582612 Practical course in microarray data analysis

Requirements for credits

You will form 5 groups of 3 students each (the assigning to groups will be done during the lecture on 22 Feb). Each student will open a new account in Expression Profiler called Helsinki_*yourname*_Group*n* (where n is 1, 2, …, 5), … , depending on your group. My colleagues and me will be able and to see the contents of these accounts (please don't treat them as private).

Each group will be given to analyse one dataset from ArrayExpress. The analysis will consist on normalization, finding differentially expressed genes, clustering and subsequent analysis. The analysis will be conducted in the Expression Profiler new interface beta version (http://www.ebi.ac.uk/microarray-srv/EPdev/cgi-bin/ep_ui.pl). You will have to comment each major action using the commenting facility in Expression Profiler, such as what is the main observation, or if something has not worked as you expected.

Each group member will replicate the end portion of the analysis using a different normalized dataset – the processed dataset as given in ArrayExpress, and two different normalisations of your own (e.g., RMA and GCRMA) – each student in the group can choose their own.

As the Expression Profiler software you will be using is still under development, there may be occasional problem, which you will report by e-mail to kolais@ebi.ac.uk (and cc to brazma@ebi.ac.uk) with subject 'Helsinki Course'.

Although everybody has to perform the actions in your own account, the students in the groups are encouraged to communicate and to discuss the problems.

In addition to commenting your actions in the Expression Profiler interface you are also asked to write 2-4 page report of the observations and suggestions to improve Expression Profiler interface. This should be submitted to Esa Pitkänen epitkane@cs.helsinki.fi by 25.03.2007.

In summary, to successfully pass the course, you need to

- complete at least 20 work steps from the 25 described below
- comment your actions in Expression Profiler
- return a 2-4 page report where you describe each work step, explain the choices you have made and suggest improvements to the Expression Profiler interface

The datasets for groups 1-5 are

1. E-MEXP-420
2. E-TABM-34
3. E-MEXP-774
4. E-MEXP-438

5. E-MEXP-433

The steps to try out (a modification from lab 3 on February 22, 23):

Obtaining the data

1. Find your experiment E-xxxx-yy in the ArrayExpress Repository. Explore its properties. Explore array annotation (you will see probe names and annotations for the probes). Describe the data set briefly in the final report.
2. Download raw data from this experiment to your PC.
3. Generate a Processed data matrix and save it to you PC. Include
    a. all experimental conditions
    b. normalized signal intensities for all probes in all conditions
    c. some meaningful annotations for the genes

    Comment your choices in the final report.

Data normalization

4. Now go to Expression profiler (log in, unless you've already done so).
5. Upload raw expression data into Expression profiler.
6. Explore the intensity distribution plots and the box plots. What do you see?
7. Normalize the data – explore the expression value distribution plots after the normalisation. Compare these to those from raw data. What has changed? What normalization did you use? Why?
8. Try out also other normalization methods. Explain why you would want to use different normalization methods.
9. Upload the normalised data (processed data from ArrayExpress).
10. Compare the different normalized data sets. What differences are there?

The study question

11. The remaining steps you can do using any of the normalised datasets you have obtained.
12. Define the study question. What is compared to what? Describe the study question in the final report.
13. Based on the study question convert the absolute intensities (single arrays) of the normalized data to ratios (log2-ratios of treatment and control arrays vs. a baseline) to obtain a normalized relative dataset. Describe your relative data set in the final report.
14. Explore the intensity distribution plot after the conversion and compare it to the earlier one (absolute intensities). What has changed? Why?

Differentially expressed genes

15. Use t-test to select the most differentially expressed genes between your treatment and control groups. How many differentially expressed genes are there? Do you observe some general trends (up/down regulation) among the differentially expressed genes?

16. Select some of the differentially expressed genes for closer analysis. Take both up-regulated and down-regulated differentially expressed genes to make the analysis more interesting.
17. Search for the annotations for these genes in ArrayExpress. Can you find some genes that are in Ensembl? What are the functions of these genes? Describe you findings in the final report.
18. Construct a small data set (~ 500 hundred genes) from genes that are similar to your chosen set of differentially expressed genes. (Select by similarity and type in the probset names for several most differentially expressed genes)

Clustering the small data set, Gene Ontology (GO)

19. Cluster the small data set constructed in step 18, using at least two clustering methods (different clustering methods, distance measures, options). Explore the clusters and compare results from different clustering methods. Describe you findings in the final report.
20. Select one interesting cluster and explore it further. For example, see what Gene ontology categories are overrepresented in this cluster. Describe your findings in the final report.

PCA and correspondence analysis on the small data set

21. Try out PCA and correspondance analysis with the small data set constructed in step 18. (Ordination-menu). Describe your findings in the final report.

Between group analysis on the small data set

22. Try our between group analysis for the small data set constructed in step 18.
23. First you will need to define a new factor (Define new factors button) based on your study question – for example a factor with two groups ‘normal/treatment’.
24. Use this new factor for correspondence analysis or PCA. Explore the output – how well the two groups are separated and what are the separating genes? Describe your findings in the final report.

Compare results with other group members

25. Compare your results from steps 12-24 with members of your group. Explain differences in results.

Alvis Brazma, Merja Oja, Esa Pitkänen