Signal peptide prediction

Timothy Bergström

January 1, 2018

Contents

Abstract

Detecting signal peptides (SPs) in proteins computationally is . . . A neural network model was created and trained with X amount of data, which yielded a model with a classification accuracy of 92% of the validation set. The model was used to predict the total amount of SPs in the X and Y organism, which yielded Z and Q SPs respectively, showing that the model is . . .

1 Introduction

Course bla.

The goal is to create a classifier that can accurately classify proteins with signal peptides (SPs) in an organisms proteome.

SPs are short regions in peptides that are on average between 16 to 30 amino acids long, but can be as short as 8 and as long as 65 amino acids [2]. SPs function is to translocate proteins to different parts of the cell, such as insertion into membranes, moving proteins to organelles or to help the cell excrete the proteins [3].

1.1 The structure of a signal peptide

An SP is always located near the N-terminal of a protein and have three regions; the n-region, the h-region and the c-region.

The n-region is the first region, which usually contains positively charged amino acids, making the region very polar and hydrophilic. Then comes the h-region, which contains mostly hydrophobic amino acids that forms an α-helix and lastly the c-region, which is usually a sequence of 3 to 8 non-charged polar amino acids. After the c-region, there is usually a cleavage site that have a tendency to be surrounded by small, uncharged amino acids.

When analyzing an organisms proteome, the SPs can give an indication where in the cell the proteins are translocated to which could be of an interest when researching an organism. Not all proteins contain SPs, which is why signal peptide prediction is useful for determining the layout of the cell and for annotating organisms proteomes. The problem with signal peptide prediction is that polar regions in trans-membrane proteins can mistakenly be predicted as the h-region of an SP, which makes predicting SPs in trans-membrane proteins difficult.

2 Methods and models

2.1 Design choices

Design choices, python, keras etc.

2.2 Data set

Scraped from websites, extracted etc. POS and NEG, see Table 1

Type	Positive	Negative
TM	10	30
non-TM	1	2
Unknown	a	b

Table 1: Data table.

Typical signal peptide from dataset:

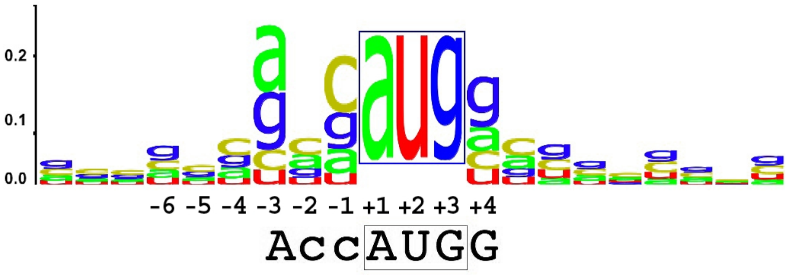

Figure 1: Extracted data from all positive samples

How you handled data, noise, ignore etc for bad samples. Seq files `.fasta` file containing sequences

See Figure 1

2.3 Neural Network structure

LSTM, CNN and dropouts. Reference self to text generator?

Model picture and maybe summary?

2.4 Related works

There are multiple softwares and websites that predicts signal peptides, such as SignalP, SignalBlast and Predisi to name a few [4] [5] [6]. Neural Networks (NN) are the most used methods while some sites use Hidden Markov Models (HMM) instead. Similar methods are used to predict other biological problems, such as motifs and trans-membrane regions in proteins.

3 Evaluating model

3.1 Model performance

ROC-curve, CM-matrix for full data full data.

3.2 Signal peptide detection in transmembrane proteins

Evaluate transmembrane and non transmembrane acc/loss. CM for TM and non-TM

3.3 Signal peptide detection in proteomes

Test two organisms, correct for CM

3.4 Generative model

A generative model was created that can create SPs. See in appendix 5.3.

4 Conclusion

Is it good enough? Can the model be improved? Can you use other methods (ex HMM, CNN, grouping etc...)

References

[1] Mrfool *Readings compiled for History 21.479.* 1991.

[2] Henrik Nielsen http://www.cbs.dtu.dk/services/SignalP-1.1/sp_lengths.html

[3] Signal peptides wikipedia https://en.wikipedia.org/wiki/Signal_peptide

[4] Henrik Nielsen, *Predicting Secretory Proteins with SignalP*, In Kihara, D (ed): Protein Function Prediction (Methods in Molecular Biology vol. 1611), pp. 59-73, Springer 2017, doi: 10.1007/978-1-4939-7015-5_6, PMID: 28451972, http://www.cbs.dtu.dk/services/SignalP/

[5] Karl Frank; Manfred J. Sippl, *High Performance Signal Peptide Prediction Based on Sequence Alignment Techniques*, Bioinformatics, 24, pp. 2172-2176 (2008), http://sigpep.services.came.sbg.ac.at/signalblast.html

[6] Karsten Hiller, *PrediSi*, Institute for Microbiology Technical University of Braunschweig,http://www.predisi.de/

5 Appendix

5.1 Training the model

Train time, gpu used, specs, time to train etc... Gpu is needed to improve training speed. Add epoch time for gpu and cpu

Everything that the person reading don't want to know

5.2 Pitfalls when training a model

Concept of under and overfitting and how to prevent it. Example pictures of bad models.

Overfitting . . .

other stuff too

more stuff

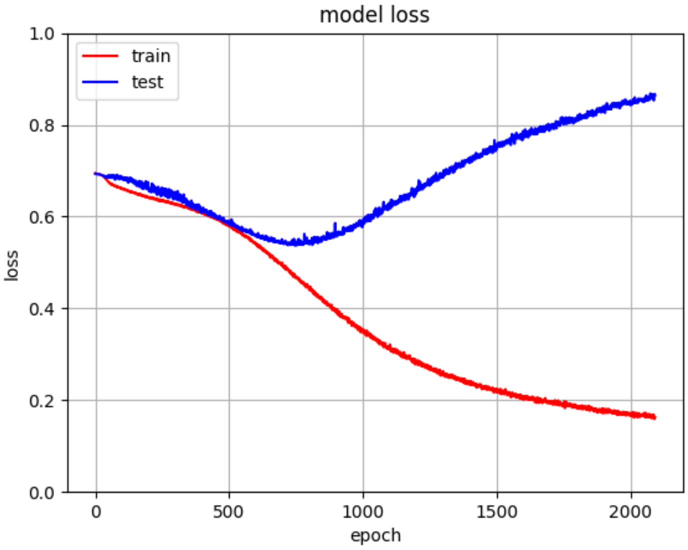

Figure 2: Overfitted

Too high learning rate . . .

other stuff too

more stuff

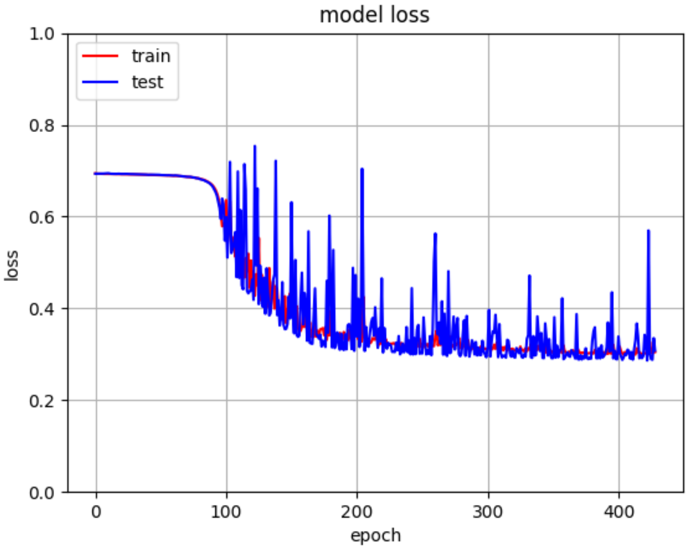

Figure 3: Too high learning rate

5.3 Generative Model