Bioinformatics Research Centre
Aarhus University

Predicting Protein Secondary Structure using Artificial Neural Networks with a focus on RNNs

Mathias Byskov Nielsen, 201506038

Supervisor: Christian Nørgaard Storm Pedersen
Master's Thesis
June 2020

ABSTRACT

Protein secondary structure prediction is a huge field and very important in order to determine a protein's overall three-dimensional structure. There has been a rapid development within the fields of genomics and proteomics the last decades, which makes the computational and statistical methods for structure-prediction more important than before [1]. This thesis investigates the field within protein secondary structure prediction. The dataset used to train all networks in this project is the publicly available CB513 dataset [2].

First, feedforward neural networks are presented along with the implementations and experiments conducted. The implementation presented is inspired by Qian and Sejnowski [3], which is a simple one-layer feedforward neural network. They claim to have obtained an accuracy of 64%. The implementation presented here obtains accuracy in the same range ($63\% - 65\%$).

Next, a recurrent neural network implementation is presented. Various different structures were investigated and the results are presented in chapter 7. The final implementation inspired by Heffernan and Yang [4] which involves a bidirectional LSTM-network obtained an accuracy of approximately 70%. The article by Heffernan and Yang claims to have obtained an accuracy of 83.9% on the CB513 dataset, although they trained their implementation on a dataset much larger than the original CB513.

CONTENTS