Springboard Data Science Capstone Three: Predicting binding sites of an RNA-binding protein (RBP)

Grace Nye

Problem Statement: How can we predict whether or not a given RNA-binding protein will bind an RNA sequence using a convolutional neural network (CNN) with a large-scale CLIP-seq dataset with an accuracy greater than 90% in the next month?

Context:

RNA-binding proteins (RBPs) are common throughout eukaryotic genomes (around 5-10% of the proteome), and they have been found to have many key roles in biological processes. Experimental methods to detect which RNA sequences bind to given RBPs are costly and time-consuming. Therefore, we propose using high-throughput CLIP-seq data to gather datasets about known RBPs and constructing a CNN to predict whether or not the protein will bind to a given RNA sequence.

Criteria for Success:

Success will be defined by whether the trained neural network can determine whether a given RNA binds to a given RBP with at least 90% accuracy.

Scope of the Solution Space:

Solution space will include existing CLIP-seq datasets that contain known RBPs and RNA sequences whose binding to the protein has been experimentally determined. RNA sequences with positive or negative labels as well as secondary structure information that can be computationally calculated about each RNA sequence, CNNs of various architectures will be developed to provide a binary classification about whether or not a given RNA will bind to the protein.

Constraints within solution space:

Factors that can influence the success of this project include the size of each dataset (as CNNs are dependent on the size of the dataset for their performance), as well as the computational expense that accompanies the calculations.

Stakeholders to provide key insight:

Key stakeholders include my Springboard mentor, Blake Arensdorf, as well as other members of the Springboard community.

Key data sources:

https://github.com/xypan1232/iDeepE/blob/master/ideepe.py
https://github.com/liyu95/Deep_learning_examples/tree/master/8.RBP_prediction_CNN

Pan, X., Shen, H. Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics,* **2018,** 34, 20, 3427-3436. https://doi.org/10.1093/bioinformatics/bty364