

BioSys PhD | Earthsystems PhD

Statistics 1

Normalization Methods in Microarray Data

Lisete Sousa

lmsousa@fc.ul.pt

Room: 6.4.25

*Departamento de Estatística e Investigação Operacional
Faculdade de Ciências da Universidade de Lisboa*

2017

Summary

- 1 Experimental Design
- 2 Image Processing
- 3 Normalization

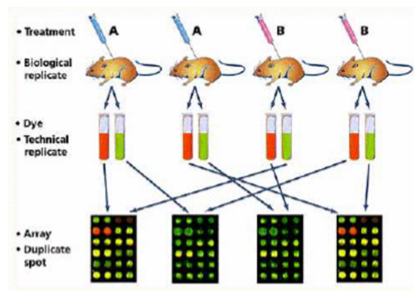
Experimental Design and Replicates

Experimental design objective:

Make the analysis of the data and the interpretation of the results as simple and as powerful as possible, given the purpose of the experiment and the constraints.

Experimental design scheme (example):

Three layers of design in a simple microarray experiment.



Churchill (2002)

Types of replicates:

Microarray experiments can be replicated at many different levels. Fundamentally there are two types of replicates:

Biological replicates

Biological replicates are taken at the level of the population being studied. In the previous figure, each rat is a biological replicate.

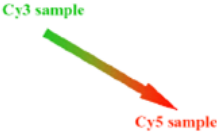
Technical replicates

Technical replicates are taken at the level of experimental apparatus. The purpose of technical replicates is to account for variability in the experimental setup. Technical replicates can be at many levels:

- Replicate features on the array, which can account for differential printing or hybridization
- Replicate arrays hybridized with the same sample
- Replicate sample preparation; for example, two dye-reversed labelings.

Note that if the experiment was of sufficient high quality, then technical replicates would not be required at all.

Important points in experimental design:

- Use adequate biological replication.
- Use dye swapping:
 
- Make direct comparisons between samples that are of most interest.

(i) Common reference design (ii) All-pairs design

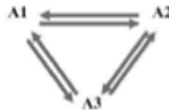


Image Processing

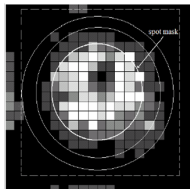
- The image of the microarray generated by the scanner is the raw data of the experiment.
- Computer algorithms, known as feature extraction software, convert the image into the numerical information that quantifies gene expression (this is the first step of data analysis).
- The image processing involved in feature extraction has a major impact on the quality of the data and the interpretation we can place on it.

Various image processing techniques may be applied to read and interpret the outputs of Microarrays.

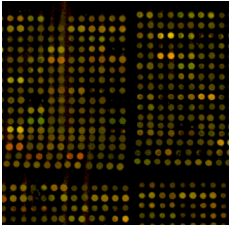
Feature extraction:

The first step in the computational analysis of microarray data is to convert the digital images of hybridization intensity generated by the scanner into numerical measures of hybridization intensity of each channel on each feature. This step is known as **feature extraction**. There are four steps:

- 1 Identify the positions of the features on the microarray.
- 2 For each feature, identify the pixels on the image that are part of the feature.
- 3 For each feature, identify nearby pixels that will be used for background calculation.
- 4 Calculate numerical information for the intensity of the feature and the intensity of the background.



Spot quality problems:

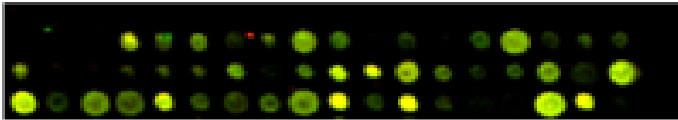


Uneven grid positions

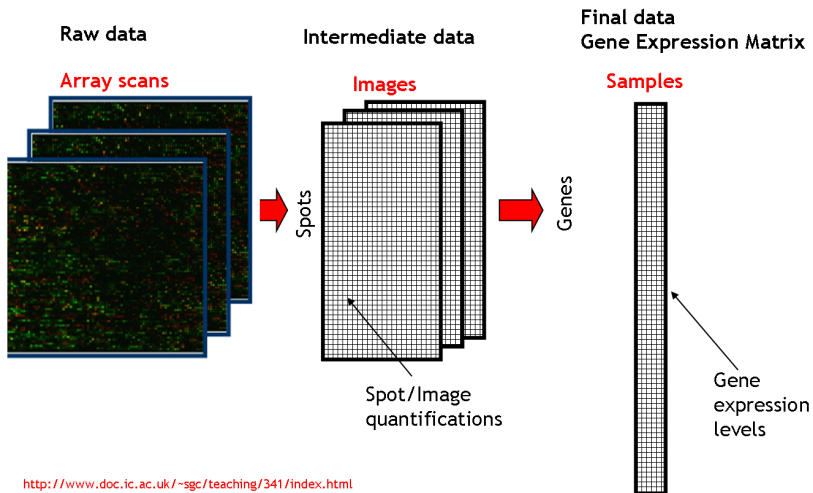
Curves within a grid

Variable spot size or shape

Variable distance between spots



From microarray images to gene expression matrices:



<http://www.doc.ic.ac.uk/~sgc/teaching/341/index.html>

Normalization

Normalization is a general term for a collection of methods that are directed at resolving the **systematic errors** and **bias** introduced by the microarray experimental platform.

We will give a general understanding of why we need to normalize microarray data and the **methods for normalization** that are most commonly used, arranged into three sections:

- **Data cleaning and transformation:** looks at cleaning and transforming the data generated by the feature extraction software.
- **Within-array normalization:** methods that allow for the comparison of Cy3 and Cy5 channels of a two color microarray.

- **Between-array normalization:** methods that allow for the comparison of measurements on different arrays – applicable both to two-color and single channel arrays.

Data cleaning and transformation:

The microarray data generated by the feature extraction software is typically in the form of one or more text files.

Before using the data to answer scientific questions, there are a number of steps that are commonly taken to ensure that the data is of high quality and suitable for analysis.

There are three stages of data cleaning and transformation:

- Removing flagged features.
- Background subtraction.
- Taking logarithms.

Removing flagged features:

- Remove flagged features from the data.

Disadvantage: remove potentially valuable data.

- Refer back to the original image of every flagged feature. Identify the problem. Perform a new feature extraction on the flagged feature to obtain a more reliable measure.

Disadvantage: Requires time, resources and may not always be practical.

Background subtraction:

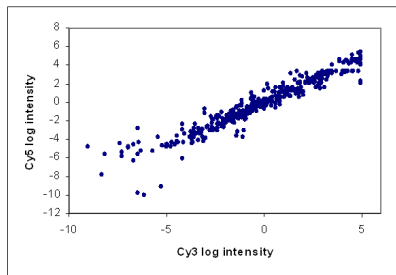
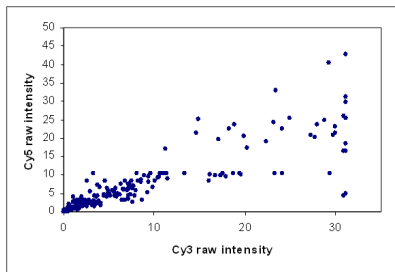
Subtract the background signal from the feature intensity. However, when the background intensity is higher than the feature intensity the result would be a negative value. There are three approaches to deal with this:

- Remove these features from the analysis. The unusually high background is taken to represent a local problem with the array. The intensity of the feature is regarded as unreliable.
- Use the lowest available signal-intensity as the background-subtracted intensity. If the background intensity is higher than the feature intensity, it represents a gene with no or very low expression.
- Use more sophisticated (Bayesian) algorithms to estimate the true feature intensity. Based on the assumption that the true feature intensity is higher than the background intensity. High background intensity represents some type of experimental error.

Taking logarithms:

It is a common practise to transform DNA microarray data from the raw intensities into log intensities before proceeding with the analysis:

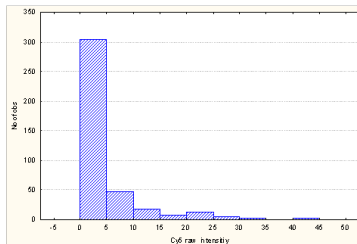
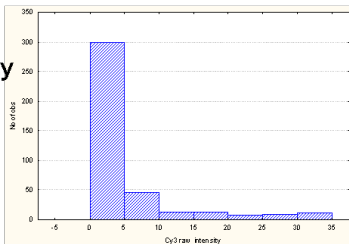
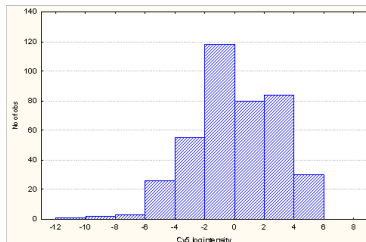
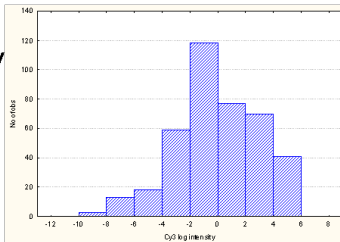
- 1 There should be a reasonably even spread of features across the intensity range.
- 2 The variability should be constant at all intensity levels.



- 3 The distribution of the intensities should be approximately bell-shaped.

Cy3


Cy5

Raw
intensityLog
intensity

It is usual in microarray data analysis to use **logarithms to base 2**.

- ➊ The **ratio** of the raw Cy3 and Cy5 intensities is transformed into the **difference** between the logs of the intensities.
- ➋ 2-fold up-regulated genes correspond to a log ratio of **+1**.
- ➌ 2-fold down-regulated genes correspond to a log ratio of **-1**.
- ➍ Genes that are not differentially expressed have a log ratio of 0.

$$\text{Ratio} = \frac{RED}{GREEN}$$



$$\text{LogRatio} = \log_2 \left(\frac{RED}{GREEN} \right) = \log_2(RED) - \log_2(GREEN)$$

Within-array normalization

We need to be able to **compare the Cy3 and the Cy5 intensities** on an equal footing, when measuring differential expression between two samples. This is achieved by eliminating sources of systematic bias, as:

- 1 Different abundance (Cy3 and Cy5 labels).
- 2 Different emission responses (laser).
- 3 Different measurements (photomultiplier tube).

Methods:

- Linear regression of Cy5 against Cy3
- Linear regression of log ratio against average intensity
- Nonlinear (Loess) regression of log ratio against average intensity

Linear regression of log ratio against average intensity:

An alternative and very useful approach to visualizing and normalizing the data is to produce a scatterplot of the log ratio against the average intensity of each feature (MA plot).

If the two channels are responding similarly, then the data should appear symmetrically about a horizontal line through zero.

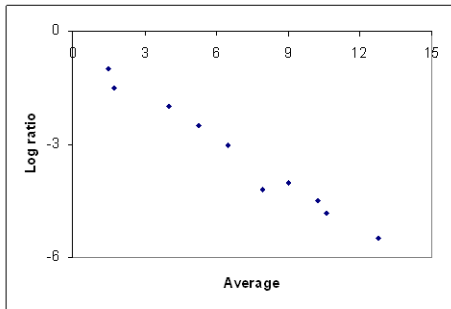
Normalization procedure:

- Construct the average log intensity and log ratio for each feature.
- Produce de MA plot.
- Perform a linear regression.
- For each feature calculate the

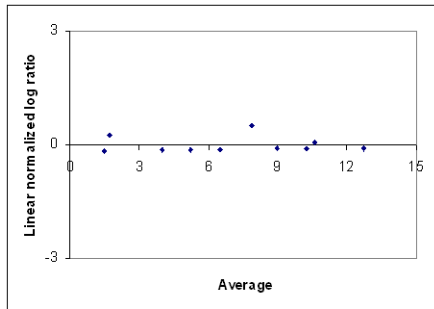
$$\text{normalized ratio} = \text{fitted logratio} - \text{raw logratio}$$

Example (Cont.)

<i>average</i>	1.5	1.75	4	5.25	6.5	7.9	9	10.25	10.6	12.75
<i>log ratio</i>	-1	-1.5	-2	-2.5	-3	-4.2	-4	-4.5	-4.8	-5.5
<i>normal. log ratio</i>	-0.16	0.24	-0.14	-0.13	-0.12	0.53	-0.11	-0.10	0.06	-0.08



Slope: -0.393
Intercept: -0.567



Slope: -0.021
Intercept: -0.071

Nonlinear (loess) regression of log ratio against average intensity:

It is common with microarray data that the relationship between the two channels is nonlinear. When that is the case, linear regression may not produce the best answers.

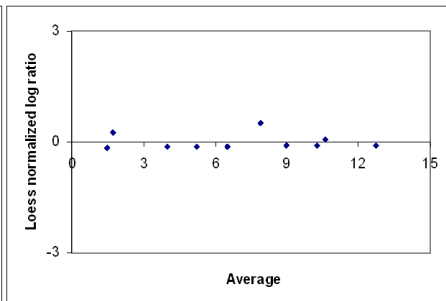
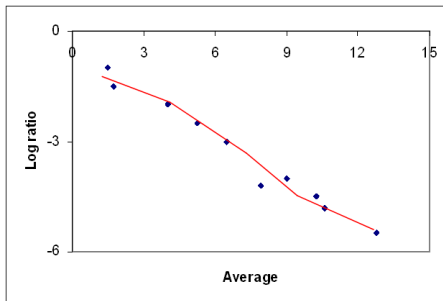
The most commonly used method for nonlinear regression with microarray data is called Loess regression. Loess stands for **locally weighted polynomial regression**.

Normalization procedure:

- Construct the average log intensity and log ratio for each feature.
- Produce de MA plot.
- Apply the Loess regression to the data.
- For each feature calculate the

$$\text{normalized ratio} = \text{fitted logratio} - \text{raw logratio}$$

Example (Cont.)



Between-array normalization

Now, we look at normalization methods that allow to make **comparisons between samples** hybridized to different arrays, which could be either two-color arrays or Affymetrix arrays.

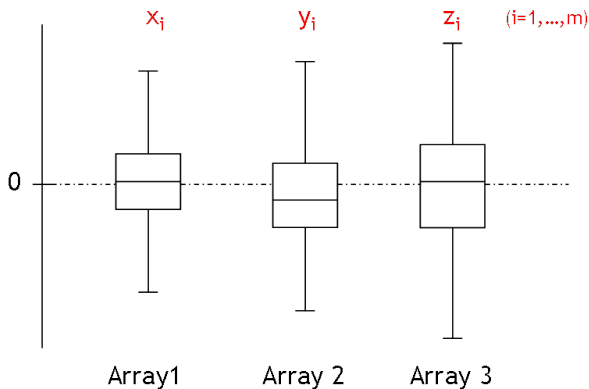
In order to compare the samples hybridized to different arrays on an equal footing, it is necessary to correct for variability introduced by using multiple arrays.

There are essentially two methods for normalizing data:

- Scaling
- Centering

Visualizing the data – box plots:

The box plot is an excellent method for comparing the distributions of log intensities or log ratios of genes on several microarrays.

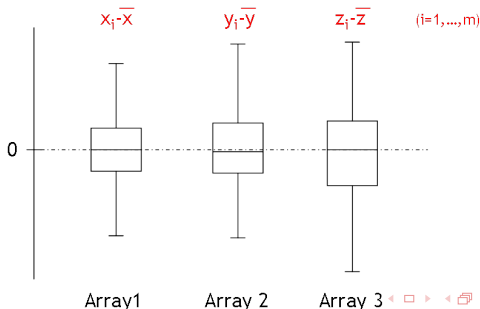


1. Scaling

Data is scaled to ensure that the means of all distributions are equal.

Method: Subtract the mean log ratio (or log intensity) of all of the data on the array from each log ratio (or log intensity) measurement on the array.

Alternative: Using the median, provides a more robust measure of the average intensity on an array in situations where there are outliers or the intensities are not normally distributed.



2. Centering (most commonly used)

Data is centered to ensure that the means and the standard deviations of all distributions are equal.

Method: For each measurement on the array subtract the mean of the array and divide by the standard deviation.

Alternative: Using the median and the median absolute deviation, provides a more robust measure.

