Handin 1

Søren Meldgaard 201303712, Malthe Bisbo 201303718

November 23, 2017

1 Status of work

2 Model structure

In order to translate the annotations into sequences of hidden states, we have used the following strategy. All the annotations start with an N, which is immediately translated into the hidden state 0. Then we look at the next 4 annotations. If we encounter 'NCCC' we look in the genome file and assign '0XYZ' where 'XYZ' is the corresponding start-code. If we do not recognize the start-code we simply label the next coding part as non-coding. If we have encountered a start-code we then look for 'CCC' and label it '[10, 11, 12]' until we find one of the three stop-codes. After looking for 'NCCC' we look for 'NRRR' and the procceed in the same way as for 'NCCC'. If none of the two have been found it must be an 'N' which we then label as '0'. The process is then repeated for the next base. This means that we would miss 'RRRCCC', since we look for 'NCCC' and then 'NRRR', but we have assumed that it has neglible effect.

3 Gene prediction

In order to predict the gene structures for the sequences without annotation we have trained our model by counting. Then the Viterbi-algorithm is used to decode the sequence (X) provinding a sequence of hidden states (Z). This is then easily translated into an annotation, since hidden state '0' corresponds to 'N', 1 to 21 is 'C' and 22 to 42 is 'R'.

4 Performance of gene predictor

In order to train the model 5 genomes with known annotations are available. To check the performance the HMM is trained on 4 of them and then used to predict the one left out. This

prediction is then compared to the correction annotation. This is done for all five permutations, with the results for each stage of cross validation shown below.

Validating on genome1:
Cs (tp=658123, fp=166731, tn=301163, fn=58582): Sn = 0.9183, Sp = 0.7979, AC = 0.5985
Rs (tp=546592, fp=121250, tn=306046, fn=53699): Sn = 0.9105, Sp = 0.8184, AC = 0.6480
Both (tp=1204715, fp=287981, tn=247464, fn=112281): Sn = 0.9147, Sp = 0.8071, AC = 0.4359

Validating on genome2:
Cs (tp=756778, fp=154730, tn=307365, fn=57869): Sn = 0.9290, Sp = 0.8302, AC = 0.6330
Rs (tp=796436, fp=138307, tn=305126, fn=60108): Sn = 0.9298, Sp = 0.8520, AC = 0.6527
Both (tp=1553214, fp=293037, tn=247257, fn=117977): Sn = 0.9294, Sp = 0.8413, AC = 0.4527

Validating on genome3:

Validating on genome4:
Cs (tp=590520, fp=139488, tn=331346, fn=66225): Sn = 0.8992, Sp = 0.8089, AC = 0.6226 Rs (tp=545012, fp=124255, tn=317385, fn=80186): Sn = 0.8717, Sp = 0.8143, AC = 0.6015 Both (tp=1135532, fp=263743, tn=251160, fn=146411): Sn = 0.8858, Sp = 0.8115, AC = 0.4084
Validating on genome5:
Cs (tp=928852, fp=142151, tn=442271, fn=110668): Sn = 0.8935, Sp = 0.8673, AC = 0.6587
Rs (tp=824584, fp=236489, tn=472544, fn=80395): Sn = 0.9112, Sp = 0.7771, AC = 0.6047
Both (tp=1753436, fp=378640, tn=361876, fn=191063): Sn = 0.9017, Sp = 0.8224, AC = 0.4336

5 Model used on unknown structures

Finally the model is trained on all 5 known genomes and then used to predict 5 unknown genomes. The results are seen below.

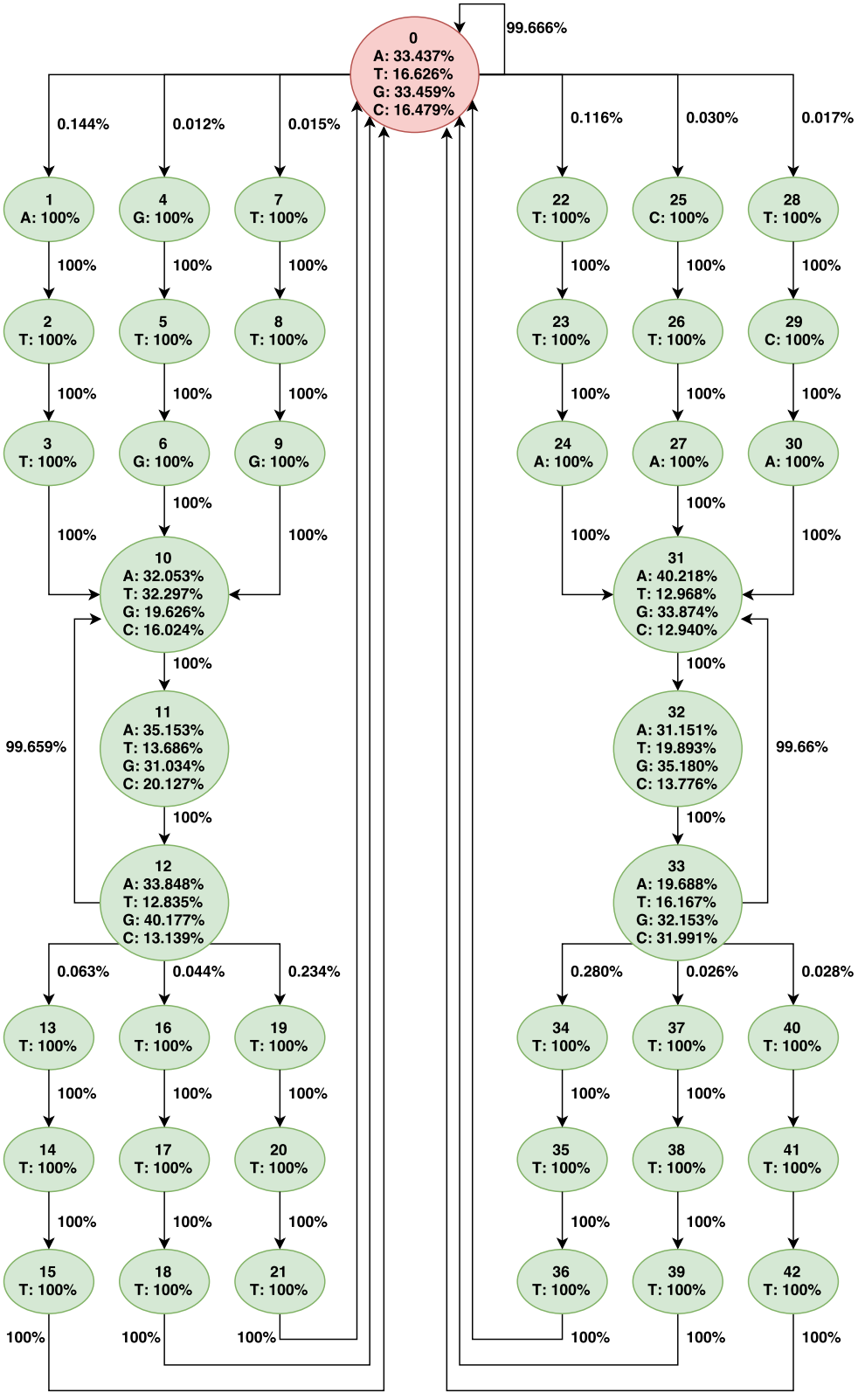
0
A: 33.437%
T: 16.626%
G: 33.459%
C: 16.479%
99.666%
0.144%
0.012%
0.015%
0.116%
0.030%
0.017%
1
A: 100%
4
G: 100%
7
T: 100%
22
T: 100%
25
C: 100%
28
T: 100%
2
T: 100%
5
T: 100%
8
T: 100%
23
T: 100%
26
T: 100%
29
C: 100%
3
T: 100%
6
G: 100%
9
G: 100%
24
A: 100%
27
A: 100%
30
A: 100%
10
A: 32.053%
T: 32.297%
G: 19.626%
C: 16.024%
31
A: 40.218%
T: 12.968%
G: 33.874%
C: 12.940%
11
A: 35.153%
T: 13.686%
G: 31.034%
C: 20.127%
32
A: 31.151%
T: 19.893%
G: 35.180%
C: 13.776%
99.659%
99.66%
12
A: 33.848%
T: 12.835%
G: 40.177%
C: 13.139%
33
A: 19.688%
T: 16.167%
G: 32.153%
C: 31.991%
0.063%
0.044%
0.234%
0.280%
0.026%
0.028%
13
T: 100%
16
T: 100%
19
T: 100%
34
T: 100%
37
T: 100%
40
T: 100%
14
T: 100%
17
T: 100%
20
T: 100%
35
T: 100%
38
T: 100%
41
T: 100%
15
T: 100%
18
T: 100%
21
T: 100%
36
T: 100%
39
T: 100%
42
T: 100%
100%

Figure 5.1: