settings, the inclusion of the latent variable indicators into the model is done via the likelihood rather than a prior model on regression coefficients.

In the 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 different platforms, together with priors that capture biological information on the variables. We look in particular at statistical procedures that aim at inferring biological networks of high dimensionality, where microRNAs, small RNAs, are supposed to down-regulate mRNAs, also called targets, and where sequence and structure information is integrated into the model via the prior formulation.

All modeling settings we describe in this chapter exploit variable selection techniques and utilize prior constructions that cleverly incorporate biological knowledge about structural dependencies among the variables. The rest of the chapter is organized as follows. In Section 13.2, we describe Bayesian models that integrate external information in the analysis of gene expression data. In Section 13.3, we address models that integrate data from different platforms.

13.2 Models That Integrate External Information With Experimental Data

The flexibility of the prior models for variable selection and the fact that the inferential methods can handle the"large p–small n" paradigm have made these techniques particularly relevant for the analysis of genomic studies, where high-throughput technologies allow thousands of variables to be measured on individual samples. Here we focus on model developments that utilize prior constructions that incorporate external biological information, as typically available via online databases, into the chosen model for the experimental data.

13.2.1 Linear Models for Pathway and Gene Selection

Stingo et al. (2011) considered the problem of finding genes that relate to a response variable. In their approach the authors take into account that recent interest in biology has moved from the analysis of single genes to the analysis of known groups of genes, called pathways. Many databases exist nowadays where information on gene-pathway memberships and on gene–gene networks can be retrieved.

Say we have available information on K pathways for a total of p genes. For example, we can interrogate the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to obtain pathway memberships and the gene-to-gene interaction network. Let S be the $K \times p$ matrix indicating membership of genes in

pathways, with element $s_{kj} = 1$ if gene j belongs to pathway k, and $s_{kj} = 0$ otherwise. Let R be the $p \times p$ matrix describing relationships between genes, with element $r_{lj} = 1$ if genes l and j have a direct link in the gene network, and $r_{lj} = 0$ otherwise. Stingo et al. (2011) studied the association between the response variable and the pathways by defining measures of the "pathway expression" summarizing the group behavior of included genes within pathways. In their proposed model formulation, these pathway "scores" are defined via partial least squares (PLS) regression of Y on selected subsets of genes from the given pathway. The first PLS components are then used as predictors in a linear regression model. A latent binary vector θ is introduced for pathway selection. In addition, gene selection is performed by introducing a $p \times 1$ binary vector γ, where $\gamma_j = 1$ if gene j is selected, and $\gamma_j = 0$ otherwise. Their linear regression model is defined as

$$Y = \mathbf{1}\alpha + \sum_{k=1}^{K_\theta} T_{k(\gamma)}\beta_{k(\gamma)} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}), \tag{13.1}$$

where $T_{k(\gamma)}$ corresponds to the first latent PLS component generated based on the expression levels of selected genes belonging to pathway k, that is, using the X_j's corresponding to $s_{kj} = 1$ and $\gamma_j = 1$, and where $K_\theta = \sum_{k=1}^{K} \theta_k$ is the number of selected pathways.

In the Bayesian paradigm, variable selection can be achieved by imposing mixture priors on the regression coefficients of model (13.1) as

$$\beta_k|\theta_k, \sigma^2 \sim (1 - \theta_k)\delta_0(\beta_k) + \theta_k N(0, h_k\sigma^2), \tag{13.2}$$

where $\delta_0(\cdot)$ is the Dirac function at zero and the h_k's hyperparameters to be chosen. With this prior, if $\theta_k = 0$, then β_k is set to 0, whereas if $\theta_k = 1$, a nonzero estimate of β_k corresponds to an important predictor (pathway). Independent Bernoulli priors can be specified for the parameters θ_k. In addition, conjugate priors can be imposed on α and σ^2,

$$\alpha|\sigma^2 \sim N(\alpha_0, h_0\sigma^2) \tag{13.3}$$

$$\sigma^2 \sim IG(\nu/2, \lambda/2) \tag{13.4}$$

with α_0, h_0, ν, and λ to be chosen. Mixture priors of type (13.2) for univariate linear regression models were made popular by George and McCulloch (1993, 1997), Clyde et al. (1996), Geweke (1996), Smith and Kohn (1996), and Raftery et al. (1997). Brown et al. (1998, 2002) extended the construction to multivariate linear regression models with q response variables. Reviews of special features of the selection priors and on computational aspects can be found in Chipman et al. (2001) and Clyde and George (2004).

Common choices of the hyperparameters h_k's in the prior model (13.2) assume that the β_k's are a priori independent given θ, for example, by choosing $h_k = c$ for every k. Brown et al. (1998) investigated the case of h_k chosen to be proportional to the k-th diagonal element of $(\mathbf{T}'\mathbf{T})^{-1}$, whereas Smith and Kohn (1996) proposed the use of a full Zellner's g-prior. This type of priors has an intuitive interpretation as they use the design matrix of the current experiment. Liang et al. (2008) investigated formulations that use a fully Bayesian approach by imposing mixtures of g-priors on c.

Priors that Incorporate Biological Information

Model (13.1) is completed by specifying a prior on γ. Here, it makes sense to use priors that encode the gene–gene network information available to us via the KEGG database. These relations can be modeled using a Markov random field (MRF), where variables are represented by nodes and relations between them by edges. One possible parametrization, used in Stingo et al. (2011), of the MRF is represented by the following probabilities:

$$p(\gamma_j|\mu, \eta, \gamma_l, l \in N_j) = \frac{\exp(\gamma_j F(\gamma_j))}{1 + \exp(F(\gamma_j))}, \tag{13.5}$$

where $F(\gamma_j) = \mu + \eta \sum_{l \in N_j}(2\gamma_l - 1)$ and N_j is the set of direct neighbors of variable j in the MRF. The global distribution on the MRF is given by

$$p(\gamma|\mu, \eta) \propto \exp(\mu\ n_1 - \eta\ n_{01}), \tag{13.6}$$

where n_1 is the number of selected variables and n_{01} is the number of edges linking nodes with different values of γ_j (i.e., edges linking included and nonincluded nodes),

$$n_1 = \sum_{j=1}^{p} \gamma_j, \qquad n_{01} = \frac{1}{2} \sum_{l=1}^{p} \left[\sum_{j=1}^{p} r_{lj} - \left| \sum_{j=1}^{p} r_{lj}(1 - \gamma_l) - \sum_{j=1}^{p} r_{lj}\gamma_j \right| \right].$$

The parameter μ controls the sparsity of the model, whereas higher values of η result in neighboring variables taking on the same γ_j value. If a variable does not have any neighbor, its prior distribution reduces to an independent Bernoulli with parameter $\exp(\mu)/[1 + \exp(\mu)]$, which is a logistic transformation of μ. Telesca et al. (2012) also used these MRF priors in a model for the identification of differentially expressed genes that takes into account the dependence structure among genes. Another parametrization of the MRF, recently used by Li and Zhang (2010), corresponds to the following distribution for γ

$$p(\gamma|D, G) \propto \exp(D'\gamma + \gamma' G \gamma), \tag{13.7}$$

with $D = d1_p$, 1_p the unit vector of dimension p, and G a matrix with elements $\{g_{ij}\}$ usually set to some constants. Although d plays the same role as μ in (13.6), G and η affect differently the probability of selection of a variable. This is evident from the conditional probability

$$P(\gamma_j|d, g, \gamma_l, l \in N_j) = \frac{\exp(\gamma_j(d + g\sum_{l\in N_j}\gamma_l))}{1 + \exp(d + g\sum_{l\in N_j}\gamma_l)}, \tag{13.8}$$

which can only increase as a function of the number of selected neighboring genes. In contrast, with the parametrization in (13.5), the prior probability of selection for a variable does not decrease if none of the neighbors are selected. Although the parametrization is somewhat arbitrary, some care is needed in deciding whether to put a prior distribution on G. Allowing G to vary can lead to a phase transition problem; that is, the expected number of variables equal to 1 can increase massively for small increments of G.

Stingo et al. (2011) imposed some constraints on the joint distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ to ensure both interpretability and identifiability of the model. Both prior specification and MCMC moves are specified to avoid the creation of empty pathways, the creation of orphan genes, and the selection of identical subsets of genes by different pathways, a situation that generates identical values $T_{k(\gamma)}$ and $T_{k'(\gamma)}$ to be included in the model.

Posterior Inference

Methods for posterior inference benefit from the conjugate prior setting on the model parameters. Efficient MCMC can be designed to sample only the selection parameters γ and θ, in addition to the MRF parameters if parametrization (13.6) is used. When a large number of predictors makes the full exploration of the model space unfeasible, MCMC methods can be used as stochastic searches to quickly and efficiently explore the posterior distribution looking for "good" models, that is, models with high posterior probability; see George and McCulloch (1997). The MCMC steps to fit the model of Stingo et al. (2011) consist of (1) sampling the pathway and gene selection indicators from $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\text{rest})$; and (2) sampling the MRF parameter from $p(\eta|\text{rest})$. In the first step of the algorithm, the parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ are updated using a Metropolis-Hastings algorithm in a two-stage sampling scheme. The pathway-gene relationships are used to structure the moves and account for interpretability and identifiability constraints. In step 2, techniques that deal with unknown normalized constants need to be implemented. In addition, although we have focused on the regression setting, where the response is continuous, the models and priors for variable selection described earlier can be easily applied to linear settings with other types of response, for example, probit models for binary

and multinomial responses (Sha et al., 2003, 2004) and accelerated failure time models for survival responses (Sha et al., 2006). Such cases use data augmentation approaches and require the sampling of additional latent variables in the MCMC procedures. Stingo et al. (2011), in particular, used this approach to analyze which pathways and genes can predict the time to distant metastasis for breast cancer patients.

The MCMC procedure results in a list of visited models, with included pathways indexed by $\boldsymbol{\theta}$ and selected genes indexed by $\boldsymbol{\gamma}$, and corresponding relative posterior probabilities. Pathway selection can be based on the marginal posterior probabilities $p(\theta_k|\boldsymbol{T}, Y)$. A simple strategy is to compute Monte-Carlo estimates by counting the number of appearances of each pathway across the visited models. Relevant pathways are identified as those with largest marginal posterior probabilities. Then relevant genes from these pathways are identified based on their marginal posterior probabilities conditional on the inclusion of a pathway of interest, $p(\gamma_j|\boldsymbol{T}, Y, I\{\theta_k s_{kj} = 1\})$.

Application to Microarray Data

We consider the breast cancer microarray data of van't Veer et al. (2002).[1] We focus on the 76 sporadic lymph node–negative patients, 33 of whom developed distant metastasis within 5 years, whereas the remaining 43 did not; the latter are viewed as censored cases. We randomly split the patients into a training set of 38 samples and a test set of the same size. The goal is to identify a subset of pathways and genes that can predict time to distant metastasis for breast cancer patients.

The gene network and pathway information were obtained from the KEGG database. A total of 196 pathways and 3,592 corresponding probes were included in the analysis. There is a many-to-many correspondence between pathways and genes; that is, each pathway contains multiple genes, and most genes are associated with several pathways. The selected pathways are clearly indicated in the marginal posterior probability plots displayed in Figure 13.1. A subset of the selected pathways along with islands, that is, sets of genes directly connected in the gene network, and singletons are displayed in Figure 13.2. Several of the identified pathways are known to be involved in tumor formation and progression. For instance, the mitogen-activated protein kinase (MAPK) signaling pathway, which is involved in various cellular functions, including cell proliferation, differentiation, and migration, has been implicated in breast cancer metastasis (Lee et al., 2007; Keyse, 2008). The KEGG pathway in cancers was also selected with high posterior probability. Other interesting

[1] Available at www.rii.com/publications/2002/vantveer.htm.

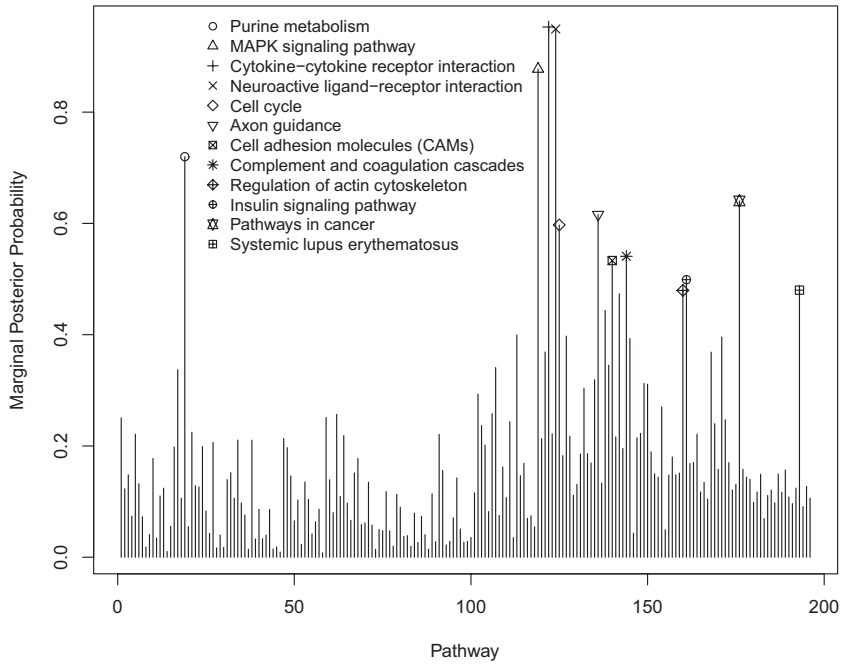

Figure 13.1 Breast cancer microarray data: marginal posterior probabilities for pathway selection, $p(\theta_k|\boldsymbol{T}, Y)$. The 12 pathways with largest probabilities are marked with symbols.

pathways are the insulin signaling pathway, which has been linked to the development, progression, and outcome of breast cancer, and purine metabolism, which is involved in nucleotide biosynthesis and affects cell cycle activity of tumor cells.

In addition, several genes with known association to breast cancer were also selected. One of these is protein kinase C alpha (PKCA), which belongs to the MAPK pathway and the KEGG pathways in cancer. PKCA has been reported to play roles in many different cellular processes, including cell functions associated with breast cancer progression. It has been shown to be overexpressed in some antiestrogen-resistant breast cancer cell lines and to be involved in the growth of tamoxifen-resistant human breast cancer cells (Frankel et al., 2007). Patients with PKCA-positive tumors have also been shown to have worse survival than patients with PKCA-negative tumors, independently of other factors. Prostaglandin-endoperoxide synthase-2 (PTGS2, also known as cyclooxygenase-2, or COX2) has also been related to breast cancer. Denkert et al. (2004) observed COX2 overexpression in breast cancer and strong

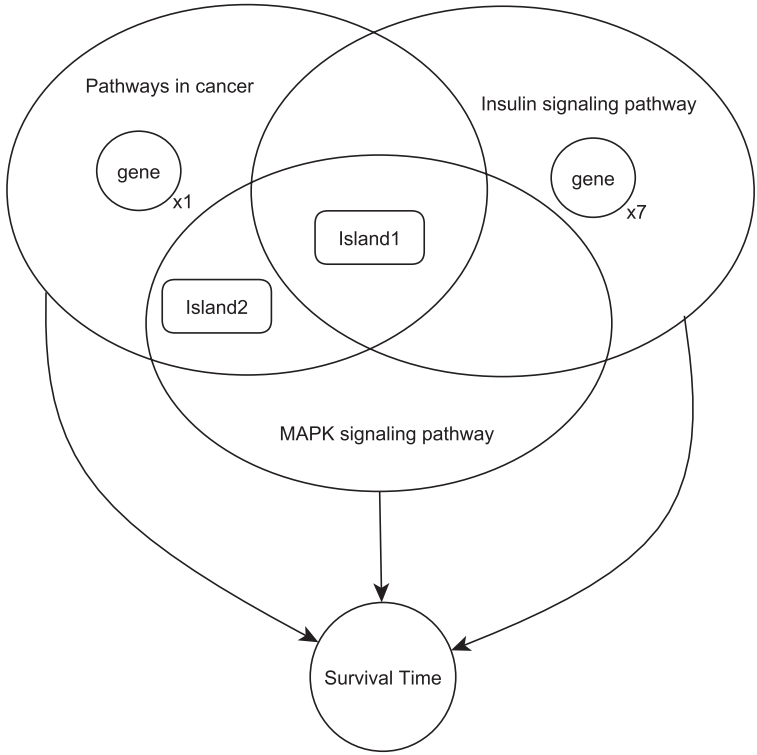

Figure 13.2 Breast cancer microarray data: graphical representation of a subset of selected pathways with islands and singletons.

association with indicators of poor prognosis, such as lymph node metastasis, poor differentiation, and large tumor size. This was further confirmed by Gupta et al. (2007), who showed that the expression of COX2 in human breast cancer cells facilitates the assembly of new tumor blood vessels, the release of tumor cells into the circulation, and the breaching of lung capillaries by circulating tumor cells to seed pulmonary metastasis. This is an important finding, as the majority of breast cancer deaths result from metastases rather than from direct effects of the primary tumor itself.

The model also showed good predictive performance. Sha et al. (2006) already analyzed these data using an AFT model with 3,839 probes as predictors and obtained a predictive MSE of 1.9317 using the 11 probe sets with highest marginal probabilities. The model incorporating pathway information achieved a predictive MSE of 1.4497 on the validation set, using 12 selected pathways and 41 probe sets with highest posterior probabilities.

13.2.2 Biomarker Selection in Mixture Models

We now address variable selection in a different modeling context, that is, mixture models for pattern recognition. We treat in particular the unsupervised framework, known in the statistical literature as clustering, and then describe an adaptation to the simpler supervised framework, known as discriminant analysis. For both model formulations, we borrow ideas from the linear settings treated in the previous section. For example, a latent binary vector γ is introduced for variable selection, and stochastic search MCMC techniques are used to explore the space of variable subsets. However, building a variable selection mechanism into mixture models is more challenging than the linear settings. In clustering, for example, there is no observed response to guide the selection, and the elements of the matrix $\mathbf{X}$ are viewed as random variables. The inclusion of the latent indicators into the models, therefore, cannot be done like in the linear modeling context, where γ is used to induce mixture priors on regression coefficients.

An approach to variable selection for model-based clustering was put forward by Tadesse et al. (2005), who formulated the clustering in terms of a finite mixture of Gaussian distributions with an unknown number of components and then introduced latent variables to identify discriminating variables. The authors used a reversible jump MCMC technique to allow for the creation and deletion of clusters. A similar model was considered by Raftery and Dean (2006). Kim et al. (2006) proposed an alternative modeling approach that uses infinite mixture models via Dirichlet process priors. Hoff (2006) adopted a mixture of Gaussian distributions where different clusters are identified by mean shifts and Bayes factors are computed to identify discriminating variables. This method allows separate subsets of variables to discriminate different groups of observations.

In the finite mixture model formulation of Tadesse et al. (2005), the data are viewed as coming from a mixture of distributions

$$f(\mathbf{x}_i|\mathbf{w}, \phi) = \sum_{k=1}^{K} w_k f(\mathbf{x}_i|\phi_k), \tag{13.9}$$

where $f(\mathbf{x}_i|\phi_k)$ is the density of sample $\mathbf{x}_i$ from group k and $\mathbf{w} = (w_1, \ldots, w_K)^T$ are the cluster weights ($\sum_k w_k = 1, w_k \geq 0$); see McLachlan and Basford (1988). Here K is assumed finite but unknown. Latent variables $\mathbf{c} = (c_1, \ldots, c_n)^T$, with $c_i = k$ if the i-th sample comes from group k, are introduced to identify the cluster from which each observation is drawn. The sample allocations, c_i, are assumed to be independently and identically distributed with probability mass function $p(c_i = k) = w_k$. We assume that the mixture

distributions are multivariate normal with component parameters $\phi_k = (\mu_k, \Sigma_k)$. Thus, for sample i, we have

$$\mathbf{x}_i | c_i = k, \mathbf{w}, \phi \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{13.10}$$

For variable selection, a latent binary vector γ is used to identify the discriminating variables. More specifically, variables indexed by a $\gamma_j = 1$, denoted $\mathbf{X}_{(\gamma)}$, define the mixture distribution, whereas variables indexed by $\gamma_j = 0$, $\mathbf{X}_{(\gamma^c)}$, favor one multivariate normal distribution across all samples. The distribution of sample i is then given by

$$\begin{aligned} \mathbf{x}_{i(\gamma)} | c_i = k, \mathbf{w}, \phi, \gamma &\sim \mathcal{N}(\mu_{k(\gamma)}, \Sigma_{k(\gamma)}) \\ \mathbf{x}_{i(\gamma^c)} | \psi, \gamma &\sim \mathcal{N}(\eta_{(\gamma^c)}, \Omega_{(\gamma^c)}), \end{aligned} \tag{13.11}$$

where $\psi = (\eta, \Omega)$.

Priors on γ can be specified similarly to what was discussed for the linear settings, via MRF distributions. For the vector of component weights, a symmetric Dirichlet prior can be specified. For the unknown number of components, K, a truncated Poisson or a discrete Uniform prior on $[1, \ldots, K_{\max}]$, where $K_{\max}$ is chosen arbitrarily large, are suitable choices. An efficient sampler can be implemented by working with a marginalized likelihood where the model parameters are integrated out. The integration is facilitated by taking conjugate Normal-Wishart priors on both ϕ and ψ. Some care is needed in the choice of the hyper-parameters. In particular, the variance parameters need to be specified within the range of variability of the data. The MCMC procedure is described in Tadesse et al. (2005) and requires a sampler that jumps between different dimensional spaces, generalizing the reversible jump approach of Richardson and Green (1997).

Discriminant Analysis

We now show an adaptation of the method to the simpler supervised setting, where, in addition to the observed vectors $\mathbf{x}_i$'s, the number of groups G and the classification labels c_i's are also available and where the aim is to derive a classification rule that will assign further cases to their correct groups. When the distribution of $\mathbf{X}$ conditional on the group membership is assumed normal, then this statistical methodology is known as discriminant analysis.

In discriminant analysis, given the selected variables, the predictive distribution of a new observation $\mathbf{x}^f$ is used to classify every new sample into one of the possible G groups. This distribution is a multivariate T-student, see Brown (1993) among others. The probability that a future observation, given

the observed data, belongs to the group k is then given by

$$\pi_k(c^f|\mathbf{X}) = p(c^f = k|\mathbf{x}^f, \mathbf{X}), \tag{13.12}$$

where c^f is the group indicator of $\mathbf{x}^f$. By estimating the prior probability that one observation comes from group k as $\hat{\pi}_k = n_k/n$, the previous distribution can be written in closed form as

$$\pi_k(c^f|\mathbf{X}) = \frac{p_k(\mathbf{x}^f)\hat{\pi}_k}{\sum_{i=1}^{G} p_i(\mathbf{x}^f)\hat{\pi}_i},$$

where $p_k(\mathbf{x}^f)$ indicates the predictive T-student distribution. A new observations is then assigned to the group with the highest posterior probability.

As in the clustering setting, we introduce a latent binary vector γ to perform the selection. As done by Raftery and Dean (2006), extending the approach of Tadesse et al. (2005) to avoid any independence assumptions, the following likelihood can be used to separate the discriminant variables from the noisy ones as

$$L(X, c; \cdot) = \prod_{i=1}^{n} p(\mathbf{x}_{i(\gamma^c)}|\mathbf{x}_{i(\gamma)}) \prod_{k=1}^{G} w_k^{n_k} \prod_{j=1}^{n_k} p_k(\mathbf{x}_{j(\gamma)}).$$

The first factor of the likelihood refers to the non-important variables, whereas the second is formed by variables able to classify observations into the correct groups. Under the normality assumption, the likelihood becomes

$$\prod_{i=1}^{n} N_{|\gamma^c|}(\mathbf{x}_{i(\gamma^c)} - \beta\mathbf{x}_{i(\gamma)}; \eta_{(\gamma^c)}, \Sigma_{(\gamma^c)}) \prod_{k=1}^{G} w_k^{n_k} \prod_{j=1}^{n_k} N_{|\gamma|}(\mathbf{x}_{j(\gamma)}; \mu_{k(\gamma)}, \Sigma_{k(\gamma)}).$$

As in the unsupervised case, conjugate Normal-Wishart priors can be specified on the model parameters. Again, the corresponding MCMC algorithm benefits of this parametrization, because it is possible to integrate out means and variances and design Metropolis steps that depend only on the selected and proposed variables.

An Application to Microarray Data

We show an application from Stingo and Vannucci (2011) on the widely used leukemia data of Golub et al. (1999) that comprise a training set of 38 patients and a validation set of 34 patients. An MRF prior on γ is used to capture knowledge on the gene network structure, as extracted from KEGG. Note that some of the genes do not have neighbors. The training set consists of bone marrow samples obtained from acute leukemia patients, whereas the validation set consists of 24 bone marrow samples and 10 peripheral blood samples.