Statistical analysis of gene expression data

Alex Lewin

Department of Epidemiology and Public Health, Imperial College

BBSRC- EPSRC funded Project

Flexible Bayesian clustering and Partition models for gene expression data

- Biostatistics at Imperial College (St Mary's):

Prof Richardson (PI and coordinator), Dr Marshall, Dr Lewin, Dr Hein

- Statistics at Bristol University:

Prof Green (PI), Dr Ambler

- Microarray Centre at Hammersmith Hospital:

Prof Aitman, Dr Causton

http://www.stats.bris.ac.uk/BGX/

Outline

Low-level Model

↓

Normalisation

↙ ↘

Differential Expression | Clustering Model

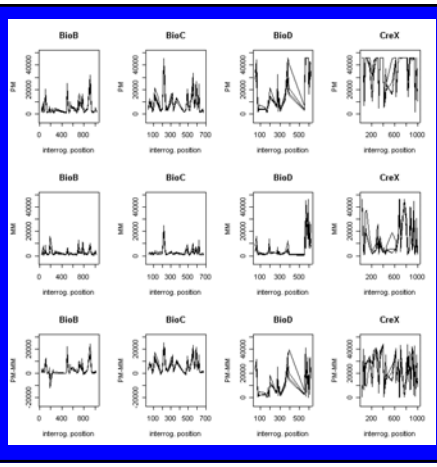
Low-level Exploratory Work

- Affymetrix chips: each gene has a number of probe pairs on one array
- Probe pair: perfect match (pm) and mismatch (mm)
- Pm's and mm's are combined into an expression value for the gene.

Issues

- Consistency of probe pairs within one gene
- Different levels of variability

→ Model for combining pm and mm values



Normalisation

- Overall level of expression on different arrays can be different
- Normalisation brings arrays into line with each other
- Often carried out as a pre-processing step
- We include normalisation as part of a model for differential expression

A simple additive model for normalisation

Notation

- y_{gr} = gene expression measurements for gene g, array r

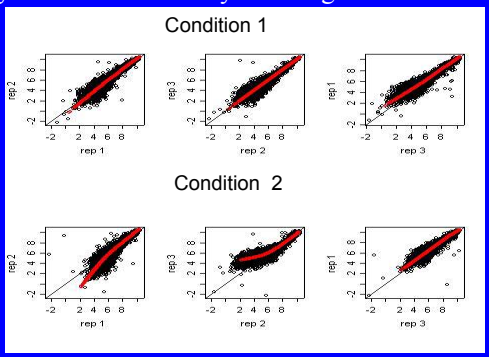
Simple additive model:

- $\ln(y_{gr}) = x_{gr} = \mu_g + c_r + \varepsilon_{gr}$

Here μ_g is the expression level of the g^{th} gene, c_r is the effect of the r^{th} array (normalisation term) and ε_{gr} an error term.

- This model can be elaborated to include main effects of experimental conditions, interaction terms, etc.
- Non linear array effect: $c_r \rightarrow c_r(\mu_g)$

Affymetrix data: Two conditions, 3 replicates each
'Array effect' estimated by local regression techniques



Expression data has been log transformed

Differential expression

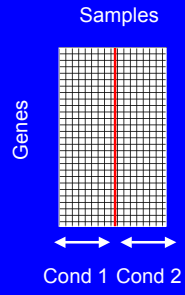
Gene expression data matrix

One common question:

'How does gene expression change under different experimental conditions?'

Analyses strategies rely on:

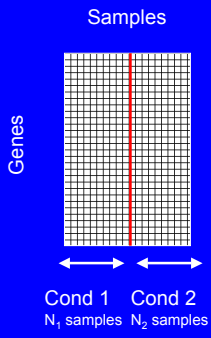
- Hypothesis testing procedures
- Mixture model procedures



Notation:

Log expression for gene g, sample r, condition s (s=1,2) is denoted by: x_{gsr}

$\bar{x}_{gs}$ The average expression level for gene g under Condition s over the N_s samples



Hypothesis testing

- Statistics usually defined from differences or standardised differences

$$d_g^* = \bar{x}_{g2} - \bar{x}_{g1}$$

$$d_g = (\bar{x}_{g2} - \bar{x}_{g1}) / s_g$$

- Define a set of Null Hypotheses

  H_{0g}: $E(d_g) = 0$

Questions:
- How to estimate the variances?
- Control of multiple testing

Estimation of gene variability

- Standard variance estimates based on few replications are highly variable.
- Tusher et al (SAM) propose to add a constant:

  $s_g \rightarrow s_g + s_0$ in the denominator of d_g
- Bayesian estimation: posterior variance estimate is a weighted average of prior mean and sample estimate, informed by the data

Bayesian Estimate of Variance

- Bayesian hierarchical model for the variances

$$\begin{aligned} x_{gr}|\mu_g, c_r, \sigma_g^2 &\sim N(\mu_g + c_r, \sigma_g^2) \\ \sigma_g^2|\beta &\sim IG(1, \beta) \\ \beta &\sim G(\epsilon, \epsilon) \end{aligned}$$

- Variances shrink towards average variance
- Variances are estimated using information from all 8000 x 3 measurements (rather than 3)

Bayesian Model Checking

- Different assumptions for gene variances give very different results
- Bayesian model-checking
  - Gene-specific variance good fit to data
  - Equal variance model has too little variability for the data

Multiple testing problem

- When carrying out a very large number of tests, the probability of at least one Type I error increases sharply
- Usual approach (Bonferroni correction) very conservative
- Other p-adjustment accounting for dependence between genes (Westfall and Young) still conservative

Mixture model approaches

Finite normal mixtures with an unknown number of states (Broet, Richardson, Radvanyi, 2002)

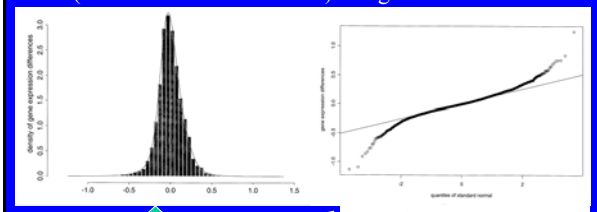
- A gene can be in G different states: down regulated, …, unaffected, …, up-regulated

Bayesian estimation of posterior distribution:

- for number of states
- for the allocation of genes to the states

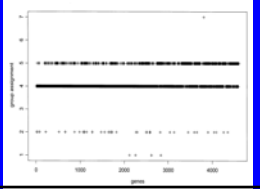
➡ classification of genes in the 'extreme components' based on their posterior probability

Comparison of gene expression in two bladder cancer cell lines (transfected versus control) using mixture model



Distribution of d_g and QQ plot

Classification of the 4608 genes: 1 Down reg. 5 + 45 Up reg.



Broet, Richardson, Radvanyi, 2002

Summary

- Probe pairs highly variable within one gene, mm mirrors pm
- Use information from all genes to estimate gene variances (not just 3 measurements)
- Model checking shows we need gene-specific variances
- Mixture models useful for overcoming multiple-testing problems