bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima

Gang Li[1], Kersten S Rabe[2], Jens Nielsen[1,3], Martin KM Engqvist[1*]

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.
[2] Institute for Biological Interfaces 1 (IBG 1), Karlsruhe Institute of Technology (KIT), Group for Molecular Evolution, Karlsruhe, Germany
[3] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

* Corresponding author
E-mail: martin.engqvist@chalmers.se

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

The optimal growth temperature (OGT) of organisms is an important index to estimate the stability of enzymes encoded in their genomes. However, experimental determination of OGT for microorganisms that cannot be cultivated is difficult. Here, we report on the development of a machine learning model that can accurately predict OGT directly from proteome-wide 2-mer amino acid composition. We make use of this model to predict OGTs for 1,438 microorganisms. In a subsequent step we combine OGT data with amino acid composition of individual enzymes to develop a second machine learning model for prediction of enzyme temperature optima (T_{opt}). The resulting model is far superior to using OGT alone for estimating T_{opt} in a dataset of 2,609 enzymes. Finally, we predict T_{opt} for 6.5 million enzymes, covering 4,447 EC numbers, and make the resulting dataset available for researchers, enabling simple identification of enzymes that are potentially functional at extreme temperatures.

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The optimal growth temperature (OGT) of microorganisms is an important physiological parameter. Thus, OGT has been widely used to understand the strategies organisms use to adapt their genomes and proteomes to different environmental conditions[1–3]. OGT has also been an important tool for identifying thermostable enzymes[4] in the field of molecular biology (for example the Taq DNA polymerase) and in industrial applications (for example enzymes for biorefinery processes)[5–7].

Determining the OGT of a microorganism is a laborious process that requires cultivation in temperature-controlled conditions. Furthermore, the number of microorganisms that can be cultured in the lab is only a small fraction of the total diversity in nature[8]. Consequently, the OGT for the vast majority of microbial organisms is currently unknown, and an easy way to computationally estimate the OGT of microbes is in demand. For such computational estimations to be feasible there must be general trends for how quantifiable biological properties change with growth temperature.

Previous studies have revealed many genomic and proteomic features that are strongly correlated with OGT. Examples include the existence of thermophile-specific enzymes[9], the presence or absence of certain dinucleotides[10], the GC content of structural RNAs[11], as well as amino acid composition of the proteome[1,12]. Examples such as these indicate that estimating OGT directly from genomes or proteomes may indeed be feasible.

Statistical tools, such as regression and classification, have been used to model the correlation between OGT and biological features. For example, the OGT of 22 bacteria could be predicted using a linear combination of either dinucleotide or amino acid composition[10]. Additionally, Zeldovich found that the sum fraction of the seven amino acids I, V, Y, W, R, E and L showed a correlation coefficient as high as 0.93 with OGT in a dataset consisting of 204 proteomes of archaea and bacteria[12]. Jensen et al developed a Bayesian classifier to distinguish three thermophilicity classes (thermophiles, mesophiles and psychrophiles) based on 77 bacteria with known OGT[13]. Training datasets containing the OGTs for a large number of organisms have been hard to obtain, something which has prevented the development of state-of-the-art machine learning models for OGT prediction.

Here we use a recently published dataset which contains OGTs for 21,498 microorganisms[14], and combine this with genomic resources to build a machine learning model. The model predicts OGT from a file containing all proteins encoded by an

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

organism's genome. We show that the optimal growth temperature of a microorganism can be accurately predicted solely by the amino acid sequences of its proteome. The model was then used to assign an OGT value for those microorganisms for which we had no experimental values. In a second step we make use of the resulting OGT dataset to improve the prediction of enzyme temperature optima (T_{opt}) by another machine learning approach. The resulting model was used to estimate T_{opt} of enzymes in BRENDA[15]. Finally, to make easy use of our OGT model and generated datasets we develop a computational method which we call Tome (Temperature optima for microorganisms and enzymes) and make it freely available for reuse (https://github.com/EngqvistLab/Tome). This method has two applications. First, it can be used to predict OGT from the amino acid composition of proteins encoded by an organism's genome. Second, it can be used to find enzyme homologs with a predicted T_{opt} in a given temperature range. Both these applications will be highly valuable for the microbiology, protein engineering and biotechnology communities.

Results

Collection of optimal growth temperature and proteomes of microorganisms

To build a machine learning model that can predict OGT from the amino acid composition of proteins encoded by an organism's genome we first established a training dataset. To this end, we downloaded an OGT dataset (https://doi.org/10.5281/zenodo.1175608), which contains data for 21,498 microorganisms, including bacteria, archaea and eukarya[14]. Using this dataset, all proteins from 5,761 organisms from RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) and Ensembl genomes (http://ensemblgenomes.org/) could be associated with an OGT value (we refer this as the annotated dataset), while proteins from an additional 1,803 organisms could not be associated with an OGT value (we refer this as the unannotated dataset) (Fig. 1a).

For each organism in both the annotated and unannotated dataset we calculated the global amino acid monomer and dipeptide frequencies. However, some organisms in the dataset contain only a small number of protein sequences, as a consequence the amino acid composition obtained from those sequences may not represent the true amino acid composition of the complete proteome. To address this problem we applied a filtering step. As it was unclear how many protein sequences are required to obtain a stable amino acid composition we designed three different metrics (see Methods for details) to test how much protein sequence data was needed to obtain a stable amino

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

acid composition. (Fig. 1b). For each organism in the annotated dataset the three metrics were calculated for every protein sequence added in order to observe at which point the values stop fluctuating. Using this analysis on amino acid monomer frequencies we found that at least 10^5 amino acids are needed to get a stable amino acid composition (Fig. 1c, d, e). Repeating this analysis for amino acid dipeptides resulted in the same threshold (Supplementary Fig. 1).

A further concern was that the order in which proteins appear in the input files may affect our cutoff analysis. For this reason, proteins from 17 organisms with different sizes of available proteomes were randomly selected. For each of these organisms the order in which protein sequences appear was shuffled and the three metrics were calculated. The shuffling, with subsequent analysis, was repeated 100 times. As expected, the analysis shows a high initial variability, where few sequences have been analyzed, but with increasing numbers of averaged proteins the values stabilized and converged (Fig. 1f, g, h). From this analysis it is clear that the arrangement of proteins in a proteome has a negligible effect when the proteome size is larger than 10^5, and we therefore only chose organisms with at least 10^5 amino acids in the dataset for further analysis.

This approach resulted in a dataset with 5,532 organisms annotated with OGT and 1,438 un-annotated organisms. In the annotated dataset the magnitude of protein number in each organism follows a normal distribution centered around 3,000 (Fig. 1i). The OGT distribution is, however, highly skewed with the majority of organisms having an OGT in the range 25-30℃ and at 37℃ (Fig. 1j). The number of organisms in the data set with an OGT higher than 40℃ is 425 (340 for higher than 50°C). The majority of microorganisms in the dataset are bacteria, with around 12% being archaea and eukarya (Fig. 1k).

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

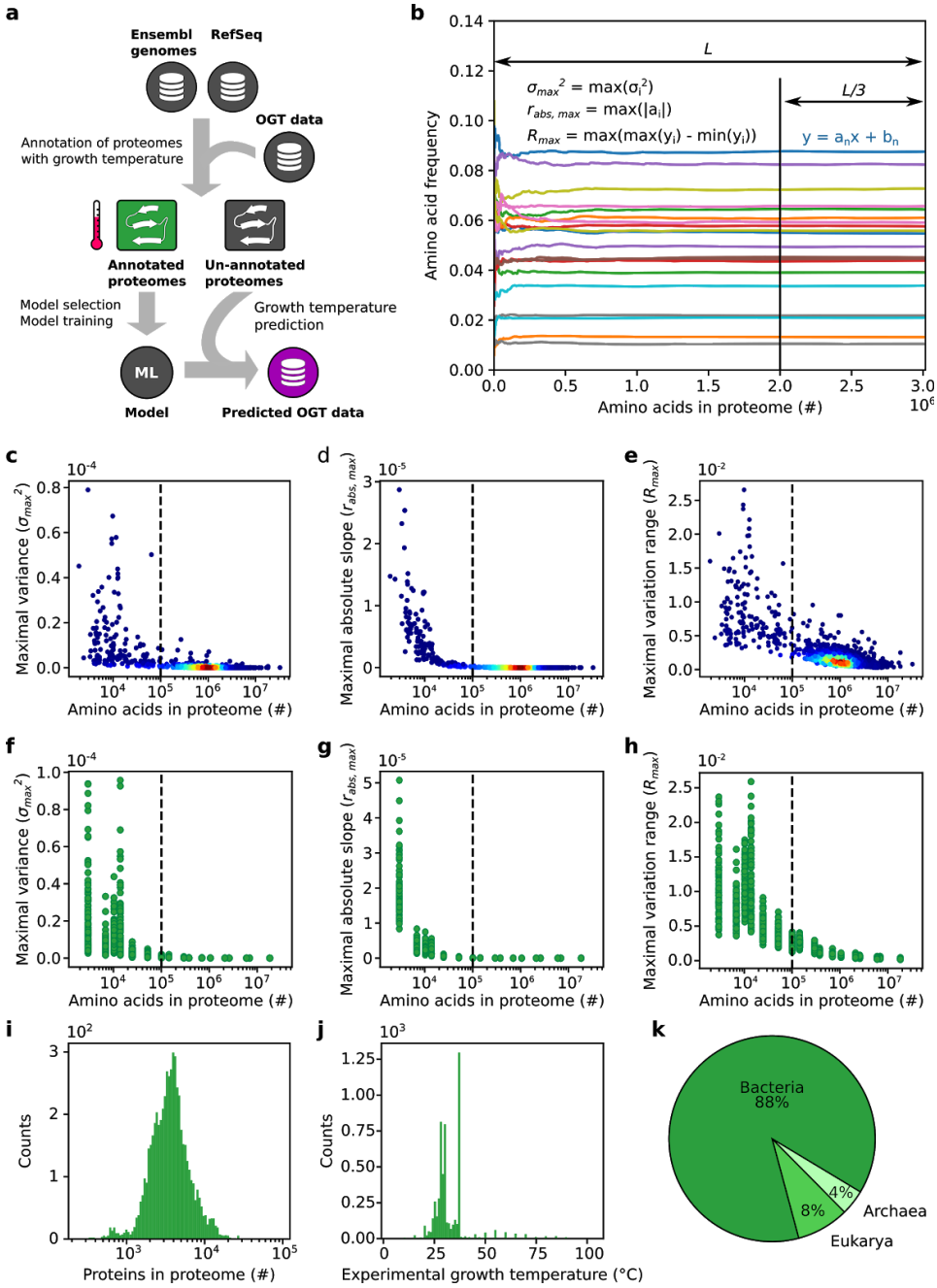

Figure 1 | Variability in amino acid frequencies decrease log-linearly with proteome size. (**a**) Schematic overview of process to build a machine learning model to predict OGT. Protein records from Ensembl genomes (bacteria and fungi) and RefSeq (bacteria, archaea and fungi) were downloaded. Sequences were annotated with the growth temperature of the organism from which they originate. Sequences from organisms that could not be annotated, i.e. for which there is no available information about the OGT for the organism, were retained in a separate un-annotated dataset. Amino acid frequencies of the annotated sequences were used to train a statistical model. This model was in turn used to predict growth temperatures for the un-annotated dataset. OGT: optimal growth temperature. (**b**) The frequency of each amino acid was plotted against the number of amino acids used to calculate the

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

frequency. The final third part was fitted to a linear model to get the absolute slope value ($|a_i|$), as well as its frequency variance (σ_i^2) and varying range (R). The maximal $|a_i|$, σ_i^2 and R of 20 amino acids of each proteome ($r_{abs,max}$, σ^2_{max} and R_{max}) give measures of whether frequencies were stable. The calculated (**c**) σ^2_{max}, (**d**) $r_{abs,max}$ and (**e**) R_{max} of all species in the dataset were plotted against the number of amino acids in the proteome. The dashed line indicates the cutoff for the selection of proteomes based on size. Effect of protein order on (**f**) σ^2_{max}, (**g**) $r_{abs,max}$ and (**h**) R_{max}. 17 proteomes with different size were randomly selected. Proteins in each proteome were shuffled 100 times and the three metrics for each shuffled proteome were calculated. (**i**) Distribution of proteome sizes in the annotated dataset after filtering. (**j**) Distribution of growth temperatures in the annotated dataset after filtering. (**k**) Proportion of species belonging to the three different taxonomic superkingdoms in the filtered dataset.

OGT can be accurately predicted from amino acid composition of the proteome

Protein amino acid composition is strongly correlated with OGT[10,12]. For this reason we decided to train machine learning models using the amino acid composition as features. For each organism in the annotated dataset we calculated the global amino acid monomer frequencies (20 features) as well as amino acid dipeptide frequencies (400 features). To get the best feature set and statistical model for the prediction of OGT, we tested six different regression models and compared their performance on the monomer dataset and the dipeptide dataset. As shown in Fig. 2a, a 5-fold cross-validation was applied to evaluate the performance of different regression models. Using the 20 amino acid frequencies, non-linear models (SVR and Random forest) perform much better than linear models (Linear, Elastic net, Bayesian ridge regression). The superior performance of non-linear models suggests that there are important non-linear relationships between amino acid frequencies and OGT. In contrast, all models except decision tree show an almost identical performance when using dipeptide frequencies. Since models trained on each of the two datasets individually show good performance we reasoned that models trained on the combined datasets may be even better performing. However, contrary to this expectation the six models trained using both monomer dataset and dipeptide dataset together do not show improvement (Fig. 2a).

Validation of the SVR model for growth temperature prediction

In our comparison of different models we found that an SVR model trained on the dipeptide dataset produces the best OGT prediction results. The final SVR model was trained on the whole dipeptide dataset and stored for further use. This model can explain an astounding 95% (88% by cross-validation) of the variance in OGT (Fig. 2b). Leveraging this model, the OGT of 1,438 organisms in the unannotated dataset were predicted (Fig. 1a).

The OGT predictions were validated using two separate approaches. First, we

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

performed a manual literature search to find experimentally obtained OGTs for a subset of the organisms (for which no experimental OGT was present in our original dataset). We randomly sampled 54 of the organisms with predicted OGTs, in a manner that ensured even spread across temperatures. For 45 of the 54 organisms, OGT values could indeed be found in published peer-reviewed articles (Supplementary Table 1). The agreement between the predicted OGT and the ones collected from literature is very high, with a Pearson's correlation coefficient of 0.96 (Fig. 2c). Second, we seized on the fact that the average temperature optimum of catalysis (T_{opt}) of at least five enzymes from an organism shows a Pearson correlation above 0.75 with growth temperature[14]. Essentially, the catalytic optimum of an enzyme tends to be close to the organism growth temperature. The correlation between OGT and T_{opt} provides a way to validate predicted OGTs using experimental data for enzyme temperature optima obtained from the BRENDA database (https://www.brenda-enzymes.org/). Of the 1,438 organisms with predicted OGT only 23 were found to have at least five enzymes with T_{opt} available in BRENDA. Plotting the mean enzyme optima against the predicted OGT for these organisms reveals a strong correlation, with a Pearson's correlation coefficient of 0.77 (Fig. 2d). Indeed, this correlation is the same as that obtained with experimentally determined organism OGTs[14], again showing that the predicted OGTs are very accurate.

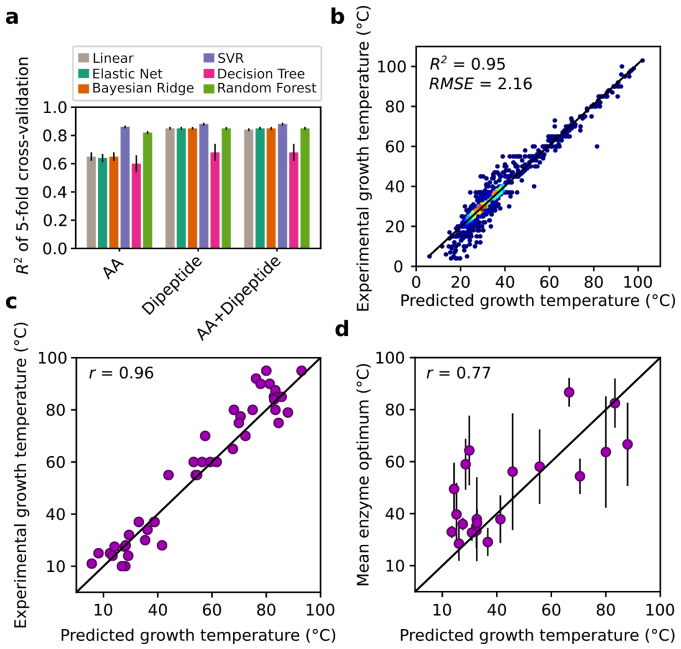

Figure 2 | Model development for OGT prediction. (**a**) R^2 score obtained by a 5-fold cross-validation for six different regression models. Error bars represent the standard derivation of R^2 scores. (**b**) Performance of the final SVR (support vector regression) model trained on dipeptide data. The correlation between predicted organism growth temperatures and those present in the original annotated dataset

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was evaluated. RMSE: root mean square error. Colors indicate the density of the points. (**c**) Correlation between literature values for growth temperatures and predicted growth temperatures. Species for unannotated dataset were sampled at random, but with ensuring equal coverage over the temperature range. Growth temperatures for these organisms were obtained by manually searching the primary scientific literature. (**d**) Correlation between the mean enzyme temperature optima and predicted growth temperatures for each species present in both datasets. Only organisms with optima for at least five enzymes are shown. Error bars show the standard deviation. In (**c** and **d**) *r* denotes Pearson's correlation coefficient.

Improved estimation of enzyme temperature optima using machine learning

In biotechnology and protein engineering OGT is often used to guide the discovery of thermostable enzymes[7]. Even though this strategy has proven very useful, OGT yields an imprecise estimation of enzyme temperature optima (T_{opt}), as indicated by the Pearson correlation between the T_{opt} of individual enzymes and OGT being as low as 0.48[14]. We hypothesized that the accuracy of this estimation could be improved using machine learning. We collected 2,609 enzymes that have both T_{opt} and protein sequence data in the BRENDA database[15], and that could also be mapped to organisms with experimental OGT using the annotated dataset (Fig. 3a, b). We first tested the accuracy of using OGT as an estimation of T_{opt} and found that only 25% of the enzyme T_{opt} variance could be explained (Fig. 3c, black bar). Then, to improve the accuracy of this estimation using machine learning, we extracted three feature sets from the enzyme sequences, namely amino acid frequencies, dipeptide frequencies and other basic protein properties like length, isoelectric point etc. (See Methods). Six regression models were trained and tested on these feature sets individually, as well as the two and three sets combined, with a 5-fold cross-validation approach. As shown in Fig. 3c, the best model (SVR) trained on amino acid frequencies achieved a slightly improved accuracy compared to OGT, as quantified by an R^2 score of around 30%. Using dipeptide frequencies alone in combination with amino acid frequencies did not further improve the accuracy.

Since OGT and sequence-derived features each produce estimates of similar accuracy (25% and 30%, respectively) we tested whether their combined use could boost predictive power. Surprisingly, the best model (random forest) trained on the combination of amino acid frequencies and OGT almost doubled the model predictive accuracy to over 50%. Further inclusion of other basic enzyme properties (see Methods) did not further improve the accuracy (Fig. 3c).

To generate a final model for the prediction of enzyme temperature optima the random forest model was re-trained using the full set of amino acid frequencies and OGT data

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Fig. 3d). In this final model, OGT of the source organism is the most informative individual feature, whereas the 20 amino acid frequencies combined contribute over half of the predictive power of the model (Fig. 3e).

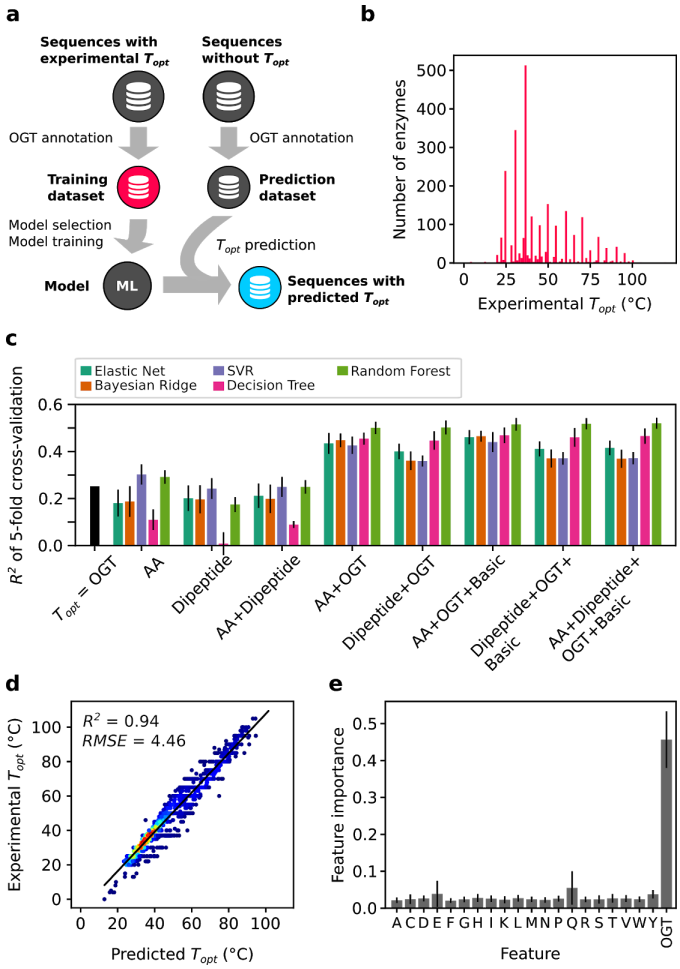

Figure 3 | Model development for prediction of enzyme temperature optima. (**a**) Schematic overview of process to build a T_{opt} prediction model. (**b**) The distribution of enzyme temperature optima in training dataset. (**c**) 5-fold cross-validation results for five regression models on different feature sets. The T_{opt} = OGT bar shows the explained variance when using OGT as the estimation of enzyme T_{opt}. Error bars shows the standard deviation of R^2 scores obtained in 5-fold cross validation. AA, amino acid frequencies; Dipeptide, dimer frequencies; OGT, optimal growth temperature of source organism; Basic, basic information of proteins, like length, isoelectric point etc., see details in Methods section. (**d**) Performance of the final random forest model trained on AA+OGT data. The correlation between predicted and experimental T_{opt} was evaluated. RMSE: root mean square error. Colors indicate the density of the points. (**e**) The feature importance in the final random forest model. Error bars indicates the standard deviation of feature importances of 1,000 estimators.

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Annotating enzymes in BRENDA using OGT and predicted T_{opt}

Currently, a main resource for T_{opt} data is the BRENDA database[15]. However, there are approximately 12 million native protein sequences in BRENDA while there are only about 33,000 T_{opt} records, many of which are not connected to a protein sequence. Due to the very small number of features (20 amino acid frequencies plus the OGT of the source organism) required to predict T_{opt} using our random forest model, it is computationally feasible to carry out that prediction for millions of enzymes in BRENDA.

First, experimentally determined OGTs[14] and the 1,438 OGTs predicted with the SVR model (Fig. 2b) were combined to generate a dataset containing the OGT of 22,936 microorganisms. Using this combined OGT dataset 6,507,076 out of 12,115,011 enzymes (54%) in BRENDA could be annotated with the OGT value of their source organism, of which 909,954 enzymes (14%) were contributed by the predicted OGT values. In a second step, these OGT annotations were combined with the amino acid frequencies extracted from each enzyme sequence. Our random forest model (Fig. 3c) was applied to this data to estimate the T_{opt} of each individual enzyme. This prediction dramatically added to the experimental T_{opt} values in BRENDA, increasing them 197-fold (Fig. 4a) and covering 4,447 different EC numbers (Fig. 4b). Moreover, the temperature coverage, i.e. the minimal and maximal T_{opt} for an enzyme class, of the vast majority these EC numbers (3,721 of 4,447) were expanded (Fig. 4b). The predicted enzyme T_{opt} and annotated OGT values of these enzymes are freely available for download and re-use (https://zenodo.org/record/2539114, https://doi.org/10.5281/zenodo.2539114).

As can be seen in Fig. 4c, many of the predicted enzyme T_{opt} values differ significantly from the OGT of the source organism. For enzymes from organisms with OGT below 40°C many have T_{opt} higher than the OGT. In contrast, enzymes from thermophiles generally have a lower T_{opt} than the OGT. These results are in good agreement with previous findings comparing experimental OGT of organisms with average enzyme T_{opt}[14]. For three representative organisms we show that the distribution of predicted T_{opt} values are indeed consistent with experimental values (Supplementary Fig. 4).

Tome: a command line tool for OGT prediction and identification of enzyme homologues with different T_{opt}

To ensure easy access to the OGT predictive model for the scientific community, as well as the enzyme data with estimated T_{opt}, we developed the command line tool Tome (Temperature optima for microorganisms and enzymes). This tool is simple to use and has two fundamental applications: (1) prediction of OGT from a file containing protein sequences encoded by an organism's genome; (2) identification of functional

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

homologues within a specified temperature range for an enzyme of interest. For the prediction of OGT, a list of proteomes in fasta format[16] is provided as input and the temperature predictions are returned as an output. While this tool will perform predictions on any input given, we stress that the tool has been trained on bacteria, archaea and a only small set of eukarya - mostly fungi and protists. Predictions on organisms which do not fall into these categories may result in inaccurate results. For the identification of enzyme functional homologs with different estimated T_{opt}, one can either simply specify an EC number and temperature range of interest to get all enzyme sequences from BRENDA matching the criteria. Alternatively, the sequence of an enzyme of interest can be provided in fasta format. The algorithm will then perform a protein BLAST[17] and an additional output file will be generated containing only homologous enzymes (default e-value cutoff is 10^{-10}) within the specified temperature range. Full instructions regarding installation and usage of the Tome tool is available online (https://github.com/EngqvistLab/Tome).

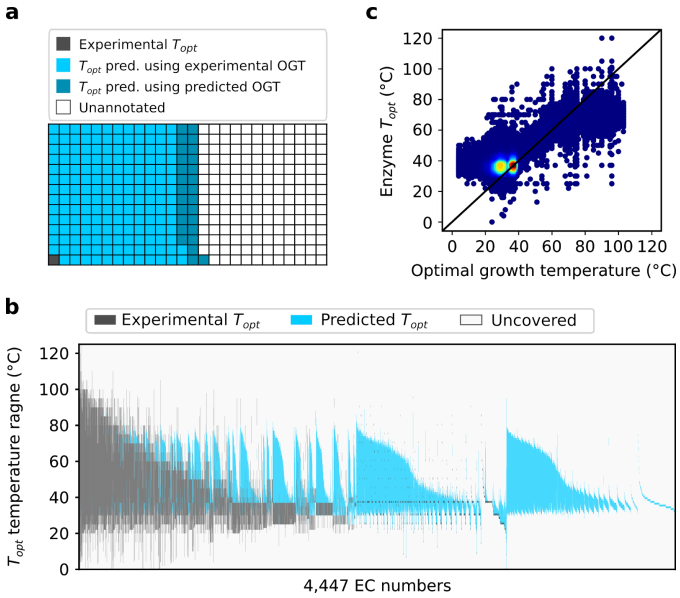

Figure 4 | Prediction of enzyme temperature optima. (**a**) Visual representation of the number of the enzymes with experimental T_{opt} in BRENDA and the number of enzymes for which T_{opt} was predicted leveraging experimental and predicted OGTs. Each box represents ~33,050 enzymes. There are 12,115,011 enzymes in total. Pred. is an abbreviation of predicted. (**b**) A visual representation of the T_{opt} temperature coverage for each EC number after annotation. The span between the highest and lowest T_{opt} for each enzyme is indicated. Experimental (BRENDA) and predicted T_{opt} values are shown in different colors. (**c**) Comparison between OGT of source organism and predicted and experimental T_{opt} values of enzymes. Colors indicate the density of the points.

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

A main finding of this study is that the OGT of microorganisms from the three domains of life can, using non-linear regression models, be accurately predicted from the sequence information of proteins encoded by their genomes. The fact that the same predictive model can be used for widely different organisms implies that thermoadaptation in proteins follow generalizable evolutionary trends. What those trends are is the topic of a forthcoming study by these authors. The OGT prediction model generated in this study should prove useful in assigning growth temperatures for un-culturable organisms for which sequence data is available, and even for those microorganisms where OGT has been determined but where the information is absent from major databases.

Our OGT prediction model is the most accurate model to date when compared to other published models (Supplementary Fig. 2 and Supplementary Fig. 3). We propose that the high predictive accuracy seen in our OGT prediction model results from two features of our approach; the size and quality of the training data used, and the use of non-linear regression models. Our training dataset consisted of 5,533 microorganisms, including 4,974 bacteria, 222 archaea and 337 eukarya. This training set is much larger than those used in other approaches, such as 22 bacteria[10], 77 bacteria[13], 204 prokaryotes[12]. As a direct consequence of the increased size of the dataset, we could train models that are more general applicable. We find that in general non-linear models outperform linear models when using amino acid frequencies (Fig. 2a). This suggests that the linear models used previously such as that from Nakashima et al.[10] might be further improved by non-linear regression to correlate the amino acid frequencies to OGT.

While there are only a few published models predicting organism OGT, there are many methods for the estimation of protein stability. These methods fall into two main categories; predicting the stability of whole proteins, and predicting the stability change in a protein upon amino acid substitutions. Machine learning has been used extensively for the prediction of stability change upon amino acid substitutions[18–22], while only a few methods have been developed for the prediction of stability of whole protein empirically[23–26]. The desire to accurately predict protein stability and stability changes largely stems from a real-world need to engineer proteins with increased thermostability, as well as identifying natural ones that are already stable, for use in industrial applications[6]. However, computational prediction of protein stability is challenging since it usually needs an accurate calculation of Gibbs-free energy change of protein unfolding process[25,26], which relies mainly on high-quality protein structures, which are

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

limited in number.

Another important finding of this study is that T_{opt} values of enzymes can be predicted from a combination of enzyme sequence information and physiological parameters of the source organisms (OGT in the present study). OGT has been used to estimate protein stability[4], which is based around the fact that a typical protein should be functional at the growth temperature optimum of its parent organism, something which is supported by a strong correlation between OGT and average enzyme T_{opt} from the same organism[14,27]. However, we found that OGT by itself can only explain 25% of individual enzyme T_{opt} variance when tested on a dataset containing 2,609 enzymes from BRENDA (Fig. 3b, c). To improve the accuracy of this estimation, a second machine learning approach was applied. Machine learning models trained on sequence features are not significantly better than using OGT as estimation (Fig. 3c). However, the combined use of OGT and sequence features almost doubled the explained variance to 51% (Fig. 3c). This is a remarkable result that demonstrates a clear importance of physiological parameters in the estimation of protein properties. We speculate that inclusion of more samples and extraction of more descriptive features (both from sequence and physiological parameters) in conjunction with advanced machine learning models, like deep learning[28] may further improve the prediction of enzyme T_{opt}. The R^2 score of 51% obtained for T_{opt} predictions in this study could be used as a benchmark accuracy for future model development.

In addition to our scientific findings we used experimental as well as predicted OGT to predict T_{opt} for enzymes in the BRENDA database. We could assign T_{opt} for 54% (6,507,076) of the all enzyme sequences and up to 59% of all EC numbers in the database. The number of sequences annotated with T_{opt} computed using predicted OGT is 909,954. This is remarkable given that only around 33,000 enzymes in BRENDA has an experimentally determined T_{opt}. The annotation expanded both lower and upper bound of T_{opt} for almost every EC number (Fig. 4b). This T_{opt} annotation should prove very useful for identification of both thermostable enzymes[7] as well as cold-active enzymes[29] for applications in protein engineering and biotechnology. We provide a data file with all these annotations (https://zenodo.org/record/2539114, https://doi.org/10.5281/zenodo.2539114) as well as a script with which to query the data for unrestricted re-use as part of our Tome package (https://github.com/EngqvistLab/Tome).

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Proteome dataset

The bulk of protein sequence data used in this work was obtained from Ensembl Genomes release 37, obtained in September 2017 (http://ensemblgenomes.org/). For all archaea and bacteria listed at ftp://ftp.ensemblgenomes.org/pub/bacteria/release-37/fasta/ fasta files containing protein sequences were downloaded. Similarly, fasta files containing protein sequences for all fungi listed at ftp://ftp.ensemblgenomes.org/pub/fungi/release-37/fasta/ were downloaded. As a complement to the Ensembl Genome data we made use of protein data from RefSeq release 87, obtained in March 2017 (https://www.ncbi.nlm.nih.gov/refseq/). Fasta files containing a nonredundant set of protein sequences for each organism were downloaded from ftp://ftp.ncbi.nlm.nih.gov/refseq/release/ for archaea, bacteria, fungi and protozoa.

In many cases the Ensembl Genomes and RefSeq datasets both contained information for the same organism, or for several strains of the same organism. Therefore, to combine the two datasets, the following steps were followed: First, where multiple strains from the same organism were present in the Ensembl Genomes dataset, the strain with the largest file size, indicating the greatest number of amino acids in the downloaded fasta file, was selected for analysis. Other strains for that organism were discarded. Second, where the same organism was present in both the Ensembl Genomes and RefSeq datasets the one from Ensembl Genomes was retained and the one from RefSeq was discarded. In this way a protein dataset comprising protein sequence data for 7,565 microorganisms was obtained. Of these 5,325 originated from Ensembl Genomes and 2,240 originated from RefSeq.

For each organism in the protein dataset we attempted to annotate it with its optimal growth temperature. In this annotation procedure organism names were stemmed to the species level (ignoring strain designations) and cross-referenced with a published dataset containing growth temperatures for 21,498 microorganisms (https://doi.org/10.5281/zenodo.1175608). Growth temperatures could be associated with the protein sequence data from 5,762 organisms, whereas 1,803 were left unannotated.

Estimation of threshold

For each proteome, the total length of each protein was calculated. Then the amino acid frequencies and the total number of residues of the first n proteins (n = 1, 2, ..N, were N

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

is the total number of proteins) were calculated sequentially. The data points in the last one-third of all residues added were used to measure the stability of the calculated amino acid frequencies. Three different metrics were designed: (1) the absolute slope value $|a_i|$ in the linear regression between the number of residues and amino acid frequency; (2) frequency variance of these selected frequencies (σ_i^2) and (3) varying range (R), the difference between maximal frequency and minimal frequency. Ideally, 0 was expected for all these three metrics if there is an absolutely stable amino acid frequency in a given proteome. Finally, for each proteome, the maximal $|a_i|$, σ_i^2 and R of 20 amino acids of each proteome ($r_{abs,max}$, σ^2_{max} and R_{max}) were used to measure whether frequencies were stable.

To test the effect of the protein order in a proteome in the above analysis, a shullfing strategy was applied. Firstly, equal coverage over the $\log_{10}$-transformed proteome size range 3-7.5 was ensured by performing the random sampling in 20 bins. One proteome was randomly selected for each bin and this resulted in 17 selected proteomes as there is no proteome in 3 of these bins. The order of the proteins in each proteome was randomly shuffled and then $r_{abs,max}$, σ^2_{max} and R_{max} were calculated. Each proteome was shuffled for 100 times.

Machine learning workflow for OGT model

20 amino acid frequencies and 400 dipeptide frequencies were extracted for each proteome. Then, each of these features were normalized by $x_{N,i} = \frac{x_i - u_i}{\delta_i}$, where x_i is the values of feature i, u_i and δ_i are mean and standard derivation of x_i, respectively. The following six models were selected and their performance were tested on the annotated and filtered proteome dataset using single amino acid frequencies (AA), dipeptide frequencies (Dipeptide) or the two together (AA+Dipeptide): Linear regression (Linear), bayesian ridge, elastic net, decision tree, support vector regression (SVR) and random forest. 5-fold cross-validation was used for the calculation of R^2 scores. For SVR, elastic net, decision tree and random forest models, an additional 3-fold internal cross-validation were used to optimize the hyperparameters. The model with the highest R^2 score was selected and trained, without cross-validation, on the whole dataset. For the prediction of OGT for those un-annotated organisms, dipeptide frequencies were normalized by $x_{N,i} = \frac{x_i - u_i}{\delta_i}$, where x_i is the values of feature i. u_i and δ_i are mean and standard derivation of feature i in the training dataset, respectively.

OGT Model validation

For validating the OGT prediction model we sampled 54 species with predicted growth temperatures (for which no growth temperatures were available in the original dataset)

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

at random. Equal coverage over the temperature range 0-100℃ was ensured by performing the random sampling in 10 bins, each spanning a 10℃ temperature range. The primary scientific literature was then manually searched to obtain documented experimental growth temperatures for the sampled organisms. For 45 organisms a documented growth temperature could be found, for 9 organisms it could not. The accuracy of predicted OGT was assessed by computing the Pearson correlation with experimental OGT.

In a second approach to validating the OGT prediction model we used Python scripts and the Zolera SOAP package (https://pypi.python.org/pypi/ZSI/) to extract all available experimentally determined enzyme temperature optima from the BRENDA enzyme database release 2018.2 (July 2018). Data coming from the same enzyme was de-duplicated by averaging temperature optima from records with the same EC number and originating from the same organism. For each organism with catalytic optima for more than five enzymes the arithmetic mean of those optima were calculated. Those organisms present in both the BRENDA enzyme data as well as the dataset with predicted OGT were identified through cross-referencing species names. The accuracy of predicted OGT was assessed by computing the Pearson correlation between predicted OGT and mean catalytic optima of enzymes.

Machine learning workflow for T_{opt} model

UniProt identifiers for proteins with an experimentally determined catalytic optimum were obtained from the “TEMPERATURE OPTIMUM” table in the web pages of the BRENDA database, release 2018.2 (July 2018). These identifiers were filtered to retain only those associated with an organism with experimentally determined OGT. After further filtering to remove sequences containing “X” (unknown amino acid), a dataset with 2,609 enzymes was generated. The protein sequences for each of these identifiers were downloaded from the UniProt database in fasta format.

The following features were extracted for each enzyme: (1) 20 amino acid frequencies (AA); (2) 400 dipeptide frequencies (Dipeptide); (3) OGT of its source organism; (4) Basic features including protein length, isoelectric point, molecular weight, aromaticity[30], instability index[31], gravy[32] and fraction of three secondary structure units: helix, turn and sheet. These features were extracted with the module Bio.SeqUtils.ProtParam.ProteinAnalysis in Biopython (version 1.70)[33]. Additionally, siix binary features were extracted: EC=1, 2, 3, 4, 5, 6. These numbers represent the first digit in a EC number. All features except binary features were normalized as described in section “Machine learning workflow for OGT model”. The following five models were tested on the resulting dataset: bayesian ridge, elastic net, decision tree, support vector

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

regression (SVR) and random forest. The linear model was not used due to its poor performance on any datasets containing dipeptide frequences (negative R^2 scores by cross-validation). The performance of the five regression models was tested using the same cross-validation strategy as for OGT. In addition, to test the accuracy of using OGT of the organism as an estimation of enzyme T_{opt}, the R^2 score between each enzymes T_{opt} and associated OGT was calculated. The model with the highest R^2 score was chosen and trained on the full training dataset.

BRENDA annotation

Protein sequence data for each EC class was obtained by downloading comma-separated flatfiles from the BRENDA database version 2018.2 (July 2018). Each sequence in these files contain information regarding source organism as well as unique UniProt identifiers. Where possible, each protein sequence was associated with an OGT value by mapping the source organism name to the OGT dataset from (https://doi.org/10.5281/zenodo.1175608). Those sequences were firstly mapped to the existing T_{opt} values in BRENDA by matching EC-UniProt id pair. For those enzymes without any experimental T_{opt} values, the amino acid frequencies were calculated (ignore all 'X' in the sequence). All 20 amino acid frequencies as well as the OGT variable were normalized by $x_{N,i} = \frac{x_i - u_i}{\delta_i}$, where x_i is the values of feature i. u_i and δ_i are mean and standard derivation of feature i in the original training dataset, respectively. Finally, the normalized values were used for the prediction of T_{opt} by the previously generated random forest regressor trained on the AA+OGT datasets. The predicted enzyme T_{opt} and annotated OGT values of these enzymes are freely available for download and re-use (https://zenodo.org/record/2539114, https://doi.org/10.5281/zenodo.2539114).

Software

All machine learning analysis were conducted with scikit-learn package (version 0.19.1)[34] using Python version 2.7.14. The module and model hyperparameters used are listed in Supplementary Table S2.

Code availability

Python code for proteome analysis, machine learning and data visualization are available from the authors upon request. The source code for the Tome package is available under a permissive GPLv3 license at GitHub (https://github.com/EngqvistLab/Tome).

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

The computations were performed on resources at Chalmers Centre for Computational Science and Engineering (C3SE) provided by the Swedish National Infrastructure for Computing (SNIC). We thank Pia Schwitters for her assistance with performing the manual literature search for organism growth temperatures.

Author contributions

GL, JN and MKME conceptualized the research. JN acquired funding to support the project. MKME generated the proteome and BRENDA datasets and performed data curation. GL performed the computational and statistical data analysis. GL, KSR and MKME interpreted results. GL wrote the computer code for the Tome package. GL and MKME created the publication figures. GL and MKME wrote the initial draft of the paper. GL, KSR, JN and MKME carried out revisions on the initial draft and wrote the final version.

Competing interests

The authors declare no competing interests.

Funding

GL and JN have received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie program, project PAcMEN (grant agreement No 722287). We also acknowledge funding from the Novo Nordisk Foundation (grant no. NNF10CC1016517) and the Knut and Alice Wallenberg Foundation. KSR acknowledges financial support by the Helmholtz program "BioInterfaces in Technology and Medicine".

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Hickey, D. A. & Singer, G. A. C. 10.1186/gb-2004-5-10-117. Genome Biol 5, 117 (2004).
2. Saunders, N. F. W. et al. Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea Methanogenium frigidum and Methanococcoides burtonii. Genome Res. 13, 1580–1588 (2003).
3. Venev, S. V. & Zeldovich, K. B. Thermophilic Adaptation in Prokaryotes Is Constrained by Metabolic Costs of Proteostasis. Mol. Biol. Evol. 35, 211–224 (2018).
4. Pezeshgi Modarres, H., Mofrad, M. R. & Sanati-Nezhad, A. ProtDataTherm: A database for thermostability analysis and engineering of proteins. PLoS One 13, e0191222 (2018).
5. Demirjian, D. C., Morís-Varas, F. & Cassidy, C. S. Enzymes from extremophiles. Curr. Opin. Chem. Biol. 5, 144–151 (2001).
6. Turner, P., Mamo, G. & Karlsson, E. N. Potential and utilization of thermophiles and thermostable enzymes in biorefining. Microb. Cell Fact. 6, 9 (2007).
7. Vieille, C. & Zeikus, G. J. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol. Mol. Biol. Rev. 65, 1–43 (2001).
8. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. Annu. Rev. Microbiol. 57, 369–394 (2003).
9. Forterre, P. A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet. 18, 236–237 (2002).
10. Nakashima, H., Fukuchi, S. & Nishikawa, K. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. J. Biochem. 133, 507–513 (2003).
11. Galtier, N. & Lobry, J. R. Relationships between genomic G+C content, RNA secondary

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. 44, 632–636 (1997).

12. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput. Biol. 3, e5 (2007).
13. Jensen, D. B., Vesth, T. C., Hallin, P. F., Pedersen, A. G. & Ussery, D. W. Bayesian prediction of bacterial growth temperature range based on genome sequences. BMC Genomics 13 Suppl 7, S3 (2012).
14. Engqvist, M. K. M. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. BMC Microbiology 18, (2018).
15. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. (2018). doi:10.1093/nar/gky1048
16. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U. S. A. 85, 2444–2448 (1988).
17. Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009).
18. Fariselli, P., Martelli, P. L., Savojardo, C. & Casadio, R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. Bioinformatics 31, 2816–2821 (2015).
19. Quan, L., Lv, Q. & Zhang, Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 32, 2936–2946 (2016).
20. Dehouck, Y. et al. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25, 2537–2543 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

21. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. 320, 369–387 (2002).
22. Chen, C.-W., Lin, J. & Chu, Y.-W. iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinformatics 14 Suppl 2, S5 (2013).
23. Ku, T. et al. Predicting melting temperature directly from protein sequences. Comput. Biol. Chem. 33, 445–450 (2009).
24. Dill, K. A., Ghosh, K. & Schmit, J. D. Physical limits of cells and proteomes. Proc. Natl. Acad. Sci. U. S. A. 108, 17876–17882 (2011).
25. Murphy, K. P. & Freire, E. Structural energetics of protein stability and folding cooperativity. J. Macromol. Sci. Part A Pure Appl. Chem. 65, 1939–1946 (1993).
26. Oobatake, M. & Ooi, T. Hydration and heat stability effects on protein unfolding. in Computer Aided Innovation of New Materials II 1307–1310 (1993).
27. Dehouck, Y., Folch, B. & Rooman, M. Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. Protein Eng. Des. Sel. 21, 275–278 (2008).
28. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).
29. Santiago, M., Ramírez-Sarmiento, C. A., Zamora, R. A. & Parra, L. P. Discovery, Molecular Mechanisms, and Industrial Applications of Cold-Active Enzymes. Front. Microbiol. 7, 1408 (2016).
30. Lobry, J. R. & Gautier, C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 22, 3174–3180 (1994).
31. Guruprasad, K., Reddy, B. V. & Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from

bioRxiv preprint doi: https://doi.org/10.1101/522342; this version posted January 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

its primary sequence. Protein Eng. 4, 155–161 (1990).

32. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132 (1982).

33. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423 (2009).

34. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).