Using AI to Predict Protein Structural Stability

Mir Sultan: Bioinformatics & Computational Biology
Advisor: Professor Dmitry Korkin

ABSTRACT

Alternative splicing contributes to proteome diversity, but many classified alternatively spliced proteins lack proper verification. To verify the human proteome, a Positive-Unlabeled classification algorithm was implemented to predict protein existence with structural stability as the deciding factor. Using features of structural data, it was tested/trained on known stable and unknown proteins, and then predicted the stability of known stable and unstable proteins from genes CFTR and TP53. Improvements are necessary, but the testing/training and predictions gave reliable results.

PU LEARNING

Positive-Unlabeled learning/PU learning, is when a "learner" only has access to positive and unlabeled data where the unlabeled data may contain positive and negative samples [6]. It uses unlabeled data in the learning process and specializes in standard semi-supervised learning [6].

- $P(s=1|x)$
- $P(y=1|s=1)$
- $P(s=1|k)$
- $P(s=1|k)/P(s=1|y=1)$

BACKGROUND

- Alternative Splicing, known as AS, is the result of RNA splicing during different stages of development being regulated such that some splicing signals/sequences are ignored resulting in different mRNA molecules that encode related proteins with sequence and function overlap and distinction [1].

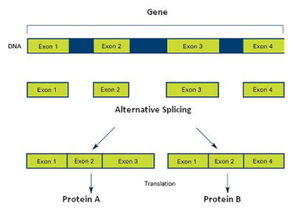


Figure 1. A visual of AS that shows how from the same gene, distinct exon mixes are made allowing one gene to code for many proteins [7].

- Protein Structural Stability is important because of protein degradation, which has 3 functions [2].
  1. Store nutrients for use in times of metabolic need [2].
  2. Eliminate abnormal proteins harmful to the cell [2].
  3. Regulation of cellular metabolism [2].

- Relevant Past Research examples would be
  - ProTstab predicts protein stability by predicting the melting temperature (T_M) of input proteins [3].
  - iStable 2.0 combines different algorithms together, uses PDB as a source for its stable data, and goes for ΔΔG [4].
  - SCooP predicts ΔG and T_M, uses PDB, but the input must have the 3D structure of the target protein [5].

METHODOLOGY

The project objective was to predict the true structural stability of a given protein isoform/variant with a machine learning algorithm

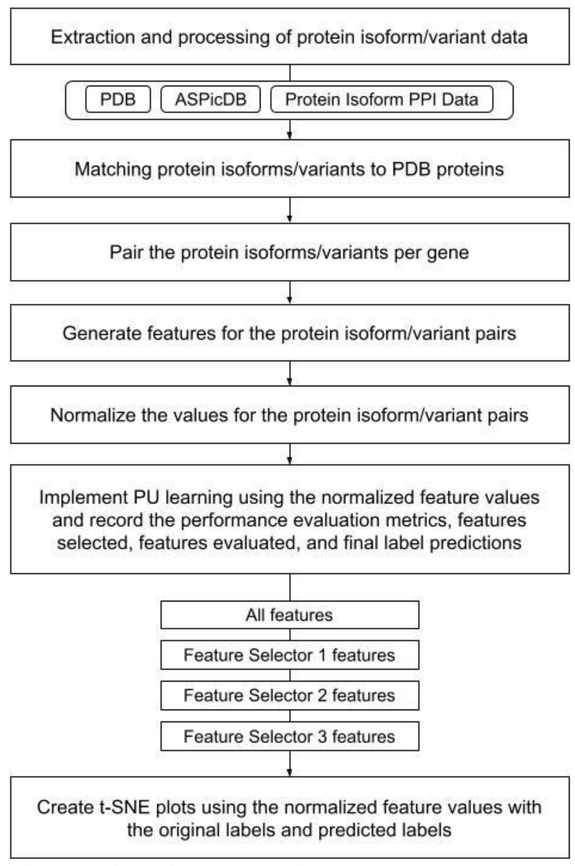
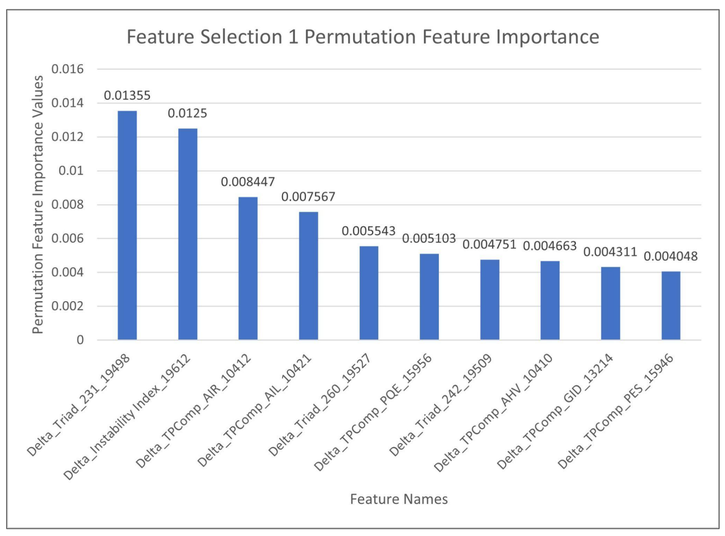
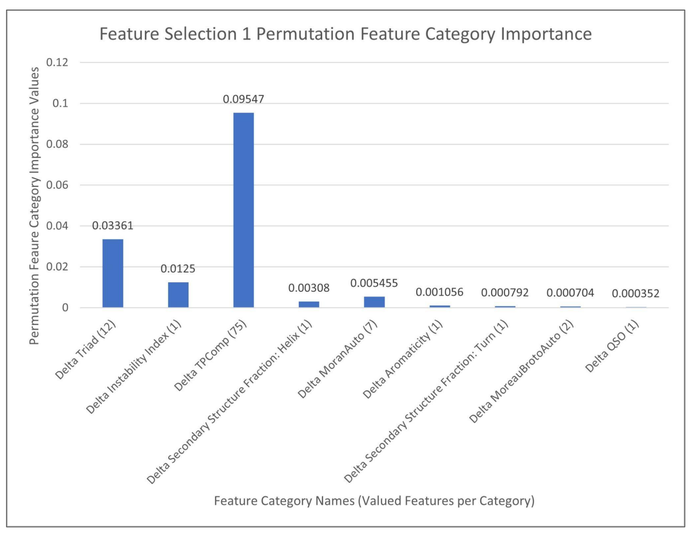


Figure 2. The Protein Data Processing Pipeline. A flowchart showing the chronological steps for processing the protein isoform data in this project.

RESULTS

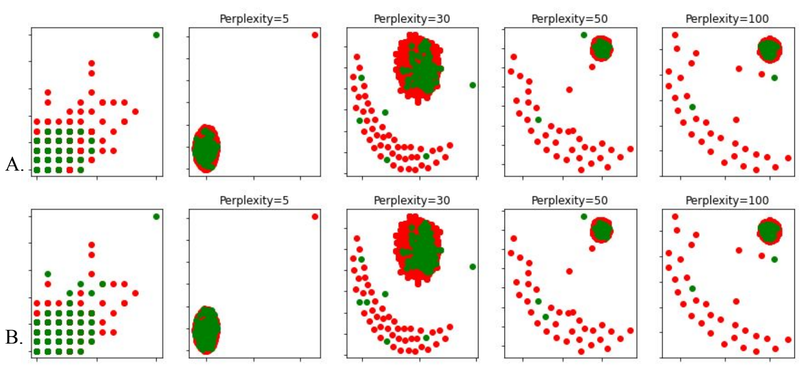


A.



B.

Figure 6. Permutation feature importances for data with features from first feature selector



A.

B.

Figure 5. t-SNE plot of data with features from the first feature selector. Panel A is the plot with the original labels. Panel B is the plot with the labels predicted by the PU learning model.

Table 6. Metric Estimations for Data with Features from First Feature Selector

	k-fold value Average	k-fold Standard Deviation	Final Value
Recall	0.9535	0.0140	1.0
BC Accuracy estimate	1.0	0	1.0
BC Balanced Accuracy	1.0	0	1.0
BC F score	1.0	0	1.0
BC Matthews Correlation Coefficient	1.0	0	1.0

Case Studies

Table 8. Evaluation Metrics for CFTR proteins

	Full Feature Selection	Feature Selection 1	Feature Selection 2	Feature Selection 3
F1 Score	66.67%	66.67%	66.67%	66.67%
ROC	75.00%	75.00%	75.00%	75.00%
Recall	50.00%	50.00%	50.00%	50.00%
Precision	100.00%	100.00%	100.00%	100.00%

- Feature selection had no effect on results of CFTR case study. All four selections incorrectly predicted a stable protein to be unstable

Figure 13. 3D structure of WT CFTR protein. The red section shows the deletion in F508del CFTR

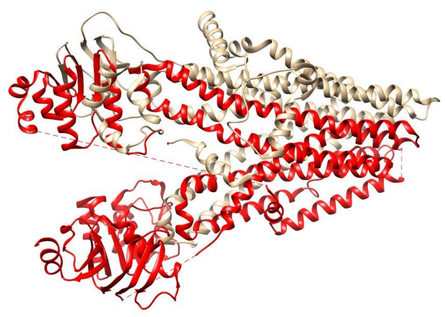


Figure 14. 3D Structure of WT CFTR protein. The red section shows the deletion in G542X

Table 9. Evaluation Metrics for TP53 proteins

	Full Feature Selection	Feature Selection 1	Feature Selection 2	Feature Selection 3
F1 Score	80.00%	66.67%	100.00%	80.00%
ROC	50.00%	75.00%	100.00%	50.00%
Recall	100.00%	50.00%	100.00%	100.00%
Precision	66.67%	100.00%	100.00%	66.67%

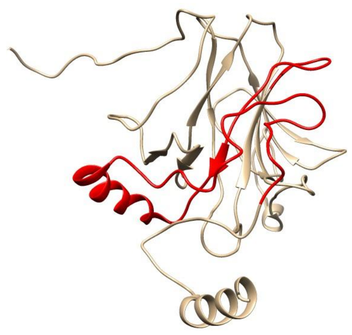


Figure 17. 3D Structure of p53. The red section shows the deletion in Δp53

- Feature Selection 2 had perfect results for the TP53 case study. This along with the differing predictions from the other feature selections proves that the feature selection matters.

DISCUSSION

- The key findings of this project was that despite a lack of negatively labeled data and the PU learning model constantly overvaluing features less indicative of structural stability, the model performed well.
- It was known and acknowledged that there would be a negative bias for the PU learning model due to how this implementation of the traditional PU learning approach treats unlabeled data as negative.
- It was discouraging to see features less relevant to structural stability valued higher than features that have great relevance to structural stability, but such irrelevant features do not consistently provide computationally useful values and the model recognized this.

CONCLUSION

Overall, this project showcased that despite some disadvantages, this PU learning model can be of applicable use in the real world and provide accurate predictions on the stability of protein isoforms/variants.

For future work, it is recommended to test/train this PU model using training/testing sets composed of data samples with known labels where the negative samples and a random subset of the positive samples are unlabeled before the actual testing/training [6]. Next, the selection of features should be managed more directly. And finally, physically verify more proteins as the data would help testing/training.

REFERENCES

[1] de Klerk E, & 't Hoen P. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends in Genetics, 31(3), 128–139. https://doi.org/10.1016/j.tig.2015.01.001
[2] Voet D, Voet J, Pratt C. (2016). Fundamentals of Biochemistry: Life at the Molecular Level (5th ed.). Wiley.
[3] Yang Y, Ding X, Zhu G, Niroula A, Lv Q, Vihinen M. (2019). ProTstab – predictor for cellular protein stability. BMC Genomics, 20(1). https://doi.org/10.1186/s12864-019-6138-7
[4] Chen C, Lin M, Liao C, Chang H, Chu, Y. (2020). iStable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules. Computational and Structural Biotechnology Journal, 18, 622–630. https://doi.org/10.1016/j.csbj.2020.02.021
[5] Pucci F, Kwasigroch J, Rooman M. (2017). SCooP: an accurate and fast predictor of protein stability curves as a function of temperature. Bioinformatics (Oxford, England), 33(21), 3415–3422. https://doi.org/10.1093/bioinformatics/btx417
[6] Bekker J, Davis J. (2020). Learning from positive and unlabeled data: a survey. Machine Learning, 109(4), 719–760. https://doi.org/10.1007/s10994-020-05877-5
[7] Alternative Splicing of Genes: Definition, Mechanism & Regulation. (2015). Retrieved from https://study.com/academy/lesson/alternative-splicing-of-genes-definition-mechanism-regulation.html.