Group Report - Team 2

Handin 3: Hidden Markov Models

November 24. 2017

Students:

William Lund Sommer (201303715)

Emil Hartvig Pedersen (201303680)

Anders Hein Hansen (201305292)

Machine Learning, ML, 2017

Department of Engineering, Aarhus University

1 Introduction

In this report we present the results of a Hidden Markov Model (HMM) for gene annotation. It achieved an average approximate correlation of 0.37 for five unknown genomes.

2 Theory

2.0.1 Explanation of model:

The model is inspired by the codon model, and has the same start and stop sequences as that model. To reduce computational requirements the main forward and backwards coding sequences are replaced with sequences of three non-specified observables. See appendix A for the model. The states in the model account for the fact that the genome has genes in both directions. The model starts of assuming 7 different start codons. These are 'ATG', 'ATC', 'ATA', 'ATT', 'GTG', 'GTT' and 'TTG' for the forward coding sequences, and three different stopping codons 'TAG', 'TAA' and 'TCA' (see appendix A for possible stopping and starting codons for the reverse direction).

2.0.2 Training the model:

The model was trained by counting the number of transitions and emissions from different states and dividing by the total number to get the probabilities. In order to do this we had to translate the annotations of C, N and R. This was done using indexing. We made a class with different dictionaries corresponding to the different sequences (forward start, forward stopping, forward coding, reverse coding) making strings to indexes. From here we defined different strings in the dictionaries that was connected to certain numbers (indexes) in such a way that each string had it's own specific number (that goes for both the forward and backward direction). The strings and numbers was in pairs of threes. Such that the forward starting values for example 'ATG' was associated to the numbers [1,2,3]. Thus to translate the annotation to indices, we looped through the annotation string together with the genome data, and set up if-statements to match the different sequences of annotations with the genome data, to obtain the indices.

2.0.3 Illustration of model:

For an illustration of the model, a transition diagram are to be found in Appendix A. This diagram shows all the different states in the model together with the transitions made in the model. Also transition and emission probabilities are shown. Allowing for different start- and stop-codons gives more flexibility in the model. This of course, takes up extra computing time, however with the result of a more accurate model.

2.0.4 Prediction of the gene structure for the unannotated sequences:

In order to predict the 5 unannotated genomes we apply the best model (trained on the 5 annotated genomes) and use the Viterbi algorithm with the subsequent backtracking. The results for this are found in the "Results and Data Analysis" section in table 2.

2.0.5 Transition diagram:

A transition diagram of our model can be found in Appendix A.

3 Results and Data Analysis

3.0.1 Approximate correlations (AC):

The empirical results for the predictions on the first five genomes are to be found in table 1 below:

3.0.2 Model results for cross validation (training on annotated genomes):

In order to obtain a better estimate of our model accuracy we use cross validation, to evaluate the different performances. The results for the cross validation is displayed in the table below:

Approximate correlation (AC) for the cross validation			
Specific genome	Only Cs	Only Rs	Both
Genome 1	0.5608	0.6324	0.3877
Genome 2	0.6052	0.6272	0.4052
Genome 3	0.6314	0.6153	0.4308
Genome 4	0.5868	0.5806	0.3619
Genome 5	0.6370	0.5745	0.3927

Table 1: Approximate correlations (AC) between the predictions on the genome 1 to 5 and their true annotation in the cross validation. The gene predictor was trained by the remaining four genomes respectively. The training method used for this is "Training by Counting".

3.0.3 Model results using Viterbi decoding and backtracking (predicting unannotated) genes):

Below one can see the results for predicting the last 5 genomes using the model trained on all the first 5 genomes. The average of the approximate performance, was found to be 0.3477, which is an okay result, however it is possible to improve this performance by a significant amount, by using for example a full codon model.

Genome 6	Genome 7	Genome 8	Genome 9	Genome 10	Average
0.4024	0.4023	0.3711	0.2630	0.2996	0.3477

Table 2: Individual approximate correlations (AC) between the predictions and the true annotations, including the average

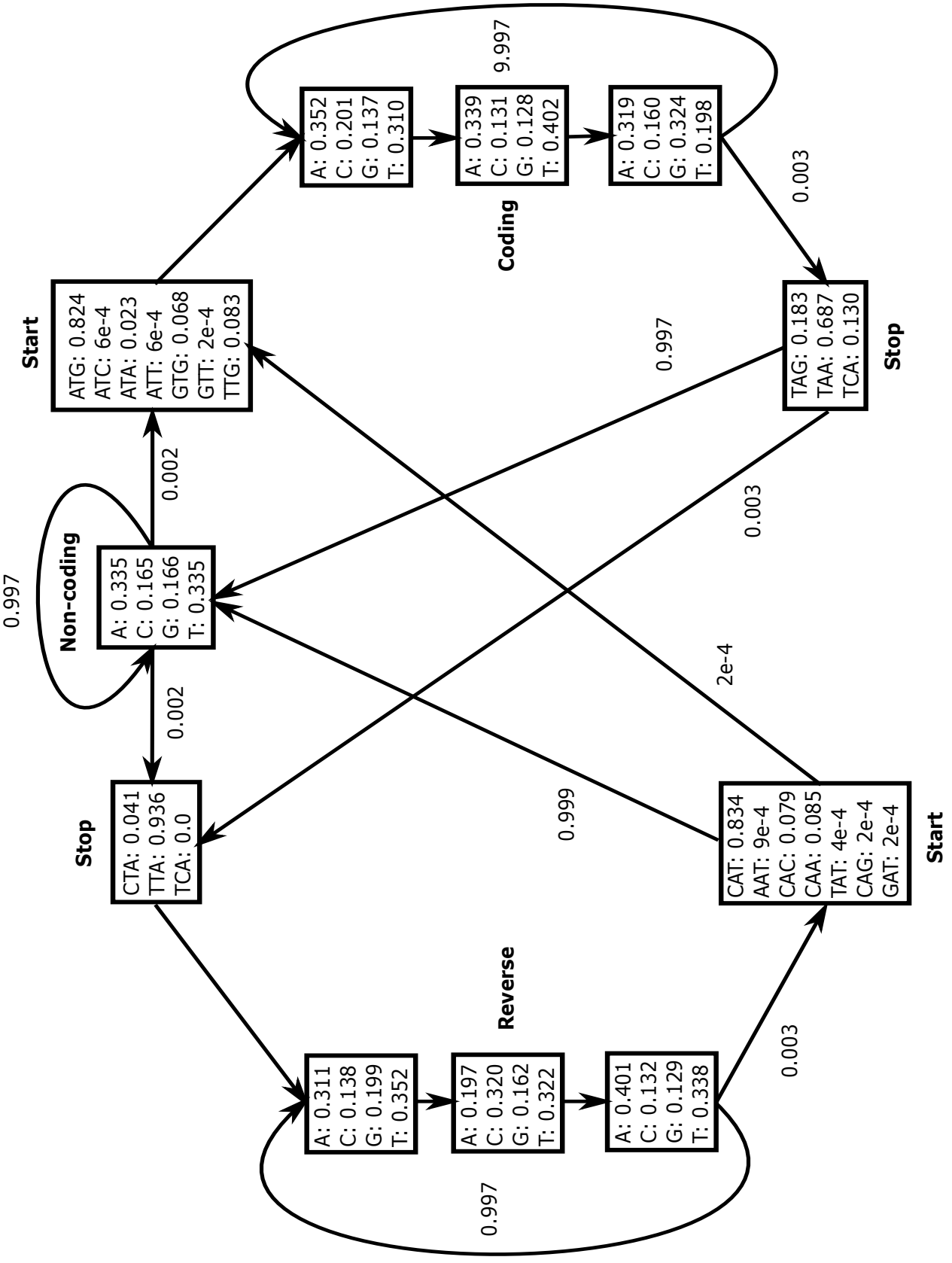
4 Discussion

All in all the hidden Markov model we constructed worked as it was supposed to, however it didn't achieve top performance. This is probably due to the fact that we didn't apply the full cordon model, in our hidden markov model framework, but only used one of the models inspired from the lecture. Our model didn't treat all the different combinations of the different gene triplets as separate states when in the main coding sequences. For better performance, one should have tried the full codon model. However this would have required a lot more computing power, taking up more time.

5 Conclusion

Our model was able to provide an average approximation correlation of 0.3477 when considering the predictions vs. the true annotations, of the genome 6 to 10. The computing time of the whole script took around 2,5 hours to run. (cross validation + unannotated genomes)

6 Appendix A


A: 0.352
C: 0.201
G: 0.137
T: 0.310
A: 0.339
C: 0.131
G: 0.128
T: 0.402
A: 0.319
C: 0.160
G: 0.324
T: 0.198
9.997
Coding
0.003
Start
ATG: 0.824
ATC: 6e-4
ATA: 0.023
ATT: 6e-4
GTG: 0.068
GTT: 2e-4
TTG: 0.083
0.997
TAG: 0.183
TAA: 0.687
TCA: 0.130
Stop
0.002
0.997
0.003
Non-coding
A: 0.335
C: 0.165
G: 0.166
T: 0.335
0.997
0.002
2e-4
Stop
CTA: 0.041
TTA: 0.936
TCA: 0.0
0.999
CAT: 0.834
AAT: 9e-4
CAC: 0.079
CAA: 0.085
TAT: 4e-4
CAG: 2e-4
GAT: 2e-4
Start
A: 0.311
C: 0.138
G: 0.199
T: 0.352
A: 0.197
C: 0.320
G: 0.162
T: 0.322
A: 0.401
C: 0.132
G: 0.129
T: 0.338
Reverse
0.003
0.997