T-61.5050 High-throughput bioinformatics

Exam 19.12.2014

There are 6 questions (remember to turn the page). You can reach 30 points in total, 5 points for each question.

Question 1 (5p): Term Definitions

Define the following concepts (1-2 phrases for each concept):

a) Interactomics (1p)
b) SNP (1p)
c) Volcano plot (1p)
d) p-value (1p)
e) Alternative splicing (1p)

Question 2 (5p): Sequencing

a) Briefly describe the difference between synchronous and asynchronous extension in sequencing? Name one example technology for each! (2p)
b) Explain what are paired-reads and what are they used for. (1p)
c) Briefly describe and compare de-novo sequencing and resequencing approaches. (2p)

Question 3 (5p): Burrows-Wheeler transform

a) Construct the suffix array and the Burrows-Wheeler transform for the string "CCATAGAT$". Describe the construction procedure in 1-2 sentences. (2p)
b) Construct the original sequence, knowing that its Burrows-Wheeler transform is "R$KEPKOOBEE" (show and explain the intermediary steps of this back-transformation). (2p)
c) For which bioinformatics task is the Burrows-Wheeler transform used and what is the benefit from using it? (1p)

Question 4 (5p): Transcriptomics

a) Name 2 of the main approaches that are nowadays used to measure gene expression on the transcriptomic level. What kind of information do you get from the measurement and how do the approaches differ in that respect? (2p)
b) Choose one of the 2 approaches named in part (a) and list the general steps in its pipeline from biological samples to gene expression values. (Focus on the description of the main steps and use relevant keywords, rather than going into details.) (2p)
c) Describe the quantile normalization and its purpose. (1p)

Question 5 (5p): Enrichment analysis

a) Define the concept "enrichment of a gene set S in a list of genes L". (1p)
b) Name and briefly describe 2 resources that can be used for assigning genes to biological processes. (1p)

c) Name and define the three types of ontologies provided by GO. (1p)

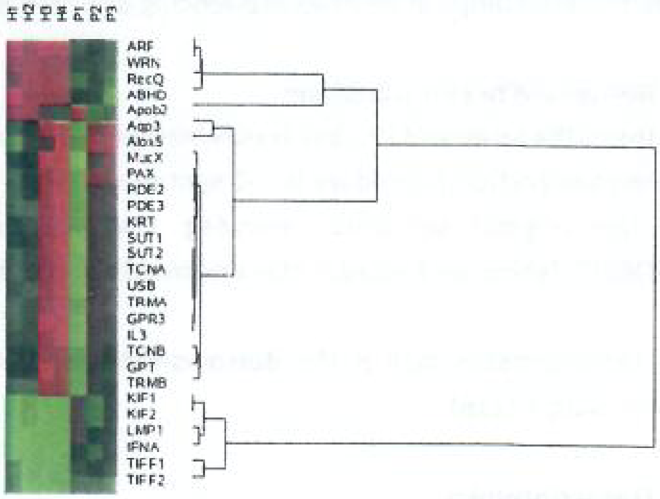
d) Choose one of the main approaches for performing enrichment analysis and describe it in details. What statistical test is used for deciding the significance of the enrichment? (Focus on the description of the main steps and use relevant keywords.) (2p)

Question 6 (5p): Learning methods

a) Werner syndrome (WS) is a premature aging disease that begins in adolescence or early adulthood and results in the appearance of old age by 30-40 years of age. It is known that the Werner syndrome is caused by defects in the WRN gene. We search for other gene(s) that might be responsible for WS using gene expression data. We have cells available from four healthy controls (denoted as H1, H2, H3, H4) and three patients with WS (denoted as P1, P2, P3) and we measure the gene expression in G1 phase of the cell cycle using micro arrays. The figure shows a clustering of absolute gene expression from 28 genes from healthy individuals and WS patients. Red colors indicate high gene expression, black close-to-average gene expression and green low gene expression. Assume that all patients (P1-P3) have the same defect in the promoter of WRN gene.

i. What is the effect of this defect in the WRN expression? (1p)

ii. According to this clustering of gene expression, which gene(s) other than WRN can be responsible for Werner syndrome? Explain in one sentence. (1p)



b) What is the motivation for biclustering and how can it help in analyzing gene expression data? (1p)

c) Use complete and average linkage agglomerative clustering to group the data described by the following distance matrix. Show the 2 dendrograms you obtain and the intermediary steps in obtaining them (2p).

	A	B	C	D
A	0			
B	1	0		
C	5	2	0	
D	5	6	4	0