The aim of the analysis is to identify genes whose expression discriminates acute lymphoblastic leukaemia (ALL) patients from acute myeloid leukaemia (AML) patients. Following Dudoit et al. (2002), expression measures were truncated beyond the threshold of reliable detection at 100 and 16,000, and probe sets with intensities such that $max/min \leq 5$ and $max - min \leq 500$ were removed. This left 3,571 genes for the analysis. The expression readings were log-transformed, and each variable was rescaled by its range.

This data set was also analyzed by Kim et al. (2006) using a mixture model for cluster analysis. As in Kim et al. (2006), we assume that the nonsignificant variables are marginally independent of the significant ones. The hyper-parameters are taken to be $\delta = 3$, $h_1 = 10$, $h_0 = 100$, $\Omega_1 = 0.6^{-1} \cdot I_{|\gamma|}$, and $k_0 = 10^{-1}$. The hyper-parameters of the MRF prior, parameterized according to equation (13.8), were set to $d = -2.5$ and $g = 0.5$. Two samplers were run, with randomly selected starting models that had 10 and 2 included variables, respectively, with 150,000 iterations of which the first 50,000 were used as burn-in. Final inference was performed by pooling the two chains together.

A threshold of 0.85 on the marginal probability of inclusion selected 29 genes that were able to correctly classify 33 of the 34 samples. Lowering the threshold to 0.5 selected 72 significant variables that were able to correctly classify 30 of the 34 patients of the validation set. A heatmap of the 29 selected genes is given in Figure 13.3. As described in Golub et al. (1999), the validation set includes a much broader range of samples, including samples from peripheral blood rather than bone marrow, from childhood AML patients, and from different reference laboratories that used different sample preparation protocols. Their method made a correct prediction for 29 of the 34 samples, and the authors considered this result a “notable success” also because some observations came from one laboratory that used a very different protocol for sample preparation. The prediction results indicate that the selection of the top genes is not affected by the different protocol used in one laboratory or by other confounding effects.

Some of the selected genes were already known to be implicated with the differentiation or progression of leukemia cells. For example, Secchiero et al. (2005) found that COX2, selected with posterior probability of 0.93, increases tumorigenic potential by promoting resistance to apoptosis, and Peterson et al. (2007) found that the CD44 gene, selected with posterior probability of 0.98, is involved in the growth and maintenance of the AML blast/stem cells.

13.3 Models That Integrate Data From Different Platforms

In this second part of the chapter, we focus on Bayesian models that achieve an even greater type of integration, by incorporating into the modeling data from

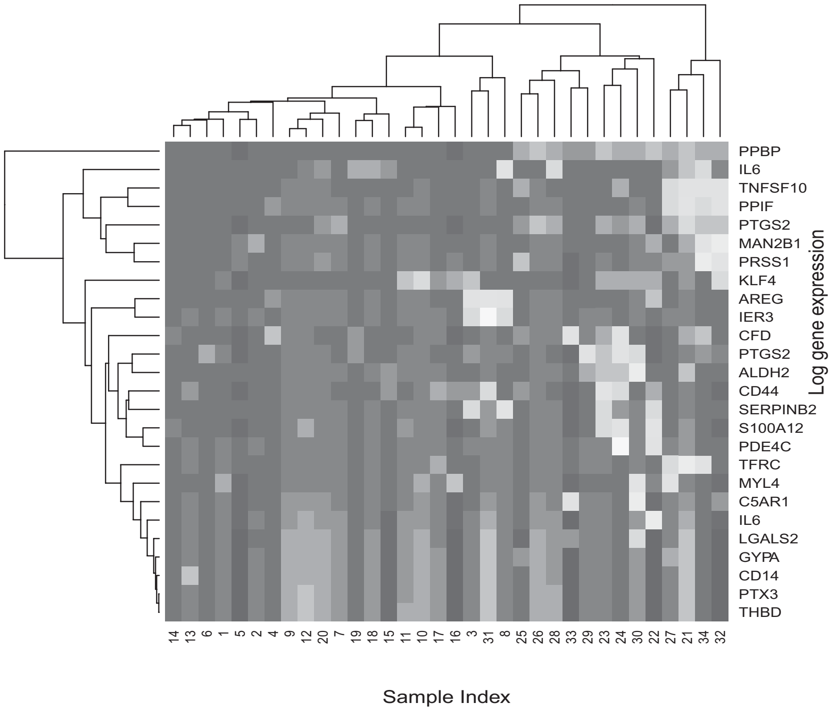

Figure 13.3 Golub data: heatmap of the 29 selected genes with a dendrogram of the clustering on the observations (on top) and a dendrogram of the clustering on the selected genes (on the left-hand side). (See color plate).

different platforms, together with priors that capture biological information on the variables. We look again at linear settings, via directed graphical models that integrate two kinds of expression data to infer a biological network.

13.3.1 Graphical Models to Infer Regulatory Networks

Graphical models focus on identifying latent graphical structures that encode conditional independencies. A graph is formed by nodes and arcs; nodes represent random variables, and the lack of arcs represents conditional independence. Hence graphical models provide a compact representation of joint probability distributions. Arcs can be undirected or directed. Undirected graphical models are also called Markov random field (MRF) models. Directed graphical models are also called Bayesian network (BN). Directed acyclic graph (DAG), in particular, do not allow for the presence of cycles. Conditional independencies in a DAG depend on the ordering of the variables. When the joint distribution is a multivariate normal, the model is called graphical Gaussian model (GGM). Nodes that are directly connected to node j and precede j in the ordering are called parents of j. In a Bayesian network, X_j is independent, given its parents, of the set of all the other variables in the graph, except its parents.

When the goal of the analysis is to recover the structure of a directed graphical model, with the ordering of the variables known a priori, it is possible to write the model in terms of a system of linear equations and therefore employ the spike and slab prior formulation (13.2) for the regression coefficients to achieve variable selection (Jones et al., 2005). Exploiting this idea, Stingo et al. (2010) put forward a graphical model formulation of a multivariate regression model that is used to infer a biological network of very high dimensionality, where microRNAs, small RNAs, are supposed to downregulate mRNAs, also called target genes. The main goal of the model is to understand which elements of the network are connected and which ones are not. In addition, specific biological characteristics/constraints need to be considered. Their model formulation includes constraints on the regression coefficients and selection priors that incorporate biological knowledge. The variable selection formulation they adopt overcomes the somehow rigid structure of the model in Brown et al. (1998), which does not allow to select different predictors for different responses. See also Monni and Tadesse (2009) for an approach based on partition models.

Briefly, Stingo et al. (2010) defined a DAG and imposed an ordering of the variables such that each target gene can be affected only by the miRNAs and that the miRNAs can affect only the targets. Let $\mathbf{Z} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_G, \mathbf{X}_1, \ldots, \mathbf{X}_M)$, with $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_G)$ the matrix representing the targets and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ the miRNAs. In their application, the data consist of $G = 1,297$ targets and $M = 23$ miRNAs observed on $N = 11$ units. Matrix $\mathbf{Z}$ is assumed to be a matrix-variate normal variable with zero mean and a variance matrix Ω for its generic row; that is, following the notation of Dawid (1981), $\mathbf{Z} - \mathbf{0} \sim \mathcal{N}(I_N, \Omega)$. In addition, the assumption that the target genes are independent conditionally on the miRNAs, that is, $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Y}_j | \mathbf{X}_1, \ldots, \mathbf{X}_M$, is made. Note that assumptions on the marginal distribution of $(\mathbf{X}_1, \ldots, \mathbf{X}_M)$ do not affect the regulatory network. In a Bayesian network framework, these assumptions imply an ordering of the nodes and, consequently, a likelihood factorization of the type

$$f(\mathbf{Z}) = \prod_{g=1}^{G} f(\mathbf{Y}_g | \mathbf{X}) \prod_{m=1}^{M} f(\mathbf{X}_m), \tag{13.13}$$

where $f(\mathbf{Y}_g|\mathbf{X}) \sim N(\mathbf{X}\beta_g, \sigma_g I_N)$ and $f(\mathbf{X}_m) \sim N(0, \sigma_m I_N)$, with $\beta_g = \Omega_{\mathbf{XX}}^{-1}\Omega_{\mathbf{XY}_g}$ and $\sigma_g = \omega_{gg} - \Omega_{\mathbf{XY}_g}^T \Omega_{\mathbf{XX}}^{-1}\Omega_{\mathbf{XY}_g}$. Here ω_{gg} indicates the g-th diagonal element of Ω, and $\Omega_{\mathbf{XX}}$, $\Omega_{\mathbf{XY}}$ are the appropriate blocks of the covariance matrix. For $m = 1, \ldots, M$, we have $\sigma_m = \omega_{mm}$. This graphical model formulation is equivalent to a system of G linear regression models.

Incorporating Biological Information

Knowledge about the fact that miRNAs downregulate gene expression can be incorporated into the model by specifying negative regression coefficients via the prior choice, that is $(\tilde{\beta}_{gm}|\sigma_g) \sim Ga(1, c\,\sigma_g)$ and $\sigma_g^{-1} \sim Ga((\delta + M)/2, d/2)$, with $\beta_{gm} = -\tilde{\beta}_{gm}$. Furthermore, the underlying regulatory network can be completely encoded introducing a $(G \times M)$ association matrix $\mathbf{R}$ with elements $r_{gm} = 1$ if the m-th miRNA is included in the regression of the g-th target and $r_{gm} = 0$ otherwise. The regression coefficient parameters are then stochastically independent, given the regulatory network $\mathbf{R}$, and have the following mixture prior distribution

$$\pi(\tilde{\beta}_{gm}|\sigma_g, r_{gm}) = r_{gm} Ga(1, c\,\sigma_g) + (1 - r_{gm})\delta_0(\tilde{\beta}_{gm}). \qquad (13.14)$$

Prior distributions for $\mathbf{R}$ can be specified by taking into account biological information encoded by sequence/structure databases available on the internet. Scores of possible gene-miRNA pair associations that come from these sources can be integrated into the model by defining the prior probability of selecting the edge between a gene g and an miRNA m as

$$P(r_{gm} = 1|\tau) = \frac{\exp[\eta + \tau_1 s_{gm}^1 + \tau_2 s_{gm}^2 + \cdots + \tau_J s_{gm}^J]}{1 + \exp[\eta + \tau_1 s_{gm}^1 + \tau_2 s_{gm}^2 + \cdots + \tau_J s_{gm}^J]},$$

with $\tau = (\tau_1, \ldots, \tau_J)$ and where the s_{gm}^j's, with $j = 1, \ldots, J$, denote the J available scores.

For posterior inference, the regression coefficients can be integrated out, reducing the computational complexity of the MCMC algorithm to the sampling of the models space, $\mathbf{R}$, the data integration parameters, τ_j, and the variances, σ_g. See Stingo et al. (2010) for details.

Application to Gene-miRNA Data

Stingo et al. (2010) considered experimental data from a study on a very well-known developmental toxicant causing neural tube defects, hyperthermia. The analyzed data consist of 23 mouse miRNAs and a total of 1,297 potential target genes. There are $N = 11$ i.i.d. observations under the control status and

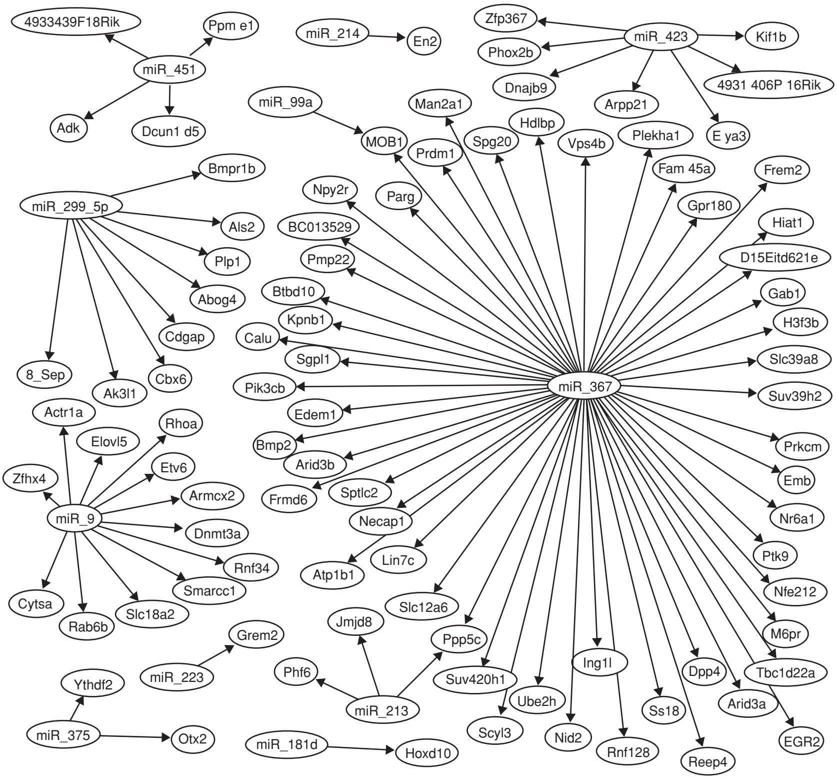

Figure 13.4 Gene-miRNA inferred network for the hyperthermia group using a threshold of 0.8 on the posterior probability.

$N = 12$ i.i.d. observations under hyperthermia. To goal of the analysis is to infer their regulatory network under two different treatment conditions. Four of the most widely used algorithms, miRanda, TargetScan, PITA, and PicTar, are used in the analysis to retrieve the sequence and structure information for target prediction. Here s_{gm}^{j}'s, with $j = 1, \dots, 5$, denote the PicTar, miRanda, aggregate TargetScan, total TargetScan, and PITA scores, respectively.

Important pairs of target genes and miRNAs can be selected as those corresponding to arrows with highest posterior probabilities. For example, by exploring the regulatory network as a function of this posterior probability of the arrows, the authors found that a posterior probability cutoff of 0.8 selected 88 arrows between 88 target genes and 7 miRNAs, for the hyperthermia group, see Figure 13.4. A close look at the pairs of target genes and miRNAs with high posterior probabilities reveals that some of the regulatory relationships

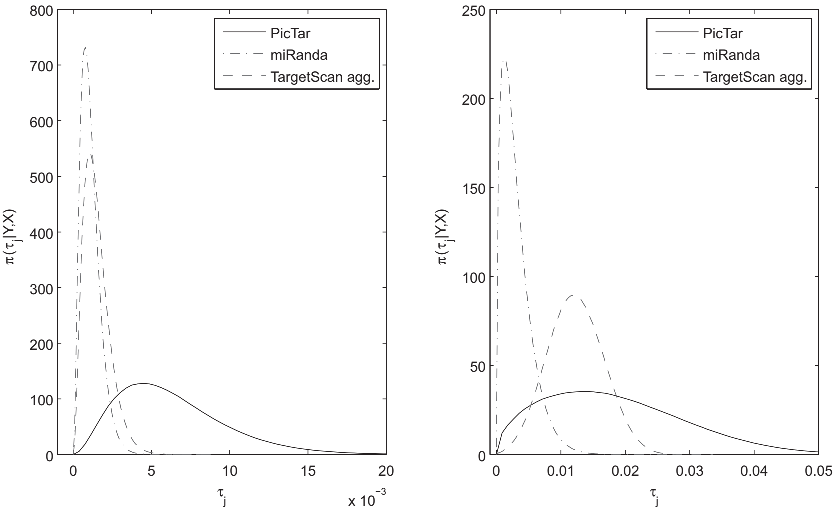

Figure 13.5 Gene-miRNA network inference. Kernel density estimates of the τ parameters using the time-independent model for the control group (left panel) and the time-dependent model for the hyperthermia group (right panel).

seem plausible and warrant future investigation. For example, links between miR-367 and target Egr2 and Mob1, selected with posterior probability of 0.97 and 1, are particularly interesting. A total of 108 of the gene targets identified were associated with miR-367, a pluripotency-specific marker in human and mouse ES cells (Li et al., 2009), whereas 27 of the gene targets were associated with miR-423, which has previously been shown to be expressed in the adult and/or developing brain (Zhang and Pan, 2009). Expression of both miR-367 and miR423 decreased over time in control and hyperthermia-treated embryos, which is consistent for a marker of pluripotency in a differentiating embryo. Although decreasing, the expression levels of these miRNAs were higher after hyperthermia exposure when compared with controls, which may indicate a delay in the differentiation program.

The authors also evaluated the inference on $\tau_1, \ldots, \tau_5$. Their results suggested that the information extracted from PicTar is the most influential on the posterior inference. MiRanda and Target Scan aggregate also contribute somehow to the selection process, whereas TargetScan total and PITA do not affect the posterior inference. Figure 13.5 shows the posterior inference for the three most influential algorithms when $\eta = -3$. The other two algorithms resulted in posterior densities that were very concentrated around zero (plot not shown). For details on the time-dependent coefficients model, see Stingo et al. (2010).