A Fully Bayesian Model to Cluster Gene Expression Profiles

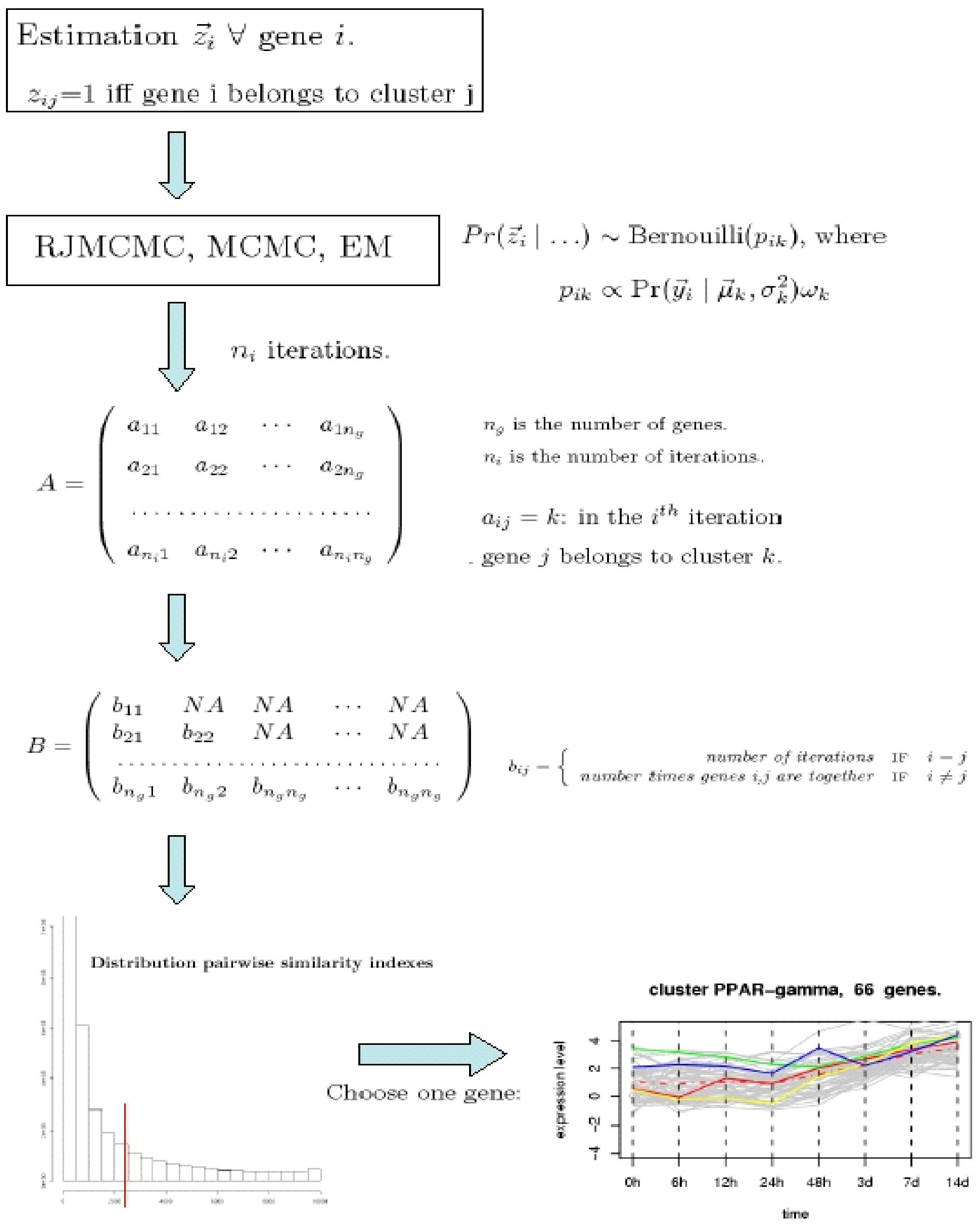
F. Sanchez-Cabo[1,2], H. Hackl[2], S. Hubbard[1], G. Stocker[2], Z. Trajanoski[2], O. Wolkenhauer[3], C. Vogl[4]

[1]Department of Biomolecular Sciences, UMIST, PO Box 88, Manchester M60 1QD, UK
[2]Institute for Genomics and Bioinformatics, Graz University of Technology, 8010 Graz, Austria
[3] Systems Biology and Bioinformatics Group, University of Rostock, 18051 Rostock, Germany
[4] Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna, A-1210 Vienna, Austria

1. Introduction and Objectives

Clustering methods are popular screening tools for microarray data in order to identify subgroups of genes that share common regulatory elements, a common function or a common cellular origin. But the most popular clustering algorithms, *e.g.* K-means, require *a priori* determination of the number of clusters. Results strongly depend on this choice. Additionally, microarray data are inherently noisy and many measurements are missing, which results in the loss of a great amount of information with most earlier methods. Therefore, we propose a probabilistic model in which the number of clusters and missing values are treated as random variables that can be estimated from the available data using the *Reversible Jump Markov Chain Montecarlo (RJMCMC)* simulation scheme [1].

2. Methods



Estimation $\vec{z}_i$ ∀ gene i.

z_{ij}=1 iff gene i belongs to cluster j

RJMCMC, MCMC, EM

$Pr(\vec{z}_i \mid \ldots) \sim \text{Bernouilli}(p_{ik})$, where

$$p_{ik} \propto \Pr(\vec{y}_i \mid \vec{\mu}_k, \sigma_k^2)\omega_k$$

n_i iterations.

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n_g} \\ a_{21} & a_{22} & \cdots & a_{2n_g} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n_i1} & a_{n_i2} & \cdots & a_{n_in_g} \end{pmatrix}$$

n_g is the number of genes.
n_i is the number of iterations.

$a_{ij} = k$: in the i^{th} iteration
. gene j belongs to cluster k.

$$B = \begin{pmatrix} b_{11} & NA & NA & \cdots & NA \\ b_{21} & b_{22} & NA & \cdots & NA \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{n_g1} & b_{n_g2} & b_{n_gn_g} & \cdots & b_{n_gn_g} \end{pmatrix}$$

$$b_{ij} - \begin{cases} \textit{number of iterations} & \text{IF } i - j \\ \textit{number times genes i,j are together} & \text{IF } i \neq j \end{cases}$$

Distribution pairwise similarity indexes

Choose one gene:

Figure 1: Sketch of the RJMCMC-clustering algorithm. The key is the estimation of the posterior distribution of the indicator variable z_i. If the RJMCMC simulation scheme is used, the number of clusters does not need to be specified in advance. Otherwise MCMC and EM algorithm [2] can be applied for estimation.

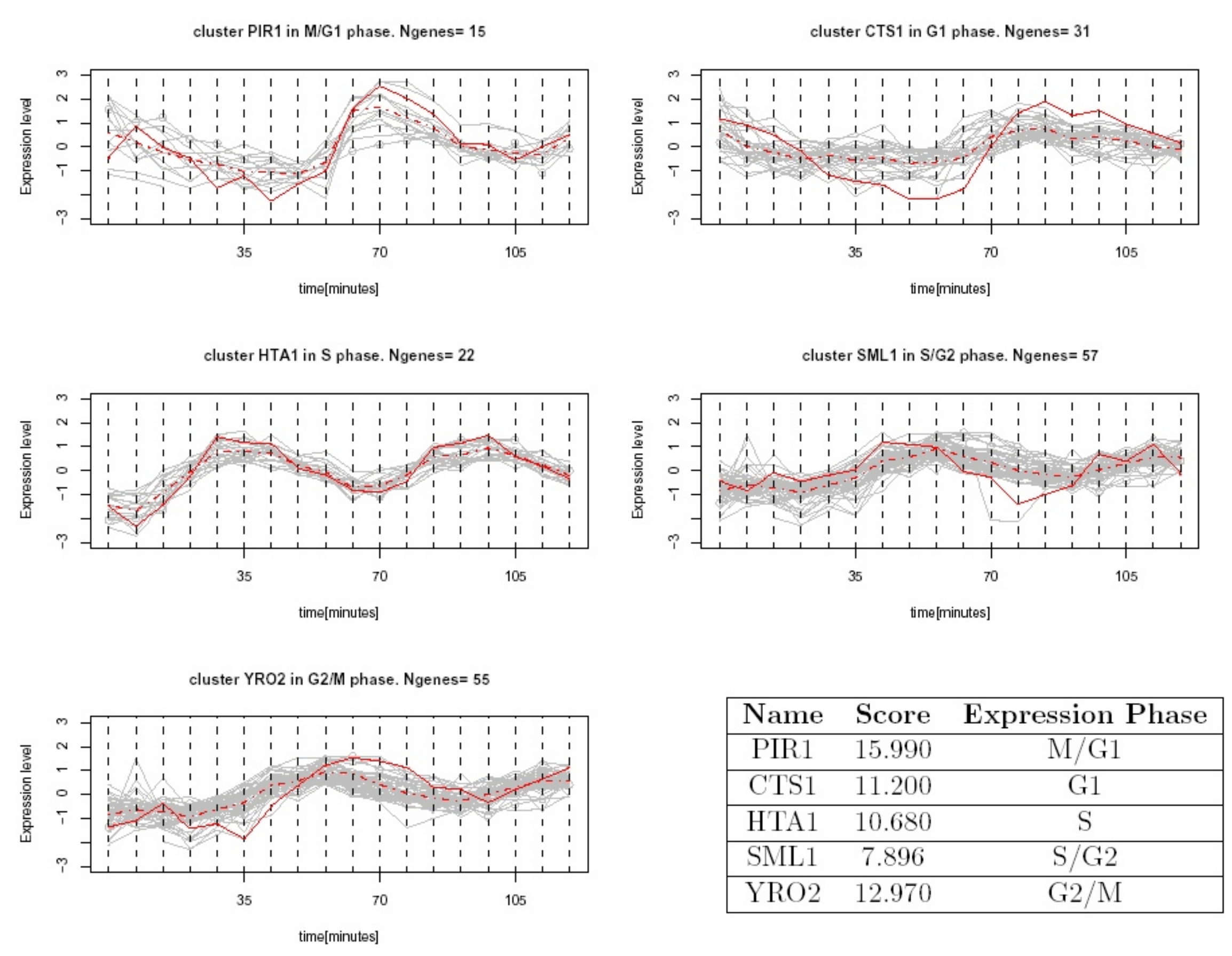
3. Results

The methodology described in brief in Section 2 was applied to two data sets: A public data set performed to discover cell-cycle regulated genes in the yeast *S.Cerevisae* [3] and the data from a microarray experiment carried out to discover new targets and transcription factors involved in adipogenesis [4].

Yeast Cell Cycle regulated experiment [3]

Cluster	♯ genes	EM50	EM40	kmeans=40	MC50	MC40	RJMCMC
MCM	38	14	13	12	**21**	18	19
CLB2	36	**29**	23	23	28	28	**29**
MET	20	13	14	8	15	13	**17**
Hist	9	**9**	**9**	**9**	**9**	**9**	**9**
Y	31	**30**	**30**	29	**30**	**30**	**30**
CLN2	58	38	**55**	34	**55**	36	**55**
SIC1	27	15	17	21	**27**	**27**	21

Table 1: Comparison of different deterministic (k-means), frequentist (ECM) and bayesian (MCMC, RJMCMC) clustering algorithms. From the clusters of biologically related genes found by Spellman *et. al* [3] the table displays how many of them were found together with each one of the clustering algorithm studied.



Name	Score	Expression Phase
PIR1	15.990	M/G1
CTS1	11.200	G1
HTA1	10.680	S
SML1	7.896	S/G2
YRO2	12.970	G2/M

Figure 2: RJMCMC-clusters generated by the five genes exhibiting the highest cell-cycle regulated score in the different cell-cycle phases according to Spellman *et al*. [3] (red solid line). The five "main genes" and their score are displayed in the table. The red broken line is the centroid of the cluster.

Transcriptional profiling of adipogenesis [4]

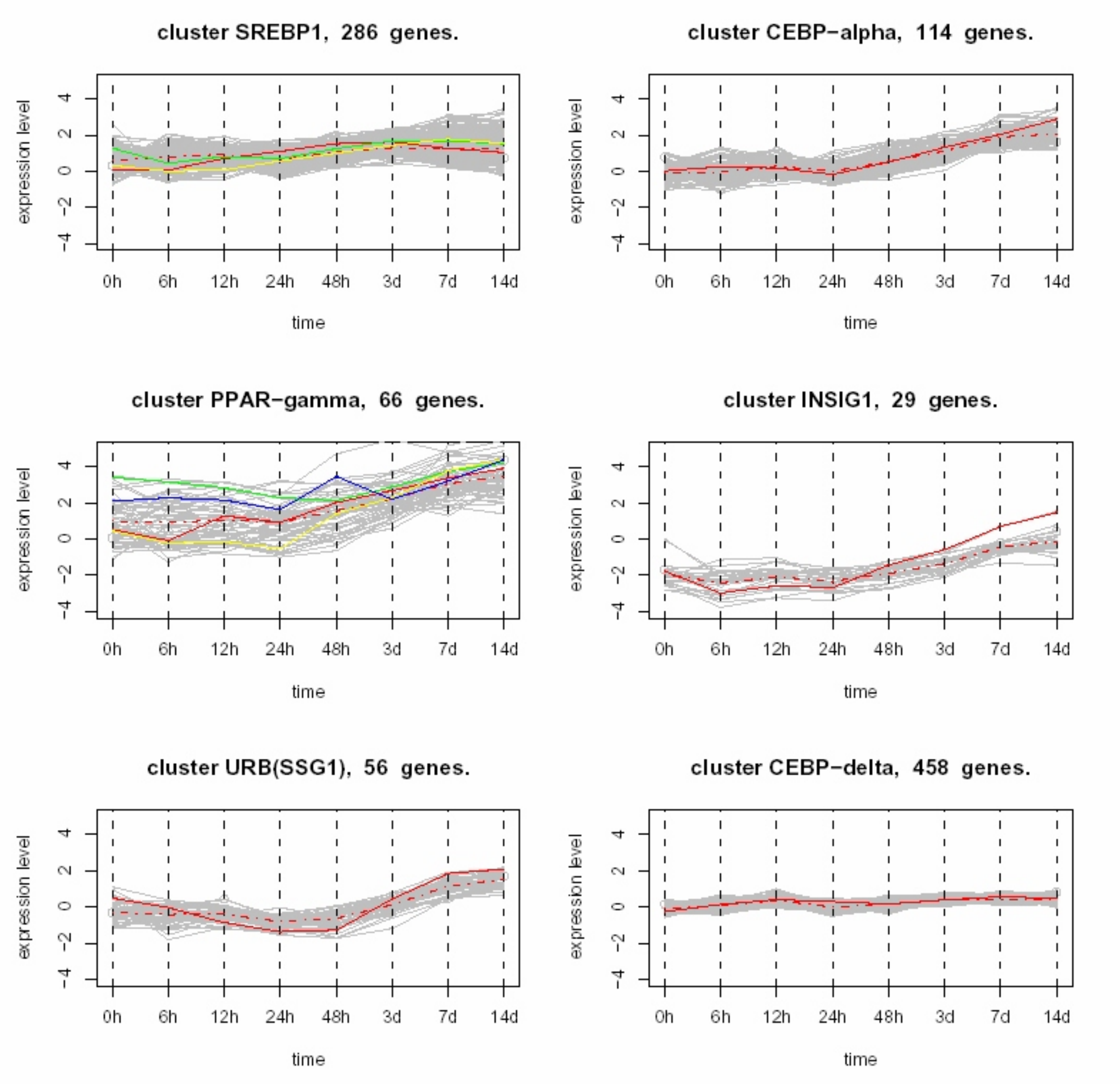


Figure 3: Clusters generated by the genes known to be inducing adipogenesis (red solid line), *i.e.* PPARγ, C/EBPα and SREBP1. Some other important genes are highlighted in each cluster. CEBPδ is expressed after two hours after induction, but this time point was not covered in this experiment, hence it exhibits a flat profile.

Description (TIGR)	0 h	6 h	12 h	24 h	48 h	3 d	7 d	14 d
glutamine synthetase	NaN	4.1	4.2	NaN	4.6	NaN	NaN	NaN
glutamine synthetase	NaN	4.4	3.9	3.7	4.6	4.7	4.5	4.4
Clusterin precursor(Apo-J)	NaN	2.6	NaN	NaN	NaN	NaN	2.3	NaN
IGFII precursor	3.3	NaN	NaN	2.3	NaN	2.9	4.4	5.2
Apolipoprotein E precursor	1.4	1.5	1.3	0.7	NaN	1.7	NaN	2.5
TNFR-related death receptor-6	2.5	NaN	1.6	2.4	NaN	2.6	2.3	1.9
ALHDII	NaN	1.5	2.6	3.5	2.9	3.4	3.3	3.2
LXRα	NaN	NaN	NaN	NaN	NaN	3.8	4.3	3.6
KLF5(BTEB2)	-0.4	2.1	2.5	2.8	3.6	2.6	1.0	NaN
C/EBPα	NaN	NaN	NaN	NaN	NaN	1.9	3.7	3.3

Table 2: Genes with missing values clustered with PPARγ. Very important is the presence of the Liver Receptor (LXRα), known to play an essenial role in lipid metabolism. The high expression that it exhibits towards the end of the adipogenesis process is significative of its role in adipocyte differentiation. However, it had not been possible to study it with other clustering algorithms.

4. Discussion and conclusions

Bayesian Statistics are specially suitable for inference from microarray data because, (1) a complex problem can easily be subdivided into managable units, which increases flexibility but also allows more complex modelling, (2) they allow incorporation of *a priori* knowledge from other experiments, which often facilitates experimental design and may allow for more economical decisions. The greater degree of flexibility that this approach offers had been already exploited in the discovery of gene and protein networks,in the discovery of differentially expressed genes measured using microarrays and also, but less often, for clustering of gene expression profiles [5]. However, the algorithm presented in this poster is the first one that considers all unknowns (*i.e.* centroid, cluster dispersion, number of clusters, missing values) as random variables which posterior distribution can be estimated from the data. The clustering algorithm succesfully identified groups of similar genes without the need of determining the number of clusters beforehand. The clusters found were biologically meaningful and genes previously discarded due to missing values were correctly assigned to clusters of genes of biological affinity. The only drawback is the computational cost, what will be overcome with the advance of this field.

References:

[1] Richardson S. and Green P.J. **On Bayesian Analysis of Mixtures with an Unknown Number of Components.** *J. R. Statist. Society* 59**:** 731--792 (1997)

[2] Gelman A., Carlin J.B., Stern H.S. and Rubin D.B. **Bayesian Data Analysis.** Chapman and Hall, London (1995)

[3] Spellman P.T., Sherlock G., Zhang M. Q., Vishwanath R. I., Anders K., Eisen M.B., Brown P. O., Botstein D. and Futcher B. **Comprehensive Identification of cell cycle-regulated genes of the yeast Saccharomices cerevisae by microarray hybridization.** *Molecular Biology of the Cell* **9**: 3273-3297 (1998)

[4] Hackl, H. Transcription profiling of adipogenesis. ww.genome.tugraz.at/adipocyte

[5] Medvedovic M. and Sivaganesan S. **Bayesian infinite mixture model based clustering of gene expression profiles**. *Bioinformatics* **18**(9): 1194-

Acknowledgements:

F.S-C was supported by a BBSRC studentship and by the EU Marie Curie Training Site "Genomics of Lipid Metabolism".
Z.T., H.H. ang G.S. want to thank the Austrian GEN-AU project BIN and the SFB-Biomembrane project.
O.W. acknowledges the support received by the U.K. Department for the Environment, Food and Rural Affairs (DEFRA) and in collaboration with the Veterinary Laboratories Agency (VLA) , Weybridge

Contact: F.Sanchez-Cabo@postgrad.umist.ac.uk Claus.Vogl@vu-wien.ac.at
f.sanchezcabo@tugraz.at