Hand-in 2

Mads-Peter Verner Christiansen, au616397
Zeyuan Tang, au597881
Douwe Tjeerd Schotanus, au600876

November 20, 2018

1 Status

Our 7-state model with training by counting works.

2 Model Structure

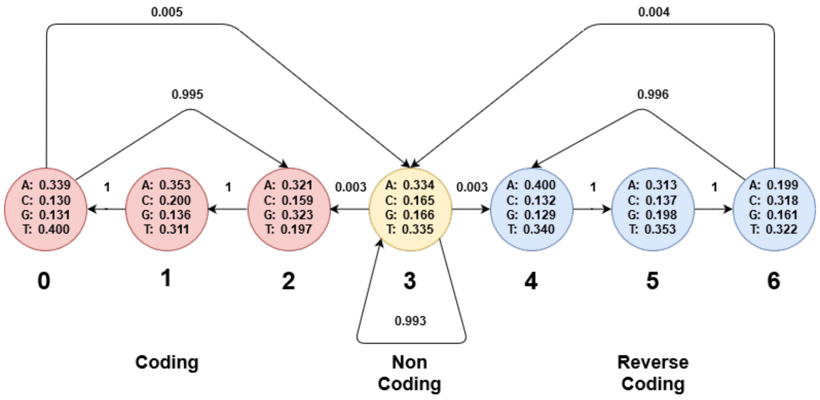
The 7 state model has been used, which does not take into account specific start and stop codons, only that genes should consists of triplets. The transmission and emission probabilities of the trained model are shown in the figure.

3 Cross validation

The 7-state model has been 5 fold cross validated on the genome1-5 datasets. The resulting approximation correlation coefficients are shown in Table 1.

	Cs	Rs	AC
1	0.5253	0.6136	0.3567
2	0.5728	0.6043	0.3706
3	0.6177	0.6069	0.4164
4	0.5659	0.5688	0.3432
5	0.6244	0.5548	0.3735

Table 1: Cross-validation table



Probabilities might not add up to 1 due to decimal precision

Figure 1: Transmission diagram showing the various transmission between hidden states and the emission probabilities of each state.

4 Gene structure prediction

The unannotated genes (6-10) were predicted using a model trained on all of the 5 gene/annotation sets with training by counting. The performance of the model is summarized in the table below.

	Cs	Rs	AC
6	0.5692	0.6066	0.3728
7	0.5804	0.5677	0.3797
8	0.5819	0.5504	0.3491
9	0.5045	0.4703	0.2422
10	0.5260	0.5159	0.2818
Avg	0.5523	0.5421	0.3251

Table 2: Prediction

The model achieves the expected accuracy compared to previous years results shown in class, but obviously performs worse than models that take more details of the sequences into account, such as specific start/stop triplets.