Prediction of orIC(Origin of Replication) using Machine Learning

Initially, we worked on predicting the origin of the replication of S.Cerevisiae using the structural properties of DNA. I used a variety of supervised learning techniques, among whom ensemble learning techniques such as gradient boosting and Xg boost were the best performing with 79% accuracy on test data.

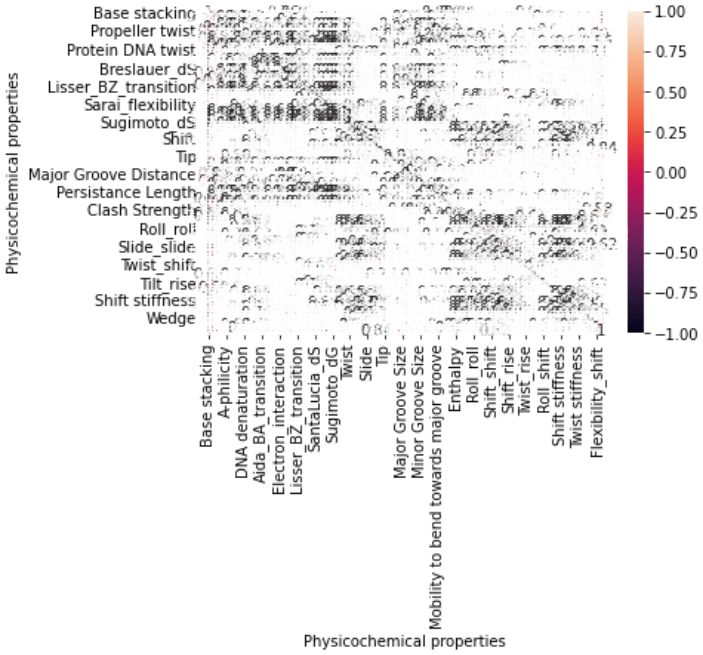

Figure 1: Showing the correlation between physicochemical properties.

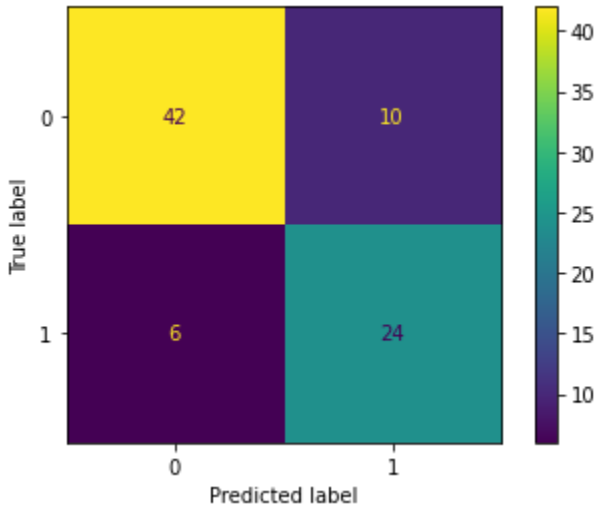

Figure 2: Showing performance of Xg boost on the test dataset.

I used deep learning-based methods later where a deep neural network couldn't perform well, but CNN based model performed much better than the rest of the models.

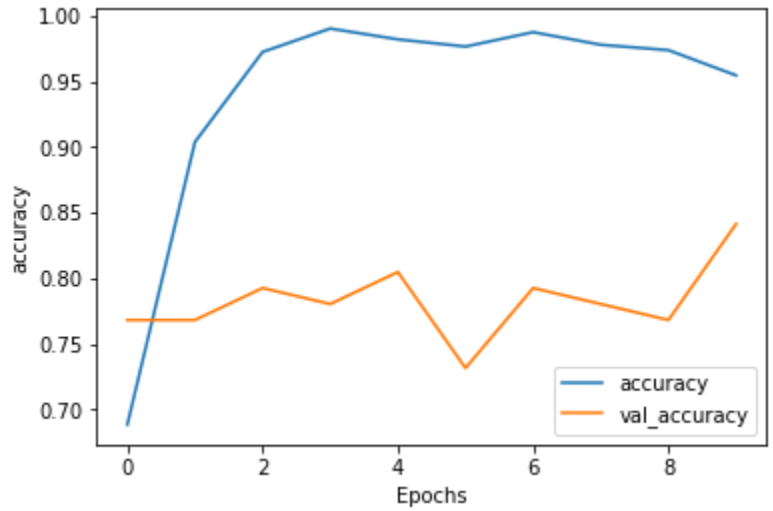

Figure 3: Showing performance of CNN-based model on the training and validation dataset.

We extend this model by using it rather than the 20 most significant features; we use all 91 DNA structural properties.

Again CNN based models perform the best among other supervised learning methods.

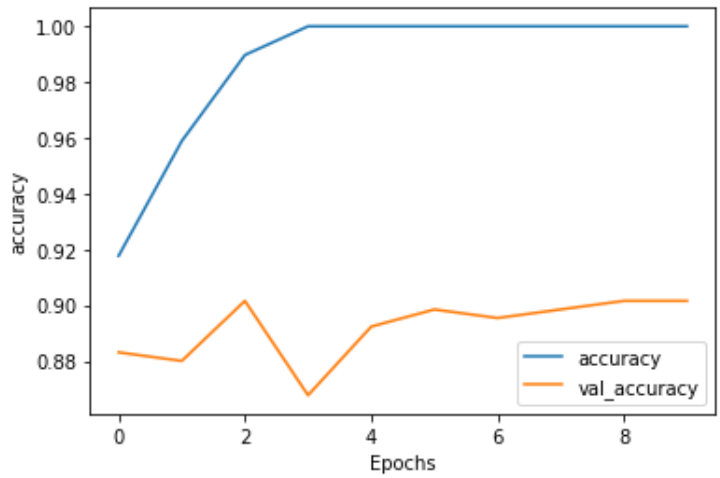

Figure 4: Showing performance of CNN-based model on the training and validation dataset.

We then used LSTM-based models to use sequence information; DNA sequences were fragmented into 1000-length sequences, and 16 dinucleotides were tokenised to integers from 1 to 16. A DNA fragment could now be represented as a sequence of numbers, one hot encoded, and the data now had the shape of 1000*16 for each fragment of length 1000 base pairs.

LSTM-based models performed well on sequence data, achieving 80% accuracy on the validation dataset.

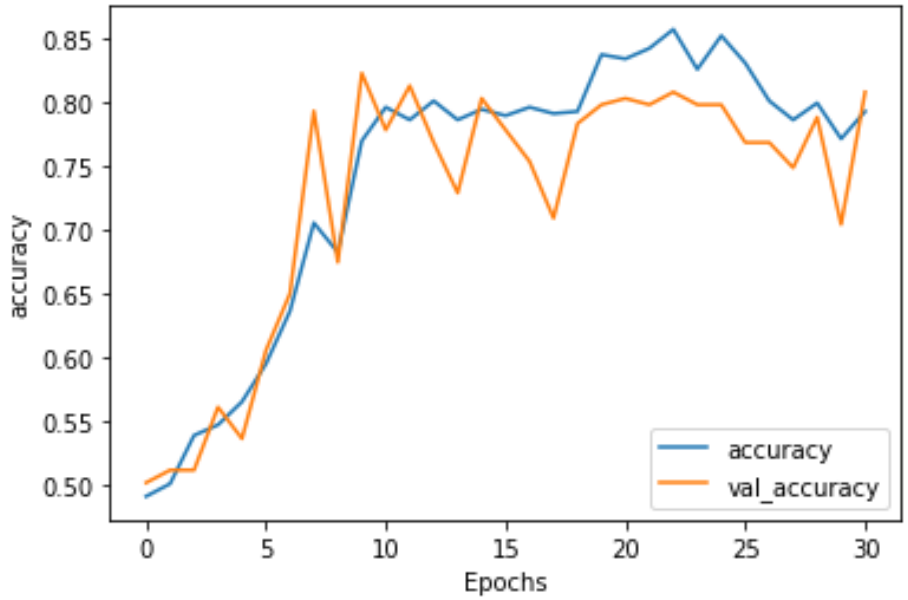

Figure 5: Showing performance of the LSTM-based model on the training and validation dataset.

MODEL TESTING

The LSTM-based model was then tested on S Pombe DNA.

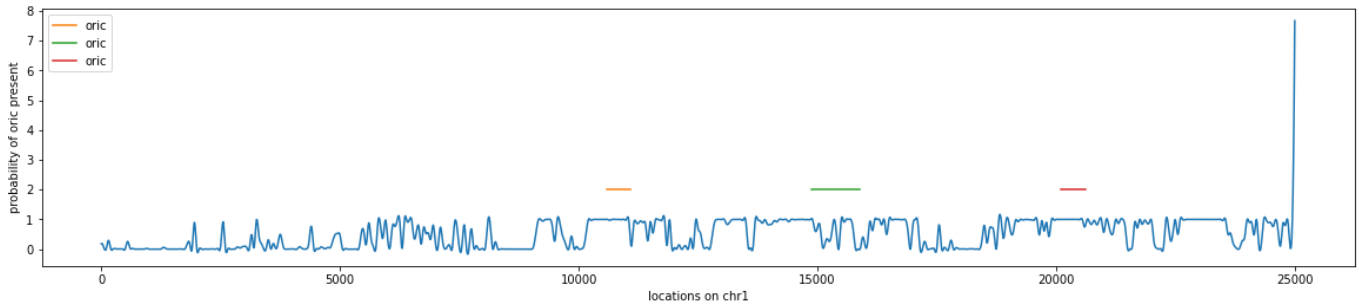

Figure 6: Showing performance of the LSTM-based model trained on the cerevisiae dataset but tested on S pombe DNA, with the model prediction of probabilities of orIC plotted and actual orIC present in the S pombe DNA also shown.

Conclusion

A variety of deep learning and supervised learning techniques were used on sequence data and numerical data extracted using DNA structural properties. On sequence data, the best-performing validation accuracy was 80% and on DNA structural properties dataset was 90%.

References

VK Singh, V Kumar, A Krishnamachari (2018). Prediction of replication sites in Saccharomyces cerevisiae genome using DNA segment properties: Multi-view ensemble learning (MEL) approach ***Biosystems***, 163, 59-69