Machine Learning: Assignment 3 - Group 01

Andreas Salling Heglingegård[1], Andreas Sørensen[2], and Christian Zhuang-Qing Nielsen[3]

[1]201405359, 201405359@post.au.dk
[2]201508158, au542705@uni.au.dk
[3]201504624, christian@czn.dk

November 23, 2017

1 Introduction

This assignment is about using a hidden Markov model for gene finding in prokaryotes. We are given 10 genomes (each with several genes). For five of them, we are given the location of these genes. Our objective is to find the location (read: structure) of the genes in the remaining five genomes. Since know that they follow a 'gene syntax' we can use the first five genomes as training data to learn how to predict the location of the remaining genes.

2 Our model

We chose a model with 43 hidden-states. This model have three start and stop `codons` for both `R` and `C`. The start `codons` for `C` are `ATG`, `GTG`, and `TTG`, the stop `codons` are `CAT`, `CAC`, and `CAA`. The start `codons` for `R` are `TTA`, `TCA`, and `CTA`, the stop codons are `TAG`, `TGA` and `TAA`. We picked them based upon the analysis from the slides. To train we did training by counting. Here we divided the states as follows: black is `N`, red is `R`, and blue is `C`. During the training we only counted the start and stop `codons` if they actually happen. If there is a switch from `N` to `C` and one of the start codons is not there, we do not count anything and jump six positions ahead and move to the 12 state directly. If there is a switch from `C` to `N` we look at previous three genes and check if it should be counted. If there is a switch from `C` to `R` or `R` to `C` we just skip it in the training. We skip quite a bit of the data however because of the size of the genomes it should not matter as much. To predict the unannotated genomes we used the Viterbi algorithm to calculate ω. We then used backtracking to predict the most optimal path.

3 Optimization

It is probably possible to optimize the code quite a bit if it is made with a smaller but as effective model or by optimizing the way we calculate ω with the Viterbi algorithm.

4 Result

The accuracy we got was not as accurate as we had hoped for. The approximate correlation coefficient is 0.25-0.35 from the genefinder-verifier and compare-ann. The inaccuracy stems from the start-codons, stop-codons, and the model we chose. We did not choose the start- and stop-codons from analysis of the genomes genome1.fa-genome5.fa, we chose the codons based upon the information from the slides. If we did an analysis of the genomes we could have increased the accuracy. We only tried one model to train the hidden Markov model, which means there is most likely a better model to predict than `N`, `C`, and `R` hidden layer.

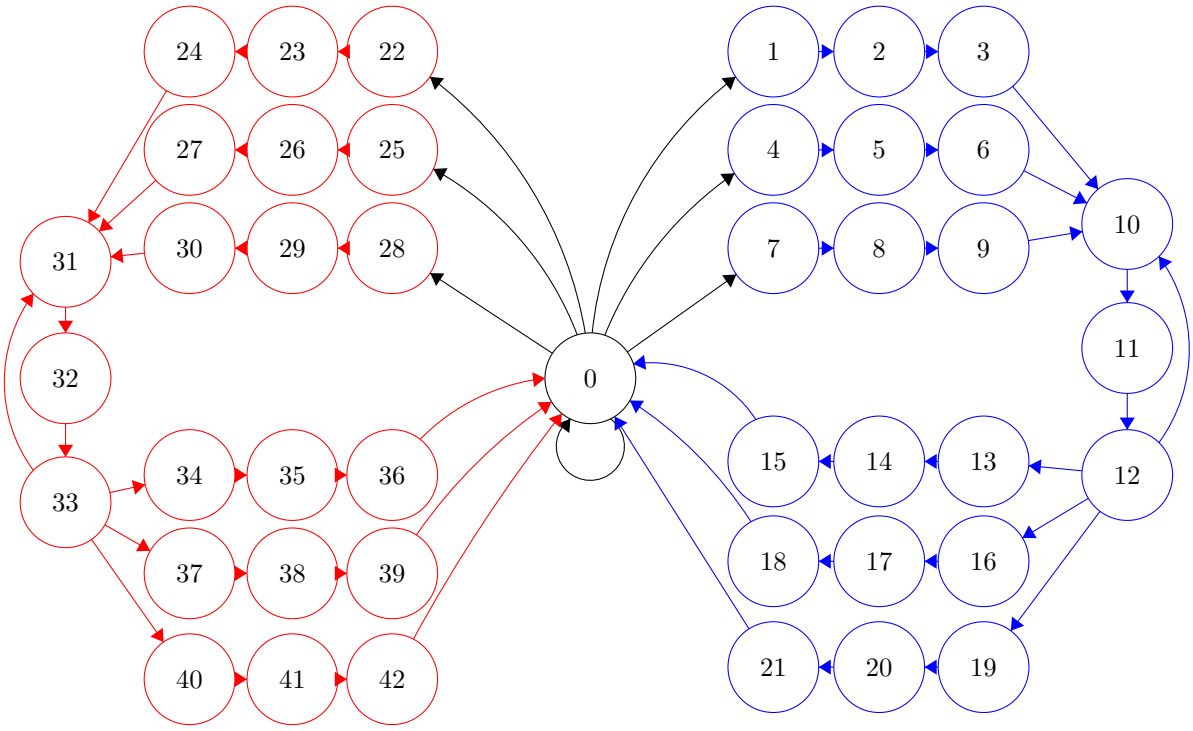

Figure 1: The model which we used

```
Genome 6
Cs   (tp=719760, fp=179337, tn=301561, fn=89368): Sn = 0.8896, Sp = 0.8005, AC = 0.5443
Rs   (tp=689405, fp=148408, tn=313240, fn=77689): Sn = 0.8987, Sp = 0.8229, AC = 0.6007
Both (tp=1409165, fp=327745, tn=223872, fn=167057): Sn = 0.8940, Sp = 0.8113, AC = 0.3419
Genome 7
Cs   (tp=817111, fp=246896, tn=517651, fn=120782): Sn = 0.8712, Sp = 0.7680, AC = 0.5635
Rs   (tp=781561, fp=258530, tn=525006, fn=113427): Sn = 0.8733, Sp = 0.7514, AC = 0.5585
Both (tp=1598672, fp=505426, tn=404224, fn=234209): Sn = 0.8722, Sp = 0.7598, AC = 0.3548
Genome 8
Cs   (tp=674059, fp=153031, tn=338304, fn=101924): Sn = 0.8687, Sp = 0.8150, AC = 0.5703
Rs   (tp=590257, fp=188540, tn=351610, fn=88618): Sn = 0.8695, Sp = 0.7579, AC = 0.5385
Both (tp=1264316, fp=341571, tn=249686, fn=190542): Sn = 0.8690, Sp = 0.7873, AC = 0.3229
Genome 9
Cs   (tp=729085, fp=203509, tn=335018, fn=127757): Sn = 0.8509, Sp = 0.7818, AC = 0.4894
Rs   (tp=738426, fp=254640, tn=346985, fn=115790): Sn = 0.8644, Sp = 0.7436, AC = 0.4673
Both (tp=1467511, fp=458149, tn=219228, fn=243547): Sn = 0.8577, Sp = 0.7621, AC = 0.2086
Genome 10
Cs   (tp=581504, fp=108653, tn=239039, fn=118251): Sn = 0.8310, Sp = 0.8426, AC = 0.5151
Rs   (tp=363221, fp=159817, tn=295158, fn=62132): Sn = 0.8539, Sp = 0.6944, AC = 0.5116
Both (tp=944725, fp=268470, tn=176907, fn=180383): Sn = 0.8397, Sp = 0.7787, AC = 0.2554
```

Figure 2: Picture of the results from genefinder

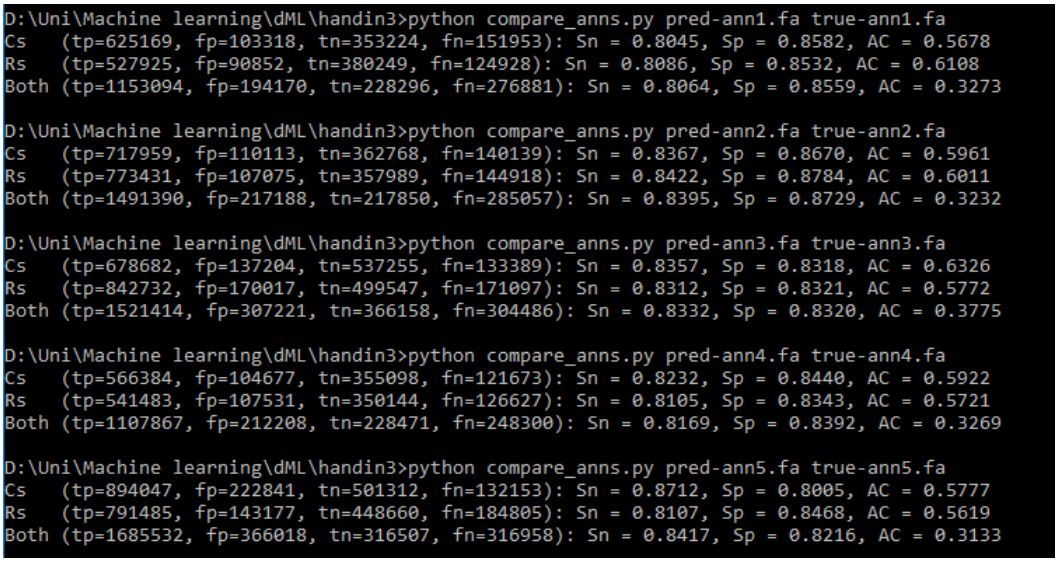

Figure 3: Picture of the results from compare-ann