Bioinformatics Supervision

Michaelmas Term 2017

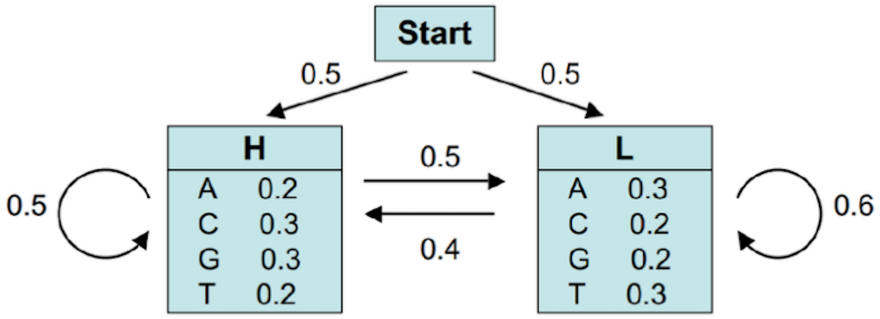
–Problem Sheet 3–

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 Sciences 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. If you hand in electronically, please name the file `group<x>_<crsid>_problemsheet<x>.<x>`

Hidden Markov Models

1. 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).



   (a) Use Forward Algorithm to find the probability sequence GGCA was generated by this model.

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

2. 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?

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

Genome assembly

4. For de Brujin graphs, why are we assigning the k-mers of a string to edges instead of the nodes?
5. 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 5a and 5b. What do you notice?
   (d) Find all contigs for the graphs in 5a and 5b. What do you notice?

Clustering

6. Explain the two high-level steps taken by the expectation-maximisation (EM) algorithm, and then show how it relates to soft k-means clustering (giving particular care to the stiffness parameter).
7. What is the output of a typical gene expression experiment, and why might one wish to do further processing on such a result?
8. Discuss the properties of the Markov clustering algorithm and the differences with respect to the k-means and hierarchical clustering algorithms.
9. 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?

Optional tasks

Feel free to complete as much as you feel confident.

One of the most popular programming language in bioinformatics is R (`https://cran.r-project.org/`).

Try to run and understand the following code:

```
head(iris) #inspecting iris dataset
str(iris)  #still inspecting
?help(iris)
dd <- dist(scale(iris[,-5]), method = "euclidean") #creating distance
    matrix
hc <- hclust(dd, method = "average") #UPGMA clustering
plot(hc)
```

10. Familiarise yourself with the iris dataset (see code above) as well as distance matrices (`help(dist)`), could you think of any situations the euclidean distance is not appropriate?
11. Familiarise yourself with the hierarchical clustering function (help(hclust)).
12. Familiarise yourself with *cutree* (`help(cutree)`). Whats the point of this function?

(a) Investigate the 3 cluster solution `(table(cutree(hc,3)))`.

(b) Compare the 3 cluster solution with the actual classification (stored column 5).
`table(cutree(hc,3), iris[,5])`.

(c) How do you interpret this table? Is this result expected?

(d) Would it improve if you took the 4 cluster solution instead? How about using the default Ward clustering instead UPGMA?

13. Try to write code to compute the euclidean distances yourself

14. Very optional: try to write code to perform UPGMA clustering yourself and run it on the distance matrix dd.