Bioinformatics Supervision

–Problem Sheet 3–

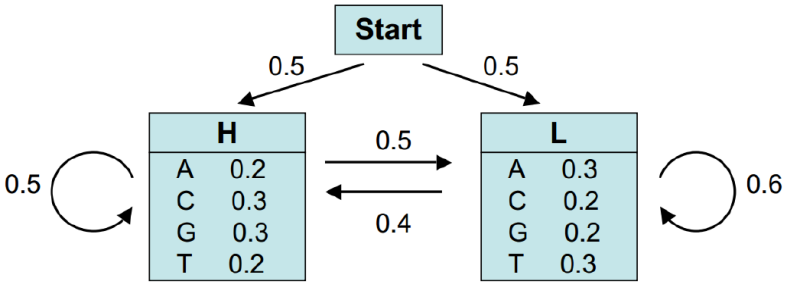
Supervisor: Sebastian Müller (Department of Plant Sciences)

Please hand in your work 24 hours prior to the supervision either to `sm934@cam.ac.uk` or at the Plant Scienes Department reception (make sure my name is on it). Feel free to team up with other group members, the main aim is to understand the material.

1. Discuss the properties of the Markov clustering algorithm and the differences with respect to the k-means and hierarchical clustering algorithms.

2. How can you evaluate the results obtained? Describe one external as well as internal validity index of your choice (this paper might get you started: `http://www.universitypress.org.uk/journals/cc/20-463.pdf`). What do you think are their limitations?

3. Consider the following Hidden Markov Model. In this approximation, H represents DNA coding segments (high GC content), whereas L represents DNA non-coding segments (low GC content).



4. Use Forward Algorithm to find the probability sequence GGCA was generated by this model.

5. Use the Viterbi Algorithm to determine the most probable path through the model for the sequence GGCACTGAA (hint: you could also use log-probabilities).

6. Describe how you would build a hidden Markov model (HMM) to identify membrane segments in amino-acid sequences. How you would assess the sensitivity and specificity of your HMM?

7. For de Brujin graphs, why are we assigning the k-mers of a string to edges instead of the nodes?

8. Given the following string s = "ATTACGGTACCCCTACA".

   (a) Construct the de Brujin graph with $k = 3$ for string s.

   (b) Construct a paired de Brujin graphs with $k = 3$ and distance $d = 1$ for string s.

(c) Find all eulerian paths for the graphs in 8a and 8b. What do you notice?

(d) Find all contigs for the graphs in 8a and 8b. What do you notice?

9. Describe briefly Viterbi learning and Baum-Welch learning. What is the main difference?