T-61.5050 High-throughput bioinformatics **Exam 11.10.2008**
Nikkilä, Auvinen, Lahti

Note that all questions do not have a unique correct answer. Be brave, but explain and justify your answers well. If content is equivalent, briefer explanations give better grades. Every question is 6 points, resulting in total of 30 points.

Remeber to give feedback of the course at
`http://www.cis.hut.fi/teaching/Kurssipalaute/kurssipalaute.shtml`.

Question 1

Define the following concepts briefly (max. 5 lines per concept, max. 1 point per concept):

a) Affinity differences between probes

b) Alternative splicing

c) Enhancer

d) Operon

e) n<<p problem in high-throughput data

f) Bonferroni correction

Question 2

You are given quality controlled and normalized gene expression data from 30 rats measured with Affymetrix arrays with about 32000 probe sets. 10 rats are of wild type strain, and 20 are of a mutated strain, which is a disease model for rheumathoid arthritis. 10 out of the 20 mutated rats are treated with a new drug for artrithis for one month (starting at age of on month), and the other 10 are treated with the same drug for 5 months (starting at age of one month). Blood samples are collected from all the mice at ages 1 month, 2 months, and 6 months. Explain and justify what methods you would use to analyze which genes are statistically significantly affected by the new drug after 1 month treatment.(6p)

Question 3

a) Describe lowess normalization. What assumption(s) does the method make? (3p)

b) Explain the ideas and motivation behind the method and the computational procedure of GC-RMA preprocessing (what it is, why it works, what are the stages, what do the symbols in formulas represent, etc.), and compare it to the cDNA microarray preprocessing, explaining whether it would be a good idea to use GC-RMA also for cDNA arrays. You can use the background information given about GC-RMA in the Appendix. Use at most 1 page. (3p)

Question 4

Regulation of gene activity.

Question 5

The aim of a research project is to study the potential subtypes of prostate cancer, and their connection to Gene Ontology classes and blood metabolism. You have available a laboratory capable of producing custom cDNA arrays, LC-MS equipment, samples from 20 healthy patients' prostata and blood, and samples from 40 cancer patients' prostata and blood. Describe (and justify) in resonable detail procedures and computational methods you would use to accomplish the aim of the study.

A APPENDIX

GCRMA background

The perfect match and mismatch values in GC-RMA are assumed to follow the model

$$\begin{aligned} PM &= O_{PM} + N_{PM} + S \\ MM &= O_{MM} + N_{MM} + \phi S, \end{aligned}$$

where O follows a log-normal distribution and that $\log(N_{PM})$ and $\log(N_{MM})$ follow a bivariate-normal distribution with means of μ_{PM} and μ_{MM} and the variance $var[\log(N_{PM}] = var[\log(N_{MM})] \equiv \sigma_2$ and correlation ρ constant across probes. We assume $\mu_{PM} \equiv h(\alpha_{PM})$ and $\mu_{MM} \equiv h(\alpha_{MM})$, with h a smooth (almost linear) function. The αs are defined as

$$\alpha = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k - j}, \text{with } \mu_{j,k} = \sum_{l=0}^{3} \beta_{j,l} k^l,$$

where $k = 1, ..., 25$ indicates the position along the probe, j indicates the base letter, b_k represents the base at position k, $1_{b_k=j}$ is an indicator function that is 1 when the k-th base is of type j and 0 otherwise, and $\mu_{j,k}$ represents the contribution to affinity of base j in position k. For fixed j, the effect $\mu_{j,k}$ is assumed to be a polynomial of degree 3.

To obtain an expression measure it is assumed that for each probe set n, the background-adjusted, normalized, and log-transformed PM and MM intensities, denoted with Y, follow a linear additive model

$$Y_{i,j,n} = \mu_{i,n} + \sigma_{j,n} + \epsilon_{i,j,n},$$

where $i = 1, ..., I,\ j = 1, ..., J,\ n = 1, ..., N.$ with σ_j a probe affinity effect, μ_i representing the log scale expression level for array i , and ϵ_{ij} representing an independent identically distributed error term with mean 0.