Aspects of Statistical Modelling & Data Analysis in Gene Expression Genomics

Mike West
Duke University
Papers, software, many links:
www.isds.duke.edu/~mw

ABS04 web site: Lecture slides, stats notes, papers, data, links:
www.isds.duke.edu/~mw/ABS04

Integrated Cancer Biology Program
icbp.genome.duke.edu

Genome Institute @ Duke
www.genome.duke.edu
#1
Genomics, Microarrays, Data:  
Big picture

#2
Bayesics - Regression and Shrinkage:  
Gene expression as predictors

#3
Patterns and Factors:  
Prediction via pattern profiling

#4
Sparse Modelling:  
Regression subset-structure uncertainty

#5
Sparse Models and Profiling:  
Gene expression as response: Designed experiments

#6
Sparse Models and Profiling:  
Gene expression as response: Latent factor models
(Yet another) history of life as we know it...

Homo Apriorius, Homo Pragmaticus, Homo Frequentistus, Homo Sapiens, Homo Bayesianus
#1

Genomics, Microarrays, Data:  
Big picture

#2

Bayesics - Regression and Shrinkage:  
Gene expression as predictors

#3

Patterns and Factors:  
Prediction via pattern profiling

#4

Sparse Modelling:  
Regression subset-structure uncertainty

#5

Sparse Models and Profiling:  
Gene expression as response: Designed experiments

#6

Sparse Models and Profiling:  
Gene expression as response: Latent factor models
Transitions in Biology: Data and Observation

Observational science

\[\downarrow\]

Molecular science

\[\downarrow\]

Genomic science

Data: Scale, Complexity -

Computational & Statistical Science
Low resolution phenotypes
“Small worlds”, small data

Breast cancer:

- Lymph node involvement
- Hormone receptor status
- Tumor size
- Visual assessment
Higher resolution
Genome scale, big data

Increased understanding

\[ p(X) \]

\[ p(Y \mid X) \]

Improved prediction
Modelling Genomic Data for Prediction

Genomic Medicine
Personalised Prognostics

Integrated Clinico-Genomic Models

(Breast cancer - Pittman et al PNAS 04)
Genomic Data: Opportunities and Challenges

Synthesis ... or Translation

Gene expression profiles: Signatures of states

Laboratory/In vitro

Laboratory/Animal models

Human Observational Studies

Human Clinical Studies
Genomic Data?

- Genotypes
- SNPs
- Haplotypes
- Serum expression
- DNA methylation
- Gene expression

Biological/disease state...

- Genotypes
- SNPs
- Haplotypes
- Serum expression
- DNA methylation
- Gene expression

‘host’...

- Serum-metabolomics
- Environmental factors
- Clinical markers
- Serum proteomics

- Metabolomics
- Proteomics
- DNA copy (CGH)
Affymetrix DNA Microarray Data

- Gene probesets
- Imaging/Scanning
- 100Mb raw data

Expression intensity estimates X+/−S
p genes, n samples

Background, noise, gross defects, ...
Cross-hybridization
Sample-sample normalisation
'Low level' data processing, analysis

West et al 2001
Wong & Li (dChip) 2001
Bolstad, Irizarry, Speed et al 2003a,b
RMA estimates - www.bioconductor.org
First Generation Microarrays: Messy Data

Multiple expression data sets

Multiple array technologies

Multiple species: genome A - B mappings

Same array platform: sample/lab/study/gene effects

Assay/batch/reagent/hybridisation sensitivities

Sporadic - Sparse
#1
Genomics, Microarrays, Data:
Big picture

#2
Bayesics - Regression and Shrinkage:
Gene expression as predictors

#3
Patterns and Factors:
Prediction via pattern profiling

#4
Sparse Modelling:
Regression subset-structure uncertainty

#5
Sparse Models and Profiling:
Gene expression as response: Designed experiments

#6
Sparse Models and Profiling:
Gene expression as response: Latent factor models
**Gene expression as covariates (predictors)**

Molecular phenotyping:

- Predict aggressive vs. benign
- Disease susceptible vs. resistance
- Drug/treatment response
- Finding genes linked to response
- Patterns of association among genes
- Signatures of effect – multiple genes

\[ p(y|x) \]
Regression Models and Shrinkage

\[ z = H\beta + \nu, \quad \nu \sim N(0, \sigma^2 I) \]

Phenotype \( z \)

\( H = \text{Subsets of genes} \)

**LSE:**

\[
\hat{\beta} = B_*^{-1} H' z \\
B_* = H' H
\]

**Prior:**

\[ \beta|C \sim N(0, C^{-1}) \]

**Posterior:**

\[ \beta|z, C \sim N(b, \sigma^2 B^{-1}) \]

**Shrinkage:**

\[
b = B^{-1} H' z \\
B = \sigma^2 C + H' H
\]

(minimal) Bayes: Shrinkage priors

Decision theory

Regularisation - Ridge regression

Key with many predictors

Relevance of zero-mean location
Degrees and Dimensions of Shrinkage

\[ \beta \sim N(0, C^{-1}) \]

\[ C^{-1} = \tau I, \quad \tau \sim \text{InvGamma} \]

LSE as limiting case - no shrinkage: \( \tau^{-1} \rightarrow 0 \)

Shrinks when it matters - weak/no association

Acts against over-fitting, improves stability and robustness in prediction

\[ C^{-1} = \text{diag}(\tau_1, \ldots, \tau_k), \quad \tau_j \sim \text{InvGamma} \]

Multiple shrinkage

\[ \beta' = (\beta_1, \ldots, \beta_k) \]

\[ \beta_j \sim N(0, \tau_j) \]

“Shrinks out” irrelevant covariates
Simulate Posterior:
Iteratively resample conditional posteriors

Sample means, histograms
MC approximation of posterior

\[ p(\beta|z, C) = N(b, \sigma^2 B^{-1}) \]

\[ p(C|z, \beta) = \prod_{j=1}^{k} p(\tau_j|\beta_j) \]
Computation: MCMC in Regression

Modules in MCMC
e.g. response error variance

\[
\begin{align*}
p(\beta | z, C, \sigma^2) &= N(b, \sigma^2 B^{-1}) \\
p(C | z, \beta, \sigma^2) &= \prod_{j=1}^{k} p(\tau_j | \beta_j) \\
p(\sigma^2 | z, \beta, C) &= \text{InvGamma}
\end{align*}
\]
Binary Regression

Binary = thresholded latent continuous
probit~normal, logit~logistic, ...

\[ Pr(y_i = 1) = \Phi(h_i^T \beta) \]

Natural model/interpretation
Computationally nice

\[ Pr(y_i = 1) = Pr(z_i > 0), \quad z_i \sim N(h_i^T \beta, 1) \]

\[ z = H\beta + \nu, \quad \nu \sim N(0, I) \]
Computation: MCMC in Binary Regression

Linear model if z known

Add module to impute latent z

MC samples for z

Easy summary, prediction

\[
p(\beta|z, C) = N(b, B^{-1})
\]

\[
p(C|z, \beta) = \prod_{j=1}^{k} p(\tau_j|\beta_j)
\]

\[
p(z|y, \beta) = \prod_{i=1}^{n} p(z_i|y_i, \beta)
\]
y=0/1 (ER -/+)  
Protein assay  
Immunohistochemical staining  
0/1 (0-3)

Basic Examples: Breast Cancer Data

ER - (O) Estrogen Receptor Status  
HER2 hormone status  
Lymph node (recurrence risk) status

Frozen tumour: Gene expression  
Higher resolution  
Future clinical tests: Pr(ER+)

ER positive tumour

IHC for Estrogen Receptor  
(~60x magnification)

nuclei of breast epithelial cells  
cytoplasm of breast epithelial cells  
brown-red & pink ~ ER+

collagen  
nucleii of stromal cells;  

SemStat05 - Warwick - Sept 11 & 12th 2005
Prediction and \{Gene, Variable, Feature\} Selection

Leave-one-out Cross-Validation (CV) analysis:

“Honest” assessment of precision

Heterogeneity, small samples

Feature/Variable selection

Critical (dominant) component of predictive assessment

(End of Document)
Predicting lymph node status

Pre-selection of 100 genes vs. “Honest” CV predictions

Large p:
Small models-
Sparsity

Variable selection,
Uncertainty

Complex
interdependencies

Multiplicities
#1 Genomics, Microarrays, Data: Big picture

#2 Bayesics - Regression and Shrinkage: Gene expression as predictors

#3 Patterns and Factors: Prediction via pattern profiling

#4 Sparse Modelling: Regression subset-structure uncertainty

#5 Sparse Models and Profiling: Gene expression as response: Designed experiments

#6 Sparse Models and Profiling: Gene expression as response: Latent factor models
Many Related Predictors: Patterns as Predictors

Patterns of coordinately expressed genes:
  • Signatures

Metagenes
PCA, SVD of expression data

Biologically selected gene subsets
Trained subset selection
Clusters

Cardiovascular disease: High/Low
(Seo, West et al 2004)
Empirical Factor Regression

**SVD:**

\[ X = ADF \]

**PCA:**

\[ XX' = AD^2 A' \]

\[ \beta_f = DA' \beta_x \]

\[ \dim(\beta_f) = n << p = \dim(\beta_x) \]

\[ z = X' \beta_x + \nu \]

\[ z = F' \beta_f + \nu \]

Genes X

Metagene factors F

Patterns-Factors “underlying” X are predictors

X variable set selection

\( p=n: \) Shrinkage priors key

F variable selection
Expression Profiles: Signatures of States

Translation of characterising genomic patterns

Predictive profiling: oncogenic pathway deregulation

(Huang et al 03, Black et al 03)
Out-of-sample prediction

Cell line derived signatures predict differences in oncogenic activity in mouse tumours

c-Myc up-expression

Metagene: gene subset & pattern as a predictor

(Huang et al 03, Black et al 03)
Oncogene Sub-Pathway Profiles: Translation

Single Oncogenes
- pathway characterisation
- potential targets

Cell lines signatures
Human lung cancers (ovarian, breast)

Clinical prognostic
- clinical evaluation
- therapeutic evaluation

(Bild et al 05)
Metagenes in Clinico-Genomic Prognostic Models

Genomic Medicine
Personalised Prognostics

Gene expression clustering
Metagene factors

Non-linear regressions - CART models

Integration:
non-genomic predictors

(Breast cancer - Pittman et al PNAS 04)
#1 Genomics, Microarrays, Data:
Big picture

#2 Bayesics - Regression and Shrinkage:
Gene expression as predictors

#3 Patterns and Factors:
Prediction via pattern profiling

#4 Sparse Modelling:
Regression subset-structure uncertainty

#5 Sparse Models and Profiling:
Gene expression as response: Designed experiments

#6 Sparse Models and Profiling:
Gene expression as response: Latent factor models
$$(z_i|\beta) \sim N(h_i^T\beta, \sigma^2)$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Variable inclusion uncertainty

Large p: parsimony

Sparsity:

$$\beta_j \sim (1 - \pi)\delta_0(\beta_j) + \pi N(\beta_j|0, \tau)$$

Augment:

$$\gamma_j \sim Ber(\pi)$$

$$\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{pmatrix}$$

MCMC computation:

$$p(\beta, \gamma, \tau, \pi|Z)$$
Large p - Shrinkage and Sparsity

Model-based, automatic shrinkage - Simultaneous "multiple tests"

Multiple shrinkage: conservative, parsimonious
Decision theory/false discovery?
Estimation versus Decision?

\[ \pi_j^* = Pr(\gamma_j = 1|Z) = Pr(\beta_j \neq 0|Z) \]
\[ p(\beta_j|\beta_j \neq 0, Z) \]

Model/subset probabilities:
\[ Pr(M_\gamma|Z) \]

Issues:
Collinearities
Multiple related models
Computation with very large p

(Clyde & George StatSci 04)
Stochastic Search Methods

MCMC “local search” inspired

Good models “near” good models

Add/drop/replace variables

Move by sampling new model

Shoot out ALL neighbours:
"local proposals"

Swiftly find high probability regions of model space

Catalogue of many “good” models

Parallelisation

KEY: easily compute

\[ \propto \Pr(M_\gamma | Z) \]

(Hans, Dobra, West 05; Rich et al 2005 - p=8400)
Brain cancer expression: \( p=8400 \)

**Survival regressions:**
- multiple related 3-5 gene subsets
- key cellular motility/infiltration genes
- regression model uncertainty in prediction

(Cancer Research, 05)
Sparsity -
Regression variable in/out probabilities

Dimension -
Implicit in Bayesian & other likelihood-based analyses
(cf. BIC)
p=8400

Cascade of regression models:
- Models to predict/explain gene expression for survival predictive genes
- and so on ...

Generate Directed Acyclic Graphical models (DAGs) of association patterns in gene expression

Exploratory data analysis, visualization uses

http://graphexplore.cgt.duke.edu

(Cancer Research, 05)
EGFR
Brain cancer gene expression
Duke Keck Center for Neurooncogenomics

(Dobra et al JMVA, 04; Jones et al Stat Sci 05)
Papers, software, many links:
www.isds.duke.edu/~mw

ABS04 web site: Lecture slides, stats notes, papers, data, links:
www.isds.duke.edu/~mw/ABS04

Integrated Cancer Biology Program
icbp.genome.duke.edu

Genome Institute @ Duke
www.genome.duke.edu