Protein features as determinants of wild-type glycoside hydrolase thermostability

Geertz-Hansen, Henrik Marcus; Kiemer, Lars; Nielsen, Morten; Stanchev, Kiril; Blom, Nikolaj; Brunak, Søren; Petersen, Thomas Nordahl

Published in:
Proteins: Structure, Function, and Bioinformatics

Link to article, DOI:
10.1002/prot.25357

Publication date:
2017

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Protein features as determinants of wild-type glycoside hydrolase thermostability

Short title: Determinants of glycoside hydrolase thermostability

#Key words: protein thermal stability, melting temperature, biofuels, biomass, cellulases, lignocellulose, bioinformatics, machine learning, prediction

#Authors:
Henrik Marcus Geertz-Hansen1,2,3, Lars Kiemer3, Morten Nielsen2,4, Kiril Stanchev2, Nikolaj Blom1,2, Søren Brunak2,5 and Thomas Nordahl Petersen2

#Affiliations:
1Novo Nordisk Foundation Center for Biosustainability, Kemitorvet, Building 220, Technical University of Denmark, DK-2800 Lyngby, Denmark
2Department of Bio and Health Informatics, Kemitorvet, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark
3Novozymes A/S, Krogshøjvej 36, DK-2880 Bagsværd, Denmark
4Instituto de Investigaciones Biotecnologicas, Universidad Nacional de San Martin, San Martin, B 1650 HMP, Buenos Aires, Argentina
5Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, Blegdamsvej 3B, University of Copenhagen, DK-2200 Copenhagen N, Denmark

#Corresponding author: Thomas Nordahl Petersen, tnp@cbs.dtu.dk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an ‘Accepted Article’, doi: 10.1002/prot.25357

© 2017 Wiley Periodicals, Inc.
Received: Mar 27, 2017; Revised: Jun 26, 2017; Accepted: Jul 21, 2017

This article is protected by copyright. All rights reserved.
Abstract

Thermostable enzymes for conversion of lignocellulosic biomass into biofuels have significant advantages over enzymes with more moderate thermostability due to the challenging application conditions. Experimental discovery of thermostable enzymes is highly cost intensive, and the development of in-silico methods guiding the discovery process would be of high value. To develop such an in-silico method and provide the data foundation of it, we determined the melting temperatures of 602 fungal glycoside hydrolases from the families GH5, 6, 7, 10, 11, 43 and AA9 (formerly GH61). We, then used sequence and homology modeled structure information of these enzymes to develop the ThermoP melting temperature prediction method. Furthermore, in the context of thermostability, we determined the relative importance of 160 molecular features, such as amino acid frequencies and spatial interactions, and exemplified their biological significance. The presented prediction method is made publicly available at http://www.cbs.dtu.dk/services/ThermoP.

Introduction

Advanced biofuels are receiving considerable attention as a sustainable and renewable source of energy due to concerns about energy security, depletion of fossil fuels and greenhouse gas emissions\(^1\). The non-food energy feedstock lignocellulose (such as agricultural waste, forestry residues and woody/herbaceous crops) is the most abundant biomass on Earth, with an estimated annual availability of one billion tons by 2030 in the US alone\(^2\). Lignocellulose is primarily composed of plant cell-wall polysaccharides such as cellulose and hemicellulose. These polymers contain hexose and pentose sugars that can be converted into substituents for current transportation fuels and other biomaterials through microbial fermentation\(^3\). Commercial production
plants are now operating in North- and South America as well as Asia and Europe, however, the recalcitrance of lignocellulose to deconstruction into fermentable sugars is still a major barrier to large-scale deployment of second-generation biofuels. Enzymatic hydrolysis is considered to have the greatest potential as an economically viable process, but techno-economic analyses suggest that the contribution of enzymes to the overall biofuel production costs must be further reduced.

Saccharification of thermo-mechanically or chemically pretreated biomass performed at elevated temperatures entails increased conversion rates resulting in shorter hydrolysis time and lower enzyme loading. In addition, it allows for greater biorefinery process flexibility, and reduces the risk of microbial contamination. For these reasons, thermophilic bacteria and fungi are receiving significant attention as sources of thermostable cellulolytic enzymes. Thermostable enzymes are produced both by thermophilic and mesophilic organisms, but are difficult to identify in the latter, as the donor organism’s growth temperature cannot be used as a direct measure of the enzymes thermostability.

Attempts to unravel the molecular mechanisms of enzyme thermostabilization have been carried out in a number of ways, including comparison of entire proteomes, proteins of a single family or a range of families from thermo- and mesophilic organisms. These studies have indicated that no single mechanism is responsible for thermostabilizing enzymes, but rather is the product of a combination of contributing factors, whose relative importance may differ for various enzyme families and structural classes. However, these earlier analyses were limited by few available structures or lack of experimentally determined enzyme stabilities, and the significant effects of salts and pH on thermostability were not considered. These factors may explain some of the difficulties to reach consistent conclusions about the
determinants of enzyme thermostability, but it might as well be true that no single or few features drives the protein stability.

This study focuses on the challenge of novel enzyme discovery for improved enzyme cocktail development. It presents, to our knowledge, the first method for prediction of wild-type glycoside hydrolase (GH) enzyme thermostability; a key characteristic for cost-effective hydrolysis of lignocellulosic biomass. We reasoned that a large, diverse set of enzymes characterized under identical conditions would enable us to better infer the determinants of protein thermostability and so, we have characterized more than 600 recombinant enzymes from seven CAZy families involved in cellulose or hemicellulose deconstruction. Using computational approaches, we demonstrate that a complex wild-type enzyme property of high industrial importance can be modeled from sequence and homology modeled structure information, and exemplify new family-specific strategies to stabilize GH enzymes.

**Online methods**

**Differential scanning fluorimetry**

Cloning in *E. coli*, expression in *A. niger* or *A. oryzae* and preparation of fungal glycoside hydrolase (GH) enzymes were performed as described elsewhere. A typical sample for differential scanning fluorimetry (DSF) was prepared by mixing 10 μl enzyme sample with 20 μl Sypro Orange protein gel stain (Invitrogen) diluted 200x in 400mM acetate buffer, pH 5.0. DSF melting curves in the temperature range of 25-95°C were obtained with a LightCycler 480 II real-time PCR instrument (Roche) at a scan rate of 200°C/h. dF/dT analysis and melting temperature calculation was done using the LightCycler 480 software (release 1.5.0 SP4, Roche).
Catalytic domain identification

To identify family and catalytic domain boundaries for the enzymes of interest, position-specific scoring matrices (PSSMs) for families GH5, 6, 7, 10, 11, 43 and AA9 were obtained from position-specific iterated (PSI)-BLAST (version 2.2.28+) searches of fungal sequences listed under the ‘Characterized’ tab on the CAZy website (http://www.cazy.org). For GH5, individual PSSMs were created for the subfamilies 2, 4, 5, 7, 9, 12, 15, 16, 22, 23, 31 and 49 for which at least one fungal sequence was listed on the website as characterized. Preference was given to seed sequences obtained from crystal structures as these typically comprise only the catalytic domain. Known signal peptides or peptides predicted using SignalP 4.1 were removed from seed sequences for subfamilies where no crystal structure existed. The used seed sequences are listed in Table 1.

Full-length sequences of the characterized enzymes were searched against a database of the obtained PSSMs using RPS-BLAST. The family and catalytic domain boundaries were assigned to the query sequences according to the alignment with the highest alignment score.

Phylogenetics

Putative GH5, 6, 7, 10, 11, 43 and AA9 amino acid sequences were obtained from the CAZy website. Sequences marked as ‘fragmented’ were excluded. Catalytic domains were identified using previously described PSSMs combined with the catalytic domains of the characterized sequences. A summary of the sequence sets is provided in Table 2. These sets were filtered for redundancy (95% sequence identity) using UCLUST and the filtered sequences were aligned with MUSCLE. Phylogenetic trees were constructed using FastTree and visualized with iTOL. Standard
parameters were used for all programs.

**Calculation of protein features from sequence and structure**

The principal mechanisms of enzyme thermostability are encoded in the structure, and the spatial interactions that keep it folded. We included sequence-derived features in our analysis as these might provide additional information about amino acid or local structure environment preferences facilitating important interactions. Protein structure homology models of the catalytic domains were obtained using CPHmodels 3.2\textsuperscript{24}, an automated, single template homology modeling server. Seven structure models from GH5 were obtained by ‘remote homology modeling’ (sequence identity between query sequence and PDB template < 30%). The minimum standard score (z-score) of these models was 15. As a z-score above 10 indicates a high reliability model, all ‘remote homology’ models were included in the study. Query specific sequence identity and coverage of PDB template are summarized in Figure S1.

Residue-specific relative solvent accessibility (RSA) and secondary structure assignment were calculated from these models using DSSP\textsuperscript{25}. All modeled amino acids were grouped into three RSA classes: buried (RSA value ≤ 9%), intermediate (9% < RSA value ≤ 36%) and exposed (RSA value > 36%). Subsequently, the distribution in percent between the classes was calculated for each amino acid.

Secondary structure assignments were also grouped into three classes: The H class comprised by DSSP class G, H and I; the E class comprised by DSSP class B and E; and the C class comprised by the remaining DSSP classes: ".", S and T.

The atom coordinates of the homology models were used to determine counts of a number of structural interactions, based solely on distance criteria (criteria obtained from Tina \textit{et al.}\textsuperscript{26}). The following interactions were considered (specific
distance criterion \( d \) is noted in parenthesis): disulfide bonds \( (d < 2.2\text{Å} \) between the sulfur atoms), salt bridges \( (d < 6\text{Å} \) between the side-chain nitrogen of Lys, the geometric midpoint of the two side-chain nitrogens of Arg or His, and the geometric midpoint of the two side-chain oxygens of Asp or Glu), hydrogen bonds \( (d < 3.5\text{Å} \) between donor oxygen and acceptor nitrogen, \( d < 4.0\text{Å} \) between acceptor thiol and donor nitrogen or oxygen), hydrophobic interactions \( (d < 5.0\text{Å} \) between side-chain carbon atoms of Ala, Val, Leu, Ile, Phe, Tyr, Pro, Trp or Met), interactions between aromatic amino acids (\( \pi \)-stacking, \( 4.5\text{Å} < d < 7.0\text{Å} \) between geometric midpoints of the aromatic ring of Tyr or Phe, or the indole ring of Trp), interactions between cations and aromatic amino acids (cation-\( \pi \) interactions, \( d < 6.0\text{Å} \) between geometric midpoints of the aromatic ring of Tyr or Phe, or the indole ring of Trp and side-chain nitrogen of Lys, the geometric midpoint of the two side-chain nitrogens of Arg or His) as well as interactions between the thiols of Cys residues and aromatic amino acids \( (d < 5.3\text{Å} \) between the sulfur atom and geometric midpoints of the aromatic ring of Tyr or Phe, or the indole ring of Trp). Hydrogen bonds were further subdivided into three bins: main chain atoms to main chain atoms (MM), main chain atoms to side chain atoms (MS) and side chain atoms to side chain atoms (SS). All structural interactions were feature-encoded as both raw counts as well as “pr. 100 residues”, producing 18 features in total. Frequencies of the 20 standard amino acids and counts of total residues of the GH domain and total atoms of the structure homology models were also calculated. The 160 features used for predictive model development are summarized in Table 3.

All features were converted from raw scores to standard scores (\( z \)-scores) using the mean \( (\mu) \) and standard deviation \( (\sigma) \) calculated from the raw scores of all sequences within a family or combination of families: \( z = (x - \mu) / \sigma \). In addition, the
target values, i.e. the melting point temperatures, were linearly transformed to the range [0:1] using:

\[ T_{m,\text{norm}} = \frac{(T_m - T_{m,\text{min}})}{(T_{m,\text{max}} - T_{m,\text{min}})} \]

where \( T_{m,\text{max}} \) and \( T_{m,\text{min}} \) is the maximal and minimal melting temperature for each family, respectively.

**BLAST based assignment of enzyme melting temperatures**

Aligning a query protein sequence against a database of all 602 glycoside hydrolases enables a simple melting temperature (Tm) prediction by inheriting Tm of the best target as the one with the highest alignment score. Alignments were made with BLAST (version 2.2.28+)\(^1\) using the E-value to identify the highest scoring target sequence.

**Sequence homology partitioning**

Enzymes within each GH family are described in terms of protein features derived from both primary, secondary and tertiary structure information. Sequences with high mutual sequence identity may contain significant differences when described in terms of protein features. For that reason we used homology partitioning of all unique sequences based on sequence identity rather than homology reduction. Within each GH family, sequences were divided into four partitions such that the maximal sequence identity between sequences from two different partitions is less than 80%. Within each partition pairwise sequence identities above 80% were allowed. Partition sizes are summarized in Table 4.

**Feature combinations and artificial neural networks**
Optimal subsets of descriptive protein features were selected for modeling of enzyme melting temperature using a two-step approach: First, simple feed-forward fully connected artificial neural networks with a single hidden layer of two units were trained on all single and pairwise combinations of the 160 protein features. The networks were trained using a four-fold nested cross-validation procedure as described in Petersen et al.\(^\text{19}\) One of four partitions was left out for evaluation while a full three-fold cross-validation was performed on the remaining partitions. In each of these runs, two partitions were used for training the networks and one for selecting the optimal number of training cycles to achieve minimal error. This procedure was repeated four times, such that all partitions were used in each of the roles, producing 12 networks in total (Figure S2). The networks were trained using a standard back-propagation procedure\(^\text{27}\) for a maximum of 3,000 training epochs and the optimal training cycle was selected according to the lowest test error. The tested feature combinations (12,880 in total) were ranked based on the minimum averaged test error from 12 networks. Second, networks with four hidden units were trained using the top 20 non-redundant features selected from the ranked list of pairs. All feature combinations up to seven features were tested (137,979 in total) and ranked according to the minimum averaged test error. The top five feature combinations were selected for further optimization. We are aware that we introduce a slight risk of over-fitting in both of these feature selection steps. However, it was necessary due to the limited number of experimental data, i.e. 602 glycoside hydrolase sequences.

High and low temperature cutoffs for training data balancing were manually selected based on the test-error distribution of the selected top five feature combinations. An example of test-error distributions used for selecting the temperature cutoffs is shown in Figure S3. Feature-target correlations close to the
extremities of the family-specific melting temperature distributions were balanced by being presented to the neural networks 2, 4 and 6 times and the optimal balancing was determined from the test set Pearson’s correlation coefficient, the mean absolute error and the predicted temperature span (difference between the highest and lowest predicted melting temperature). The reported prediction performances were obtained on the evaluation sets of a four-fold nested cross-validation procedure with the combination of selected feature combinations described above, and the obtained balanced training sets.

**Results**

*Enzyme diversity and characterization*

The melting temperatures of 602 fungal enzymes from CAZy families GH5, 6, 7, 10, 11, 43 and AA9 (formerly GH61) were determined under identical experimental conditions (Figure 1(a); see Online Methods for details). The characterized enzymes were selected to broadly cover the known fungal sequence diversity (Figure 1(b) and Figures S10-13), and comprise the largest set of thermal stability assayed enzymes ever studied. The most thermostable enzymes are found in GH5 and GH10. For bioinformatics analyses, we considered only the catalytic domain sequence, however it should be noted that for catalytic domains (CD) associated with a carbohydrate-binding module (CBM), we observed a significantly higher average melting temperature of 5-10°C in four of seven studied families, compared to CDs without associated CBMs (Figure S4).

To identify molecular factors important for enzyme thermostability, we calculated the correlation between experimentally determined melting temperatures and 160 descriptive features derived from sequences and structure models (Figure 2
and Figure S5). Additionally, a bootstrap procedure was performed with 1000 permutations of the melting temperatures for each of the 160 features within a family to assure statistical significance of the correlations. In total, 19 features were identified having significant correlations to the melting temperatures (p-value < 0.01) of one or more families (Table S1). Five of these features were identified as significant in more than two families: the frequencies of tyrosine and threonine had positive correlation in three families (GH5, GH6 and GH10), whereas a few features anti-correlated with the melting temperature. Three features were observed in more than two families, namely the frequency of buried threonine (GH5, GH10 and GH11), buried asparagine (GH6, GH7 and GH43) and the frequency of alanine (GH5, GH7 and AA9). It should also be noted that spatial interactions ranked among the most strongly correlating features in four of seven studied families. The most important spatial features are shown for all families in the supplementary Excel sheet.

Exemplifying molecular mechanisms of thermostability

Thermostability is a complex enzyme trait that may be strongly influenced by minute changes to the sequence and structure, which may not be reflected in the features used in this study. Nevertheless, the presented feature correlations can be used as guide for exemplifying the mechanisms of enzyme thermostabilization. As an example, the number of disulfide bridges in GH6 enzymes varies between two and three in the enzymes under study. Two of these disulfide bridges were found to be conserved among all 43 characterized cellobiohydrolase II enzymes, whereas 15 of 43 sequences also had a third disulfide bridge (exemplified in Figure S6). On average the latter group had melting temperatures that were 5°C higher compared to the group with only two disulfide bridges (p-value = 0.02, one-sided t-test). A similar observation
was made in family GH10 where enzymes with an additional disulfide bridge at the N-terminal (exemplified in Figure S7) had an increased melting temperature of 7°C on average (p-value = 0.002, one-sided t-test).

The β-strand propensity of phenylalanine is the second most strongly correlating feature to GH5 enzyme thermostability. Comparing the two GH5 structures with most and least Phe residues in β-strand (referred to as A and B, respectively), which ranked 12 and 64 in terms of melting temperature out of 130 GH5 enzymes, it was observed that the core forming β-barrel ties the Phe residues in A (highlighted in orange and red in Figure 3) closer together in space, compared to the sparsely distributed Phe residues in B (highlighted in cyan) that are not located in β-strand. Although the number of Phe residues in A and B are almost identical at 14 and 13, respectively, the amino acid sequence comprising structure A facilitates more than twice as many aromatic-aromatic interactions and 37% more hydrophobic interactions involving phenylalanine, compared to B.

Prediction of wild-type thermostability

Given the large number and sequence diversity of enzymes characterized, we hypothesized that the presented data set could facilitate the development of a prediction method of wild-type enzyme melting temperature, despite the convoluted nature of the trait. The ThermoP method described here integrates up to 20 protein sequence and structure features for prediction of melting temperature, out of 160 individual features included. We first mapped the predictive capacity of all pairwise combinations of the 160 features for each GH family. Typically, pairs of individually strongly correlated or anti-correlated features ranked highest (Tables S2-8; selected features highlighted on Figure 2 and Figure S5). All combinations up to seven of the
non-redundant set of 20 features were then calculated, and the five best feature combinations for each family were selected to make up the final prediction method (Tables S9-15; further model optimization is described in Supplementary Results and Figure S3). The top five combinations were selected based on test set performances as a well performing compromise for all GH-families.

For all studied families, the feature-based approach outperformed the BLAST-based assignment of melting temperature as judged from the magnitude of the Pearson correlations (Figure 4(A)) and the mean absolute prediction errors (Figure 5). However, as seen in Figure 4A, the Pearson correlation coefficients (PCC) for our feature-based method and the BLAST-based Tm assignment were quite similar for the GH7 family. Significant PPC values between the two methods were observed for all but the GH7 family. Two-sided p-values were calculated to test for statistical significance using R (library psych version 1.5.4)\textsuperscript{28}. The following p-value were obtained: GH5 \textit{p}=2.78e-02, GH6 \textit{p}=5.34e-05, GH7 \textit{p}=0.37, GH10 \textit{p}=3.18e-03, GH11 \textit{p}=1.28e-04, GH43 \textit{p}=2.89e-09, AA9 \textit{p}=1.03e-05, ALL \textit{p}=3.36e-17). Given our data set of characterized enzymes, we were best able to model the thermostability of family GH43 hemicellulases (PCC of 0.82 and error of 4.7°C) whereas the endoglucanases and cellobiohydrolases of family GH7, and the monoxygenases of AA9 proved the most difficult (PCCs of 0.60). Linear regression plots for the feature-based approach are shown in Figure S8. The largest difference in prediction performance between the two approaches was also observed for AA9, where BLAST-based assignment of melting temperature resulted in a Pearson correlation of only -0.02. Linear regression plots for the BLAST-based approach are shown in Figure S9.

The ThermoP method development was carried out using a nested cross-validation procedure to limit bias in the estimate of prediction performance on
sequences not included in our data set. As an additional validation of the feature selection, we also developed a neural network based model for GH5 from a data set with permuted target values. The top-ranked feature pairs from this training showed an increased prediction error compared to the model training on the original data set (Tables S16 and S17), and the overall prediction performance of the final model was considerably lower, with a Pearson correlation of 0.33 and mean absolute error of 8.1, compared to 0.61 and 6.6, respectively, for the model trained on the original dataset.

Discussion

The presented work provides new insights into the enzyme family-specific protein features conferring thermostability as well as the first prediction algorithm of wild-type enzyme melting temperature. Understanding the basis of biophysical enzyme properties such as thermostability is obscured by the large sequence and structural diversity demonstrated by these molecules. Individual melting temperature determinations, performed under identical experimental conditions, of 602 heterologously expressed fungal glycoside hydrolases enabled an isolated view on the determinants of this complex trait. This approach is considerably different from previous studies of protein thermostability, which were commonly based on assignment of the studied proteins as meso- or thermostable based on the growth temperature of the donor organism\(^9,12,29-31\).

While the majority of enzymes had a melting temperature in the range of 40-80°C, a number of the characterized GH5 and GH10 enzymes were able to withstand temperatures around 90°C or above. These two families share a number of stabilizing descriptive protein features, as well as the same protein fold, namely the common TIM-barrel fold, suggesting that this framework facilitates more thermostable
structures than for example the β-jelly roll fold of GH7 and GH11. The observation that catalytic domains attached to one or more carbohydrate-binding modules (CBM) are on average more thermostable than those lacking a CBM likely reflect an increased importance of the CBM in keeping the enzyme associated with the substrate at elevated temperatures.

Generally, notable differences among the top 10 strongest correlating and anti-correlating features for each of the seven investigated families indicate that no single rule for enzyme thermostabilization exists. However, the exemplified translations of feature correlations to specific molecular mechanisms of thermostabilization demonstrate the strength of a data-driven approach to study protein thermostability. Such exemplifications may serve as wild-type enzyme diversity inspired protein-design strategies for engineering of stabilized variants from the studied families. However, a protein feature correlating with melting temperature does not necessarily mean causation. For example, the neighborhood of a specific amino acid residue may influence its impact on stability, but the neighborhood is not captured within a single feature.

The melting temperature data set enabled development of a supervised prediction method for any given enzyme sequence from GH families 5, 6, 7, 10, 11, 43 and AA9. It performs its non-linear real-value predictions in a family-specific feature space, defined by a simple selection-method from the set of 160 features considered initially. The method enables rapid screening of large sequence sets for enzymes with desired melting temperature. Transfer of thermostability by sequence-similarity is limited to highly similar sequences, whereas the feature-based algorithm also performs well on more distantly related sequences. The feature-based prediction algorithm is available online at http://www.cbs.dtu.dk/services/ThermoP.
We have demonstrated here that a complex biophysical trait of fungal glycoside hydrolases can be modeled *in silico*, based on experimental data generated under identical conditions. Despite the large size of the data set, the method would benefit from even more training data. In particular, experimental data from sequence representatives of phylogenetic clades not covered by the current data set would broaden the application of the prediction algorithm. Generally, a greater understanding of enzyme properties such as thermostability and the availability of predictive models like ThermoP, benefit enzyme discovery in turn leading to even better tools for efficient conversion of biomass into renewable fuels.

**Acknowledgements**

The authors would like to thank Kim Borch, Henrik Frisner, Susanne Christensen and Nikolaj Spodsberg for providing the melting temperature data. The Novo Nordisk Foundation (grant agreement NNF14CC0001) and Novozymes A/S supported the work.

**References**


4. Klein-Marcuschamer D, Oleskowicz-Popiel P, Simmons BA, Blanch HW. The


11. Panasik N, Brenchley JE, Farber GK. Distributions of structural features


20. Edgar RC. Search and clustering orders of magnitude faster than BLAST.


30. Glyakina A V., Garbuzynskiy SO, Lobanov MY, Galzitskaya O V. Different


FIGURES LEGENDS

Figure 1. Temperature distribution and diversity of characterized sequences. (a) Family-specific summary of experimentally determined melting temperatures. The 1st and 3rd quartile outline the colored range, and whiskers extend 1.5 times the interquartile range. The number of characterized enzymes per family was: 141 for GH5, 99 for GH6, 108 for GH7, 87 for GH10, 61 for GH11, 55 for GH43 and 72 for AA9 (formerly known as GH61). (b) Phylogenetic tree of GH5 sequences (n = 2095). Clade and leaf background color indicate the taxonomic origin in terms of kingdom.
for each sequence (green = bacteria, blue = eukaryota, yellow = archaea, red = metagenomic, pink = viruses). Sequences with determined melting temperature are highlighted with black boxes at the rim of the tree. A larger version of the tree and trees for the other families studied are available in the supplement (Figures S10-16).

**Figure 2.** Descriptive protein feature correlation with enzyme melting temperature. Family-specific descriptive features ranked according to correlation with enzyme melting temperature and sorted by Pearson correlations (PCC). Features marked in red showed a direct correlation with melting temperature, whereas those marked in yellow were anti-correlated. Features also among the top 10 by the neural networks (ANN) for prediction of melting temperature are marked. The complete feature ranking is available in the supplement (Figure S5).

**Figure 3.** Structural alignment of the two GH5 enzymes with most (9 of 14) and least (0 of 13) phenylalanine residues in β-strand (ΔTm = 16°C). The β-barrel of the TIM-barrel fold is highlighted in the structure homology model of the most stable enzyme where Phe residues in β-strand are colored orange and remaining Phe residues are colored red. The phenylalanines of the least stable enzyme are colored cyan.

**Figure 4.** Thermostability prediction performance benchmark of ThermoP and a sequence similarity based method. (A) Cross-validated ThermoP performance (blue) compared to all-versus-all BLAST-based melting temperature assignment (green). A larger Pearson’s correlation coefficient indicates a better prediction performance. (B) Mean absolute prediction error of the ThermoP and BLAST methods binned according to sequence similarity between each sequence in the data set and its nearest
neighbor. The error bars represent standard error of the mean.

**Figure 5.** Performance of the artificial neural network and BLAST-based prediction models for enzyme thermostability by means of the mean absolute error (the lower bar the smaller prediction error). The error bars represent standard error of the mean.

**Table 1.** PSI-BLAST seed sequences for catalytic domain identification. PDB entries are listed with their 4-letter PDB ID and the remaining IDs correspond to Uniprot entries.

**Table 2.** Catalytic domain sequences used for phylogenetics. Sequences obtained from CAZy and sequences characterized in this work were clustered using UCLUST.

**Table 3.** Summary of features used for predictive modeling of wild-type enzyme thermostability.

**Table 4.** Sequence partitions used for the cross-validated predictive modeling of enzyme melting temperature. Sequences were partitioned such that two sequences from different partitions share a maximum of 80% sequence identity on a per family basis. Within the same partition, two sequences can share more than 80% sequence identity. For GH5, only 130 of 141 characterized enzymes were successfully structure homology modeled.
Figure 1. Temperature distribution and diversity of characterized sequences. (a) Family-specific summary of experimentally determined melting temperatures. The 1st and 3rd quartile outline the colored range, and whiskers extend 1.5 times the interquartile range. The number of characterized enzymes per family was: 141 for GH5, 99 for GH6, 108 for GH7, 87 for GH10, 61 for GH11, 55 for GH43 and 72 for AA9 (formerly known as GH61). (b) Phylogenetic tree of GH5 sequences (n = 2095). Clade and leaf background color indicate the taxonomic origin in terms of kingdom for each sequence (green = bacteria, blue = eukaryota, yellow = archaea, red = metagenomic, pink = viruses). Sequences with determined melting temperature are highlighted with black boxes at the rim of the tree. A larger version of the tree and trees for the other families studied are available in the supplement (Figures S10-16).
Figure 2. Descriptive protein feature correlation with enzyme melting temperature. Family-specific descriptive features ranked according to correlation with enzyme melting temperature and sorted by Pearson correlations (PCC). Features marked in red showed a direct correlation with melting temperature, whereas those marked in yellow were anti-correlated. Features also among the top 10 by the neural networks (ANN) for prediction of melting temperature are marked. The complete feature ranking is available in the supplement (Figure S5).
Figure 3. Structural alignment of the two GH5 enzymes with most (9 of 14) and least (0 of 13) phenylalanine residues in β-strand (ΔTm = 16°C). The β-barrel of the TIM-barrel fold is highlighted in the structure homology model of the most stable enzyme where Phe residues in β-strand are colored orange and remaining Phe residues are colored red. The phenylalanines of the least stable enzyme are colored cyan.
Figure 4. Thermostability prediction performance benchmark of ThermoP and a sequence similarity based method. (A) Cross-validated ThermoP performance (blue) compared to all-versus-all BLAST-based melting temperature assignment (green). A larger Pearson’s correlation coefficient indicates a better prediction performance. (B) Mean absolute prediction error of the ThermoP and BLAST methods binned according to sequence similarity between each sequence in the data set and its nearest neighbor. The error bars represent standard error of the mean.
Figure 5. Performance of the artificial neural network and BLAST-based prediction models for enzyme thermostability by means of the mean absolute error (the lower bar the smaller prediction error). The error bars represent standard error of the mean.
<table>
<thead>
<tr>
<th>Family (subfamily)</th>
<th>Seed sequence for PSSM generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH5_2</td>
<td>2A3H</td>
</tr>
<tr>
<td>GH5_4</td>
<td>2JEQ</td>
</tr>
<tr>
<td>GH5_5</td>
<td>1GZJ</td>
</tr>
<tr>
<td>GH5_7</td>
<td>1RH9</td>
</tr>
<tr>
<td>GH5_9</td>
<td>3N9K</td>
</tr>
<tr>
<td>GH5_12</td>
<td>Q4WXA7 (full length)</td>
</tr>
<tr>
<td>GH5_15</td>
<td>Q5B6Q3 (mature)</td>
</tr>
<tr>
<td>GH5_16</td>
<td>A2QFW8 (mature)</td>
</tr>
<tr>
<td>GH5_22</td>
<td>B5M6A4 (full length)</td>
</tr>
<tr>
<td>GH5_23</td>
<td>B5MEI8 (mature)</td>
</tr>
<tr>
<td>GH5_31</td>
<td>Q76G11 (mature)</td>
</tr>
<tr>
<td>GH5_49</td>
<td>O74799 (full length)</td>
</tr>
<tr>
<td>GH6</td>
<td>1DYS</td>
</tr>
<tr>
<td>GH7</td>
<td>1GPI</td>
</tr>
<tr>
<td>GH10</td>
<td>1K6A</td>
</tr>
<tr>
<td>GH11</td>
<td>2QZ2</td>
</tr>
<tr>
<td>GH43</td>
<td>Q9HFS9 (mature)</td>
</tr>
<tr>
<td>GH61</td>
<td>4B5Q</td>
</tr>
</tbody>
</table>

**Table 1.** PSI-BLAST seed sequences for catalytic domain identification. PDB entries are listed with their 4-letter PDB ID and the remaining IDs correspond to Uniprot entries.
Table 2. Catalytic domain sequences used for phylogenetics. Sequences obtained from CAZy and sequences characterized in this work were clustered using UCLUST.

<table>
<thead>
<tr>
<th>Family</th>
<th>CAZy</th>
<th>Characterized</th>
<th>Total seq.</th>
<th>Unique seq. (95% id.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH5</td>
<td>2849</td>
<td>141</td>
<td>2990</td>
<td>2095</td>
</tr>
<tr>
<td>GH6</td>
<td>452</td>
<td>99</td>
<td>551</td>
<td>364</td>
</tr>
<tr>
<td>GH7</td>
<td>297</td>
<td>108</td>
<td>405</td>
<td>287</td>
</tr>
<tr>
<td>GH10</td>
<td>1271</td>
<td>87</td>
<td>1358</td>
<td>1071</td>
</tr>
<tr>
<td>GH11</td>
<td>599</td>
<td>61</td>
<td>660</td>
<td>385</td>
</tr>
<tr>
<td>GH43</td>
<td>2865</td>
<td>45</td>
<td>2910</td>
<td>1837</td>
</tr>
<tr>
<td>AA9</td>
<td>237</td>
<td>72</td>
<td>309</td>
<td>289</td>
</tr>
<tr>
<td>Description</td>
<td>Number of features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------</td>
<td>--------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amino acid frequency</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative solvent accessibility (3 classes and 20 AA)</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary structure (3 classes and 20 AA)</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial interactions</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of residues and atoms</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>160</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** Summary of features used for predictive modeling of wild-type enzyme thermostability.
<table>
<thead>
<tr>
<th>Partition</th>
<th>GH5</th>
<th>GH6</th>
<th>GH7</th>
<th>GH10</th>
<th>GH11</th>
<th>GH43</th>
<th>AA9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>27</td>
<td>32</td>
<td>30</td>
<td>17</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>48</td>
<td>27</td>
<td>18</td>
<td>12</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>14</td>
<td>28</td>
<td>20</td>
<td>13</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>10</td>
<td>21</td>
<td>19</td>
<td>19</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>130</td>
<td>99</td>
<td>108</td>
<td>87</td>
<td>61</td>
<td>45</td>
<td>72</td>
</tr>
</tbody>
</table>

**Table 4.** Sequence partitions used for the cross-validated predictive modeling of enzyme melting temperature. Sequences were partitioned such that two sequences from different partitions share a maximum of 80% sequence identity on a per family basis. Within the same partition, two sequences can share more than 80% sequence identity. For GH5, only 130 of 141 characterized enzymes were successfully structure homology modeled.