**1) Theory:** Explain ‘dimensionality reduction’ in the context of PCA (possibly illustrate your explanation with a drawing, 10-20 lines)

**2) True or false**. Read the questions carefully. Motivate your answer (try to be concise and complete). The statement you have to motivate is given in italics.

a) *Hierarchical clustering with a correlation-derived distance metric applied on non-rescaled data x (data not mean centered nor variance rescaled) gives exactly the same results as hierarchical clustering with a Euclidian distance applied on the same non-rescaled data.*

or

d.correlation <- as.dist(1 - cor(x)))

d.euclidian <- dist(x, method="euclidian")

hclust(d.correlation, method = ‘complete’)

hclust(d.euclidian, method = ‘complete’)

b) Given a cancer dataset, containing the expression of 3051 genes in 38 patients. The dataset was rescaled and the Euclidean distance was calculated as follows

golub\_m = golub – rowMeans(golub)

SD = apply(golub, 1, sd, na.rm=TRUE)

golub\_r = golub\_m/SD

d.euclidian <- dist(golub\_r, method="euclidian")

#as.matrix(dist(golub\_r, method="euclidian"))

m <- as.matrix(d.euclidian)

dim(m)

[1] 3051 3051

kmeangolub10= kmeans(d.euclidian,10)

*Kmeans clustering is executed on the matrix golub to group together patients based on the similarity of their (rescaled) expression profiles. 10 patient groups are generated in total.*