# Bioinformatics Supervision

## –Problem Sheet 2–

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.

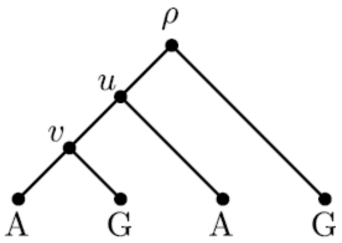
1. Considerable recent Bioinformatics research has focused on phylogenetics. What is the motivation for this work?

2. Describe the differences in complexity and performance between parsimony and two distance phylogenetic methods. Also try to find another tree construction method not mentioned in the lecture and describe it conceptually.

3. Commonly used methods for traversing a binary tree include pre-order, in-order, and post-order. Suppose we need to implement the SmallParsimony algorithm using one of these traversal methods. Which one(s) would be suitable for our implementation? Explain your choice.

4. How is the score matrix used in phylogenetic tree building techniques? Hint: A prominent example is the PAM score matrix described on page 285 of Vol.I (Compeau and Pevzner 2nd Edition).

5. You are given the tree below with single-letter sequences at its leaves. Use the SmallParsimony algorithm to find the minimum parsimony cost for the given tree and identify the optimal state assignments for each node with c(A; G) = 1 and c(A; A) = c(G; G) = 0.




6. Given the following distance matrix, calculate an evolutionary tree using UPGMA:

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | | | | | |
| B | 2 | 0 | | | | |
| C | 4 | 4 | 0 | | | |
| D | 6 | 6 | 6 | 0 | | |
| E | 6 | 6 | 6 | 4 | 0 | |
| F | 8 | 8 | 8 | 8 | 8 | 0 |

7. Given the following distance matrix, calculate an evolutionary tree using neighbour joining:

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | 5 | 0 | | | |
| C | 4 | 7 | 0 | | |
| D | 7 | 10 | 7 | 0 | |
| E | 6 | 9 | 6 | 5 | 0 |

**Optional task** Feel free to complete as much as you feel confident. One of the most popular programming langage 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)
```

1. Familiarise yourself with the iris dataset (see code above) as well as distance metrices (help(dist)), could you think of any situations the euclidean distance is not appropriate?
2. Familiarise yourself with the hirachical clustering function (help(hclust)).
3. 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?
4. Try to write code to compute the euclidean distances yourself
5. Very optional: try to write code to perform UPGMA clustering yourself and run it on the distance matrix dd.