

Optimal Design and Analysis of Microarray Experiments

Ernst Wit & John McClure

Department of Statistics, University of Glasgow

Half-day intermediate course for bioinformaticians about statistical ideas and principles involved in the design and analysis of microarray experiments using the R computing environment.

1 Motivation

Gene expression profiling has become a routine technique that can be useful to many applied life scientists in some stage of their research. The ease with which it is possible to generate thousands of gene profiles stands in no comparison to the difficult and often treacherous path of getting reliable conclusions or even interesting research directions from these data. The data are very noisy and many traditional computational techniques are not made to deal with large numbers of variables, i.e. genes. Further, appropriate statistical design is essential to make the best use of the limited number of arrays and samples that are in general available.

2 Goals

Statistical thinking is an important prerequisite in dealing with large numbers of noisy observations. In microarray studies it helps the applied scientist...

- to design optimal or near-optimal microarray experiments;
- to perform simple but effective microarray data cleaning techniques;
- to perform effective hypothesis tests with accurate error rate control.

These are the three main goals of the tutorial

3 Detailed outline of course & learning outcomes

The course is taught in three approximately 1 hour slots, each with a separate theme. The course is conducted via a mix of lecture-style explanations of statistical concepts and hands-on explorations of real microarray data using an R interface. The specific objectives are given in the form of learning outcomes.

After attending the course, the participant should be able to...

Design (70 min)

- understand the difference between technical and biological variation and explain how pooling mRNA samples reduces biological variation of gene expression.
- understand how considerations about optimal design can have a real impact on the practical set-up of microarray experiments.
- design optimal microarray experiments.

Normalization and Quality Control (60 min)

- identify a good scale on which to analyze microarray data and be aware of alternatives.
- have ways to deal with missing data, such as data imputation.
- understand the pros and cons of sequential normalization techniques and the best sequential order in which to do them.
- perform spatial, dye, within- and across-condition normalizations.
- use quality control techniques to assess whether normalizations have properly dealt with data artifacts.

Multiple Testing, Error Rates and Differential Expression (70 min)

- distinguish between different test-statistics for differential expression and understand when each of them is appropriate.
- understand different error rates involved in multiple testing, such as the False Discovery Rate (FDR), the Familywise Error Rate (FWER) and the False Positive Rate (FPR).
- control the FDR, FWER and the FPR by means of hypothesis testing.
- control the FDR using Empirical Bayes methodology.

4 Intended audience, level and prerequisites

The course is of an intermediate level aimed at bioinformaticians and statistically trained biologists. The course is conceptual rather than mathematical, but some use is made of mathematics. Prerequisites are:

- basic familiarity with the microarray platform.
- intermediate level of university statistics (knowledge of standard deviation, standard error, hypothesis testing, ANOVA).
- grasp of matrix manipulations and basic calculus.
- rudimentary familiarity with R computing environment (programming skills are not necessary).

5 Relevant teaching experience of instructors

Dr Ernst Wit is a Reader in Statistics at the University of Glasgow with over 10 years of university teaching experience. He has taught many subjects in the field of statistics to a wide spectrum of audiences, ranging from master classes for secondary school pupils to specialized topics courses for postgraduate students. Currently, he is teaching Statistics for Psychology Students, a 2nd year Probability Distributions course, and a 4th year honours Time Series and Spatial Statistics course. Dr Wit is also coordinating a postgraduate teaching in the Statistics Department at the University of Glasgow. In August 2002 he completed an ILT (Institute for Learning and Teaching in Higher Education) accredited didactic programme at the University of Glasgow. In September 2003 Dr Wit taught a 1.5 day course on microarray data analysis at the regional International Biometrics Society conference, which was repeated in May 2004 in Verona (Italy).

Dr John McClure is a Lecturer in Statistical Genetics at the University of Glasgow. He has worked in the field of bioinformatics for over three years, principally on statistical analysis methods for microarray experiments. Dr McClure has over 5 years experience in conducting “statistics laboratories” for life sciences and other students. At present he is teaching statistics to BSc Medical Science students. The aim of this teaching is to convey basic statistical ideas that are crucial to the experimental design, data handling and analysis these medical students will need in their future research careers. In 2002-2003 he taught a Practical Statistics course (S1B) for Science, Arts and Social Science students in the University of Glasgow. Dr McClure has also taught the statistics module in the Basic Bioinformatics course of the M.Sc. in IT (Bioinformatics Strand) in 2003-2004. He is at present completing an ILT accredited Postgraduate Certificate in Academic Practice.