Using Machine Learning Techniques to Detect Mutant p53 Transcriptional Activity

Azhar, Dua (School: Beehive Science and Technology Academy)
Kargi, Sanjana (School: Beehive Science and Technology Academy)

The p53 protein, known for its role in preventing cancer, can also become the cause of cancer when mutated. Finding a way to understand the protein and develop a way to reactivate it has been a focus of cancer research. By utilizing artificial intelligence and machine learning algorithms, we can predict the transcriptional activity of a p53 protein in silico, and reduce the workload that is needed to do so in vitro. p53's 2D and 3D structures can be analyzed to predict whether it is active or inactive. By using a dataset of over 30,000 p53 mutations with 5,409 features each, we can train a computer to assist in the prediction. Coding in Python, we used four different algorithms to concentrate our data set and train a model for prediction. First, with Correlation-based Feature Selection (CFS), we selected the 20 features that most influence p53's activity. Then we used the unsupervised learning technique, Random Cut Forest (RCF), to detect and label the data. We used a one-level decision tree, also known as a decision stump, to train and predict with our model. To improve performance, we supplemented the Decision Tree with the meta-algorithm AdaBoost. Between RCF and AdaBoost, our model used both bagging and boosting techniques. We started prediction with only a fraction of our attributes before eventually using all 5407. To increase the speed when working with all the features, we used deep learning techniques. With the help of neural networks, we were able to predict mutant p53's transcriptional activity with 99.75% accuracy in just over 10 minutes. p53 mutations are present in 50% of human cancers and is one of the most common causes of cancer. By identifying which p53 mutations are harmful, we can more accurately identify cancer in its early stages.