Hagenbuchner 2020). In conclusion, perfect systems do not exist, thus research must continue!

5 Acknowledgments

I would like to thank my supervisors Pere Puigbò, Antti Airola, and Manu Tamminen for their extensive support and patience. I thank Turku Collegium for Science and Medicine for funding my internship and donating a computer for analysis in summer 2019.

6 References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D., Steiner, B., Tucker, F., Vasudevan, V., Warden, F., Wicke, M., Yu, Y., Zheng, X., & Brain, G. (2016, November). *TensorFlow: A System for Large-Scale' ' Machine Learning*. Presented at the 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA.

Allaby, M. (2010). *A dictionary of ecology* (4th ed.). Oxford University Press.

Angermueller, C., Lee, H. J., Reik, W., & Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, *18*(1), 67.

Antunes, L.P., Martins, L.F., Pereira, R.V., Thomas, A.M., Barbosa, D., Lemos, L.N., Silva, G.M.M., Moura, L.M.S., Epamino, G.W.C., Digiampietri, L.A., Lombardi, K.C., Ramos, P.L., Quaggio, R.B., de Oliveira, J.C.F., Pascon, R.C., Cruz, J.B. da, da Silva, A.M., & Setubal, J.C. (2016). Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Scientific Reports*, *6*, 38915.

Berezovsky, I. N., & Shakhnovich, E. I. (2005). Physics and evolution of thermophilic adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(36), 12742–12747.

Berthold, M. R., Borgelt, C., Höppner, F., & Klawonn, F. (2010). *Guide to Intelligent Data Analysis* (p. 98). London: Springer London.

Bertrand, J.-C., Normand, P., Ollivier, B., & Sime-Ngando, T. (2018). *Prokaryotes and Evolution*. (Jean-Claude, P. Normand, B. Ollivier, & T. Sime-Ngando, eds.) (p. 5,6,9,115). Cham: Springer International Publishing.

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, *60*(2), 223–311.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*(6), 2350–2383.

Breiman, L. (2001a). Random Forests. *Machine learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231.

Buchanan, R. L., Whiting, R. C., & Damert, W. C. (1997). When is simple good enough: a comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves. *Food microbiology*, *14*(4), 313–326.

Budach, S., & Marsico, A. (2018). pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics*, *34*(17), 3035–3037.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552.

Cawley, G. C., & Talbot, N. L. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research : JMLR*, *11*.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD' ' '16* (pp. 785–794). New York, New York, USA: ACM Press.

Chollet, F. (2018). *Deep Learning With Python* (1st ed., pp. 4–11,51,52,58,70,79,84,95). Shelter Island, New York: Manning Publications.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27*(3), 326–327.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews. Genetics*, *20*(7), 389–403.

Flemming, H.-C., & Wuertz, S. (2019). Bacteria and archaea on Earth and their abundance in biofilms. *Nature Reviews. Microbiology*, *17*(4), 247–260. doi:10.1038/s41579-019-0158-9

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern recognition*, *44*(8), 1761–1776.

Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2019). Microbial genome analysis: the COG approach. *Briefings in Bioinformatics*, *20*(4), 1063–1070.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, *43*(Database issue), D261–9.

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, *49*(D1), D274–D281.

Géron, A. (2017). *Hands-on Machine Learning With Scikit-learn And Tensorflow: Concepts, Tools, And Techniques To Build Intelligent Systems* (First., pp. 52,79,119,150–161,249,327,368). Beijing: O'reilly Media.

Gevers, D., Vandepoele, K., Simillon, C., & Van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends in Microbiology*, *12*(4), 148–154.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, AP., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Campanoro, S., Curtiss, M., Deutchbauer, A., Entian, KD., Flaherty, P., Foury, F., Garfinkel, DJ., Gerstein, M., Gotte, D., Gűldener, U., Hegemann, JH., Hempel, S., Herman, Z., Jaramillo, DF., Kelly, DE., Kelly, SL., Kötter, P., LaBonte, D., Lamb, DC., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, SL., Revuelta, JL., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, DD., Sookai-Mahadeo, S., Storms, RK.,

Strathern, JN., Valle, G., Voet, M., Volckaert, G., Wang, CY., Ward, TR., Wilhermy, J., Winzeler, EA., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, JD., Snyder, M., Philippsen, P., Davis, & RW., Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, *418*(6896), 387–391.

Glansdorff, N., Xu, Y., & Labedan, B. (2008). The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biology Direct*, *3*, 29.

Goldstein, R. A. (2007). Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: insights from the quasi-chemical approximation. *Protein Science*, *16*(9), 1887–1895.

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222–2232.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. *California management review*, *61*(4), 5–14.

Hagenbuchner, M. (2020). The black box problem of AI in oncology. *Journal of Physics: Conference Series*, *1662*, 012012.

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine learning*, (45), 171–186.

Hao, J., Kim, Y., Kim, T.-K., & Kang, M. (2018). PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinformatics*, *19*(1), 510.

Hedlund, B. P., Murugapiran, S. K., Alba, T. W., Levy, A., Dodsworth, J. A., Goertz, G. B., Ivanova, N., & Woyke, T. (2015). Uncultivated thermophiles: current status and spotlight on “Aigarchaeota”. *Current Opinion in Microbiology*, *25*, 136–145.

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, AI., Ramkumar, P. N. (2020). Machine learning and artificial

intelligence: definitions, applications, and future directions. *Current reviews in musculoskeletal medicine*, *13*(1), 69–76.

Hibbing, M. E., Fuqua, C., Parsek, M. R., & Peterson, S. B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews. Microbiology*, *8*(1), 15–25.

Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C., & Stetter, K. O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, *417*(6884), 63–67.

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *106*(4), 620–630.

Jensen, D. B., Vesth, T. C., Hallin, P. F., Pedersen, A. G., & Ussery, D. W. (2012). Bayesian prediction of bacterial growth temperature range based on genome sequences. *BMC Genomics*, *13 Suppl 7*, S3.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., & Crook, D. W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiology Reviews*, *33*(2), 376–393.

Kellner, S., Spang, A., Offre, P., Szöllősi, G. J., Petitjean, C., & Williams, T. A. (2018). Genome size evolution in the Archaea. *Emerging topics in life sciences*, *2*(4), 595–605.

Kern-Isberner, G. (1998). Characterizing the principle of minimum cross-entropy within a conditional-logical framework. *Artificial intelligence*, *98*(1-2), 169–208.

Koonin, E. V., & Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, *36*(21), 6688–6719.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., & Koonin, E. V. (2011). Computational methods for Gene Orthology inference. *Briefings in Bioinformatics*, *12*(5), 379–391.

Landenmark, H. K. E., Forgan, D. H., & Cockell, C. S. (2015). An estimate of the total DNA in the biosphere. *PLoS Biology*, *13*(6), e1002168.

Lecocq, M., Groussin, M., Gouy, M., & Brochier-Armanet, C. (2021). The molecular determinants of thermoadaptation: methanococcales as a case study. *Molecular Biology and Evolution*, *38*(5), 1761–1776.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Letunic, I., & Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews. Genetics*, *16*(6), 321–332.

Liu, K., Kang, G., Zhang, N., & Hou, B. (2018). Breast Cancer Classification Based on Fully-Connected Layer First Convolutional Neural Networks. *IEEE access : practical innovations, open solutions*, *6*, 23722–23732.

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). Topical Word Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29). Association for the Advancement of Artificial Intelligence.

López-García, P., Zivanovic, Y., Deschamps, P., & Moreira, D. (2015). Bacterial gene import and mesophilic adaptation in archaea. *Nature Reviews. Microbiology*, *13*(7), 447–456.

Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B., & Koonin, E. V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Research*, *30*(2), 482–496.

Makarova, K. S., Sorokin, A. V., Novichkov, P. S., Wolf, Y. I., & Koonin, E. V. (2007). Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biology Direct*, *2*, 33.

Martin, W., & Koonin, E. V. (2006). A positive definition of prokaryotes. *Nature*, *442*(7105), 868.

MicrobesOnline. (2010). FastTree 2.1 documentation. Retrieved April 16, 2020, from http://www.microbesonline.org/fasttree/

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valín, F., & Bernardi, G. (2006). Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochemical and Biophysical Research Communications*, *347*(1), 1–3.

Osborne, J. W. (2015). *Best practices in logistic regression* (p. 8,25,26,27,28,29,30,31,246). 1 Oliver's Yard, 55 City Road London EC1Y 1SP : SAGE Publications, Ltd.

Osbourn, A. E., & Field, B. (2009). Operons. *Cellular and Molecular Life Sciences*, *66*(23), 3755–3775.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, *2*(11), 1533–1542.

Pasamontes, A., & Garcia-Vallve, S. (2006). Use of a multi-way method to analyze the amino acid composition of a conserved group of orthologous proteins in prokaryotes. *BMC Bioinformatics*, *7*, 257.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Dobourh, V., Vanderplas, J., Passos, A., Cournapeua, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*.

Pedrós-Alió, C., & Manrubia, S. (2016). The vast unknown microbial biosphere. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(24), 6585–6587.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, *26*(7), 1641–1650.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 — approximately maximum-likelihood trees for large alignments. *Plos One*, *5*(3), e9490.

Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I., & Koonin, E. V. (2014). Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biology*, *12*, 66.

Puigbò, P., Pasamontes, A., & Garcia-Vallve, S. (2008). Gaining and losing the thermophilic adaptation in prokaryotes. *Trends in Genetics*, *24*(1), 10–14.

Reimer, L. C., Vetcininova, A., Carbasse, J. S., Söhngen, C., Gleim, D., Ebeling, C., & Overmann, J. (2019). BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Research*, *47*(D1), D631–D636.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv*.

Rzhetsky, A., & Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, *10*(5), 1073–1095.

Sauer, D. B., & Wang, D.-N. (2019). Predicting the optimal growth temperatures of prokaryotes using only genome derived features. *Bioinformatics*, *35*(18), 3224–3231.

Shuang Cheng, Maozu Guo, Chunyu Wang, Xiaoyan Liu, Yang Liu, & Xuejian Wu. (2016). MiRTDL: A Deep Learning Approach for miRNA Target Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(6), 1161–1169.

Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135.

Sonnhammer, E. L. ., & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, *18*(12), 619–620.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research : JMLR*, *15*, 1929–1958.

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577.

Tatusov, R L, Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, *28*(1), 33–36.

Tatusov, R L, Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, *278*(5338), 631–637.

Tatusov, Roman L, Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, DM., Mazumber, R., Mekhedow, SL., Nikolskaya, AN., Roa, BS., Smirnov, S., Sverdlov, AV., Vasudevan, S., Wolf, YI., Yin, JJ., & Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, *4*, 41.

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition* (4th ed., pp. 7,8,9,156,276,215,216–220,231,595). Elsevier.

Tibshirani, R., Wainwright, M., & Hastie, T. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations* (p. 7,8,12,29,30,38). Chapman and Hall/CRC.

Torsvik, V., Øvreås, L., & Thingstad, T. F. (2002). Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science*, *296*(5570), 1064–1066.

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P., & Rubin, E. M. (2005). Comparative metagenomics of microbial communities. *Science*, *308*(5721), 554–557.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, 91.

Weber Zendrera, A., Sokolovska, N., & Soula, H. A. (2019). Robust structure measures of metabolic networks that predict prokaryotic optimal growth temperature. *BMC Bioinformatics*, *20*(1), 499.

Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V., & Martin, W. F. (2018). The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genetics*, *14*(8), e1007518.

Whitman, W B, Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(12), 6578–6583.

Whitman, William B. (2009). The modern concept of the procaryote. *Journal of Bacteriology*, *191*(7), 2000–5; discussion 2006.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090.

XGBoost developers. (2020a). Introduction to Boosted Trees. Retrieved March 25, 2020, from https://xgboost.readthedocs.io/en/latest/tutorials/model.html

XGBoost developers. (2020b). Notes on Parameter Tuning. Retrieved September 11, 2020, from https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html

Yang, Y. (1999). An Evaluation of Statistical Approachesto Text Categorization. *Information retrieval*, *1*, 69–90.

Zhang, J. (2019). Gradient Descent based Optimization Algorithms for Deep Learning Models Training. *arXiv*.

Zhu, L., Wu, X., Xu, B., Zhao, Z., Yang, J., Long, J., & Su, L. (2021). The machine learning algorithm for the diagnosis of schizophrenia on the basis of gene expression in peripheral blood. *Neuroscience Letters*, *745*, 135596.

Zwietering, M. H., Jongenburger, I., Rombouts, F. M., & van 't Riet, K. (1990). Modeling of the bacterial growth curve. *Applied and Environmental Microbiology*, *56*(6), 1875–1881.

Appendices

Appendi*x* 1. Manually discarded COGs, their functional category, protein function, and discard reason.

COGs	Functional Category	Protein function	Discard reason
COG1581	K	Archaeal DNA-binding protein	only in Archaea
COG0691	O	tmRNA-binding protein	only in Bacteria
COG1197	LK	Transcription-repair coupling factor (superfamily II helicase)	only in *Ba*cteria
COG0669	H	Phosphopantetheine adenylyltransferase	only in Bacteria
COG0690	U	Preprotein translocase subunit SecE	only in Bacteria
COG1311	L	Archaeal DNA polymerase II, small subunit/DNA polymerase delta, subunit B	only in three Bacteria
COG0305	L	Replicative DNA helicase	almost in all Bacteria, only in two Archaea
COG1602	S	Uncharacterized protein	only in Archaea
COG0266	L	Formamidopyrimidine-DNA glycosylase	almost in all Bacteria, only in four Archaea
COG0353	L	Recombinational DNA repair protein RecR	only in Bacteria
COG0536	DL	GTPase involved in cell partioning and DNA repair	only in Bacteria
COG0587	L	DNA polymerase III, alpha subunit	only in one Archaea, almost all Bacteria
COG0593	L	Chromosomal replication initiation ATPase DnaA	only in Bacteria
COG0629	L	Single-stranded DNA-binding protein	only in Bacteria
COG0692	L	Uracil DNA glycosylase	only in one Archean
COG0749	L	DNA polymerase I - 3'-5' exonuclease and polymerase domains	only in two archaea, almost all Bacteria
COG0776	L	Bacterial nucleoid DNA-binding protein	only in six Archaea, almost all Bacteria

COG0817	L	Holliday junction resolvasome RuvABC endonuclease subunit, few archaea	only in three Archaea, almost all Bacteria
COG1107	L	Archaea-specific RecJ-like exonuclease, contains DnaJ-type Zn finger domain	almost in all Archaea, only in six Bacteria
COG1200	L	RecG-like helicase	only in Bacteria
COG1202	L	Superfamily II helicase, archaea-specific	only in Archaea
COG1241	L	DNA replicative helicase MCM subunit Mcm2, Cdc46/Mcm family	only in one Bacteria, almost in all Archaea
COG1381	L	Recombinational DNA repair protein (RecF pathway)	only in Bacteria
COG1389	L	DNA topoisomerase VI, subunit B	almost in all Archaea, only seven Bacteria
COG1466	L	DNA polymerase III, delta subunit	only in Bacteria
COG1467	L	Eukaryotic-type DNA primase, catalytic (small) subunit	only in two Bacteria, almost in all Archaea
COG1591	L	Holliday junction resolvase, archaeal type	almost in all Archaea, only in three Bacteria
COG1599	L	ssDNA-binding replication factor A, large subunit	almost in all Archaea, only in two Bacteria
*COG*1630	L	NurA 5'-3' nuclease	only in six Bacteria
COG1697	L	DNA topoisomerase VI, subunit A	almost in all Archaea, only eight Bacteria
COG1711	L	DNA replication initiation complex subunit, GINS family	almost in all Archaea, only in three Bacteria
COG1860	FL	Uncharacterized conserved protein, UPF0179 family	only in Archaea
COG2255	L	Holliday junction resolvasome RuvABC, ATP-dependent DNA helicase subunit	almost in all Bacteria, only in three Archaea
COG2256	L	Replication-associated recombination protein RarA (DNA-dependent ATPase)	almost in all Bacteria, only in four Archaea
COG2812	L	DNA polymerase III, gamma/tau subunits	almost in all Bacteria, only in three Archaea
COG4083	L	Exosortase/Archaeosortase	only in three Bacteria

Appendix 2. Descriptions of tested classifier hyperparameters and other settings. Values marked with "-" represent a continuous range and "," represent separate values. Parameters in cursive were modified in CV. 50 different parameter combinations were tested for each classifier.

Hyperparameters of the LogisticRegression module

Hyperparameter	**Value**
penalty	l1
C, regularization strength, step 0.5	0.1-24.6
multi_class	ovr
solver	liblinear
max_iter	200
class_weight	balanced

Hyperparameters of RandomForestClassifier module

Hyperparameter	**Value**
n_estimator	500
max_depth	5,7,9,14,25, None
min_sample_split	2,3,5
min_sample_leaf	1,3,5
max_features	sqrt,0.6
bootstrap	False

Hyperparameters of the neural network model

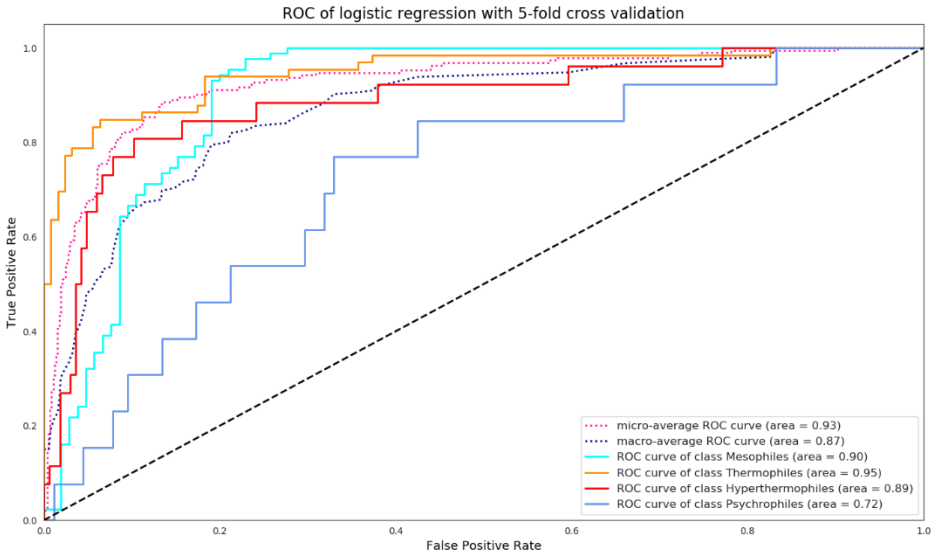
Hyperparameter	**Value**
Fully connected layer 1 nodes	110,120,130,140
Fully connected layer 1 regularization	0.001,0.003,0.05
Fully connected layer 1 activation	tf.nn.relu
Fully connected layer 2 nodes	60,70,80
Fully connected layer 2 activation	ReLU
Dropout layer drop rate	0.4,0.5,0.6
Output layer activation	tf.nn.softmax
MomentumOptimizer learning rate	0.001,0.02,0.1

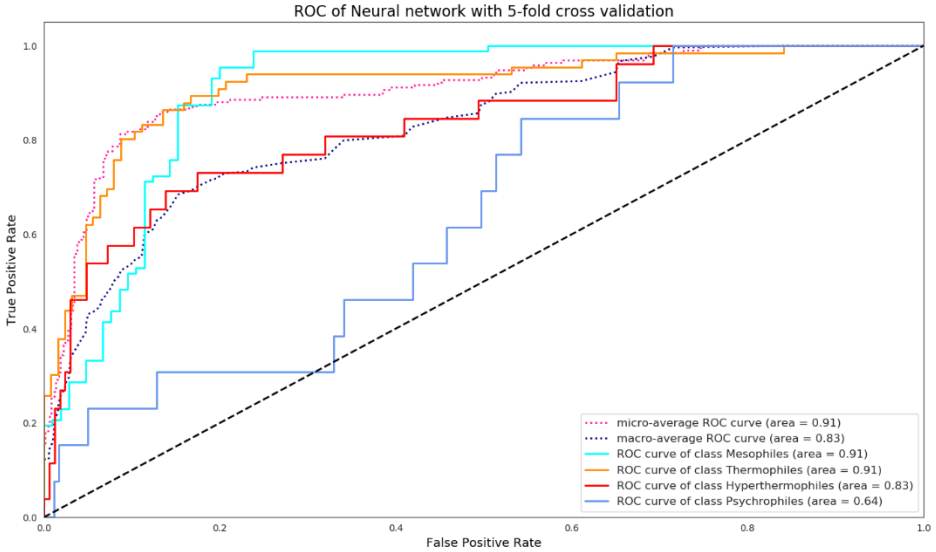
MomentumOptimizer momentum rate	0.7,0.8,0.90
Loss function	sparse_categorical_crossentropy
Batch size	80
Number of epochs	250

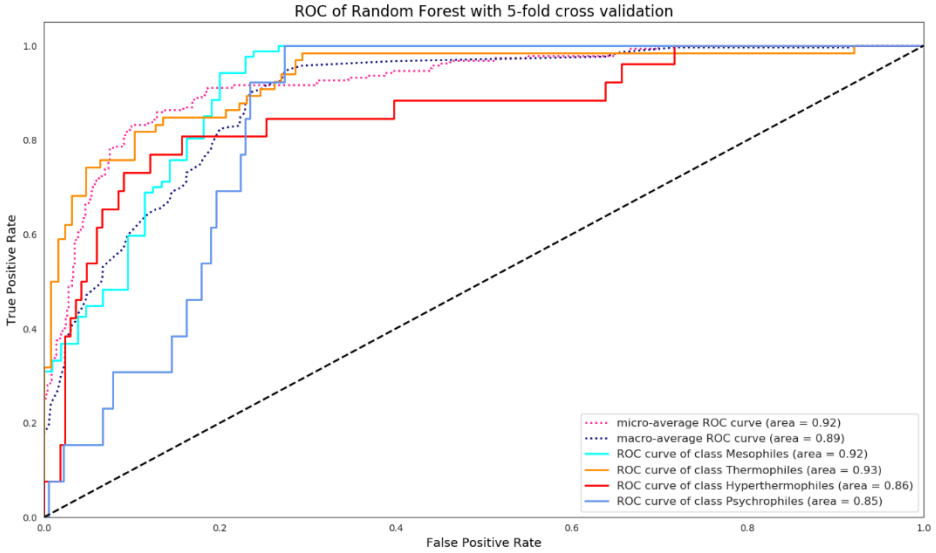
Hyperparameters of the XGBClassifier

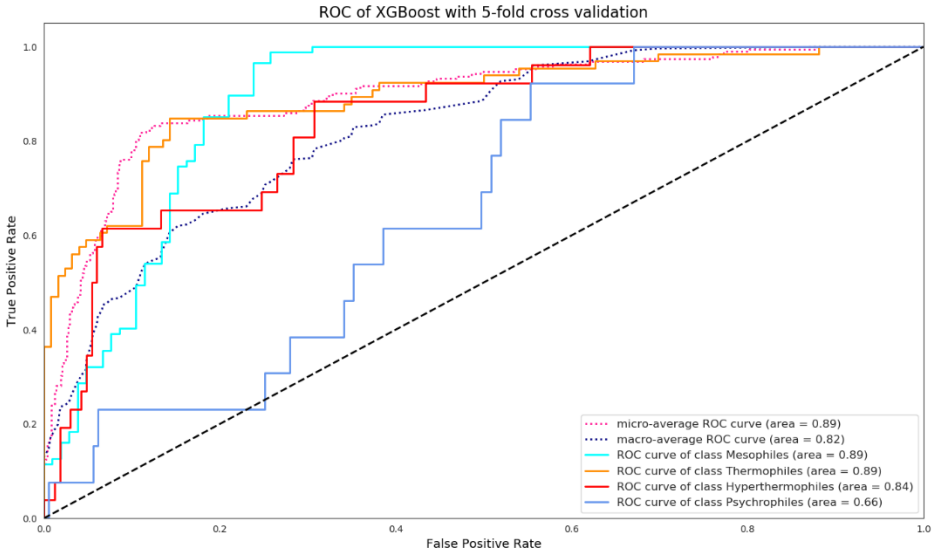
Hyperparameter	**Value**
n_estimator	150
max_depth	3,5,10,25
learning_rate	0.002,0.003,0.05,0.1
min_child_weight	2,3,5,7
gamma	0.1,0.2,0.5
reg_lamda	0.001,0.01,0.1,0.5
objective	multi:softmax
num_class	*4*
subsample	1
scale_pos_weight	1

Appendix 3. ROC curves and AUCs of each predicted class for each classifier.




ROC of Neural network with 5-fold cross validation
True Positive Rate
False Positive Rate
micro-average ROC curve (area = 0.91)
macro-average ROC curve (area = 0.83)
ROC curve of class Mesophiles (area = 0.91)
ROC curve of class Thermophiles (area = 0.91)
ROC curve of class Hyperthermophiles (area = 0.83)
ROC curve of class Psychrophiles (area = 0.64)


ROC of Random Forest with 5-fold cross validation
True Positive Rate
False Positive Rate
micro-average ROC curve (area = 0.92)
macro-average ROC curve (area = 0.89)
ROC curve of class Mesophiles (area = 0.92)
ROC curve of class Thermophiles (area = 0.93)
ROC curve of class Hyperthermophiles (area = 0.86)
ROC curve of class Psychrophiles (area = 0.85)



Appendix 4. Table of samples' predictions, kingdom, and phylum.

Species	**Prediction**	**taxid**	**kingdom**	**phyla**
Acaryochloris_marina_MBIC11017_uid58167	Mesophile	329726	Bacteria	Cyanobacteria
Acetobacter_pasteurianus_IFO_3283_01_uid59279	Mesophile	634452	Bacteria	Proteobacteria
Acetobacterium_woodii_DSM_1030_uid88073	Mesophile	931626	Bacteria	Firmicutes
Acetohalobium_arabaticum_DSM_5501_uid51423	Thermophile	574087	Bacteria	Firmicutes
Acholeplasma_laidlawii_PG_8A_uid58901	Mesophile	441768	Bacteria	Other
Achromobacter_xylosoxidans_A8_uid59899	Mesophile	762376	Bacteria	Proteobacteria
Acidaminococcus_intestini_RyC_MR95_uid74445	Thermophile	568816	Bacteria	Firmicutes
Acidianus_hospitalis_W1_uid66875	Hyperthermophile	933801	Archaea	Crenarchaeota
Acidimicrobidae_bacterium_YM16_304_uid193703	Mesophile	1313172	Bacteria	Actinobacteria
Acidimicrobium_ferrooxidans_DSM_10331_uid59215	Thermophile	525909	Bacteria	Actinobacteria
Acidiphilium_multivorum_AIU301_uid63345	Thermophile	926570	Bacteria	Proteobacteria
Acidithiobacillus_ferrooxidans_ATCC_23270_uid57649	Mesophile	243159	Bacteria	Proteobacteria
Acidobacterium_MP5ACTX9_uid50551	Mesophile	1198114	Bacteria	Acidobacteria

Acidovorax_avenae_ATCC_19860_uid42497	Mesophile	643561	Bacteria	Proteobacteria
Acinetobacter_baumannii_ATCC_17978_uid58731	Mesophile	400667	Bacteria	Proteobacteria
Actinobacillus_suis_H91_0380_uid176363	Mesophile	696748	Bacteria	Proteobacteria
Actinoplanes_friuliensis_DSM_7358_uid226110	Mesophile	1246995	Bacteria	Actinobacteria
Actinosynnema_mirum_DSM_43827_uid58951	Mesophile	446462	Bacteria	Actinobacteria
Adlercreutzia_equolifaciens_DSM_19450_uid223286	Mesophile	1384484	Bacteria	Actinobacteria
Advenella_kashmirensis_WT001_uid80859	Mesophile	1036672	Bacteria	Proteobacteria
Aerococcus_urinae_ACS_120_V_Col10a_uid64757	Mesophile	866775	Bacteria	Firmicutes
Aeromonas_salmonicida_A449_uid58631	Mesophile	382245	Bacteria	Proteobacteria
Aggregatibacter_actinomycetemcomitans_D7S_1_uid46989	Mesophile	694569	Bacteria	Proteobacteria
Agrobacterium_fabrum_C58_uid57865	Mesophile	176299	Bacteria	Proteobacteria
Agromonas_oligotrophica_S58_uid192186	Mesophile	1245469	Bacteria	Proteobacteria
Akkermansia_muciniphila_ATCC_BAA_835_uid58985	Mesophile	349741	Bacteria	Other
Alcanivorax_borkumensis_SK2_uid58169	Mesophile	393595	Bacteria	Proteobacteria
Alicycliphilus_denitrificans_K601_uid66307	Mesophile	596154	Bacteria	Proteobacteria
Alicyclobacillus_acidocaldarius_Tc_4_1_uid158681	Thermophile	1048834	Bacteria	Firmicutes
Aliivibrio_salmonicida_LFI1238_uid59251	Mesophile	316275	Bacteria	Proteobacteria
Alistipes_finegoldii_DSM_17242_uid168180	Mesophile	679935	Bacteria	Chlorobi
Alkalilimnicola_ehrlichii_MLHE_1_uid58467	Mesophile	187272	Bacteria	Proteobacteria
Alkaliphilus_metalliredigens_QYMF_uid58171	Thermophile	293826	Bacteria	Firmicutes
Allochromatium_vinosum_DSM_180_uid46083	Mesophile	572477	Bacteria	Proteobacteria
Alteromonas_SN2_uid67349	Psychrophile	715451	Bacteria	Proteobacteria
Aminobacterium_colombiense_DSM_12261_uid47083	Thermophile	572547	Bacteria	Synergistetes
Amphibacillus_xylanus_NBRC_15112_uid176453	Mesophile	698758	Bacteria	Firmicutes
Amycolatopsis_mediterranei_S699_uid158689	Mesophile	713604	Bacteria	Actinobacteria

Amycolicicoccus_subflavus_DQS3_9A1_uid67253	Mesophile	443218	Bacteria	Actinobacteria
Anabaena_cylindrica_PCC_7122_uid183339	Mesophile	272123	Bacteria	Cyanobacteria
Anaerococcus_prevotii_DSM_20548_uid59219	Mesophile	525919	Bacteria	Firmicutes
Anaeromyxobacter_dehalogenans_2CP_1_uid58989	Thermophile	455488	Bacteria	Proteobacteria
Anaplasma_phagocytophilum_HZ_uid57951	Mesophile	212042	Bacteria	Proteobacteria
Arcanobacterium_haemolyticum_DSM_20595_uid49489	Mesophile	644284	Bacteria	Actinobacteria
Arcobacter_nitrofigilis_DSM_7299_uid49001	Mesophile	572480	Bacteria	Proteobacteria
Aromatoleum_aromaticum_EbN1_uid58231	Mesophile	76114	Bacteria	Proteobacteria
Arthrobacter_chlorophenolicus_A6_uid58969	Psychrophile	452863	Bacteria	Actinobacteria
Arthrospira_platensis_NIES_39_uid197171	Mesophile	696747	Bacteria	Cyanobacteria
Aster_yellows_witches_broom_phytoplasma_AYWB_uid58297	Mesophile	322098	Bacteria	Other
Asticcacaulis_excentricus_CB_48_uid55641	Mesophile	573065	Bacteria	Proteobacteria
Atopobium_parvulum_DSM_20469_uid59195	Mesophile	521095	Bacteria	Actinobacteria
Azoarcus_KH32C_uid193704	Mesophile	748247	Bacteria	Proteobacteria
Azorhizobium_caulinodans_ORS_571_uid58905	Mesophile	438753	Bacteria	Proteobacteria
Azospirillum_brasilense_Sp245_uid162161	Mesophile	1064539	Bacteria	Proteobacteria
Azotobacter_vinelandii_DJ_uid57597	Mesophile	322710	Bacteria	Proteobacteria
Bacillus_thuringiensis_serovar_kurstaki_HD73_uid189188	Mesophile	1279365	Bacteria	Firmicutes
Bacteriovorax_marinus_SJ_uid82341	Mesophile	862908	Bacteria	Proteobacteria
Baumannia_cicadellinicola_Hc__Homalodisca_coagulata__uid58111	Mesophile	374463	Bacteria	Proteobacteria
Beijerinckia_indica_ATCC_9039_uid59057	Mesophile	395963	Bacteria	Proteobacteria
Belliella_baltica_DSM_15883_uid168182	Psychrophile	866536	Bacteria	Chlorobi
Beutenbergia_cavernae_DSM_12333_uid59047	Mesophile	471853	Bacteria	Actinobacteria
Bibersteinia_trehalosi_192_uid193709	Mesophile	1171377	Bacteria	Proteobacteria
Bifidobacterium_longum_infantis_ATCC_15697_uid159865	Mesophile	391904	Bacteria	Actinobacteria

Blastococcus_saxobsidens_DD2_uid89391	Mesophile	1146883	Bacteria	Actinobacteria
Blattabacterium__Blattella_germanica__Bge_uid41533	Mesophile	331104	Bacteria	Chlorobi
Bordetella_petrii_uid61631	Mesophile	340100	Bacteria	Proteobacteria
Brachybacterium_faecium_DSM_4810_uid58649	Mesophile	446465	Bacteria	Actinobacteria
Brachyspira_intermedia_PWS_A_uid158369	Mesophile	1045858	Bacteria	Spirochaetes
Bradyrhizobium_japonicum_USDA_6_uid158851	Mesophile	1037409	Bacteria	Proteobacteria
Brevibacillus_brevis_NBRC_100599_uid59175	Mesophile	358681	Bacteria	Firmicutes
Brevundimonas_subvibrioides_ATCC_15264_uid42117	Mesophile	633149	Bacteria	Proteobacteria
Brucella_melitensis_bv__1_16M_uid57735	Mesophile	224914	Bacteria	Proteobacteria
Burkholderia_xenovorans_LB400_uid57823	Mesophile	266265	Bacteria	Proteobacteria
Butyrivibrio_proteoclasticus_B316_uid51489	Mesophile	515622	Bacteria	Firmicutes
Calothrix_PCC_7507_uid182930	Mesophile	99598	Bacteria	Cyanobacteria
Candidatus_Accumulibacter_phosphatis_clade_IIA_UW_1_uid59207	Mesophile	522306	Bacteria	Proteobacteria
Candidatus_Amoebophilus_asiaticus_5a2_uid58963	Mesophile	452471	Bacteria	Chlorobi
Candidatus_Arthromitus_SFB_mouse_Japan_uid71379	Mesophile	1029718	Bacteria	Firmicutes
Candidatus_Azobacteroides_pseudotrichonymphae_genomovar__CFP2_uid59163	Mesophile	511995	Bacteria	Chlorobi
Candidatus_Blochmannia_pennsylvanicus_BPEN_uid58329	Mesophile	291272	Bacteria	Proteobacteria
Candidatus_Caldiarchaeum_subterraneum_uid227223	Hyperthermophile	311458	Archaea	Thaumarchaeota
Candidatus_Carsonella_ruddii_DC_uid213383	Mesophile	667013	Bacteria	Proteobacteria
Candidatus_Chloracidobacterium_thermophilum_B_uid73587	Thermophile	981222	Bacteria	Acidobacteria
Candidatus_Cloacamonas_acidaminovorans_Evry_uid62959	Thermophile	459349	Bacteria	Other
Candidatus_Desulforudis_audaxviator_MP104C_uid59067	Thermophile	477974	Bacteria	Firmicutes
Candidatus_Hamiltonella_defensa_5AT__Acyrthosiphon_pisum__uid59289	Mesophile	572265	Bacteria	Proteobacteria
Candidatus_Hodgkinia_cicadicola_Dsem_uid59311	Mesophile	573234	Bacteria	Proteobacteria
Candidatus_Kinetoplastibacterium_desouzaii_TCC079E_uid189750	Mesophile	1208919	Bacteria	Proteobacteria

Candidatus_Korarchaeum_cryptofilum_OPF8_uid58601	Hyperthermophile	374847	Archaea	Korarchaeota
Candidatus_Koribacter_versatilis_Ellin345_uid58479	Mesophile	204669	Bacteria	Acidobacteria
Candidatus_Liberibacter_solanacearum_CLso_ZC1_uid61245	Mesophile	658172	Bacteria	Proteobacteria
Candidatus_Methylomirabilis_oxyfera_uid161981	Thermophile	671143	Bacteria	Other
Candidatus_Midichloria_mitochondrii_IricVA_uid68687	Mesophile	696127	Bacteria	Proteobacteria
Candidatus_Moranella_endobia_PCVAL_uid197215	Mesophile	1234603	Bacteria	Proteobacteria
Candidatus_Nasuia_deltocephalinicola_NAS_ALF_uid214084	Mesophile	1343077	Bacteria	Proteobacteria
Candidatus_Nitrosopumilus_koreensis_AR1_uid176129	Mesophile	1229908	Archaea	Thaumarchaeota
Candidatus_Nitrososphaera_gargensis_Ga9_2_uid176707	Mesophile	1237085	Archaea	Thaumarchaeota
Candidatus_Nitrospira_defluvii_uid51175	Thermophile	330214	Bacteria	Other
Candidatus_Pelagibacter_IMCC9063_uid66305	Mesophile	1002672	Bacteria	Proteobacteria
Candidatus_Phytoplasma_australiense_uid61641	Mesophile	59748	Bacteria	Other
Candidatus_Portiera_aleyrodidarum_BT_QVLC_uid176374	Mesophile	1239881	Bacteria	Proteobacteria
Candidatus_Puniceispirillum_marinum_IMCC1322_uid47081	Mesophile	488538	Bacteria	Proteobacteria
Candidatus_Riesia_pediculicola_USDA_uid46841	Mesophile	515618	Bacteria	Proteobacteria
Candidatus_Ruthia_magnifica_Cm__Calyptogena_magnifica__uid58645	Mesophile	413404	Bacteria	Proteobacteria
Candidatus_Saccharibacteria_bacterium_RAAC3_TM7_1_uid230715	Mesophile	1394711	Bacteria	Other
Candidatus_Saccharobacterium_alaburgensis_uid203361	Mesophile	1332188	Bacteria	Other
Candidatus_Solibacter_usitatus_Ellin6076_uid58139	Thermophile	234267	Bacteria	Acidobacteria
Candidatus_Sulcia_muelleri_CARI_uid52535	Mesophile	706194	Bacteria	Chlorobi
Candidatus_Tremblaya_phenacola_PAVE_uid209173	Mesophile	1266371	Bacteria	Proteobacteria
Candidatus_Uzinura_diaspidicola_ASNER_uid186740	Mesophile	1133592	Bacteria	Chlorobi
Candidatus_Vesicomyosocius_okutanii_HA_uid59427	Mesophile	412965	Bacteria	Proteobacteria
Candidatus_Zinderia_insecticola_CARI_uid52459	Mesophile	871271	Bacteria	Proteobacteria
Capnocytophaga_canimorsus_Cc5_uid70727	Psychrophile	860228	Bacteria	Chlorobi

Cardinium_endosymbiont_cEper1_of_Encarsia_pergandiella_uid175524	Mesophile	1231626	Bacteria	Chlorobi
Carnobacterium_maltaromaticum_LMA28_uid179370	Mesophile	1234679	Bacteria	Firmicutes
Catenulispora_acidiphila_DSM_44928_uid59077	Mesophile	479433	Bacteria	Actinobacteria
Cellulomonas_fimi_ATCC_484_uid66779	Mesophile	590998	Bacteria	Actinobacteria
Cellvibrio_japonicus_Ueda107_uid59139	Mesophile	498211	Bacteria	Proteobacteria
Cenarchaeum_symbiosum_A_uid61411	Mesophile	414004	Archaea	Thaumarchaeota
Chamaesiphon_PCC_6605_uid183005	Mesophile	1173020	Bacteria	Cyanobacteria
Chelativorans_BNC1_uid58069	Mesophile	266779	Bacteria	Proteobacteria
Chitinophaga_pinensis_DSM_2588_uid59113	Psychrophile	485918	Bacteria	Chlorobi
Chlorobaculum_parvum_NCIB_8327_uid59185	Hyperthermophile	517417	Bacteria	Chlorobi
Chloroherpeton_thalassium_ATCC_35110_uid59187	Thermophile	517418	Bacteria	Chlorobi
Chromohalobacter_salexigens_DSM_3043_uid62921	Mesophile	290398	Bacteria	Proteobacteria
Chroococcidiopsis_thermalis_PCC_7203_uid183002	Mesophile	251229	Bacteria	Cyanobacteria
Citrobacter_koseri_ATCC_BAA_895_uid58143	Mesophile	290338	Bacteria	Proteobacteria
Clavibacter_michiganensis_sepedonicus_uid61577	Mesophile	31964	Bacteria	Actinobacteria
Clostridiales_genomosp__BVAB3_UPII9_5_uid46219	Mesophile	699246	Bacteria	Firmicutes
Clostridium_botulinum_A_ATCC_3502_uid61579	Mesophile	413999	Bacteria	Firmicutes
Clostridium_difficile_630_uid57679	Mesophile	272563	Bacteria	Firmicutes
Collimonas_fungivorans_Ter331_uid70793	Mesophile	1005048	Bacteria	Proteobacteria
Comamonadaceae_bacterium_CR_uid223378	Mesophile	946483	Bacteria	Proteobacteria
Comamonas_testosteroni_CNB_2_uid62961	Mesophile	688245	Bacteria	Proteobacteria
Conexibacter_woesei_DSM_14684_uid43467	Mesophile	469383	Bacteria	Actinobacteria
Coprococcus_catus_GD_7_uid197174	Mesophile	717962	Bacteria	Firmicutes
Coraliomargarita_akajimensis_DSM_45221_uid47079	Mesophile	583355	Bacteria	Other
Corallococcus_coralloides_DSM_2259_uid157997	Psychrophile	1144275	Bacteria	Proteobacteria

Coriobacterium_glomerans_PW2_uid65787	Mesophile	700015	Bacteria	Actinobacteria
Coxiella_burnetii_Dugway_5J108_111_uid58629	Mesophile	434922	Bacteria	Proteobacteria
Crinalium_epipsammum_PCC_9333_uid183113	Mesophile	1173022	Bacteria	Cyanobacteria
Croceibacter_atlanticus_HTCC2559_uid49661	Psychrophile	216432	Bacteria	Chlorobi
Cronobacter_turicensis_z3032_uid40821	Mesophile	693216	Bacteria	Proteobacteria
Cryptobacterium_curtum_DSM_15641_uid59041	Mesophile	469378	Bacteria	Actinobacteria
Cupriavidus_necator_N_1_uid68689	Mesophile	1042878	Bacteria	Proteobacteria
Cyanobacterium_PCC_10605_uid183340	Mesophile	755178	Bacteria	Cyanobacteria
Cyanobium_gracile_PCC_6307_uid182931	Mesophile	292564	Bacteria	Cyanobacteria
Cyanothece_PCC_7822_uid52547	Mesophile	497965	Bacteria	Cyanobacteria
Cyclobacterium_marinum_DSM_745_uid71485	Psychrophile	880070	Bacteria	Chlorobi
Cycloclasticus_zancles_7_ME_uid214092	Mesophile	1198232	Bacteria	Proteobacteria
Cylindrospermum_stagnale_PCC_7417_uid183111	Mesophile	56107	Bacteria	Cyanobacteria
Cytophaga_hutchinsonii_ATCC_33406_uid57651	Mesophile	269798	Bacteria	Chlorobi
Dactylococcopsis_salina_PCC_8305_uid183341	Mesophile	13035	Bacteria	Cyanobacteria
Dechloromonas_aromatica_RCB_uid58025	Mesophile	159087	Bacteria	Proteobacteria
Dechlorosoma_suillum_PS_uid81439	Mesophile	640081	Bacteria	Proteobacteria
Dehalobacter_CF_uid177714	Thermophile	1131462	Bacteria	Firmicutes
Dehalogenimonas_lykanthroporepellens_BL_DC_9_uid48131	Mesophile	552811	Bacteria	Other
Delftia_acidovorans_SPH_1_uid58703	Mesophile	398578	Bacteria	Proteobacteria
Denitrovibrio_acetiphilus_DSM_12809_uid46657	Mesophile	522772	Bacteria	Other
Desulfarculus_baarsii_DSM_2075_uid51371	Mesophile	644282	Bacteria	Proteobacteria
Desulfatibacillum_alkenivorans_AK_01_uid58913	Thermophile	439235	Bacteria	Proteobacteria
Desulfitobacterium_hafniense_Y51_uid58605	Thermophile	138119	Bacteria	Firmicutes
Desulfobacca_acetoxidans_DSM_11109_uid65785	Thermophile	880072	Bacteria	Proteobacteria

Desulfobacterium_autotrophicum_HRM2_uid59061	Thermophile	177437	Bacteria	Proteobacteria
Desulfobacula_toluolica_Tol2_uid175777	Psychrophile	651182	Bacteria	Proteobacteria
Desulfobulbus_propionicus_DSM_2032_uid62265	Thermophile	577650	Bacteria	Proteobacteria
Desulfocapsa_sulfexigens_DSM_10523_uid189952	Mesophile	1167006	Bacteria	Proteobacteria
Desulfococcus_oleovorans_Hxd3_uid58777	Thermophile	96561	Bacteria	Proteobacteria
Desulfohalobium_retbaense_DSM_5692_uid59183	Mesophile	485915	Bacteria	Proteobacteria
Desulfomicrobium_baculatum_DSM_4028_uid59217	Mesophile	525897	Bacteria	Proteobacteria
Desulfomonile_tiedjei_DSM_6799_uid168320	Thermophile	706587	Bacteria	Proteobacteria
Desulfosporosinus_orientis_DSM_765_uid82939	Thermophile	768706	Bacteria	Firmicutes
Desulfotomaculum_gibsoniae_DSM_7213_uid76945	Thermophile	767817	Bacteria	Firmicutes
Desulfurispirillum_indicum_S5_uid45897	Mesophile	653733	Bacteria	Other
Desulfurivibrio_alkaliphilus_AHT2_uid49487	Mesophile	589865	Bacteria	Proteobacteria
Dichelobacter_nodosus_VCS1703A_uid57643	Mesophile	246195	Bacteria	Proteobacteria
Dickeya_dadantii_3937_uid52537	Mesophile	198628	Bacteria	Proteobacteria
Dinoroseobacter_shibae_DFL_12_uid58707	Mesophile	398580	Bacteria	Proteobacteria
Dyadobacter_fermentans_DSM_18053_uid59049	Psychrophile	471854	Bacteria	Chlorobi
Echinicola_vietnamensis_DSM_17526_uid184076	Psychrophile	926556	Bacteria	Chlorobi
Ectothiorhodospiraceae_bacterium_M19_40_uid199898	Mesophile	1260251	Bacteria	Proteobacteria
Edwardsiella_ictaluri_93_146_uid59403	Mesophile	634503	Bacteria	Proteobacteria
Eggerthella_lenta_DSM_2243_uid59079	Mesophile	479437	Bacteria	Actinobacteria
Ehrlichia_chaffeensis_Arkansas_uid57933	Mesophile	205920	Bacteria	Proteobacteria
Elusimicrobium_minutum_Pei191_uid58949	Mesophile	445932	Bacteria	Other
Emticicia_oligotrophica_DSM_17448_uid177079	Psychrophile	929562	Bacteria	Chlorobi
Enterobacter_cloacae_ATCC_13047_uid48363	Mesophile	716541	Bacteria	Proteobacteria
Erwinia_amylovora_ATCC_49946_uid46943	Mesophile	716540	Bacteria	Proteobacteria

Erysipelothrix_rhusiopathiae_SY1027_uid206518	Mesophile	1313290	Bacteria	Firmicutes
Erythrobacter_litoralis_HTCC2594_uid58299	Mesophile	314225	Bacteria	Proteobacteria
Escherichia_coli_K_12_substr__MG1655_uid57779	Mesophile	511145	Bacteria	Proteobacteria
Ethanoligenens_harbinense_YUAN_3_uid46255	Mesophile	663278	Bacteria	Firmicutes
Eubacterium_limosum_KIST612_uid59777	Mesophile	903814	Bacteria	Firmicutes
Exiguobacterium_MH3_uid227425	Mesophile	1399115	Bacteria	Firmicutes
Faecalibacterium_prausnitzii_L2_6_uid197183	Mesophile	718252	Bacteria	Firmicutes
Ferrimonas_balearica_DSM_9799_uid53371	Mesophile	550540	Bacteria	Proteobacteria
Ferroplasma_acidarmanus_fer1_uid54095	Hyperthermophile	333146	Archaea	Euryarchaeota
Fibrella_aestuarina_uid178352	Psychrophile	1166018	Bacteria	Chlorobi
Fibrobacter_succinogenes_S85_uid41169	Mesophile	59374	Bacteria	Other
Filifactor_alocis_ATCC_35896_uid46625	Mesophile	546269	Bacteria	Firmicutes
Finegoldia_magna_ATCC_29328_uid58867	Mesophile	334413	Bacteria	Firmicutes
Flavobacterium_johnsoniae_UW101_uid58493	Psychrophile	376686	Bacteria	Chlorobi
Flexibacter_litoralis_DSM_6794_uid168257	Psychrophile	880071	Bacteria	Chlorobi
Flexistipes_sinusarabici_DSM_4947_uid68147	Thermophile	717231	Bacteria	Other
Fluviicola_taffensis_DSM_16823_uid65271	Psychrophile	755732	Bacteria	Chlorobi
Francisella_tularensis_SCHU_S4_uid57589	Mesophile	177416	Bacteria	Proteobacteria
Frankia_EAN1pec_uid58367	Mesophile	298653	Bacteria	Actinobacteria
Frankia_symbiont_of_Datisca_glomerata_uid46257	Mesophile	656024	Bacteria	Actinobacteria
Frateuria_aurantia_DSM_6220_uid81775	Mesophile	767434	Bacteria	Proteobacteria
Gallibacterium_anatis_UMN179_uid66567	Mesophile	1005058	Bacteria	Proteobacteria
Gallionella_capsiferriformans_ES_2_uid51505	Mesophile	395494	Bacteria	Proteobacteria
Gardnerella_vaginalis_ATCC_14019_uid55487	Mesophile	525284	Bacteria	Actinobacteria
Geitlerinema_PCC_7407_uid183007	Mesophile	1173025	Bacteria	Cyanobacteria

Gemmatimonas_aurantiaca_T_27_uid58813	Thermophile	379066	Bacteria	Other
Geobacillus_thermoleovorans_CCB_US3_UF5_uid82949	Thermophile	1111068	Bacteria	Firmicutes
Geodermatophilus_obscurus_DSM_43160_uid43725	Mesophile	526225	Bacteria	Actinobacteria
Gloeocapsa_PCC_7428_uid183112	Mesophile	1173026	Bacteria	Cyanobacteria
Gluconacetobacter_diazotrophicus_PAl_5_uid61587	Mesophile	272568	Bacteria	Proteobacteria
Gluconobacter_oxydans_H24_uid179202	Mesophile	1224746	Bacteria	Proteobacteria
Gordonia_KTR9_uid174812	Mesophile	337191	Bacteria	Actinobacteria
Gordonibacter_pamelaeae_7_10_1_b_uid197167	Mesophile	657308	Bacteria	Actinobacteria
Gramella_forsetii_KT0803_uid58881	Psychrophile	411154	Bacteria	Chlorobi
Granulibacter_bethesdensis_CGDNIH1_uid58661	Mesophile	391165	Bacteria	Proteobacteria
Granulicella_mallensis_MP5ACTX8_uid49957	Mesophile	682795	Bacteria	Acidobacteria
Hahella_chejuensis_KCTC_2396_uid58483	Psychrophile	349521	Bacteria	Proteobacteria
Halanaerobium_hydrogeniformans_uid60191	Mesophile	656519	Bacteria	Firmicutes
Haliangium_ochraceum_DSM_14365_uid41425	Mesophile	502025	Bacteria	Proteobacteria
Haliscomenobacter_hydrossis_DSM_1100_uid66777	Psychrophile	760192	Bacteria	Chlorobi
Halobacillus_halophilus_DSM_2266_uid162033	Mesophile	866895	Bacteria	Firmicutes
Halobacteroides_halobius_DSM_5150_uid184862	Thermophile	748449	Bacteria	Firmicutes
Haloferax_volcanii_DS2_uid46845	Mesophile	309800	Archaea	Euryarchaeota
Halomonas_elongata_DSM_2581_uid52781	Mesophile	768066	Bacteria	Proteobacteria
Haloquadratum_walsbyi_C23_uid162019	Mesophile	768065	Archaea	Euryarchaeota
Halorhabdus_tiamatea_SARL4B_uid214082	Mesophile	1033806	Archaea	Euryarchaeota
Halorubrum_lacusprofundi_ATCC_49239_uid58807	Mesophile	416348	Archaea	Euryarchaeota
Halothece_PCC_7418_uid183338	Mesophile	65093	Bacteria	Cyanobacteria
Halothermothrix_orenii_H_168_uid58585	Thermophile	373903	Bacteria	Firmicutes
Halothiobacillus_neapolitanus_c2_uid41317	Mesophile	555778	Bacteria	Proteobacteria

Halyomorpha_halys_symbiont_uid222821	Mesophile	1235990	Bacteria	Proteobacteria
Herbaspirillum_seropedicae_SmR1_uid50427	Mesophile	757424	Bacteria	Proteobacteria
Herminiimonas_arsenicoxydans_uid58291	Mesophile	204773	Bacteria	Proteobacteria
Herpetosiphon_aurantiacus_DSM_785_uid58599	Thermophile	316274	Bacteria	Other
Hirschia_baltica_ATCC_49814_uid59365	Mesophile	582402	Bacteria	Proteobacteria
Hydrogenobaculum_Y04AAS1_uid58857	Thermophile	380749	Bacteria	Aquificae
Hyphomicrobium_MC1_uid68453	Mesophile	717785	Bacteria	Proteobacteria
Hyphomonas_neptunium_ATCC_15444_uid58433	Mesophile	228405	Bacteria	Proteobacteria
Ignavibacterium_album_JCM_16511_uid162097	Thermophile	945713	Bacteria	Chlorobi
Ilyobacter_polytropus_DSM_2926_uid59769	Mesophile	572544	Bacteria	Fusobacteria
Intrasporangium_calvum_DSM_43043_uid61729	Mesophile	710696	Bacteria	Actinobacteria
Isoptericola_variabilis_225_uid67501	Mesophile	743718	Bacteria	Actinobacteria
Isosphaera_pallida_ATCC_43644_uid62207	Thermophile	575540	Bacteria	Planctomycetes
Jannaschia_CCS1_uid58147	Mesophile	290400	Bacteria	Proteobacteria
Janthinobacterium_Marseille_uid58603	Mesophile	375286	Bacteria	Proteobacteria
Jonesia_denitrificans_DSM_20603_uid59053	Mesophile	471856	Bacteria	Actinobacteria
Kangiella_koreensis_DSM_16069_uid59209	Mesophile	523791	Bacteria	Proteobacteria
Ketogulonicigenium_vulgare_Y25_uid59581	Mesophile	880591	Bacteria	Proteobacteria
Kineococcus_radiotolerans_SRS30216_uid58067	Mesophile	266940	Bacteria	Actinobacteria
Kitasatospora_setae_KM_6054_uid77027	Mesophile	452652	Bacteria	Actinobacteria
Klebsiella_pneumoniae_342_uid59145	Mesophile	507522	Bacteria	Proteobacteria
Kocuria_rhizophila_DC2201_uid59099	Mesophile	378753	Bacteria	Actinobacteria
Kribbella_flavida_DSM_17836_uid43465	Mesophile	479435	Bacteria	Actinobacteria
Krokinobacter_4H_3_7_5_uid66593	Psychrophile	983548	Bacteria	Chlorobi
Kytococcus_sedentarius_DSM_20547_uid59071	Mesophile	478801	Bacteria	Actinobacteria

Lacinutrix_5H_3_7_4_uid68067	Psychrophile	983544	Bacteria	Chlorobi
Lactobacillus_plantarum_ZJ316_uid188689	Mesophile	1284663	Bacteria	Firmicutes
Laribacter_hongkongensis_HLHK9_uid59265	Mesophile	557598	Bacteria	Proteobacteria
Lawsonia_intracellularis_N343_uid186598	Mesophile	1234378	Bacteria	Proteobacteria
Leadbetterella_byssophila_DSM_17132_uid60161	Psychrophile	649349	Bacteria	Chlorobi
Leifsonia_xyli_cynodontis_DSM_46306_uid221294	Mesophile	1389489	Bacteria	Actinobacteria
Leisingera_methylohalidivorans_DSM_14336_uid232356	Mesophile	999552	Bacteria	Proteobacteria
Leptolyngbya_PCC_7376_uid182928	Mesophile	111781	Bacteria	Cyanobacteria
Leptospira_biflexa_serovar_Patoc__Patoc_1__Paris__uid58993	Mesophile	456481	Bacteria	Spirochaetes
Leptospirillum_ferriphilum_ML_04_uid175904	Thermophile	1048260	Bacteria	Other
Leptothrix_cholodnii_SP_6_uid58971	Mesophile	395495	Bacteria	Proteobacteria
Leptotrichia_buccalis_C_1013_b_uid59211	Mesophile	523794	Bacteria	Fusobacteria
Leuconostoc_kimchii_IMSNU_11154_uid48589	Mesophile	762051	Bacteria	Firmicutes
Listonella_anguillarum_M3_uid217771	Mesophile	882944	Bacteria	Proteobacteria
Lysinibacillus_sphaericus_C3_41_uid58945	Mesophile	444177	Bacteria	Firmicutes
Macrococcus_caseolyticus_JCSC5402_uid59003	Mesophile	458233	Bacteria	Firmicutes
Magnetococcus_MC_1_uid57833	Thermophile	156889	Bacteria	Proteobacteria
Magnetospirillum_magneticum_AMB_1_uid58527	Mesophile	342108	Bacteria	Proteobacteria
Mahella_australiensis_50_1_BON_uid66917	Thermophile	697281	Bacteria	Firmicutes
Mannheimia_haemolytica_M42548_uid198769	Mesophile	1316932	Bacteria	Proteobacteria
Maricaulis_maris_MCS10_uid58689	Mesophile	394221	Bacteria	Proteobacteria
Marinobacter_adhaerens_HP15_uid162009	Mesophile	225937	Bacteria	Proteobacteria
Marinomonas_MWYL1_uid58715	Mesophile	400668	Bacteria	Proteobacteria
Marivirga_tractuosa_DSM_4126_uid60837	Psychrophile	643867	Bacteria	Chlorobi
Megamonas_hypermegale_uid197163	Thermophile	657316	Bacteria	Firmicutes

Megasphaera_elsdenii_DSM_20460_uid71135	Hyperthermophile	1064535	Bacteria	Firmicutes
Meiothermus_silvanus_DSM_9946_uid49485	Thermophile	526227	Bacteria	Deinococcus-Thermus
Melioribacter_roseus_P3M_uid170941	Thermophile	1191523	Bacteria	Chlorobi
Melissococcus_plutonius_ATCC_35311_uid66803	Mesophile	940190	Bacteria	Firmicutes
Mesoplasma_florum_W37_uid224253	Mesophile	1406864	Bacteria	Other
Mesotoga_prima_MesG1_Ag_4_2_uid52599	Thermophile	660470	Bacteria	Thermotogae
Methanocella_arvoryzae_MRE50_uid61623	Mesophile	351160	Archaea	Euryarchaeota
Methanococcoides_burtonii_DSM_6242_uid58023	Mesophile	259564	Archaea	Euryarchaeota
Methanohalobium_evestigatum_Z_7303_uid49857	Mesophile	644295	Archaea	Euryarchaeota
Methanolobus_psychrophilus_R15_uid177925	Mesophile	1094980	Archaea	Euryarchaeota
Methanomassiliicoccus_Mx1_Issoire_uid207287	Hyperthermophile	1295009	Archaea	Euryarchaeota
Methanoplanus_petrolearius_DSM_11571_uid52695	Mesophile	679926	Archaea	Euryarchaeota
Methanoregula_formicicum_SMSP_uid184406	Mesophile	593750	Archaea	Euryarchaeota
Methanosalsum_zhilinae_DSM_4017_uid68249	Mesophile	679901	Archaea	Euryarchaeota
Methanospirillum_hungatei_JF_1_uid58181	Mesophile	323259	Archaea	Euryarchaeota
Methylacidiphilum_infernorum_V4_uid59161	Hyperthermophile	481448	Bacteria	Other
Methylibium_petroleiphilum_PM1_uid58085	Mesophile	420662	Bacteria	Proteobacteria
Methylobacillus_flagellatus_KT_uid58049	Mesophile	265072	Bacteria	Proteobacteria
Methylobacterium_nodulans_ORS_2060_uid59023	Mesophile	460265	Bacteria	Proteobacteria
Methylocella_silvestris_BL2_uid59433	Mesophile	395965	Bacteria	Proteobacteria
Methylocystis_SC2_uid174072	Mesophile	187303	Bacteria	Proteobacteria
Methylomicrobium_alcaliphilum_uid77119	Mesophile	1091494	Bacteria	Proteobacteria
Methylomonas_methanica_MC09_uid67363	Mesophile	857087	Bacteria	Proteobacteria
Methylophaga_JAM1_uid162947	Mesophile	754476	Bacteria	Proteobacteria
Methylovorus_glucosetrophus_SIP3_4_uid59367	Mesophile	582744	Bacteria	Proteobacteria

Micavibrio_aeruginosavorus_ARL_13_uid73585	Mesophile	856793	Bacteria	Proteobacteria
Microbacterium_testaceum_StLB037_uid62789	Mesophile	979556	Bacteria	Actinobacteria
Micrococcus_luteus_NCTC_2665_uid59033	Mesophile	465515	Bacteria	Actinobacteria
Microcoleus_PCC_7113_uid183114	Mesophile	1173027	Bacteria	Cyanobacteria
Microcystis_aeruginosa_NIES_843_uid59101	Mesophile	449447	Bacteria	Cyanobacteria
Microlunatus_phosphovorus_NM_1_uid68055	Mesophile	1032480	Bacteria	Actinobacteria
Micromonospora_aurantiaca_ATCC_27029_uid42501	Mesophile	644283	Bacteria	Actinobacteria
Mobiluncus_curtisii_ATCC_43063_uid49695	Mesophile	548479	Bacteria	Actinobacteria
Modestobacter_marinus_uid167487	Psychrophile	477641	Bacteria	Actinobacteria
Moraxella_catarrhalis_BBH18_uid48809	Mesophile	749219	Bacteria	Proteobacteria
Morganella_morganii_KT_uid180867	Mesophile	1124991	Bacteria	Proteobacteria
Muricauda_ruestringensis_DSM_13258_uid72479	Psychrophile	886377	Bacteria	Chlorobi
Myxococcus_xanthus_DK_1622_uid58003	Psychrophile	246197	Bacteria	Proteobacteria
Nakamurella_multipartita_DSM_44233_uid59221	Thermophile	479431	Bacteria	Actinobacteria
Natrinema_J7_uid171337	Mesophile	406552	Archaea	Euryarchaeota
Natronobacterium_gregoryi_SP2_uid74439	Mesophile	797304	Archaea	Euryarchaeota
Nautilia_profundicola_AmH_uid59345	Thermophile	598659	Bacteria	Proteobacteria
Neorickettsia_sennetsu_Miyayama_uid57965	Mesophile	222891	Bacteria	Proteobacteria
Niastella_koreensis_GR20_10_uid83125	Psychrophile	700598	Bacteria	Chlorobi
Nitratifractor_salsuginis_DSM_16511_uid62183	Thermophile	749222	Bacteria	Proteobacteria
Nitrobacter_hamburgensis_X14_uid58293	Mesophile	323097	Bacteria	Proteobacteria
Nitrosococcus_halophilus_Nc4_uid46803	Mesophile	472759	Bacteria	Proteobacteria
Nitrosomonas_Is79A3_uid68745	Mesophile	261292	Bacteria	Proteobacteria
Nitrosospira_multiformis_ATCC_25196_uid58361	Mesophile	323848	Bacteria	Proteobacteria
Nocardia_brasiliensis_ATCC_700358_uid86913	Mesophile	1133849	Bacteria	Actinobacteria

Nocardioides_JS614_uid58149	Thermophile	196162	Bacteria	Actinobacteria
Nocardiopsis_alba_ATCC_BAA_2165_uid174334	Mesophile	1205910	Bacteria	Actinobacteria
Novosphingobium_PP1Y_uid67383	Mesophile	702113	Bacteria	Proteobacteria
Oceanimonas_GK1_uid81627	Mesophile	511062	Bacteria	Proteobacteria
Ochrobactrum_anthropi_ATCC_49188_uid58921	Mesophile	439375	Bacteria	Proteobacteria
Odoribacter_splanchnicus_DSM_20712_uid63397	Mesophile	709991	Bacteria	Chlorobi
Oenococcus_oeni_PSU_1_uid59417	Mesophile	203123	Bacteria	Firmicutes
Oligotropha_carboxidovorans_OM5_uid59155	Mesophile	504832	Bacteria	Proteobacteria
Olsenella_uli_DSM_7084_uid51367	Mesophile	633147	Bacteria	Actinobacteria
Opitutus_terrae_PB90_1_uid58965	Thermophile	452637	Bacteria	Other
Orientia_tsutsugamushi_Ikeda_uid58869	Mesophile	334380	Bacteria	Proteobacteria
Ornithobacterium_rhinotracheale_DSM_15997_uid168256	Mesophile	867902	Bacteria	Chlorobi
Oscillatoria_PCC_7112_uid183110	Mesophile	179408	Bacteria	Cyanobacteria
Oscillibacter_valericigenes_uid73895	Mesophile	693746	Bacteria	Firmicutes
Owenweeksia_hongkongensis_DSM_17368_uid82951	Mesophile	926562	Bacteria	Chlorobi
Paenibacillus_mucilaginosus_KNP414_uid68311	Mesophile	1036673	Bacteria	Firmicutes
Paludibacter_propionicigenes_WB4_uid60725	Mesophile	694427	Bacteria	Chlorobi
Pandoraea_pnomenusa_3kgm_uid229878	Mesophile	1416914	Bacteria	Proteobacteria
Pantoea_At_9b_uid55845	Mesophile	592316	Bacteria	Proteobacteria
Parabacteroides_distasonis_ATCC_8503_uid58301	Mesophile	435591	Bacteria	Chlorobi
Parachlamydia_acanthamoebae_UV7_uid68335	Mesophile	765952	Bacteria	Chlamydiae
Paracoccus_denitrificans_PD1222_uid58187	Mesophile	318586	Bacteria	Proteobacteria
Parvibaculum_lavamentivorans_DS_1_uid58739	Mesophile	402881	Bacteria	Proteobacteria
Parvularcula_bermudensis_HTCC2503_uid51641	Mesophile	314260	Bacteria	Proteobacteria
Pectobacterium_SCC3193_uid193707	Mesophile	1166016	Bacteria	Proteobacteria

Pediococcus_claussenii_ATCC_BAA_344_uid81103	Mesophile	701521	Bacteria	Firmicutes
Pedobacter_heparinus_DSM_2366_uid59111	Psychrophile	485917	Bacteria	Chlorobi
Pelagibacterium_halotolerans_B2_uid74393	Mesophile	1082931	Bacteria	Proteobacteria
Pelobacter_propionicus_DSM_2379_uid58255	Mesophile	338966	Bacteria	Proteobacteria
Pelodictyon_phaeoclathratiforme_BU_1_uid58173	Hyperthermophile	324925	Bacteria	Chlorobi
Persicivirga_dokdonensis_DSW_6_uid186842	Psychrophile	592029	Bacteria	Chlorobi
Phaeobacter_gallaeciensis_DSM_26640_uid232357	Mesophile	1423144	Bacteria	Proteobacteria
Phenylobacterium_zucineum_HLK1_uid58959	Mesophile	450851	Bacteria	Proteobacteria
Phycisphaera_mikurensis_NBRC_102666_uid157331	Mesophile	1142394	Bacteria	Planctomycetes
Pirellula_staleyi_DSM_6068_uid43209	Mesophile	530564	Bacteria	Planctomycetes
Planctomyces_brasiliensis_DSM_5305_uid60583	Mesophile	756272	Bacteria	Planctomycetes
Pleurocapsa_PCC_7327_uid183006	Mesophile	118163	Bacteria	Cyanobacteria
Polaromonas_JS666_uid58207	Mesophile	296591	Bacteria	Proteobacteria
Polymorphum_gilvum_SL003B_26A1_uid65447	Mesophile	991905	Bacteria	Proteobacteria
Polynucleobacter_necessarius_asymbioticus_QLW_P1DMWA_1_uid58611	Mesophile	312153	Bacteria	Proteobacteria
Porphyromonas_gingivalis_TDC60_uid67407	Thermophile	1030843	Bacteria	Chlorobi
Prevotella_ruminicola_23_uid47507	Mesophile	264731	Bacteria	Chlorobi
Propionibacterium_acidipropionici_ATCC_4875_uid179069	Mesophile	1171373	Bacteria	Actinobacteria
Prosthecochloris_aestuarii_DSM_271_uid58151	Thermophile	290512	Bacteria	Chlorobi
Proteus_mirabilis_HI4320_uid61599	Mesophile	529507	Bacteria	Proteobacteria
Providencia_stuartii_MRSN_2154_uid162193	Mesophile	1157951	Bacteria	Proteobacteria
Pseudanabaena_PCC_7367_uid183004	Mesophile	82654	Bacteria	Cyanobacteria
Pseudoalteromonas_atlantica_T6c_uid58283	Psychrophile	342610	Bacteria	Proteobacteria
Pseudogulbenkiania_NH8B_uid73423	Mesophile	748280	Bacteria	Proteobacteria
Pseudonocardia_dioxanivorans_CB1190_uid65087	Mesophile	675635	Bacteria	Actinobacteria

Pseudovibrio_FO_BEG1_uid82373	Mesophile	911045	Bacteria	Proteobacteria
Pseudoxanthomonas_spadix_BD_a59_uid75113	Psychrophile	1045855	Bacteria	Proteobacteria
Psychrobacter_G_uid210641	Mesophile	571800	Bacteria	Proteobacteria
Pusillimonas_T7_7_uid66391	Mesophile	1007105	Bacteria	Proteobacteria
Rahnella_Y9602_uid62715	Mesophile	741091	Bacteria	Proteobacteria
Ramlibacter_tataouinensis_TTB310_uid68279	Mesophile	365046	Bacteria	Proteobacteria
Raoultella_ornithinolytica_B6_uid198431	Mesophile	1286170	Bacteria	Proteobacteria
Rhizobium_leguminosarum_bv__viciae_3841_uid57955	Mesophile	216596	Bacteria	Proteobacteria
Rhodanobacter_2APBS1_uid74431	Mesophile	666685	Bacteria	Proteobacteria
Rhodobacter_sphaeroides_KD131_uid59277	Mesophile	557760	Bacteria	Proteobacteria
Rhodococcus_jostii_RHA1_uid58325	Mesophile	101510	Bacteria	Actinobacteria
Rhodoferax_ferrireducens_T118_uid58353	Mesophile	338969	Bacteria	Proteobacteria
Rhodomicrobium_vannielii_ATCC_17100_uid43247	Mesophile	648757	Bacteria	Proteobacteria
Rhodospirillum_centenum_SW_uid58805	Mesophile	414684	Bacteria	Proteobacteria
Riemerella_anatipestifer_RA_CH_1_uid175469	Mesophile	1228997	Bacteria	Chlorobi
Rivularia_PCC_7116_uid182929	Psychrophile	373994	Bacteria	Cyanobacteria
Robiginitalea_biformata_HTCC2501_uid58285	Psychrophile	313596	Bacteria	Chlorobi
Roseburia_intestinalis_XB6B4_uid197179	Mesophile	718255	Bacteria	Firmicutes
Roseiflexus_RS_1_uid58523	Thermophile	357808	Bacteria	Other
Roseobacter_litoralis_Och_149_uid54719	Psychrophile	391595	Bacteria	Proteobacteria
Rothia_dentocariosa_ATCC_17931_uid49331	Mesophile	762948	Bacteria	Actinobacteria
Rubrivivax_gelatinosus_IL144_uid158163	Mesophile	983917	Bacteria	Proteobacteria
Ruegeria_pomeroyi_DSS_3_uid57863	Psychrophile	246200	Bacteria	Proteobacteria
Ruminococcus_albus_7_uid51721	Thermophile	697329	Bacteria	Firmicutes
Runella_slithyformis_DSM_19594_uid68317	Psychrophile	761193	Bacteria	Chlorobi

Saccharomonospora_viridis_DSM_43017_uid59055	Mesophile	471857	Bacteria	Actinobacteria
Saccharophagus_degradans_2_40_uid57921	Mesophile	203122	Bacteria	Proteobacteria
Saccharopolyspora_erythraea_NRRL_2338_uid62947	Mesophile	405948	Bacteria	Actinobacteria
Saccharothrix_espanaensis_DSM_44229_uid184826	Mesophile	1179773	Bacteria	Actinobacteria
Salinarchaeum_laminariae_Harcht_Bsk1_uid207001	Mesophile	1333523	Archaea	Euryarchaeota
Salinibacter_ruber_M8_uid47323	Mesophile	761659	Bacteria	Chlorobi
Salinispora_arenicola_CNS_205_uid58659	Mesophile	391037	Bacteria	Actinobacteria
Sanguibacter_keddieii_DSM_10542_uid40845	Mesophile	446469	Bacteria	Actinobacteria
Saprospira_grandis_Lewin_uid89375	Mesophile	984262	Bacteria	Chlorobi
Sebaldella_termitidis_ATCC_33386_uid41865	Mesophile	526218	Bacteria	Fusobacteria
Segniliparus_rotundus_DSM_44985_uid49049	Mesophile	640132	Bacteria	Actinobacteria
Selenomonas_ruminantium_lactilytica_TAM6421_uid157247	Mesophile	927704	Bacteria	Firmicutes
Serratia_plymuthica_S13_uid210642	Mesophile	1348660	Bacteria	Proteobacteria
Shigella_dysenteriae_1617_uid229875	Mesophile	754093	Bacteria	Proteobacteria
Sideroxydans_lithotrophicus_ES_1_uid46801	Mesophile	580332	Bacteria	Proteobacteria
Simiduia_agarivorans_SA1_uid177713	Psychrophile	1117647	Bacteria	Proteobacteria
Simkania_negevensis_Z_uid68451	Mesophile	331113	Bacteria	Chlamydiae
Singulisphaera_acidiphila_DSM_18658_uid81777	Thermophile	886293	Bacteria	Planctomycetes
Slackia_heliotrinireducens_DSM_20476_uid59051	Mesophile	471855	Bacteria	Actinobacteria
Sodalis_glossinidius__morsitans__uid58553	Mesophile	343509	Bacteria	Proteobacteria
Solibacillus_silvestris_StLB046_uid168516	Mesophile	1002809	Bacteria	Firmicutes
Solitalea_canadensis_DSM_3403_uid81783	Psychrophile	929556	Bacteria	Chlorobi
Sorangium_cellulosum_So0157_2_uid210741	Mesophile	1254432	Bacteria	Proteobacteria
Sphaerochaeta_pleomorpha_Grapes_uid82365	Mesophile	158190	Bacteria	Spirochaetes
Sphingobacterium_21_uid64755	Psychrophile	743722	Bacteria	Chlorobi

Sphingobium_japonicum_UT26S_uid47077	Mesophile	452662	Bacteria	Proteobacteria
Sphingomonas_wittichii_RW1_uid58691	Mesophile	392499	Bacteria	Proteobacteria
Sphingopyxis_alaskensis_RB2256_uid58351	Mesophile	317655	Bacteria	Proteobacteria
Spiribacter_UAH_SP71_uid226111	Mesophile	1335757	Bacteria	Proteobacteria
Spirochaeta_smaragdinae_DSM_11293_uid51369	Mesophile	573413	Bacteria	Spirochaetes
Spiroplasma_chrysopicola_DF_1_uid205053	Mesophile	1276227	Bacteria	Other
Spirosoma_linguale_DSM_74_uid43413	Psychrophile	504472	Bacteria	Chlorobi
Stackebrandtia_nassauensis_DSM_44728_uid46663	Psychrophile	446470	Bacteria	Actinobacteria
Stanieria_cyanosphaera_PCC_7437_uid183115	Mesophile	111780	Bacteria	Cyanobacteria
Starkeya_novella_DSM_506_uid48815	Mesophile	639283	Bacteria	Proteobacteria
Stenotrophomonas_maltophilia_K279a_uid61647	Mesophile	522373	Bacteria	Proteobacteria
Stigmatella_aurantiaca_DW4_3_1_uid158509	Mesophile	378806	Bacteria	Proteobacteria
Strawberry_lethal_yellows_phytoplasma__CPA__NZSb11_uid203392	Mesophile	980422	Bacteria	Other
Streptobacillus_moniliformis_DSM_12112_uid41863	Mesophile	519441	Bacteria	Fusobacteria
Streptomyces_bingchenggensis_BCW_1_uid82931	Mesophile	749414	Bacteria	Actinobacteria
Streptosporangium_roseum_DSM_43021_uid42521	Mesophile	479432	Bacteria	Actinobacteria
Sulfobacillus_acidophilus_TPY_uid68841	Thermophile	1051632	Bacteria	Firmicutes
Sulfuricella_denitrificans_skB26_uid170240	Mesophile	1163617	Bacteria	Proteobacteria
Sulfuricurvum_kujiense_DSM_16994_uid60789	Mesophile	709032	Bacteria	Proteobacteria
Sulfurimonas_autotrophica_DSM_16294_uid53043	Mesophile	563040	Bacteria	Proteobacteria
Sulfurospirillum_barnesii_SES_3_uid168117	Mesophile	760154	Bacteria	Proteobacteria
Sulfurovum_NBC37_1_uid58863	Mesophile	387093	Bacteria	Proteobacteria
Synechococcus_PCC_6312_uid182934	Hyperthermophile	195253	Bacteria	Cyanobacteria
Synergistetes_bacterium_SGP1_uid197182	Thermophile	651822	Bacteria	Synergistetes
Syntrophobacter_fumaroxidans_MPOB_uid58177	Mesophile	335543	Bacteria	Proteobacteria

Syntrophobotulus_glycolicus_DSM_8271_uid63343	Thermophile	645991	Bacteria	Firmicutes
Syntrophomonas_wolfei_Goettingen_uid58179	Thermophile	335541	Bacteria	Firmicutes
Syntrophus_aciditrophicus_SB_uid58539	Mesophile	56780	Bacteria	Proteobacteria
Tannerella_forsythia_ATCC_43037_uid83157	Mesophile	203275	Bacteria	Chlorobi
Taylorella_equigenitalis_MCE9_uid62103	Mesophile	937774	Bacteria	Proteobacteria
Tepidanaerobacter_acetatoxydans_Re1_uid184827	Thermophile	1209989	Bacteria	Firmicutes
Teredinibacter_turnerae_T7901_uid59267	Mesophile	377629	Bacteria	Proteobacteria
Terriglobus_roseus_DSM_18391_uid168183	Mesophile	926566	Bacteria	Acidobacteria
Tetragenococcus_halophilus_uid74441	Mesophile	945021	Bacteria	Firmicutes
Thalassobaculum_L2_uid182483	Mesophile	1193729	Bacteria	Proteobacteria
Thalassolituus_oleivorans_MIL_1_uid195604	Mesophile	1298593	Bacteria	Proteobacteria
Thauera_MZ1T_uid58987	Mesophile	85643	Bacteria	Proteobacteria
Thermincola_potens_JR_uid48823	Thermophile	635013	Bacteria	Firmicutes
Thermoanaerobacterium_thermosaccharolyticum_M0795_uid184821	Thermophile	698948	Bacteria	Firmicutes
Thermobacillus_composti_KWC4_uid74021	Mesophile	717605	Bacteria	Firmicutes
Thermobaculum_terrenum_ATCC_BAA_798_uid42011	Thermophile	525904	Bacteria	Other
Thermobifida_fusca_YX_uid57703	Mesophile	269800	Bacteria	Actinobacteria
Thermobispora_bispora_DSM_43833_uid48999	Mesophile	469371	Bacteria	Actinobacteria
Thermodesulfobium_narugense_DSM_14796_uid66601	Thermophile	747365	Bacteria	Firmicutes
Thermofilum_1910b_uid215374	Hyperthermophile	1365176	Archaea	Crenarchaeota
Thermomonospora_curvata_DSM_43183_uid41885	Thermophile	471852	Bacteria	Actinobacteria
Thermoplasmatales_archaeon_BRNA1_uid195930	Mesophile	1054217	Archaea	Euryarchaeota
Thioalkalimicrobium_cyclicum_ALM1_uid67391	Mesophile	717773	Bacteria	Proteobacteria
Thioalkalivibrio_nitratireducens_DSM_14787_uid184011	Mesophile	1255043	Bacteria	Proteobacteria
Thiocystis_violascens_DSM_198_uid74025	Mesophile	765911	Bacteria	Proteobacteria

Thioflavicoccus_mobilis_8321_uid184343	Mesophile	765912	Bacteria	Proteobacteria
Thiomonas_3As_uid178369	Mesophile	426114	Bacteria	Proteobacteria
Tistrella_mobilis_KA081020_065_uid167486	Mesophile	1110502	Bacteria	Proteobacteria
Tolumonas_auensis_DSM_9187_uid59395	Mesophile	595494	Bacteria	Proteobacteria
Trichodesmium_erythraeum_IMS101_uid57925	Mesophile	203124	Bacteria	Cyanobacteria
Truepera_radiovictrix_DSM_17093_uid49533	Thermophile	649638	Bacteria	Deinococcus-Thermus
Tsukamurella_paurometabola_DSM_20162_uid48829	Mesophile	521096	Bacteria	Actinobacteria
Turneriella_parva_DSM_21527_uid168321	Mesophile	869212	Bacteria	Spirochaetes
Variovorax_paradoxus_B4_uid218005	Mesophile	1246301	Bacteria	Proteobacteria
Veillonella_parvula_DSM_2008_uid41927	Mesophile	479436	Bacteria	Firmicutes
Verminephrobacter_eiseniae_EF01_2_uid58675	Mesophile	391735	Bacteria	Proteobacteria
Verrucosispora_maris_AB_18_032_uid66297	Thermophile	263358	Bacteria	Actinobacteria
Waddlia_chondrophila_WSU_86_1044_uid49531	Mesophile	716544	Bacteria	Chlamydiae
Weeksella_virosa_DSM_16922_uid63627	Psychrophile	865938	Bacteria	Chlorobi
Weissella_koreensis_KACC_15510_uid68837	Mesophile	1045854	Bacteria	Firmicutes
Wolbachia_endosymbiont_of_Culex_quinquefasciatus_Pel_uid61645	Mesophile	570417	Bacteria	Proteobacteria
Wolbachia_wRi_uid59371	Mesophile	66084	Bacteria	Proteobacteria
Xanthobacter_autotrophicus_Py2_uid58453	Mesophile	78245	Bacteria	Proteobacteria
Xanthomonas_oryzae_PXO99A_uid59131	Mesophile	360094	Bacteria	Proteobacteria
Xenorhabdus_nematophila_ATCC_19061_uid49133	Mesophile	406817	Bacteria	Proteobacteria
Xylanimonas_cellulosilytica_DSM_15894_uid41935	Mesophile	446471	Bacteria	Actinobacteria
Zobellia_galactanivorans_uid70621	Psychrophile	63186	Bacteria	Chlorobi
Zunongwangia_profunda_SM_A87_uid48073	Psychrophile	655815	Bacteria	Chlorobi
Zymomonas_mobilis_NCIMB_11163_uid41019	Mesophile	622759	Bacteria	Proteobacteria
archaeon_Mx1201_uid196597	Mesophile	1236689	Archaea	Euryarchaeota

candidate_division_SR1_bacterium_RAAC1_SR1_1_uid230714	Mesophile	1394709	Bacteria	Other
candidate_division_WWE3_bacterium_RAAC2_WWE3_1_uid230713	Mesophile	1394710	Bacteria	Other
secondary_endosymbiont_of_Ctenarytaina_eucalypti_uid172737	Mesophile	1199245	Bacteria	Proteobacteria
secondary_endosymbiont_of_Heteropsylla_cubana_Thao2000_uid172738	Mesophile	134287	Bacteria	Proteobacteria
syncytium_symbiont_of_Diaphorina_citri_uid213384	Mesophile	669502	Bacteria	Proteobacteria
uncultured_Termite_group_1_bacterium_phylotype_Rs_D17_uid59059	Thermophile	471821	Bacteria	Other

Appendix 5. Names, coefficients, functional categories, and biological functions of features chosen by logistic regression models.

Prediction of mesophiles.

COGs	**Coefs**	**Category**	**Function**
COG2249	1.751	R	Putative NADPH-quinone reductase (modulator of drug activity B)
COG0207	1.154	F	Thymidylate synthase
COG3794	0.834	C	Plastocyanin
COG1621	0.794	G	Sucrose-6-phosphate hydrolase SacC GH32 family
COG0280	0.787	C	Phosphotransacetylase
COG3299	0.785	X	Uncharacterized phage protein gp47/JayE
COG3315	0.737	Q	O-Methyltransferase involved in polyketide biosynthesis
COG3905	0.733	K	Predicted transcriptional regulator
COG2770	0.726	T	HAMP domain
COG2039	0.675	O	Pyrrolidone-carboxylate peptidase (N-terminal pyroglutamyl peptidase)
COG1570	0.664	L	Exonuclease VII large subunit
COG0582	0.652	LX	Integrase
COG0120	0.620	G	Ribose 5-phosphate isomerase
COG4166	0.612	E	ABC-type oligopeptide transport system periplasmic component
COG0588	0.606	G	Phosphoglycerate mutase (BPG-dependent)
COG2183	0.561	K	Transcriptional accessory protein Tex/SPT6

COG2945	0.520	R	Alpha/beta superfamily hydrolase
COG0262	0.467	H	Dihydrofolate reductase
COG4372	0.322	S	Uncharacterized conserved protein contains DUF3084 domain
COG0556	0.321	L	Excinuclease UvrABC helicase subunit UvrB
COG0500	0.315	QR	SAM-dependent methyltransferase
COG0571	0.309	K	dsRNA-specific ribonuclease
COG0178	0.301	L	Excinuclease UvrABC ATPase subunit
COG1502	0.290	I	Phosphatidylserine/phosphatidylglycerophos-phate/cardiolipin synthase or related enzyme
COG2230	0.277	I	Cyclopropane fatty-acyl-phospholipid synthase and related methyltransferases
COG0188	0.268	L	DNA gyrase/topoisomerase IV subunit A
COG0834	0.264	ET	ABC-type amino acid transport/signal transduction system periplasmic component/domain
COG2978	0.258	H	p-Aminobenzoyl-glutamate transporter AbgT
COG0652	0.257	O	Peptidyl-prolyl cis-trans isomerase (rotamase) - cy-clophilin family
COG1164	0.242	E	Oligoendopeptidase F
COG0575	0.238	I	CDP-diglyceride synthetase
COG4188	0.223	R	Predicted dienelactone hydrolase
COG3864	0.220	R	Predicted metal-dependent peptidase
COG1453	0.213	R	Predicted oxidoreductase of the aldo/keto reductase family
COG1178	0.198	P	ABC-type Fe3+ transport system permease compo-nent
COG0229	0.192	O	Peptide methionine sulfoxide reductase MsrB
COG1346	0.190	M	Putative effector of murein hydrolase
COG0534	0.189	V	Na+-driven multidrug efflux pump
COG0322	0.177	L	Excinuclease UvrABC nuclease subunit
COG3153	0.176	R	Predicted N-acetyltransferase YhbS
COG3467	0.175	V	Nitroimidazol reductase NimA or a related FMN-containing flavoprotein pyridoxamine 5'-phosphate oxidase superfamily

COG0398	0.173	S	Uncharacterized membrane protein YdjX TVP38/TMEM64 family SNARE-associated domain
COG0765	0.173	E	ABC-type amino acid transport system permease component
COG1509	0.173	E	L-lysine 23-aminomutase (EF-P beta-lysylation pathway)
COG4122	0.158	R	Predicted O-methyltransferase YrrM
COG1722	0.155	L	Exonuclease VII small subunit
COG4251	0.149	T	Bacteriophytochrome (light-regulated signal transduction histidine kinase)
COG0551	0.142	L	ssDNA-binding Zn-finger and Zn-ribbon domains of topoisomerase 1
COG0576	0.141	O	Molecular chaperone GrpE (heat shock protein)
COG3049	0.135	MR	Penicillin V acylase or related amidase Ntn superfamily
COG1696	0.132	M	D-alanyl-lipoteichoic acid acyltransferase DltB MBOAT superfamily
COG0389	0.109	L	Nucleotidyltransferase/DNA polymerase involved in DNA repair
COG0733	0.101	R	Na+-dependent transporter SNF family
COG0484	0.097	O	DnaJ-class molecular chaperone with C-terminal Zn finger domain
COG0514	0.094	L	Superfamily II DNA helicase RecQ
COG0605	0.093	P	Superoxide dismutase
COG0553	0.091	KL	Superfamily II DNA or RNA helicase SNF2 family
COG0225	0.091	O	Peptide methionine sulfoxide reductase MsrA
COG3549	0.089	V	Plasmid maintenance system killer protein
COG1511	0.088	S	Uncharacterized membrane protein YhgE phage infection protein (PIP) family
COG2227	0.087	H	2-polyprenyl-3-methyl-5-hydroxy-6-metoxy-14-benzoquinol methylase
COG1285	0.071	S	Uncharacterized membrane protein YhiD involved in acid resistance

COG0187	0.062	L	DNA gyrase/topoisomerase IV subunit B
COG0443	0.062	O	Molecular chaperone DnaK (HSP70)
COG1704	0.058	S	Uncharacterized conserved protein
COG0591	0.052	E	Na+/proline symporter
COG1301	0.048	C	Na+/H+-dicarboxylate symporter
COG1283	0.047	P	Na+/phosphate symporter
COG0786	0.046	E	Na+/glutamate symporter
COG2002	0.031	KV	Bifunctional DNA-binding transcriptional regulator of stationary/sporulation/toxin gene expression and antitoxin component of the YhaV-PrlF toxin-antitoxin module
COG0783	0.030	PV	DNA-binding ferritin-like protein (oxidative damage protectant)
COG1253	0.029	R	Hemolysin or related protein contains CBS domains
COG1063	0.009	ER	Threonine dehydrogenase or related Zn-dependent dehydrogenase
COG0272	0.002	L	NAD-dependent DNA ligase
COG4148	-0.002	P	ABC-type molybdate transport system ATPase component
COG1703	-0.004	O	Putative periplasmic protein kinase ArgK or related GTPase of G3E family
COG1801	-0.005	S	Uncharacterized conserved protein YecE DUF72 family
COG2327	-0.005	M	Polysaccharide pyruvyl transferase family protein WcaK
COG0455	-0.011	DN	MinD-like ATPase involved in chromosome partitioning or flagellar assembly
COG2177	-0.018	D	Cell division protein FtsX
COG3225	-0.020	N	ABC-type uncharacterized transport system involved in gliding motility auxiliary component
COG0314	-0.030	H	Molybdopterin synthase catalytic subunit
COG1399	-0.030	S	Uncharacterized metal-binding protein YceD DUF177 family

COG1129	-0.037	G	ABC-type sugar transport system ATPase component
COG3450	-0.037	R	Predicted enzyme of the cupin superfamily
COG0411	-0.039	E	ABC-type branched-chain amino acid transport system ATPase component
COG1743	-0.039	L	Adenine-specific DNA methylase contains a Zn-ribbon domain
COG0374	-0.045	C	NiFe-hydrogenase I large subunit
COG1172	-0.047	G	Ribose/xylose/arabinose/galactoside ABC-type transport system permease component
COG1609	-0.056	K	DNA-binding transcriptional regulator LacI/PurR family
COG1950	-0.063	S	Uncharacterized membrane protein YvlD DUF360 family
COG3654	-0.063	X	Prophage maintenance system killer protein
COG2096	-0.065	H	Cob(I)alamin adenosyltransferase
COG2896	-0.071	H	Molybdenum cofactor biosynthesis enzyme MoaA
COG0303	-0.072	H	Molybdopterin biosynthesis enzyme
COG1079	-0.072	R	ABC-type uncharacterized transport system permease component
COG1032	-0.075	R	Radical SAM superfamily enzyme YgiQ UPF0313 family
COG1397	-0.082	O	ADP-ribosylglycohydrolase
COG2905	-0.092	T	Signal-transduction protein containing cAMP-binding CBS and nucleotidyltransferase domains
COG0005	-0.093	F	Purine nucleoside phosphorylase
COG2086	-0.096	C	Electron transfer flavoprotein alpha and beta subunits
COG0428	-0.099	P	Zinc transporter ZupT
COG4603	-0.119	R	ABC-type uncharacterized transport system permease component
COG1986	-0.131	FV	Non-canonical (house-cleaning) NTP pyrophosphatase all-alpha NTP-PPase family
COG3439	-0.154	S	Uncharacterized conserved protein DUF302 family

COG1540	-0.157	R	Lactam utilization protein B (function unknown)
COG0490	-0.160	P	K+/H+ antiporter YhaU regulatory subunit KhtT
COG1055	-0.161	P	Na+/H+ antiporter NhaD or related arsenite permease
COG2120	-0.182	G	N-acetylglucosaminyl deacetylase LmbE family
COG4454	-0.186	R	Uncharacterized copper-binding protein cupredoxin-like subfamily
COG2025	-0.190	C	Electron transfer flavoprotein alpha subunit
COG0038	-0.191	P	H+/Cl- antiporter ClcA
COG0819	-0.216	H	Thiaminase
COG1518	-0.224	V	CRISPR/Cas system-associated endonuclease Cas1
COG0315	-0.242	H	Molybdenum cofactor biosynthesis enzyme
COG2971	-0.249	G	BadF-type ATPase related to human N-acetylglucosamine kinase
COG3956	-0.259	R	Uncharacterized conserved protein YabN contains tetrapyrrole methylase and MazG-like pyrophosphatase domain
COG1410	-0.261	E	Methionine synthase I cobalamin-binding domain
COG3677	-0.267	X	Transposase
COG1647	-0.267	Q	Esterase/lipase
COG0123	-0.270	BQ	Acetoin utilization deacetylase AcuC or a related deacetylase
COG3206	-0.272	M	Uncharacterized protein involved in exopolysaccharide biosynthesis
COG2723	-0.275	G	Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase
COG0296	-0.276	G	14-alpha-glucan branching enzyme
COG1526	-0.301	C	Formate dehydrogenase assembly factor FdhD
COG1349	-0.305	KG	DNA-binding transcriptional regulator of sugar metabolism DeoR/GlpR family
COG4989	-0.324	R	Predicted oxidoreductase
COG3653	-0.326	Q	N-acyl-D-aspartate/D-glutamate deacylase

COG3301	-0.350	P	Formate-dependent nitrite reductase membrane component NrfD
COG0709	-0.355	E	Selenophosphate synthase
COG1922	-0.369	M	UDP-N-acetyl-D-mannosaminuronic acid transferase WecB/TagA/CpsF family
COG0182	-0.373	E	Methylthioribose-1-phosphate isomerase (methionine salvage pathway) a paralog of eIF-2B alpha subunit
COG4942	-0.383	D	Septal ring factor EnvC activator of murein hydrolases AmiA and AmiB
COG1792	-0.386	D	Cell shape-determining protein MreC
COG0243	-0.391	C	Anaerobic selenocysteine-containing dehydrogenase
COG2068	-0.393	H	CTP:molybdopterin cytidylyltransferase MocA
COG3959	-0.394	G	Transketolase N-terminal subunit
COG1583	-0.407	V	CRISPR/Cas system endoribonuclease Cas6 RAMP superfamily
COG1228	-0.413	Q	Imidazolonepropionase or related amidohydrolase
COG1814	-0.444	P	Predicted Fe2+/Mn2+ transporter VIT1/CCC1 family
COG1085	-0.469	G	Galactose-1-phosphate uridylyltransferase
COG4242	-0.481	QR	Cyanophycinase and related exopeptidases
COG1702	-0.492	T	Phosphate starvation-inducible protein PhoH predicted ATPase
COG1203	-0.528	V	CRISPR/Cas system-associated endonuclease/helicase Cas3
COG0698	-0.535	G	Ribose 5-phosphate isomerase RpiB
COG1077	-0.551	D	Actin-like ATPase involved in cell morphogenesis
COG0644	-0.588	C	Dehydrogenase (flavoprotein)
COG2981	-0.593	E	Uncharacterized protein involved in cysteine biosynthesis
COG0421	-0.594	E	Spermidine synthase
COG3958	-0.746	G	Transketolase C-terminal subunit
COG1318	-0.747	K	Predicted transcriptional regulator

COG2759	-0.863	F	Formyltetrahydrofolate synthetase
COG2986	-0.915	E	Histidine ammonia-lyase
COG1883	-0.916	C	Na+-transporting methylmalonyl-CoA/oxaloacetate decarboxylase beta subunit
COG1501	-1.007	G	Alpha-glucosidase glycosyl hydrolase family GH31
COG3167	-1.064	NW	Tfp pilus assembly protein PilO
COG1110	-1.291	L	Reverse gyrase
COG2930	-1.325	I	Lipid-binding SYLF domain
COG2382	-1.440	P	Enterochelin esterase or related enzyme
COG1857	-1.471	V	CRISPR/Cas system-associated protein Cas7 RAMP superfamily

Prediction of thermophiles.

COGs	**Coefs**	**Category**	**Function**
COG1583	1.197	V	CRISPR/Cas system endoribonuclease Cas6 RAMP superfamily
COG1550	0.552	S	Uncharacterized conserved protein YlxP DUF503 family
COG0846	0.468	O	NAD-dependent protein deacetylase SIR2 family
COG2316	0.408	R	Predicted hydrolase HD superfamily
COG1866	0.379	C	Phosphoenolpyruvate carboxykinase ATP-dependent
COG1743	0.296	L	Adenine-specific DNA methylase contains a Zn-ribbon domain
COG1387	0.269	ER	Histidinol phosphatase or related hydrolase of the PHP family
COG1085	0.256	G	Galactose-1-phosphate uridylyltransferase
COG1937	0.165	K	DNA-binding transcriptional regulator FrmR family
COG1993	0.158	T	PII-like signaling protein
COG2805	0.148	NW	Tfp pilus assembly protein PilT pilus retraction ATPase

COG1894	0.140	C	NADH:ubiquinone oxidoreductase NADH-binding 51 kD subunit (chain F)
COG1922	0.136	M	UDP-N-acetyl-D-mannosaminuronic acid transferase WecB/TagA/CpsF family
COG1765	0.092	R	Uncharacterized OsmC-related protein
COG1884	0.088	I	Methylmalonyl-CoA mutase N-terminal domain/subunit
COG3391	0.079	R	DNA-binding beta-propeller fold protein YncE
COG1487	0.073	R	Predicted nucleic acid-binding protein contains PIN domain
COG2894	0.057	D	Septum formation inhibitor-activating ATPase MinD
COG0421	0.056	E	Spermidine synthase
COG3959	0.051	G	Transketolase N-terminal subunit
COG1905	0.043	C	NADH:ubiquinone oxidoreductase 24 kD subunit (chain E)
COG4636	0.039	R	Endonuclease Uma2 family (restriction endonuclease fold)
COG3958	0.023	G	Transketolase C-terminal subunit
COG1774	0.016	T	Cell fate regulator YaaT PSP1 superfamily (controls sporulation competence biofilm development)
COG0432	0.001	H	Thiamin phosphate synthase YjbQ UPF0047 family
COG0670	-0.007	R	Integral membrane protein interacts with FtsH
COG0454	-0.009	KR	N-acetyltransferase GNAT superfamily (includes histone acetyltransferase HPA2)
COG0326	-0.010	O	Molecular chaperone HSP90 family
COG0560	-0.032	E	Phosphoserine phosphatase
COG0400	-0.040	R	Predicted esterase
COG2183	-0.058	K	Transcriptional accessory protein Tex/SPT6
COG1243	-0.065	KB	Histone acetyltransferase component of the RNA polymerase elongator complex
COG1340	-0.070	S	Uncharacterized coiled-coil protein contains DUF342 domain

COG0633	-0.072	C	Ferredoxin
COG1881	-0.089	R	Uncharacterized conserved protein phosphatidyl-ethanolamine-binding protein (PEBP) family
COG2227	-0.133	H	2-polyprenyl-3-methyl-5-hydroxy-6-metoxy-14-benzoquinol methylase
COG0575	-0.144	I	CDP-diglyceride synthetase
COG0443	-0.145	O	Molecular chaperone DnaK (HSP70)
COG3356	-0.165	I	Predicted membrane-associated lipid hydrolase neutral ceramidase superfamily
COG0484	-0.165	O	DnaJ-class molecular chaperone with C-terminal Zn finger domain
COG0708	-0.169	L	Exonuclease III
COG0076	-0.190	E	Glutamate or tyrosine decarboxylase or a related PLP-dependent protein
COG1056	-0.194	H	Nicotinamide mononucleotide adenylyltransferase
COG0262	-0.203	H	Dihydrofolate reductase
COG0318	-0.228	IQ	Acyl-CoA synthetase (AMP-forming)/AMP-acid ligase II
COG0212	-0.230	H	5-formyltetrahydrofolate cyclo-ligase
COG0652	-0.233	O	Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclophilin family
COG4591	-0.350	M	ABC-type transport system involved in lipoprotein release permease component
COG0605	-0.380	P	Superoxide dismutase
COG0229	-0.381	O	Peptide methionine sulfoxide reductase MsrB
COG0464	-0.383	MDT	AAA+-type ATPase SpoVK/Ycf46/Vps4 family
COG0431	-0.458	C	NAD(P)H-dependent FMN reductase
COG0207	-1.254	F	Thymidylate synthase

Prediction of hyperthermophiles.

COGs	**Coefs**	**Category**	**Function**
COG0863	1.531	L	DNA modification methylase

COG2014	1.432	S	Uncharacterized conserved protein contains DUF4213 and DUF364 domains
COG2723	1.210	G	Beta-glucosidase/6-phospho-beta-gluco-sidase/beta-galactosidase
COG2034	1.094	S	Uncharacterized membrane protein
COG0464	1.067	MDT	AAA+-type ATPase SpoVK/Ycf46/Vps4 family
COG2703	1.000	T	Hemerythrin
COG0121	0.925	R	Predicted glutamine amidotransferase
COG1472	0.922	G	Periplasmic beta-glucosidase and related glyco-sidases
COG1209	0.851	M	dTDP-glucose pyrophosphorylase
COG1619	0.842	M	Muramoyltetrapeptide carboxypeptidase LdcA (peptidoglycan recycling)
COG2084	0.826	I	3-hydroxyisobutyrate dehydrogenase or related beta-hydroxyacid dehydrogenase
COG5557	0.692	C	Ni/Fe-hydrogenase 2 integral membrane subunit HybB
COG0003	0.678	P	Anion-transporting ATPase ArsA/GET3 family
COG1063	0.649	ER	Threonine dehydrogenase or related Zn-dependent dehydrogenase
COG0633	0.572	C	Ferredoxin
COG3339	0.535	S	Uncharacterized membrane protein YkvA DUF1232 family
COG0170	0.522	O	Dolichol kinase
COG3158	0.491	P	K+ transporter
COG1892	0.489	G	Phosphoenolpyruvate carboxylase
COG2410	0.463	R	Predicted nuclease (RNAse H fold)
COG1005	0.427	C	NADH:ubiquinone oxidoreductase subunit 1 (chain H)
COG3023	0.409	M	N-acetyl-anhydromuramyl-L-alanine amidase AmpD
COG1271	0.401	C	Cytochrome bd-type quinol oxidase subunit 1
COG1725	0.396	K	DNA-binding transcriptional regulator YhcF GntR family

COG4152	0.356	R	ABC-type uncharacterized transport system ATPase component
COG3404	0.341	E	Formiminotetrahydrofolate cyclodeaminase
COG3385	0.335	X	IS4 transposase
COG1836	0.313	S	Uncharacterized membrane protein
COG3876	0.312	S	Uncharacterized conserved protein YbbC DUF1343 family
COG1403	0.285	V	5-methylcytosine-specific restriction endonuclease McrA
COG1324	0.283	P	Uncharacterized protein involved in tolerance to divalent cations
COG3012	0.280	S	Uncharacterized conserved protein YchJ contains N- and C-terminal SEC-C domains
COG3255	0.274	I	Putative sterol carrier protein
COG3934	0.250	G	Endo-14-beta-mannosidase
COG1647	0.245	Q	Esterase/lipase
COG1986	0.243	FV	Non-canonical (house-cleaning) NTP pyrophos-phatase all-alpha NTP-PPase family
COG2164	0.227	S	Uncharacterized protein
COG1091	0.227	M	dTDP-4-dehydrorhamnose reductase
COG1489	0.222	GT	DNA-binding protein stimulates sugar fermenta-tion
COG2010	0.201	C	Cytochrome c mono- and diheme variants
COG0075	0.191	EF	Archaeal aspartate aminotransferase or a related aminotransferase includes purine catabolism pro-tein PucG
COG3328	0.190	X	Transposase (or an inactivated derivative)
COG1463	0.185	M	ABC-type transporter Mla maintaining outer mem-brane lipid asymmetry periplasmic component MlaD
COG1950	0.177	S	Uncharacterized membrane protein YvlD DUF360 family
COG0286	0.174	V	Type I restriction-modification system DNA methylase subunit

COG2270	0.171	R	MFS-type transporter involved in bile tolerance Atg22 family
COG1401	0.171	V	5-methylcytosine-specific restriction endonuclease McrBC GTP-binding regulatory subunit McrB
COG1850	0.148	G	Ribulose 15-bisphosphate carboxylase large subunit or a RuBisCO-like protein
COG2971	0.140	G	BadF-type ATPase related to human N-acetylglucosamine kinase
COG0699	0.128	L	Replication fork clamp-binding protein CrfC (dynamin-like GTPase family)
COG1881	0.125	R	Uncharacterized conserved protein phosphatidylethanolamine-binding protein (PEBP) family
COG1392	0.119	S	Uncharacterized conserved protein YkaA distantly related to PhoU UPF0111/DUF47 family
COG0475	0.118	P	Kef-type K+ transport system membrane component KefB
COG1804	0.113	I	Crotonobetainyl-CoA:carnitine CoA-transferase CaiB and related acyl-CoA transferases
COG2062	0.104	T	Phosphohistidine phosphatase SixA
COG1654	0.093	K	Biotin operon repressor
COG1166	0.089	E	Arginine decarboxylase (spermidine biosynthesis)
COG0490	0.082	P	K+/H+ antiporter YhaU regulatory subunit KhtT
COG1148	0.080	C	Heterodisulfide reductase subunit A (polyferredoxin)
COG5012	0.077	C	Methanogenic corrinoid protein MtbC1
COG1930	0.075	P	ABC-type cobalt transport system periplasmic component
COG1526	0.074	C	Formate dehydrogenase assembly factor FdhD
COG1853	0.069	C	NADH-FMN oxidoreductase RutF flavin reductase (DIM6/NTAB) family
COG5662	0.059	K	Transmembrane transcriptional regulator (anti-sigma factor RsiW)
COG0515	0.056	T	Serine/threonine protein kinase
COG1874	0.051	G	Beta-galactosidase GanA

COG1528	0.042	P	Ferritin
COG1702	0.033	T	Phosphate starvation-inducible protein PhoH predicted ATPase
COG2083	0.029	S	Uncharacterized protein UPF0216 family
COG0345	0.028	E	Pyrroline-5-carboxylate reductase
COG1492	0.027	H	Cobyric acid synthase
COG0318	0.025	IQ	Acyl-CoA synthetase (AMP-forming)/AMP-acid ligase II
COG0560	0.008	E	Phosphoserine phosphatase
COG3894	0.008	S	Uncharacterized 2Fe-2 and 4Fe-4S clusters-containing protein contains DUF4445 domain
COG1646	0.007	I	Heptaprenylglyceryl phosphate synthase
COG1058	0.0004	R	Predicted nucleotide-utilizing enzyme related to molybdopterin-biosynthesis enzyme MoeA
COG1237	-0.0002	R	Metal-dependent hydrolase beta-lactamase superfamily II
COG0760	-0.0005	O	Parvulin-like peptidyl-prolyl isomerase
COG0717	-0.003	F	Deoxycytidine triphosphate deaminase
COG2834	-0.004	M	Outer membrane lipoprotein-sorting protein
COG1048	-0.005	C	Aconitase A
COG2063	-0.005	N	Flagellar basal body L-ring protein FlgH
COG0413	-0.006	H	Ketopantoate hydroxymethyltransferase
COG0568	-0.010	K	DNA-directed RNA polymerase sigma subunit (sigma70/sigma32)
COG1765	-0.011	R	Uncharacterized OsmC-related protein
COG0819	-0.012	H	Thiaminase
COG2159	-0.012	R	Predicted metal-dependent hydrolase TIM-barrel fold
COG0134	-0.013	E	Indole-3-glycerol phosphate synthase
COG1295	-0.014	S	Uncharacterized membrane protein BrkB/YihY/UPF0761 family (not an RNase)
COG3595	-0.019	S	Uncharacterized conserved protein YvlB contains DUF4097 and DUF4098 domains

COG1160	-0.022	R	Predicted GTPases
COG1176	-0.029	E	ABC-type spermidine/putrescine transport system permease component I
COG1476	-0.032	K	DNA-binding transcriptional regulator XRE-family HTH domain
COG1198	-0.032	L	Primosomal protein N' (replication factor Y) - superfamily II helicase
COG0295	-0.040	F	Cytidine deaminase
COG1894	-0.041	C	NADH:ubiquinone oxidoreductase NADH-binding 51 kD subunit (chain F)
COG0781	-0.042	K	Transcription termination factor NusB
COG0493	-0.044	ER	NADPH-dependent glutamate synthase beta chain or related oxidoreductase
COG0478	-0.046	T	RIO-like serine/threonine protein kinase fused to N-terminal HTH domain
COG0624	-0.047	E	Acetylornithine deacetylase/Succinyl-diaminopimelate desuccinylase or related deacylase
COG1301	-0.048	C	Na+/H+-dicarboxylate symporter
COG1132	-0.052	V	ABC-type multidrug transport system ATPase and permease component
COG1510	-0.056	K	DNA-binding transcriptional regulator GbsR MarR family
COG0592	-0.060	L	DNA polymerase III sliding clamp (beta) subunit PCNA homolog
COG1706	-0.063	N	Flagellar basal body P-ring protein FlgI
COG0307	-0.071	H	Riboflavin synthase alpha chain
COG1757	-0.071	C	Na+/H+ antiporter NhaC
COG1378	-0.081	K	Sugar-specific transcriptional regulator TrmB
COG1893	-0.094	H	Ketopantoate reductase
COG0206	-0.095	D	Cell division GTPase FtsZ
COG3951	-0.097	N	Rod binding protein domain
COG1636	-0.102	R	Predicted ATPase Adenine nucleotide alpha hydrolases (AANH) superfamily

COG1396	-0.104	K	Transcriptional regulator contains XRE-family HTH domain
COG1031	-0.105	R	Radical SAM superfamily enzyme with C-terminal helix-hairpin-helix motif
COG0502	-0.119	H	Biotin synthase or related enzyme
COG0846	-0.121	O	NAD-dependent protein deacetylase SIR2 family
COG1585	-0.143	O	Membrane protein implicated in regulation of membrane protease activity
COG0785	-0.144	CO	Cytochrome c biogenesis protein CcdA
COG0745	-0.151	TK	DNA-binding response regulator OmpR family contains REC and winged-helix (wHTH) domain
COG1592	-0.157	C	Rubrerythrin
COG1985	-0.159	H	Pyrimidine reductase riboflavin biosynthesis
COG1158	-0.160	K	Transcription termination factor Rho
COG0840	-0.164	NT	Methyl-accepting chemotaxis protein
COG1134	-0.177	GM	ABC-type polysaccharide/polyol phosphate transport system ATPase component
COG0618	-0.191	F	nanoRNase/pAp phosphatase hydrolyzes c-di-AMP and oligoRNAs
COG3174	-0.192	S	Uncharacterized membrane protein DUF4010 family
COG1242	-0.198	R	Radical SAM superfamily enzyme
COG2336	-0.198	T	Antitoxin component of the MazEF toxin-antitoxin module
COG1177	-0.204	E	ABC-type spermidine/putrescine transport system permease component II
COG1527	-0.216	C	Archaeal/vacuolar-type H+-ATPase subunit C/Vma6
COG1253	-0.225	R	Hemolysin or related protein contains CBS domains
COG0322	-0.235	L	Excinuclease UvrABC nuclease subunit
COG0190	-0.242	H	510-methylene-tetrahydrofolate dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase
COG2770	-0.253	T	HAMP domain

COG2250	-0.257	S	HEPN domain
COG0583	-0.259	K	DNA-binding transcriptional regulator LysR family
COG0650	-0.269	C	Formate hydrogenlyase subunit 4
COG1256	-0.275	N	Flagellar hook-associated protein FlgK
COG0392	-0.280	S	Uncharacterized membrane protein YbhN UPF0104 family
COG1351	-0.282	F	Thymidylate synthase ThyX
COG0736	-0.299	I	Phosphopantetheinyl transferase (holo-ACP synthase)
COG0513	-0.303	L	Superfamily II DNA and RNA helicase
COG3639	-0.308	P	ABC-type phosphate/phosphonate transport system permease component
COG0765	-0.325	E	ABC-type amino acid transport system permease component
COG0248	-0.341	FTP	Exopolyphosphatase/pppGpp-phosphohydrolase
COG4907	-0.344	S	Uncharacterized membrane protein
COG1865	-0.349	H	Adenosylcobinamide amidohydrolase
COG0317	-0.353	TK	(p)ppGpp synthase/hydrolase HD superfamily
COG3620	-0.374	K	Predicted transcriptional regulator with C-terminal CBS domains
COG0514	-0.377	L	Superfamily II DNA helicase RecQ
COG1230	-0.379	P	Co/Zn/Cd efflux system component
COG1682	-0.403	GM	ABC-type polysaccharide/polyol phosphate export permease
COG1278	-0.424	K	Cold shock protein CspA family
COG1238	-0.431	S	Uncharacterized membrane protein YqaA SNARE-associated domain
COG1569	-0.440	R	Predicted nucleic acid-binding protein contains PIN domain
COG0823	-0.450	U	Periplasmic component of the Tol biopolymer transport system
COG1670	-0.458	JO	Protein N-acetyltransferase RimJ/RimL family
COG1232	-0.464	H	Protoporphyrinogen oxidase

COG0571	-0.464	K	dsRNA-specific ribonuclease
COG0341	-0.469	U	Preprotein translocase subunit SecF
COG1033	-0.495	R	Predicted exporter protein RND superfamily
COG2431	-0.495	S	Uncharacterized membrane protein YbjE DUF340 family
COG0466	-0.526	O	ATP-dependent Lon protease bacterial type
COG1126	-0.534	E	ABC-type polar amino acid transport system ATPase component
COG1866	-0.545	C	Phosphoenolpyruvate carboxykinase ATP-dependent
COG4870	-0.567	O	Cysteine protease C1A family
COG0342	-0.577	U	Preprotein translocase subunit SecD
COG0738	-0.590	G	Fucose permease
COG0022	-0.654	C	Pyruvate/2-oxoglutarate/acetoin dehydrogenase complex dehydrogenase (E1) component
COG1524	-0.689	R	Predicted pyrophosphatase or phosphodiesterase AlkP superfamily
COG0534	-0.697	V	Na+-driven multidrug efflux pump
COG1205	-0.703	L	ATP-dependent helicase YprA contains C-terminal metal-binding DUF1998 domain
COG1326	-0.750	R	Uncharacterized archaeal Zn-finger protein
COG1484	-0.784	L	DNA replication protein DnaC
COG1071	-0.841	C	TPP-dependent pyruvate or acetoin dehydrogenase subunit alpha
COG0579	-0.858	G	L-2-hydroxyglutarate oxidase LhgO
COG1505	-0.864	E	Prolyl oligopeptidase PreP S9A serine peptidase family
COG0582	-0.869	LX	Integrase
COG0232	-0.873	F	dGTP triphosphohydrolase
COG3391	-0.877	R	DNA-binding beta-propeller fold protein YncE
COG0703	-0.897	E	Shikimate kinase
COG4221	-0.898	C	NADP-dependent 3-hydroxy acid dehydrogenase YdfG
COG2317	-0.930	E	Zn-dependent carboxypeptidase M32 family

COG1624	-0.943	T	Diadenylate cyclase (c-di-AMP synthetase) DisA_N domain
COG1704	-0.944	S	Uncharacterized conserved protein
COG2026	-1.210	V	mRNA-degrading endonuclease RelE toxin component of the RelBE toxin-antitoxin system
COG1275	-1.341	V	Tellurite resistance protein TehA and related permeases

Prediction of psychrophiles.

COGs	**Coefs**	**Category**	**Function**
COG2376	1.628	G	Dihydroxyacetone kinase
COG2382	1.549	P	Enterochelin esterase or related enzyme
COG2930	1.321	I	Lipid-binding SYLF domain
COG3677	1.236	X	Transposase
COG2356	1.095	L	Endonuclease I
COG3620	0.976	K	Predicted transcriptional regulator with C-terminal CBS domains
COG0819	0.920	H	Thiaminase
COG2068	0.894	H	CTP:molybdopterin cytidylyltransferase MocA
COG2509	0.874	R	Uncharacterized FAD-dependent dehydrogenase
COG2335	0.764	R	Uncaracterized surface protein containing fasciclin (FAS1) repeats
COG4928	0.712	R	Predicted P-loop ATPase KAP-like
COG3530	0.690	S	Uncharacterized conserved protein DUF3820 family
COG3651	0.632	S	Uncharacterized conserved protein DUF2237 family
COG1398	0.627	I	Fatty-acid desaturase
COG2986	0.614	E	Histidine ammonia-lyase
COG3369	0.545	S	Uncharacterized protein contains Zn-finger domain of CDGSH type
COG1652	0.483	S	Nucleoid-associated protein YgaU contains BON and LysM domains

COG3296	0.465	S	Uncharacterized conserved protein Tic20 family
COG3201	0.451	H	Nicotinamide riboside transporter PnuC
COG2303	0.390	IR	Choline dehydrogenase or related flavoprotein
COG1246	0.377	E	N-acetylglutamate synthase or related acetyl-transferase GNAT family
COG1876	0.373	M	LD-carboxypeptidase LdcB LAS superfamily
COG1501	0.356	G	Alpha-glucosidase glycosyl hydrolase family GH31
COG1741	0.352	R	Redox-sensitive bicupin YhaK pirin superfamily
COG4067	0.342	S	Uncharacterized conserved protein
COG1397	0.338	O	ADP-ribosylglycohydrolase
COG1975	0.317	O	Xanthine and CO dehydrogenase maturation factor XdhC/CoxF family
COG0386	0.316	VI	Glutathione peroxidase house-cleaning role in reducing lipid peroxides
COG0429	0.314	R	Predicted hydrolase of the alpha/beta-hydrolase fold
COG2130	0.273	QR	NADPH-dependent curcumin reductase CurA
COG1479	0.265	S	Uncharacterized conserved protein contains ParB-like and HNH nuclease domains
COG3344	0.213	X	Retron-type reverse transcriptase
COG1526	0.211	C	Formate dehydrogenase assembly factor FdhD
COG0729	0.203	M	Outer membrane translocation and assembly module TamA
COG3653	0.193	Q	N-acyl-D-aspartate/D-glutamate deacylase
COG4977	0.156	K	Transcriptional regulator GlxA family contains an amidase domain and an AraC-type DNA-binding HTH domain
COG0400	0.116	R	Predicted esterase
COG3409	0.116	M	Peptidoglycan-binding (PGRP) domain of peptidoglycan hydrolases
COG3167	0.114	NW	Tfp pilus assembly protein PilO
COG1705	0.109	MN	Flagellum-specific peptidoglycan hydrolase FlgJ

COG1835	0.108	M	Peptidoglycan/LPS O-acetylase OafA/YrhL contains acyltransferase and SGNH-hydrolase domains
COG3287	0.099	S	Uncharacterized conserved protein contains FIST_N domain
COG2761	0.094	O	Predicted dithiol-disulfide isomerase DsbA family
COG2836	0.077	P	Sulfite exporter TauE/SafE
COG2360	0.069	O	Leu/Phe-tRNA-protein transferase
COG1022	0.063	I	Long-chain acyl-CoA synthetase (AMP-forming)
COG4148	0.045	P	ABC-type molybdate transport system ATPase component
COG4176	0.035	E	ABC-type proline/glycine betaine transport system permease component
COG3203	0.032	M	Outer membrane protein (porin)
COG2049	0.028	E	Allophanate hydrolase subunit 1
COG2326	0.012	C	Polyphosphate kinase 2 PPK2 family
COG2127	0.003	O	ATP-dependent Clp protease adapter protein ClpS
COG0226	-0.001	P	ABC-type phosphate transport system periplasmic component
COG0608	-0.004	L	Single-stranded DNA-specific exonuclease DHH superfamily may be involved in archaeal DNA replication intiation
COG0668	-0.010	M	Small-conductance mechanosensitive channel
COG0574	-0.010	G	Phosphoenolpyruvate synthase/pyruvate phosphate dikinase
COG0632	-0.013	L	Holliday junction resolvasome RuvABC DNA-binding subunit
COG0463	-0.015	M	Glycosyltransferase involved in cell wall bisynthesis
COG0395	-0.020	G	ABC-type glycerol-3-phosphate transport system permease component

COG1155	-0.020	C	Archaeal/vacuolar-type H+-ATPase catalytic subunit A/Vma1
COG0581	-0.021	P	ABC-type phosphate transport system permease component
COG1253	-0.021	R	Hemolysin or related protein contains CBS domains
COG0783	-0.027	PV	DNA-binding ferritin-like protein (oxidative damage protectant)
COG0553	-0.028	KL	Superfamily II DNA or RNA helicase SNF2 family
COG2002	-0.031	KV	Bifunctional DNA-binding transcriptional regulator of stationary/sporulation/toxin gene expression and antitoxin component of the YhaV-PrlF toxin-antitoxin module
COG3158	-0.036	P	K+ transporter
COG3189	-0.044	S	Uncharacterized conserved protein YeaO DUF488 family
COG1008	-0.078	C	NADH:ubiquinone oxidoreductase subunit 4 (chain M)
COG1108	-0.087	P	ABC-type Mn2+/Zn2+ transport system permease component
COG1708	-0.096	R	Predicted nucleotidyltransferase
COG1166	-0.097	E	Arginine decarboxylase (spermidine biosynthesis)
COG0500	-0.120	QR	SAM-dependent methyltransferase
COG1156	-0.129	C	Archaeal/vacuolar-type H+-ATPase subunit B/Vma2
COG1283	-0.131	P	Na+/phosphate symporter
COG1007	-0.135	C	NADH:ubiquinone oxidoreductase subunit 2 (chain N)
COG2192	-0.146	R	Predicted carbamoyl transferase NodU family
COG1164	-0.151	E	Oligoendopeptidase F
COG0717	-0.182	F	Deoxycytidine triphosphate deaminase
COG1078	-0.192	R	HD superfamily phosphohydrolase

COG3211	-0.201	R	Secreted phosphatase PhoX family
COG3190	-0.204	N	Flagellar biogenesis protein FliO
COG1310	-0.208	O	Proteasome lid subunit RPN8/RPN11 contains Jab1/MPN domain metalloenzyme (JAMM) motif
COG1354	-0.211	L	Chromatin segregation and condensation protein Rec8/ScpA/Scc1 kleisin family
COG1005	-0.213	C	NADH:ubiquinone oxidoreductase subunit 1 (chain H)
COG0616	-0.218	O	Periplasmic serine protease ClpP class
COG0636	-0.221	C	FoF1-type ATP synthase membrane subunit c/Archaeal/vacuolar-type H+-ATPase subunit K
COG2268	-0.225	S	Uncharacterized membrane protein YqiK contains Band7/PHB/SPFH domain
COG1180	-0.235	O	Pyruvate-formate lyase-activating enzyme
COG0334	-0.241	E	Glutamate dehydrogenase/leucine dehydrogenase
COG1269	-0.253	C	Archaeal/vacuolar-type H+-ATPase subunit I/STV1
COG0396	-0.253	O	Fe-S cluster assembly ATPase SufC
COG0719	-0.269	O	Fe-S cluster assembly scaffold protein SufB
COG0483	-0.274	G	Archaeal fructose-16-bisphosphatase or related enzyme of inositol monophosphatase family
COG0649	-0.283	C	NADH:ubiquinone oxidoreductase 49 kD subunit (chain D)
COG1264	-0.288	G	Phosphotransferase system IIB components
COG2190	-0.300	G	Phosphotransferase system IIA component
COG0598	-0.343	P	Mg2+ and Co2+ transporter CorA
COG0163	-0.348	H	3-polyprenyl-4-hydroxybenzoate decarboxylase
COG1121	-0.362	P	ABC-type Mn2+/Zn2+ transport system ATPase component
COG2317	-0.389	E	Zn-dependent carboxypeptidase M32 family
COG0736	-0.405	I	Phosphopantetheinyl transferase (holo-ACP synthase)

COG0586	-0.407	S	Uncharacterized membrane protein DedA SNARE-associated domain
COG1828	-0.446	F	Phosphoribosylformylglycinamidine (FGAM) synthase PurS component
COG0575	-0.450	I	CDP-diglyceride synthetase
COG1143	-0.454	C	Formate hydrogenlyase subunit 6/NADH:ubiquinone oxidoreductase 23 kD subunit (chain I)
COG0302	-0.489	H	GTP cyclohydrolase I
COG0852	-0.504	C	NADH:ubiquinone oxidoreductase 27 kD subunit (chain C)
COG2013	-0.507	S	Uncharacterized conserved protein AIM24 family
COG0221	-0.512	CP	Inorganic pyrophosphatase
COG1387	-0.517	ER	Histidinol phosphatase or related hydrolase of the PHP family
COG1871	-0.536	NT	Chemotaxis receptor (MCP) glutamine deamidase CheD
COG0827	-0.554	L	Adenine-specific DNA methylase
COG0561	-0.567	HR	Hydroxymethylpyrimidine pyrophosphatase and other HAD family phosphatases
COG2132	-0.576	DPM	Multicopper oxidase with three cupredoxin domains (includes cell division protein FtsP and spore coat protein CotA)
COG0543	-0.591	HC	NAD(P)H-flavin reductase
COG1713	-0.694	R	HD superfamily phosphohydrolase YqeK (fused to NMNAT in mycoplasms)
COG0610	-0.720	V	Type I site-specific restriction-modification system R (restriction) subunit and related helicases ...
COG0043	-0.745	H	3-polyprenyl-4-hydroxybenzoate decarboxylase
COG4166	-0.861	E	ABC-type oligopeptide transport system periplasmic component
COG1254	-0.903	C	Acylphosphatase
COG0551	-0.968	L	ssDNA-binding Zn-finger and Zn-ribbon domains of topoisomerase 1

COG0392	-0.989	S	Uncharacterized membrane protein YbhN UPF0104 family
COG2039	-1.187	O	Pyrrolidone-carboxylate peptidase (N-terminal pyroglutamyl peptidase)
COG0648	-1.275	L	Endonuclease IV
COG0588	-1.315	G	Phosphoglycerate mutase (BPG-dependent)
COG1324	-1.360	P	Uncharacterized protein involved in tolerance to divalent cations