Postprint: Protein Denaturation Temperature Prediction Based on Multilayer Perceptron

Authors: Ding Xuesong, Huang Liqun, Zhang Buzhong, Yang Yang, Lü Qiang

Date: 2018-05-24T00:00:00+00:00

Abstract

Accurately predicting protein denaturation temperature is of significant importance in fields such as protein engineering and drug development. Using global features and sequence features as the initial feature vector, dimensionality reduction is performed on the initial feature vector through the proposed weight-based dimensionality reduction algorithm, and the reduced features are fed into a multilayer perceptron model to predict protein denaturation temperature. On a blind test dataset, the PCC value between the predicted results of this method and the experimentally measured results increased from 0.77 before dimensionality reduction to 0.8, the RMSE value decreased from 0.17 before dimensionality reduction to 0.16, and the classification accuracy of the predicted protein denaturation temperature values showed a significant improvement compared with existing methods.

Full Text

Predicting Protein Melting Temperature Using Multi-layer Perceptron

Ding Xuesong[1], Huang Liqun[1], Zhang Buzhong[1], Yang Yang[1,2]†, Lyu Qiang[1,2] [1]School of Computer Science & Technology, [2]Jiangsu Provincial Key Laboratory for Information Processing Technologies, Soochow University, Suzhou, Jiangsu 215006, China

Abstract: Accurate prediction of protein melting temperature is of great significance in protein engineering and drug design. This study employs global and sequential features as the initial feature vector and applies a novel weight-based dimensionality reduction algorithm to reduce the feature space. The reduced features are then fed into a multi-layer perceptron (MLP) model to predict protein melting temperature. On a blind test dataset, the Pearson correlation coefficient (PCC) between predicted and experimentally determined melting temperatures

increased from 0.77 to 0.8, while the root mean square error (RMSE) decreased from 0.17 to 0.16. The classification accuracy of predicted melting temperatures showed significant improvement compared to existing methods.

Keywords: protein melting temperature; multi-layer perceptron; regression prediction

0 Introduction

Protein stability refers to a protein' s ability to resist thermal denaturation in high-temperature environments and represents an intrinsic property for maintaining optimal activity. Protein melting temperature serves as a crucial metric for determining whether protein function is lost and constitutes one measure of protein kinetic stability. Consequently, predicting protein melting temperature holds vital importance in both scientific research and pharmaceutical applications. Currently, protein melting temperature is primarily determined through experimental methods such as differential scanning calorimetry, circular dichroism, and Fourier transform infrared spectroscopy. However, these experimental approaches suffer from high costs, complex procedures, and long cycles.

In recent years, computational methods based on mathematical statistics and machine learning have gained widespread application for predicting protein melting temperature. Ku et al. [3] employed a statistical estimation method to establish correlations between dipeptide content and protein melting temperature, but this approach could only estimate temperature ranges rather than predict specific values. Pucci et al. [1] predicted stability curves of homologous proteins using temperature-dependent statistical potentials, a process requiring numerous experimentally determined protein properties such as flexibility [4], hydrophilicity [5], and hydrogen bonding [6], making the prediction procedure rather complex. Gorania et al. [2] constructed artificial neural network and adaptive neuro-fuzzy inference system models based on sequence information to predict melting temperature by analyzing the complex nonlinear relationship between amino acid sequences and protein melting temperature. However, their dataset was too small to fully capture the associations between protein characteristics and melting temperature.

Deep learning [7] has demonstrated outstanding performance in speech recognition [8], machine translation [9], and other domains, attracting increasing attention and adoption. This paper proposes a prediction method based on multilayer perceptron (MLP) models for protein melting temperature. By combining global and sequential features as the initial feature vector and applying a weight-based method for dimensionality reduction, our approach achieved an RMSE of 0.16 and PCC of 0.8 on the test dataset, demonstrating superior performance compared to experimental results reported in literature [3].

1 Methods

1.1 Dataset

Our dataset, sourced from Leuenberger et al. [10], contains 3,520 protein sequences with corresponding global melting temperatures from four organisms: *E. coli* (729 entries), *S. cerevisiae* (709 entries), *Thermus thermophilus* (1,073 entries), and human cervical cancer cells (1,009 entries). We first extracted 300 proteins as a test set according to their distribution ratios: 60 from *E. coli*, 60 from *S. cerevisiae*, 90 from *Thermus thermophilus*, and 90 from human cervical cancer cells. The remaining 3,220 proteins were used as the training set. Table 1 details the specific numbers for the training and test sets.

1.2 Feature Engineering

1.2.1 Global Features For each protein, we extracted 1,644-dimensional global features comprising: 1,437 structural and physicochemical features calculated from amino acid and protein sequences using ProFEAT [11]; 140 electronic features derived from protein charge density based on atomic information using Protein_recon [12]; 19 functional and three-dimensional features based on amino acid sequences using ProtDCal [13]; and 48 features including protein length, relative molecular mass, and the count and proportion of each amino acid obtained from ExPASy [14].

1.2.2 Sequential Features Sequential features consist of two components: amino acid classification and dipeptide bond information. Based on physicochemical properties, all amino acids in the dataset were categorized into six classes [15]: hydrophobic (V, I, L, F, M, W, Y, C), negatively charged (D, E), positively charged (R, K, H), conformationally special (G, P), polar (N, Q, S), and others (A, T). We used the count and proportion of these six amino acid classes in each protein as 12 features. Unknown amino acids beyond the 20 standard types were designated as X. Finally, we constructed 882 features by counting the number and proportion of dipeptide bonds based on the 20 amino acids plus X.

1.2.3 Initial Feature Vector Our method concatenates the global and sequential features to form an initial 2,538-dimensional feature vector. All features and labels were normalized.

1.2.4 Weight-Based Dimensionality Reduction Algorithm High dimensionality in machine learning leads to high time and space complexity, and excessive features may introduce noise or redundancy, thereby reducing accuracy. Moreover, extracting certain biological features consumes substantial resources, making dimensionality reduction necessary to filter noise, improve generalization, enhance precision, and reduce feature count.

We construct an MLP model for fitting and output the weight relationships between each input node and the subsequent hidden layer. We consider that weights—whether positive or negative—play an active role in the neural network if their absolute values are large (large absolute negative values indicate strong inhibition). Therefore, for each node' s weight set, we take the absolute value of each element and apply a threshold (0.0285). If the count of values exceeding the threshold is more than half (i.e., greater than 10, given the first hidden layer has 20 nodes), we retain the input feature as important for melting temperature; otherwise, we eliminate it. Applying Algorithm 1 to the initial feature vector yields a 541-dimensional feature vector. We then build a new MLP regression model using the selected features for retraining.

Algorithm 1 Weight-Based Feature Dimensionality Reduction
Step 1: Build MLP model
Step 2: Input 2,538 features to fit protein melting temperature
Step 3: Output the weight matrix
For i = 1 to 3,220 (number of proteins):
  count = 0
  For j = 1 to 2,538 (number of protein features):
    If abs(weight matrix) > 0.0285:
      count += 1
    If count > 10:
      Save the feature index
Step 4: From Step 3, obtain feature index matrix and select new features for each protein accordingly. Build new MLP model. Input the new protein features to the new MLP model to fit protein melting temperature.

1.2.5 Evaluation Metrics Root Mean Square Error (RMSE) measures the square root of the average squared differences between observed and true values. For the i-th protein, $X_{\text{predicted},i}$ represents the model' s predicted value while $X_{\text{observed},i}$ represents the experimental value, with n being the total number of proteins:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{\text{predicted},i} - X_{\text{observed},i})^2}$$

Pearson Correlation Coefficient (PCC) is a linear correlation measure describing the strength of linear association between two variables, with larger absolute values indicating stronger correlation. Here, N represents the sample size, while X and Y denote observed and true melting temperatures, respectively:

$$PCC = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

1.3 Model Architecture and Training

We constructed an MLP regression model for protein melting temperature using the sklearn platform, as illustrated in Figure 1 [Figure 1: see original paper]. The model comprises three hidden layers with 20 nodes each, using the ReLU activation function. The multi-layer perceptron employs multiple hidden layers suitable for nonlinear function fitting and utilizes the backpropagation algorithm [18] with gradient descent to minimize the loss function:

$$C(w,b) = \frac{1}{2n}\sum_{x}\|y(X) - a\|^2$$

The optimization objective determines weights (w) and biases (b) that minimize the loss function, making network outputs increasingly approach true values. The weight and bias update formulas are:

$$w_k = w_k - \eta\frac{\partial C}{\partial w_k}, \quad b_k = b_k - \eta\frac{\partial C}{\partial b_k}$$

where represents the learning rate.

2 Results

2.1 Performance Analysis

Using the initial 2,538-dimensional feature vector (Section 1.2.3) as input to the MLP model (Section 1.3) yielded test set results of PCC = 0.772347 and RMSE = 0.1874. After dimensionality reduction to 541 features, the retrained MLP model achieved PCC = 0.80559 and RMSE = 0.1638 on the test set. Table 2 compares the performance before and after dimensionality reduction.

Table 3 presents a comparison between predicted and experimental melting temperatures for selected proteins from our test set. Figure 2 [Figure 2: see original paper] shows scatter plots of predicted versus experimental values: the left panel displays results using the original 2,538 features, while the right panel shows results after reduction to 541 features. The latter demonstrates better clustering around the $y = x$ line, indicating improved prediction accuracy.

2.2 Comparative Evaluation

Among existing computational methods for predicting melting temperature, only Ku et al. [3] provide a web service (http://tm.life.nthu.edu.tw/). We submitted our test set from Section 1.1 to this service. Since their method only provides categorical predictions (>65°C, 55–65°C, <55°C), we compared classification accuracy using the experimental melting temperature categories as the gold standard. Table 4 compares the classification accuracy between our method (Section 1.2.4) and Ku' s method.

3 Conclusion

This study develops a predictive model for protein melting temperature to provide auxiliary evidence for bioengineering, thereby reducing the time and economic costs of biological experiments. Our MLP-based approach employs an initial 2,538-dimensional feature vector, which is reduced to 541 features through weight-based selection, achieving enhanced predictive performance. Comparative experiments demonstrate that our model not only predicts specific protein melting temperature values but also outperforms reported methods in categorical prediction. The primary challenge lies in identifying more representative feature attributes, and future work will focus on mining such properties.

References

[1] Pucci F, Rooman M. Stability curve prediction of homologous proteins using temperature-dependent statistical potentials [J]. *PLoS Computational Biology*, 2014, 10(7): e1003689.

[2] Gorania M, Seker H, Haris P I, et al. Predicting a protein' s melting temperature from its amino acid sequence [C]// *Proc of Annual International Conference of Engineering in Medicine and Biology Society*. 2010: 1820-1823.

[3] Ku Tienhsiung, Lu Peiyu, Chan Chenhsiung, et al. Predicting melting temperature directly from protein sequences [J]. *Computational Biology & Chemistry*, 2009, 33(6): 445-450.

[4] Vihinen M. Relationship of protein flexibility to thermostability [J]. *Protein Engineering, Design & Selection*, 1987, 1(6): 477-480.

[5] Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions [J]. *Proteins*, 1994, 19(2): 141-149.

[6] Prevost M, Wodak S J, Tidor B, et al. Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96→Ala mutation in barnase [J]. *Proceedings of the National Academy of Sciences of the USA*, 1991, 88(23): 10880-10884.

[7] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. *Science*, 2006, 313(5786): 504-507.

[8] Graves A, Mohamed A, Hinton G E. Speech recognition with deep recurrent neural networks [C]// *Proc of IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013: 6645-6649.

[9] Sutskever I, Vinyals O, Le Quoc V. Sequence to sequence learning with neural networks [J/OL]. (2014-10-14). https://arxiv.org/pdf/1409.3215.

[10] Leuenberger P, Ganscha S, Kahraman A, et al. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability [J]. *Science*, 2017, 355(6327): eaai7825.

[11] Zhang Peng, Tao L, Zeng Xian, et al. PROFEAT update: a protein features web server with added facility to compute network descriptors for studying Omics-derived networks [J]. *Journal of Molecular Biology*, 2016, 429(3): 534-535.

[12] Zaretzki J, Bergeron C, Rydberg P, et al. RS-Predictor: a new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4 [J]. *Journal of Chemical Information & Modeling*, 2011, 51(7): 1667-1675.

[13] Ruiz-Blanco Y B, Paz W, Green J, et al. ProtDCal: a program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins [J]. *BMC Bioinformatics*, 2015, 16(1): 162.

[14] Gasteiger E, Hoogland C, Gattiker A, et al. Protein identification and analysis tools on the ExPASy server [M]// *The Proteomics Protocols Handbook*. [S.l.]: Humana Press, 2005: 531.

[15] Yang Yang, Abhishek N, Shen Bairong, et al. PON-Sol: prediction of effects of amino acid substitutions on protein solubility [J]. *Bioinformatics*, 2016, 32(13): 2032-2034.

[16] Wang Juan, Wu Xianxiang, Cao Yanling. Multi-layer perceptron using hybrid differential evolution and biogeography-based optimization [J]. *Application Research of Computers*, 2017, 34(3): 693-696.

[17] Kruse R, Borgelt C, Klawonn F, et al. Multi-layer perceptrons [M]// *Computational Intelligence*. London: Springer, 2013: 47-81.

[18] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [J]. *Journal of Machine Learning Research*, 2010, 9: 249-256.

Note: Figure translations are in progress. See original paper for figures.