EMBO03 Lab 3: Introduction to Bioconductor

marray Packages

Sandrine Dudoit and Robert Gentleman

March 14, 2003

In this lab, we demonstrate the main functions in the marray suite of packages for diagnostic plots and normalization of two-color spotted microarray data. Efforts are underway to interact (read and write) with MAGE-ML documents. A brief description of the four main marray packages is given next.

marrayClasses. This package contains class definitions and associated methods for pre- and post-normalization intensity data for batches of arrays. Methods are provided for the creation and modification of microarray objects, basic computations, printing, subsetting, and class conversions.

marrayInput. This package provides functionality for reading microarray data into R, such as intensity data from image processing output files (e.g., .spot and .gpr files for the Spot and GenePix packages, respectively) and textual information on probes and targets (e.g., from .gal files and god lists). tcltk widgets are supplied to facilitate and automate data input and the creation of microarray-specific R objects for storing these data.

marrayPlots. This package provides functions for diagnostic plots of microarray spot statistics, such as boxplots, scatterplots, and spatial color images. Examination of diagnostic plots of intensity data is important in order to identify printing, hybridization, and scanning artifacts that can lead to biased inferences concerning gene expression.

marrayNorm. This package implements robust adaptive location and scale normalization procedures, which correct for different types of dye biases (e.g., intensity, spatial, plate biases) and allow the use of control sequences spotted onto the array and possibly spiked into the mRNA samples. Normalization is needed to ensure that observed differences in intensities are indeed due to differential expression and not experimental artifacts; fluorescence intensities should therefore be normalized before any analysis that involves comparisons among gene expression measures within or between arrays.
To load the packages

```r
> library(marrayNorm)
```

For a more detailed introduction, consult the package vignettes which can be listed by the command `openVignette()`. A demo for `marrayPlots` can also be accessed by `demo(marrayPlots)`. We will work with the sample dataset `swirl`; for a description of swirl, type `? swirl`. To load this dataset

```r
> data(swirl)
```

1 Basic classes and methods: `marrayClasses` package

One of the main classes in `marrayClasses` is the `marrayLayout` class; it is used to keep track of important layout parameters, such as the total number of spotted probe sequences on the array, the dimensions of the spot and grid matrices, the plate origin of the probes, information on spotted control sequences. For details on this class consult the help file, `? marrayLayout`. Two other important classes are `marrayRaw` and `marrayNorm`, which represent, respectively, pre-normalization and post-normalization intensity data for a batch of spotted microarrays. Methods for manipulating instances of these classes are also described in the help files.

The object `swirl` is an instance of the class `marrayRaw`. Try the following commands to obtain information on this object

```r
> class(swirl)
[1] "marrayRaw"
```

```r
> slotNames(swirl)
[1] "maRf"   "maGf"   "maRb"   "maGb"   "maW"   "maLayout"
[7] "maGnames" "maTargets" "maNotes"
```

```r
> swirl
Pre-normalization intensity data: Object of class marrayRaw.

Number of arrays: 4 arrays.

A) Layout of spots on the array:
Array layout: Object of class marrayLayout.

Total number of spots: 8448
Dimensions of grid matrix: 4 rows by 4 cols
Dimensions of spot matrices: 22 rows by 24 cols

Currently working with a subset of 8448 spots.

Control spots:
There are 2 types of controls:
Control N

768 7680

Notes on layout:
No Input File

B) Samples hybridized to the array:
Object of class marrayInfo.

<table>
<thead>
<tr>
<th>maLabels</th>
<th># of slide</th>
<th>Names</th>
<th>experiment</th>
<th>Cy3 experiment</th>
<th>Cy5</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81</td>
<td>81 swirl.1.spot</td>
<td>swirl</td>
<td>wild type</td>
<td>2001/9/20</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>82</td>
<td>82 swirl.2.spot</td>
<td>wild type</td>
<td>swirl</td>
<td>2001/9/20</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>93</td>
<td>93 swirl.3.spot</td>
<td>swirl</td>
<td>wild type</td>
<td>2001/11/8</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>94</td>
<td>94 swirl.4.spot</td>
<td>wild type</td>
<td>swirl</td>
<td>2001/11/8</td>
<td></td>
</tr>
</tbody>
</table>

comments
1 NA
2 NA
3 NA
4 NA

Number of labels: 4
Dimensions of maInfo matrix: 4 rows by 6 columns

Notes:
C:/GNU/R/rw1041/library/marrayInput/data/SwirlSample.txt

C) Summary statistics for log-ratio distribution:

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>swirl.1.spot</td>
<td>-2.73</td>
<td>-0.79</td>
<td>-0.58</td>
<td>-0.48</td>
<td>-0.29</td>
<td>4.42</td>
</tr>
<tr>
<td>swirl.2.spot</td>
<td>-2.72</td>
<td>-0.15</td>
<td>0.03</td>
<td>0.03</td>
<td>0.21</td>
<td>2.35</td>
</tr>
<tr>
<td>swirl.3.spot</td>
<td>-2.29</td>
<td>-0.75</td>
<td>-0.46</td>
<td>-0.42</td>
<td>-0.12</td>
<td>2.65</td>
</tr>
<tr>
<td>swirl.4.spot</td>
<td>-3.21</td>
<td>-0.46</td>
<td>-0.26</td>
<td>-0.27</td>
<td>-0.06</td>
<td>2.90</td>
</tr>
</tbody>
</table>

D) Notes on intensity data:

To access individual slots
> maLayout(swirl)

Array layout: Object of class marrayLayout.

Total number of spots: 8448
Dimensions of grid matrix: 4 rows by 4 cols
Dimensions of spot matrices: 22 rows by 24 cols

Currently working with a subset of 8448 spots.

Control spots:
There are 2 types of controls:
Control  N
    768  7680

Notes on layout:
No Input File

> maGnames(swirl)

Object of class marrayInfo.

     maLabels  "ID"  "Name"
   1   geno1 control  geno1
   2   geno2 control  geno2
   3   geno3 control  geno3
   4    3XSSC control    3XSSC
   5    3XSSC control    3XSSC
   6      EST1 control      EST1
   7   geno1 control  geno1
   8   geno2 control  geno2
   9   geno3 control  geno3
  10    3XSSC control    3XSSC
...

Number of labels: 8448
Dimensions of maInfo matrix: 8448 rows by 2 columns

Notes:
C:/GNU/R/rw1041/library/marrayInput/data/fish.gal

As with other microarray objects in Bioconductor packages, you can use subsetting commands for marrayRaw objects. For data on the first 100 genes in the second array in the swirl batch
> sw <- swirl[1:100, 2]
> class(sw)

[1] "marrayRaw"

> sw

Pre-normalization intensity data: Object of class marrayRaw.

Number of arrays: 1 arrays.

A) Layout of spots on the array:
Array layout: Object of class marrayLayout.

Total number of spots: 8448
Dimensions of grid matrix: 4 rows by 4 cols
Dimensions of spot matrices: 22 rows by 24 cols

Currently working with a subset of 100 spots.

Control spots:
There are 2 types of controls:
Control N
  48  52

Notes on layout:
No Input File

B) Samples hybridized to the array:
Object of class marrayInfo.

            maLabels # of slide Names experiment Cy3 experiment Cy5 date comments
2       82     82 swirl.2.spot      wild type       swirl 2001/9/20 2 NA

Number of labels: 1
Dimensions of maInfo matrix: 1 rows by 6 columns

Notes:
C:/GNU/R/rw1041/library/marrayInput/data/SwirlSample.txt
C) Summary statistics for log-ratio distribution:

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>swirl.2.spot</td>
<td>-1.08</td>
<td>-0.26</td>
<td>-0.12</td>
<td>-0.12</td>
<td>0</td>
<td>1.72</td>
</tr>
</tbody>
</table>

D) Notes on intensity data:

You can access red and green foreground and background intensities, and log ratios as follows

```r
> Gb <- maGb(swirl)
> dim(Gb)
[1] 8448 4

> Gb[1:5, ]
[1,] 182  175  86   97
[2,] 171  183  86   85
[3,] 153  183  86   85
[4,] 153  142  71   87
[5,] 153  142  71   87

> Rf <- maRf(swirl)
> dim(Rf)
[1] 8448 4

> Rf[1:5, ]
swirl.1.spot swirl.2.spot swirl.3.spot swirl.4.spot
[1,] 19538.470  16138.720   2895.160  14054.5400
[2,] 23619.820  17247.670   2976.6230 20112.2600
[3,] 21579.950  17317.150   2735.6190 12945.8500
[4,]  8905.143   6794.381   318.9524   524.0476
[5,]  8676.095   6043.542   780.6667   304.6190

> M <- maM(swirl)
> dim(M)
[1] 8448 4
```
2 Reading in data: marrayInput package

This package provides functionality for reading microarray data into R, such as intensity data from image processing output files (e.g., .spot and .gpr files for the Spot and GenePix packages, respectively) and textual information on probes and targets (e.g., from .gal files and god lists). tcltk widgets are supplied to facilitate and automate data input and the creation of microarray-specific R objects for storing these data. See for example ?read.marrayRaw or ?widget.marrayRaw.

3 Diagnostic plots: marrayPlots package

The marrayPlots package provides functions for diagnostic plots of microarray spot statistics, such as boxplots, scatterplots, and spatial color images. To produce a spatial image of background intensities for the Cy3 channel in the third array

```r
> tmp <- maImage(swirl[, 3], x = "maGb", bar = FALSE)

To produce a spatial image of log ratios for the first array in the batch
```
To produce boxplots of log ratios by sector for the first array in the batch

```r
> maBoxplot(swirl[, 1])
```
To produce boxplots of log ratios by plate for the second array in the batch

```r
> maPlate(swirl) <- maCompPlate(swirl, n = 384)
> maBoxplot(swirl[, 2], x = "maPlate", names = NULL)
```
For boxplots of log ratios for all four arrays

> maBoxplot(swirl)
4 Normalization: marrayNorm package

The marrayNorm package implements robust adaptive location and scale normalization procedures, which correct for different types of dye biases (e.g., intensity, spatial, plate biases). The main location and scale normalization function is maNormMain. Simpler wrapper functions are provided in maNorm and maNormScale. The functions operate on objects of class marrayRaw (or possibly marrayNorm, if normalization is performed in several steps) and return objects of class marrayNorm. For within-print-tip-group loess location normalization of the batch swirl

> swirl.norm <- maNormMain(swirl)

For boxplots of post-normalization log ratios

> maBoxplot(swirl.norm[, 1])
> maBoxplot(swirl.norm)
5 Miscellaneous tools: marrayTools package

The marrayTools package provides additional functions for handling two-color spotted microarray data, including a number of user-friendly wrapper functions for performing standard analyses.

The spotTools and gpTools functions in the development version of marrayTools start from Spot (.spot and .gal) and GenePix (.gpr and .gal) image analysis output files, respectively, and automatically read in these data into R, perform standard normalization (within print-tip-group loess), and create a directory with a standard set of diagnostic plots (jpeg format), excel files of quality measures, and tab delimited files of normalized log ratios M and average log intensities A. In addition, an object of class marrayRaw or marrayNorm is returned. The package also includes functions for computing various gene statistics and for generating HTML pages for gene lists (htmlPage).

```r
> datadir <- system.file("data", package="marrayInput")
```
> normdata <- spotTools(path=datadir, quality=FALSE)