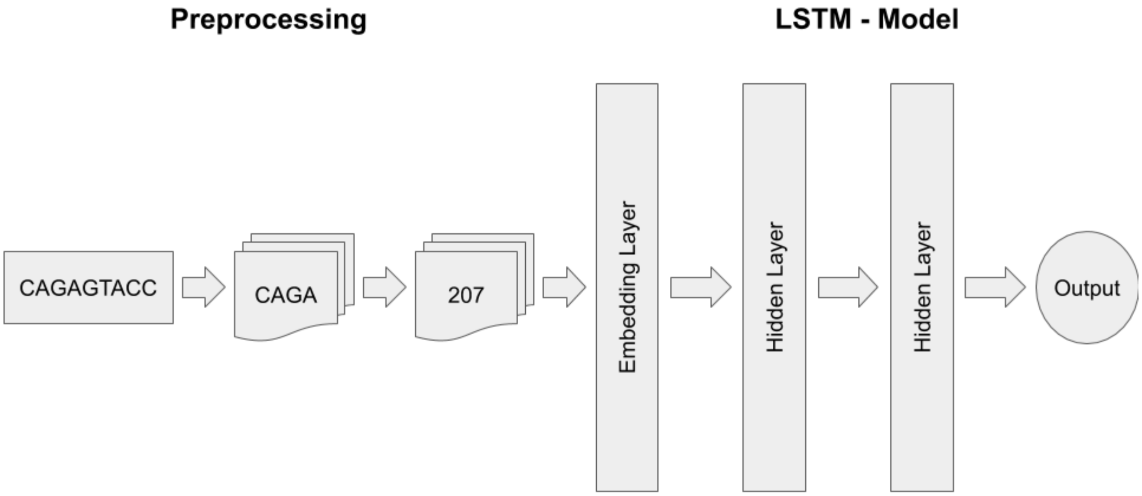
the course research project. Elaborate on the implementation methods, problems faced, figures generated (e.g., figures you may be trying to reproduce from the paper you are extending), experiments run and results obtained. Also include details on the direction the project is currently taking, benchmarks for experiments you wish to achieve and figures you wish to generate before the final project presentation/report in the week starting Nov 4th.

Solution:

Current Status: I created a LSTM Neural Network using Keras, for predicting binding probabilities of DNA sequences with one specific protein. A further description of the model can be found in the following. So far we downloaded one Transcription Factor binding dataset for the A549 cell, containing Human-DNA sequences, the dataset was created out of an assembly in December 2013 and can be found at https://genome.ucsc.edu/cgi-bin/hgTables. The dataset contains in total 11014 DNA sequence with an average length of 472 nucleotides, the longest sequence is 1813 nucleotides long. These sequences can be used as positive samples for training and testing the created model, for each of these sequences a negative samples was generated by matching the size and GC-content of the positive sample.



Before feeding a sample in our network we split it up in it's k-mers and transform these k-mers to unambiguous numerical values. Then the Model reads those values sequentially to predict the binding capability of the sequence. Our LSTM Neural Network currently consist of an Embedding Layer, two Hidden Layers and a logistic Sigmoid function to generate the final output. This output ranges between 0 and 1, where a higher value can be interpreted as an higher binding probability. We can compare the output with thresholds to give a boolean answer if we think that a protein will bind to this specific DNA sequence. We used different techniques like dropout and gradually reducing of the learning rate to boost the performance of our model.

Plan for the last week: By the end of the project we hope to be able to achieve good results on the described dataset and other suitable datasets we still need to find. We hope to generate comparable results to the work of Zhen Shen et al[1]. For this we will need to deploy our model to a server and train it with multiple different parameters, depending on the results we need to adapt the structure of the model to perform better. If time allows and no major changes within the core model are necessary we would like to create a model for multi class classification, which learns from data of multiple proteins and can make predictions about the binding properties of a DNA sequence with every of the proteins at the same time.

Problems: We hoped to access the Test and Training Data used by Zhen Shen et al[1] as well as the model they used. With this we would have cut a lot of effort in reimplementing a comparable model on our own. Sadly the authors didn't share any code or data publicly and did not reply to us when we tried to reach out for them via Email. Because of this we can not hold up with our original plan where implementing our own comparable model was only the first step, followed by further improvements such as using an attention mechanism for an early identification of the key parts for protein binding within a long DNA sequence. Possible comparisons of the result of our model with different published work can also only be made with caution since we have to generate our own datasets for training and testing. We hope that these datasets have the same properties as the data used by Zhen Shen et al[1].

[1] Zhen Shen, Wenzheng Bao, De-Shuang Huang(2018)
Recurrent Neural Network for Predicting Transcription Factor Binding Sites.

3. (2 points) Please find below the link to the post-midterm feedback form for the course CS6024. Use your smail ids to access: `https://forms.gle/ReSWabMNeVkBGDcH6`. Kindly provide feedback by Assignment deadline. Constructive criticism is encouraged and requested.