Comparison of Support Vector Regression Models of Transcription Factors E2F1 and E2F4's Binding Specificities to DNA Sequences

Although all cells in the human body contain the same DNA, each one plays a unique functional role in the body because proteins called transcription factors (TFs) control how DNA information is interpreted, thus regulating which genes are expressed in the cell. In particular, the TFs E2F1 and E2F4 play crucial roles in determining whether normal healthy cells become benign or malignant tumors. Although TFs such as E2F1 and E2F4 belong to the same family and share very similar structures, experiments have shown that these TFs bind to different sequences. Despite this, current research has not yet explored the reasons for this observation. Therefore, the goal of our project was to investigate the relationships between nucleotides in the TF binding sites to see how they affect E2F1 and E2F4 binding preferences.

By using a machine learning technique called support vector regression (SVR), we used experimental data from E2F1 and E2F4 to computationally train models that predict the tendency of each TF to bind to different DNA sequences. Using these models, we found the sequence features that the TFs had a higher affinity for.

By analyzing these features, we found patterns in the DNA sequences that TFs E2F1 and E2F4 preferred. First, we found that E2F1 had a greater affinity to sequences with stretches of three adjacent A or T nucleotides than E2F4, thus supporting the findings in current literature on TF binding. In addition, we found that the flanks that were six bases long on either side of the immediate binding region had the greatest effect on TF binding. In general, we noticed that the models were less able to predict TF binding for E2F4 than for E2F1, suggesting that external factors, such as cofactors, had a greater effect on E2F4's binding specificities.

The results of this project provided a more comprehensive understanding of how TFs with very similar protein structures bind differently to DNA sequences. With these conclusions, we can aid future research on cell proliferation and the development of cancerous tumors.