

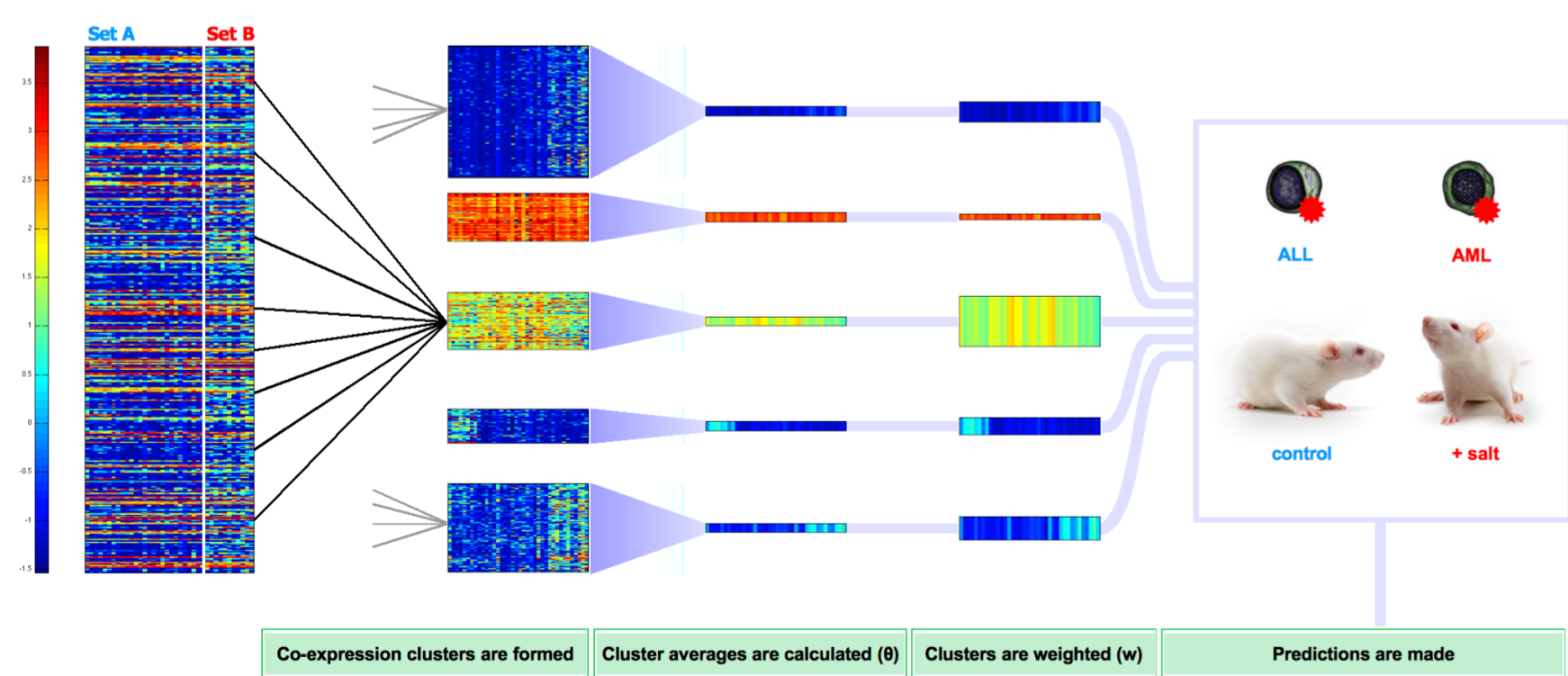
We have developed an alternative method for gene selection that combines model based clustering and binary classification. By averaging the covariates within the clusters obtained from model based clustering, we define “meta-covariates” and use them to build a probit regression model, thereby selecting clusters of similarly behaving genes, aiding interpretation. This simultaneous learning task is accomplished by an EM algorithm that optimises a single likelihood function which rewards good performance at both classification and clustering. We explore the performance of our methodology on a well known leukaemia dataset and use the Gene Ontology to interpret our results.

## Introduction

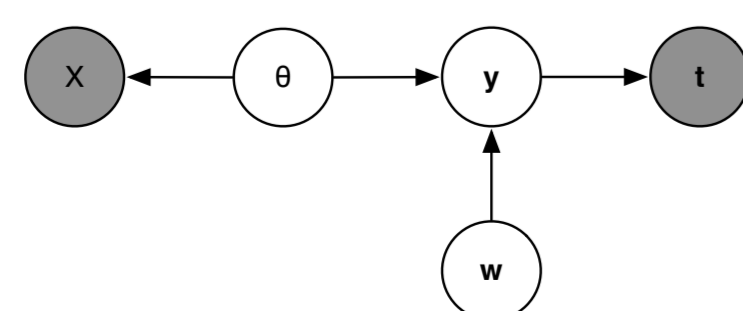
**Microarrays** allow the simultaneous measurement of gene expression in a biological sample. In analysing these data with a view to making predictions, there are significant statistical challenges: the transcription levels of tens of thousands of genes are measured often in a very small number of replicates and many of the genes will have similar expression patterns. Patterns of ‘co-expression’ can be exploited to reduce the dimensionality of the data by forming **clusters of similarly behaving genes**, with predictions being made on some average of each cluster [2; 3].

Here, we describe a novel procedure that potentially improves the **classification of gene expression profiles through coupling with the method of model based clustering**. By combining inter-predictor correlations with predictor-outcome correlations, we can potentially improve prediction performance. Furthermore, **we obtain clusters of co-expressed genes that are relevant to the response** (see Figure 1).

## The model



**Figure 1:** An overview of the method. Co-expression clusters are identified and represented by a cluster mean. Each cluster is assigned a weight according to its ability to distinguish between set A and set B data. Prediction performance is used to update the clustering structure and the regression weights.



**Figure 2:** Graphical representation of the conditional dependences within the meta-covariate classification model (see joint distribution in the Model details)

## Model details

$X$	Design matrix	$N \times D$	$X = [x_1, \dots, x_N]^T$
$t$	Response	$N \times 1$	$t_n \in \{-1, 1\}$
$\theta$	Cluster means	$K \times N$	the ‘meta-covariates’
$z$	Clustering latent variables	$D \times K$	
$\gamma$	Clustering responsibilities	$D \times K$	equivalent to $E(z_{dk})$
$w$	Regression coefficients	$K \times 1$	
$y$	Classification auxiliary variable	$N \times 1$	$+ve$ if $t_n = 1$ , $-ve$ if $t_n = -1$

**Joint distribution**  $p(t, y, X, \theta, w) = p(t, y | \theta, w) p(X | \theta) p(\theta) p(w)$ . (see Figure 2).

**Classification model**  $t_n = \begin{cases} 1 & \text{if } y_n > 0 \\ 0 & \text{otherwise.} \end{cases}$

$$y_n = w^T \theta_n + \epsilon_n \text{ where } \epsilon_n \sim \mathcal{N}(0, 1).$$

**Clustering model**  $\Rightarrow p(x) = \frac{1}{K} \sum_{k=1}^K \mathcal{N}(x | \theta_k, I)$ .

**Prior distributions**  $p(\theta) = \prod_{k=1}^K \mathcal{N}(\theta_k | \theta_0, hI)$ ,  $p(w) = \mathcal{N}(w | 0, I)$ .

## EM Algorithm

**E-step**

$$\gamma(z_{dk}) = \frac{\exp\{-\frac{1}{2}\|x_d - \theta_k\|^2\}}{\sum_{j=1}^K \exp\{-\frac{1}{2}\|x_d - \theta_j\|^2\}}, \quad E(y_n) = \begin{cases} w^T \theta_n + \frac{\phi(-w^T \theta_n)}{1 - \Phi(-w^T \theta_n)} & \text{if } t_n = 1 \\ w^T \theta_n - \frac{\phi(-w^T \theta_n)}{\Phi(-w^T \theta_n)} & \text{otherwise.} \end{cases}$$

**M-step**

$$\theta_k = \frac{(E(y) - \theta^T w_{-k}) w_k + X \gamma_k + \frac{1}{h} \theta_0}{w_k^2 + \sum_{d=1}^D \gamma(z_{dk}) + \frac{1}{h}}, \quad w = \left( \theta \theta^T + \frac{1}{l} I \right)^{-1} \theta E(y).$$

## A test set: Leukaemia data

Bone marrow or peripheral blood samples obtained from 72 AML and ALL patients. A training set of 38 samples (27 ALL and 11 AML) and a test set of 34 samples (20 ALL and 14 AML) were used [1].

$K = 21$  is used as it gives the minimal test error. The clusters (Table 1) give perfect discrimination in the training set and **competitive performance in the test set (two misclassifications)** (c.f., [2; 3]).

Note the **limited influence of control probes**; the **differing functional repertoires of the clusters**; and the association of **metal ion binding** and **receptor activity** with positively weighted clusters. Important genes **Zyxin** and **Cystatin C** are maximally influential.

Future work includes a **more general covariance structure** in the clustering model; extension to **multinomial classification**; development of a **Bayesian sampler** to sample from the marginal and conditional distributions and **more sophisticated functional analysis** of the resulting clusters.

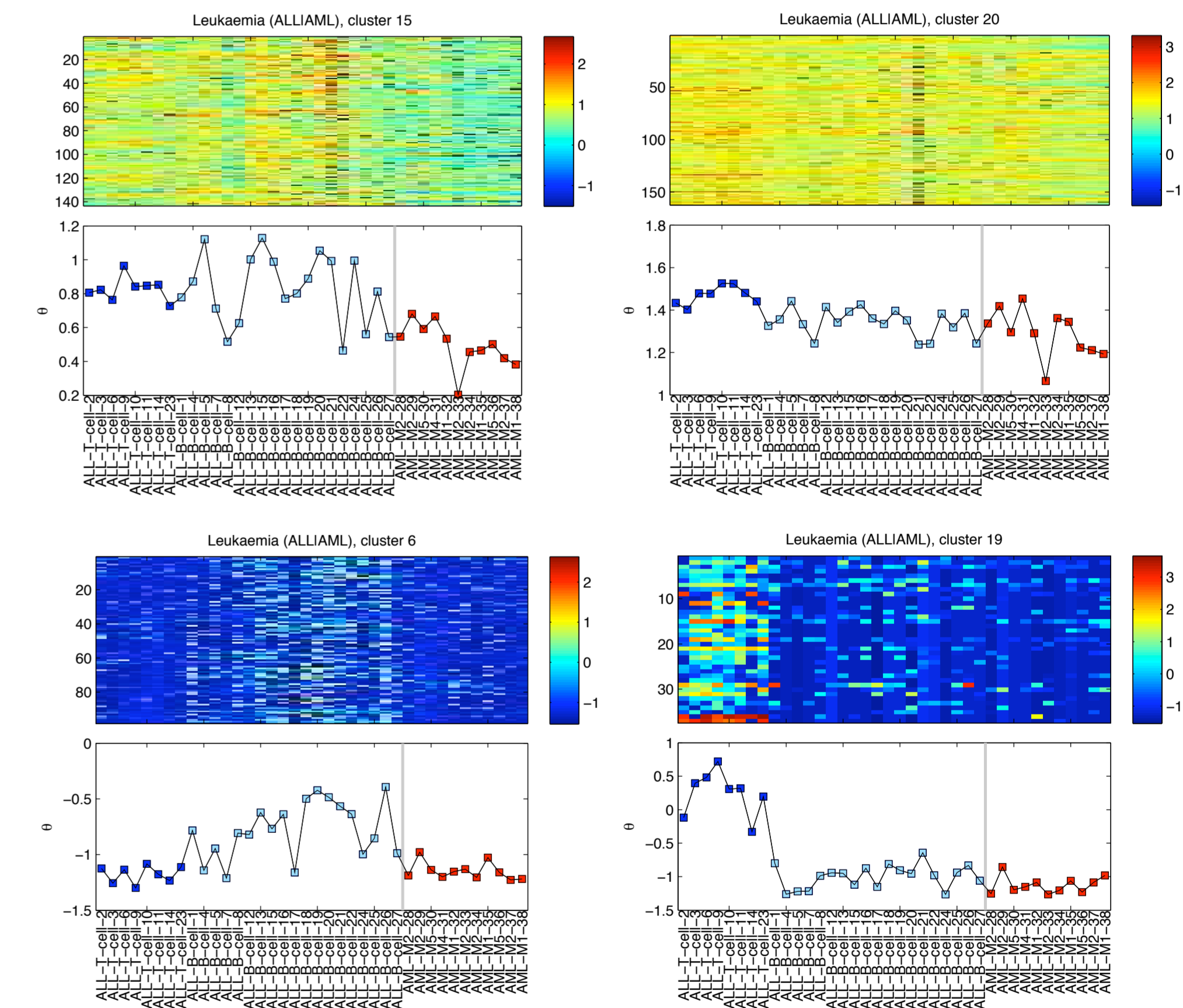
**Table 1:** The best clusters ( $K = 21$ )

Cluster	Probes	Controls	w
15	61	0	3.87
20	101	1	-3.79
16	240	0	3.21
6	240	1	-3.08
10	230	0	-2.66
21	210	3	2.46
4	253	0	-1.88
3	20	0	-1.22
11	189	1	-1.10
1	20	0	1.00
18	17	0	-0.95
12	210	1	0.88
13	228	0	0.79
19	267	1	0.75
5	182	0	0.55
14	230	0	0.55
17	213	0	-0.50
7	110	2	-0.37
2	486	4	0.22
8	60	4	0.16
9	4	4	-0.15

**Table 2:** Characterising the clusters functionally

Cluster	w	MIB	D/RB	RA	ER	RB	KA	TMT	TRR
15	3.87	.	.	.	.	.	.	.	.
16	3.21	.	.	.	.	.	.	.	.
21	2.46	.	.	.	.	.	.	.	.
1	1.00	.	.	.	.	.	.	.	.
11	-1.10	.	.	.	.	.	.	.	.
3	-1.22	.	.	.	.	.	.	.	.
4	-1.88	.	.	.	.	.	.	.	.
10	-2.66	.	.	.	.	.	.	.	.
6	-3.08	.	.	.	.	.	.	.	.
20	-3.79	.	.	.	.	.	.	.	.

MIB = metal ion binding  
RA = receptor activity  
RB = receptor binding  
TMT = transmembrane transport  
D/RB = DNA or RNA binding  
ER = enzyme regulation  
KA = kinase activity  
TRR = transcription regulation  
• = over-representation  
• = under-representation  
• = conflicting results



## References

- [1] T. R. Golub *et al.* *Science*, 286:531–537, 1999.
- [2] B. Hanczar *et al.* *SIGKCC Explorations*, 5(2):23–30, 2003.
- [3] M. Y. Park *et al.* *Biostatistics*, 8:212–227, 2007.