CS221 Final Project: Proposal

Ruben Krueger, Marco Mora, Josh Wolff

May 3, 2019

Introduction

Proteins are one of the four fundamental classes of macromolecules, with a variety of biological functions. These functions arise from the structure of the protein, and the structure of proteins are based on the protein's sequence of amino acids, which are the building blocks of proteins.

A protein is functional when it is "folded" and is non-functional when "unfolded." Proteins can become unfolded when temperature rises, and protein thermostability is based on composition and sequence (Ku et. al, 2009). Thus, deriving the melting temperature of proteins is of interest, particularly in drug development (Gorania et. al 2010). Current techniques, including differential scanning calorimetry, circular dichroism, and Fourier transform infrared spectroscopy, are expensive (Gorania et. al 2010). An easy, computerized way to estimate this characteristic would be useful.

Task

Given a protein's amino acid sequence, predict the T_m, the temperature at which 50% of the protein is unfolded.

Dataset

There is not one centralized dataset for protein sequences and melting temperature. Our dataset will consist of data compiled from multiple sources, including but not limited to the following journal articles:

- Table 1 of "Some thermodynamic implications for the thermostability of proteins" (Rees, et. al.)
- Table 1 of "Predicting melting temperature directly from protein sequences" (Ku, et. al.)
- Supplementary Data of "Towards an accurate prediction of the thermal stability of homologous proteins" (Pucci, et. al.)

Input and Output

Example: Staphylococcal nuclease
Staphylococcal nuclease is a protein found in Staphylococcus aureus.
Input: Amino Acid Sequence
MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSCLVLTLV
VVSSLSSSANASQTDNGVNRSGSEDPTVYSATSTKKLHKEPATLIKAIDGDT
VKLMYKGQPMTFRLLLVDTPETKHPKKGVEKYGPEASAFTKKMVENAKKIEV
EFDKGQRTDKYGRGLAYIYADGKMVNEALVRQGLAKVAYVYKPNNTHEQHLR
KSEAQAKKEKLNIWSEDNADSGQ
Output: Melting Temperature
$T_m = 327^\circ K$

Baseline

We implemented three baselines. The first baseline simply considered the length of the amino acid. Using this as the only feature, we achieved $r^2 = -0.091$. We then tested a simple baseline that considered the counts of polar, nonpolar, and "special," amino acids, and we reported $r^2 = 0.005$. Finally, we had the most success by considering groupings of amino acids. First, we considered the frequency of individual letters in amino acids (e.g. 'Q') and observed $r^2 = -0.272$. Then, we considered the frequency of groupings of two (e.g. 'QR') and reported $r^2 = 0.055$. Finally, groupings of three resulted in $r^2 = 0.017$. Our data consisted of 36 proteins.

Oracle

There is no oracle for this task because it is not solved. This problem is intractable for humans: someone cannot predict the protein's melting temperature given the amino acid sequence for a protein. Current methods to determine melting temperatures are labratory methods and thus no easy "shortcut" exists.

We did, however, attempt to create an oracle. Because all of the proteins were derived from particular organisms, our oracle considered a prediction function that returned twice the organism's typical body temperature or optimal growth temperature for the T_m of the protein of interest. The hypothesis behind this attempt was that proteins evolve with the organism, and they should function best at the organism's preferred temperature. However, our r^2 value was 0.008, and thus, this did not work.

Evaluation

Evaluation was performed using r^2 error and we will continue to report using this metric.

Related Work and Going Forward

There have been a variety of techniques to predict thermostability based on amino acid sequence. Gorania et. al (2010) used artificial neural networks (ANN) and an adaptive network-fuzzy inference system (ANFIS). Wu et. al (2009) used a decision tree. We think exploring some type of search algorithm or neural network might be helpful. We will explore using q-learning to estimate the rewards (in terms of temperature) and transitions probabilities of each "choice" of amino acid. We can then use this model to estimate the total reward of a given path.

Conclusion and Challenges

While we found a clear need for the development of an algorithm that could accurately predict the melting temperature of a given protein, previous attempts from various research groups to predict this data have yielded limited results. For our CS221 project, we hope to leverage a combination of different AI methods to improve the accuracy of these predictions. While there isn't necessarily a concrete oracle to strive for, we aim our project to be as accurate as possible so that researchers are able to predict this information and use it to rapidly iterate on protein design.

The challenges largely consider the complexity of the problem. There are several variables at play, such as the fact that T_m is also a function of the solution's pH and salinity at which the T_m was measured. We hope to capture as many of these features as possible, but these are outside of the scope of sequence. Doing so would require us to expand our input, or at least make some assumptions, such as a pH of 7.0 for example. The interplay of the environment, complexity, and physics with our algorithm is likely to be the greatest challenge.

Works Cited

1. Ku, Tienhsiung, Peiyu Lu, Chenhsiung Chan, Tsusheng Wang, Szuming Lai, Pingchiang Lyu, and Naiwan Hsiao. "Predicting Melting Temperature Directly from Protein Sequences." Computational Biology and Chemistry 33, no. 6 (2009): 445-50. doi:10.1016/j.compbiolchem.2009.10.002.

2. Gorania, M., H. Seker, and P. I. Haris. "Predicting a Proteins Melting Temperature from Its Amino Acid Sequence." 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010. doi:10.1109/iembs.2010.5626421.

3. De novo structure prediction with deep-learning based scoring R.Evans, J.Jumper, J.Kirkpatrick, L.Sifre, T.F.G.Green, C.Qin, A.Zidek, A.Nelson, A.Bridgland, H.Penedones, S.Petersen, K.Simonyan, S.Crossan, D.T.Jones, D.Silver, K.Kavukcuoglu, D.Hassabis, A.W.Senior In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4 December 2018.

4. Wu, Li-Cheng, Jian-Xin, Lee, Hsien-Da, Huang, Baw-Juine, Liu, and Jorng-Tzong Horng. "An Expert System to Predict Protein Thermostability Using Decision Tree." Expert Systems with Applications 36, no. 5 (2009): 9007-014. doi:10.1016/j.eswa.2008.12.020.

5. Rees, D C, and A D Robertson. "Some thermodynamic implications for the thermostability of proteins." Protein science : a publication of the Protein Society vol. 10,6 (2001): 1187-94. doi:10.1110/ps.180101

6. Pucci, Fabrizio, and Marianne Rooman. "Towards an accurate prediction of the thermal stability of homologous proteins." Journal of Biomolecular Structure and Dynamics. vol 33,5. (2015): 1132-1142. doi:10.1080/07391102.2015.1073631