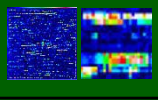
Bayesian Two-way Clustering for Gene Expression Data

Graeme Ambler and Peter Green
University of Bristol
12 July 2003



University of BRISTOL

Motivation: Obvious potential for Bayesian and EB methods in gene expression analysis: can they be made to work?

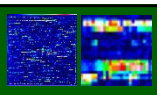
BGX project, BBSRC funded

with [illegible], Clare [illegible]ewin and Anne-Mette H[illegible] (Imperial), in collaboration with Helen Causton and Tim Aitman and colleagues (CSC/IC Microarray Centre)

Model-based, flexible approach to gene expression analysis

University of BRISTOL

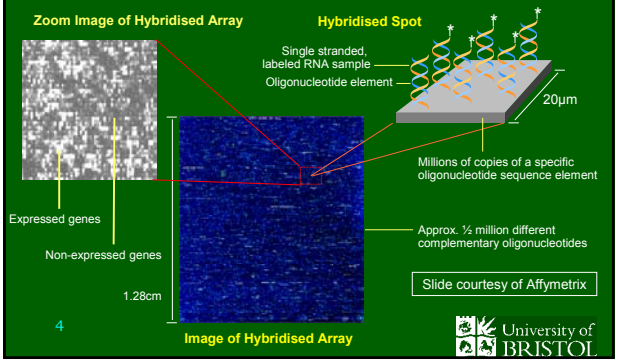
Plan



- Variation and uncertainty in gene expression
- Hierarchical models
- Simultaneous inference
- Common framework, including clustering
- Initial experiments with layer models

University of BRISTOL

Gene expression using Affymetrix chips



University of BRISTOL

Variation and uncertainty

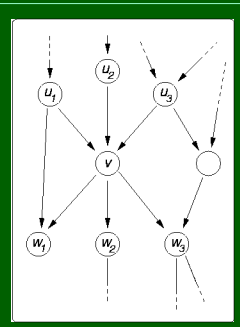
Gene expression data (e.g. Affymetrix™) is the result of multiple sources of variability

- condition/treatment
- biological
- array manufacture
- imaging
- technical
- within/between array variation
- gene-specific variability

University of BRISTOL

Hierarchical models

Variables at several levels - allows modelling of complex systems



University of BRISTOL

Bayesian hierarchical models

One of the most important benefits of the Bayesian approach has nothing much to do with having real quantitative prior information

- it has more to do with the structures connecting variables
- especially when there is uncertainty at more than one level

University of BRISTOL

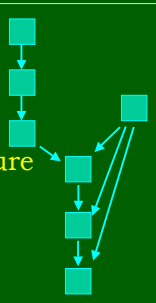
The Bayes orthodoxy

- Should avoid a plug-in approach -- all sources of variation should be assimilated
- Propagates uncertainty
- 'Borrows strength' - shares out information - according to principle
- Avoids over-optimistic inference

University of BRISTOL

Gene expression is a hierarchical process

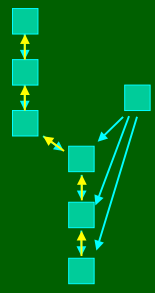
- Substantive question
- Experimental design
- Sample preparation
- Array design & manufacture
- Gene expression matrix
- Probe level data
- Image level data



University of BRISTOL

Bayes in hierarchical models

- The arrows represent (top down) model specification, not the order in which operations are performed
- Once specified, model unknowns should be estimated simultaneously
- (We cannot yet claim all of this is practical in gene expression)



University of BRISTOL

Additive models for (log-) gene expression

The simplest model: gene + sample

$$y_{gs} = \alpha_g + \beta_s + \varepsilon_{gs}$$

g=gene
s=sample/condition

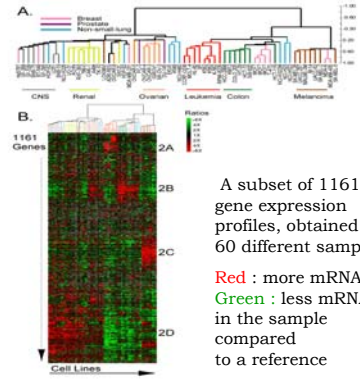
Under standard conditions, the (least-squares) estimates of gene effects are

$$\hat{\alpha}_g = \bar{y}_{g.} - \bar{y}_{..}$$

The model generates the method, and in this case performs a simple form of normalisation

University of BRISTOL

Hierarchical clustering of samples



A subset of 1161 gene expression profiles, obtained in 60 different samples

Red : more mRNA
Green : less mRNA in the sample compared to a reference

The gene expression profiles cluster according to tissue of origin of the samples

Ross et al, Nature Genetics, 2000

University of BRISTOL

Non-model-based clustering

- Many clustering algorithms have been developed and used for exploratory purposes
- They rely on a measure of 'distance' (dissimilarity) between gene or sample profiles, e.g. Euclidean
- Hierarchical clustering proceeds in an agglomerative manner: single profiles are joined to form groups using the distance metric, recursively
- Good visual tool, but many arbitrary choices → care in interpretation!

University of BRISTOL

Model-based clustering

- Build the cluster structure into the model, rather than estimating gene effects (say) first, and post-processing to seek clusters
- Bayesian setting allows use of real prior information where it is exists (biological understanding of pathways, etc, previous experiments, …)

University of BRISTOL

A common framework for specifying gene expression models

For ease of exposition, consider only gene expression matrix y_{gs}

g=gene s=sample/condition

with no structure to samples

(although incorporating experimental structure is a key goal for later)

University of BRISTOL

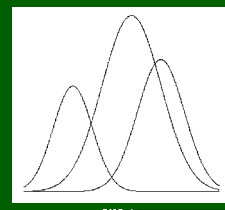
Clustering via additive model

(single sample first!)

$$y_g = \alpha + \varepsilon_g$$

g=gene

$$y_g = \alpha + \gamma_{T_g} + \varepsilon_g$$



T_g= unknown cluster to which gene g belongs
This is a mixture model

University of BRISTOL

Clustering via additive model

(multiple samples)

$$y_{gs} = \alpha_g + \beta_s + \varepsilon_{gs}$$

s=sample/condition
g=gene

$$y_{gs} = \alpha_g + \beta_s + \gamma_{T_g s} + \varepsilon_{gs}$$

T_g= unknown cluster to which gene g belongs
→ clustering of gene profiles

University of BRISTOL

Clustering via additive model

$$y_{gs} = \alpha_g + \beta_s + \gamma_{T_g s} + \varepsilon_{gs}$$

T_g=cluster to which gene g belongs

$$y_{gs} = \alpha_g + \beta_s + \delta_{gU_s} + \varepsilon_{gs}$$

U_s=cluster to which sample s belongs

University of BRISTOL

Two-way Clustering via additive model

$$y_{gs} = \alpha_g + \beta_s + \gamma_{T_g s} + \varepsilon_{gs}$$

$$y_{gs} = \alpha_g + \beta_s + \delta_{gU_s} + \varepsilon_{gs}$$

$$y_{gs} = \alpha_g + \beta_s + \gamma_{T_g s} + \delta_{gU_s} + \varepsilon_{gs}$$

or

$$y_{gs} = \alpha_g + \beta_s + \gamma_{T_g U_s} + \varepsilon_{gs}$$

University of BRISTOL

Lazzeroni and Owen 'Plaid' model

$$y_{gs} = \alpha_g + \beta_s + \gamma_{T_g s} + \varepsilon_{gs}$$

Now write ρ_{gh}=1 if and only if T_g=h, 0 otherwise

$$y_{gs} = \alpha_g + \beta_s + \sum_h \rho_{gh}\gamma_s^{(h)} + \varepsilon_{gs}$$

h denotes a 'cluster', 'block' or 'layer' - and now we allow them to overlap

.... continued over ....

University of BRISTOL

'Plaid' model

$$y_{gs} = \alpha_g + \beta_s + \sum_h \rho_{gh}\gamma_s^{(h)} + \varepsilon_{gs}$$

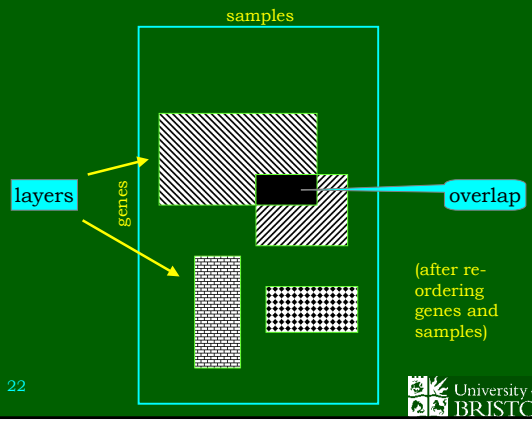
$$y_{gs} = \sum_h \rho_{gh}\kappa_{sh}\gamma_{gs}^{(h)} + \varepsilon_{gs}$$

h denotes a 'cluster', 'block' or 'layer' – pathway?

ρ_{gh}= 0 or 1 and κ_{sh}= 0 or 1

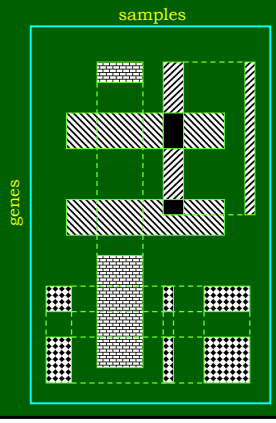
$$\gamma_{gs}^{(h)} = \mu^{(h)} + \alpha_g^{(h)} + \beta_s^{(h)}$$

University of BRISTOL



(after re-ordering genes and samples)

University of BRISTOL



University of BRISTOL

MacKay and Miskin model

Instead of $$y_{gs} = \sum_h \rho_{gh}\kappa_{sh}\gamma_{gs}^{(h)} + \varepsilon_{gs}$$

where h denotes a 'cluster', 'block' or 'layer';

ρ_{gh}= 0 or 1 and κ_{sh}= 0 or 1

MacKay and Miskin take simply

$$y_{gs} = \sum_h a_s^{(h)} b_g^{(h)} + \varepsilon_{gs}$$

University of BRISTOL

Markov chain Monte Carlo (MCMC) computation

- Fitting of Bayesian models hugely facilitated by advent of these simulation methods
- Produce a large sample of values of all unknowns, ≈ from posterior given data
- Easy to set up for hierarchical models
- BUT can be slow to run (for many variables!)
- and can fail to converge reliably

University of BRISTOL

Simultaneous inference

- An important example of the flexibility of MCMC computation in a Bayesian model: inference about several unknowns at once.
- e.g. not only 'which gene has the biggest estimated differential effect?', but also 'how probable is it that this gene has the biggest differential effect?'

University of BRISTOL

Contact details

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

Graeme.Ambler@bristol.ac.uk

P.J.Green@bristol.ac.uk

University of BRISTOL