Microarray preprocessing

See Bioinformatics and Computational Biology Solutions Using R and BioConductor

Eds. Gentleman, Carey, Huber, Irizarry, Dudoit

Springer

This lecture will focus on the Affymetrix GeneChip oligonucleotide microarrays and consider the following

Preprocessing tasks

- Image analysis
- Data import
- Background adjustment
- Normalization
- Summarization
- Quality assessment

Image analysis

- Arrays scanned to produce a high-resolution image
- Image analysis consists of as:
  - work out which pixels in the image correspond to each probe
  - summarize the intensity of all pixels associated with the probe
- Details of image analysis will not be covered in this course
- Typically carried out using manufacturer software

The CEL file

- Stores the results of intensity calculations from image analysis
- Contains an estimated intensity for each feature on the array
  - cell location
  - mean intensity
  - standard deviation
  - number of pixels used to calculate these values
- Different versions of the CEL format exist (text and binary)

The CDF file

- Chip Design File
- Describes what’s on the chip
- Required in order to read the data using BioConductor

Types of probes

- Perfect Match (PM)
- Mismatch (MM)
- Genomic
- Anti-genomic

Data import

- CEL files can be read using the ReadAffy( ) function from the BioConductor library affy
- By default ReadAffy( ) will read all the .CEL files in the current working directory
- Returns an object of class affyBatch

RMA – Robust Multichip Averaging

We will use the function RMA( ) from the affy library to do the next three steps: background subtraction, normalization and summarization

RMA is based on the following statistical model (additive multiplicative model):

$$Y = B + \alpha S$$

$$\log(S) = \theta + \phi + \varepsilon$$

B: Background noise; *S:* Signal, composed of θ the true abundance, φ, probe affinity effect and ε measurement error.

Aside

Additive multiplicative model is written in some sources as:

$$y_{ij} = b_j + \theta_i \phi_j + \varepsilon$$

B: Background noise; *S:* Signal, composed of θ the true abundance, φ, probe affinity effect and ε measurement error.

RMA Background adjustment

Required because part of the signal results from

- ‘noise’ in the image
- non-specific hybridization

There are many methods of carrying out background subtraction

Here we consider only the default background subtraction method used by RMA

Background subtraction in RMA (model-based method):

- consider the intensity as a sum of normally distributed noise and exponentially distributed signal
- estimate the parameters of these distributions across the array and use that to infer the signal from the probe

RMA: Normalization

- Uses quantile normalization across arrays
- Causes all arrays to have the same distribution of signal intensities

Quantile normalization algorithm

1. Form a matrix with each column of the matrix representing probe expression levels in one sample
2. Sort the columns of the matrix independently
3. For each row of the sorted matrix replace each value in the row with the row average
4. Return the columns to their original orders

Quantile normalization hands-on

Carry out quantile normaliztion on the following vectors

a1: 3 8 9 2 1 7

a2: 30 18 9 11 7 9

a3: 66 4 3 8 2 2

a4: 7 2 3 3 2 6

When finished make a quantile-quantile plot of each pair using the R function qqplot( ) to confirm that you have quantile normalized them correctly.

RMA: Summarization

- Uses a robust method (median polish) to fit estimates of the expression level of each probeset