Prediction of CTCF binding sites using a convolutional neural network

Marie Cornellier - 260805212
COMP 561: Computational Biology Methods and Research
November 5th, 2020

Intro

Transcription factors (TFs) are proteins that recognize and bind regions of DNA to regulate gene expression. It is crucial to be able to understand TF binding specificities in order to create better models of gene regulation and provide insights into genetic diseases such as cancer and autoimmune disorders. In this project I explored the usefulness of a convolutional neural network in predicting DNA sequences that are bound by TFs, using data for the most abundant transcription factor in mammalian cells. CTCF is a key transcription factor that is highly conserved throughout the evolution of mammals. Through ChIP-seq experiments, a total of 150 thousand binding sites have been identified on the human genome for CTCF. The position weight matrix (PWM) for CTCF is provided in the ENCODE Factorbook database, representing the frequencies of nucleotides at each position of the binding site. For any particular TF, there will be thousands of locations across the genome that match the PWM but that are not actually bound in vivo. The goal of this project is to differentiate between bound and unbound sites using deep learning and sequence data. Because of the flexibility of the protein to bind with many different motifs, it makes sense that the exact specificity of the sequence cannot be predicted by a single matrix. This is where a convolutional neural network (CNN) is very useful, the first layer of the network is analogous to a collection of matrices, each recognizing different motifs, that all work together to come to a final output. In more simple machine learning methods such as linear regression or decision trees, it has been shown that calculating the 3D shape of the DNA increases the ability of machine learning algorithms to predict TF binding (Mathelier et al, 2016). This is supported by evidence that transcription factor binding specificity is affected by shape as well as sequence (Maher, 2009). However, since the shape data is calculated directly from the sequence, my hypothesis was that a convolutional neural network (CNN) will be able to predict just as well with or without shape data, because it captures motifs in the sequence using different filters and kernels. After testing two different types of CNNs using sequence and shape data, I have concluded that adding shape data does not improve accuracy. However, the accuracy was already so high that it is possible that the error is simply due to sites that were missed by ChIP-seq.

Methods

The ENCODE Factorbook database provides transcription factor binding positions and scores as well as position weight matrices for various TFs. In total for these experiments, 150 thousand actual binding sites for CTCF were collected from human genome sequence data. Using a PWM and a threshold of -20, searching within known regulatory regions, another 150 thousand examples of non-binding sites were collected so that the amount of bound and unbound sites would be about equal in the training data. An extra 10 nucleotides were read on both sides of the TF binding site because the TF may be sensitive to the surrounding region, resulting in a sequence length of 35, which was reduced to 31 after calculating the shape data. Four features of DNA shape were predicted using the pentamer dictionary method (Zhou et al., 2013). The shape features were visualized using box plots to look for possible patterns or differences between bound and non-bound sequences (supplementary figure on the last page). Interestingly, there are obvious variations in the shape directly at the location where CTCF binds, but not in the surrounding extra 10 nucleotides. This is expected, as those regions have similar sequences and the shape data is calculated directly from the sequence.

Two types of convolutional neural network were designed for this project. One of them a regression type which predicts 0 for non-bound sites and a range between 0.3-1 for bound sites (using relu activation with a minimum threshold and max value). The other a binary classifier which predicts 0 for unbound and 1 for bound. The reason for the quantitative predictor lies in the nature of the target values, visualized in figure 1. What is interesting about this distribution is the peak of middle-strength binding sites. It could be that lower affinity binding sites are evolutionarily favourable because it allows for greater plasticity, and faster changing of gene expression profiles in response to environmental stimuli.

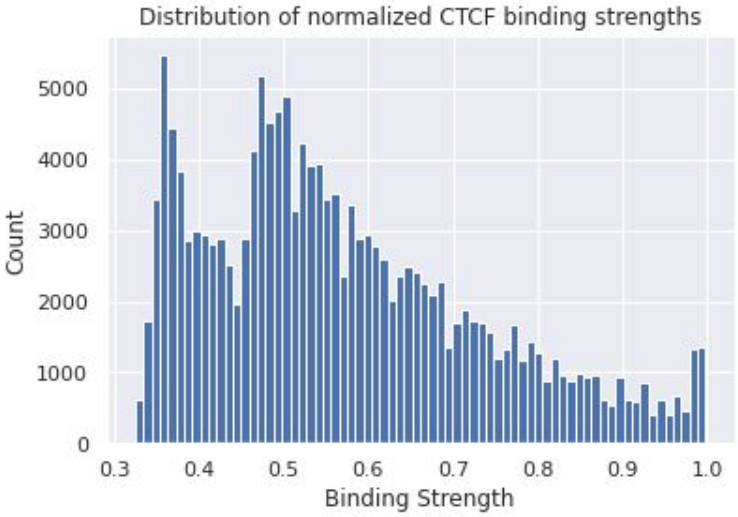


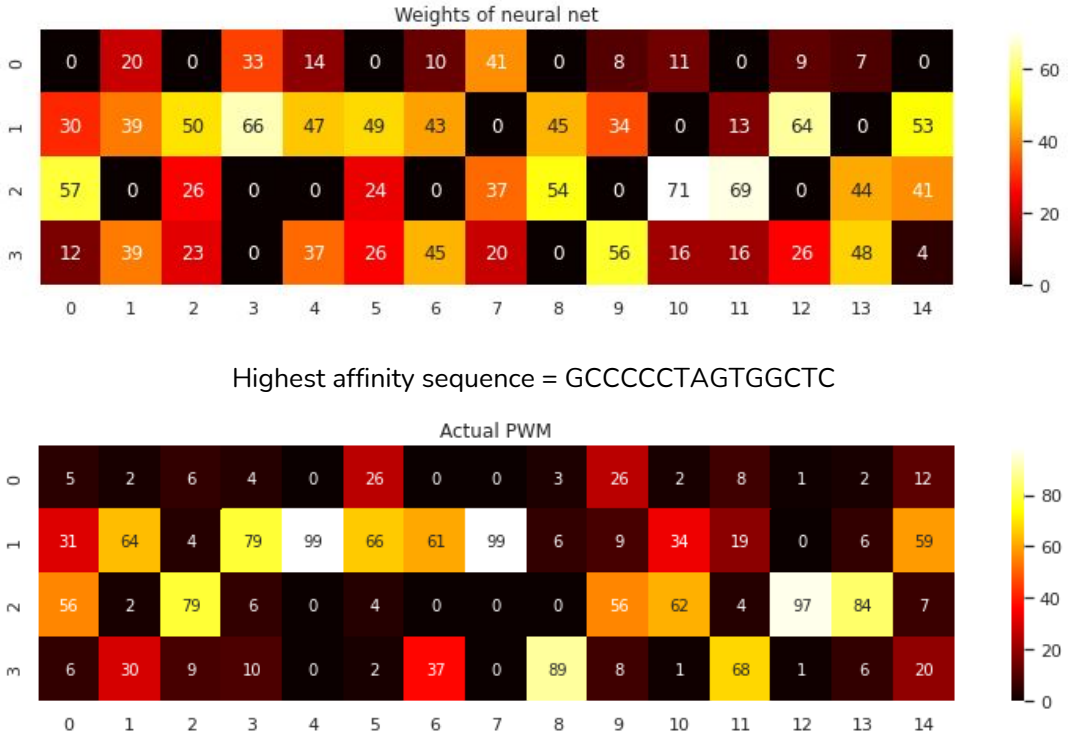
Figure 1: Histogram showing the distribution of scores for CTCF binding positions after normalizing to between 0 and 1, and removing the lowest 0.001% of outliers.
Notice that there is a threshold (1.3) below which there are no sites, thus if a machine learning algorithm can predict binding affinity on a scale of 0 to 1, there is a threshold below which it can squash values to 0 using relu activation, providing a method to both classify and quantify binding sites and affinities

After playing around with a few neural network structures I found that the accuracy was best when using 1 convolution layer with 128 filters of kernel size 25. This structure was used for all the neural networks. However, the hyper-parameters such as learning rate and batch size were tuned separately depending on the input and output data.

Results

Both binary and quantitative classification proved to be highly accurate, achieving >90% accuracy with either method. Convolutional neural networks are especially interesting for analyzing sequence data because the weights can be interpreted and visualized. By training a 1D convolutional classifier with only 1 filter with a kernel the size of the binding site, predicted against the TF score produces a matrix very similar to the actual PWM for CTCF (figure 2), and serves as a moderate regression analysis for the relative affinity of a certain binding site (figure 3A). This highlights the importance of certain sequence motifs for predicting the relative affinity of CTCF for a binding site. This can be used to produce a highest affinity binding sequence which is very similar to the one produced by the PWM.

Figure 2. Weights of neural net compared to weights of PWM, with corresponding highest activation sequence shown below each



Highest affinity sequence = GCCCCCTAGTGGCTC

Most likely to bind: GCGCCCCCTGGTGGC

The same CNN model can be used to differentiate between bound and non-bound sites by setting the target values of non-bound sites to 0, and using a relu activation layer with a threshold. This already achieves a high accuracy, however it was not as accurate as a simple binary classifier with softmax activation on the output.

Figure 3A. Regression plot for CNN used for regression with 1 layer and 1 filter

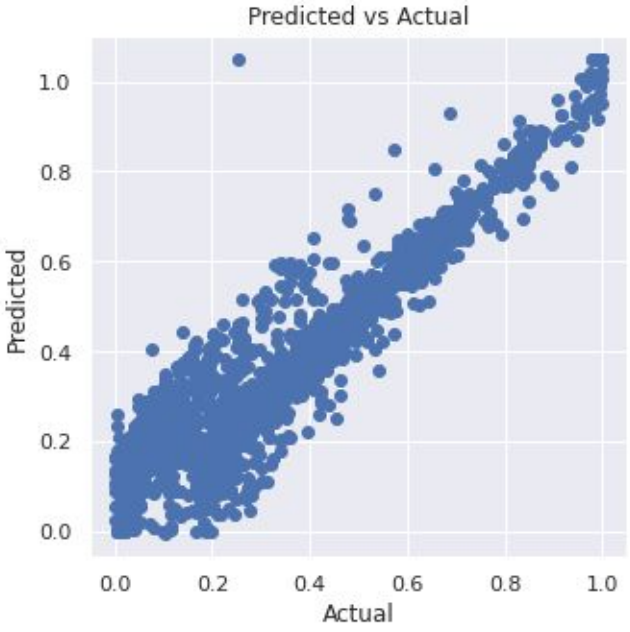
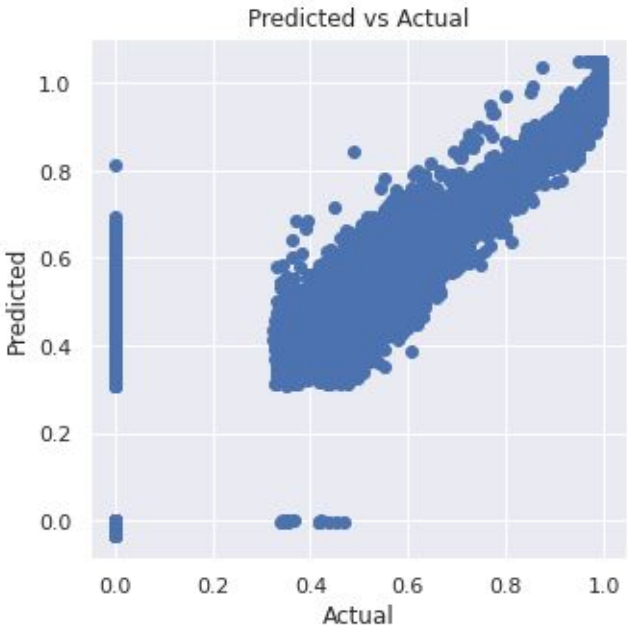


Figure 3B. Regression plot for CNN with relu activation for squashing below a threshold



Using the threshold method does achieve high accuracy (confusion plot shown in figure 4A). However, this accuracy can be improved even more by making the network a binary classifying using soft max (4B).

Figure 4A.
Confusion matrix for threshold method

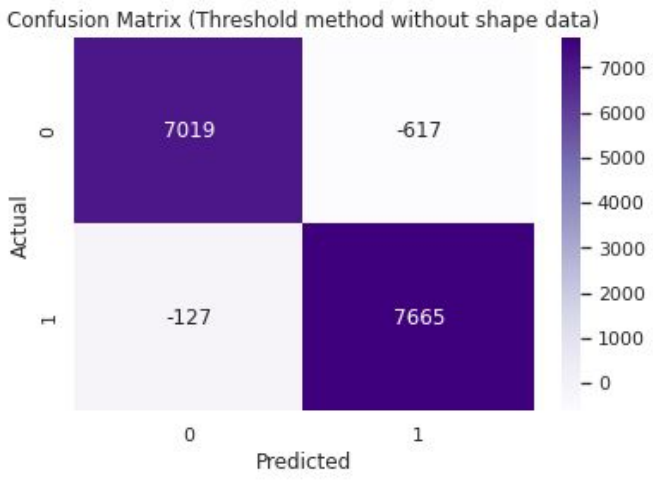
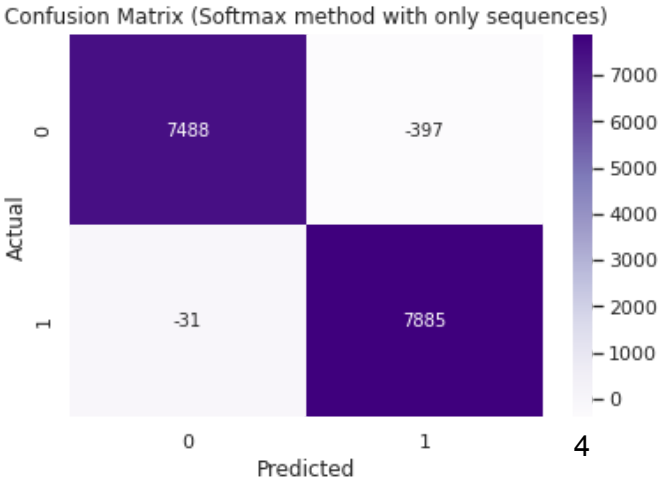
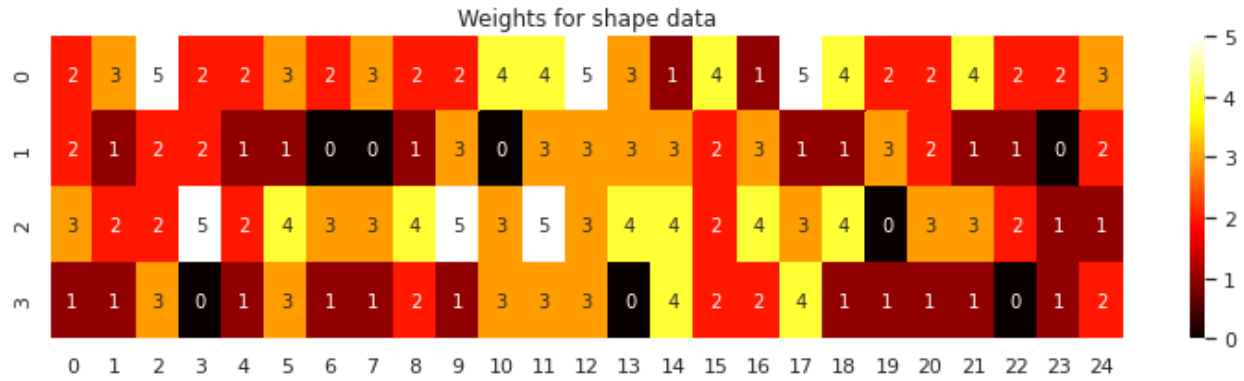


Figure 4B.
Confusion matrix for softmax method



Adding shape data did not improve accuracy for the softmax method, and was not tested on the threshold method. It is possible that the algorithm was already at the maximum accuracy it could achieve due to regions missed by ChIP-seq that are actually real bounding regions, given that most of the error is due to false-positives. Furthermore, the mean of the weights of the first layer filter was very low (figure 6), suggesting possible gradient disappearance, or perhaps that the shape data is irrelevant when the sequence data is already known. It does seem that the weights are higher in the center of the sequence, possibly due to greater importance of shape at the center of the binding position.

Figure 6. Average of the weights of the convolution layer for the shape data scaled times 1000



Discussion and Future work

The experiments in this paper contribute to recent debate concerning the relative utility of DNA sequence-based models and DNA shape-based models representing TF specificity. CNN proved to be a very accurate and fast algorithm for classifying TF binding regions. It achieved over >90% accuracy in 10 epochs, taking only a few minutes to train, and a few seconds to predict thousands outputs. Furthermore, analyzing the weights of the CNN can provide insights into the specific features and regions that are important for TF binding. Although the accuracy may be lower on a smaller dataset, deep learning neural network algorithms are likely to be the best choice for learning about DNA sequences, due to their ability to understand features without us having to manually extract or calculate them. If I had more time to continue this project I would explore other deep learning methods, as well as similar experiments with other proteins. It would also be interesting to create a multi-label CNN that can predict the binding of multiple proteins in a region of 100bp.

References

Mathelier, A., Xin, B., Chiu, T., Yang, L., Rohs, R., & Wasserman, W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. Cell Systems, 3(3). doi:10.1016/j.cels.2016.07.001

Maher, J. (2009). Faculty Opinions recommendation of The role of DNA shape in protein-DNA recognition. Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. doi:10.3410/f.1166633.629043

Zhou, T., Yang, L., Lu, Y., Dror, I., Machado, A. C., Ghane, T., . . . Rohs, R. (2013). DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Research, 41(W1). doi:10.1093/nar/gkt437

Supplementary figure