Identifying patterns in predicted binding probabilities of different proteins with Xist lncRNA

Arsh Madhani, Megan Kratz, Keriayn Smith

Introduction

The functions of lncRNAs are mediated by intermolecular interactions. Detailed mapping of interactions through laboratory experimentation is limited by reagent availability and other resources. To overcome these limitations, machine learning approaches can be used to predict interactions between RNAs and the proteins that interact with them for regulation of the RNAs, and/or mediation of their functions.

Objectives

The aim of this project is to develop prototype analysis methods for identifying predicted RNA-protein binding. This includes developing a distance metric that describes how similarly two proteins bind to an RNA and developing a method for clustering proteins into groups based on how they bind to an RNA.

Methods

To determine the most optimal set of methods that would output the accurate protein pairs, seven filters and two correlation techniques were tested. Figure 1 shows a workflow diagram that describes the filers and method that the data was processed through to determine the best technique.

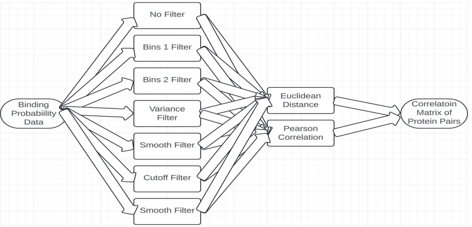
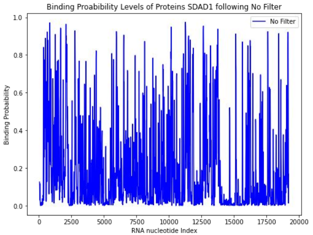
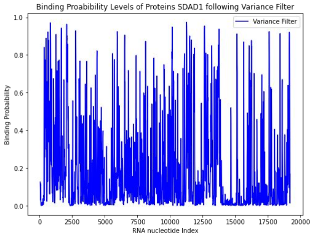


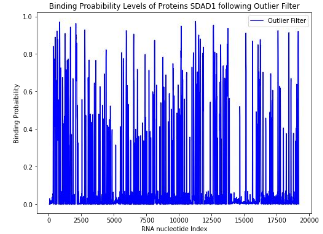
Figure 1: The first set of nodes represents the predicted binding probabilities of each protein with lncRNA Xist. The second set of nodes represent the different filtering methods used on the data. The third set of nodes describe the two different correlation methods employed, and the final output is a correlation matrix between each protein.

The filters investigated:

1. Bins 1 Filter: Bins the numerical data in into discrete intervals with custom bin edges focusing more on the edges/outliers: [0.0, 0.05, 0.1, 0.2, 0.3, 0.7, 0.8, 0.9, 0.95, 1.0].
2. Bins 2 Filter: Bins the numerical data into discrete intervals with bin edges from 0 to 1 in 0.1 increments.
3. Variance filter: Filters out low-variance features from the data using a threshold of 0.05.
4. Smooth filter: Applies a moving average smoothing technique with a window size to 3 to the data.
5. Cutoff Filter: Filters out extreme values from the numerical data based on 20% percentile thresholds for lower and upper bounds.
6. Outlier Filter: Identifies outliers in the numerical data data using z-scores and threshold of 0.5 to replace them with zeros.







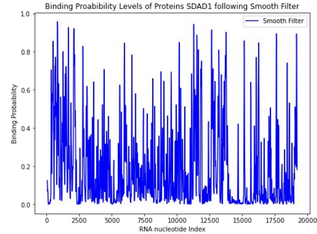


Figure 2: Displays the binding probabilities of the Protein SDAD1 following No filter, Variance Filter, Outlier Filter, and Smooth Filter.

The correlations tested:

1. Euclidean Distance: Finds the area between two curves by finding the difference in y-values at aligned x-values.
2. Pearson Correlation: Measures the strength of the linear relationship between two variables.

Results

Most correlated protein pairs for Euclidean distance:

- No Filter: SDAD1 and ZNF800
- Bins 1 Filter: DGCR8 and LSM11
- Bins 2 Filter: SDAD1 and ZNF800
- Variance Filter: SDAD1 and ZNF800
- Smooth Filter: SDAD1 and ZNF800
- Cutoff Filter: SDAD1 and ZNF800
- Outlier filter: SDAD1 and ZNF800

Most correlated protein pairs for Pearson Correlation:

- No Filter: DDX42 and XRN2
- Bins 1 Filter: DDX42 and XRN2
- Bins 2 Filter: DDX42 and XRN2
- Variance Filter: DDX42 and XRN2
- Smooth Filter: DDX42 and XRN2
- Cutoff Filter: ABCF1 and NIP7
- Outlier filter: DDX42 and XRN2

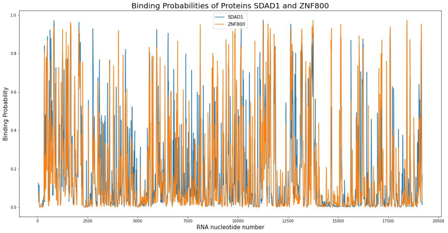


Figure 3a: This graph compares the probability binding levels between two proteins: SDAD1 and ZNF800, colored as blue and orange, respectively. This is plotted over the RNA nucleotide number and represents the popular most correlated protein pair for Euclidean distance.

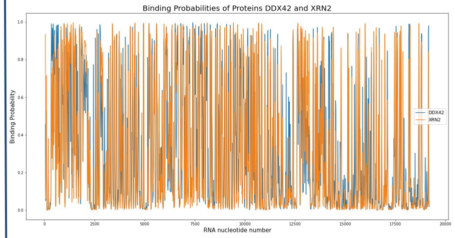


Figure 3b: This graph compares the probability binding levels between two proteins: DDX42 and XRN2, colored as blue and orange, respectively. This is plotted over the RNA nucleotide number and represents the popular most correlated protein pair for Pearson correlation.

Least correlated protein pairs for Euclidean distance:

- No Filter: HNRNPL and HNRNPM
- Bins 1 Filter: HNRNPL and HNRNPM
- Bins 2 Filter: HNRNPL and HNRNPM
- Variance Filter: BCCIP and KHDRBS1
- Smooth Filter: HNRNPL and HNRNPM
- Cutoff Filter: HLTF and ILF3
- Outlier Filter: HLTF and ILF3

Least correlated protein pairs for Pearson Correlation:

- No Filter: HLTF and HNRNPK
- Bins 1 Filter: HLTF and HNRNPK
- Bins 2 Filter: HLTF and HNRNPK
- Variance Filter: ABCF1 and HNRNPA1
- Smooth Filter: HLTF and HNRNPC
- Cutoff Filter: HLTF and HNRNPC
- Outlier filter: HLTF and HNRNPC

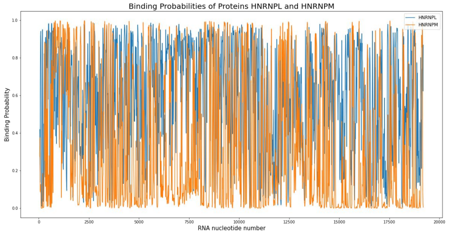


Figure 4a: This graph compares the probability binding levels between two proteins: HNRNPLand HNRNPM, colored as blue and orange, respectively. This is plotted over the RNA nucleotide number and represents the popular least correlated protein pair for Euclidean distance.

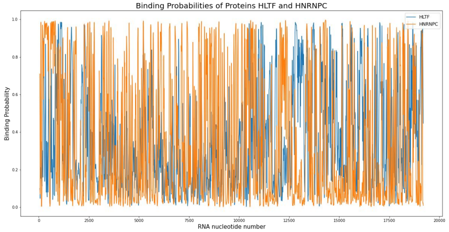


Figure 4b: This graph compares the probability binding levels between two proteins: HLTF and HNRNPC, colored as blue and orange, respectively. This is plotted over the RNA nucleotide number and represents the popular least correlated protein pair for Pearson correlation.

Conclusions

- Significantly correlated protein pairs following filtering methods applied to the binding probabilities were found to match up with outputted graphs, showing veritability of the methods.
- The highly correlated protein pair outputted using the Pearson correlation has greater similarity than the highly correlated protein pair outputted using Euclidean distance.
  - The same can be said for the lowly correlated pair, showing Pearson correlation as the better fit.
  - These conclusions are made based on Figures 3 and 4.
- The four methods that outputted distinct protein pairs were Standard, Outlier, Variance, and Smooth filters.
  - The binding probabilities following filtering are displayed in Figure 2, showing Smooth filter to accentuate the extreme values the most.
- Background research needs to be conducted on each protein pair to determine if there is a significant correlation in its function.
  - This can also be done using STRING, a database of known and predicted protein-protein interactions.

References

Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research, 2019, 47(D1):D607-D613.