Machine Learning E16 - Handin 3
Hidden Markov Model for Gene Finding in Prokaryotes

Group 22
Mark Medum Bundgaard, Morten Jensen, Martin Sand Nielsen
December 5, 2016, Aarhus University

Model

Our hidden markov model for annotating genome coding areas is depicted in Figure 1. Since the genome can contain both forward and reverse regions, the model has two distinct branches. One could have opted for creating just decoding of the forward coding area, and then perform the algorithm again on the end-wise reversed and AT CG switched string. The two branches allow for small deviations in transition and emission probabilities, even though they would be expected to have same probability distribution.

All coding states in the model, only allow for three letter *codon* outputs. This could be approximated by expanding every such codon emitting state to three letter emitting states. To only allow some codons, a combination af many three letter emitting states configurations in parallel could have been used. Such a model, that could express all those probabilities derived from counting occurrences would have to have nearly 500 states.

To avoid such many states, we opted for this simple 7 state model. To complexity was then shifted towards the implementation of the viterbi algorithm, since it would need to accommodate variable length outputs.

Approximating probabilities by counting occurrences

Emission- and transitions probabilities was approximated by counting every occurrence of emission output or state change. The 68 possible emissions span a set of the single letters 'A', 'C', 'G', 'T' for noncoding, and the 64 three letter codon combinations hereof for coding regions. The states representing the start and stop codons seem to be from a very limited set of codons. The prior probabilities of starting in a certain state, π, was set to start in the noncoding, since nothing else has occurred in the dataset given.

Decoding genomes

The Viterbi algorithm has been tried implemted for decoding a genome with our HMM described above. But it seems that the probabilities for being in some conding states allways are lower than just staying in the noncoding state.

0.1 Performance

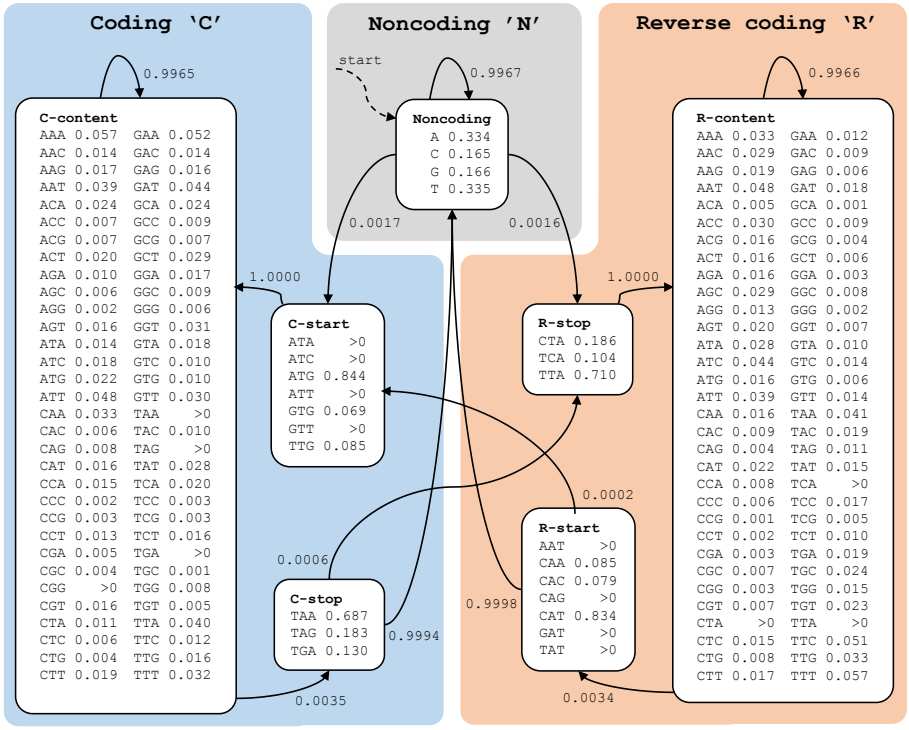

Figure 1: *Transition diagram and state emissions for our HMM for genome annotation by explicitly matching codons within coding areas, and enforcing certain start and stop codons. Transition and emission probabilities approximated by counting annotated data.*

Table 1: *text*

	Approximate correlation coefficient		
validation	C	R	Both
Genome 1			
Genome 2			
Genome 3			
Genome 4			
Genome 5			
Average			