Predicting the thermostability of wild-type glycoside hydrolases

Henrik M. Geertz-Hansen[1,2,3*], Thomas Nordahl Petersen[1], Morten Nielsen[1,4], Nikolaj Blom[1], Jesper Salomon[3] and Lars Kiemer[3]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark,

[2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Fremtidsvej 3, DK-2970 Hørsholm, Denmark,

[3]Novozymes A/S, Krogshøjvej 36, DK-2880 Bagsværd, Denmark,

[4]Instituto de Investigaciones Biotecnologicas, Universidad Nacional de San Martin, San Martin, B 1650 HMP, Buenos Aires, Argentina

[*]**Correspondance: hmgh@biosustain.dtu.dk**

Objective

Enable prediction of enzyme melting temperature from amino acid sequence.

Based on a large dataset of melting temperatures (Tm) of fungal glycoside hydrolase (GH) enzymes, determined under identical conditions, we have developed Tm prediction methods for 7 GH families. As an example of its application, the prediction method was used to analyze the stability of GH enzymes found in 265 genomes obtained from the 1000 Fungal Genome Project at JGI (http://jgi.doe.gov/fungi).

Melting temperature of 434 glycoside hydrolase enzymes

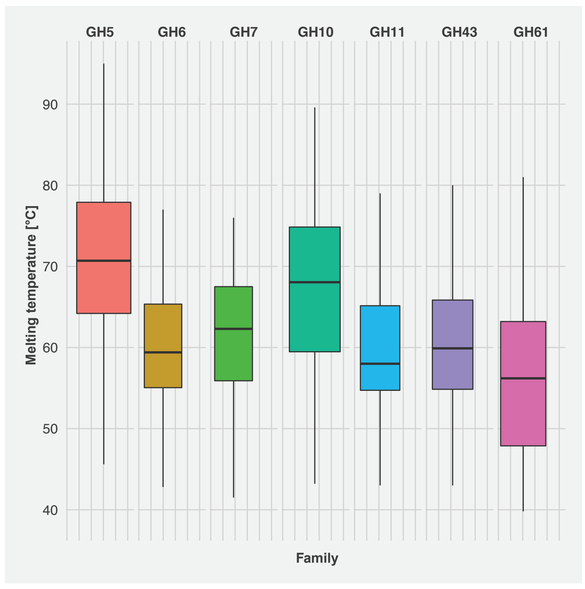


Figure 1: Five-number summaries of melting temperatures distributed by glycoside hydrolase family. The reported T_m values were all determined under identical experimental conditions.

Methods

Data set: Melting temperatures of 434 wild-type glycoside hydrolase enzymes of fungal origin were provided by Novozymes A/S. All enzymes were individually characterized under identical experimental conditions using a thermal shift assay at pH 5.

Molecular features: Homology models of all sequences were obtained using the CPHmodels 3.2 prediction server[2]. The following features were calculated from the sequence or structure: amino acid frequencies, secondary structure propensities (helix, strand and coil), relative solvent accessibility propensities (buried, intermediate and exposed) and spatial interactions (hydrophobic interactions, salt bridges, main-main chain and main-side chain hydrogen bonds, disulphide bridges and aromatic interactions).

Machine learning: Sequences were homology partitioned into 4 sets sharing a maximum of 80% sequence identity between sets. Artificial neural networks were trained for two rounds of feature selection on minimizing the error using 4-fold nested cross-validation (Figure 2). Final prediction performance was thus calculated from an independent evaluation set (Table 1). The ThermoP method is publicly available at http://www.cbs.dtu.dk/services/ThermoP (manuscript in preparation).

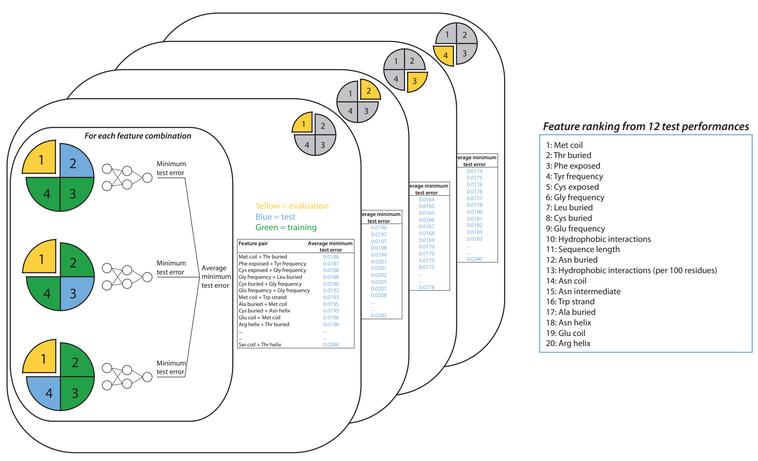


Figure 2: Illustration of the artificial neural network training procedure. Through two rounds of feature selection a combination of 7 features resulting in the best test set prediction performance were selected. This combination was used to obtain an independent evaluation set performance.

Results

Prediction performance: The prediction performance is summarized in Table 1 in terms of the Pearson's correlation coefficients (PCC) and mean absolute prediction error (MAE) for selected families. Furthermore, a benchmark against a BLAST-based prediction model is shown, in which the Tm of the nearest neighbor is transferred to the query sequence.

	Neural network (ANN)		BLAST	
Family	PCC	MAE [°C]	PCC	MAE [°C]
GH5	0.61	6.6	0.44	8.0
GH6	0.65	4.6	0.27	6.0
GH7	0.59	5.7	0.54	6.1

Table 1: Summary of melting temperature prediction model performance and a comparison against a BLAST-based model. The neural network model outperforms the sequence similarity-based model for all families.

Application on 265 fungal genomes: A translated gene catalogue of 13.4 mio. predicted genes from 265 genomes obtained from the 1000 Fungal Genome Project at JGI were searched for GH enzymes. From the seven families GH5, 6, 7, 10, 11, 43 and 61 a total of 11.716 enzymes were identified from the catalogue (Figure 3).

Identified GH enzymes in 265 fungal genomes

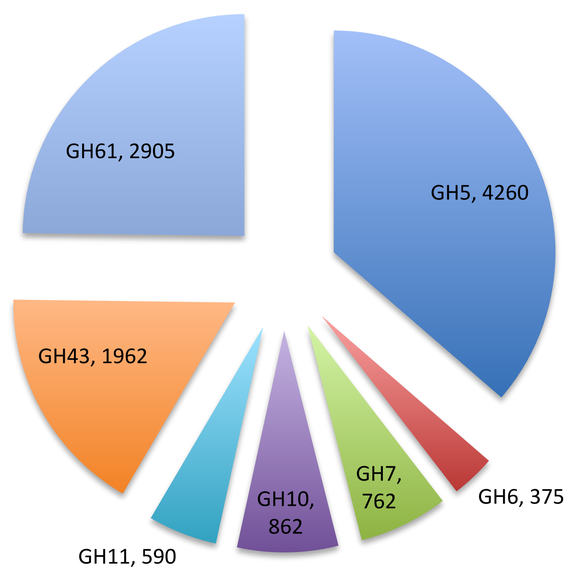


Figure 3: Identified fungal glycoside hydrolase enzymes in 265 genomes from the 1000 Fungal Genome project at JGI, distributed across family.

Of the 11.716 identified GH enzymes, the melting temperature could be predicted for 10.895 sequences, using the ThermoP webserver (shown below). A homology model could not be obtained for the remaining 821 sequences.

Predicted melting temperature of 10.895 WT fungal GH enzymes

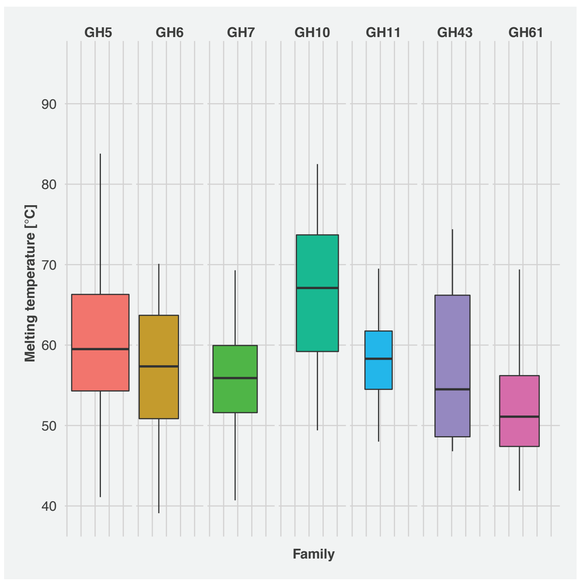


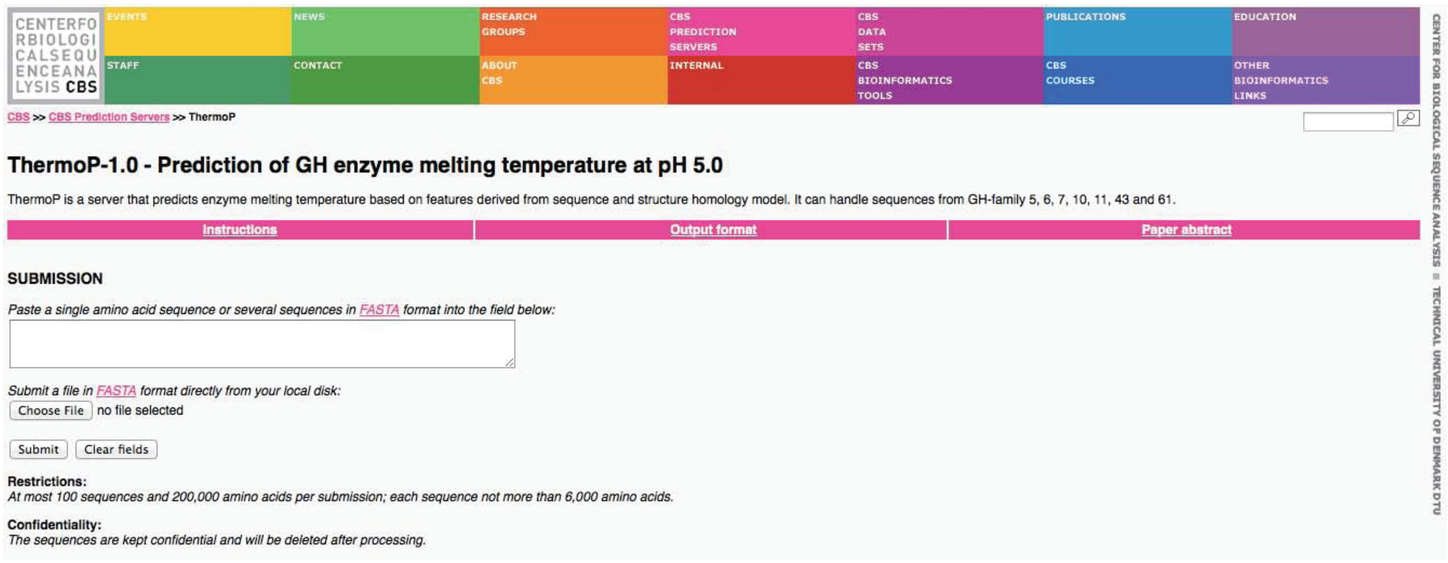
Figure 4: Five-number summaries of predicted melting temperatures distributed by family. The enzymes were discovered within 265 fungal genomes obtained from the 1000 Fungal Genome project at JGI. The predictions were completed within 24h.

Perspectives

- **ThermoP: fast primary screening tool for thermostable, fungal GH enzymes**
- **Extendable to bacterial enzymes as characterization data become available**
- **Applications in experimental design and large scale gene selction for industrial applications**
- **Approach to predictive model development could be used for other enzyme characteristics such as pH stability**

Acknowledgements

This study was co-financed by Novozymes A/S and The Novo Nordisk Foundation Center for Biosustainability.



Technical University of Denmark

novozymes® Rethink Tomorrow