Home assignment (April 26th, 2007)

1 Requirements

You will need to complete (and pass) this homework assignment in order to pass the course. This assignment will be graded (scale 0-5). Those who will pass with a grade 4 or 5 will get 1 or 2 extra points to the exam. This may lead to a higher overall grade from the course.

The deadline for returning the homework assignment is Sunday the 27th of May. The deadline is strict. Next opportunity for returning the homework assignment will be in the Autumn on the date of the next exam.

To complete this assignment you will need to complete the tasks listed below and to write a report. The report should be written in the form of a tiny research article. Please use the formating template provided by the journal Bioinformatics.[1] The report should not exceed 6 pages in this format.

Hand in the code used to complete the assignment as a separate attachment. You may use any programming language, but we recommend using R and the Bioconductor. They provide many ready made functions for microarray data preprocessing and analysis.

2 Introduction

This assignment will lead you through all the major steps in a typical gene expression data analysis. The main focus of this work is for you to understand why each subtask is done and what can be achieved by it. By writing a good report you can show that you understood what you did and were able interpret the results from each step of the analysis. Programming is not supposed to be a major part of the assignment: you can reuse R-code fragments from the answers to exercises in many subtasks of this assignment.

2.1 Data set

If you have a gene expression data set of your own, it is recommended that you use that. Contact the lecturer and the assistant to ensure the applicability of the data set for this assigment.

If you do not have a data set of your own, you can use the public data set provided by the course. The data set is a gene expression data set measured with Affymetrix arrays (Singh et al., 2002). Data consists of several human tissue samples from both normal and prostata cancer tumors. More information can be found from (Singh et al. 2002). You can access the paper and the data set through the links in the course web-site.

[1] Word and latex templates can be found from
http://www.oxfordjournals.org/bioinformatics/for_authors/submission_online.html

3 Tasks

1. *Download the data set*

2. *Define the study question* What is compared to what? How many samples do you have?

3. *Preprocess the microarray data* Preprocess the microarray data using appropriate preprocessing techniques. Justify the choice of the preprocessing method on technical level, and use also some other alternative normalization method for the data.

4. *Quality control* Visualize the intensity distributions (histogram, boxplot, MA-plot) of the arrays before and after normalization. Check and plot also quality measures for Affymetrix arrays: RNA degradation, NUSE, RLE, intensity plots for the arrays with dubious quality values. Is everything ok? Discard the samples with bad quality.

5. *Find differentially expressed genes* Find a set of genes that are significantly differentially expressed between the conditions, and also have a reasonable change in the log-ratio. Visualize the genes with a volcano plot. How many differentially expressed genes do you find? Remember to use some p-value correction method. Describe the chosen test and correction method, and justify the choices.

6. *Visualization of samples* Visualize the samples (arrays) using principal component analysis (PCA) and multidimensional scaling (MDS). Color the samples according to their classes. Interpret the PCA results (gene loadings on the few first principal components). Which genes have high loadings? Are these genes differentially expressed?

7. *Clustering the samples and the differentially expressed genes* Cluster both the samples and the set of differentially expressed genes with an appropriate clustering method. Visualize the clusterings. Describe the clustering method, and justify the choice of method. What kinds of clusters do you see? Are there any natural clusters in the data?

8. *Interpretation of a cluster of genes* Select a suitable representative number of clusters from your clustering. Study these clusters with respect to Gene Ontology (GO). Are some GO categories enriched in your clusters? Use Fisher's exact test or some other statistical test to check the enrichment. Describe the test.

9. *Classification of samples* Use either cross-validation (CV) or bootstrap to test the classification. Learn a classifier on training data and test it with the independent test data (left-out data set in CV, or the samples not in the bootstrap set in bootstrap). Select an appropriate

classification method and feature selection scheme. How good is your classifier? Compare the results to the results presented in the original paper.

10. *Write the report* The report should include the following sections.

    - Introduction
    - Data
    - Methods (a brief list of the methods used in the assignment and reasons why you chose those methods)
    - Results with following subsections (required figures in parentheses)
        - Preprocessing and quality control (2 relevant figures)
        - PCA (visualization of PCA of genes and samples with class coloring)
        - Differentially expressed genes (volcano plot with thresholds)
        - Clusterings (figure of sample clustering, table of most significantly enriched GO classes in clusters)
        - Classification of samples
    - Conclusions
    - References