Tampere University of Technology
Department of Signal Processing
SGN-6176 Microarray Data Analysis / J. Kesseli

Exam 1.10.2012

No calculator allowed.

Write to each of your answering sheets
- course name, date, number of the sheet / total number of the sheets
- family name, given name, student id, signature

You can answer either in English or in Finnish. You are permitted to take this problem sheet with you when you leave the exam.

1. a) How does TopHat work? Explain briefly the sequence of operations. (4p)

b) For what purpose can multidimensional scaling be used? How is it different from principal component analysis (PCA)? (2p)

2. a) What is plotted in an MA plot? (2p)

b) In RNA-Seq data, read counts are commonly modeled with a Poisson distribution with \( \lambda = Np \). Why is this model intuitively attractive? What do \( N \) and \( p \) represent? (3p)

c) Quantile normalize the following data matrix. Each column corresponds to a sample and each row to a gene. (3p)

\[
\begin{bmatrix}
3 & 2 & 6 \\
5 & 3 & 2 \\
1 & 4 & 3
\end{bmatrix}
\]

5. What new information is ChIP-Seq giving in large quantities due to developments in sequencing technology? How does this help in understanding the functionality of the genes? (5p)
3. a) A gene is tested for differential expression between two classes of samples, and a p-value 0.0045 is obtained from the test. What does this value mean, i.e. what is the definition of a p-value? (2p)

b) Given that you have tested N genes for differential expression and obtained p-values \( p_1, p_2, \ldots, p_N \) with the histogram of p-values as shown in Fig. 1, how can you estimate False Discovery Rate (FDR) for a given significance threshold \( t \)? Explain the method in brief and draw a schematic figure to represent the feature or features of interest for the method in Fig. 1. (4p)

4. Explain briefly the following:
   a) Paired-end read in sequencing (2p)
   b) Color-space read (1p)
   c) Consensus clustering (2p)
   d) Position Weight Matrix (PWM) (2p)

5. What new information is ChIP-Seq giving in large quantities due to developments in sequencing technology? How does this help in understanding the functionality of the genome? (5p)