T-61.5050 High-throughput bioinformatics

Exam 12.12.2014

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) Copy number variation (1p)
b) Transcriptomics (1p)
c) Boxplot (1p)
d) Genome assembly (1p)
e) p-value (1p)

Question 2 (5p): Sequencing

a) Describe the difference between synchronous and asynchronous extension in sequencing? Name one example technology for each! (2p)
b) Briefly define the following terms: contig, scaffold, coverage. (1p)
c) Briefly 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 "CGAAGCAT$". Describe the construction procedure in 1-2 sentences. (2p)
b) Construct the original sequence, knowing that its Burrows-Wheeler transform is "IPSSM$PISSII" (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 types of approaches that are nowadays used to measure gene expression on the transcriptomic level. Then, briefly compare these 2 approaches with each other, mentioning the main advantages and drawbacks. (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) Explain the difference between biological replicates and technical replicates, as well as their purposes! (1p)

Question 5 (5p): Enrichment analysis

a) Define the concept "enrichment of a gene set S within a list of genes L". (1p)

b) Name and briefly state the differences between the 3 main approaches for enrichment analysis? (2-3 phrases for each method) (3p)

c) Choose one of the 3 main approaches for performing enrichment analysis and describe it in more details. (Focus on the description of the main steps and use relevant keywords.) (1p)

Question 6 (5p): Learning methods

a) Suppose you are given a gene expression dataset containing samples from 1000 patients with a specific disease. Knowing there are several subtypes of that particular disease, describe a possible approach to obtain and visualize an appropriate clustering of the genes (distance measure, name of the algorithm, setup of your computational experiment). Briefly motivate your choices. How do you evaluate the obtained clusters of genes? (Note: You do not have to explain how the basic algorithms work; assume you use ready-made programming functions for those!) (2p)

b) How does unsupervised learning differ from supervised learning? (1p)

c) Use single and complete linkage agglomerative clustering to group the data described by the following distance matrix. Show the 2 dendrograms you obtain (2p).

	A	B	C	D
A	0	1	4	5
B	1	0	2	6
C	4	2	0	3
D	5	6	3	0