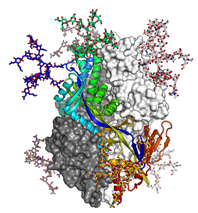


Predicting Protein Thermostability Based on Amino Acid Sequence

Marco Mora-Mendoza, Ruben Krueger, Josh Wolff

Task Definition

Given a protein's amino acid sequence, what is the protein's melting temperature?

Input: "TKLQQAAAKK" (String of Amino acids)
Output: 350 K (Temperature)

Background

Proteins are macromolecules with a variety of biological functions. These functions arise from their structure, which is based on the protein's sequence of amino acids. Proteins can become non-functional with temperature changes, as the structure changes. Thus, deriving the melting temperature of proteins is of great interest, particularly in drug development (Gorania et. al 2010). Current techniques, including differential scanning calorimetry, are expensive (Gorania et. al 2010).

Data

We collected protein sequences and melting temperature data from previous academic papers. We gathered a total of 245 proteins with sequence and melting temperature data. We split the data into test and training sets.

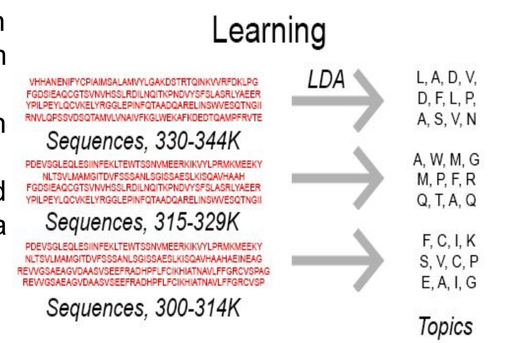
Features

Feature	Amino acid counts	Hydrophobic amino acids	N-grams	Hydrophilic amino acids	(E + K)/(Q + H)
Description	Counts of individual amino acids	Number of hydrophobic acids	Frequency counts of n-groups of amino acids	Number of hydrophobic acids	Ratio of glutamic acid and lysine to glutamine and histidine
Example	Lysine: 1, Glycine: 3	23	AA	47	0.67

Models

Latent Dirichlet Allocation Model (LDA)

- A latent dirichlet allocation model extracts topics from text.
- We use this to get hidden topics from proteins.
- We classify proteins, based on sequence input, into a melting temperature range.
- Library used: Gensim

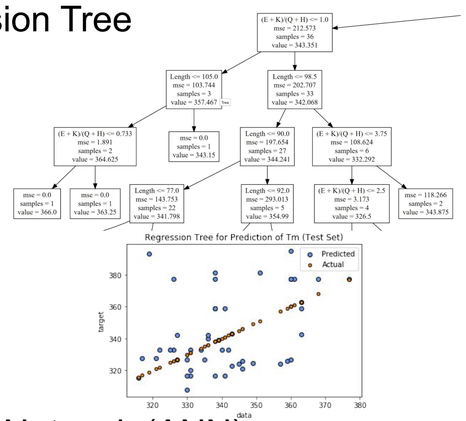


Truth

Predicted	315-329	330-344	345-359	360-374
315-329	9	15	9	3
330-344	9	12	2	1
345-359	1	4	3	0
360-374	1	5	3	0
None	13	22	11	7

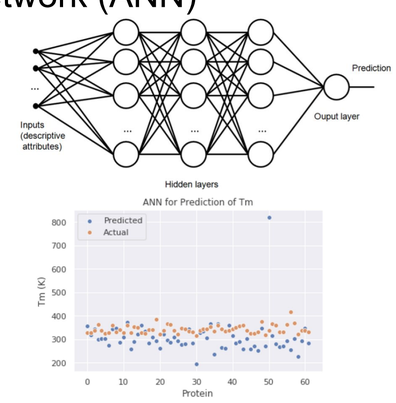
Regression Tree

- The input sequence is passed into an extractor function.
- Each level of the tree divides the data according to a specific feature.
- The order of the features depends on the minimization of an error function (mean squared error).



Artificial Neural Network (ANN)

- Input from one layer of nodes is input to the next layer
- Each node maintains own weight vector
- ReLu activation function determines the output from one node
- Sequential model from Keras
- Trained over 100 epochs
- Adam optimizer



Results

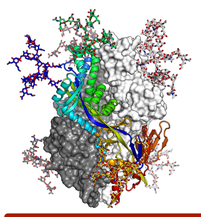
- The LDA properly classified 31.9% of the proteins for which it gave an output. This does not include "No output."
- Over the testing set, the ANN had an average percentage error of 16.9%
- The regressive tree correctly had an average percentage error of 10.2%

Conclusions

- The models would benefit from more data.
- Our models are not pH-dependent, and this creates some error in our data as the melting temperature varies with pH.
- More developments in determining structure from sequence will help with determining the melting temperature computationally as well.
- Protein structure data, as inferred from sequence, would improve model accuracy.

References

1. 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.
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. 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.
4. https://www.researchgate.net/figure/Neural-network-with-three-hidden-layers-and-one-output-neuron-for-price-prediction_fig4_329663610



Predicting Protein Melting Temperature Based on Amino Acid Sequence

Marco Mora-Mendoza, Ruben Krueger, Josh Wolff

Task Definition

Given a protein's amino acid sequence, what is the protein's melting temperature?
Input: TKLQQAAAKKK (Amino acids)
Output: 350 K (Temperature)

Background

Proteins are macromolecules with a variety of biological functions. These functions arise from the structure of the protein, which is based on the protein's sequence of amino acids. Proteins can become non-functional with temperature changes, as the structure changes. Thus, deriving the melting temperature of proteins is of great interest, particularly in drug development (Gorania et. al 2010). Current techniques, including differential scanning calorimetry, are expensive (Gorania et. al 2010)

Data

While databases exist (e.g., PDB) on a variety of protein information, there is not one centralized database of protein sequences and melting temperature. Thus, we collected protein sequences and melting temperature data from academic papers on the topic. We gathered a total of 245 proteins with sequence and melting temperature data. We split this into test and training sets.

Models

Latent Dirichlet Allocation (LDA)

An LDA generates a distribution over topics from text. We generate temperature ranges (e.g. 300-314K) and amalgamate the protein sequences that have a melting temperature in this range into one text sequence. From this text, we generate a distribution over topics. When presented with an unknown protein, we generate a distribution over topics for this protein, and return the range that most closely corresponds to the protein we predict.

Artificial Neural Network (ANN)

- An ANN consistents of nodes
-

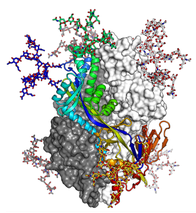
Our ANN contained three hidden layers

Regression Tree

Comparison

Conclusions

References



Predicting Protein Melting Temperature Based on Amino Acid Sequence

Marco Mora-Mendoza, Ruben Krueger, Josh Wolff

Task Definition

Given a protein's amino acid sequence, what is the protein's melting temperature?
Input: TKLQQAAAKKK (Amino acids)
Output: 350 K (Temperature)

Background

- Proteins are macromolecules with biological functions
- Functions arise from the structure of the protein
- Protein structure of amino acids

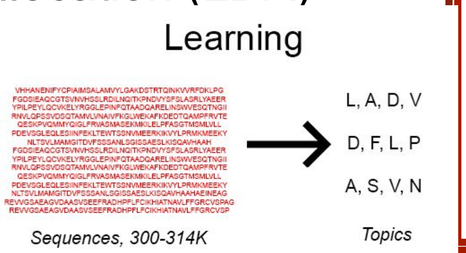
Data

While databases exist (e.g., PDB) on a variety of protein information, there is not one centralized database of protein sequences and melting temperature. Thus, we collected protein sequences and melting temperature data from academic papers on the topic. We gathered a total of 245 proteins with sequence and melting temperature data.

Models

Latent Dirichlet Allocation (LDA)

- Generates a distribution of topics from text (gensim)
- Our algorithm:
  - Amalgamate protein sequences into ranges (e.g. 300-314K)
  - Generate a distribution o topics for a range
  - With new protein, get topics and compare to learned topics. Return most probabilistic range.



Artificial Neural Network (ANN)

An ANN consistents of nodes

Our ANN contained three hidden layers

Regression Tree

Results

Error Analysis

Conclusions

References

1. Gorania, M., H. Seker, and P. I. Haris. "Predicting a Proteins Melting Temperature from Its Amino Acid Sequence." 2010