Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Dr. Pedro Feijão

Algorithms in Genome Research
Winter 2015/2016

Exercises

Number 7, Discussion: 2016 January 08

1. Given the following multiple DNA sequence alignment:

```
G A T T - - A T
C A T A G C A T
C A A G G C T A
G A A - - C - T
```

   (a) Use the alignment to construct a PSSM without Pseudocounts.

   (b) Same as in (a), but add Pseudocounts of 0.25 for each base.

   (c) Find the most probable match(es) to the genomic sequence fragment

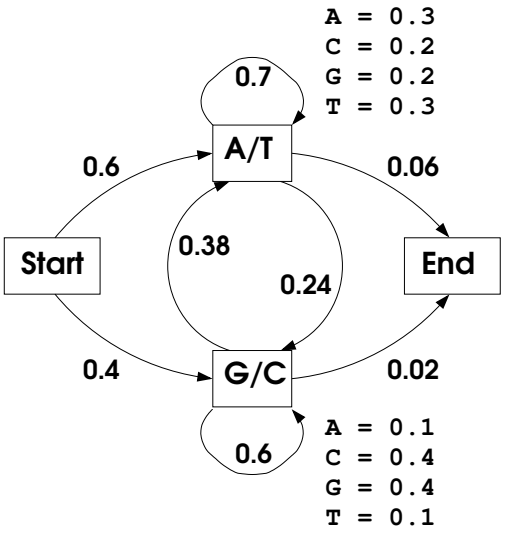
```
GCTGAAGATCATTGTCAAGT
```

   using your PSSMs from (a) and (b).

2. Given the following HMM and an emitted sequence `GAT`, compute

   (a) the overall probability to observe the emitted sequence.

   (b) the most probable sequence of hidden states.



3. How does an interpolated Markov model differ from a hidden Markov model?
   How does an interpolated context model differ from an interpolated Markov model?

4. What is the mRNA-Codon used in the incorporation of Selenocysteine? Does this codon also code for something else? How is this possible?

5. What are the main issues why eukaryotic gene finding is more difficult than prokaryotic gene finding?

   What are the algorithmic/computational techniques that are used in order to account for this increased difficulty?

6. Discuss the issue of obtaining test data sets for HMM training in gene finding.

7. What kind of signals does one typically expect in non-protein coding DNA regions?

   Are there differences between prokaryotes and eukaryotes?

8. What is the application scenario of phylogenetic footprinting?