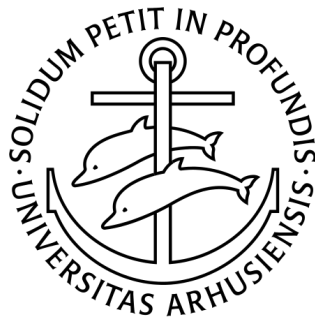


Bioinformatics Research Centre
Aarhus University

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

1	INTRODUCTION	1
I THEORETICAL BACKGROUND		
2	PROTEINS	4
2.1	General	4
2.2	Structure	5
2.3	Structure Determination	8
2.3.1	X-Ray Crystallography	8
2.3.2	NMR Spectroscopy	9
2.3.3	Electron Microscopy	9
2.4	Determining Secondary Structure from Atomic Coordinates	9
3	REPRESENTATION OF STRUCTURE	12
3.1	Primary Structure Representation	12
3.2	Secondary Structure Representation	13
4	PROTEIN SECONDARY STRUCTURE PREDICTION	15
4.1	Prediction Accuracy Measures	15
4.2	Hidden Markov Models	16
4.3	Support Vector Machines	17
4.4	Neural Networks	18
II METHODS, IMPLEMENTATION & RESULTS		
5	DATASET	21
5.1	General information of the CB513 dataset	21
5.2	Distributions in the CB513 dataset	21
6	FEEDFORWARD NEURAL NETWORK	23
6.1	Methodology	23
6.1.1	Training a feedforward neural network	25
6.2	Data Parsing	26

6.3	Implementation	28
6.4	Results	30
6.4.1	Input Wrangling	31
6.4.2	Architecture	34
6.4.3	Optimizers	36
6.4.4	Regularization	40
6.4.5	Summary of FFNN Results	42
7	RECURRENT NEURAL NETWORK	43
7.1	Methodology	43
7.2	Data Parsing	44
7.3	Implementation	46
7.4	Results	48
7.4.1	Bidirectional RNN	50
7.4.2	Gated RNNs	52
7.4.3	Structure proposed by Heffernan and Yang	55
7.4.4	Summary of RNN results	57
8	CONCLUSION	58

III APPENDIX

A	APPENDIX	61
A.1	Amino Acid Letter Code	61
A.2	Amino Acid Distribution of CB513	62
A.3	Mathematical formulation of activation functions	63
A.4	Adam Algorithm: The update steps	63
A.5	Feedforward NN: Optimizers and batch-size	64
A.6	Calculating gradients in a RNN setting	65
A.7	Calculations for the LSTM unit	66
A.8	Gated Recurrent Unit	66