

1

Preliminaries

1.1 Using the R Computing Environment

The R package is a mathematical computer language and a multifaceted computing environment for statistics and graphics. Although there are several such packages available, R is one of the best and, unlike almost all others, it is *free*. It has a GNU General Public License. All you need to do to use it is to visit the R website (<http://www.r-project.org/>) and download the package from there.

Because of its mathematical diversity and its free availability, R has become a very popular tool for analysing microarrays. It is based on the S language developed at Bell Laboratories and therefore shares many similarities with the commercial package S-PLUS.

Although the methods we discuss in this book can be implemented in other packages, all the plots and analyses in the book were obtained using R. In the book, we give information on how to implement many of the procedures. The website

<http://www.stats.gla.ac.uk/~microarray/book/>

that accompanies the book has more detailed information on how to use each procedure in R, though it assumes a working knowledge of the package. Many of the new procedures programmed by the authors have been bundled together in the *smida* library. This can be downloaded from the book's website or from the R website. Follow the instructions on how to download the latest version of R for your platform from one of the CRAN mirrors (see <http://www.r-project.org/>).

If you are not familiar with R, *Introductory Statistics with R* (Dalgaard 2002) is a useful guide. Another excellent book is *Modern Applied Statistics with S-PLUS* (Venables and Ripley 1999), which gives an overview of many implementations of cutting-edge statistical techniques. Although this is written primarily for S-PLUS, the similarities between the two packages mean that it is equally relevant for R

users. The book's associated MASS library is available in R. Other R implementations of microarray-related algorithms are available from the Bioconductor website (<http://www.bioconductor.org/>).

Note that throughout this book, whenever we refer to R commands or libraries, we use typewriter font, for example, `smida`. All the R functions mentioned in this book are included in the `smida` package, unless explicitly stated.

1.1.1 Installing `smida`

This book not only covers many traditional statistical techniques but also proposes new methodology and algorithms for the analysis of microarray data. Many of these algorithms have been implemented by the authors of this book in R. They have been bundled into an add-on library called `smida`. This library is available from the book's website, and from any of the CRAN mirrors at the R website. The library has to be installed only once on your computer.

For Windows users: the easiest way to install the library in R is as follows:

1. download `smida` as a zip file (ends `.zip`), either from the book website or from one of the CRAN mirrors, and store it in a convenient folder;
2. open R on your computer;
3. click on Packages drop down menu and select the Install package(s) from local zip files... option;
4. search for the folder you chose in the window that appears, double click on `smida.zip` and then click on the Open button.

For Unix and Linux users: An easy way to install the library in R is as follows:

1. download `smida` as a file (ends `.tgz`) file, either from the book website or from one of the CRAN mirrors, and store it in a convenient folder;
2. type

```
R INSTALL smida.tgz
```

in a local client window in the same directory where the `smida` file has been saved.

For more information on how to install packages, see the section on *R Add-On Packages* in the *Frequently Asked Questions* page of the R online help. You might need system administrator privileges to install an R library. If your computer is part of a network, you may need to ask your system administrator for help in doing this.

Other platforms users: follow the instructions in your version of R.

Installing other libraries

Some of the algorithms in `smida` require sub-algorithms from other libraries, such as `MASS` or `lars`. Some are pre-installed in R, such as `MASS`, but others are not, such as `lars`. A list of required libraries can be found on the book website. When you load `smida` (see below) it will tell you which libraries need to be installed. These libraries are available from the book's website as well as from the CRAN mirrors and can be installed in the same way as the `smida` library.

1.1.2 Loading `smida`

Once you have installed the library in R, you can then load `smida` by simply typing

```
library(smida)
```

inside R. You need to load a library once per session. If you need to load another library, you simply replace `smida` with the name of that library. Note that `smida` is written using version 1.8 and therefore might not work properly with older versions of R.

R is a command line-driven programme. This makes investigating data sets highly interactive and flexible. On the other hand, it does require some familiarity with the S language used by R. Even for experienced users, it is extremely helpful to begin each R session by typing

```
help.start()
```

in the R command line. It launches a HTML page in your browser that automatically lists the help files for all the libraries that you have installed in R.

1.2 Data Sets from Biological Experiments

Throughout this book, we use five data sets from original microarray experiments. They are all of a very different nature and are therefore an excellent testing ground for many statistical concepts. In the *Arabidopsis* experiment, the relationship between potassium starvation in the plants and gene expression of transporter and other genes is of interest. The skin cancer experiment is a comparative microarray study of gene expression in a cancerous cell line and a normal cell line. The breast cancer experiment is a clinical study of DNA amplification and deletion patterns, using microarray technology. Its aim is to study the relationship between DNA amplification patterns (rather than gene expression) and the severity of the breast cancer, as measured by several clinical indicators on the patients. The mammary gland experiment studied the relationship between gene expression in the

mammary gland of mice and the developmental stage of the mice. The tuberculosis experiment is also a time-course experiment, measuring the effect of starvation stress on the gene expression of the tuberculosis bacterium.

1.2.1 Arabidopsis experiment: Anna Amtmann

Dr Amtmann is a plant scientist at the University of Glasgow. She is interested in how plants regulate nutrient uptake. Plants cannot move and therefore must be able to adjust their growth and development to an ever-changing environment. Inorganic ions in the soil (e.g. potassium K^+ and calcium Ca^{2+}) are essential components of a plant's nutrition. However, both ion supply and demand fluctuate considerably depending on water status of the soil, light, time of the day, season, developmental stage and tissue. Plants overcome this fluctuation by temporarily storing the nutrients in leaves when they are in large supply and transporting them back to essential cells when they are in short supply.

In order to study the nature, regulation and physiological integration of plant ion transport, a variety of techniques can be used. Electrophysiology measures minute electrical currents passing through individual membrane proteins. Such live recordings have led to the discovery of ion channels responsible for the uptake of essential nutrients as well as toxic ions into plants. Structure–function analysis involves cloning and heterologous expression of genes to elucidate important protein regions for particular functional activities. Finally, microarrays can reveal which genes are involved in the regulation of ion transport.

By varying the physical growth environment for the *Arabidopsis thaliana* plant, Dr Amtmann triggers nutrition transport in the plant. By comparing the expression of a control plant with a plant in which nutrition reallocation takes place, it is possible to pick up differences in the gene expressions. However, the question 'which are the genes that are responsible for the regulation of ion transport?' is in fact a bit too general. It is difficult to compare the situation of nutrition transport with no nutrition transport *per se*. The situation of nutrition transport can be the result of two very different situations, and different genes may be responsible for each. Either the plant has been starved and is moving nutrition from storage sites into growing tissues or there is an excess of nutrients that requires the plant to store them in, for example, the leaves. The question of interest for the microarray experiment therefore is 'which are the genes that show significant changes when the plant is put under different kinds of stress to provoke nutrient transport within the plant?' The types of stress considered are varying lengths of potassium starvation and resupply.

Dr Amtmann is part of a larger group of scientists interested in a number of different nutrients, and some of their findings are published in Maathius et al. (2003).

The main aim of the experiment was to find which transporter genes are involved when Arabidopsis is starved of potassium. A number of different genes

working together are needed to move potassium around the plant, and the eventual aim is to discover the regulatory network that controls potassium homeostasis under varying conditions of external supply.

Collaborators of Dr Amtmann have found that calcium and sodium transporters were strongly affected by nutrient starvation (Maathius et al. 2003). Analysis of this particular experiment found that few known potassium transporter genes responded to the potassium stress. The results could indicate that the majority of potassium transporters in *Arabidopsis* are regulated post-transcriptionally rather than at the transcript level.

Also, the plants used for the experiment were mature and thus were not growing quickly. Mature plants already have stores of potassium, and therefore starvation might not have induced a lot of potassium transport. Further investigations of potassium starvation will concentrate on young *Arabidopsis* seedlings, where changes in potassium are likely to have greater effects.

Physical aspects of the experiment

The *Arabidopsis* plants used in the experiment were grown hydroponically under the same laboratory conditions, initially with no mineral starvation. Once fully grown, some plants were then given a mineral starvation treatment; others were used as controls.

For the extraction of RNA, the whole roots were used. The roots are the first point of contact with external nutrient solution and so are the first part of the plant affected by a change in the surrounding conditions. The root samples are then immediately shock-frozen in liquid nitrogen. The RNA was then extracted from the roots using Qiagen RNeasy products. Sometimes spiking controls were added. A protocol from MWG (Ebersberg, Germany) was used for hybridization, using direct labelling with Cy dyes during reverse transcription and a formamide-based hybridization buffer. ScanArray Lite was used to scan the arrays, and the image analysis was carried out using QuantArray.

The arrays used were custom-made two-channel cDNA chips (MWG) containing 50-mer probes representing 1,153 genes and 57 control sequences. Each of these 1,250 oligonucleotide probes are replicated twice on the arrays. For more information, see Maathius et al. (2003).

Organizational aspects of the experiment

Eleven two-channel arrays were used, each with one channel assigned to a treatment sample and the other to a control. In seven of the arrays, the treatments involved starving the plants for different lengths of time, ranging from 5 h to 4 d. The aim of this was to give insight into which genes were responsible, in the root, for re-allocating nutrients from storage areas (such as leaves) to essential cells. The treatment samples used for the other four arrays had potassium reintroduced for 5 or 24 h after a period of starvation of 24 or 96 h. This was done to look at the recovery process of the plant after K^+ starvation. Table 1.1 details the 11 arrays'

Table 1.1 Details of the treatment samples applied to each array used in the Arabidopsis experiment, together with the dye assignment for treatment and control.

Array	Treatment type		Repeat	Cy3 channel sample	Cy5 channel sample
	K ⁺ starvation (Hours)	K ⁺ re-addition (Hours)			
1	5	None	1	Control	Treatment
2	10	None	1	Control	Treatment
3	10	None	2	Treatment	Control
4	24	None	1	Control	Treatment
5	24	None	2	Treatment	Control
6	96	None	1	Control	Treatment
7	96	None	2	Treatment	Control
8	24	5	1	Control	Treatment
9	96	5	1	Control	Treatment
10	96	5	2	Treatment	Control
11	96	24	2	Treatment	Control

treatments and the channels to which these treatments were assigned. The control sample, also known as the *reference* sample, is Arabidopsis RNA from a plant that was not starved and was harvested at the same time as the treatment plants.

Note that repeats were not taken at the same time of day and so cannot be considered exact replicates.

1.2.2 Skin cancer experiment: Nighean Barr

Dr Nighean Barr is a researcher at the Cancer Research UK Beatson Laboratories in Glasgow. One experiment that she has carried out investigated differences between gene expressions in cancerous and normal fibroblast cells. These cells are the main constituent of connective tissue within the body and make fibres and the extracellular matrix. In the skin, these cells are susceptible to UV radiation from sunlight, and through this they can become cancerous. By finding which genes are differentially expressed in the cancerous versus normal cells, one can focus research into treatments for cancer. The data have not yet been published.

Physical aspects of the experiment

The fibroblast tissue used was created *in vitro* from two cell lines—one cancerous and one normal. From each of these two cell lines, four separate technical replicates were created. Pairs of cancerous and normal replicates were then hybridized to two-channel cDNA arrays. Each array contained 4,608 genes replicated twice.

Organizational aspects of the experiment

As the RNA comes from two cell lines, the conclusions of this experiment can only be generalized to these specific cell lines. They do not necessarily have something to say about the expression in the general population of unaffected and cancerous fibroblast cells. In this sense, these replicates are sometimes referred to as technical replicates.

The experiment is a direct comparison of cancerous fibroblast cells with normal cells using four technical replicates of each of the cell lines. Four arrays were used in the experiment. The solution hybridized to each array contains one replicate of the cancerous tissue and one replicate of the normal tissue. For two of the arrays, the normal tissue was stained with Cy3 dye and Cy5 was used for the cancerous tissue; on the other two arrays the dye assignments were swapped. The design details are given in the Table 1.2.

1.2.3 Breast cancer experiment: John Bartlett

Dr John Bartlett is a senior lecturer in the Division of Cancer Sciences and Molecular Pathology of the University of Glasgow at the city's Royal Infirmary. He is interested in finding out the key genetic determinants associated with aggressive and non-aggressive breast cancer.

Rather than looking at gene expression of RNA in tissue, he investigates differences in the genomic DNA of breast cancer patients compared to that in controls. In cancer cells, the genome can undergo changes, such as obtaining additional copies of certain genes and losing genetic material from other genes. An increase in the *gene number* is known as *gene amplification*. When fewer copies are present compared to the genome of normal cells, this is known as *gene deletion*. An established method, comparative genomic hybridization (CGH) (Kallioniemi et al. 1992), allows regions of the genome where changes occur to be identified. Until recently, only fairly large regions of the genome could be considered—around 5 to 10 million base pairs. Using microarrays for conducting CGH, Dr Bartlett can consider specific genes rather than the much larger genomic regions previously looked at.

Table 1.2 Details of the dye assignments of samples applied to each array used in the skin cancer experiment.

Array	Cy3 sample	Cy5 sample
1	Cancer	Normal
2	Normal	Cancer
3	Cancer	Normal
4	Normal	Cancer

In order to link gene amplification/deletion information to the aggressiveness of the tumours in this experiment, clinical information is available about each of the patients, such as their Nottingham prognostic index (NPI). This is used to help classify the tumours into different severity groups while controlling for non-genomic influences.

Dr Bartlett is working with Dr Caroline Witton of the University of Glasgow at the Royal Infirmary and with Steven Seelig, Walter King and Teresa Ruffalo of Vysis Inc., Downers Grove, Chicago, who produced the arrays used in the experiment.

The work has resulted in a sub-classification of breast cancers and it has suggested genes that have an effect on the aggressiveness of the cancer. The results of the experiment are detailed in Witton et al. (2002). The chips used in this experiment have since been superseded by larger chips from Vysis Inc. John Bartlett and Caroline Witton continue to work on breast cancer classification with these new arrays.

Physical aspects of the experiment

Genomic DNA from cancer patients is extracted from stored frozen tumour tissue and from female reference DNA. The protocol used for the extraction, PCR amplification, labelling and hybridization are given in Pestova et al. (2004).

The arrays contain 59 clones, each spotted three times. 57 genes are represented by these 59 clones, since two genes have both 5' and 3' versions included. In each of the two-channel arrays, reference female DNA is used as a control in one channel. Rather than Cy dyes, the dyes used are Alexa 488, a green dye, and Alexa 594, a red one. Further details about the image analysis and arrays used are given in Pestova et al. (2004).

Organizational aspects of the experiment

The experiment involves the genetic material from 62 breast cancer patients. To measure the gene profiles in all tumours, 62 arrays were used in the experiment. In all the cases, the control samples were coloured with Alexa 594 and the patients' samples with Alexa 488. For each patient, there is also a variety of clinical information available, including the following:

- their survival time (in years) after the tissue was removed;
- their age at diagnosis (in years);
- the size of their tumour (in mm);
- whether they died from breast cancer;
- whether they are still alive;

- the severity grade of their breast cancer: 1 (low) to 3 (high);
- their NPI score.

1.2.4 Mammary gland experiment: Gusterson group

Professor Barry Gusterson and Dr Torsten Stein of the University of Glasgow's Division of Cancer Science and Molecular Pathology are interested in the molecular mechanisms of breast cancer development. They have conducted a large microarray time-series experiment looking at changes in gene expression in healthy mammary tissue during development and pregnancy.

The interest in healthy breast morphogenesis for cancer research is partly due to a desire to understand what changes in gene expression occur in non-cancerous tissue before looking at pathological tissue. Even more importantly, the morphological changes that happen during normal development show many characteristics that are also found during breast cancer development, including invasive growth, high proliferation and tissue remodelling. However, genes that regulate these changes in healthy mammary tissue are often deregulated in cancerous tissue. The identification of these genes could therefore enable the development of treatments to target such genes in cancerous patients. The time frame considered was from puberty through adult virgin, early, mid- and late pregnancy, lactation and involution, which is the period after lactation stops and the milk secreting cells die.

Given the ethical infeasibility of using human tissue, mice were used as a model organism for all experiments. In order to find the regulatory genes for the different morphological changes, it is important to identify those genes that are only expressed at certain time points of development. However, it is known that similar processes can also occur at multiple stages, and therefore similar sets of genes can be expressed at different times. Therefore, it can be important to find the genes that are expressed in certain combinations of stages.

For example, after birth, mouse pups do not suckle immediately. During this time, the mammary gland starts to go through a process that resembles early involution. Thus, it is of interest to see what genes are up- or down-regulated during both the beginning of lactation and of involution.

As a result of this project, a number of putative regulatory genes have been identified at various developmental stages. Also, around 100 genes have been found that are specifically up-regulated during involution, many of which have been independently verified. This work has been published (Stein et al. 2004); however, further analysis will be done on this large experiment.

Physical aspects of the experiment

The whole mammary gland of a mouse in the experiment was used for each array, after removing the major lymph node from the gland. The RNA from each gland

was extracted after preparation and liquid nitrogen freezing using Trizol reagents and Qiagen RNeasy-columns. Following the Affymetrix guidelines, the labelled RNA was hybridized to the chip for 16 hours at 42 °C. Afterwards, the chip was washed using an Affymetrix Fluidics station. An Agilent scanner, together with Affymetrix's MAS 5.0 software, was used for the image analysis.

The chips were Affymetrix mouse arrays containing 12,488 probe sets, representing around 8,600 genes. The number of probes in each set varied between 12 and 20. Housekeeping genes were used as controls along with 4 bacterial gene spikes. Each probe is 25 nucleotides in length. Affymetrix arrays are single-channel chips and so only one condition or tissue can be probed per array.

Organizational aspects of the experiment

For each mammary gland, tissue from a different mouse is used. In the experiment, 18 time points during the development of a female mouse were considered. Each time point was replicated three times. Altogether, 54 arrays were produced. Table 1.3 details the design of the experiment.

1.2.5 Tuberculosis experiment: B μ G@S group

The bacterial microarray group at St. George's hospital in London (B μ G@S), led by Professor Philip Butcher, has created a number of different types of two-channel microarrays. Each array type is suitable for a different type of bacteria. The arrays have been produced not only for their own group but also for the wider scientific community. One array type is for *mycobacterium tuberculosis* (*M. tuberculosis*), the bacterium that is the cause of tuberculosis. We consider a time-course experiment conducted on *M. tuberculosis* arrays by the senior scientist at B μ G@S, Jason Hinds.

The aim of the experiment was to understand the developmental changes in gene expression of *M. tuberculosis* under stressed growth conditions, where resources

Table 1.3 Table showing the stage of growth and time point within this stage of each array in the mouse mammary gland experiment.

Arrays	State	Time point	Arrays	State	Time point
1–3	Virgin	Week 6	28–30	Pregnancy	Day 17.5
4–6	Virgin	Week 10	31–33	Lactation	Day 1
7–9	Virgin	Week 12	34–36	Lactation	Day 3
10–12	Pregnancy	Day 1	37–39	Lactation	Day 7
13–15	Pregnancy	Day 2	40–42	Involution	Day 1
16–18	Pregnancy	Day 3	43–45	Involution	Day 2
19–21	Pregnancy	Day 8.5	46–48	Involution	Day 3
22–24	Pregnancy	Day 12.5	49–51	Involution	Day 4
25–27	Pregnancy	Day 14.5	52–54	Involution	Day 20

are limited. A related experiment done by the B μ G@S group has been done on the *M. tuberculosis* trcS mutant and has found a number of genes differentially expressed in this mutant compared to a reference strain (Wernisch et al. 2003).

Physical aspects of the experiment

Samples of *M. tuberculosis* strains were grown in flasks with a fixed amount of growth medium. Initially, the *M. tuberculosis* grows abundantly; when resources become scarce the bacterium shuts down many of its functions. This is thought to occur at some point before day 30, the end of the experiment.

The arrays used were two-channel chips specifically designed and made by the B μ G@S group in collaboration with the National Institute for Medical Research, with funding from the Medical Research Council. The array used the H37Rv strain annotation described in Cole et al. (1998). Using PCR amplification, 3,924 target genes were spotted on the arrays. These were placed on the array in 4 \times 4 arrangement of 16 sub-grids.

mRNA was harvested from samples at four different time periods, thought to represent different stages of development for the bacterium in the given medium. This signal sample was converted to cDNA labelled with either the Cy3 or Cy5 dye, mixed with a reference sample and hybridized to the array. On each of the arrays, genomic *M. tuberculosis* DNA was used as the reference and was labelled with the other dye. Commonly used reference samples are typically isolated from a single representative RNA source or pooled mixtures of RNA derived from several sources. Genomic DNA offers an alternative reference nucleic acid with a number of potential advantages, including stability, reproducibility and a potentially uniform representation of all genes, as each unique gene should have equal representation in the genome (Kim et al. 2002).

Table 1.4 Table showing the time point of each array in the tuberculosis experiment together with dye assignments for the signal (mRNA) and reference (gDNA) channels.

Time Point	Replicates	Cy3 sample	Cy5 sample
Day 6	1–3	mRNA	gDNA
Day 6	4	gDNA	mRNA
Day 14	1–3	mRNA	gDNA
Day 14	4	gDNA	mRNA
Day 20	1,2,4	mRNA	gDNA
Day 20	3	gDNA	mRNA
Day 30	1–3	mRNA	gDNA
Day 30	4	gDNA	mRNA

Organizational aspects of the experiment

The four time points investigated were 6, 14, 20 and 30 days after the bacterium's introduction to the growth medium. Four different replicate samples were used at each of the time points, with one sample being used per array. Consequently, 16 arrays were used in total.

Table 1.4 gives the details of the dye assignments used for each array. Note that, although dye assignments for the signal (mRNA) and reference (gDNA) mixtures were swapped, this was not done evenly. Three of the four replicates at each time points used Cy3 for the signal and only the remaining replicate used Cy5 for the signal.