Team Rap: Rasmus Jørgensen (201909451), Mathias Weller (201907948) & Anna Blume (201907691)

Machine Learning assignment 3

- The status of the work.

It works.

- An explanation and illustration of your model structure (i.e. a transition diagram). This should include an explanation of how you model start and stop codons, and an explanation of how you have trained your model.

We have used the seven-state model from the practical exercises. Our model forces the number of symbols to be a multiple of 3, by having states with 'forced' transmissions to the next state. We do not model specific start- and stop-codons. We trained our model with training by counting with 5-fold cross validation. This means that we set aside a fifth of the data at a time, train on the remaining data and validate on the data we set aside. This is done 5 times, and we choose the model with the best accuracy on the validation data.
We translated the meta-states (C, N and R) like this: every N is translated to state 3, our only non-coding state. The C's and the R's always come in multiples of 3, so if we see a C and have just been in state 2, we know that we are in state 1, etc. If we go directly from C to R, we know that we go to state 4, and state 2 for the reverse case. Only one of these occur in the model (from 0 to 4).

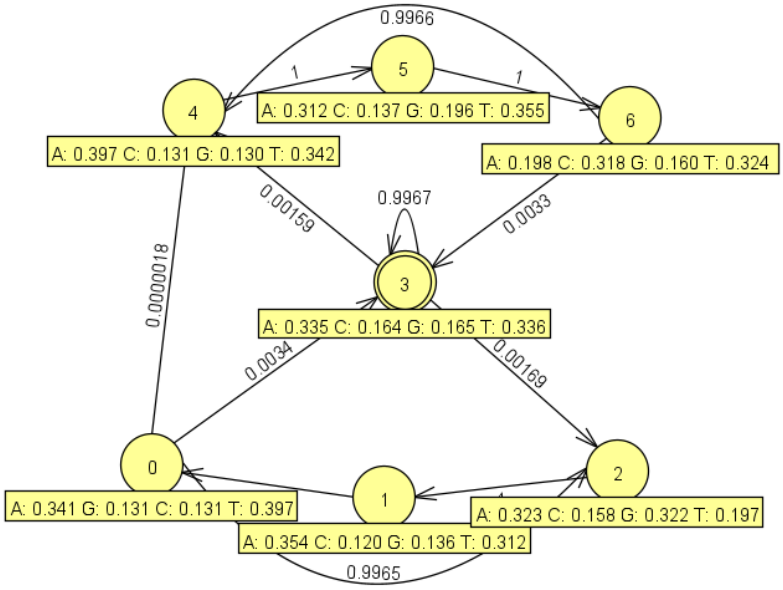
Transition probabilities (from state k_i to state k_j):

	0	1	2	3	4	5	6
0	0	0	0.9965	0.0035	2E-06	0	0
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	0.0017	0.9967	0.0016	0	0
4	0	0	0	0	0	1	0
5	0	0	0	0	0	0	1
6	0	0	0	0.0034	0.9966	0	0

Emission probabilities (from state k_i to symbol x_j):

	A	C	G	T
0	0.3407	0.1315	0.1309	0.3969
1	0.3536	0.1989	0.1361	0.3115
2	0.3228	0.1579	0.3219	0.1974
3	0.335	0.1644	0.165	0.3355
4	0.3971	0.1306	0.1303	0.342
5	0.3122	0.1369	0.1961	0.3548
6	0.198	0.3184	0.1598	0.3238

Team Rap: Rasmus Jørgensen (201909451), Mathias Weller (201907948) & Anna Blume (201907691)



- An explanation of how you have predicted the gene structure for the 5 genomes with unknown gene structure. You should comment on how you translate a most likely sequence of hidden states as returned by Viterbi decoding into a sequence of C, N, and R's.

We predicted the gene structure by using a Viterbi decoding on the sequence, which gives us the most likely sequence of states in the model. Then we translate the states to meta-states like this: state 0, 1 and 2 are coding, state 4, 5, and 6 are reverse coding, and state 3 is non-coding, so those states can be translated directly into C’s, R’s and N’s respectively.

- The result of your 5-fold cross validation on the 5 genomes with known gene structure.

Validation set	Accuracy
1	0.7579895931908223
2	0.7799040011575932
3	0.7598675458002088
4	0.7437142637710744
5	0.7549037901091801

- The result of comparing your predictions on the 5 genomes with unknown gene structure against their true structures via the www-service GeneFinder Verifer.

Results

```
Genome 6
Cs   (tp=731750, fp=170152, tn=319028, fn=82398): Sn = 0.8988, Sp = 0.8113, AC = 0.5785
Rs   (tp=690828, fp=133683, tn=317890, fn=83536): Sn = 0.8921, Sp = 0.8379, AC = 0.6129
Both (tp=1422578, fp=303835, tn=235492, fn=165934): Sn = 0.8955, Sp = 0.8240, AC = 0.3714
Genome 7
Cs   (tp=837583, fp=245135, tn=539222, fn=111762): Sn = 0.8823, Sp = 0.7736, AC = 0.5858
Rs   (tp=777600, fp=231229, tn=530174, fn=120810): Sn = 0.8655, Sp = 0.7708, AC = 0.5735
Both (tp=1615183, fp=476364, tn=418412, fn=232572): Sn = 0.8741, Sp = 0.7722, AC = 0.3784
Genome 8
Cs   (tp=685073, fp=146695, tn=357556, fn=98310): Sn = 0.8745, Sp = 0.8236, AC = 0.5958
Rs   (tp=588801, fp=169680, tn=362704, fn=93162): Sn = 0.8634, Sp = 0.7763, AC = 0.5583
Both (tp=1273874, fp=316375, tn=264394, fn=191472): Sn = 0.8693, Sp = 0.8011, AC = 0.3528
Genome 9
Cs   (tp=760708, fp=210203, tn=352998, fn=110280): Sn = 0.8734, Sp = 0.7835, AC = 0.5228
Rs   (tp=733524, fp=220722, tn=340833, fn=122445): Sn = 0.8570, Sp = 0.7687, AC = 0.4841
Both (tp=1494232, fp=430925, tn=230553, fn=232725): Sn = 0.8652, Sp = 0.7762, AC = 0.2438
Genome 10
Cs   (tp=593796, fp=112914, tn=246331, fn=110309): Sn = 0.8433, Sp = 0.8402, AC = 0.5300
Rs   (tp=361895, fp=145240, tn=293376, fn=63264): Sn = 0.8512, Sp = 0.7136, AC = 0.5281
Both (tp=955691, fp=258154, tn=183067, fn=173573): Sn = 0.8463, Sp = 0.7873, AC = 0.2809
```

The average is 0.32546 over the 5 genomes.