Prediction of Optimal Growth Temperature using only Genome Derived Features

David B. Sauer[1]* and Da-Neng Wang[1]*

[1]Department of Cell Biology, and The Helen L. and Martin S. Kimmel Center for Biology and Medicine, Skirball Institute of Biomolecular Medicine, New York University School of Medicine, New York, New York, United States of America

*Corresponding authors

E-mail: david.sauer@med.nyu.edu (D.B.S), da-neng.wang@med.nyu.edu (DN.W.)

Abstract

Optimal growth temperature is a fundamental characteristic of all living organisms. Knowledge of this temperature is central to the study the organism, the thermal stability and temperature dependent activity of its genes, and the bioprospecting of its genome for thermally adapted proteins. While high throughput sequencing methods have dramatically increased the availability of genomic information, the growth temperatures of the source organisms are often unknown. This limits the study and technological application of these species and their genomes. Here, we present a novel method for the prediction of growth temperatures of prokaryotes using only genomic sequences. By applying the reverse ecology principle that an organism's genome includes identifiable adaptations to its native environment, we can predict a species' optimal growth temperature with an accuracy of 4.69 °C root-mean-square error and a correlation coefficient of 0.908. The accuracy can be further improved for specific taxonomic clades or by excluding psychrophiles. This method provides a valuable tool for the rapid calculation of organism growth temperature when only the genome sequence is known.

Author Summary

The optimal growth temperature is a fundamental characteristic of all living organisms. It is the temperature at which the organism grows at the greatest rate, and is a consequence of adaptations of that organism to its native environment. These adaptations are contained within the genome of the organism, and therefore species from varying environments have distinct genomic characteristics. Here we use those genomic characteristics to predict a species' optimal growth temperature. This provides a novel tool for describing a key parameter of the species' native environment when it is otherwise unknown. This is particularly valuable as the rate of genome sequencing has increased, while the determination of growth temperature remains laborious.

Introduction

Growth conditions of an organism are essential to its characterization. However, these values may be unknown in organisms which are difficult to culture, “unculturable”, or otherwise poorly characterized. Reverse ecology posits that the evolutionary effects of an organism’s native environment is reflected by adaptations in its genome [1]. Therefore, an organism’s native environment can be identified by comparing its genome to the genomes of other organisms from a range of environments. Notably, this is done without experimental manipulation or interrogation of the organism beyond genome sequencing. Such reverse ecology strategies have been successful in studying adaptation to soil conditions [2], salinity [3], and temperature [4].

Of these environmental pressures, temperature, being a description of the internal energy of the environment, is a particularly strong driving force for adaptation. Prokaryotes are often viable over a range of temperatures, which varies by species. For a particular organism, increasing temperature beyond it’s growth range, corresponding to increased internal energy, can lead to loss of protein and nucleic acid structure. Conversely, a sub-optimal temperature leads to reduced enzyme kinetics and stiffening lipid membranes. Each of these biological consequences may be deleterious to un-adapted organisms. Therefore, it is perhaps not surprising that an organism’s optimal growth temperature (OGT) correlates to quantifiable properties (features) in the organism’s nucleotide and protein sequences. Features correlated with OGT can be identified in the genomic [5], tRNA [6,7], rRNA [6–8], open reading frame [9,10], and in the proteomic sequences [10–13]. Correlations between OGT and tRNA G+C content [6,7] or the charged versus polar amino acid ratios [14] are well known.

Clearly, OGT is a necessary parameter for analyzing physiological processes of an organism or activities of its genes and proteins. [15,16]. However, the experimental determination of OGT is laborious [17,18], and sometimes unattainable [19]. Also, recorded OGT or environmental temperature may be inconsistently measured, particularly in genetic samples not obtained from pure culture [20]. Further, for metagenomic samples the conditions during collection may significantly differ from the originating species' growth environment. This can be due to the organism or its genetic material being found distant from its originating environment [21], or the collected genomic material may be from organisms which are inviable [22]. Even in pure culture in the laboratory, the experimental growth conditions can vary greatly [23] and may not be at the source organisms' OGT [24].

While many previous studies have aimed to identify genes and proteins [25], mutations [16], and mechanisms [15] that drive thermal adaptation, there is also great value in using these adaptive differences to provide data of an organism's native environment when it may not be otherwise known or well-described. A number of parameters have been identified which correlate with OGT [14]. However, those correlations are often weak and therefore of limited predictive value alone. Here, we aim to predict a prokaryotic species' OGT only from its genomic sequence. We set out to develop a novel tool for the ecological characterization of a species based solely on its genome, the study of thermoadaptation, and bioprospecting for thermoadapted genes.

Results

Prokaryote genome redundancy is highly skewed

Of the initial 8270 prokaryotic species with a reported OGT, genome sequences were available for 2708 species. These sequenced species were composed of 2538 Bacteria and 170 Archaea, with OGTs ranging from 4 to 103 °C. A total of 36,721 sequenced genomes for these species were downloaded, indicating multiple genomes for each species on average. However, the number of genomes per species was highly skewed, with great redundancy for model organisms and pathogens (Fig S1C). To avoid having these relatively few species dominate the analysis, features were averaged by species and all regressions were done by species rather than by genome.

Individual genome derived features correlate with OGT

Based on the reverse genomics principle that an organism's adaptation to its environment is reflected within its genome, we hypothesize that a species' OGT could be predicted based on characteristics of its genome and genome derived sequences. This hypothesis was supported by previous noted correlations between OGT and individual features of the genomic [5,6,26], tRNA [6,7], rRNA [6–8], open reading frame [7,9,10,27,28], and proteomic or protein sequences [10–14,29–35]. These features are quantifiable properties of the sequence, such as G+C content, length, and nucleotide or amino acid fraction. Of the features calculated, 42 were found in this work to be correlated with OGT in the present dataset by the Pearson correlation coefficient with |r| > 0.3 (Fig 1, Table S1). However, these correlations to OGT were often weak and therefore insufficient for the calculation of a species' growth temperature. Furthermore, there was a strong association among many features (Fig S2). We therefore decided to consider them simultaneously, using multiple linear regression, with features added individually to minimize multicollinearity. We started by classifying features based on the source sequences (genomic, tRNA, rRNA, open reading frames, and proteome). Multiple linear regressions were calculated, progressively increasing the number of feature classes used in the regression.

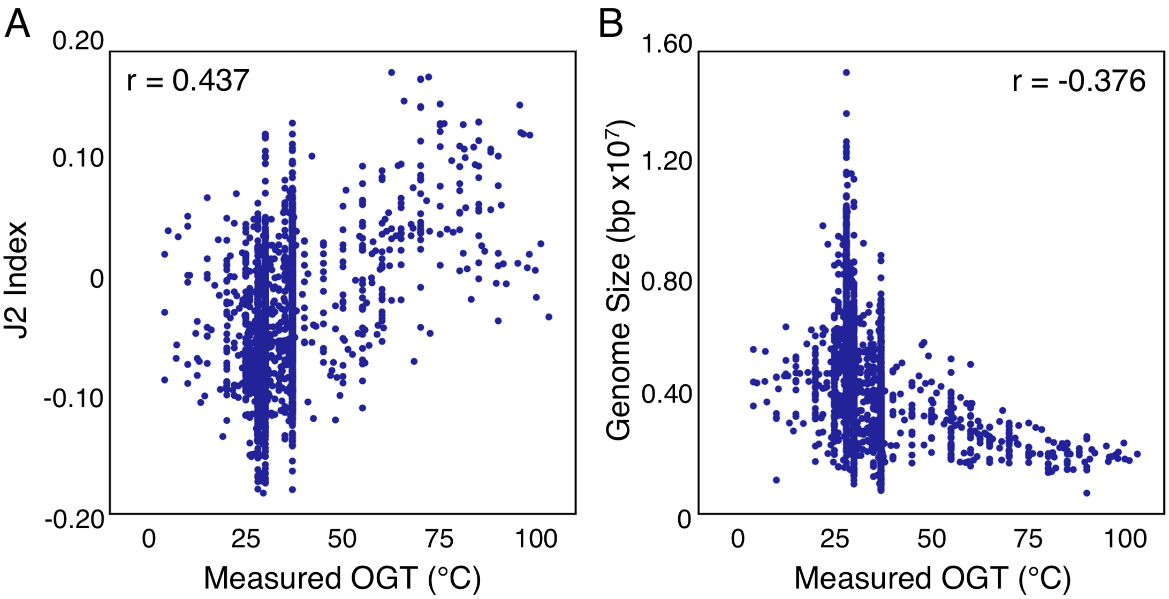

Figure 1. Individual genome derived features correlate weakly with the originating species' OGT. Measure optimal growth temperature for each species versus J2 index of genomic dinucleotide fractions (A) and total genome size (B).

A regression using only genomic sequence based features is weakly predictive of OGT

The genomic sequence provides information about the nucleotide content, nucleotide order, and chromosomal structure of an organism's hereditable genetic material. In the absence of any other knowledge, this sequence still reflects adaptations to the particular thermal environment of the organism. For example, total genome size has been shown to be negatively correlated with a species' OGT [26]. Accordingly, it has been proposed that the reduced time and energy of genomic replication offers selective advantages at higher temperatures. Additionally, the necessity of maintaining genomic structure with increased temperature is thought to be reflected in a species' genomic dinucleotide fractions [36], which is quantified in the J2 index [5].

In the present dataset, individual nucleotide and dinucleotide fractions of the genome, the J2 index, the G+C content, and total size were calculated for each genome. Of these features, the J2 index, genome size, and the CT and AG dinucleotide fractions correlated with OGT, but only weakly. Using these poorly correlated and collinear input features for regression, the resulting multiple linear regression is poor at predicting OGT with a root mean squared error (RMSE) of 9.86 °C (r = 0.469) (Fig S3).

tRNA and rRNA sequences improve OGT prediction

tRNA and rRNA are nucleic acids whose structure, and enzymatic activity in the case of rRNA, are essential to cell viability. Therefore, the direct correlation of OGT to G+C content of tRNAs [6,7] and rRNAs [8,37] is thought to reflect the necessary increase in base pair hydrogen bonding needed to maintain the structure of these nucleic acids at elevated temperatures. While a subset of the previously analyzed genomic sequence, we hypothesized that features derived from these tRNA and rRNA sequences might be more strongly correlated with OGT. To this end, we identified their sequences bioinformatically. tRNA and 16S rRNA sequences were identified in 100% and 98% of the species respectively, reflecting the highly conserved nature of these genes.

Using these identified tRNA and rRNA sequences, nucleotide fractions and G+C content were calculated for each. All calculated features for tRNA and rRNA sequences were correlated with OGT. Calculating a new linear regression with the OGT using tRNA features, in addition to genomic features, improved accuracy (RMSE = 7.30 °C, r = 0.757) (Fig 2A). Similarly, a regression calculated with rRNA and genomic features also improved accuracy (RMSE = 6.99 °C, r = 0.784) (Fig 2B). By using all available tRNA, rRNA, and genomic features, a still more accurate linear regression was calculated (RMSE = 6.71 °C, r = 0.802) (Fig 2C).

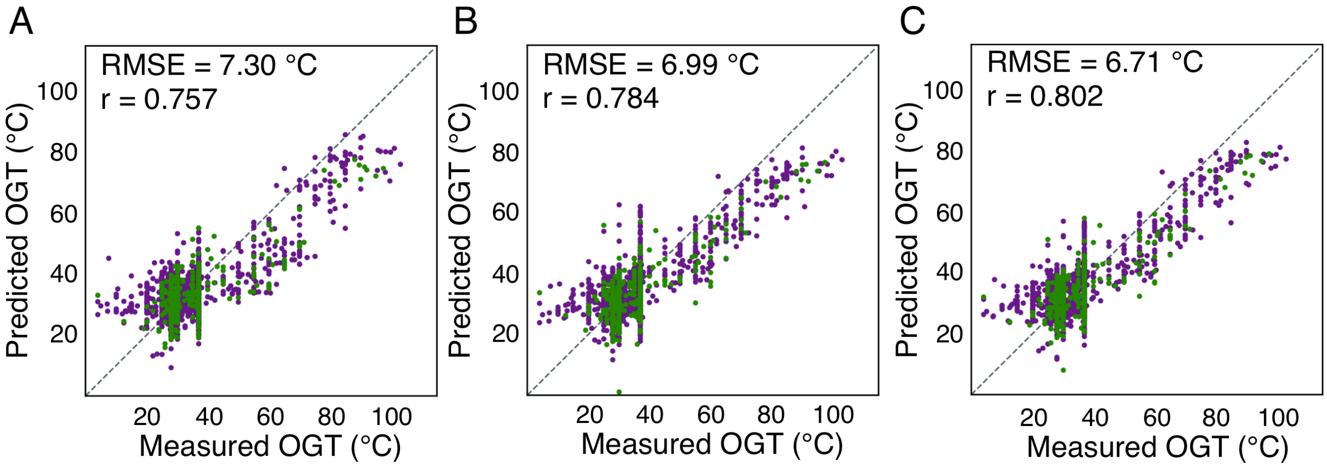

Figure 2. Using genomic and genic sequences improve OGT prediction accuracy. Predicted versus measured OGT for each species, using linear regressions with features derived from genomic and (A) tRNA, (B) rRNA, or (C) tRNA and rRNA sequences. Species used for regression and evaluation are shown in purple and green, respectively. The dotted line indicates a perfect prediction.

ORF sequences improve OGT prediction

As tRNA and rRNA features clearly improve the ability to predict a species' OGT, we examined if other gene sequences might also improve the regression. In particular open reading frames, which code for proteins but exclude the non-coding regions of the genome, were considered. We hypothesized that using coding regions alone would increase sensitivity to changes in OGT. Additionally, codon biases have previously been reported to correlate with OGT [13], likely reflecting both amino acid differences and the necessity of maintaining proper codon-anticodon pairing in differing thermal environments. Furthermore, the greater number of ORFs in a genome, relative to tRNAs and rRNAs, make the features of ORFs less sensitive to single gene aberrations or mispredictions. Therefore, ORF derived features were hypothesized to more sensitively and accurately report on the thermal environment than tRNA or rRNA sequences.

We identified ORFs within the genomic sequences bioinformatically. From these ORFs, a number of derived features were calculated including nucleotide and dinucleotide fractions, codon fractions, start and stop codon fractions, the coding ratio and fraction of the genome, the ORF density of the genome, G+C and A+G content, and average length. Of these, nine were found to be correlated with OGT. These include the A+G content, codon and dinucleotide fractions, and the fraction of the alternative start codon TTG. These ORF derived features, in addition to the genomic, tRNA, and rRNA features, were used to calculate a new multiple linear regression with significantly improved accuracy (RMSE = 5.77 °C, r= 0.857) (Fig 3).

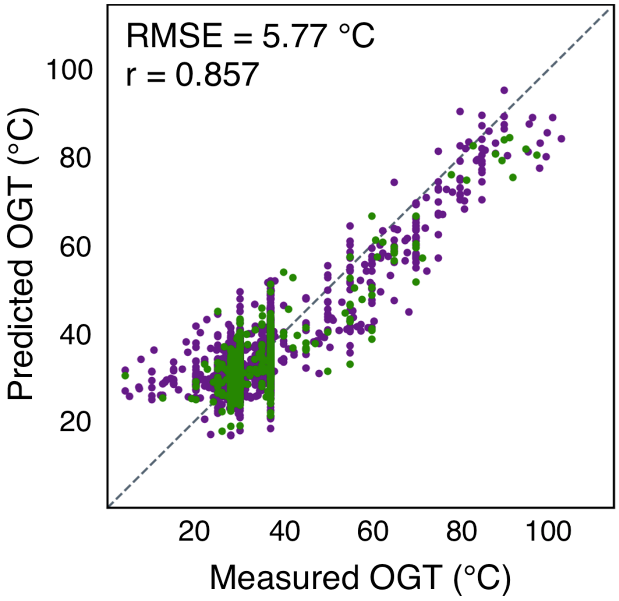

Figure 3. Open reading frame sequences further improve OGT prediction accuracy. Predicted versus measured OGT for each species, using a linear regression with features derived from

genomic, tRNA, rRNA, and ORF sequences. Species used for regression and evaluation are shown in purple and green, respectively. The dotted line indicates a perfect prediction.

Including proteome features significantly improves OGT prediction

While ORF feature correlation to OGT partially reflects the adaptation of the coding regions and mRNAs to the thermodynamic environment, it has been suspected that this correlation also reflected adaptations in each species' proteome to OGT. Temperature is known to correlate with protein folding, biochemistry, and enzyme kinetics, all of which are essential to organismal viability [10,14,32]. Based on these biological consequences, proteome derived features were hypothesized to be especially sensitive to thermal environment. Therefore, the proteome was translated from each species' ORFs, and features calculated from the proteome's primary sequence. These features included amino acid fractions, the fraction of the proteome to be charged or thermolablile, and the EK/QH, LK/Q, Polar/Charged, and Polar/Hydrophobic amino acid ratios.

Supporting this hypothesis, proteome derived features were found to have the strongest correlation to OGT (Table S1), with the greatest correlation being the fraction of the proteome composed of the amino acids ILVWYGERKP [13]. The linear regression of OGT using proteome features, in addition to previously described features, significantly improved accuracy (RMSE = 4.69 °C, r = 0.908). (Fig 4, Eq S1).

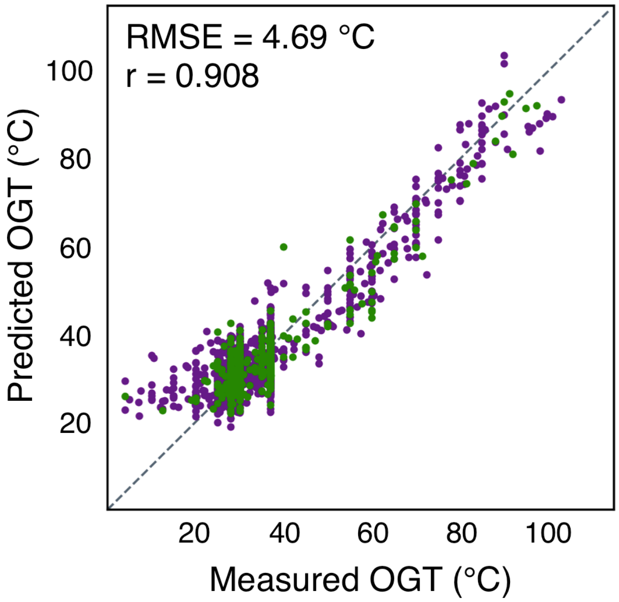

Figure 4. Proteome derived features significantly improve OGT prediction accuracy. Predicted versus measured OGTs for each species, using a linear regression with features derived from genomic, tRNA, rRNA, ORF, and proteome sequences. Species used for regression and evaluation are shown in purple and green, respectively. The dotted line indicates a perfect prediction.

Taxonomic clade specific regressions are the most accurate

The regressions described up to this point were all made using all prokaryotic species. However, we had noted that the number of individual features correlated with OGT was much higher in Archaea than Bacteria (Table S1). In addition, we hypothesized that the magnitude of the response of each feature to OGT may be distinct in each superkingdom.

Based on these distinctions, we tested whether superkingdom specific regressions would be more accurate than the regression of all prokaryotes (Fig. 5). Using the NCBI taxonomic assignment for each species, an Archaea-only regression dramatically improved accuracy for these species (RMSE = 3.21 °C, r = 0.995) (Eq S2). However, the Bacteria-only regression only showed only a slight improvement (RMSE = 4.61 °C, r = 0.816) (Eq S3). This likely reflects bias of the general prokaryotic regression, due to the numerical majority of bacterial species and the greater diversity of bacterial species.

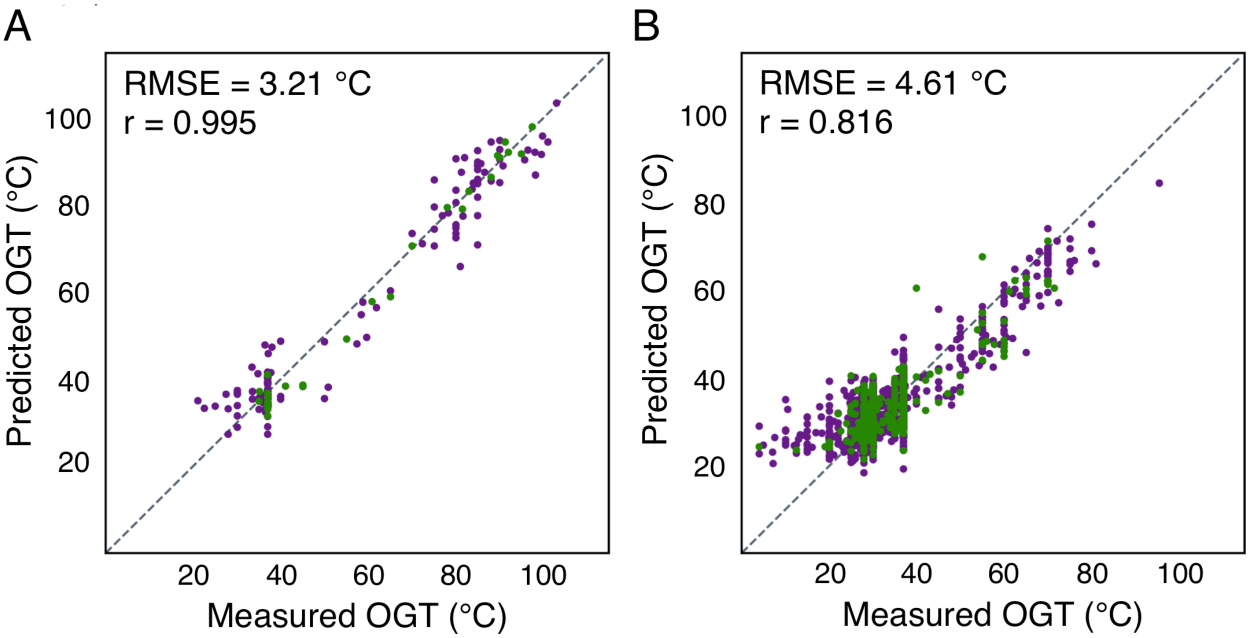

Figure 5. Taxon specific linear regressions are most accurate. Predicted versus measured OGT for each species using superkingdom specific linear regressions for Archaea (A) and Bacteria (B). Species used for regression and evaluation are shown in purple and green, respectively. The dotted line indicates a perfect prediction.

Addressing this diversity in bacteria, the taxonomic specific regression can be further improved when the data is separated by phylum or class. OGT regression was limited to clades where the number of species (N) was greater than 50 to ensure the significance of the regression. Of the individual phyla, the most accurate regressions are found in the *Firmicutes* (RMSE = 4.88 °C, r = 0.831), *Actinobacteria* (RMSE = 2.90 °C, r = 0.818), *Bacteroidetes* (RMSE = 1.58 °C, r =

0.964), and *Euryarchaeota* (RMSE = 4.00 °C, r = 0.985) (Fig S4). In contrast, the *Proteobacteria* regression had much more weakly correlated predicted and reported OGTs (RMSE = 4.10 °C, r = 0.569), though the small RMSE likely reflects the narrow OGT range of this phylum. Further subdivision of the *Proteobacteria* into classes (Fig S5) resulted in significant correlation of the *Betaproteobacteria* (RMSE = 2.94 °C, r = 0.789), and *Deltaproteobacteria* regressions (RMSE = 2.04 °C, r = 0.761). However, no correlation was found in regressions for the *Proteobacteria* classes of *Alphaproteobacteria* or *Gammaproteobacteria*.

Discussion

Knowing an organism's optimal growth characteristics is central to addressing basic biological questions about how organisms adapt to a particular environmental niche. Further, the systematic study of adaption often requires the optimal growth conditions of the species of origin for each species and gene or protein examined. Additionally, proteins from organisms adapted to particular environmental niches are often particularly suited for structural biology [38–40] and industrial applications [41,42].

However, if the growth characteristics of already sequenced organisms are uncharacterized, the physiochemical properties of these genes that otherwise might be inferred are lost [20]. Consequently, this limits the use of these genomes in academic study and mining for biotechnology applications. Exacerbating this issue, high throughput sequencing has enabled rapid growth in the number of available genomic, metagenomic, and derived proteomic sequences. This growth in genetic information is likely to outpace the laborious experimental task of characterizing the growth conditions of each species, leading to an increasing number of genomic sequences with unknown growth characteristics. This is already apparent by those organisms which have been ‘unculturable’ to date, but which have been sequenced by metagenomics.

To satisfy the need for growth condition data when only genomic sequences are available, here we demonstrate a novel reverse ecology tool to accurately predict the OGT using solely the genomic sequence as input. Our method can predict the OGT for sequenced Archaea and bacteria with an accuracy of 3.21 °C and 4.61 °C, respectively.

OGT can be accurately predicted using only genome derived parameters

Genome classification is clearly essential to the most accurate prediction of OGT. The programs used for tRNA, rRNA, and ORF identification all require some level of taxonomic classification. When applying the general prokaryotic regression, this is only requires the relatively simple exclusion of eukaryotic samples prior to sequencing [43]. However, the most accurate OGT regressions are taxon specific, and therefore genomic samples require further classification. This assignment is routinely addressed *in silico*, using specialized bioinformatic tools which can easily assign taxonomic clade to genomic material [44,45].

As a simple proof-of-concept, the prokaryotic genomes were also classified by superkingdom using the best scoring 16S rRNA hidden Markov model in Barrnap (Fig S6). These regressions were of similar accuracy to those using NCBI superkingdom assignments.

Excluding genome size does not alter the regression accuracy

While prokaryote genome size is strongly correlated with OGT, it is unique among all features used here in requiring a complete genome for calculation. Therefore, this feature might not be available in metagenomic samples, or otherwise incompletely assembled genomes. Excluding this feature has only a minor impact on the regression for all prokaryotes (RMSE = 5.07 °C, r = 0.891), or the separate regressions for Bacteria (RMSE = 4.97 °C, r = 0.783), or Archaea (RMSE = 3.21 °C, r = 0.995) (Fig S7).

Psychrophiles are poorly fit

While the final regressions of prokaryotes and Bacteria were generally accurate, species with optimal growth temperatures less than approximately 25 °C are clearly poorly fit. This outcome is unsurprising, as few psychrophilic sequences are present in the dataset (Fig. S1), and the mechanisms of thermoadaptation to higher and lower temperatures are not equivalent [46]. Excluding those species with an OGT of less than 25 °C yields a slightly better general prokaryotic regression (RMSE = 4.42 °C, r = 0.916) (Fig. S6). The archaeal regression was slightly worse (RMSE = 3.12 °C, r = 0.993), while the bacterial regression improved (RMSE = 4.26 °C, r = 0.832), reflecting the known OGT ranges of each superkingdom.

Improvements over comparable methods

Our method significantly expands and improves upon the individual features previously described to correlate with OGT. By studying a much larger set of genomes, a more precise correlation between each feature and OGT can be calculated. Further, by using multiple features, more accurate and predictive regression models have been calculated. Notably, our method improves on previously reported analyses requiring particular genes being present in the genome, thereby making the method more general in application [47]. Also, this method quantitatively predicts an OGT rather than using classification (psychrophile, mesophile, thermophile, or hyperthermophile). This improves on methods which predict OGT ranges [47–50], where classification necessarily limited accuracy.

The most comparable method is reported by Zeldovich *et al.* calculating OGT from the proteome as OGT = 937F – 335, where F is the sum of the proteome fraction for the amino acids IVYWREL [13]. Using the current larger dataset, we calculate a lower correlation (r = 0.726) and accuracy (RMSE = 10.5 °C) than previously reported. This is likely a consequence of more genomic sequences being available, and our keeping of individual species separate rather than averaging those with the same OGT. By considering more features derived from the source organism's genome, the prokaryotic regression presented here clearly advances upon this previous method improving in both correlation and accuracy. While we focus on growth temperature, the same principle could be readily applied to other quantifiable characteristics of an organism's optimal growth environment, such as pH, salinity, osmolarity, or oxygen concentration.

Application and validation

Applying these regressions, we predicted OGTs for those species with a genomic sequence available, but without a reported OGT in Sauer *et al.* (2015), using the most taxon specific linear regression available. Only the *Betaproteobacteria* and *Deltaproteobacteria* classes of *Proteobacteria* were predicted, excluding the *Alphaproteobacteria*, *Gammaproteobacteria*, and other *Proteobacteria* due to the poor predictive values of those taxon specific regressions. In total, 482 species' OGTs were predicted (Table S2). Of the species with newly predicted OGTs, a more recent literature search revealed reported OGTs for 36 species [51–87]. The predicted and measured OGTs were strongly correlated (RMSE = 6.94 °C, r = 0.857), validating the predictive value of this method (Fig S9).

Materials and Methods

Source data and sequence extraction

Experimentally measured OGTs of various prokaryotic species were used as previously published without modification [88]. Taxonomic assignments for each species were collected from NCBI [89]. All available top level genome sequences for each species were downloaded from Ensembl [90]. tRNA sequences were identified with tRNAScan-SE 1.3.1 [91] with general settings. Ribosomal RNA genes were identified with Barrnap 0.8 [92] using superkingdom specific hidden Markov models, and rRNA sequences extracted from the genome using BEDtools 2.26.0 [93]. Open reading frame sequences were identified with GenemarkS 4.32 [94] using the default settings. ORFs were also translated into protein sequences using the standard genetic code. Features were calculated for each genome and derived proteome, ignoring ambiguous nucleotides and amino acids. All calculated features were averaged by species. Twenty percent of the species with available genomes were set aside as a test set and never used for regression, only evaluation.

Multiple linear regression

Only individual features linearly correlated with OGT ($|r| > 0.3$) were used for multiple linear regression. To minimize multicollinearity, the initial regression input feature set consisted of only the feature most correlated with OGT. To this set all other correlated features were added individually, and multiple linear regressions were calculated. If the correlation between measured and predicted OGTs increased for any regression, the input feature which most increased the correlation was added to the input set. This was repeated until the correlation did not increase.

Regression evaluation and prediction

The test set was only used for evaluation of the multiple linear regressions, comparing the calculated and measured OGTs. Regressions were evaluated by comparing the predicted and reported OGT using the Pearson correlation coefficient and root mean square error.

De novo OGT prediction and validation

All top level genomes in Ensembl Bacteria were downloaded for each species where there was not a reported OGT in the Sauer *et al.* (2015) dataset. Taxonomic assignment and feature calculation were preformed as described above. The most taxonomic specific regression available, using genomic, tRNA, ORF, and proteome features, was used to predict the OGT for each species. For these newly predicted species, Pubmed was searched using the binomial name and “optimal growth” as keywords. From the returned publications, OGTs were manually collected where available.

Analyses were carried out using custom Python scripts using Biopython 2.7.12 [95], NumPy 1.13.3 [96], SciPy 1.0.0, Scikit-learn 0.19.1 [97], and MatPlotLib 2.1.0 [98].

Acknowledgements

The authors thank Jennifer Marden for discussion and critical review of this manuscript.

This work was financially supported in part by an American Cancer Society Postdoctoral Fellowship (16-A1-00-005739 to D.B.S), the Department of Defense (W81XWH-16-1-0153 to D.B.S.), and NIH (R01-GM121994, R01-DK099023, and R01-GM093825 to D-N.W). This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs, through the Peer Reviewed Cancer Research Program under Award No. W81XH-16-1-0153. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

References

1. Li YF, Costello JC, Holloway AK, Hahn MW. "Reverse ecology" and the power of population genomics. Evol Int J Org Evol. 2008;62: 2984–2994. doi:10.1111/j.1558-5646.2008.00486.x

2. Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet. 2010;42: 260–263. doi:10.1038/ng.515

3. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 2010;6: e1000862. doi:10.1371/journal.pgen.1000862

4. Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, et al. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Natl Acad Sci U S A. 2011;108: 2831–2836. doi:10.1073/pnas.1014971108

5. Kawashima T, Amano N, Koike H, Makino S, Higuchi S, Kawashima-Ohya Y, et al. Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. Proc Natl Acad Sci U S A. 2000;97: 14257–14262. doi:10.1073/pnas.97.26.14257

6. Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol. 1997;44: 632–636.

7. Hurst LD, Merchant AR. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci. 2001;268: 493–497. doi:10.1098/rspb.2000.1397

8. Khachane AN, Timmis KN, dos Santos VAPM. Uracil content of 16S rRNA of thermophilic and psychrophilic prokaryotes correlates inversely with their optimal growth temperatures. Nucleic Acids Res. 2005;33: 4016–4022. doi:10.1093/nar/gki714

9. Lynn DJ, Singer GAC, Hickey DA. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res. 2002;30: 4272–4277.

10. Singer GAC, Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene. 2003;317: 39–47.

11. Lobry JR, Chessel D. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J Appl Genet. 2003;44: 235–261.

12. Tekaia F, Yeramian E, Dujon B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene. 2002;297: 51–60.

13. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 2007;3: e5. doi:10.1371/journal.pcbi.0030005

14. Suhre K, Claverie J-M. Genomic correlates of hyperthermostability, an update. J Biol Chem. 2003;278: 17198–17202. doi:10.1074/jbc.M301327200

15. Nguyen V, Wilson C, Hoemberger M, Stiller JB, Agafonov RV, Kutter S, et al. Evolutionary drivers of thermoadaptation in enzyme catalysis. Science. 2017;355: 289–294. doi:10.1126/science.aah3717

16. Perl D, Mueller U, Heinemann U, Schmid FX. Two exposed amino acid residues confer thermostability on a cold shock protein. Nat Struct Biol. 2000;7: 380–383. doi:10.1038/75151

17. Elliott RP. Temperature-Gradient Incubator for Determining the Temperature Range of Growth of Microorganisms. J Bacteriol. 1963;85: 889–894.

18. Honglin Z, Yongjun L, Haitao S. Determination of thermograms of bacterial growth and study of optimum growth temperature. Thermochim Acta. 1993;216: 19–23. doi:10.1016/0040-6031(93)80377-M

19. Stewart EJ. Growing unculturable bacteria. J Bacteriol. 2012;194: 4151–4160. doi:10.1128/JB.00345-12

20. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev MMBR. 2008;72: 557–578, Table of Contents. doi:10.1128/MMBR.00009-08

21. Rose M, Landman D, Quale J. Are community environmental surfaces near hospitals reservoirs for gram-negative nosocomial pathogens? Am J Infect Control. 2014;42: 346–348. doi:10.1016/j.ajic.2013.12.025

22. Cangelosi GA, Meschke JS. Dead or alive: molecular assessment of microbial viability. Appl Environ Microbiol. 2014;80: 5884–5891. doi:10.1128/AEM.01763-14

23. Hearing J, Hunter E, Rodgers L, Gething MJ, Sambrook J. Isolation of Chinese hamster ovary cell lines temperature conditional for the cell-surface expression of integral membrane glycoproteins. J Cell Biol. 1989;108: 339–353.

24. Hashimoto H, Moritani N, Saito TR. Comparative study on circadian rhythms of body temperature, heart rate, and locomotor activity in three species hamsters. Exp Anim. 2004;53: 43–46.

25. Wang Q, Cen Z, Zhao J. The survival mechanisms of thermophiles at high temperatures: an angle of omics. Physiol Bethesda Md. 2015;30: 97–106. doi:10.1152/physiol.00066.2013

26. Sabath N, Ferrada E, Barve A, Wagner A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome Biol Evol. 2013;5: 966–977. doi:10.1093/gbe/evt050

27. Li W, Zou H, Tao M. Sequences downstream of the start codon and their relations to G + C content and optimal growth temperature in prokaryotic genomes. Antonie Van Leeuwenhoek. 2007;92: 417–427. doi:10.1007/s10482-007-9170-6

28. Zheng H, Wu H. Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species. BMC Bioinformatics. 2010;11 Suppl 11: S7. doi:10.1186/1471-2105-11-S11-S7

29. Burra PV, Kalmar L, Tompa P. Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. PloS One. 2010;5: e12069. doi:10.1371/journal.pone.0012069

30. Robinson-Rechavi M, Alibés A, Godzik A. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of Thermotoga maritima. J Mol Biol. 2006;356: 547–557. doi:10.1016/j.jmb.2005.11.065

31. Puigbò P, Pasamontes A, Garcia-Vallve S. Gaining and losing the thermophilic adaptation in prokaryotes. Trends Genet TIG. 2008;24: 10–14. doi:10.1016/j.tig.2007.10.005

32. Cambillau C, Claverie JM. Structural and genomic correlates of hyperthermostability. J Biol Chem. 2000;275: 32383–32386. doi:10.1074/jbc.C000497200

33. Saelensminde G, Halskau Ø, Helland R, Willassen N-P, Jonassen I. Structure-dependent relationships between growth temperature of prokaryotes and the amino acid frequency in their proteins. Extrem Life Extreme Cond. 2007;11: 585–596. doi:10.1007/s00792-007-0072-3

34. Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. 2001;29: 1608–1615.

35. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. Proc Natl Acad Sci U S A. 1999;96: 3578–3583.

36. Amano N, Ohfuku Y, Suzuki M. Genomes and DNA conformation. Biol Chem. 1997;378: 1397–1404.

37. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. Science. 1999;283: 220–221.

38. Yernool D, Boudker O, Jin Y, Gouaux E. Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. Nature. 2004;431: 811–818. doi:10.1038/nature03018

39. Jiang Y, Lee A, Chen J, Cadene M, Chait BT, MacKinnon R. Crystal structure and mechanism of a calcium-gated potassium channel. Nature. 2002;417: 515–522. doi:10.1038/417515a

40. Karpowich NK, Wang D-N. Assembly and mechanism of a group II ECF transporter. Proc Natl Acad Sci U S A. 2013;110: 2534–2539. doi:10.1073/pnas.1217361110

41. Acharya S, Chaudhary A. Bioprospecting thermophiles for cellulase production: a review. Braz J Microbiol Publ Braz Soc Microbiol. 2012;43: 844–856. doi:10.1590/S1517-83822012000300001

42. Koskinen PEP, Lay C-H, Beck SR, Tolvanen KES, Kaksonen AH, Örlygsson J, et al. Bioprospecting Thermophilic Microorganisms from Icelandic Hot Springs for Hydrogen and Ethanol Production. Energy Fuels. 2008;22: 134–140. doi:10.1021/ef700275w

43. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304: 66–74. doi:10.1126/science.1093857

44. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15: R46. doi:10.1186/gb-2014-15-3-r46

45. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26: 1721–1729. doi:10.1101/gr.210641.116

46. Yang L-L, Tang S-K, Huang Y, Zhi X-Y. Low Temperature Adaptation Is Not the Opposite Process of High Temperature Adaptation in Terms of Changes in Amino Acid Composition. Genome Biol Evol. 2015;7: 3426–3433. doi:10.1093/gbe/evv232

47. Jensen DB, Vesth TC, Hallin PF, Pedersen AG, Ussery DW. Bayesian prediction of bacterial growth temperature range based on genome sequences. BMC Genomics. 2012;13 Suppl 7: S3. doi:10.1186/1471-2164-13-S7-S3

48. Taylor TJ, Vaisman II. Discrimination of thermophilic and mesophilic proteins. BMC Struct Biol. 2010;10 Suppl 1: S5. doi:10.1186/1472-6807-10-S1-S5

49. Li Y, Middaugh CR, Fang J. A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. BMC Bioinformatics. 2010;11: 62. doi:10.1186/1471-2105-11-62

50. Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. J Microbiol Methods. 2011;84: 67–70. doi:10.1016/j.mimet.2010.10.013

51. Ogg CD, Patel BKC. *Thermotalea metallivorans* gen. nov., sp. nov., a thermophilic, anaerobic bacterium from the Great Artesian Basin of Australia aquifer. Int J Syst Evol Microbiol. 2009;59: 964–971. doi:10.1099/ijs.0.004218-0

52. Zhao W, Zeng X, Xiao X. *Thermococcus eurythermalis* sp. nov., a conditional piezophilic, hyperthermophilic archaeon with a wide temperature range for growth, isolated from an oil-immersed chimney in the Guaymas Basin. Int J Syst Evol Microbiol. 2015;65: 30–35. doi:10.1099/ijs.0.067942-0

53. Puente-Sánchez F, Sánchez-Román M, Amils R, Parro V. *Tessaracoccus lapidicaptus* sp. nov., an actinobacterium isolated from the deep subsurface of the Iberian pyrite belt. Int J Syst Evol Microbiol. 2014;64: 3546–3552. doi:10.1099/ijs.0.060038-0

54. Debnath R, Saikia R, Sarma RK, Yadav A, Bora TC, Handique PJ. Psychrotolerant antifungal *Streptomyces* isolated from Tawang, India and the shift in chitinase gene family. Extrem Life Extreme Cond. 2013;17: 1045–1059. doi:10.1007/s00792-013-0587-8

55. Chen Z, Feng D, Zhang B, Wang Q, Luo Y, Dong X. Proteomic insights into the temperature responses of a cold-adaptive archaeon *Methanolobus psychrophilus* R15. Extrem Life Extreme Cond. 2015;19: 249–259. doi:10.1007/s00792-014-0709-y

56. Pivovarova TA, Kondrat'eva TF, Batrakov SG, Esipov SE, Sheĭchenko VI, Bykova SA, et al. Phenotypic features of *Ferroplasma acidiphilum* strains Yt and Y-2. Mikrobiologiia. 2002;71: 809–818.

57. Dsouza M, Taylor MW, Ryan J, MacKenzie A, Lagutin K, Anderson RF, et al. *Paenibacillus darwinianus* sp. nov., isolated from gamma-irradiated Antarctic soil. Int J Syst Evol Microbiol. 2014;64: 1406–1411. doi:10.1099/ijs.0.056697-0

58. Sukweenadhi J, Kim Y-J, Lee KJ, Koh S-C, Hoang V-A, Nguyen N-L, et al. *Paenibacillus yonginensis* sp. nov., a potential plant growth promoting bacterium isolated from humus soil of Yongin forest. Antonie Van Leeuwenhoek. 2014;106: 935–945. doi:10.1007/s10482-014-0263-8

59. Kwon YM, Yang S-H, Kwon KK, Kim S-J. *Nonlabens antarcticus* sp. nov., a psychrophilic bacterium isolated from glacier ice, and emended descriptions of *Nonlabens marinus* Park et al. 2012 and *Nonlabens agnitus* Yi and Chun 2012. Int J Syst Evol Microbiol. 2014;64: 400–405. doi:10.1099/ijs.0.056606-0

60. Stieglmeier M, Klingl A, Alves RJE, Rittmann SK-MR, Melcher M, Leisch N, et al. *Nitrososphaera viennensis* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum *Thaumarchaeota*. Int J Syst Evol Microbiol. 2014;64: 2738–2752. doi:10.1099/ijs.0.063172-0

61. Cui H-L, Tohty D, Liu H-C, Liu S-J, Oren A, Zhou P-J. *Natronorubrum sulfidifaciens* sp. nov., an extremely haloalkaliphilic archaeon isolated from Aiding salt lake in Xin-Jiang, China. Int J Syst Evol Microbiol. 2007;57: 738–740. doi:10.1099/ijs.0.64651-0

62. Itoh T, Yamaguchi T, Zhou P, Takashina T. *Natronolimnobius baerhuensis* gen. nov., sp. nov. and *Natronolimnobius innermongolicus* sp. nov., novel haloalkaliphilic archaea isolated from soda lakes in Inner Mongolia, China. Extrem Life Extreme Cond. 2005;9: 111–116. doi:10.1007/s00792-004-0426-z

63. Xin H, Itoh T, Zhou P, Suzuki K, Nakase T. *Natronobacterium nitratireducens* sp. nov., a aloalkaliphilic archaeon isolated from a soda lake in China. Int J Syst Evol Microbiol. 2001;51: 1825–1829. doi:10.1099/00207713-51-5-1825

64. Kern T, Fischer MA, Deppenmeier U, Schmitz RA, Rother M. *Methanosarcina flavescens* sp. nov., a methanogenic archaeon isolated from a full-scale anaerobic digester. Int J Syst Evol Microbiol. 2016;66: 1533–1538. doi:10.1099/ijsem.0.000894

65. Sun L, Toyonaga M, Ohashi A, Tourlousse DM, Matsuura N, Meng X-Y, et al. *Lentimicrobium saccharophilum* gen. nov., sp. nov., a strictly anaerobic bacterium representing a new family in the phylum *Bacteroidetes*, and proposal of *Lentimicrobiaceae* fam. nov. Int J Syst Evol Microbiol. 2016;66: 2635–2642. doi:10.1099/ijsem.0.001103

66. Baek K, Choi A, Kang I, Lee K, Cho J-C. *Kordia antarctica* sp. nov., isolated from Antarctic seawater. Int J Syst Evol Microbiol. 2013;63: 3617–3622. doi:10.1099/ijs.0.052738-0

67. Surendra V, Bhawana P, Suresh K, Srinivas TNR, Kumar PA. *Imtechella halotolerans* gen. nov., sp. nov., a member of the family *Flavobacteriaceae* isolated from estuarine water. Int J Syst Evol Microbiol. 2012;62: 2624–2630. doi:10.1099/ijs.0.038356-0

68. Birkenbihl RP, Neef K, Prangishvili D, Kemper B. Holliday junction resolving enzymes of archaeal viruses SIRV1 and SIRV2. J Mol Biol. 2001;309: 1067–1076. doi:10.1006/jmbi.2001.4761

69. Castillo AM, Gutiérrez MC, Kamekura M, Ma Y, Cowan DA, Jones BE, et al. *Halovivax asiaticus* gen. nov., sp. nov., a novel extremely halophilic archaeon isolated from Inner Mongolia, China. Int J Syst Evol Microbiol. 2006;56: 765–770. doi:10.1099/ijs.0.63954-0

70. Gutiérrez MC, Castillo AM, Kamekura M, Ventosa A. *Haloterrigena salina* sp. nov., an extremely halophilic archaeon isolated from a salt lake. Int J Syst Evol Microbiol. 2008;58: 2880–2884. doi:10.1099/ijs.0.2008/001602-0

71. Cui H-L, Tohty D, Zhou P-J, Liu S-J. *Haloterrigena longa* sp. nov. and *Haloterrigena limicola* sp. nov., extremely halophilic archaea isolated from a salt lake. Int J Syst Evol Microbiol. 2006;56: 1837–1840. doi:10.1099/ijs.0.64372-0

72. Gutiérrez MC, Castillo AM, Pagaling E, Heaphy S, Kamekura M, Xue Y, et al. *Halorubrum kocurii* sp. nov., an archaeon isolated from a saline lake. Int J Syst Evol Microbiol. 2008;58: 2031–2035. doi:10.1099/ijs.0.65840-0

73. Hong H, Kim S-J, Min U-G, Lee Y-J, Kim S-G, Jung M-Y, et al. *Geosporobacter ferrireducens* sp. nov., an anaerobic iron-reducing bacterium isolated from an oil-contaminated site. Antonie Van Leeuwenhoek. 2015;107: 971–977. doi:10.1007/s10482-015-0389-3

74. Söderholm H, Derman Y, Lindström M, Korkeala H. Functional csdA is needed for effective adaptation and initiation of growth of *Clostridium botulinum* ATCC 3502 at suboptimal temperature. Int J Food Microbiol. 2015;208: 51–57. doi:10.1016/j.ijfoodmicro.2015.05.013

75. Davidova IA, Wawrik B, Callaghan AV, Duncan K, Marks CR, Suflita JM. *Dethiosulfatarculus sandiegensis* gen. nov., sp. nov., isolated from a methanogenic paraffin-degrading enrichment culture and emended description of the family *Desulfarculaceae*. Int J Syst Evol Microbiol. 2016;66: 1242–1248. doi:10.1099/ijsem.0.000864

76. Abin CA, Hollibaugh JT. *Desulfuribacillus stibiiarsenatis* sp. nov., an obligately anaerobic, dissimilatory antimonate- and arsenate-reducing bacterium isolated from anoxic sediments, and emended description of the genus *Desulfuribacillus*. Int J Syst Evol Microbiol. 2017;67: 1011–1017. doi:10.1099/ijsem.0.001732

77. An TT, Picardal FW. *Desulfocarbo indianensis* gen. nov., sp. nov., a benzoate-oxidizing, sulfate-reducing bacterium isolated from water extracted from a coal bed. Int J Syst Evol Microbiol. 2014;64: 2907–2914. doi:10.1099/ijs.0.064873-0

78. Hahnke S, Langer T, Koeck DE, Klocke M. Description of *Proteiniphilum saccharofermentanssp*. nov., *Petrimonas mucosasp*. nov. and *Fermentimonas caenicolagen*. nov., sp. nov., isolated from mesophilic laboratory-scale biogas reactors, and emended description of the genus *Proteiniphilum*. Int J Syst Evol Microbiol. 2016;66: 1466–1475. doi:10.1099/ijsem.0.000902

79. Hahnke S, Striesow J, Elvert M, Mollar XP, Klocke M. *Clostridium bornimense* sp. nov., isolated from a mesophilic, two-phase, laboratory-scale biogas reactor. Int J Syst Evol Microbiol. 2014;64: 2792–2797. doi:10.1099/ijs.0.059691-0

80. Xu Y, Zhou P, Tian X. Characterization of two novel haloalkaliphilic archaea *Natronorubrum bangense* gen. nov., sp. nov. and *Natronorubrum tibetense* gen. nov., sp. nov. Int J Syst Bacteriol. 1999;49 Pt 1: 261–266. doi:10.1099/00207713-49-1-261

81. Yang S-H, Seo H-S, Woo J-H, Oh H-M, Jang H, Lee J-H, et al. *Carboxylicivirga* gen. nov. in the family *Marinilabiliaceae* with two novel species, *Carboxylicivirga mesophila* sp. nov. and *Carboxylicivirga taeanensis* sp. nov., and reclassification of *Cytophaga fermentans* as *Saccharicrinis fermentans* gen. nov., comb. nov. Int J Syst Evol Microbiol. 2014;64: 1351–1358. doi:10.1099/ijs.0.053462-0

82. Lee G-H, Rhee M-S, Chang D-H, Kwon KK, Bae KS, Yang S-H, et al. *Bacillus solimangrovi* sp. nov., isolated from mangrove soil. Int J Syst Evol Microbiol. 2014;64: 1622–1628. doi:10.1099/ijs.0.058230-0

83. Dunlap CA, Kwon S-W, Rooney AP, Kim S-J. *Bacillus paralicheniformis* sp. nov., isolated from fermented soybean paste. Int J Syst Evol Microbiol. 2015;65: 3487–3492. doi:10.1099/ijsem.0.000441

84. Dunlap CA, Saunders LP, Schisler DA, Leathers TD, Naeem N, Cohan FM, et al. *Bacillus nakamurai* sp. nov., a black-pigment-producing strain. Int J Syst Evol Microbiol. 2016;66: 2987–2991. doi:10.1099/ijsem.0.001135

85. Kim S-J, Dunlap CA, Kwon S-W, Rooney AP. *Bacillus glycinifermentans* sp. nov., isolated from fermented soybean paste. Int J Syst Evol Microbiol. 2015;65: 3586–3590. doi:10.1099/ijsem.0.000462

86. Shi W, Takano T, Liu S. *Anditalea andensis* gen. nov., sp. nov., an alkaliphilic, halotolerant bacterium isolated from extreme alkali-saline soil. Antonie Van Leeuwenhoek. 2012;102: 703–710. doi:10.1007/s10482-012-9770-7

87. Chu Y, Zhu Y, Chen Y, Li W, Zhang Z, Liu D, et al. aKMT Catalyzes Extensive Protein Lysine Methylation in the Hyperthermophilic *Archaeon Sulfolobus* islandicus but is Dispensable for the Growth of the Organism. Mol Cell Proteomics MCP. 2016;15: 2908–2923. doi:10.1074/mcp.M115.057778

88. Sauer DB, Karpowich NK, Song JM, Wang D-N. Rapid Bioinformatic Identification of Thermostabilizing Mutations. Biophys J. 2015;109: 1420–1428. doi:10.1016/j.bpj.2015.07.026

89. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2017;45: D37–D42. doi:10.1093/nar/gkw1070

90. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, et al. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res. 2016;44: D574-580. doi:10.1093/nar/gkv1209

91. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25: 955–964.

92. Seemann T. Barrnap [Internet]. Available: https://github.com/tseemann/barrnap

93. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

94. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29: 2607–2618.

95. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinforma Oxf Engl. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

96. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Comput Sci Eng. 2011;13: 22–30. doi:10.1109/MCSE.2011.37

97. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

98. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9: 90–95. doi:10.1109/MCSE.2007.55

Supporting Information Captions

Figure S1. The genomes available are dominated by mesophiles, bacteria, and repetitively sequenced organisms.

Figure S2. Features are often highly associated.

Figure S3. Using only genomic sequence features is poorly predictive of OGT.

Figure S4. Phylum specific regressions are often strongly predictive.

Figure S5. Class specific regressions can be strongly predictive.

Figure S6. Bioinformatic classification allows for accurate OGT prediction.

Figure S7. Genome size is not necessary for OGT prediction accuracy.

Figure S8. Excluding psychrophiles improves OGT prediction.

Figure S9. OGT prediction validated using previously unknown species-OGT values.

Equation S1. Features and coefficients for the prediction of the OGT for a prokaryote.

Equation S2. Features and coefficients for the prediction of the OGT for an Archaea.

Equation S3. Features and coefficients for the prediction of the OGT for a Bacterium.

Table S1. Correlation of features to OGT.

Table S2. *De novo* predicted OGT for species without a measured OGT in Sauer *et al.* 2015